

eman ta zabal zazu



Universidad
del País Vasco

Euskal Herriko
Unibertsitatea

NIFPTML: NETWORK INVARIANT, INFORMATION FUSION,
PERTURBATION THEORY, MACHINE LEARNING STUDY OF
DUAL ANTIBACTERIAL DRUGS-NANOPARTICLES (DADNP)
SYSTEMS - METABOLIC NETWORKS INTERACTION

MEMORIA PRESENTADA POR
Karel Diéguez Santana
PARA OPTAR AL GRADO DE DOCTOR

Leioa, 2022

© 2022 Karel Diéguez Santana

eman ta zabal zazu



Universidad
del País Vasco

Euskal Herriko
Unibertsitatea

NIFPTML: NETWORK INVARIANT, INFORMATION FUSION,
PERTURBATION THEORY, MACHINE LEARNING STUDY OF
DUAL ANTIBACTERIAL DRUGS-NANOPARTICLES (DADNP)
SYSTEMS - METABOLIC NETWORKS INTERACTION

to obtain the degree of PhD in Synthetic and Industrial Chemistry at the University of
Basque Country (UPV/EHU)

Karel Diéguez Santana

(c)2022 KAREL DIÉGUEZ SANTANA

Thesis Supervisor:

Dr. Humberto González Díaz,

- Department of Organic and Inorganic Chemistry, University of Basque Country UPV/EHU, 48940 Leioa, Spain.
- BIOFISIKA, Basque Center for Biophysics CSIC-UPVEH, 48940 Leioa, Spain.
- IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Biscay, Spain.

University tutor:

Prof. Dr. María Nuria Sotomayor Anduiza

- Department of Organic and Inorganic Chemistry, Faculty of Science and Technology, University of the Basque Country (UPV/EHU). Bilbao (Spain).

ACKNOWLEDGMENTS

There are many people without whom this work would not have been possible. First, I would like to thank my director, Humberto González Díaz. This work could not have been conceived without his guidance and support. I would also like to thank my friends Gerardo Casañola Martín and Oscar Miguel Borroto Rivera for getting me involved in this training process. I would like to thank Sonia Arrasate for her support in the Basque sections and the other members of the Synthetic and Industrial Chemistry Ph.D. Program for the knowledge they contributed to my training.

Finally, I would also like to dedicate this work to my family. It would have been impossible without their strength and support at every moment of this stage, to the support and energy of my daughters (Laura and Liz), my wife (Liliana), and my parents (Victor A. and Teresita). Also, to all those friends who have been watching my progress and encouraging me to complete this process, this work is especially for all of them (family and friends) for trusting me all this time, for their unconditional love and support.

Thank you for all the good moments that have contributed so much to this thesis and to me as a person.

Karel Diéguez Santana

March 2022.

ABSTRACT

The combination of classical computational approaches with machine learning techniques (MLT) is gaining traction in academia and industry. MLTs are used in chemoinformatics processes to predict the activity of an unknown drug and thus discover new potential antibacterial drugs (ADs). This thesis focuses on the design of a methodology based on the application of perturbation theory (PT) combined with machine learning (ML) methods and information fusion to predict the antibacterial activity of drugs at the design stage from preclinical assay information, chemical structures, nanoparticles (NP), and variations of metabolic networks (MN) of multiple microorganisms. First, an exploration of the state of the art on bacterial resistance, targets and mechanisms of action, databases useful for computational modeling, MLT, and performance evaluation algorithms applied in the field of antibacterial drugs was performed. Subsequently, the study was continued with the creation of a computational model to study the connectivity (structure) of a metabolite in the MN reactions of a query organism. Once the main nodes (metabolites) were identified, MN of > 40 bacterial species were quantitatively related to chemical and preclinical data from the ChEMBL database. Next, a model was developed for the prediction of the biological activities of AD functionalized with NP systems. Finally, an analysis and mapping of DADNP (AD + NP) systems against MN of pathogenic bacterial species was performed using Network Invariance Information Fusion, Machine Learning with Perturbation Theory (NIFPTML = N + IF + PT + ML) as an application of AI/ML methods in the search for AD that cope with the emergence of multidrug-resistant strains. The additive NIFPMTL strategy may become a useful tool to aid in the design and discovery of new DADNP systems.

LABURPENA

Mundu akademikoan eta industrian ohiko konputazio ikuspegiak eta Ikasketa Automatikoaren Teknikak (IAT, ingelesez MLT)) konbinazioak gero eta kide gehiago irabazten ari du. IATak kimioinformatikako prozesuetan erabiltzen dira farmako ezezagunen aktibitatea aurreratzeko eta bakterio-kontrako farmako berriak (BF, ingelesez AD) izan daitezkeenak aurkitzeko. Doktorego-tesi honetan ikasketa automatikoaren metodoekin (IA, ingelesez ML) eta informazio fusioarekin konbinatuta dagoen perturbazio teorian (PT) datuan metodologia garatu da. Hala, diseinu fasean dauden farmakoaren bakterio-aurkako aktibitatea aurreran daiteke entsegu aurre klinikotatik, egitura kimikoetatik, nano partikuletatik eta hainbat mikroorganismoen erreakzio metaboliko sareen aldaketetatik abiatuz.

Lehenengo, bakterioen erresistentzia, ituak eta mekanismoak, konputazio-ereduetarako datu sorta egokiak, ikasketa automatikoaren teknikak eta etekina ebaluatzeko algoritmoak aztertu ziren bakterio-aurkako arloan aplikatzeko. Ondoren, azterketarekin aurrera egiteko, konputazio ereduak eraiki zen organismo batean metabolito batek erreakzio metabolikoetan duen konektagarritasuna (egitura) aztertzeko. Nodo (metabolito) nagusienak eta 40 bakterio espezie bako bakoaren erreakzio metabolikoak behin identifikatuta, ChEMBL datu-basean dauden datu kimikoekin eta aurre klinikoen kuantitatiboki erlazionatu ziren.

Jarraian, nano partikula sistemekin funtzionalizatutako bakterio-kontrako aktibitate biologikoa duten farmakoak aurreratzeko ereduak garatu ziren. Azkenik, DBFNP (BF + NP), ingelesez DADNP (AD + NP), sistemen analisia eta mapaketa burutu zen MN bakterio espezie patogenoen kontra, sare inbariantzaren (SI, ingelesez, NI), Informazio Fusioaren (IF) egindako Ikasketa Automatikoaren (IA, ingelesez ML) bidez eta Perturbazio Teoria erabiliz (PT) (SIFPTIA, ingelesez NIFPTML=N+IF+PT+ML). Hori, bakterio-kontrako farmakoaren aurkikuntzan IA/ML metodoen aplikazio zuzena izango litzateke andui multiresistenteen aurre egiteko. SIFPTIA (ingelesez NIFPTML) batukortasun estrategia, DBFNP (ingelesez DADNP) sistema berriak diseinatuz eta aurkitzeko erraztasun erabilgarria bilaka daiteke.

RESUMEN

La prevalencia de la resistencia a los antibióticos en los patógenos supera con creces nuestra capacidad para desarrollar nuevos fármacos antibacterianos (en inglés, *antibacterial drugs*, AD). Muchas corporaciones farmacéuticas evitan desarrollar sustancias antibacterianas innovadoras debido a la gran posibilidad de fracaso. Sin embargo, se necesitan con urgencia nuevos antibióticos, especialmente para las bacterias resistentes. Esto ha requerido que los científicos encuentren métodos rápidos, accesibles y económicos para descubrir nuevos fármacos y dianas moleculares contra los microorganismos infecciosos. Uno de los enfoques más empleados por las compañías farmacéuticas es el descubrimiento de fármacos asistidos por computadora. La combinación de enfoques computacionales clásicos con técnicas de aprendizaje automático (MLT, siglas en inglés de *Machine Learning Technique*) está ganando adeptos en el mundo académico y en la industria. Las MLT se utilizan en procesos de quimioinformática para predecir la actividad de un fármaco desconocido y descubrir así nuevos fármacos antibacterianos potenciales.

Por otra parte, la literatura muestra que muchos sistemas de AD funcionalizados con nanopartículas (NP) recopilados muestran que tienen actividad de amplio espectro y pueden ser promisorios para el tratamiento de infecciones bacterianas. La inhibición de cepas de diversos microorganismos como *S. aureus*, *P. aeruginosa*, *E. Faecium*, *E. Coli*, *E. faecalis*, *S. epidermidis*, *B. subtilis*, *A. Baumannii*, *S. enterica serovar Typhimurium*, *S. mutans*, *E. faecium*, *M. luteus* y *K. Pneumoniae* (algunas son resistentes a fármacos, por ejemplo, MRSA, MDR y VRE) han sido estudiadas, lo que demuestra que los sistemas duales de AD+NP (DADNP) se han centrado en la búsqueda de inhibidores del crecimiento de patógenos de gran interés en el campo de las infecciones bacterianas.

Los DADNP tienen la capacidad intrínseca de penetrar las barreras de la membrana celular bacteriana y llegar a sitios específicos con un mayor nivel de precisión y estabilidad que las moléculas de antibióticos libres. Muchas de estas combinaciones han ejercido efectos sinérgicos o aditivos en comparación con el uso de antibióticos en sus formas moleculares, que pueden contribuir a combatir muchas bacterias resistentes y apoyar tratamientos en infecciones clínicas. La mayoría de los estudios presentan efectos sinérgicos o aditivos, a diferencia del uso independiente de fármacos y nanopartículas. Esto significa que los sistemas DADNP pueden aumentar la eficacia y la velocidad de la muerte bacteriana. Otra ventaja de los sistemas DADNP es que tienen una alta ajustabilidad y una amplia gama de adaptabilidad para hacer frente a diversos escenarios, como células persistentes en macrófagos e infecciones de biopelículas, y esta integración podría ser una solución rentable. En ese sentido, el diseño integrado de sistemas de nanoantibióticos puede estar dotado de una variedad de funcionalidades, por ejemplo, capacidades de focalización, penetración y absorción mejoradas, modificación del microambiente infeccioso y combinación con otras técnicas de tratamiento. En consecuencia, existe un gran potencial para que los nanomateriales demuestren su capacidad para mejorar la eficacia terapéutica de los antibióticos.

Por otro lado, los ADs solos, las NPs o el sistema DADNP tienen que interactuar con el microorganismo. En este sentido, la comprensión del metabolismo de los patógenos juega un papel importante. Las redes metabólicas (en inglés, *metabolic network*, MN) están representadas por el conjunto de rutas metabólicas, que a su vez son una serie de reacciones bioquímicas en las que el producto (salida) de una reacción sirve como sustrato (entrada) para otra reacción. En este sentido, algunos estudios de Barabási han demostrado la influencia de los cambios en las MNs en la supervivencia de diferentes microorganismos.

Estos DADNP podrían considerarse como sistemas complejos para el análisis desde la perspectiva del modelado computacional. La incorporación de varios sistemas con diferentes

condiciones puede analizarse como un problema de ML en el descubrimiento de nuevos ADs, con aplicaciones de NP y de rutas metabólicas al mismo tiempo. La existencia de bases de datos públicas como ChEMBL con miles de informes de ensayos preclínicos de posibles ADs, un número creciente de informes experimentales de NP con acción antibacteriana y un informe previo de consenso de MN para múltiples bacterias patógenas hace muy interesante su estudio, pero la mayoría de los modelos de ML de fármacos antibacterianos y NP no tienen etiquetas múltiples. Esto forma un sistema complejo AD + NP + Agente de recubrimiento + Proteína + MN + etc, y se puede analizar como un todo o por partes (modelo aditivo de información de subsistemas). Estas partes se pueden agregar gradualmente para ver la solidez de la técnica (enfoque por bloques). Para realizar el análisis en su conjunto no hay suficientes datos. En el caso del análisis por partes, algunas partes tienen demasiados datos y otras muy pocos. Por tanto, como solución al problema, se pueden descomponer en partes o subsistemas.

La solución al problema metodológico podría ser la estrategia basada en los algoritmos NIFPTML (el acrónimo es: Invariantes de redes (NI) + Fusión de información (IF) + Teoría de perturbación (PT) + Aprendizaje automático (ML)). Esta técnica de aprendizaje automático puede abordar este tipo de desafío de múltiples etiquetas y códigos de entrada. En la primera fase del algoritmo NIFPTML se puede utilizar la teoría de redes complejas para estudiar sistemas biomoleculares (fármacos, proteínas, redes metabólicas, etc.). Las redes se pueden representar como gráficos a través de conjuntos de nodos y ejes. Un ejemplo es el gráfico molecular donde los nodos y ejes corresponden a los átomos y enlaces químicos de una molécula de fármaco. Otro ejemplo es la red de una proteína donde los nodos son aminoácidos y los ejes la secuencia y/o interacción/proximidad espacial entre los aminoácidos. Los parámetros numéricos llamados *Network Invariants* (NI) se pueden extraer de estas redes y se utilizan para cuantificar la estructura de estos sistemas. Estos parámetros o índices numéricos de redes o Redes (N) pueden ser correlacionados con las propiedades biológicas de dichos sistemas mediante técnicas de Inteligencia Artificial (AI) y/o Aprendizaje Automático (ML).

Por otro lado, en muchos problemas de interés es necesario fusionar información sobre varios de estos sistemas al mismo tiempo. Las técnicas de Fusión de Información (IF) de diversas fuentes permiten obtener un conjunto de datos enriquecido. Los operadores de la Teoría de Perturbación (PT) permiten cuantificar las perturbaciones/desviaciones en las variables estructurales con respecto a los valores esperados para diferentes subconjuntos de variables categóricas. Finalmente, los métodos de IA/ML permiten encontrar modelos predictivos de las propiedades biológicas de los sistemas (fármacos, proteínas, etc.). Por lo tanto, en esta tesis proponemos utilizar la estrategia NIFPTML para estudiar problemas que involucran uno o más de estos sistemas al mismo tiempo. El enfoque aditivo NIFPTML es compatible con este tipo de análisis (AD + MN + NP + Agente de recubrimiento). Permite trabajar con múltiples salidas, se pueden tratar múltiples condiciones y se pueden realizar varios problemas o estudios parciales con el enfoque NIFPTML. En ese sentido, la información disponible calculada en estudios previos de sistemas de cepas mutantes de AD, NP y MN, para mejorar el descubrimiento de aplicaciones de vías metabólicas de AD, NP y MN, al mismo tiempo. Además, el enfoque NIFPTML puede probar la reutilización de fármacos conocidos como AD y/o co-terapia con diferentes NP y simular la actividad de DADNP en diferentes bacterias (o MN).

En el primer capítulo de esta tesis se presentó una introducción general. Esta incluyó una revisión exhaustiva de la aplicación del aprendizaje automático en el descubrimiento de nuevos fármacos antibacterianos. Este estado del arte se enfocó en los principales temas de este trabajo: resistencia bacteriana, principales antibióticos, dianas proteicas, mecanismos de acción, bases de datos de ensayos preclínicos y clínicos, y otras fuentes de información útiles para modelado computacional, técnicas de aprendizaje automático y algoritmos de métricas de evaluación del desempeño aplicados en el campo de los medicamentos antibacterianos. La primera sección

describió los antecedentes del descubrimiento y la resistencia a los antibióticos. Luego, se analizó el rol de las proteínas diana en los fármacos antibacterianos. En la tercera sección se recopilaron las bases de datos disponibles públicamente que se utilizan con frecuencia en la investigación de aprendizaje automático en el campo del descubrimiento de fármacos antibacterianos. Además, se proporcionó un breve análisis de ChEMBL antibacteriano preclínico. En la cuarta parte del capítulo se presentó un resumen de bases de datos públicas y privadas gratuitas que contienen información sobre ensayos clínicos antibacterianos. Además, se realiza un análisis de ClinicalTrials.gov" y de la Plataforma de Registro Internacional de Ensayos Clínicos. Además, se describieron las técnicas de aprendizaje automático y los estudios centrados en el campo de la investigación de fármacos antibacterianos, esbozando cómo se han utilizado con éxito varias MLT en los campos de descubrimiento de fármacos antibacterianos. Se incluyó información relativa a las redes neuronales, las máquinas de vectores de soporte, los árboles de decisión, los predictores de conjunto, los clasificadores bayesianos, el aprendizaje profundo, etc. Por último, se abordó acerca de los sistemas DADNP. En particular, se destacaron las vías futuras para los avances científicos y tecnológicos en el descubrimiento de fármacos antibacterianos, por lo que el ML podría utilizarse para predecir la actividad de compuestos desconocidos y descubrir nuevos agentes antibacterianos.

En segundo capítulo de esta tesis se establecieron los antecedentes y objetivos de la investigación. En el caso del primero se presentó en dos secciones, una para exponer el problema práctico y la otra para el problema metodológico. Posteriormente, se expusieron cuáles pueden ser las potenciales soluciones al problema metodológico para otros sistemas biomoleculares complejos, y se realizó una retrospectiva de los Modelos NIFPTML=N+IF+PT+ML, como una aplicación de los métodos de IA/ML, anteriormente publicados para enfrentar problemas similares. Luego, se explicaron las causas de como los Modelos NIFPTML pueden resolver el problema de la investigación de este trabajo y se abordó el enfoque de esta tesis. En la segunda parte de este capítulo se presentaron los objetivos (metodológicos y prácticos), y se argumentó el desarrollo de cada uno de ellos en el documento. En el caso de los objetivos metodológicos fueron evaluar la viabilidad del modelo lineal aditivo de información del subsistema con el enfoque NIFPTML en este problema y evaluar la solidez de la metodología NIFPTML utilizando un enfoque de bloques de información de subsistemas para este problema. Por su parte, los objetivos prácticos trazados fueron: desarrollar un modelo computacional para analizar la conectividad (estructura) de un metabolito en las redes de reacción metabólicas de un organismo de consulta, desarrollar, con técnicas de aprendizaje automático lineales y no lineales, una metodología de "fusión de información y teoría de la perturbación (PT) basada en quimioinformática" que permita relacionar cuantitativamente datos químicos y preclínicos con datos de redes metabólicas y crear un modelo que prediga la actividad biológica de fármacos antibacterianos funcionalizados con sistemas de nanopartículas utilizando el método NIFPTML.

Basado en la importancia de las redes metabólicas en la actividad biológica de los fármacos antibacterianos y la escasez de estudios de modelos computacionales de las mismas. El tercer capítulo de esta tesis trató la temática de la comprobación de la conectividad (estructura) de los complejos modelos de Redes de Reacción Metabólica propuestos para nuevos microorganismos con propiedades prometedoras que es un objetivo importante para la biología química. En principio, se realizó una comprobación manual (*Manual Curation* en inglés). Sin embargo, esta es una tarea difícil debido al elevado número de combinaciones de pares de nodos (posibles reacciones metabólicas). En capítulo se utilizaron técnicas de Combinatoria (C), Teoría de la Perturbación (PT) y Aprendizaje Automático (ML), para buscar un modelo CPTML y analizar el conjunto de datos de MN publicado por el grupo de Barabási, que incluyó el número de nodos (metabolitos), enlaces de entrada-salida (reacciones metabólicas), grado de nodo, índices topológicos y nombres completos y códigos de más de 40 especies bacterianas. En primer lugar,

se cuantificó la estructura local de un conjunto muy grande de nodos en cada MN utilizando una nueva clase de índice de nodos denominada índices lineales de Markov. A continuación, se calcularon los operadores CPT para 150000 combinaciones de nodos de consulta y referencia de las MN. Por último, se utilizaron estos operadores CPT como entradas de diferentes algoritmos ML. El modelo lineal CPTML obtenido mediante el algoritmo LDA fue capaz de discriminar los nodos (metabolitos) con asignación correcta de reacciones de los nodos no correctos con valores de precisión, especificidad y sensibilidad en el rango del 85-100% tanto en las series de datos de entrenamiento como de validación externa. Mientras tanto, los modelos CPTML basados en la red bayesiana, el árbol de decisión J48 y los algoritmos *Random Forest* fueron identificados como los tres mejores modelos no lineales con una precisión superior al 97,5%.

Los fármacos antibacterianos (AD) modifican el estado metabólico de las bacterias, contribuyendo a su muerte. Sin embargo, la resistencia a los antibióticos y la aparición de cepas bacterianas multirresistentes aumentan el interés por comprender las mutaciones de la red metabólica (MN) y la interacción de la AD vs. la MN. En el cuarto capítulo de la tesis, se propuso una metodología de "fusión de información y teoría de la perturbación basada en quimioinformática" para relacionar cuantitativamente los conjuntos de datos químicos y preclínicos con los datos de la red metabólica. Este análisis se realizó para comprender mejor la interacción de las redes metabólicas previamente analizadas con los ensayos preclínicos de fármacos antibacterianos. El análisis incluyó una etapa de preprocesamiento de datos preclínicos de actividad antibacteriana. Se analizó un gran conjunto de datos de la base de datos ChEMBL. Después de la curación de datos, se determinó que el conjunto de datos de actividad antibacteriana de ChEMBL contiene los valores de > 300 parámetros (MIC, IC50, etc.) para > 165 000 ensayos biológicos de > 50 000 compuestos frente a > 25 especies de bacterias con > 90 cepas. Posteriormente, se aplicó la Fusión de información de fármacos antibacterianos e información de redes metabólicas. Se obtuvieron valores de actividad para las diferentes propiedades biológicas de los dos subsistemas (AD y MN). Luego se preprocesaron todos los valores observados con diferentes unidades, escalas, grados de incertidumbre, etc. para obtener funciones adimensionales que caracterizan el sistema como un todo, los casos AD vs. MN. Finalmente, se realizó el modelado, se obtuvo un modelo NIFPTML de la fusión de varios métodos quimioinformáticos. Se utilizaron operadores de promedio móvil (MA) para expresar perturbaciones en los ensayos y operadores multiplicadores de PT (PTO) para realizar la fusión de datos y la reducción de dimensiones. Se construyeron modelos de Análisis Discriminante Lineal (LDA, siglas en inglés de Linear Discriminant Analysis) y 17 modelos ML basados en el índice lineal basado en átomos para predecir los compuestos antibacterianos. El modelo NIFPTML-LDA presentó los siguientes resultados para el subconjunto de entrenamiento: especificidad (S_p) = 76,1%, sensibilidad (S_n) = 72,3% y precisión (A_c) = 74,3%. Entre los modelos no lineales, el k Nearest Neighbors (KNN) mostró los mejores resultados con S_n = 99,2%, S_p = 95,5%, A_c = 97,4% y AUROC = 0,998 para los conjuntos de entrenamiento y validación. En general, los modelos lineales y no lineales del NIFPTML de los fármacos antibacterianos frente a las redes metabólicas presentaron buenos parámetros estadísticos, y podrían contribuir a encontrar nuevas mutaciones metabólicas en la resistencia a los antibióticos y a reducir el tiempo/costes en la investigación de fármacos antibacterianos.

Por su parte, en el quinto capítulo de la tesis, se utilizó el modelo NIIFPTML por primera vez para estudiar un gran conjunto de datos de sistemas DADNP putativos compuestos por > 165000 ensayos antibacterianos de la base de datos preclínicos de ChEMBL y 300 Ensayos de NP frente a múltiples especies de bacterias. Se entrenaron modelos alternativos con análisis discriminante lineal (LDA), redes neuronales artificiales (ANN), redes bayesianas (BNN), K-vecino más cercano (KNN) y otros algoritmos. El modelo NIFPTML-LDA fue más simple con valores de

$Sp \approx 90\%$ y $Sn \approx 74\%$ tanto en la serie de entrenamiento ($>124K$ casos) como en la de validación ($>41K$ casos). Los modelos IFPTML-ANN y KNN son notablemente más complicados aun cuando están más balanceados $Sn \approx Sp \approx 88.5\% - 99.0\%$ y $AUROC \approx 0.94 - 0.99$ en ambas series. También se realizó una simulación (con >1900 cálculos) del comportamiento esperado para DADNP putativos en 72 ensayos biológicos diferentes. Los supuestos DADNP estudiados están formados por 27 fármacos diferentes con múltiples clases de NP y tipos de cubiertas. Además, se probó la validez del modelo aditivo obtenido con 80 complejos DADNP sintetizados experimentalmente y probados biológicamente (informados en > 45 artículos). Todos estos DADNPs muestran valores de MIC $< 50 \mu\text{g}\cdot\text{mL}^{-1}$ (punto de corte utilizado) mejores que los MIC de AD y NP solos (efecto sinérgico o aditivo). Los ensayos involucran complejos DADNP con 10 tipos de NP, 6 materiales de recubrimiento, rango de tamaño de NP de 5 a 100 nm frente a 15 antibióticos diferentes y 12 especies de bacterias. El modelo NIFPTML-LDA clasificó correctamente el 100 % (80 de 80) de los complejos DADNP como biológicamente activos. La estrategia aditiva NIFPMTL puede convertirse en una herramienta útil para ayudar en el diseño de sistemas DADNP para la terapia antibacteriana teniendo en cuenta solo la información sobre los componentes AD y NP por separado.

El sexto capítulo de esta tesis doctoral y, a diferencia de los trabajos anteriores se incorporaron tres subsistemas MN, AD y NP. Este estudio de la interacción entre AD (ChEMBL), NP y redes metabólicas intentó comprender los mecanismos potenciales de cepas multirresistentes (MDR) con redes metabólicas perturbadas (MN). Para mapear los sistemas DADNP (AD + NP) versus sistemas MN de especies bacterianas patógenas se utilizó un análisis NIFPTML. En consecuencia, se seleccionó el algoritmo NIFPTML para buscar modelos predictivos basados en un conjunto de datos ChEMBL de $> 160\ 000$ ensayos AD enriquecidos con 300 NP y > 25 ensayos MN de diferentes especies bacterianas. NIFPTML usa el proceso IF para unir los tres conjuntos de datos, creando un modelo NIFPTML de análisis discriminante lineal (LDA) con $Sp \approx 90\%$ y $Sn \approx 80\%$ y el mejor modelo de red neuronal artificial (ANN) encontrado presentó buenos resultados, con $Sp \approx Sn \approx 95\%$ en las series de entrenamiento y validación, por lo que podría ser útil para el descubrimiento de sistemas DADNP. También se realizaron simulaciones de $> 140\ 000$ puntos de sistemas DADNP putativos contra cepas bacterianas de tipo salvaje y knockout (KO) generadas computacionalmente. Los modelos NIFPTML lineales y aditivos fueron capaces de predecir 102 casos experimentales de DADNP complejos con un alto grado de variedad estructural y biológica. Esto podría contribuir con la vigilancia tecnológica (en la mejora de los fármacos antibacterianos) hacia la aparición de cepas resistentes a múltiples fármacos (MDR) con redes metabólicas perturbadas.

Finalmente, el último capítulo agrupó las conclusiones y trabajos futuros sugeridos en el tema de investigación. En este capítulo 7, se expusieron los principales hallazgos del trabajo realizado durante esta tesis doctoral. Se destacaron las vías futuras para los avances científicos y tecnológicos en el descubrimiento de nuevos fármacos antibacterianos, y se concluyó que ML/AI podría usarse para predecir la actividad de compuestos desconocidos y descubrir nuevos agentes antibacterianos. Los enfoques de ML/AI están abriendo cada vez más nuevas regiones de espacio químico para la exploración. Además, se describen otros enfoques prometedores como la reutilización de fármacos y la combinación de sistemas de NP antibacterianos, aunque existen obstáculos entre ellos y el éxito.

Table of contents

CHAPTER 1. GENERAL INTRODUCTION	1
1. ANTIBIOTICS	3
2. TARGET PROTEIN FOR ANTIBIOTICS AND ANTI-BACTERIAL DRUGS	9
3. PRECLINICAL AND CLINICAL DATABASES OF ANTIBACTERIAL DRUGS	14
3.1 General public datasets of Preclinical assays	14
3.2 Clinical assays of antibacterial drugs. General clinical trial	15
3.2.1 “ClinicalTrials.gov” and the International Clinical Trials Registry Platform” (ICTRP) analysis	16
4. 4. AI/ML MODELS FOR ANTIBIOTIC DISCOVERY	17
4.1 Background of AI/ML models.	17
4.2 Machine learning techniques (MLT)	17
4.3 Model validation metrics and applications	22
5. ANTIBACTERIAL DRUG DISCOVERY USING MACHINE LEARNING	23
5.1 A brief background of ML in AD research	23
5.2 Examples of the application of ML in antibacterial studies	23
6. REFERENCES	28
CHAPTER 2. BACKGROUND AND OBJECTIVES	41
1. BACKGROUND	43
1.1 Practical problem	43
1.2 The methodological problem	44
1.3 The solution to the methodological problem for other complex biomolecular systems	44
1.4 Previous NIFPTML models for similar problems	45
1.5 Previous NIFPTML models for the present problem	46
1.6 The focus of this thesis	47
2.OBJECTIVES	47
2.1 Methodological objectives	48
2.2 Practical objectives	48
2.3 Objective’s development	48
3. REFERENCES	50
CHAPTER 3. PREDICTING METABOLIC REACTION NETWORKS WITH PERTURBATION-THEORY MACHINE LEARNING (PTML) MODELS	53
1. INTRODUCTION	55

2. METHODS	58
2.1 Dataset of complex networks.	58
2.2 Markov linear indices for complex networks.	59
2.3 PT operators for MRNs of different organisms.	60
2.4 PTML linear model for MRNs of different organisms.	61
2.5 PTML non-linear models.	62
3. RESULTS AND DISCUSSION	63
3.1 PTML linear models.	63
3.2 PTML non-linear models.	65
4. CONCLUSIONS	67
5. REFERENCES	67
CHAPTER 4. MACHINE LEARNING MAPPING OF METABOLIC NETWORKS VS. CHEMBL DATA OF ANTIBACTERIAL COMPOUNDS	74
1. INTRODUCTION	76
2. MATERIALS AND METHODS	78
2.1 ChEMBL data set of antibacterial compounds	78
2.2 IFPTML analysis steps	79
2.3 IFPTML linear model	80
2.4 IFPTML non-linear models	81
3. RESULTS AND DISCUSSION	83
3.1 IFPTML linear model.	83
3.2 IFPTML Non-Linear models.	87
3.3 Comparison with other heterogeneous series of compounds approaches	92
4. CONCLUSIONS	94
5. REFERENCES	94
CHAPTER 5. TOWARDS MACHINE LEARNING DISCOVERY OF DUAL ANTIBACTERIAL DRUG-NANOPARTICLE SYSTEMS	100
1. INTRODUCTION	102
2. MATERIALS AND METHODS	104
2.1 IFPTML DADNP data analysis phases	104
2.2 ChEMBL and NP datasets	105
2.3 IF step for observed biological parameters	106
2.4 IF step for function of reference	107
2.5 Shannon-entropy scaling of physicochemical information	108
2.6 PT data preprocessing	108
2.7 IF step and design of training and validation subsets	109
2.8 IFPTML additive cross-over linear model	109

2.9 IFPTML models training and validation	110
3. RESULTS AND DISCUSSION	110
3.1 IFPTML DADNP additive linear model	110
3.2 IFPTML-ANN linear vs. non-linear models	112
3.3 IFPTML-WEKA Non-Linear models	115
3.4 Comparison to previous models	117
3.5 IFPTML DADNP simulation experiment	118
3.6 DADNP experimental cases simulation	119
4. CONCLUSIONS	122
5. REFERENCES	123
CHAPTER 6. TOWARDS RATIONAL NANOMATERIAL DESIGN BY PREDICTION OF DRUG-NANOPARTICLE SYSTEMS INTERACTION VS. BACTERIA METABOLIC NETWORKS	134
1. INTRODUCTION	136
2. MATERIALS AND METHODS	138
2.1 IFPTML analysis steps	138
3. RESULTS AND DISCUSSION	145
3.1 IFPTML DADNP additive linear model.	145
3.2 IFPTML-ANN linear vs. non-linear models.	146
3.3 IFPTML-WEKA AI/ML models.	149
3.4 IFPTML compared previous models.	154
3.5 IFPTML mapping of DADNP vs. MN of strains.	156
3.6 DADNP experimental cases simulation.	161
4. CONCLUSIONS	164
5. REFERENCES	164
CHAPTER 7. CONCLUDING REMARKS	174
1. CONCLUSIONS	176
2. FUTURE WORKS	177
3. REFERENCES	179
SCIENTIFIC OUTCOME	182
1. PUBLICATIONS	182
2. OTHER PUBLICATIONS	182

Table index

Table 1.1. Timeline of antibiotics, resistance, and key characteristics.	6
Table 1.2. Antibacterial classes, target, and resistance mechanism.	11
Table 1.3. Free tools about antibiotic resistance	14
Table 1.4. Public data sets containing information about substances and their biological activity are useful for training machine learning models (Reported numbers were obtained in September 2021).	15
Table 1.5. Free Databases offering information about CT (Reported numbers were obtained in November 2021).	16
Table 1.6. Classification performance evaluation in the case of 2×2 confusion matrix	22
Table 1.7. Chemoinformatic approaches for the development of novel antibacterial compounds.	27
Table 3.1. Details of the metabolic networks of >40 organisms	59
Table 3.2. Average values of f_k for the metabolic networks of >40 organisms	64
Table 3.3. Results of CPTML model for metabolic networks of >40 organisms	65
Table 3.4. Results of CPTML-non lineal models for metabolic networks of >40 organisms	66
Table 4.1. IFPTML workflow variables model.	83
Table 4.2. IFPTML linear model results for ChEMBL AD vs.MNs	87
Table 4.3. IFPTML-Non-linear AD vs. MN systems models	87
Table 4.4. Chemoinformatic approaches for the development of novel antibacterial compounds (Heterogeneous Series of compounds, Drug family >10)	93
Table 5.1. IFPTML DADNP model results summary	111
Table 5.2. IFPTML-ANN DADNP systems models	113
Table 5.3. IFPTML- WEKA Non-Linear models	116
Table 5.4. IFPTML study of experimentally tested DANP complexes	120

Table 6.1. IFPTML DADNP vs. MN model results summary	146
Table 6.2. IFPTML-ANN DADNP vs. MN models	148
Table 6.3. IFPTML-WEKA AI/ML models	150
Table 6.5. ML models of AD compounds, MN of bacteria, and/or NP antibacterial systems	155
Table 6.6. IFPMTL probabilities for AD vs. BB(0, -0.1) mutants (selected examples)	159
Table 6.7. Examples of IFPTML study of experimentally tested DADNP complexes	163

Figure index

Figure 1.1. Examples of approved antibiotic drugs, 1930-1950 (1-5)	3
Figure 1.2. Examples of approved antibiotic drugs, 1950-1980 (6-15)	4
Figure 1.3. Examples of approved antibiotic drugs, 1980-2000 (16-20)	5
Figure 1.4. Examples of more recent antibiotics (21-25)	8
Figure 1.5. Machine learning tools and their antibacterial drug discovery applications.	18
Figure 2.1. Relationship between the database, chapter, and other published works.	47
Figure 3.1. Workflow of the PTML method applied to mrns checking.	58
Figure 3.2. PTML non-linear model performance for metabolic networks of >40 organisms.	67
Figure 4.1. IFPTML model for AD vs. MN development workflow.	78
Figure 4.2. IFPTML information processing detailed workflow.	81
Figure. 4.3. Detailed score for the training set considering 17 ML techniques applied.	90
Figure. 4.4. Detailed score for the test set considering 17 ML technique applied.	91
Figure 4.5. William´s plot of residuals versus leverages for AD vs MN in the test and external validation sets.	92
Figure 5.1. IFPTML algorithm workflow for DADNP systems.	105
Figure 5.2. IFPTML detailed information processing workflow.	107
Figure 5.3. Auroc analysis of IFPTML-MLP and IFPTML-LNN models.	114
Figure 5.4. IFPTML-ANN model input variable sensitivity analysis for AD&NP, AD, and NP subsystems.	115
Figure 5.5. IFPTML-LDA DADNP systems simulation (selected results).	118
Figure 5.6. MIC($\mu\text{g}\cdot\text{ml}^{-1}$) surface scatterplot vs. Histograms of NP size and AD hydrophobicity distribution.	120

Figure 6.1. IFPTML workflow for DADNP vs. MNS mapping.	139
Figure 6.2. IFPTML information processing detailed workflow.	140
Figure 6.3. IFPTML ROC curve analysis.	147
Figure 6.4. Values of the accuracy in the y-randomization test, from the different training data divisions.	153
Figure 6.5. William´s plot of residuals versus leverages for DADNP vs MN in the training and test sets.	154
Figure 6.6. Effect of ko of gene over MNs topology.	157
Figure 6.7. IFPTML mapping of ko mutants.	160

LIST OF ABBREVIATIONS

Ac	Accuracy
AD	Antibacterial drugs
Alg	Alginate
ADMET	Absorption distribution metabolism elimination and toxicity
AI	Artificial Intelligence
AME	Aminoglycoside-modifying enzymes
AMP	Ampicillin
AMR	Antimicrobial resistance
ANN	Artificial neural network
APSn	Average of Size of Nanoparticle
ARDB	Antibiotic Resistance Genes Database
AUROC	Area Under Receiver Operating Curve
BLAST	Basic Local Alignment Search Tool
BLR	Binary logistic regression
BN	Bayesian Network
CARD	Comprehensive Antibiotic Resistance Database
CAT	Acetyltransferases
ChEMBL	Chemical European Molecular Biology Laboratory database
CRAB	Carbapenem-resistant <i>A. baumannii</i>
Cil	Cylindrical
CIP	Ciprofloxacin
CT	Clinical trial
CHL	Chloramphenicol
CZD	Ceftazidime
CTX	Cefotaxime
CXM	Cefuroxime
Ch	Chitosan
CV	Cross Validation
DL	Deep Learning
DNA	Deoxyribonucleic acid
DADNP	Dual Antibacterial Drug-Nanoparticle
DT	Decision tree
EP	Efflux Pumps
ENR	Elastic Net Regression
FDA	Food and Drug Administration
FP	False positive
FN	False negative
HGT	Horizontal Gene Transfer
GEN	Gentamycin
IFPTML	Information fusion perturbation-theory machine learning approach
IMI	Imipenem
ISE	Iterative stochastic elimination
KAN	Kanamycin
KNN	K-Nearest-Neighbor
LASSO	Least Absolute Shrinkage and Selection Operator Regression
LDA	Linear discriminant analysis
LOGP	Logarithm of the n-Octanol/Water Partition coefficient
LR	Logistic Regression
MBS	Multiple bacterial strain

MCC	Matthew's correlation coefficient
MD	Molecular Descriptors
MER	Meropenem
MGE	Mobile genetic elements
ML	machine learning
MLSB	Macrolides Lincosamides and Streptogramin B antibiotics
MIC	Minimal inhibitory concentration
MBC	Minimal Bactericide Concentration
MDR	Multi drug resistance
MeNP	Metal Nanoparticle
MeOxNP	Metal Oxide Nanoparticle
MLP	Multilayer Perceptron
MN	Metabolic network
MPNN	Message Passing Neural Network
MRSA	Methicillin-resistant Staphylococcus aureus
MRN	Metabolic reaction network
MRSE	Methicillin-resistant Staphylococcus epidermidis
MDRSE	Multidrug resistant Staphylococcus epidermidis
MER	Meropenem
MW	Molecular weight
NBN	Naïve Bayes classifiers based on Bayes' theorem
Nf	Nanoflakes
NN	Neural Network
NP	Nanoparticle
NVLR	Number of Violations of Lipinski's Rule
OFL	Ofloxacin
PDRAB	Pandrug-resistant A. baumannii
PDA	Polydopamine
PEG	Polyethylene glycol
POL	Polymyxin B
PVP	Polyvinylpyrrolidone
PSA	Topological Polar Surface Area
QSAR	Quantitative structure-activity relationship
RIF	Rifampicin
Sn	Sensibility
Sp	Specificity
SOM	Self-organizing map (Kohonen)
RF	Random Forest
RR	Ridge Regression
SVM	Support Vector Machines
Sph	Spherical
SWOT	Strength-Weaknesses-Opportunities-Threats
TB	Tuberculosis
TEG	Triethylene Glicol
TET	Tetracycline
TGA	Thioglycolic acid
TIG	Tigecycline
TOB	Tobramycin
TOPS-MODE	TOPological Substructural MOlecular Design
TOMOCOMD-CARDD	Topological Molecular COMputer Design
TP	True positive

TN	True negative
VAN	Vancomycin
VRE	Vancomycin-resistant enterococci
WHO	World Health Organization.

CHAPTER 1. GENERAL INTRODUCTION

1. ANTIBIOTICS

Paul Ehrlich, a German immunologist, found in the late 1890s that some dyes inhibited parasite cell multiplication but had little or no effect on human cells. This discovery resulted in the invention of Salvarsan, a treatment for syphilis. Ehrlich coined the term ‘magic bullets’ for these compounds in 1906 due to their capacity to specifically target infected host cells.¹ Alexander Fleming found in 1928 that contaminated mold produced a substance that prevented the cultivation of *Staphylococcus aureus*.² Protonsil, a sulfonamide prodrug having antibacterial action, was created by Bayer in the 1930s after extensive screening.³

Selman Waksman began a systematic examination of bacteria as producers of antibiotic compounds in the late 1930s in response to the discovery of penicillin and tyrocidine. Waksman found filamentous actinomycetales (‘actinomycetes’) as a prolific producer of antibacterial compounds.⁴ Neomycin and Streptomycin, the first antibiotics developed to combat tuberculosis, were among his discoveries.^{4,5} **Figure 1.1** shows examples of approved antibiotic drugs (1930-1950).

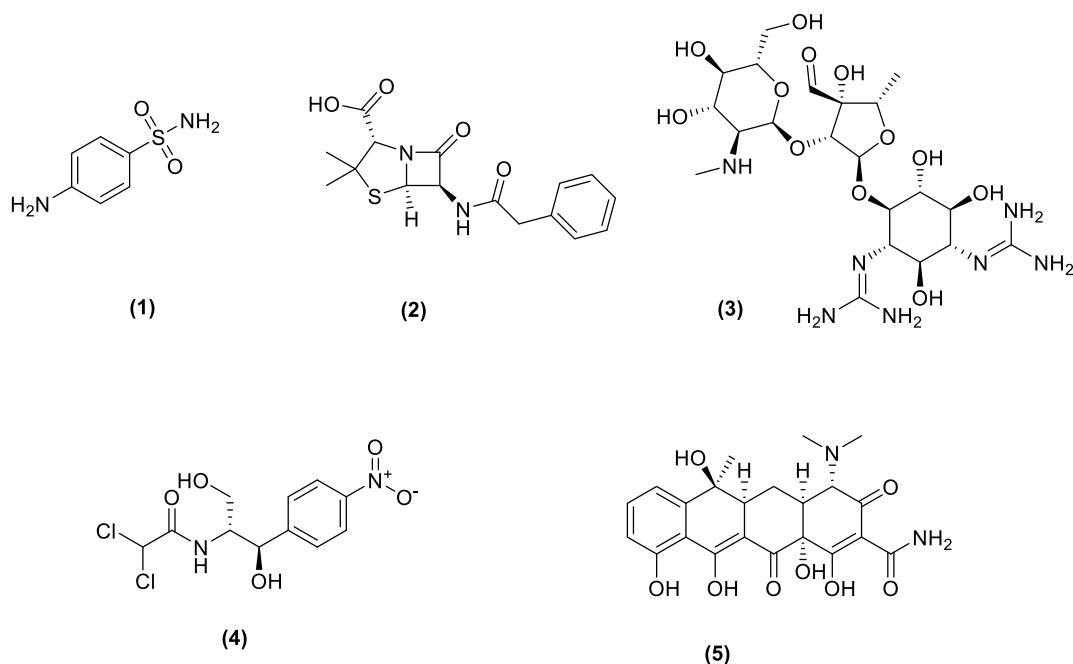


Figure 1.1. Examples of approved antibiotic drugs, 1930-1950 (1-5)

Notes: 1. Sulfanilamide, 2. Penicillin, 3. Streptomycin, 4. Chloramphenicol, 5. Tetracycline

The aminoglycosides, Streptomycin and Neomycin, were discovered in the 1940s, followed by rifampicin, nitrofurans (nitrofurantoin), and glycopeptides (Vancomycin) in the 1950s, quinolones and lincosamides in the 1960s, and mupirocin in the 1970s. The use of these drugs in clinical practice resulted in a decrease in morbidity and mortality associated with bacterial infections. Since the prevalence of antimicrobial resistance (AMR) has increased, the majority of these antibiotics are still in clinical use.⁶ The rapid and very low-cost discovery of multiple classes (and variants) of natural product antibiotics resulted in their misuse. Additionally, since the 1970s, the antibiotic discovery pipeline has been blocked, resulting in a dearth of novel antibiotics in clinical trials.^{5,6} See **Figure 1.2**, examples of approved antibiotic drugs, 1950-1980.

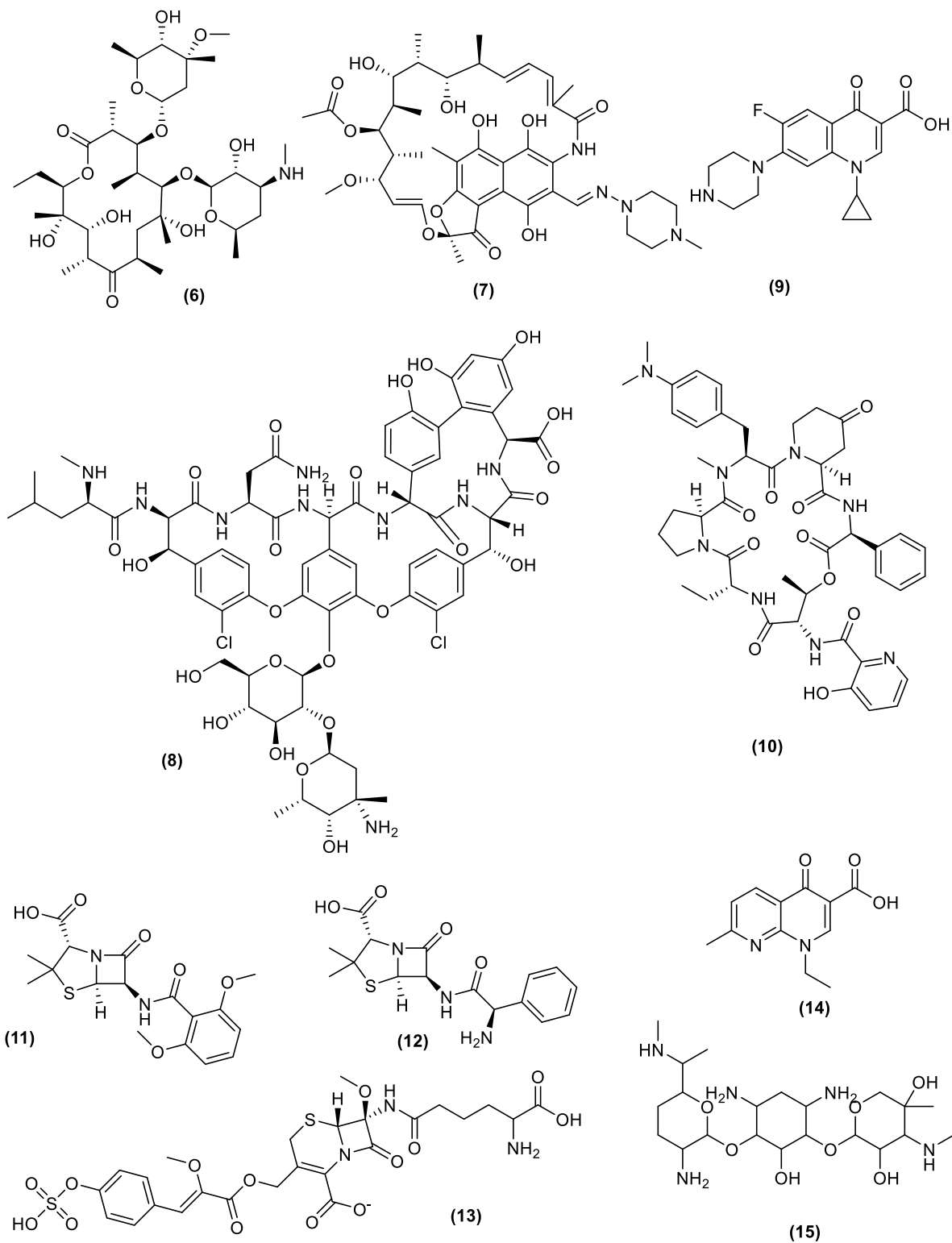


Figure 1.2. Examples of approved antibiotic drugs, 1950-1980 (6-15)

Notes: 6. Erythromycin, 7. Rifampicin, 8. Vancomycin, 9. Ciprofloxacin, 10. Streptogramin B, 11. Methicillin, 12. Ampicillin, 13. Cephamycin, 14. Nalidixic acid, 15. Gentamicin

In the 1980s, the final antibacterial classes were discovered.⁷ Due to the fact that the majority of new antibiotics belong to well-established classes, they are less effective in combating rapidly increasing resistant bacteria.⁸ **Figure 1.3** shows examples of approved antibiotic drugs, 1980-2000.

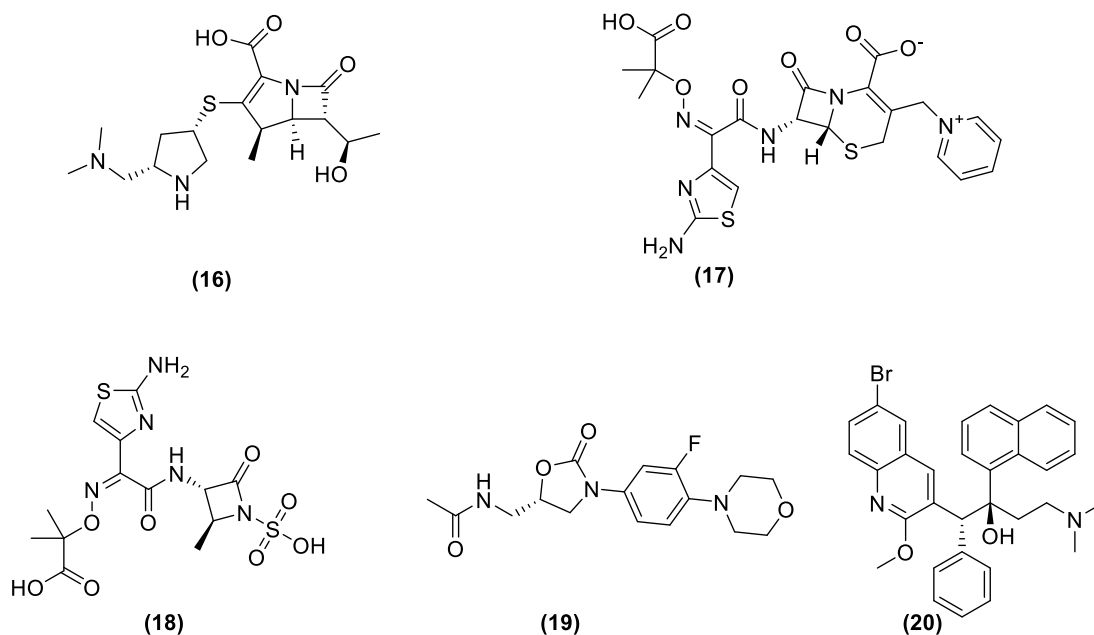


Figure 1.3. Examples of approved antibiotic drugs, 1980-2000 (16-20)

Notes: 16. Meropenem, 17. Ceftazidime, 18. Aztreonam, 19. Linezolid, 20. Bedaquiline

Parallel to the drop in the discovery of antibacterial drugs (~1970s), multidrug resistance in bacteria was immediately recognized. Inadequate use of these drugs facilitated the development of resistance. For instance, penicillin was introduced in 1946 and streptomycin was released in 1959; however, both were commercially available in 1943.⁹ (See **Table 1.1**)

Table 1.1. Timeline of antibiotics, resistance, and key characteristics.

No	Family	Antibiotic Example	Year	CRO ^a	Mechanism of Action	of	Activity or target species	Ref.
1	Sulfonamides	Sulfanilamide	1932	1942	Inhibition of dihydropteroate synthetase	of	Gram-positive bacteria	⁹
2	β -actams	Penicillin	1943	1946	Inhibition of cell wall biosynthesis		Broad-spectrum activity	⁹
3	Aminoglycosides	Streptomycin	1943	1959	Binding of 30S ribosomal subunit		Broad-spectrum activity	⁹
4	Amphenicols	Chloramphenicol	1947	1959	Binding of 50S ribosomal subunit		Broad-spectrum activity	⁹
5	Tetracyclines	Tetracycline	1948	1953	Binding of 30S ribosomal subunit		Broad-spectrum activity	⁹⁻
6	Macrolides	Erythromycin	1951	1988	Binding of 50S ribosomal subunit		Broad-spectrum activity	⁹
7	Rifamycins	Rifampicin	1958	1962	Binding of RNA polymerase β -subunit		Gram-positive bacteria	¹⁰
8	Glycopeptides	Vancomycin	1956	1988	Inhibition of cell wall biosynthesis (D-Ala-D-Ala termini of lipid II)		Gram-positive bacteria	⁹
9	(Fluoro)quinolones	Ciprofloxacin	1968	1968	Inhibition of DNA synthesis (DNA gyrase, and topoisomerase IV)		Broad-spectrum activity	¹⁰
10	Streptogramins	Streptogramin B	1963	1964	Binding of 50S ribosomal subunit		Gram-positive bacteria	¹⁰
11	β -lactams	Methicillin	1960	1961	Inhibition of cell wall biosynthesis		Gram-positive bacteria	⁹

Chapter 1

12	β-actams	Ampicillin	1961	1973	Inhibition of cell wall biosynthesis	Broad-spectrum activity	9
13	Cephalosporin	Cephamycin	1964	1964	Inhibition of cell wall biosynthesis	Gram-positive aerobic bacteria	9
14	Quinolones	Nalidixic acids	1962	1962	Inhibition of DNA synthesis (DNA gyrase)	Gram-negative bacteria	11
15	Aminoglycosides	Gentamicin	1963	1971	Binding of 30S ribosomal subunit	Broad-spectrum activity	12
16	Carbapenems	Meropenem	1985	1991	Cell wall synthesis: penicillin-binding proteins	Broad-spectrum activity	5
17	Cephalosporins	Ceftazidime	1981	1986	Inhibition of cell wall biosynthesis	Broad-spectrum activity	13
18	Monobactams	Aztreonam	1986	1988	Cell wall synthesis: penicillin-binding proteins	Gram negative bacteria.	14
19	Oxazolidinones	Linezolid	2000	2001	Protein synthesis: 50S ribosomal subunit	Gram-positive bacteria	15
20	Diarylquinolines	Bedaquiline ^b	1997	2006	Inhibition of F1 FO-ATPase	Narrow-spectrum activity (<i>Mycobacterium tuberculosis</i>)	10, 16
21	Lipopeptides	Daptomycin	2003	2003	Depolarization of cell membrane	Gram-positive bacteria	17
22	Pleuromutilins	Retapamulin ^c	2007	2007	Protein synthesis: 50S ribosomal subunit	Broad-spectrum activity	18
23	Macrolides	Fidaxomicin	2011	2011	Inhibition of RNA polymerase	Gram-positive bacteria (<i>Clostridium difficile</i>)	19
24	Cephalosporins	Ceftaroline ^d	2011	2011	Inhibition of cell wall biosynthesis	Broad-spectrum activity	20
25	Antimycobacterials	Pretomanid ^e	2019	?	Inhibition of cell wall biosynthesis	<i>Mycobacterium tuberculosis</i>	21

Notes: ^a CRO = Clinical resistance observed, ^b Bedaquiline was approved to combined therapy for MDR-TB treatment; ^c Retapamulin (topical use only) resistance was observed in clinical

isolates of *S. aureus* without previous exposure to pleuromutilins; ^d Resistance due to mutations to the mutated PBP2a; ^e Pretomanid was recently released to MDR-TB treatment. This table was modified from ²² and from the original report of ⁹.

Unfortunately, the development of new antibiotics has become more challenging. Dereplication is a problem in natural product discovery, where the same molecules are identified repeatedly,²³ and since the “golden age”, only three new classes of antibiotics active against Gram-positive bacteria, such as methicillin-resistant *Staphylococcus aureus* (MRSA), were discovered and approved: oxazolidinones (Linezolid in 2001 and Tedizolid in 2014), Daptomycin in 2006 (a cyclic lipopeptide) and Fidaxomicin in 2011 (a macrocycle drug for *C. difficile*). However, a high number of analogs of existing classes and antibiotic combinations have reached the market.²⁴ **Figure 1.4** shows examples of more recent antibiotics.

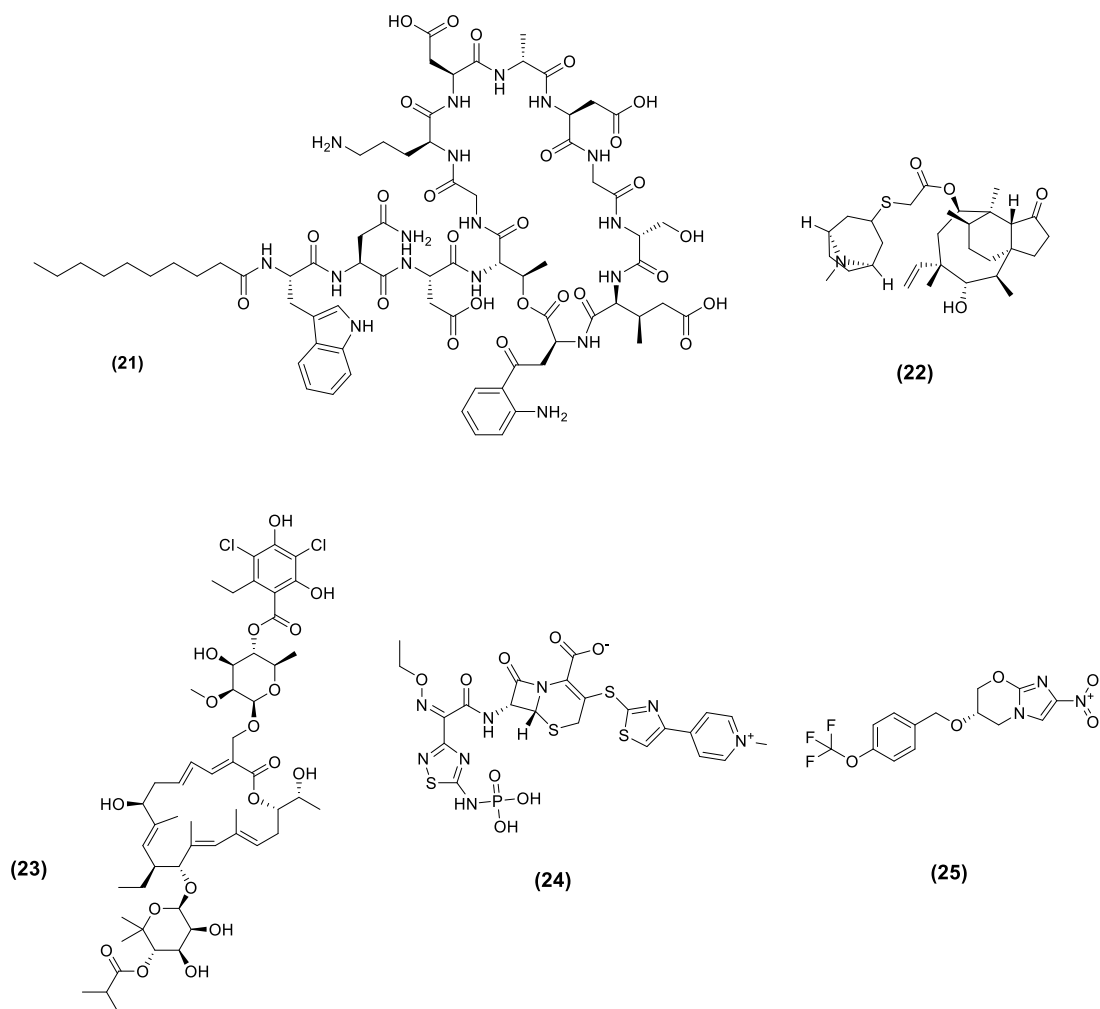


Figure 1.4. Examples of more recent antibiotics (21-25)

Notes: 21. Daptomycin, 22. Retapamulin, 23. Fidaxomicin, 24. Ceftaroline, 25. Pretomanid.

2. TARGET PROTEIN FOR ANTIBIOTICS AND ANTI-BACTERIAL DRUGS

Cell wall biosynthesis: The majority of bacteria rely on the development and maintenance of their cell walls to survive changes in osmotic pressure, which would otherwise cause the cells to lyse. Cell wall formation takes place in two stages: intracellular and extracellular. Penicillin and cephalosporin, as well as the glycopeptide antibiotic Vancomycin (VAN), all interfere with extracellular cross-linking steps, reducing the mechanical strength of the peptidoglycan layer.^{25,26} The mycobacterial cell wall has been recognized as a target for anti-TB drugs due to the critical significance of cell wall formation and assembly.²⁷

Cell membrane integrity: A basic requirement for bacterial cells is the integrity of the cell membrane as a barrier to uncontrolled ion and small molecule leaks into and out of cells. The lipopeptide antibiotic daptomycin (DAP) appears to work by affecting membrane integrity.²⁶ Furthermore, lanthionine-containing peptides like nisin, which are used as food preservatives to kill bacteria and inhibit spore outgrowth from spore-forming bacteria like clostridia, can form transient membrane pores during the membrane phase of peptidoglycan assembly by complexing with a peptide-glycolipid intermediate.^{28, 29}

Protein biosynthesis: Bacteria produce thousands of proteins during each cell generation, which perform a wide range of functions such as survival, growth, and division. Antibiotics can interfere with hundreds of stages, including the selection and activation of amino acid building blocks, the chaperoning of aminoacyl-tRNAs to the ribosome, peptide bond condensation, and chain elongation and termination steps on the ribosome. Bacterial cell death is caused by a disruption in protein production. The most well-known antibiotics that inhibit bacterial protein synthesis target either the small ribosomal subunit (tetracyclines and aminoglycosides) or the large subunit (tetracyclines and aminoglycosides) (Erythromycins, Streptogramins, and Lincosamide).²⁶

Many antibiotics interfere with protein synthesis, which is carried out by a macromolecular process known as the ribosome. The ribosome is composed of two ribonucleoprotein subunits: 30S and 50S. Translation has four stages: start, elongation, termination, and ribosome recycling.³⁰ Antibiotics (aminoglycosides like Streptomycin and Kanamycin) impede ribosome assembly by binding to the 30S or 50S subunits (chloramphenicol). Macrolides like Erythromycin stop nascent chain elongation, while peptidyl-transferase inhibitors like puromycin stop protein synthesis.³¹ For example, peptide translocation is inhibited by fusidic acid.³² Antibiotics commonly target the bacterial ribosome, with most clinically used antibiotics targeting either the decoding site on the small ribosomal subunit (30S subunit) or the peptidyl-transferase center on the large subunit (50S subunit).³³

Most antibiotics that target the 30S subunit decrease protein synthesis by either blocking tRNA binding to the ribosome or preventing tRNA transport across the ribosome during translocation. The majority of antibiotics that target the 50S subunit inhibit protein synthesis by either interfering with the binding of aminoacylated-tRNAs at the A- or P-sites or by preventing the nascent polypeptide chain from passing through the ribosomal tunnel. Bacterial antibiotic resistance mechanisms include efflux, decreased influx, drug modification and degradation, as well as mutation, alteration, or overexpression of the target. The bulk of ribosome-targeting antibiotics in clinical trials are semi-synthetic versions of naturally occurring chemicals, but more research will be needed to generate antibiotics that target unique ribosome locations.²⁶

DNA and RNA metabolism: The selective suppression of information transfer in bacterial cells' macromolecular metabolism can go beyond protein production inhibition. Interfering with one

or more of the several processes in DNA replication or RNA transcription should also be bactericidal. The synthetic antibacterial fluoroquinolones, which have long been the most often prescribed antibiotic class, target a deconcatenation phase at the conclusion of DNA replication, which is carried out by the enzyme DNA gyrase. The rifamycin class of natural compounds bind to bacterial RNA polymerase at the active site and so prevent DNA transcription into RNA.^{34, 35}

Folate biosynthesis: Unlike the previous four examples, the fifth historically important set of targets for clinically useful antibiotics is not directly involved in cellular macromolecule biosynthesis, but is rather a pathway for constructing and recycling the coenzyme form of the vitamin folic acid. The tetrahydro oxidation state of folate's pteridine ring is essential for the conversion of the uracil moiety in 20-dUMP to the 5-methyl (thymidine) moiety in dTMP, one of the four building blocks for DNA formation.²⁶ The sulfonamide antibiotics, in continuous use for the past 80 years, act as competitive inhibitors for para-aminobenzoate in the maturation of the folate scaffold. In addition, an enzyme that functions in a crucial readjustment of the folate oxidation state, from dihydro to tetrahydro, is selectively inhibited by trimethoprim. For decades, the combination of trimethoprim and a sulfonamide was commonly used to inhibit bacterial DNA production. Members of each of these antibiotic groups have found application in certain therapeutic niches. Anderson *et al.*³⁶ discuss them, as well as pharmacokinetic and pharmacodynamic characteristics, side effects, and limiting toxicities.

Although good antibacterial targets are scarce in number, they are virtually uniformly implicated in macromolecular production pathways. The majority of them are directed at ribosomal RNA, stages in cell wall formation, or membranes. Only β -lactams and fluoroquinolones target enzymes, and each of these drugs targets at least two enzymes.³⁷ ADs have targets with important bacterial functions that cannot be met by feeding them intermediates. Only a handful of the key antibiotic classes used in systemic monotherapy target important enzymes. The majority of them are directed at ribosomal RNA, stages in cell wall formation, or membranes. **Table 1.2** shows antibacterial classes, target, and the resistance mechanism (including examples of clinically important resistance to antibacterial).

Table 1.2. Antibacterial classes, target, and resistance mechanism.

Drug class	Target	Resistance type	Resistance mechanism	Examples resistance	clinical Ref.
β-Lactams	Multiple PBPs	Altered PBP.	PBP 2a	<i>mecA</i> en <i>S. aureus</i> , <i>S. pneumoniae</i> ,	33
		Enzymatic degradation	Penicillinase per ambler class	Gram Negative	33
Glycopeptides	Lipid II	Altered target	D-alanyl D-alanine is changed to D-alanyl D-lactate	VRE (<i>faecium</i> and <i>faecalis</i>)	33
Macrolides	50S RNA	Altered target	Methylation of ribosome active site with reduced binding	<i>erm</i> methylases in <i>S. aureus</i> , <i>S. pneumoniae</i> , <i>S. pyogenes</i>	33
		Efflux pumps	Mef type pumps	<i>S.pneumoniae</i> , <i>S. pyogenes</i>	33
Oxazolidinones	50S RNA	Altered target	Mutation leading to reduced binding to active site	<i>E. faecium</i> and <i>S. aureus</i>	33,38
Chloramphenicol	50S RNA	Antibiotic inactivation/	Chloramphenicol acetyl transferase	CAT in <i>S. pneumoniae</i>	33,39
		Efflux pumps	New membranes transporters	<i>cml A</i> gene and <i>flo</i> gene efflux in <i>E. coli</i>	33
Streptogramins	50S RNA	Altered target	ribosome methylation,	HGT MLS _B	40
Lincomycins	50S RNA	efflux,	modification	HGT MLS _B	40
		Efflux	New membranes transporters	<i>ten</i> gene encoding efflux protein in Gram positive, Gram negative	33,41
Tetracyclines	30S RNA	Altered target	Production of protein that bind to the ribosome and alter the conformation of the active site	tet(M) or tet (O) in in Gram positive, Gram negative	33,41
		Decrease uptake.	Change in outer membrane AGE's	<i>Pseudomonas</i> , Gram negative	33,41

Drug class	Target	Resistance type	Resistance mechanism	Examples resistance	clinical Ref.
		Enzymatic modification			
Fluoroquinolones	Gyrase, Topo IV	Altered targets	Mutations leading to reduce binding to active site. Stepwise two or more than two target mutations, efflux	Mutation in <i>gyr A</i> in enteric Gram Negative. and <i>S. aureus</i>	33, 41
		Efflux	Membrane's transporters	Mutation in <i>gyr A</i> and <i>par C</i> in <i>S. pneumoniae</i> . Nor-A in <i>S. aureus</i>	33
Polymyxins	Membranes	Modification	Modification of the lipid A or Kdo with aminoarabinose	Gram negative (<i>Salmonella enterica</i> , <i>Klebsiella pneumoniae</i> , etc)	29,42
Sulfadrog		Altered target	Mutations of genes encoding DHPS	<i>E.coli</i> , <i>S.aureus</i> , <i>S. pneumoniae</i>	33
Rifampicin	RNA polymerase	Changes in <i>rpoB</i> at many sites	Change in the β subunit of bacterial RNA polymerase	Mutation β , <i>E. coli</i> , <i>Mycobacterium tuberculosis</i>	29

Notes: PBPs, Penicillin binding proteins; HGT, horizontal gene transfer; LPS, lipopolysaccharide. DHFR, Dihydrofolate reductase; DHPS, dihydropteroate synthase; HGT, horizontal gene transfer, AGE's=Aminoglycosides modifying enzymes, CAT= Chloramphenicol acetyl transferase, MLS_B=Macrolide-Lincosamide-Streptogramin B group, Kdo= 3-deoxy-D-manno-oct-2-ulosonic acid. Mef MFS-type transporters that confer macrolide resistance in *Streptococcus pneumoniae*. Mef and Mel in the novel efflux-mediated macrolide resistance system in *S. pneumoniae* and other gram-positive bacteria. The Cfr rRNA Methyltransferase Confers Resistance to Phenicol, Lincosamides, Oxazolidinones, Pleuromutilins, and Streptogramin A Antibiotics.⁴⁰

Tool for drug target identification and resistance mechanism prediction

Table 1.3 shows some tools for drug target identification and the prediction of antibacterial resistance mechanisms. One of the main tools is the Basic Local Alignment Search Tool (BLAST). This tool finds local similarities between sequences, compares nucleotide or protein sequences to databases, and calculates statistical significance. BLAST can help you find gene families and infer functional and evolutionary relationships between sequences. The tool includes information, denoted by the letters *nt* for nucleotides and *np* for proteins. For example, BLASTp (protein-protein BLAST) is a protein sequence similarity search tool, which can be very useful in determining the structure and biological function of proteins of interest in antibacterial discovery.⁴³ Another NIH tool is Genbank, which, unlike BLAST, employs record references rather than a similarity algorithm. This is an annotated collection of all publicly available DNA sequences.⁴⁴

On the other hand, the research of the predictive genotype-to-phenotype studies based on the identification of AMR genes in genomes can employ various web servers such as ResFinder, Resfams, ARDB, or CARD. ARDB is a consolidated database of antibiotic resistance data, and it has proposed the idea of standardizing resistance annotation in freshly sequenced organisms to help find and characterize novel genes.⁴⁵ ResFinder identifies acquired genes and/or finds chromosomal mutations mediating antimicrobial resistance in the whole or part of the DNA sequence of bacteria.⁴⁶ Resfams is a curated database of protein families and associated profile hidden Markov models, confirmed for antibiotic resistance function and organized by ontology.⁴⁷ It also analyzes the antibiotic resistance gene composition of over 6000 sequenced microbial genomes.

Another database is UniProt (released in 2020), which contains over 189 million sequence records with >292,000 proteomes of different organisms. It includes completely sequenced viral, bacterial, archaeal, and eukaryotic genomes. It is available through the UniProtKB Proteomes portal (<https://www.uniprot.org/proteomes/>).⁴⁸ Finally, the Comprehensive Antibiotic Resistance Database (CARD) (<https://card.mcmaster.ca>) is a curated resource providing reference DNA, protein sequences, detection models, and bioinformatics tools on the molecular basis of bacterial antimicrobial resistance (AMR). It comprises 263 significant pathogens in terms of computer-generated resistome predictions. Sequence variants not previously described in the scientific literature are included, as are prevalence figures for AMR genes in infections, genomes, and plasmids. The CARD also includes bioinformatic tools for identifying antibiotic resistance genes in whole- or partial-genome sequence data, including unannotated raw sequence assembly contigs. The end result is a rigorously curated database in a user-friendly style that assembles over 1,600 known antibiotic resistance genes, allowing sophisticated antibiotic resistance analysis and query in a way that will benefit the broader biomedical research community.⁴⁹

From previous knowledge of biological mechanisms in these databases, antibiotic resistance genes could be identified. Some studies have included approaches for genotyping resistance characteristics, including whole genome sequence building,⁵⁰ as well as BLAST and genomic assembly⁵¹. Bradley *et al.*⁵² incorporated classification models for species identification, phylogenetic branch, and resistance profiling from sequence files into the Mykrobe prediction software tool for well-known pathogens such as *M. TB* and *S.aureus*,^{50,52} although they were not effective for novel resistance mechanisms.

Table 1.3. Free tools about antibiotic resistance.

Tool Name ^a	Type of tool	Organization/Country	Link	Ref.
BLAST	Software/ Database	National Center for Biotechnology Information, USA	https://blast.ncbi.nlm.nih.gov/Blast.cgi	43
CARD	Database	McMaster University, Ontario, Canada	https://card.mcmaster.ca/	49
ARDB	Database	Center for Bioinformatics and Computational Biology University of Maryland College Park, USA	http://ardb.cbcb.umd.edu/	45
Protein database	Database	National Center for Biotechnology Information, USA	https://www.ncbi.nlm.nih.gov/protein/	43
Uniprotkb	Database	UniProt consortium, EU	https://www.uniprot.org/help/uniprotkb	48
ResFinder	Database	Center for Genomic Epidemiology, National Food Institute, Technical University of Denmark	https://cge.cbs.dtu.dk/services/ResFinder/	46
Resfams	Database	Dantas Lab, Washington University	http://www.dantaslab.org/resfams	47

Notes: ^a Database. ARDB=Antibiotic Resistance Genes Database, BLAST=Basic Local Alignment Search Tool, CARD= Comprehensive Antibiotic Resistance Database.

3. PRECLINICAL AND CLINICAL DATABASES OF ANTIBACTERIAL DRUGS

3.1 General public datasets of Preclinical assays

The construction of MLT-based models involves a training set of known compounds and a validation set of compounds with activity values for a specified endpoint prediction. In this way, research organizations typically start with a previously reported sequence of compounds in order to develop a model that may be used to discover new hit compounds. Publicly available data sets are also frequently used in MLT investigations, and various useful sources might be cited as examples. **Table 1.4** shows a number of available compounds in each repository and their usefulness. ChEMBL and BindingDB are the two most popular of these datasets. ChEMBL is a manually curated library of bioactive molecules that contains roughly 2.1 million compounds with over 14,000 reported targets of action (ranging from enzymatic assays to cell

lines and microorganisms).⁵³ BindingDB is a supplementary database that reports on the binding affinities of 1.01 million chemicals against 8,600 protein targets.⁵⁴ Another intriguing database is AntibioticDB,⁵⁵ which contains approximately 1,100 compounds currently in pre-clinical development, clinical trials in phases I-III, clinical trials in phase 4 either awaiting approval or recently approved, and compounds that have been discontinued. Additionally, Drug Repurposing Hub is a curated and annotated collection of FDA-approved medications, clinical trial medications, and pre-clinical tool compounds, as well as a companion information resource. Finally, Zinc15 is a free virtual screening library of commercially available chemicals. Zinc15 has about 750 million chemicals that are commercially available (230 million purchasable compounds in ready-to-dock and 3D formats).⁵⁶

Table 1.4. Public data sets containing information about substances and their biological activity are useful for training machine learning models.

Database	Compound Number ^a	Usage	Link	Ref
PubChem	111 M	Computational Chemistry	https://pubchem.ncbi.nlm.nih.gov/	57
ChEMBL	2.1 M	Drug Discovery	https://www.ebi.ac.uk/chembl/	53
BindingDB	1.01 M	Drug Discovery	https://www.bindingdb.org/	54
DrugBank	13 K	Drug Discovery	https://www.drugbank.ca/	58
AntibioticDB	1.1 K	Computational Chemistry	https://www.antibioticdb.com/	55
Drug Repurposing Hub	6.8 K	Drug Discovery	https://clue.io/repurposing	59
ZINC	750 M	Virtual Screening	https://zinc.docking.org/	56

Notes: ^a (Reported numbers were obtained in September 2021).

3.2 Clinical assays of antibacterial drugs. General clinical trial

Over the last decade, a number of clinical trial (CT) registries have been established. Trial registration is governed by European and United States federal legislation, as well as international agreements (WHO).^{60, 61} All interventional CTs must be registered in the European Union (EU) and the United States (US), as required by an international consortium of medical journal editors.⁶²

Unlike scientific publications, which indicate scientific interest in a given technology, the number of CTs indicates how many attempts have been made to move the technology from the bench to the clinical phase.⁶³ However, a comprehensive database analysis of clinical studies in bacterial infections has not been performed, and published information is limited. Therefore, we have focused on performing a brief analysis of CTs on bacterial infections to generate an overview of the actual number and content involving potential ADs (**Section 3.2**). Several research studies have focused on trends in CTs, including bacterial infections; nevertheless, findings are scarce. For example, Bliziotis *et al.* reviewed the clinical evidence of rifampicin as adjuvant therapy in treating Gram-positive infections.⁶⁴ Another work associated with Gram-negative bacterial infections was performed by Long *et al.*⁶⁵ They analyzed the current status of investigational cephalosporins in early CTs (phase I and phase II development) to treat these bacterial infections.⁶⁵ An evaluation of CTs involving individuals with acute bacterial infections of the skin and skin structure was conducted to determine the efficacy and safety of these novel antibiotics.^{66, 67} On the other hand, Shepshelovich *et al.*⁶⁸ sought to analyze

the consistency of reporting in systemic antibiotic therapy trials to resolve discrepancies between ClinicalTrials.gov entries and matched articles previously described in general medicine.⁶⁸ **Table 1.5** present a summary of free public and private databases that contain information about CTs.

Table 1.5. Free Databases offering information about CT.

Database ^a	Compound Number ^b	Country	Link	Ref.
ClinicalTrials.gov	>396 K	USA	http://clinicaltrials.gov	69
ICTRP	~19.5 K	WHO	http://apps.who.int/trialsearch/Default.aspx/	70
EudraCT	>41 K	European Union	https://www.clinicaltrialsregister.eu/	71
ISRCTN	>21 K	UK	https://www.isrctn.com/search?q=	72
ANZCTR	>16 K	Australia/New Zealand	https://www.anzctr.org.au/	73
JMACTR	~0.44 K	Japan	https://dbcentre3.jmacct.med.or.jp/jmactr/Default_Eng.aspx/	74

Notes: ^a Database. ICTRP= International Clinical Trials Registry Platform, EudraCT=European Union Drug Regulating Authorities Clinical Trials Database, ISRCTN= International Standard Randomised Controlled Trials Number, ANZCTR= Australia, New Zealand Clinical Trials Registry, JMACTR=Japan Medical Association Clinical Trials Register. ^b Compound Number. (Reported numbers were obtained in November 2021).

Of the numerous databases on CTs (**Table 1.5**), the ClinicalTrials.gov database stands out as the largest one. It is provided by the US National Library of Medicine and is represented in more than 220 countries around the world.⁶⁹ Another database of note is the WHO International Clinical Trials Registry Platform (ICTRP), which aggregates CT information from various databases and implements a uniform access mechanism to the CT data stored in them.⁷⁰ In doing so, it attempts to address one of the difficulties of the various sources, which is the absence of a single comprehensive international registry of CTs.

3.2.1 “ClinicalTrials.gov” and the International Clinical Trials Registry Platform” (ICTRP) analysis

The comprehensive analysis of CTs involving ADs was carried out in November 2021 using ClinicalTrials.gov and the International Clinical Trials Registry Platform (ICTRP). They can be combined to provide a better picture of clinical studies employing ADs.⁷⁰ The systematic analysis of these databases focused on the date, research type (observational versus interventional), the number of patients enrolled, *etc.* In the case of ClinicalTrials.gov, the terms “antibacterial” OR “anti-bacterial” were used in the search and 6,469 studies were found.

In parallel, a search was conducted on the International Clinical Trials Registry Platform (ICTRP) using the same terms. This database incorporates information from numerous national and international databases. The ICTRP portal site added 250 additional CTs to the final analysis after excluding 106 duplicated search results obtained from ClinicalTrials.gov.

4. 4. AI/ML MODELS FOR ANTIBIOTIC DISCOVERY

4.1 Background of AI/ML models.

Antibiotic resistance in pathogens is surpassing our ability to generate new antibiotics.⁷⁵ Because of the high risk of failure, many pharmaceutical companies avoid researching novel antibacterial agents. Nonetheless, new medicines are desperately needed, particularly for resistant microorganisms.⁷⁶ To cope, antibiotic discovery needs to be accelerated while reducing the associated costs. Computer-assisted drug discovery methods have gained ground in this regard. The identification of new structural classes of antibiotics by algorithmic prediction of molecular properties with machine learning can be a solution for the aforementioned difficulties.⁷⁷ In this way, early in silico drug discovery can be achieved by exploring vast chemical spaces that are beyond the reach of current experimental approaches. Combining classical computational approaches with machine learning techniques is gaining traction in academia and industry. MLTs such as Artificial neural networks (ANNs), Support vector machines (SVMs), Decision trees (DTs), Ensemble predictors, and Bayesian classifiers could be used in cheminformatic pipelines to predict unknown drug activity and thus uncover new potential antibacterial drugs.⁷⁸

4.2 Machine learning techniques (MLT)

Machine learning algorithms include classification, regression, and clustering. They can forecast a compound's biological response based on its chemical description.^{79, 80} MLTs' non-linear nature enables them to identify hidden patterns in massive amounts of data that would otherwise go unreported. In the field of antibiotic discovery, diverse classification MLTs such as ANNs, DTs, Random forests (RFs), SVMs, k-nearest neighbors (kNNs), and Linear discriminant analysis (LDA) have been used. **Figure 1.5** shows ML tools and their antibacterial drug discovery applications.

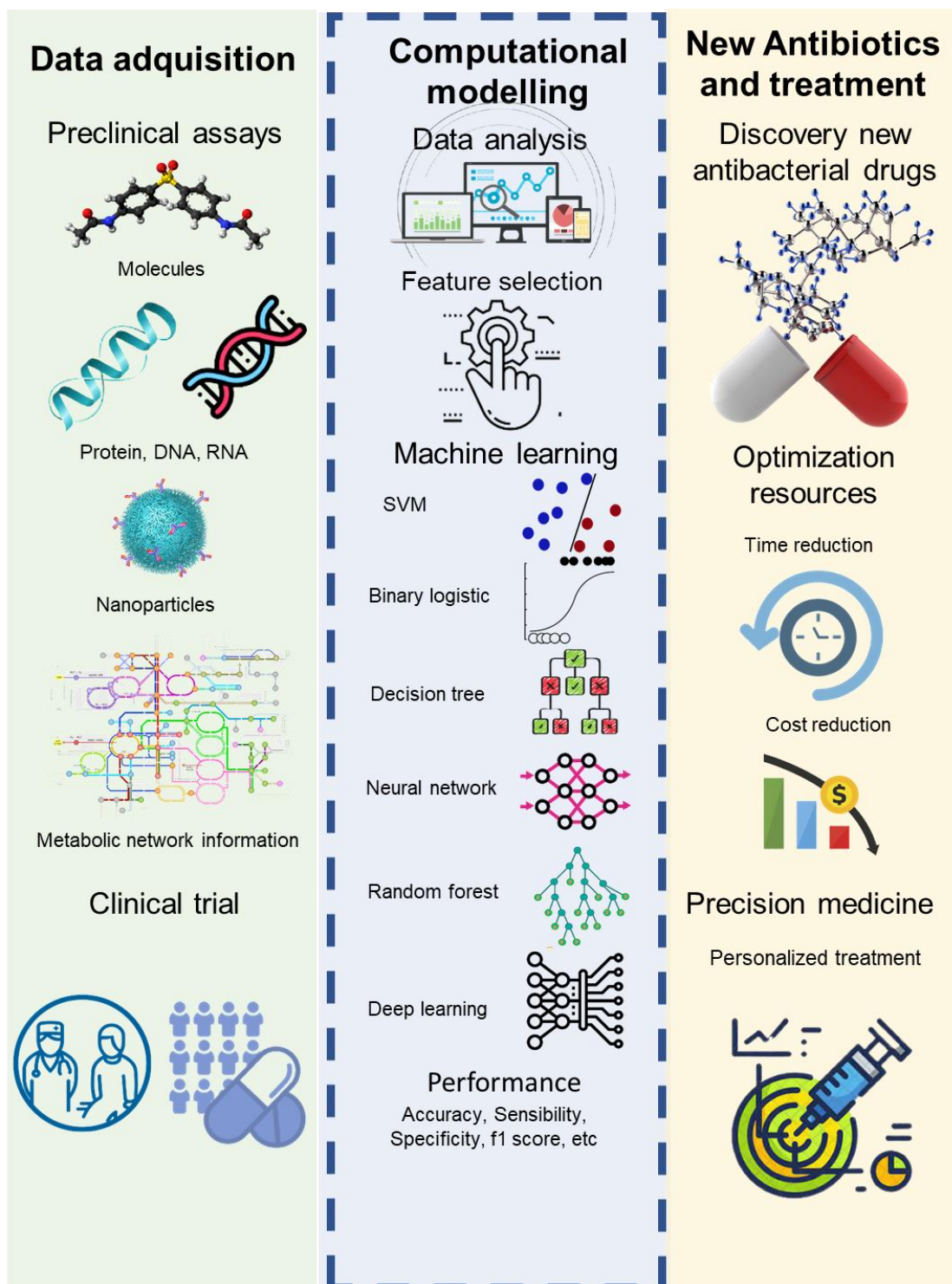


Figure 1.5. Machine learning tools and their antibacterial drug discovery applications.

4.2.1 Artificial neural networks (ANNs) and deep learning (DL)

ANN and DL can provide resilient solutions to difficult real-world problems without making any prior statistical assumptions about the probability distribution of the input data. An ANN is a machine learning approach that is inspired by the way biological nervous systems process

information, in which highly linked neurons work together to learn from incoming signals. ANNs are made up of an input layer of neurons connected to one or more hidden layers of neurons, which are then connected to an output layer.²² ANNs have the ability to generalize, i.e. correctly predict future (unseen) cases based on what they have previously predicted. An ANN's fault tolerance allows it to handle noisy or uncertain data as well as possibly incomplete data. Also, neural networks are self-adaptive, meaning they can adjust to changes in data. Multilayer Perceptrons (MLPs) are a well-known example of a learning technique based on gradient descent optimization,⁸¹ MLPs are organized into three different types of layers: input, hidden, and output.^{82, 83}

This algorithm has been widely applied in antibacterial studies. For instance, Cronin *et al.*⁸⁴ analyzed the structure-based classification of antibacterial activity, Chersakov *et al.*⁸⁵ applied inductive QSAR descriptors to distinguish compounds with antibacterial activity, and Speck-Planche *et al.*⁸⁶ developed an in silico model for virtual screening of potent and safe anti-pseudomonas agents. Durrant *et al.*⁸⁷ used ANN techniques to predict how certain antibiotics work against bacteria and to predict ligand-based and receptor-based binding. A neural network (NN) that essentially reconstructs the brain's cellular architecture made these predictions. Then they used the NN to predict the activity of other possible ligands not included in the training set. These strategies allowed NNs to anticipate how certain antibiotics worked against bacteria, decreasing the need for human and animal testing, thereby lowering research costs.

DL models are NNs that can represent progressively more complex functions by layering and unit layering. A network's depth increases as the number of layers increases.^{88, 89} Feedforward neural networks with two or three layers are considered traditional (or shallow), whereas deep neural networks feature hundreds of layers. The primary motivation for developing deep-learning approaches was the inability of conventional machine learning algorithms to analyze unstructured natural input. To discover patterns in the input, feature extractors were developed to convert the raw data to a suitable internal representation. Each layer of a deep-learning neural network improves its ability to abstractly represent incoming input. As a result, a multi-level abstraction was created that can be used to construct high-level concepts.⁹⁰

Deep learning is also commonly employed in drug discovery. This field has seen a lot of deep-learning research. For example, Gawehn *et al.*⁹¹ detailed current machine learning methods used to calculate QSAR models. These models are "shallow" because they only have one layer of feature modifications. The article explained a deep neural network and presented a brief history of deep learning. Recently, Stokes *et al.*⁹² applied a feed-forward deep neural network using Morgan fingerprints as the molecular representation for Predicting New Antibiotic Candidates.

4.2.2 Decision trees

A decision tree is constructed based on the recurrent partitioning principle in which the function space is partitioned into areas that include data with comparable answer values.⁹³ The model displays a flowchart-like tree structure where each internal node represents a test on a selected variable, branch denotes test outcomes, and each terminal (leaf) node assigns a class label. The differences are in the tree structure, the splitting criteria, the pruning method, and the way missing values are handled.^{94, 95} The popularity of DTs is largely due to their outcomes being easy to interpret, beginning with the fact that the tree structure can be quickly translated into a collection of 'ifthen' rules.⁹³ For instance, Durant and Amaro⁸⁷ remarked that the DT is an important MLT applied to antibacterial drug discovery. In Xue *et al.*'s study,⁹⁶ the DT presented results of 94% accuracy in classifying antibacterial compounds.

Recently, El Zahed *et al.*⁹⁷ reported two decision tree-based machine learning models to investigate molecular descriptors (MDs) governing Gram-negative permeation and efflux evasion. This study screened 4,500 small compounds in efflux-compromised *E. coli* to find novel Gram-negative antibiotics. This method indicated that only efflux-compromised *E. coli* may use hydrophobic and planar small compounds with low molecular stability. Similarly, Suay-García *et al.*⁹⁸ employed a tree-based classification system to predict antibiotic efficacy against *E. coli*. The model is a hierarchical decision tree where a discrete index is used to organize compounds based on their index values. The model screened the DrugBank database and identified 134 antimicrobial candidates. This study indicated that DT techniques could be a viable alternative to standard ways for obtaining prediction models, and the use of DTs provides exciting novel drug candidates for further investigation as repurposed antibacterial treatments.

4.2.3 Ensemble methods

Ensemble approaches integrate numerous individual (base) predictive models to get a more accurate model. These base predictors are often produced by running the same algorithm with multiple training conditions. Different approaches can be used to combine multiple individual decisions, such as average values or voting procedures.⁹⁹ Using an ensemble model instead of a single model offers many advantages. A single predictor is less accurate and less robust. Second, ensembles can break down a difficult task into smaller, more manageable chunks. Ensemble methods can also be used to analyse big datasets because the data can be divided into smaller subsets that are used to train different models and then blended. Furthermore, since an ensemble deals with multiple hypotheses at once, it reduces the danger of selecting a bad model.¹⁰⁰ In drug discovery research, different ensemble methods have been employed. Among them are meta-algorithms that aim to combine the abilities of weak learners, such as bagging, boosting, voting, and stacking.

Bagging: Bagging methods are used to reduce the variance of a base estimator (e.g., the decision tree) before building an ensemble from it. They are a fast and simple technique for improving a single model without changing the fundamental base algorithm.¹⁰¹ It can be applied with a CART implementation (SimpleCart) based on classifier trees, etc.¹⁰²

Boosting: Boosting algorithms are able to transform weak learners into strong ones. Intuitively, a weak learner yields little more than a random guess, while a strong learner yields almost perfection.¹⁰¹ Adaboost, LogitBoost, and MultiBoosting are three representative algorithms of this family of algorithms.¹⁰³ These models can be built together with entropy-based classifier trees (DecisionStump).

Random forests (RFs): These are a well-known ensemble technique frequently used in drug development.⁹³ Each node is divided according to the best of a randomly chosen set of descriptors in each bootstrap sample, with each unpruned tree growing to its greatest extent. The prediction is made in RF regression by averaging the predictions of individual trees. An excellent description of RF theory can be found in the literature.¹⁰⁴

Gradient boosting: The basic idea behind boosting techniques is to combine several learning models with low predictive power to create a model with very high predictive power. Freund and Schapire¹⁰⁵ and Friedman¹⁰⁶ proposed the gradient descent boosting technique, which interprets boosting as a function estimation problem. Unlike random forests, gradient descent boosting creates an ensemble (usually formed of DTs) by greedily minimizing an objective (loss) function using gradient descent.¹⁰⁶ Recently, Khaledian and Broschat¹⁰⁷ designed a good ensemble gradient boosting model to predict more than 6,000 putative antibacterial peptides (area under the curve ~0.98).

4.2.4 Other machine learning techniques

Support Vector Machine (SVM): One of the most powerful supervised learning systems, the SVM is widely used in drug development.¹⁰⁸ The theoretical basis of SVMs was described in statistical learning theory and consists of transforming a nonlinear feature space into a linear one by mapping the input data onto a high-dimensional feature space and fitting a linear model in the feature space.¹⁰⁹ This method derives an optimal separating hyperplane from a training data set of n points (x_i, y_i) , where y_i belongs to class y . The SVM objective function is given by **Eq. 1**.¹¹⁰

$$\begin{aligned} \text{Min}_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^k \xi_i, \text{ subject to} \\ y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i \end{aligned} \quad (1)$$

The previous model comprises the error function minimization by modifying the C parameter in order to increase or decrease the penalty for classification errors.¹¹¹ In the last decade, the use of SVMs in drug discovery has been increasing. For instance, Niehaus *et al.*⁷⁵ used SVMs to better predict tuberculosis antibiotic resistance. The model developed has a classification accuracy of 93% for predicting resistance to isoniazid, a critical first-line antibiotic for *M. TB*.

Bayesian classifiers: Bayesian classifiers are MLTs that make use of probability and statistical information based on Bayes' theorem. This theorem is also widely applied in the field of drug development.¹¹² Specifically, a Bayesian Network (BN) is a probabilistic network that graphically represents variables and their conditional dependencies via a directed acyclic graph $G=(X, P)$. The joint distribution probability function for a BN is given by **Eq. 2**

$$P(x_1 \dots x_n | Y) = \prod_{i=1}^n P(x_i | Y) \quad (2)$$

where, $P(x_i|Y)$ is the empirical conditional probability of the value of variable x_i in the current data instance given that the instance belongs to class Y .

In Naive Bayes, a probabilistic classifier, the features (variables) representing the data are statistically independent.¹¹³ Its learning phase involves estimating prior and class condition probabilities from training data. However, if the count of a feature given a class is 0, then the class cannot be predicted. The generalized Laplace correction¹¹⁴ substitutes the zero probability with a tiny constant. The main advantage of Naive Bayes is its computational efficiency, as it simplifies the estimation problem by requiring only the probability of each attribute given the class, independently of the rest. This approach also performed well despite a possible breach of its conditional independence requirement.¹¹⁴ The Naive Bayesian algorithm has been used to determine new treatments for tuberculosis with dose-response data for both whole-cell antitubercular activity and Vero cell cytotoxicity.¹¹⁵

Bayesian classifiers have a number of significant advantages over alternative techniques. To begin with, they can easily deal with missing values by averaging the values accessible for the corresponding feature in the training set. Second, Bayesian approaches enable the classification to be built by integrating prior knowledge about the domain with knowledge from other sources (e.g., different training data). Singh *et al.* Singh, Chaudhury, Liu, AbdulHameed, Tawa and Wallqvist¹¹⁶ constructed a Bayesian classification model using structural fingerprints and physicochemical property descriptors. They achieved an accuracy of 84% and precision of 86% on an independent test set in identifying antibacterial compounds.

Rule-based classifiers: Different methods have been applied in the field of antibacterial drug discovery.¹¹⁷ Two examples are the PART algorithm, a decision list that builds a partial C4.5 decision tree at each iteration and transforms the best leaf into a rule,¹¹⁸ and the Ripple-Down Rule (Ridor) learner that generates a default rule and then the exceptions to the default with the lowest (weighted) error rate. The exceptions are a set of rules that predict classes other than those chosen by the default,¹¹⁹ and the Fuzzy Unordered Rules Induction Algorithm (FURIA) is a novel fuzzy rule-based classification method introduced by Hühn and Hüllermeier.¹²⁰

Binary Logistic Regression (BLR): Also called a logit model, this is a simple classification method typically used to predict the probability of a dichotomous sample.¹²¹ The probability (P) that an observation falls into one of two categories of a dichotomous dependent variable Y is based on one or more independent variables $\{X=X_1, X_2, \dots, X_n\}$ that can be either continuous or categorical. The logistic regression model has the form given in **Eq. 3**, where LR is the logit or linear predictor function that assigns the value $Y=1$ (active) if $P(X)>0.5$, or $Y=0$ (inactive) otherwise.¹²²

$$P(X) = \frac{1}{1+e^{(LR(X))}} \quad (3)$$

Applications of these approaches can be extensively found in the prediction of antibacterial compounds.¹²³

K-Nearest Neighbors (kNNs): An instance-based learning classifier was developed using the Euclidean distance function. This algorithm is based on estimating the probability densities by finding k vectors that are closest to an unclassified vector according to a distance measure.¹²⁴ Therefore, the unlabelled vector is predicted as the modal class of the retrieved training vectors.¹²⁵ This method has been applied to the prediction of antibacterial compounds. Xue *et al.*⁹⁶ and Ding *et al.*¹²⁶ developed a similarity-based algorithm akin to Nearest Neighbors to learn and predict new targets (proteins), drugs, and target-drug interactions. First, drug and target spaces were projected onto two low-dimensional regions. Then, we estimated the drug–target interactions in low-dimensional spaces. This was inefficient because it used three matrices with random values.

4.3 Model validation metrics and applications

Internal validation: The main procedures include 5- or 10-fold cross-validation.

10-fold cross-validation: The original data set is randomly split into 10 equally sized subsets (folds); each subset contains approximately the correct proportion of the class values. Afterward, 9 out of 10 subsets are used to train the given model, whereas the remaining subset (left out) is used for testing. This process is repeated 10 times until each subset is left out once. The 10 evaluation results are then averaged to produce a single classification quality outcome.

5-fold cross-validation: Repeat the same procedure for a total of just 5 times.

External validations: The classification quality of models overtraining and test data sets were assessed through statistical parameters from Medicinal Chemistry literature, such as Accuracy (Acc), Matthews' correlation coefficient (MCC), Sensitivity (Sn), Specificity (Sp), negative predictive value (also called sensitivity of negatives), the false positive rate (or false alarm rate), and Area Under the Curve (AUC).^{121, 127} **Table 1.6** shows the computational metrics that are widely used in the literature to assess the accuracy of classification studies.

Table 1.6. Classification performance evaluation in the case of 2×2 confusion matrix.

Metric	Formula	Reference
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$	121, 128

MCC	$\frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP * FN) * (TN + FP) * (TN + FN)}}$	121, 128
Specificity	$\frac{TN}{TN + FP}$	121, 128
Sensitivity (Recall)	$\frac{TP}{TP + FN}$	128, 129
Precision	$\frac{TP}{TP + FP}$	121, 129
F1 Score	$\frac{2 * Precision * Recall}{Precision + Recall}$	130, 131
AUC	The area under the ROC curve	130, 131

Notes: TP: True positive, TN: True negative, FP: False positive, FN: False negative

5. ANTIBACTERIAL DRUG DISCOVERY USING MACHINE LEARNING

5.1 A brief background of ML in AD research

Several overviews of ML in the antibacterial field have been published as a result of the rapid expansion of ML applications in medicinal chemistry. Macesic *et al.*¹³² examined the current literature on machine learning for exploring antimicrobial resistance (AMR). ML has been utilized to predict antimicrobial susceptibility genotypes/phenotypes, design AMR clinical DT, uncover novel antimicrobial agents, and optimize antimicrobial therapy. Nourani *et al.*¹³³ evaluated computational approaches for predicting pathogen-host protein-protein interactions in molecular biology topics. Computational approaches generally make use of sequence data, protein structure, and known interactions. When there are enough known interactions to use as training data, classic IA/ML approaches are used. It is a text that adds information on the usage of various ML methods to predict pathogen-host protein-protein interactions. The problem of drug discovery, on the other hand, has been addressed in the paper entitled “Machine-learning techniques applied to antibacterial drug discovery”, where Durrant *et al.*⁸⁷ proposed that computer-aided drug discovery could uncover novel antibiotics more quickly and cheaply, resulting in higher hit rates and faster preclinical and clinical testing times. They demonstrated how NNs and DTs have been used to identify antibiotics that have been tested in laboratories.⁸⁷ Most of the overviews concerning ML are based on its applications and explore it from a cheminformatic, medical, or biological perspective.

Recently, Serafim *et al.*²² suggested that ANNs, DTs, and RFs are more commonly utilized in predictive model creation than traditional classification and regression methods. These MLTs can use whole-genome sequencing data to explore and discover new resistance mechanisms in bacteria populations, ultimately assisting in the prioritization and discovery of molecular targets.²² As a result, Ivanenkov *et al.*¹³⁴ emphasized the urgent need for novel antibiotics, particularly for resistant microorganisms. They created an *in silico* model capable of identifying several compounds that can exert antibacterial activity on *E. coli* among 140,000 chemicals.¹³⁴ Stokes *et al.*⁹² used a trained deep neural network to predict antibiotic efficacy in molecules that are structurally different from known antibiotics, such as Halicin, which is effective in mice against broad-spectrum bacterial infections.⁹²

5.2 Examples of the application of ML in antibacterial studies

ML has been used in the development of antibacterial active drugs in a variety of research projects published during the last decade. Below are a few significant examples. **Table 1.7** summarizes ML approaches to modelling antibacterial activity from the literature.

One of the first examples of the use of ML or statistical modelling to predict antimicrobial activity was published by Garcia-Domenech *et al.*¹³⁵. These researchers realized a study of pattern recognition to detect microbiological activity in a group of heterogeneous compounds. The structural descriptors utilized were topological connectivity indexes, while the methodologies were linear discriminant analysis and artificial neural networks (nonlinear analysis). Although both methods can distinguish between active and inactive chemicals, the artificial neural network outperformed the linear discriminant analysis, with a prediction success rate of 98% vs. 92%. To distinguish between active and inactive compounds in antibacterial drugs, Tomas-Vert *et al.* suggested a new topological technique. This method uses neural networks with training algorithms and several artificial intelligence concepts and methods with topological descriptors. After the training, the network's data can be interpreted using QSAR.¹³⁶

At a similar time, Mishra *et al.*¹³⁷ reported the use of discriminant functions of antibacterial activity based on physicochemical and topological parameters. They tried to find a discriminant function for antibacterial activity by combining semiempirical (quantum chemical) computations and topological indices. It appears that one of the maxima and minima vibrational frequencies is involved in antibacterial action.¹³⁷ Subsequently, Cronin *et al.*⁸⁴ devised a simple QSAR for antibacterial activity classification and prediction. In this study, 661 compounds were classified using linear discriminant and binary logistic regression. In the same group, 3D Molecular Descriptors (MD) were used to classify antibacterial and nonantibacterial activity.¹²³ A set of 661 organic molecules was modelled by utilizing hydrophobicity (log Kow) and AM1-level MD expressing geometric, electrostatic, nucleophilic, and electrophilic properties. LDA and BLR achieved an overall categorization rate of roughly 90%.

Mut-Ronda *et al.*¹³⁸ used molecular topology to classify antibacterial chemicals (quinolones). They employed high-accuracy connection functions and discriminant equations (>90%). In the same year, topological approaches were used to classify 972 antibacterial and non-antibacterial medicines and identify new prospective antibacterial agents.¹³⁹ They use pharmacological distribution diagrams to visualize the selection of new antibacterial drugs. This group also added an MLP or ANN model and pharmacological distribution diagrams.¹⁴⁰ The results validated the topological descriptors' discriminative capacity. The application of LDA and MLP in a guided search and selection of novel antibacterial structures was highly successful in *in vitro* activity and toxicity testing.

Molina *et al.*¹⁴¹ reported on the application of Topological Substructural Molecular Design (TOPS-MODE) to classify antibacterial drugs using computer-aided molecular design. The LDA model obtained a 91% global classification of 'good'. Another approach to the discovery of antibacterials was proposed by González-Díaz *et al.*¹⁴². They included the Markovian chemical in silico design (MARCH-INSIDE) descriptors (2.5D indices). A basic stochastic approach to the idea of electronegativity equalization (Sanderson's principle) was employed to train the classification model. In training sets, the 2.5D-QSAR model correctly distinguished between antibacterial and non-antibacterial chemicals (accuracy ~93%). The antibacterial activity of the novel compound 2-bromo-3-(furan-2-yl)-3-oxo-propionamide against *Pseudomonas aeruginosa* ATCC 27853 and *E. coli* ATCC 27853 was discovered.¹⁴² Similarly, Marrero-Ponce *et al.*¹⁴³ introduced the Topological Molecular Computer Design (TOMOCOMD-CARDD) for the classification and design of antibacterial drugs using computer-aided molecular design. This study employed a complete data set of 1,006

antibacterial drugs to simulate antibacterial activity. The models (non-stochastic and stochastic indices) accurately categorized over 90% of 1525 compounds in the training sets. In external test sets, these models correctly classified 92.8% and 89.3% of 505 compounds, respectively. In general, these approaches have become a useful tool for the *in silico* discovery of antibacterial agents.¹⁴³

Chersakov *et al.*⁸⁵ described linear and non-linear modelling inductive descriptors for antibacterial chemicals. No linear models were created using ANNs. Non-linear ML algorithms separate substances with and without antibacterial activity 93% of the time (in a set of 657 structurally heterogeneous compounds including 249 antibiotics and 408 general drugs).¹⁴⁴ used kNNs to classify compounds, comparing them using the Minkowski distance $L(p)$. The data collection contained 4,346 chemicals (including 520 antibiotics, 562 bacterial metabolites, 958 drugs, 1,202 drug-like compounds, and an additional 1,104 human metabolites). They found that kNN ML outperformed linear models (LDA, MLR) and was comparable to the ANN methods. Xue *et al.*⁹⁶ reported SVM, k-NN, and C4.5 DT algorithms to predict antibiotic chemicals (230 antibacterial and 381 nonantibacterial compounds). SVM had the highest prediction accuracy for ADs at 96.66%, 98.15%, and 99.50%, 98.02% for nonantibacterial compounds, respectively.

From the work of Speck-Planche *et al.*¹⁴⁵ comes the application of multispecies models in the antibacterial field. These models have emerged recently; however, some of them predict biological activity only for the same genus or within a subgroup of bacteria (**21-25** and **27-34 models**). The unified multitasking (Mtk) QSAR model was used to predict both anti-streptococci action and toxicity in biological models such as *Mus musculus* and *Rattus norvegicus*. With over 11,500 instances in the database, the Mtk-QSAR ANN model is a promising method for virtually screening strong and safe anti-streptococci drugs. The same researchers presented a multitasking model (Mtk-QSBER) for the simultaneous prediction of anti-tuberculosis activity and toxicological profiles of medicines.¹⁴⁶ The Mtk-QSBER LDA model classified more than 90% of the cases in the total database (almost 12,000 cases), making it an extremely powerful tool for computer-assisted drug screening. Additionally, it was utilized to determine the ADMET (absorption, distribution, metabolism, elimination, and toxicity) properties of pharmaceuticals and/or chemicals under a variety of experimental conditions.¹⁴⁷ The LDA model, which was developed using about 37,800 samples of data, achieved an overall accuracy of more than 95% in both the training and prediction (validation) sets. The Mtk-QSBER model was used to predict avarofloxacin (AVX) properties under 260 different experimental conditions. The results confirmed AVX's extraordinary anti-*E. coli* activity and safety. According to these studies, the Mtk-QSBER model is a viable computational technique for the virtual screening of anti-bacterial drugs that might be extended to safer pharmaceuticals with defined pharmacological activity.

Wang *et al.*¹⁴⁸ suggested using *in silico* ML models to discover new agents active against methicillin-resistant *S. aureus* (MRSA) based on 5,451 cell-based anti-MRSA assay data. They developed four machine learning methods (Naïve Bayesian, SVM, Recursive Partitioning (RP), and k-NN). The overall predictive accuracies of models exceeded 80% for both training and test sets. The best model was used for virtual MRSA screening, confirming 12 new anti-MRSA agents (MIC values ranging from 4 to 64 mg L⁻¹). However, no evidence was shown concerning cytotoxicity to eukaryotic cells.

Speck-Planche *et al.*⁸⁶ developed an Mtk-QSBER model to predict anti-Pseudomonas and ADMET properties of organic compounds. The Mtk-QSBER model, which was created using a large and diverse dataset (around 54,000 cases), achieved greater than 90% accuracy in both the training and prediction sets. The researchers demonstrated the applicability of the Mtk-

QSBER model using the experimental drug delafloxacin. The predictions for numerous biological effects of this drug were extremely similar to the experimental results. In another work, Speck-Planche *et al.*⁸⁶ developed an Mtk-QSBER model to forecast antibacterial activity and ADMET properties against microorganisms associated with neglected diseases, specifically noma. The Mtk-QSBER model was developed by utilizing a large and diverse chemical dataset (30,181 pairs) and has an Acc of more than 90% in both the training and prediction sets. The experimental results for the antibacterial medicine delafloxacin converged with the model's various features. This was the first model that emphasized the search for virtual anti-noma agents. Castillo-Garit *et al.*¹⁴⁹ presented a classification study of 2,230 drugs (1,006 with antibacterial activity) based on TOMOCOMD-CARDD descriptors. The non-stochastic and stochastic bilinear indices were 86.3% and 83.6%, respectively.

Antimicrobial Peptide discovery and virtual screening were conducted using the Mtk approach. Kleandrova *et al.*¹⁵⁰ focused on the simultaneous prediction of antibacterial and cytotoxic peptides. This work classified/predicted peptides using 3,592 examples and achieved 96% accuracy in both the training and prediction (test) sets. The alanine scanning method was used to calculate the quantitative contributions of amino acids to the biological effects of a specific peptide (at their respective sequence positions). They used the Mtk-computational model to generate a small library of ten peptides. All of the peptides were anticipated to possess a broad range of antibacterial and anti-cytotoxic activities.¹⁵⁰ Moreover, Speck-Planche *et al.*¹⁵¹ discussed studies on antibacterial peptides. They developed a multitarget chemo-bioinformatic model for predicting peptide antibacterial activity against a variety of Gram-positive bacterial strains. The model was constructed by comparing 2,488 AMP sequences to 50 Gram-positive bacterial strains. Both the training and prediction (test) sets of this mt-chemo-bioinformatic model correctly classified over 90% of the samples.¹⁵¹

Masalha, *et al.*¹⁵² composed a study with the goal of indexing natural products in order to facilitate the discovery of less expensive antibacterial therapeutic drugs. They made use of the iterative stochastic elimination algorithm to build a model of the 628 antibacterial drugs and 2,892 natural products. The AUC was 0.957, demonstrating a discriminative and robust prediction model. To achieve this 72% enrichment factor, the study used a virtual screening method that included both active and inactive compounds. The proposed indexing methodology identified ten natural compounds as promising antibacterial medication candidates. According to PubMed searches, two of the ten compounds (caffeine and ricinine) have antibacterial action. The other eight phytochemicals are still being tested. The proposed prediction model's efficiency and speed could be used to virtually screen vast chemical databases for AD candidates.¹⁵²

Recently, a remarkable usage of ML in antibacterial activity against *E. coli* was published by Ivanenkov *et al.*¹³⁴ They used an in silico approach to locate compounds with antibacterial activity in a large dataset of over 140,000 molecules. They also tested six in silico approaches, including kNN, SVM, and RF. The Kohonen Self-Organizing Maps (SOM) showed the strongest prediction power (~75.5%). Experiments with selected chemicals have shown high effectiveness against *E. coli*. Additionally, the CC50 values against eukaryotic cell lines were calculated in order to estimate the selectivity index for the most promising drugs.¹³⁴

Nocedo-Mena *et al.*¹⁵³ developed the first NIFPTML model for antibacterial activity modeling by combining perturbation theory, machine learning, and information fusion approaches. They employed preclinical antimicrobial activity assays from the ChEMBL database and Metabolic Networks (MNs) developed by the Barabási's group. The training set comprised 83,605 instances of more than 25 different bacterial species. In training/validation series, the best linear model obtained had an Sp of ~90.3% and an Sn of ~88.1%.¹⁵³

Table 1.7. Chemoinformatic approaches for the development of novel antibacterial compounds.

m ^a	Cmpd. Type ^a	n ^b	n _{Act.}	Var. _b	Tech. ^c	Acc (%) ^d	MultiSpecies ^e	Drug Family ^f	MO ^g	Val. _h	Ref.
1	HSC	111	60	7	LDA	94.0	No	3	No	i	135
2	HSC	111	60	7	ANN	89.0	No	3	No	i	135
3	HSC	664	249	62	ANN	94.8	No	8	No	i	136
4	HSC	59	24	17	LDA	85	No	3	No	i	137
5	HSC	661	249	6	LDA	92.6	No	8	No	ii	84
6	HSC	661	249	6	BLR	94.7	No	8	No	ii	84
7	HSC	661	249	62	ANN	-	No	8	No	iii	84
8	HSC	661	249	3	LDA	90.1	No	8	No	ii	123
9	HSC	661	249	3	BLR	92.1	No	8	No	ii	123
10	HSC	294		8	LDA	> 90	No	-	No	i	138
11	HSC	972	241	8	LDA	86.8	No	> 5	No	i	139
12	HSC	433	217	2	LDA	~ 85	No	-	No	i	140
13	HSC	351	213	7	LDA	91.0	No	9	No	i	141
14	HSC	657	249	34	ANN	92.9	No	8	No	i	85
15	HSC	667	363	7	LDA	92.9	No	8	No	i	142
16	HSC	2030	1006	8	LDA	90.4	No	8	No	i	143
17	HSC	4346	520	62	kNN	95	No	8	No	i	144
18	HSC	611	230	36	SVM	100	No	8	No	i	96
19	HSC	611	230	36	kNN	97.7	No	8	No	i	96
20	HSC	611	230	36	DT	98.6	No	8	No	i	96
21	HSC	1157 6	4208	4	ANN	97.0	<i>St</i>	>10	Yes	i	145
22	ATD	1209 6	5437	4	LDA	90.0	<i>Myc</i>	>10	Yes	i	146
23	HSC	7517	2066	21	kNN	99.3	MRSA	>10	Yes	i	148
24	HSC	7517	2066	21	SVM	92.9	MRSA	>10	Yes	i	148
25	HSC	3783 4	13203	5	LDA	95.0	No	>10	Yes	i	147
26	HSC	2230	1051	3	LDA	86.3 1	No	>10	No	i	149
27	HSC	3018 1	12474	6	LDA	90.0	<i>FN/PI</i>	>10	Yes	i	86
28	HSC	5468 2	19912	6	ANN	90.0	<i>PA</i>	>10	Yes	i	86
29	Peptide	3592	1404	4	LDA	96.0	MBS	>10	Yes	i	150
30	Peptide	2488	922	6	LDA	90.0	<i>G+</i>	>10	Yes	i	151
31	HSC	3500	628	4	ISE	94.6	MBS	>10	Yes	i	152
32	HSC	7456 7	8724	6	SOM	75.5	<i>EC</i>	>10	Yes	i	134
33	HSC	8360 5	10030	6	LDA	88.6	MBS	>10	Yes	i	153
34	HSC	2335	1760	-	MPNN	89.6	<i>E. coli</i>	>10	Yes	i	92

Notes. ^a Cmpd Type: Compound type. HSC = Heterogeneous Series of compounds, anti-TB drug = antituberculosis drugs. ^b n: Total number of cases in training and/or validation series, n_{Act.}: active cases and Var. = Variables in the model. ^c Technique: LDA = Linear discriminant analysis, ANN= artificial

neural network, BLR=Binary logistic regression, BN=Bayesian Network, DT=Decision tree, ISE=Iterative stochastic elimination, SOM=Self-organizing map (Kohonen), MPNN=Message Passing Neural Network, RF=Random Forest., KNN=K-Nearest-Neighbor. ^d Acc(%): Accuracy of training series. ^e Multi Species: Multiple bacterial strain (MBS), St=*Streptococcus spp*, Myc=*Mycobacterium spp*, EC=*Escherichia coli*, FN=*Fusobacterium necrophorum*, PI=*Prevotella intermedia*, PA=*Pseudomonas aeruginosa*, MRSA=Methicillin-resistant *Staphylococcus aureus*. G+=Gram + bacteria. ^f Drug Family: Only largely represented families were considered. ^gMO = Multi Output: multi-output models are those able to predict more than one type of biological activity (MIC, IC₅₀, MBC, etc.). ^h Val. =Validation methods: i) external validation series, ii) leave-30%-out cross validation, and iii) 100-times-averaged re-substitution technique. Furthermore, note that methods ii and iii are cross-validation methods.

Deep-learning algorithms are revolutionizing fields such as antibiotic discovery. As stated previously, deep neural networks are beneficial for automatically generating higher-order functions for ML models. Given the issues with microbial resistance and the limitations in antibacterial biological activity research, it is surprising that they have not yet been widely used for modeling antibacterial activity. Nonetheless, Stokes *et al.*⁹² discovered one drug (Halicin) with bactericidal efficacy against a broad phylogenetic spectrum of pathogens, including *M. TB* and Carbapenem-resistant *Enterobacteriaceae*. The drug also treated *C. difficile* and *A. baumannii* infections in mice. Indeed, it is the first antibiotic found by AI. This study shows how DL may be used to discover new ADs with specific structural features.

Finally, we can recognize that the application of MLTs may be effective for dealing with nonlinear data, perceiving patterns, and providing predictions that traditional classification or regression algorithms are unable to. MLT could, in some situations, perform better than other methods or adopt a new strategy for exploring other spaces in comparison with conventional methods.

In summary, of the techniques applied, LDA stands out; it was used by 17 out of 24 studies of those reported in **Table 1.7**. Among the non-linear techniques, ANN stands out and appears in Models 2,¹³⁵ 3,¹³⁶ 7,⁸⁴ 14,⁸⁵ 21,¹⁴⁵ and 28.⁸⁶ These were the most popular ML methods used for antibacterial activity prediction. In contrast, few studies successfully implemented other in silico techniques, for example BLR,^{84, 123} SVM,^{96, 148} kNN,^{96, 144, 148} DT,⁹⁶ and Iterative Stochastic Elimination (ISE).¹⁵² Meanwhile, powerful and high-performance MLTs that had not been applied to antibacterial before were introduced: Kohonen-based SOM¹³⁴ and Message Passing Neural Network (MPNN),⁹² respectively.

6. REFERENCES

1. Gradmann, C. Magic bullets and moving targets: Antibiotic resistance and experimental chemotherapy, 1900-1940. *Dynamis*. **2011**, 31 (2), 305-321, Review. DOI: 10.4321/S0211-95362011000200003 Scopus.
2. Geddes, A. 80th Anniversary of the discovery of penicillin. An appreciation of Sir Alexander Fleming. *International Journal of Antimicrobial Agents*. **2008**, 32 (5), 373, Editorial. DOI: 10.1016/j.ijantimicag.2008.06.001 Scopus.
3. Aminov, R. I. A brief history of the antibiotic era: Lessons learned and challenges for the future. *Frontiers in Microbiology*. **2010**, 1 (DEC), Article. DOI: 10.3389/fmicb.2010.00134 Scopus.
4. Waksman, S. A.; Schatz, A.; Reynolds, D. M. Production of antibiotic substances by actinomycetes. *Annals of the New York Academy of Sciences*. **2010**, 1213, 112-124. DOI: 10.1111/j.1749-6632.2010.05861.x From NLM.
5. Hutchings, M. I.; Truman, A. W.; Wilkinson, B. Antibiotics: past, present and future. *Current Opinion in Microbiology*. **2019**, 51, 72-80. DOI: 10.1016/j.mib.2019.10.008.

6. Katz, L.; Baltz, R. H. Natural product discovery: past, present, and future. *Journal of industrial microbiology & biotechnology*. **2016**, *43* (2-3), 155-176. DOI: 10.1007/s10295-015-1723-5 From NLM.
7. Durand, G. A.; Raoult, D.; Dubourg, G. Antibiotic discovery: history, methods and perspectives. *International Journal of Antimicrobial Agents*. **2019**, *53* (4), 371-382, Review. DOI: 10.1016/j.ijantimicag.2018.11.010 Scopus.
8. WHO. *Antibacterial Agents in Clinical Development – An Analysis of the Antibacterial Clinical Development Pipeline, including Mycobacterium tuberculosis*; Geneva, 2017.
9. Palumbi, S. R. Humans as the world's greatest evolutionary force. *Science*. **2001**, *293* (5536), 1786-1790, Review. DOI: 10.1126/science.293.5536.1786 Scopus.
10. Lewis, K. Platforms for antibiotic discovery. *Nature reviews. Drug discovery*. **2013**, *12* (5), 371-387. DOI: 10.1038/nrd3975 From NLM.
11. Ronald, A. R.; Turck, M.; Petersdorf, R. G. A critical evaluation of nalidixic acid in urinary-tract infections. *The New England journal of medicine*. **1966**, *275* (20), 1081-1089, Article. DOI: 10.1056/NEJM196611172752001 Scopus.
12. Chen, C.; Chen, Y.; Wu, P.; Chen, B. Update on new medicinal applications of gentamicin: Evidence-based review. *Journal of the Formosan Medical Association*. **2014**, *113* (2), 72-82. DOI: 10.1016/j.jfma.2013.10.002.
13. Seginkova, Z.; Krcmery, V.; Knothe, H. Ceftazidime resistance in *Pseudomonas aeruginosa*: Transduction by a wild-type phage. *Journal of Infectious Diseases*. **1986**, *154* (6), 1049-1050, Letter. DOI: 10.1093/infdis/154.6.1049 Scopus.
14. Parry, M. F. Aztreonam susceptibility testing. A retrospective analysis. *The American journal of medicine*. **1990**, *88* (3c), 7S-11S; discussion 38S-42S. DOI: 10.1016/0002-9343(90)90080-w From NLM.
15. Tsiodras, S.; Gold, H. S.; Sakoulas, G.; Eliopoulos, G. M.; Wennersten, C.; Venkataraman, L.; Moellering Jr, R. C.; Ferraro, M. J. Linezolid resistance in a clinical isolate of *Staphylococcus aureus*. *Lancet*. **2001**, *358* (9277), 207-208, Article. DOI: 10.1016/S0140-6736(01)05410-1 Scopus.
16. Andries, K.; Villellas, C.; Coeck, N.; Thys, K.; Gevers, T.; Vranckx, L.; Lounis, N.; De Jong, B. C.; Koul, A. Acquired resistance of *Mycobacterium tuberculosis* to bedaquiline. *PLoS ONE*. **2014**, *9* (7), Article. DOI: 10.1371/journal.pone.0102135 Scopus.
17. Lewis, J. S.; Owens, A.; Cadena, J.; Sabol, K.; Patterson, J. E.; Jorgensen, J. H. Emergence of daptomycin resistance in *Enterococcus faecium* during daptomycin therapy. *Antimicrob Agents Chemother*. **2005**, *49*, 1664-1665, Article. Scopus.
18. Gentry, D. R.; McCloskey, L.; Gwynn, M. N.; Rittenhouse, S. F.; Scangarella, N.; Shawar, R.; Holmes, D. J. Genetic characterization of Vga ABC proteins conferring reduced susceptibility to pleuromutilins in *Staphylococcus aureus*. *Antimicrobial Agents and Chemotherapy*. **2008**, *52* (12), 4507-4509, Article. DOI: 10.1128/AAC.00915-08 Scopus.
19. Goldstein, E. J. C.; Citron, D. M.; Sears, P.; Babakhani, F.; Sambol, S. P.; Gerding, D. N. Comparative susceptibilities to fidaxomicin (OPT-80) of isolates collected at baseline, recurrence, and failure from patients in two phase III trials of fidaxomicin against *Clostridium difficile* infection. *Antimicrobial Agents and Chemotherapy*. **2011**, *55* (11), 5194-5199, Article. DOI: 10.1128/AAC.00625-11 Scopus.
20. Long, S. W.; Olsen, R. J.; Mehta, S. C.; Palzkill, T.; Cernoch, P. L.; Perez, K. K.; Musick, W. L.; Rosato, A. E.; Musser, J. M. PBP2a mutations causing high-level ceftaroline resistance in clinical methicillin-resistant *Staphylococcus aureus* isolates. *Antimicrobial Agents and Chemotherapy*. **2014**, *58* (11), 6668-6674, Article. DOI: 10.1128/AAC.03622-14 Scopus.

21. FDA. *FDA approves new drug for treatment-resistant forms of tuberculosis that affects the lungs*. 2019. <http://www.fda.gov/news-events/press-announcements/fda-approves-new-drug-treatment-resistant-forms-tuberculosis-affects-lungs> (accessed 2021 September 17).
22. Serafim, M. S. Á. M.; Kronenberger, T.; Oliveira, P. R.; Poso, A.; Honório, K. M.; Mota, B. E. F.; Maltarollo, V. G. The application of machine learning techniques to innovative antibacterial discovery and development. *Expert Opinion on Drug Discovery*. **2020**, *15* (10), 1165-1180, Review. DOI: 10.1080/17460441.2020.1776696 Scopus.
23. Cox, G.; Sieron, A.; King, A. M.; De Pascale, G.; Pawlowski, A. C.; Koteva, K.; Wright, G. D. A Common Platform for Antibiotic Dereplication and Adjuvant Discovery. *Cell Chemical Biology*. **2017**, *24* (1), 98-109, Article. DOI: 10.1016/j.chembiol.2016.11.011 Scopus.
24. Coates, A. R.; Halls, G.; Hu, Y. Novel classes of antibiotics or more of the same? *Br J Pharmacol*. **2011**, *163* (1), 184-194. DOI: 10.1111/j.1476-5381.2011.01250.x From NLM.
25. Schneider, T.; Sahl, H.-G. An oldie but a goodie – cell wall biosynthesis as antibiotic target pathway. *International Journal of Medical Microbiology*. **2010**, *300* (2), 161-169. DOI: 10.1016/j.ijmm.2009.10.005.
26. Walsh, C.; Wencewicz, T. *Antibiotics: challenges, mechanisms, opportunities*; John Wiley & Sons, 2020.
27. Abrahams, K. A.; Besra, G. S. Mycobacterial cell wall biosynthesis: a multifaceted antibiotic target. *Parasitology*. **2018**, *145* (2), 116-133. DOI: 10.1017/S0031182016002377 From Cambridge University Press Cambridge Core.
28. Choi, U.; Lee, C.-R. Distinct Roles of Outer Membrane Porins in Antibiotic Resistance and Membrane Integrity in Escherichia coli. **2019**, *10* (953), Original Research. DOI: 10.3389/fmicb.2019.00953.
29. Hancock, R. E. W. Peptide antibiotics. *The Lancet*. **1997**, *349* (9049), 418-422. DOI: 10.1016/S0140-6736(97)80051-7.
30. Wilson, D. N. Ribosome-targeting antibiotics and mechanisms of bacterial resistance. *Nature Reviews Microbiology*. **2014**, *12* (1), 35-48. DOI: 10.1038/nrmicro3155.
31. Aviner, R. The science of puromycin: From studies of ribosome function to applications in biotechnology. *Computational and Structural Biotechnology Journal*. **2020**, *18*, 1074-1083. DOI: 10.1016/j.csbj.2020.04.014.
32. Fernandes, P. Fusidic Acid: A Bacterial Elongation Factor Inhibitor for the Oral Treatment of Acute and Chronic Staphylococcal Infections. *J Cold Spring Harbor Perspectives in Medicine*. **2016**, *6* (1). DOI: 10.1101/cshperspect.a025437.
33. Kapoor, G.; Saigal, S.; Elongavan, A. Action and resistance mechanisms of antibiotics: A guide for clinicians. *J Anaesthesiol Clin Pharmacol*. **2017**, *33* (3), 300-305. DOI: 10.4103/joacp.JOACP_349_15 PubMed.
34. Walsh, C. Where will new antibiotics come from? *Nature Reviews Microbiology*. **2003**, *1* (1), 65-70. DOI: 10.1038/nrmicro727.
35. Gewirtz, D. A critical evaluation of the mechanisms of action proposed for the antitumor effects of the anthracycline antibiotics adriamycin and daunorubicin. *Biochemical Pharmacology*. **1999**, *57* (7), 727-741. DOI: 10.1016/S0006-2952(98)00307-4.
36. Anderson, R.; Groundwater, P. W.; Todd, A.; Worsley, A. *Antibacterial agents: chemistry, mode of action, mechanisms of resistance and clinical applications*; John Wiley & Sons, 2012.
37. Silver, L. L. Appropriate Targets for Antibacterial Drugs. *Cold Spring Harb Perspect Med*. **2016**, *6* (12), a030239. DOI: 10.1101/cshperspect.a030239 PubMed.

38. Long, K. S.; Vester, B. Resistance to linezolid caused by modifications at its binding site on the ribosome. *Antimicrob Agents Chemother.* **2012**, *56* (2), 603-612. DOI: 10.1128/aac.05702-11 From NLM.
39. Schwarz, S.; Kehrenberg, C.; Doublet, B.; Cloeckaert, A. Molecular basis of bacterial resistance to chloramphenicol and florfenicol. *FEMS microbiology reviews.* **2004**, *28* (5), 519-542. DOI: 10.1016/j.femsre.2004.04.001 From NLM.
40. Long, K. S.; Poehlsgaard, J.; Kehrenberg, C.; Schwarz, S.; Vester, B. The Cfr rRNA Methyltransferase Confers Resistance to Phenicol, Lincosamides, Oxazolidinones, Pleuromutilins, and Streptogramin A Antibiotics. *Antimicrob Agents Chemother.* **2006**, *50* (7), 2500-2505. DOI: doi:10.1128/AAC.00131-06.
41. Munita, J. M.; Arias, C. A. Mechanisms of Antibiotic Resistance. *Microbiology spectrum.* **2016**, *4* (2), 10.1128/microbiolspec.VMBF-0016-2015. DOI: 10.1128/microbiolspec.VMBF-0016-2015 PubMed.
42. Olaitan, A. O.; Morand, S.; Rolain, J.-M. Mechanisms of polymyxin resistance: acquired and intrinsic resistance in bacteria. *Frontiers in microbiology.* **2014**, *5*, 643-643. DOI: 10.3389/fmicb.2014.00643 PubMed.
43. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. BLAST+: architecture and applications. *BMC Bioinformatics.* **2009**, *10* (1), 421. DOI: 10.1186/1471-2105-10-421.
44. Benson, D. A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. GenBank. *Nucleic Acids Res.* **2013**, *41* (Database issue), D36-42. DOI: 10.1093/nar/gks1195 From NLM.
45. Liu, B.; Pop, M. ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Res.* **2009**, *37* (Database issue), D443-447. DOI: 10.1093/nar/gkn656 From NLM.
46. Bortolaia, V.; Kaas, R. S.; Ruppe, E.; Roberts, M. C.; Schwarz, S.; Cattoir, V.; Philippon, A.; Allesoe, R. L.; Rebelo, A. R.; Florensa, A. F.; et al. ResFinder 4.0 for predictions of phenotypes from genotypes. *Journal of Antimicrobial Chemotherapy.* **2020**, *75* (12), 3491-3500. DOI: 10.1093/jac/dkaa345 %J Journal of Antimicrobial Chemotherapy (accessed 12/2/2021).
47. Gibson, M. K.; Forsberg, K. J.; Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **2015**, *9* (1), 207-216. DOI: 10.1038/ismej.2014.106 PubMed.
48. Consortium, T. U. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research.* **2020**, *49* (D1), D480-D489. DOI: 10.1093/nar/gkaa1100 %J Nucleic Acids Research (accessed 12/2/2021).
49. Alcock, B. P.; Raphenya, A. R.; Lau, T. T. Y.; Tsang, K. K.; Bouchard, M.; Edalatmand, A.; Huynh, W.; Nguyen, A. V.; Cheng, A. A.; Liu, S.; et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic acids research.* **2020**, *48* (D1), D517-d525. DOI: 10.1093/nar/gkz935 From NLM.
50. Gordon, N. C.; Price, J. R.; Cole, K.; Everitt, R.; Morgan, M.; Finney, J.; Kearns, A. M.; Pichon, B.; Young, B.; Wilson, D. J.; et al. Prediction of staphylococcus aureus antimicrobial resistance by whole-genome sequencing. *Journal of Clinical Microbiology.* **2014**, *52* (4), 1182-1191, Article. DOI: 10.1128/JCM.03117-13 Scopus.
51. Leopold, S. R.; Goering, R. V.; Witten, A.; Harmsen, D.; Mellmann, A. Bacterial whole-genome sequencing revisited: Portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. *Journal of Clinical Microbiology.* **2014**, *52* (7), 2365-2370, Article. DOI: 10.1128/JCM.00262-14 Scopus.
52. Bradley, P.; Gordon, N. C.; Walker, T. M.; Dunn, L.; Heys, S.; Huang, B.; Earle, S.; Pankhurst, L. J.; Anson, L.; De Cesare, M.; et al. Rapid antibiotic-resistance predictions

- from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nature Communications*. **2015**, 6, Article. DOI: 10.1038/ncomms10063 Scopus.
53. Gaulton, A.; Hersey, A.; Nowotka, M. L.; Patricia Bento, A.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; et al. The ChEMBL database in 2017. *Nucleic Acids Research*. **2017**, 45 (D1), D945-D954, Article. DOI: 10.1093/nar/gkw1074 Scopus.
 54. Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*. **2016**, 44 (D1), D1045-D1053, Article. DOI: 10.1093/nar/gkv1072 Scopus.
 55. Farrell, L. J.; Lo, R.; Wanford, J. J.; Jenkins, A.; Maxwell, A.; Piddock, L. J. V. Revitalizing the drug pipeline: AntibioticDB, an open access database to aid antibacterial research and development. *Journal of Antimicrobial Chemotherapy*. **2018**, 73 (9), 2284-2297, Article. DOI: 10.1093/jac/dky208 Scopus.
 56. Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*. **2015**, 55 (11), 2324-2337. DOI: 10.1021/acs.jcim.5b00559.
 57. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research*. **2018**, 47 (D1), D1102-D1109. DOI: 10.1093/nar/gky1033 %J Nucleic Acids Research (accessed 9/21/2021).
 58. Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*. **2018**, 46 (D1), D1074-D1082, Article. DOI: 10.1093/nar/gkx1037 Scopus.
 59. Corsello, S. M.; Bittker, J. A.; Liu, Z.; Gould, J.; McCarren, P.; Hirschman, J. E.; Johnston, S. E.; Vrcic, A.; Wong, B.; Khan, M.; et al. The Drug Repurposing Hub: A next-generation drug library and information resource. *Nature Medicine*. **2017**, 23 (4), 405-408, Letter. DOI: 10.1038/nm.4306 Scopus.
 60. Tse, T.; Williams, R. J.; Zarin, D. A. Update on Registration of Clinical Trials in ClinicalTrials.gov. *CHEST*. **2009**, 136 (1), 304-305. DOI: 10.1378/chest.09-1219 (accessed 2021/11/26).
 61. Cihoric, N.; Tsikkinis, A.; Miguelez, C. G.; Strnad, V.; Soldatovic, I.; Ghadjar, P.; Jeremic, B.; Dal Pra, A.; Aebbersold, D. M.; Lössl, K. Portfolio of prospective clinical trials including brachytherapy: an analysis of the ClinicalTrials.gov database. *Radiation Oncology*. **2016**, 11 (1), 48. DOI: 10.1186/s13014-016-0624-8.
 62. Palma, D. A.; Zietman, A. Clinical Trial Registration: A Mandatory Requirement for Publication in the Red Journal. *International Journal of Radiation Oncology, Biology, Physics*. **2015**, 91 (4), 685-686. DOI: 10.1016/j.ijrobp.2014.12.002 (accessed 2021/11/26).
 63. Deinsberger, J.; Reisinger, D.; Weber, B. Global trends in clinical trials involving pluripotent stem cells: a systematic multi-database analysis. *npj Regenerative Medicine*. **2020**, 5 (1), 15. DOI: 10.1038/s41536-020-00100-4.
 64. Bliziotis, I. A.; Ntziora, F.; Lawrence, K. R.; Falagas, M. E. Rifampin as adjuvant treatment of Gram-positive bacterial infections: a systematic review of comparative clinical trials. *European Journal of Clinical Microbiology & Infectious Diseases*. **2007**, 26 (12), 849. DOI: 10.1007/s10096-007-0378-1.

65. Long, T. E.; Williams, J. T. Cephalosporins currently in early clinical trials for the treatment of bacterial infections. *Expert Opinion on Investigational Drugs*. **2014**, *23* (10), 1375-1387. DOI: 10.1517/13543784.2014.930127.
66. Talbot, G. H.; Powers, J. H.; Fleming, T. R.; Siuciak, J. A.; Bradley, J.; Boucher, H.; Team, o. b. o. t. C.-A. P. Progress on Developing Endpoints for Registrational Clinical Trials of Community-Acquired Bacterial Pneumonia and Acute Bacterial Skin and Skin Structure Infections: Update From the Biomarkers Consortium of the Foundation for the National Institutes of Health. *Clinical Infectious Diseases*. **2012**, *55* (8), 1114-1121. DOI: 10.1093/cid/cis566 %J Clinical Infectious Diseases (accessed 11/25/2021).
67. Corey, G. R.; Stryjewski, M. E. New Rules for Clinical Trials of Patients With Acute Bacterial Skin and Skin-Structure Infections: Do Not Let the Perfect Be the Enemy of the Good. *Clinical Infectious Diseases*. **2011**, *52* (suppl_7), S469-S476. DOI: 10.1093/cid/cir162 %J Clinical Infectious Diseases (accessed 11/25/2021).
68. Shepshelovich, D.; Yelin, D.; Gafter-Gvili, A.; Goldman, S.; Avni, T.; Yahav, D. Comparison of reporting phase III randomized controlled trials of antibiotic treatment for common bacterial infections in ClinicalTrials.gov and matched publications. *Clinical Microbiology and Infection*. **2018**, *24* (11), 1211.e1219-1211.e1214. DOI: <https://doi.org/10.1016/j.cmi.2018.02.010>.
69. USA National Library of Medicine. *ClinicalTrials.gov*. 2021. <https://clinicaltrials.gov/ct2/search/advanced> (accessed 2021 november 07).
70. ICTRP. *International Clinical Trials Registry Platform (ICTRP) Vol. 2021*. WHO, 2021. <https://www.who.int/ictrp/en/> (accessed 2021 November 22).
71. EMA. *EudraCT (European Union Drug Regulating Authorities Clinical Trials Database)*. 2021. <https://eudract.ema.europa.eu/> (accessed 2021 November 21).
72. Morgan, B.; Hejdenberg, J.; Kuleszewicz, K.; Armstrong, D.; Ziebland, S. Are some feasibility studies more feasible than others? A review of the outcomes of feasibility studies on the ISRCTN registry. *Pilot and Feasibility Studies*. **2021**, *7* (1), 195. DOI: 10.1186/s40814-021-00931-y.
73. Askie, L. M. Australian New Zealand Clinical Trials Registry: history and growth. *J Evid Based Med*. **2011**, *4* (3), 185-187. DOI: 10.1111/j.1756-5391.2011.01147.x.
74. Yamamoto, M.; Wakai, S.; Ito, M.; Okuyama, M.; Tsukioka, M. Center for Clinical Trials, Japan Medical Association and clinical trial registration: history, present situation, prospects and challenges. *Journal of the National Institute of Public Health*. **2015**, *64* (4), 322-327. CABDirect.
75. Niehaus, K. E.; Walker, T. M.; Crook, D. W.; Peto, T. E. A.; Clifton, D. A. Machine learning for the prediction of antibacterial susceptibility in Mycobacterium tuberculosis. In *2014 IEEE-EMBS International Conference on Biomedical and Health Informatics, BHI 2014*, 2014; pp 618-621. DOI: 10.1109/BHI.2014.6864440.
76. Zaengle-Barone, J. M.; Jackson, A. C.; Besse, D. M.; Becken, B.; Arshad, M.; Seed, P. C.; Franz, K. J. Copper Influences the Antibacterial Outcomes of a β -Lactamase-Activated Prochelator against Drug-Resistant Bacteria. *ACS Infectious Diseases*. **2018**, *4* (6), 1019-1029. DOI: 10.1021/acsinfecdis.8b00037.
77. Camacho, D. M.; Collins, K. M.; Powers, R. K.; Costello, J. C.; Collins, J. J. Next-Generation Machine Learning for Biological Networks. *Cell*. **2018**, *173* (7), 1581-1592, Review. DOI: 10.1016/j.cell.2018.05.015 Scopus.
78. Diéguez-Santana, K.; Casañola-Martin, G. M.; Green, J. R.; Rasulev, B.; González-Díaz, H. Predicting Metabolic Reaction Networks with Perturbation-Theory Machine Learning (PTML) Models. *Current Topics in Medicinal Chemistry*. **2021**, *21* (9), 819-827. DOI: 10.2174/1568026621666210331161144.

79. Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*. **2018**, *23* (8), 1538-1546, Review. DOI: 10.1016/j.drudis.2018.05.010 Scopus.
80. Xu, J.; Hagler, A. Chemoinformatics and drug discovery. *Molecules*. **2002**, *7* (8), 566-600, Review. Scopus.
81. Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Nature*. **1986**, *323* (6088), 533-536, Article. DOI: 10.1038/323533a0 Scopus.
82. Svozil, D.; Kvasnicka, V.; Pospichal, J. í. Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems*. **1997**, *39* (1), 43-62. DOI: 10.1016/S0169-7439(97)00061-0.
83. Nam, N.-H.; Nga, D.-V.; Hai, D. T.; Dieguez-Santana, K.; Marrero-Poncee, Y.; Castillo-Garrit, J. A.; Casanola-Martin, G. M.; Le-Thi-Thu, H. Learning from multiple classifier systems: Perspectives for improving decision making of QSAR models in medicinal chemistry. *Current topics in medicinal chemistry*. **2017**, *17* (30), 3269-3288.
84. Cronin, M. T. D.; Aptula, A. O.; Dearden, J. C.; Duffy, J. C.; Netzeva, T. I.; Patel, H.; Rowe, P. H.; Schultz, T. W.; Worth, A. P.; Voutzoulidis, K.; et al. Structure-based classification of antibacterial activity. *Journal of Chemical Information and Computer Sciences*. **2002**, *42* (4), 869-878. DOI: 10.1021/ci025501d.
85. Cherkasov, A. Inductive QSAR descriptors. Distinguishing compounds with antibacterial activity by artificial neural networks. *International Journal of Molecular Sciences*. **2005**, *6* (1-2), 63-86, Article. DOI: 10.3390/i6010063 Scopus.
86. Speck-Planche, A.; Cordeiro, M. N. D. S. Enabling virtual screening of potent and safer antimicrobial agents against noma: Mtk-QSBER model for simultaneous prediction of antibacterial activities and ADMET properties. *Mini-Reviews in Medicinal Chemistry*. **2015**, *15* (3), 194-202. DOI: 10.2174/138955751503150312120519.
87. Durrant, J. D.; Amaro, R. E. Machine-learning techniques applied to antibacterial drug discovery. *Chemical Biology and Drug Design*. **2015**, *85* (1), 14-21, Review. DOI: 10.1111/cbdd.12423 Scopus.
88. Bengio, Y.; Goodfellow, I. J.; Courville, A. *Deep Learning*; MIT Press; 2015.
89. Lipinski, C. F.; Maltarollo, V. G.; Oliveira, P. R.; Da Silva, A. B. F.; Honorio, K. M. Advances and Perspectives in Applying Deep Learning for Drug Design and Discovery. *Front Rob AI*. **2019**, *6*, Article. Scopus.
90. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature*. **2015**, *521* (7553), 436-444, Review. DOI: 10.1038/nature14539 Scopus.
91. Gawehn, E.; Hiss, J. A.; Schneider, G. Deep Learning in Drug Discovery. *Molecular Informatics*. **2016**, *35* (1), 3-14, Review. DOI: 10.1002/minf.201501008 Scopus.
92. Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackerman, Z.; et al. A Deep Learning Approach to Antibiotic Discovery. *Cell*. **2020**, *180* (4), 688-702.e613, Article. DOI: 10.1016/j.cell.2020.01.021 Scopus.
93. Breiman, L. Random forests. *Machine Learning*. **2001**, *45* (1), 5-32, Article. DOI: 10.1023/A:1010933404324 Scopus.
94. Quinlan, R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers, 1993.
95. Diéguez-Santana, K.; Rivera-Borroto, O. M.; Puris, A.; Pham-The, H.; Le-Thi-Thu, H.; Rasulev, B.; Casañola-Martin, G. M. Beyond Model Interpretability using LDA and Decision Trees for α -Amylase and α -Glucosidase Inhibitor Classification Studies. *Chemical Biology & Drug Design*. **2019**. DOI: 10.1111/cbdd.13518.

96. Xue, Y.; Yang, X. G.; Chen, D.; Wang, M.; Chen, Y. Z. Prediction of antibacterial compounds by machine learning approaches. *Journal of Computational Chemistry*. **2009**, *30* (8), 1202-1211, Article. DOI: 10.1002/jcc.21148 Scopus.
97. El Zahed, S. S.; French, S.; Farha, M. A.; Kumar, G.; Brown, E. D. Physicochemical and Structural Parameters Contributing to the Antibacterial Activity and Efflux Susceptibility of Small-Molecule Inhibitors of Escherichia coli. *Antimicrobial agents and chemotherapy*. **2021**, *65* (4), e01925-01920. DOI: 10.1128/AAC.01925-20 PubMed.
98. Suay-Garcia, B.; Falcó, A.; Bueso-Bordils, J. I.; Anton-Fos, G. M.; Pérez-Gracia, M. T.; Alemán-López, P. A. Tree-Based QSAR Model for Drug Repurposing in the Discovery of New Antibacterial Compounds Against Escherichia coli. *Pharmaceuticals (Basel, Switzerland)*. **2020**, *13* (12). DOI: 10.3390/ph13120431 From NLM.
99. Polikar, R. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*. **2006**, *6* (3), 21-44, Review. DOI: 10.1109/MCAS.2006.1688199 Scopus.
100. Martinez-Muñoz, G.; Hernández-Lobato, D.; Suarez, A. An analysis of ensemble pruning techniques based on ordered aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **2009**, *31* (2), 245-259, Article. DOI: 10.1109/TPAMI.2008.78 Scopus.
101. Zhou, Z.-H. *Ensemble methods: foundations and algorithms*; Chapman and Hall/CRC Press, 2012.
102. Breiman, L. Bagging predictors. *Machine Learning*. **1996**, *24* (2), 123-140. DOI: 10.1007/BF00058655.
103. Hastie, T.; Tibshirani, R.; Friedman, J. H. *The elements of statistical learning: Data mining, inference, and prediction*; Springer open, 2008.
104. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*. **2003**, *43* (6), 1947-1958. DOI: 10.1021/ci034160g.
105. Freund, Y.; Schapire, R. E. Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning*. **1996**, 148-156, Article. Scopus.
106. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*. **2001**, *29* (5), 1189-1232, Article. DOI: 10.1214/aos/1013203451 Scopus.
107. Khaledian, E.; Broschat, S. L. Sequence-Based Discovery of Antibacterial Peptides Using Ensemble Gradient Boosting. *Proceedings*. **2020**, *66* (1), 6. DOI: 10.3390/proceedings2020066006.
108. Colkesen, I.; Sahin, E. K.; Kavzoglu, T. Susceptibility mapping of shallow landslides using kernel-based Gaussian process, support vector machines and logistic regression. *Journal of African Earth Sciences*. **2016**, *118*, 53-64. DOI: 10.1016/j.jafrearsci.2016.02.019.
109. Vapnik, V. *The nature of statistical learning theory*; Springer science & business media, 2013.
110. Ivanciuc, O. Applications of Support Vector Machines in Chemistry. In *Reviews in Computational Chemistry*, 2007; pp 291-400.
111. Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning*. **1995**, *20* (3), 273-297, journal article. DOI: 10.1023/a:1022627411411.
112. Stephenson, N.; Shane, E.; Chase, J.; Rowland, J.; Ries, D.; Justice, N.; Zhang, J.; Chan, L.; Cao, R. Survey of machine learning techniques in drug discovery. *Current Drug*

- Metabolism*. **2019**, *20* (3), 185-193, Review. DOI: 10.2174/1389200219666180820112457 Scopus.
113. Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*. **2001**, Article. Scopus.
 114. Kohavi, R.; Becker, B.; Sommerfield, D. Improving simple bayes. *Proceedings of the European Conference on Machine Learning*. **1997**, 78-87, Article. Scopus.
 115. Ekins, S.; Freundlich, J. S.; Reynolds, R. C. Are bigger data sets better for machine learning? Fusing single-point and dual-event dose response data for mycobacterium tuberculosis. *Journal of Chemical Information and Modeling*. **2014**, *54* (7), 2157-2165, Article. DOI: 10.1021/ci500264r Scopus.
 116. Singh, N.; Chaudhury, S.; Liu, R.; AbdulHameed, M. D. M.; Tawa, G.; Wallqvist, A. QSAR Classification Model for Antibacterial Compounds and Its Use in Virtual Screening. *Journal of Chemical Information and Modeling*. **2012**, *52* (10), 2559-2569. DOI: 10.1021/ci300336v.
 117. Diéguez-Santana, K.; González-Díaz, H. Towards Machine Learning Discovery of Dual Antibacterial Drug-Nanoparticle Systems. *Nanoscale*. **2021**, *13*, 17854-17870. DOI: 10.1039/D1NR04178A.
 118. Frank, E.; Witten, I. H. Generating accurate rule sets without global optimization. In *Fifteenth International Conference on Machine Learning*, San Francisco, CA, 1998; Morgan Kaufmann Publishers Inc: pp 144-151.
 119. Gaines, B. R.; Compton, P. Induction of ripple-down rules applied to modeling large databases. *Journal of Intelligent Information Systems*. **1995**, *5* (3), 211-228.
 120. Hühn, J.; Hüllermeier, E. FURIA: an algorithm for unordered fuzzy rule induction. *Data Mining Knowledge Discovery*. **2009**, *19* (3), 293-319, journal article. DOI: 10.1007/s10618-009-0131-8.
 121. Pham-The, H.; Nam, N. H.; Nga, D. V.; Hai, D. T.; Dieguez-Santana, K.; Marrero-Poncee, Y.; Castillo-Garit, J. A.; Casanola-Martin, G. M.; Le-Thi-Thu, H. Learning from Multiple Classifier Systems: Perspectives for Improving Decision Making of QSAR Models in Medicinal Chemistry. *Curr Top Med Chem*. **2018**, *17* (30), 3269-3288. DOI: 10.2174/1568026618666171212111018 From NLM.
 122. Dieguez-Santana, K.; Pham-The, H.; Rivera-Borroto, O. M.; Puris, A.; Le-Thi-Thu, H.; Casanola-Martin, G. M. A Two QSAR Way for Antidiabetic Agents Targeting Using α -Amylase and α -Glucosidase Inhibitors: Model Parameters Settings in Artificial Intelligence Techniques. *Letters in Drug Design & Discovery*. **2017**, *14* (8), 862-868. DOI: 10.2174/1570180814666161128121142.
 123. Aptula, A. O.; Kühne, R.; Ebert, R. U.; Cronin, M. T. D.; Netzeva, T. I.; Schüürmann, G. Modeling discrimination between antibacterial and non-antibacterial activity based on 3D molecular descriptors. *QSAR and Combinatorial Science*. **2003**, *22* (1), 113-128, Conference Paper. DOI: 10.1002/qsar.200390001 Scopus.
 124. Peterson, L. E. K-nearest neighbor. *Scholarpedia*. **2009**, *4* (2), 1883. DOI: 10.4249/scholarpedia.1883.
 125. Fix, E.; Hodges, J. L. *Discriminatory analysis: Non-parametric discrimination*; USAF School of Aviation Medicine, 1951.
 126. Ding, H.; Takigawa, I.; Mamitsuka, H.; Zhu, S. Similarity-based machine learning methods for predicting drug-target interactions: A brief review. *Briefings in Bioinformatics*. **2013**, *15* (5), 734-747, Article. DOI: 10.1093/bib/bbt056 Scopus.
 127. Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*. **2000**, *16* (5), 412-424, Review. Scopus.

128. Cui, G.; Fang, C.; Han, K. Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinformatics*. **2012**, *13* (7), S5. DOI: 10.1186/1471-2105-13-S7-S5.
129. Dyer, M. D.; Murali, T. M.; Sobral, B. W. Supervised learning and prediction of physical interactions between human and HIV proteins. *Infect Genet Evol*. **2011**, *11* (5), 917-923. DOI: 10.1016/j.meegid.2011.02.022 PubMed.
130. Coelho, E. D.; Arrais, J. P.; Matos, S.; Pereira, C.; Rosa, N.; Correia, M. J.; Barros, M.; Oliveira, J. L. Computational prediction of the human-microbial oral interactome. *BMC Systems Biology*. **2014**, *8* (1), 24. DOI: 10.1186/1752-0509-8-24.
131. Mei, S.; Zhu, H. AdaBoost Based Multi-Instance Transfer Learning for Predicting Proteome-Wide Interactions between Salmonella and Human Proteins. *PLOS ONE*. **2014**, *9* (10), e110488. DOI: 10.1371/journal.pone.0110488.
132. Macesic, N.; Polubriaginof, F.; Tatonetti, N. P. Machine learning: Novel bioinformatics approaches for combating antimicrobial resistance. *Current Opinion in Infectious Diseases*. **2017**, *30* (6), 511-517, Review. DOI: 10.1097/QCO.0000000000000406 Scopus.
133. Nourani, E.; Khunjush, F.; Durmuş, S. Computational approaches for prediction of pathogen-host protein-protein interactions. **2015**, *6* (94), Review. DOI: 10.3389/fmicb.2015.00094.
134. Ivanenkov, Y. A.; Zhavoronkov, A.; Yamidanov, R. S.; Osterman, I. A.; Sergiev, P. V.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Veselov, M. S.; Ayginin, A. A.; et al. Identification of novel antibacterials using machine-learning techniques. *Frontiers in Pharmacology*. **2019**, *10* (JULY), Article. DOI: 10.3389/fphar.2019.00913 Scopus.
135. García-Domenech, R.; De Julián-Ortiz, J. V. Antimicrobial activity characterization in a heterogeneous group of compounds. *Journal of Chemical Information and Computer Sciences*. **1998**, *38* (3), 445-449. DOI: 10.1021/ci9702454.
136. Tomás-Vert, F.; Pérez-Giménez, F.; Salabert-Salvador, M. T.; García-March, F. J.; Jaén-Oltra, J. Artificial neural network applied to the discrimination of antibacterial activity by topological methods. *Journal of Molecular Structure: THEOCHEM*. **2000**, *504* (1-3), 249-259, Article. DOI: 10.1016/S0166-1280(00)00366-3 Scopus.
137. Mishra, R. K.; Garcia-Domenech, R.; Galvez, J. Getting discriminant functions of antibacterial activity from physicochemical and topological parameters. *Journal of Chemical Information and Computer Sciences*. **2001**, *41* (2), 387-393, Article. DOI: 10.1021/ci000303c Scopus.
138. Mut-Ronda, S.; Salabert-Salvador, M. T.; Duart, M. J.; Antón-Fos, G. M. Search compounds with antimicrobial activity by applying molecular topology to selected quinolones. *Bioorganic and Medicinal Chemistry Letters*. **2003**, *13* (16), 2699-2702. DOI: 10.1016/S0960-894X(03)00544-4.
139. Murcia-Soler, M.; Pérez-Giménez, F.; García-March, F. J.; Salabert-Salvador, M. T.; Díaz-Villanueva, W.; Medina-Casamayor, P. Discrimination and selection of new potential antibacterial compounds using simple topological descriptors. *Journal of Molecular Graphics and Modelling*. **2003**, *21* (5), 375-390. DOI: 10.1016/S1093-3263(02)00184-5.
140. Murcia-Soler, M.; Pérez-Giménez, F.; García-March, F. J.; Salabert-Salvador, M. T.; Díaz-Villanueva, W.; Castro-Bleda, M. J.; Villanueva-Pareja, A. Artificial neural networks and linear discriminant analysis: A valuable combination in the selection of new antibacterial compounds. *Journal of Chemical Information and Computer Sciences*. **2004**, *44* (3), 1031-1041. DOI: 10.1021/ci030340e.

141. Molina, E.; Díaz, H. G.; González, M. P.; Rodríguez, E.; Uriarte, E. Designing antibacterial compounds through a topological substructural approach. *Journal of Chemical Information and Computer Sciences*. **2004**, *44* (2), 515-521. DOI: 10.1021/ci0342019.
142. González-Díaz, H.; Torres-Gómez, L. A.; Guevara, Y.; Almeida, M. S.; Molina, R.; Castañedo, N.; Santana, L.; Uriarte, E. Markovian chemicals "in silico" design (MARCH-INSIDE), a promising approach for computer-aided molecular design III: 2.5D indices for the discovery of antibacterials. *Journal of molecular modeling*. **2005**, *11* (2), 116-123. DOI: 10.1007/s00894-004-0228-3.
143. Marrero-Ponce, Y.; Medina-Marrero, R.; Torrens, F.; Martinez, Y.; Romero-Zaldivar, V.; Castro, E. A. Atom, atom-type, and total nonstochastic and stochastic quadratic fingerprints: A promising approach for modeling of antibacterial activity. *Bioorg. Med. Chem.* **2005**, *13* (8), 2881-2899, Article. DOI: 10.1016/j.bmc.2005.02.015 Scopus.
144. Karakoc, E.; Cherkasov, A.; Sahinalp, S. C. Distance based algorithms for small biomolecule classification and structural similarity search. *Bioinformatics*. **2006**, *22* (14), e243-e251, Conference Paper. DOI: 10.1093/bioinformatics/btl259 Scopus.
145. Speck-Planche, A.; Kleandrova, V. V.; Cordeiro, M. N. D. S. Chemoinformatics for rational discovery of safe antibacterial drugs: Simultaneous predictions of biological activity against streptococci and toxicological profiles in laboratory animals. *Bioorg. Med. Chem.* **2013**, *21* (10), 2727-2732. DOI: 10.1016/j.bmc.2013.03.015.
146. Speck-Planche, A.; Kleandrova, V. V.; Cordeiro, M. N. D. S. New insights toward the discovery of antibacterial agents: Multi-tasking QSBER model for the simultaneous prediction of anti-tuberculosis activity and toxicological profiles of drugs. *European Journal of Pharmaceutical Sciences*. **2013**, *48* (4-5), 812-818. DOI: 10.1016/j.ejps.2013.01.011.
147. Speck-Planche, A.; Cordeiro, M. N. D. S. Simultaneous virtual prediction of anti-escherichia coli activities and admet profiles: A chemoinformatic complementary approach for high-throughput screening. *ACS combinatorial science*. **2014**, *16* (2), 78-84. DOI: 10.1021/co400115s.
148. Wang, L.; Le, X.; Li, L.; Ju, Y.; Lin, Z.; Gu, Q.; Xu, J. Discovering new agents active against methicillin-resistant *Staphylococcus aureus* with ligand-based approaches. *Journal of Chemical Information and Modeling*. **2014**, *54* (11), 3186-3197, Article. DOI: 10.1021/ci500253q Scopus.
149. Castillo-Garit, J. A.; Marrero-Ponce, Y.; Barigye, S. J.; Medina-Marrero, R.; Bernal, M. G.; De La Vega, J. M. G.; Torrens, F.; Arán, V. J.; Pérez-Giménez, F.; García-Domenech, R.; et al. In silico antibacterial activity modeling based on the TOMOCOMD-CARDD approach. *Journal of the Brazilian Chemical Society*. **2015**, *26* (6), 1218-1226, Article. DOI: 10.5935/0103-5053.20150087 Scopus.
150. Kleandrova, V. V.; Ruso, J. M.; Speck-Planche, A.; Dias Soeiro Cordeiro, M. N. Enabling the Discovery and Virtual Screening of Potent and Safe Antimicrobial Peptides. Simultaneous Prediction of Antibacterial Activity and Cytotoxicity. *ACS Combinatorial Science*. **2016**, *18* (8), 490-498. DOI: 10.1021/acscombsci.6b00063.
151. Speck-Planche, A.; Kleandrova, V. V.; Ruso, J. M.; Cordeiro, M. N. D. S. First Multitarget Chemo-Bioinformatic Model to Enable the Discovery of Antibacterial Peptides against Multiple Gram-Positive Pathogens. *Journal of Chemical Information and Modeling*. **2016**, *56* (3), 588-598. DOI: 10.1021/acs.jcim.5b00630.
152. Masalha, M.; Rayan, M.; Adawi, A.; Abdallah, Z.; Rayan, A. Capturing antibacterial natural products with in silico techniques. *Molecular Medicine Reports*. **2018**, *18* (1), 763-770, Article. DOI: 10.3892/mmr.2018.9027 Scopus.

153. Nocado-Mena, D.; Cornelio, C.; Camacho-Corona, M. D. R.; Garza-González, E.; Waksman De Torres, N.; Arrasate, S.; Sotomayor, N.; Lete, E.; González-Díaz, H. Modeling Antibacterial Activity with Machine Learning and Fusion of Chemical Structure Information with Microorganism Metabolic Networks. *Journal of Chemical Information and Modeling*. **2019**, *59* (3), 1109-1120, Article. DOI: 10.1021/acs.jcim.9b00034 Scopus.

CHAPTER 2. BACKGROUND AND OBJECTIVES

1. BACKGROUND

1.1 Practical problem

In the last decade, the overuse of broad-spectrum antibiotics has greatly increased bacterial resistance to conventional antibiotics.¹ This has required scientists to find rapid, accessible, and inexpensive methods to discover new drugs and molecular targets against infectious microorganisms. Literature reports published from 2002 to date on NP-functionalized AD systems collected show that among the types of nanoparticles used, metallic ones (Au, Ag, Zn and Cu) stand out. Other NPs, such as metal oxides (CuO, ZnO, and Fe₃O₄), salts (AgNO₃ and MoS₂), and other materials (Bi₂Te₃), are also shown. AgNPs are the most frequent in the studies consulted and the spherical shape is the most common, and the size of the NP ranges from 1.86-180 nm. The antibacterial drugs used are from several families, where β -lactams, and aminoglycosides, respectively, stand out. These families of antibiotics have broad-spectrum activity and are very frequent in the treatment of bacterial infections. The case of the former includes three subclasses: carbapenems (Imipenem and Meropenem), cephalosporins (Ceftazidime, Cefotaxime, and Cefuroxime), and penicillins (Ampicillin). In the latter, aminoglycosides (Gentamicin, Kanamycin, and Tobramycin) are those used for DADNP, which are also widely used antibiotics. Other families are fluoroquinolones (Ofloxacin and Ciprofloxacin), antimycobacterials (Rifampicin), amphenicols (Chloramphenicol), glycopeptides (Vancomycin), polypeptides (Polymyxin B), and tetracycline (Tetracycline and Tigecycline). Different strains of various microorganisms were used, such as *S. aureus*, *P. Aeruginosa*, *E. Faecium*, *E. Coli*, *E. faecalis*, *S. epidermidis*, *B. subtilis*, *A. Baumannii*, *S. enterica serovar Typhimurium*, *S. mutans*, *E. faecium*, *M. luteus*, and *K. pneumoniae*. Many of the strains are drug-resistant, e.g., MRSA, MDR and VRE, which shows that DADNP systems have been focused on the search for growth inhibitors of pathogens of great interest in the field of bacterial infections. Some researchers have shown that the potentiating effect is higher in antibiotic-resistant strains than in antibiotic-sensitive strains,^{2,3} which makes DADNPs able to positively influence multi-resistant strains. This is because NPs can affect cell-membrane and cell-wall integrity, favoring antibiotic action and leading to a “restored” susceptibility for some antibiotic-resistant strains.

Dual Nanoparticles and antibacterial drugs or NP-functionalized drugs have the intrinsic ability to penetrate bacterial cell membrane barriers and reach specific sites with a higher level of precision and stability than free antibiotic molecules.⁴ Many of these combinations have exerted synergistic or additive effects compared to the use of antibiotics in their molecular forms, which may contribute to tackling many resistant bacteria and supporting treatments in clinical infections.⁵ Most studies present synergistic or additive effects, as opposed to drug and nanoparticle used independently. This means that DADNP systems can increase the efficacy and speed of bacterial death.⁵ As mentioned by Zaidi *et al.*⁶, NP-conjugated ADs deliver antibiotics to specific areas through a variety of antibacterial mechanisms. For example, the interaction between AgNPs and different antibiotics promotes an increase in the release of Ag ions, which concomitantly enhances bacterial growth inhibition.⁷ On the other hand, some nanoparticles in combination treatments can depolarize the cell membrane, affecting permeability and allowing the antibiotics (e.g., KAN and CLO) to reach the ribosomes inside the cell and increase their antibacterial activity, which generates a synergistic or additive effect, compared to the drug and NP alone.^{2,8} Another advantage of DADNP systems is that they have high adjustability and a wide range of adaptability to cope with various scenarios, such as persister cells in macrophages and biofilm infections, and this integration could be a cost-effective solution. In that sense, the integrated design of nanoantibiotic systems can be

endowed with a variety of functionalities, for instance targeting capabilities, enhanced penetration and uptake, modification of the infectious microenvironment, and combination with other treatment techniques. Consequently, there is great potential for nanomaterials to demonstrate their ability to improve the therapeutic efficacy of antibiotics.⁹ When it comes to treating drug-resistant bacteria that produce numerous antibacterial mechanisms, NPs outperform the effects of single or multiple medications.

An improvement in antibacterial activity due to NP-AD combinations is expected which would allow the use of antibiotics that have fallen into disuse due to bacterial resistance problems, thus providing additional treatment possibilities in the healthcare, veterinary and agricultural sectors. Therefore, nanoantibiotics have a potential impact on social and economic issues, as they can help mitigate the current crisis due to antibiotic resistance. In another hand, all the AD alone, the NP, or the DADNP system have to interact with the microorganism. In this regard, understanding the metabolism of pathogens plays an important role. Metabolic networks are represented by the set of metabolic pathways, which in turn are a series of biochemical reactions in which the product (output) of one reaction serves as a substrate (input) for another reaction.¹⁰ In this sense, some studies by Barabási's group have demonstrated the influence of changes in Metabolic Networks (MNs) on the survivability of different microorganisms.¹¹

1.2 The methodological problem

The DADNP could be considered as complex systems. The study of complex systems in cheminformatics has been addressed by Herrera *et al.*¹², which developed a methodology to analyze a complex system that included chemical and pre-clinical data with epidemiological data, to carry out "pharmaco-epidemiological" predictions of AIDS prevalence in US counties, taking into account the social determinants and structure-activity relationship of anti-HIV compounds in pre-clinical trials. Another example is Santana *et al.*¹³ In this work, a working dataset of preclinical trials of vitamin release and cancer cotherapy drugs that included anticancer compounds and vitamins, or vitamin derivatives was analyzed. The trials considered multiple continuous variables (descriptors) and categorical variables (assay conditions for drugs, vitamins, and NPs) with varied assay cell organisms and other conditions.

From the perspective of computational modeling, the case study addressed by this thesis can be analyzed as a complex system. The incorporation of several systems with different conditions can be analyzed as an ML problem in discovering new ADs, with NP and MN applications of metabolic pathways at the same time. There are public databases such as ChEMBL with thousands of reports of preclinical assays of potential ADs.¹⁴⁻¹⁸, a growing number of experimental reports of NPs with antibacterial action, and a previous report of consensus MNs for multiple pathogenic bacteria by Jeong *et al.*¹¹, but most of the AD¹⁹⁻²³ and NP.²⁴⁻³⁰ ML models are not multi-labeled. This forms a complex system AD + NP + COAT + PROT + MN + EPIDEMIOLOGICAL NET. It could be analyzed as a whole or in parts (subsystem information additive model). These parts can be added gradually to see the robustness of the technique (block-wise approach). To perform the analysis as a whole there is not enough data. In the case of piecewise analysis, some parts have too much data and others too little. Therefore, as a solution to the problem, they can be decomposed into parts or sub-systems.

1.3 The solution to the methodological problem for other complex biomolecular systems

In order to solve this problem González-Díaz et al. created the NIFPTML strategy. NIFPTML is a multi-output, input-coded multi-label machine learning technique, to address this type of challenge. Networks Invariants (NI) + Information Fusion (IF) + Perturbation Theory (PT) + Machine Learning (ML) is the acronym for the NIFPTML algorithm. In the first phase of the NIFPTML algorithm we can use Complex network theory to study of biomolecular systems (drugs, protein, metabolic networks, etc.). Networks can be represented as graphs through sets of nodes and axes. An example is the molecular graph where the nodes and axes correspond to the atoms and chemical bonds of a drug molecule. Another example is the network of a protein where the nodes are amino acids and the axes the sequence and/or interaction/spatial proximity between the amino acids. Numerical parameters called Network Invariants (NI) can be extracted from these networks are used to quantify the structure of these systems. These parameters or numerical indices of networks or Networks (N) can be correlated with the biological properties of said systems by means of Artificial Intelligence (AI) and/or Automatic Learning or Machine Learning (ML) techniques.

On the other hand, in many problems of interest it is necessary to merge information about several of these systems at the same time. Techniques for Information Fusion (IF) from various sources allow obtaining an enriched data set. The Perturbation Theory (PT) operators allow to quantify the disturbances/deviations in the structural variables with respect to the expected values for different subsets of categorical variables. Finally, AI/ML methods make it possible to find predictive models for the biological properties of systems (drugs, proteins, etc.). Therefore, in this thesis we propose to use the NIFPTML strategy to study problems that involve one or more than one of these systems at the same time. This NIFPTML strategy combines all the phases mentioned above (NI + IF + PT + AI/ML). The first phase uses complex networks numerical parameters of networks or networks (N) to quantify the structure of the systems, the IF phase merges data from multiple systems from different sources, the PT phase processes the information, and the AI/ ML finds the predictive model. In the thesis we apply the NIFPTML strategy to several complex problems with different systems (drug, protein, metabolic network, nanoparticles, coating agents).

The additive NIFPTML approach is compatible with this type of analysis (AD + MN + NP + COAT). It allows working with multiple outputs, multi-conditions can be treated, and several problems or partial studies can be performed with the NIFPTML approach. In that sense, the available information calculated in previous studies of AD, NP, and MN mutant strain systems, to enhance the discovery of AD, NP, and MN applications of metabolic pathways, at the same time. In addition, the NIFPTML approach can test the reuse of known drugs as AD and/or co-therapy with different NPs and simulate DADNP activity on different bacteria (or MN).

1.4 Previous NIFPTML models for similar problems

Previously have been reported many works including different steps of the NIFPTML strategy. Da Costa et al.³¹, for example, used the NIFPTML method to predict drug-protein interactions (DPIs) for dopamine pathway target proteins, including only the PT + ML phases. They used the linear and models trained with multiple nonlinear methods (artificial neural networks (ANN), Random Forest, Deep Learning, etc.). Munteanu developed multiple models using the NIPTML strategy including the NI + PT + ML but the IF phase is missing (Ph.D. Cristian Robert Munteanu, UDC, 2013). for drug and molecular target discovery using computer engineering classification techniques and artificial intelligence.³² Similarly, new dual-function

multi-QSAR models have been generated for the prediction of drugs and their molecular targets from Topological Indices, for the search of new neuroprotective drugs useful in the treatment of Parkinson's and Alzheimer's diseases and/or new molecular targets for drugs (Ph.D. Manuel Quintín Escobar Cubiella, USC, 2012).³³

In another example, the doctoral thesis (Ph.D. Diana Herrera Ibatá, UDC, 2015) developed models with NIFPTML strategy including all the phases NI + IF + PT + ML. These NIFPTML models employed various Artificial Intelligence (AI) techniques to study the HIV problem, which allowed quantitatively relating chemical and pre-clinical data with epidemiological data, to carry out "pharmaco-epidemiological" predictions.¹² For instance, Santana et al. used a NIFPTML strategy including only the IFPTML phases, but the NI phase is not present. He analyzed >970,000 cases with the data of preclinical assays of new cancer cotherapy drug-vitamin release nanosystems, vitamins, and anticancer compounds from the ChEMBL database.¹³

1.5 Previous NIFPTML models for the present problem

Some NIFPTML computational approaches have been applied in the field of antibacterial drug studies. For example, Nocedo *et al.*³⁴ obtained a NIFPTML model that analyzed AD + MN. Ortega *et al.*³⁵ considered NP + MN subsystems in the NIFPTML model. Speck-Planche *et al.*³⁶ included an IFPTML model to analyze only NP vs NP pairs without considering NI. The green boxes in **Figure 2.1** show the different parts that included the mentioned works. These studies are some examples of the application of NIFPTML to develop cheminformatics studies. However, no NIFPTML computational study capable of quantitatively relating and fusing information from chemical and pre-clinical (ChEMBL) data to the mechanisms of metabolic reaction networks and nanoparticles with antibacterial activity has been reported. In the field of antibacterial drugs, it is of interest to understand the interactions between preclinical assays of AD activity, MN, and complex DADNP systems. This study represents an initial exploration of the drugs involved in reaction networks in Barabási's group. Next, the interplay of MNs and preclinical ChEMBL assays is explored with the aim of creating new models to predict new antibacterial compounds. Subsequently, nanoparticles with antibacterial activity fused with antibacterial compounds from the ChEMBL preclinical assay database are screened. Finally, we study metal/metal oxide nanosystems with antibacterial compounds, considering variations in the MN of the involved microorganisms to predict the biological activity of new antibacterial and new DADNP systems.

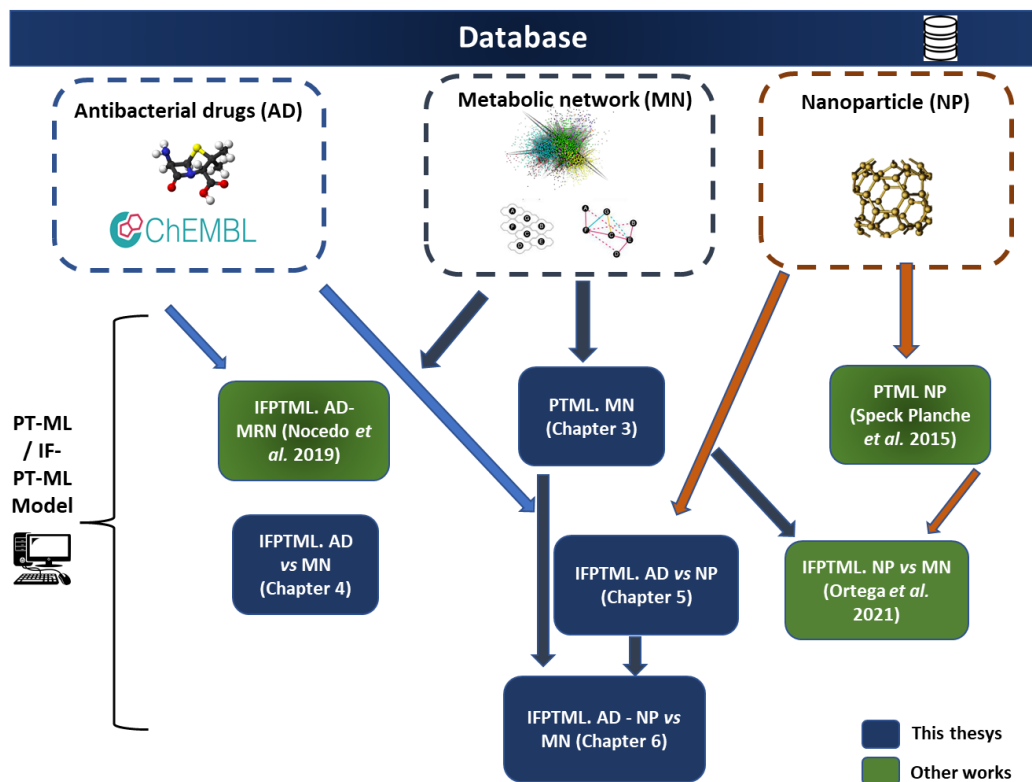


Figure 2.1. Relationship between the database, chapter, and other published works.

1.6 The focus of this thesis

There are no reports of NIFPTML models with AD, NP, and MN data, which, coupled with the fact that many reports of experimental data are inaccurate, and that AD activity, NP, and MN activity partially but not exactly match all biological activity parameters, parameter units, bacterial strains, etc. In addition, there is a very low number of experimental case reports of DADNP studies that are useful for model training and even fewer experimental case reports of MN changes in MDR bacterial strains due to AD and/or NP action. On the other hand, some risks include underestimating the synergies between subsystems using an additive approach and limiting the search to DADNP systems of known ADs, which remains a very large space for chemical exploration. Moreover, considering the additive nature of the present approach, limiting our search to known ADs is possibly less risky. However, in clinical trials, the application of ML is still limited, even though there are varied sources of information that can generate absolute and methodological data to support decision making and the deduction of risk failures in drug discovery. These approaches may hold promise in the face of the limitations of potential new antibiotic discoveries and the global threat posed by antibiotic resistant bacteria. For the time being, additive NIFPTML models may become a pragmatic solution for the time being when taking into account the increased abundance of experimental evidence for DADNP components in ADs and NPs alone.

2.OBJECTIVES

This thesis has two main objectives, one methodological and one practical.

2.1 Methodological objectives

1. Evaluating the feasibility of the subsystem information additive linear model with NIFPTML approach in this problem
2. Evaluating the robustness of the NIFPTML methodology using a subsystems information block-wise approach for this problem.

2.2 Practical objectives

1. To develop a computational model for analyzing a metabolite's connectivity (structure) in a query organism's metabolic reaction networks.
2. To develop, with linear and non-linear machine learning techniques, a "chemo-informatics-based perturbation theory (PT) and information fusion" methodology allows one to relate chemical and preclinical data with metabolic network data quantitatively.
3. To create a model that predicts the biological activity of antibacterial drugs functionalized with nanoparticle systems using the NIFPTML method.

2.3 Objective's development

The first task of this doctoral thesis was to research the state of the art on the main topics of this work: bacterial resistance, main antibiotics, protein targets, mechanisms of action, databases of preclinical and clinical trials, and other sources of information useful for computational modeling, machine learning techniques, and performance evaluation metrics algorithms applied in the field of antibacterial drugs. A summary description and the link between them are given in **Chapter 1**. Additionally, a paper (Paper 1, in Publication Listing, next section) was submitted to the Chemical Reviews Journal, which included the above topics and delved into a comprehensive review of the application of AI/ML in new antibacterial drug (AD) discovery, from classical drugs to dual antibacterial drug-nanoparticle systems (DADNP). This work also reviewed studies focused on nanoparticles used to target bacteria as an alternative to antibacterial drugs, dual antibiotic-loaded nanoparticle systems, and ML studies on nanoparticles and drugs for antibacterial activity. **Figure 2.1** shows the relationship between the database, chapter, and other published works.

Based on the importance of metabolic networks in the biological activity of antibacterial drugs and the scarcity of studies of computational models of the same. It was decided to study the connectivity (structure) of a metabolite in the metabolic reaction networks of a query organism (Objective 1). The MN dataset published by Barabási's group was analyzed, which included the number of nodes (metabolites), input-output links (metabolic reactions), node degree, topological indices, and full names and codes of > 40 bacterial species. The results were published in a research paper. The objectives achieved in this section are presented in **Chapter 3**.

In order to better understand the interaction of previously analyzed metabolic networks with preclinical antibacterial drug assays, a "chemo-informatics-based perturbation theory (PT) and information fusion" methodology was proposed to relate chemical and preclinical datasets to metabolic network data quantitatively.

Preprocessing of preclinical ChEMBL antibacterial activity data. The literature is used to obtain data for the biological activity assays. We only searched the ChEMBL database for biological activity assays of AD against organisms present in the MN dataset. After data

curation, it was determined that the ChEMBL AD activity dataset contains the values of > 300 parameters (MIC, IC₅₀, etc.) for > 155000 biological assays of > 50000 compounds vs. > 25 bacteria species with > 90 strains.

Fusion of antibacterial drug information and metabolic network information. Activity values were obtained for the different biological properties of the two subsystems (AD and MN). We then preprocessed all the observed values with different units, scales, degrees of uncertainty, etc. to obtain dimensionless functions characterizing the system as a whole, the AD vs. MN cases.

PTML model. In this case, the model allows us to predict the scoring function for the antibacterial drug and the values of the metabolic networks in the combinatorial assay conditions, taking into account the assay conditions. The NIFPTML model was obtained from the fusion of several cheminformatics methods. Initially, we proposed a linear PTML model to predict biological activity and/or classify (AD-MN) pairs as desirable or undesirable and subsequently evaluated them in several nonlinear ML techniques. We used moving average (MA) operators to express perturbations in the assays and PT multiplier operators (PTO) to perform data fusion and dimension reduction. Finally, we applied linear discriminant analysis (LDA) and nonlinear ML algorithms to find the best NIFPTML predictive model. This work was submitted/published in Paper III and is developed in **Chapter 4**.

Having analyzed the interactions between preclinical antibacterial drug assays and metabolic networks and considering the recent advances in nanomedicine, we developed the first IFPTML model to design of DADNP (Dual Antibacterial Drugs Nanoparticle) systems, including AD and NP components at the same time. We trained alternative models with Linear Discriminant Analysis (LDA), Artificial Neural Networks (ANN), Bayesian Networks (BNN), K-Nearest Neighbor (KNN), and other algorithms. We also ran a simulation of the expected behavior of putative DADNPs in 72 different biological assays (> 1900 computations). The studied putative DADNPs consist of 27 different drugs with multiple NP classes and coat types. In addition, we tested the validity of our additive model with 80 experimentally synthesized and biologically tested DADNP complexes (reported in >45 articles). All these DADNPs show MIC values < 50 $\mu\text{g} \cdot \text{mL}^{-1}$ (cutoff used) better than the MIC of AD and NP alone (synergistic or additive effect). The assays involve DADNP complexes with 10 NP types, 6 coating materials, and a NP size range of 5-100 nm against 15 different antibiotics and 12 bacterial species. The IFPTML-LDA model correctly classified 100% (80 out of 80) of the DADNP complexes as biologically active. The IFPTML additive strategy may become a useful tool to aid in designing DADNP systems for antibacterial therapy, taking into account only information about the AD and NP components separately. The work developed was collected in paper IV, and all these results are summarized in **Chapter 5**.

Once the interaction between AD (ChEMBL) and NP were analyzed, metabolic networks were incorporated to understand the potential mechanisms of multidrug resistant (MDR) strains with perturbed metabolic networks (MN). This work used an NIFPTML analysis for mapping DADNP (AD + NP) versus MN systems of pathogenic bacterial species as a new application of AI/ML methods. Accordingly, the NIFPTML algorithm was selected to search for predictive models based on a ChEMBL dataset of > 160000 AD assays enriched with 300 NP and > 25 MN assays of different bacterial species. NIFPTML uses the IF process to join the three datasets, creating an NIFPTML linear discriminant analysis (LDA) model with $Sp \approx 90\%$ and $Sn \approx 80\%$ and the best artificial neural network (ANN) model found with $Sp \approx Sn \approx 95\%$ in training/validation. The series presented good results. This type of model could be useful for the discovery of DADNP systems. We also performed simulations of > 140000 points of putative DADNP systems against computationally generated wild-type and knockout (KO)

bacterial strains. The linear and additive NIFPTML models were able to predict 102 experimental cases of complex DADNPs with a high degree of structural and biological variety. This led us to introduce the concept of MDR computational surveillance that could help detect new strains of MDR bacteria. This work was published in the journal *Env Sc: Nano* described in **Chapter 6**.

3. REFERENCES

1. Zaengle-Barone, J. M.; Jackson, A. C.; Besse, D. M.; Becken, B.; Arshad, M.; Seed, P. C.; Franz, K. J. Copper Influences the Antibacterial Outcomes of a β -Lactamase-Activated Prochelator against Drug-Resistant Bacteria. *ACS Infectious Diseases*. **2018**, *4* (6), 1019-1029. DOI: 10.1021/acscinfecdis.8b00037.
2. Vazquez-Muñoz, R.; Meza-Villezas, A.; Fournier, P. G. J.; Soria-Castro, E.; Juarez-Moreno, K.; Gallego-Hernández, A. L.; Bogdanchikova, N.; Vazquez-Duhalt, R.; Huerta-Saquero, A. Enhancement of antibiotics antimicrobial activity due to the silver nanoparticles impact on the cell membrane. *PloS one*. **2019**, *14* (11), e0224904-e0224904. DOI: 10.1371/journal.pone.0224904 PubMed.
3. Panáček, A.; Smékalová, M.; Večeřová, R.; Bogdanová, K.; Röderová, M.; Kolář, M.; Kilianová, M.; Hradilová, Š.; Froning, J. P.; Havrdová, M.; et al. Silver nanoparticles strongly enhance and restore bactericidal activity of inactive antibiotics against multiresistant Enterobacteriaceae. *Colloids and Surfaces B: Biointerfaces*. **2016**, *142*, 392-399. DOI: <https://doi.org/10.1016/j.colsurfb.2016.03.007>.
4. Jijie, R.; Barras, A.; Teodorescu, F.; Boukherroub, R.; Szunerits, S. Advancements on the molecular design of nanoantibiotics: Current level of development and future challenges. *Molecular Systems Design and Engineering*. **2017**, *2* (4), 349-369, Review. DOI: 10.1039/c7me00048k Scopus.
5. Mamun, M. M.; Sorinolu, A. J.; Munir, M.; Vejerano, E. P. Nanoantibiotics: Functions and Properties at the Nanoscale to Combat Antibiotic Resistance. *Frontiers in Chemistry*. **2021**, *9*, Review. DOI: 10.3389/fchem.2021.687660 Scopus.
6. Zaidi, S.; Misba, L.; Khan, A. U. Nano-therapeutics: A revolution in infection control in post antibiotic era. *Nanomedicine: Nanotechnology, Biology and Medicine*. **2017**, *13* (7), 2281-2301. DOI: 10.1016/j.nano.2017.06.015.
7. Deng, H.; McShan, D.; Zhang, Y.; Sinha, S. S.; Arslan, Z.; Ray, P. C.; Yu, H. Mechanistic Study of the Synergistic Antibacterial Activity of Combined Silver Nanoparticles and Common Antibiotics. *Environmental Science & Technology*. **2016**, *50* (16), 8840-8848. DOI: 10.1021/acs.est.6b00998.
8. Lok, C.-N.; Ho, C.-M.; Chen, R.; He, Q.-Y.; Yu, W.-Y.; Sun, H.; Tam, P. K.-H.; Chiu, J.-F.; Che, C.-M. Proteomic Analysis of the Mode of Antibacterial Action of Silver Nanoparticles. *Journal of Proteome Research*. **2006**, *5* (4), 916-924. DOI: 10.1021/pr0504079.
9. Wang, S.; Gao, Y.; Jin, Q.; Ji, J. Emerging antibacterial nanomedicine for enhanced antibiotic therapy. *Biomaterials Science*. **2020**, *8* (24), 6825-6839, 10.1039/D0BM00974A. DOI: 10.1039/D0BM00974A.
10. Roche-Lima, A.; Domaratzki, M.; Fristensky, B. Metabolic network prediction through pairwise rational kernels. *BMC Bioinformatics*. **2014**, *15* (1), 318, journal article. DOI: 10.1186/1471-2105-15-318.
11. Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z. N.; Barabasi, A. L. The large-scale organization of metabolic networks. *Nature*. **2000**, *407* (6804), 651-654.
12. Herrera Ibatá, D. M. Modelos Multi-escala de Inteligencia Artificial para Diseño Químico-Informático y Fármaco-Epidemiológico de Terapias anti-VIH en Condados de Estados Unidos. PhD, Universidade da Coruña, La Coruña, España, 2015.

13. Santana, R.; Zuluaga, R.; Ganan, P.; Arrasate, S.; Onieva, E.; Montemore, M. M.; Gonzalez-Diaz, H. PTML Model for Selection of Nanoparticles, Anticancer Drugs, and Vitamins in the Design of Drug-Vitamin Nanoparticle Release Systems for Cancer Cotherapy. *Mol Pharm.* **2020**, *17* (7), 2612-2627. DOI: 10.1021/acs.molpharmaceut.0c00308.
14. Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research.* **2019**, *47* (D1), D930-D940. DOI: 10.1093/nar/gky1075.
15. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; et al. The ChEMBL database in 2017. *Nucleic acids research.* **2017**, *45* (D1), D945-D954. DOI: 10.1093/nar/gkw1074.
16. Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Kruger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The ChEMBL bioactivity database: an update. *Nucleic acids research.* **2014**, *42* (Database issue), D1083-1090. DOI: 10.1093/nar/gkt1031.
17. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research.* **2012**, *40* (Database issue), D1100-1107. DOI: 10.1093/nar/gkr777 From NLM.
18. Ebejer, J. P.; Charlton, M. H.; Finn, P. W. Are the physicochemical properties of antibacterial compounds really different from other drugs? *Journal of cheminformatics.* **2016**, *8*, 30. DOI: 10.1186/s13321-016-0143-5.
19. Lane, T.; Russo, D. P.; Zorn, K. M.; Clark, A. M.; Korotcov, A.; Tkachenko, V.; Reynolds, R. C.; Perryman, A. L.; Freundlich, J. S.; Ekins, S. Comparing and Validating Machine Learning Models for Mycobacterium tuberculosis Drug Discovery. *Mol Pharm.* **2018**, *15* (10), 4346-4360. DOI: 10.1021/acs.molpharmaceut.8b00083.
20. Bueso-Bordils, J. I.; Aleman-Lopez, P. A.; Suay-Garcia, B.; Martin-Algarra, R.; Duarte, M. J.; Falco, A.; Anton-Fos, G. M. Molecular Topology for the Discovery of New Broad-Spectrum Antibacterial Drugs. *Biomolecules.* **2020**, *10* (9). DOI: 10.3390/biom10091343.
21. Serafim, M. S. M.; Kronenberger, T.; Oliveira, P. R.; Poso, A.; Honorio, K. M.; Mota, B. E. F.; Maltarollo, V. G. The application of machine learning techniques to innovative antibacterial discovery and development. *Expert Opin Drug Discov.* **2020**, *15* (10), 1165-1180. DOI: 10.1080/17460441.2020.1776696.
22. Yang, X. G.; Chen, D.; Wang, M.; Xue, Y.; Chen, Y. Z. Prediction of antibacterial compounds by machine learning approaches. *J Comput Chem.* **2009**, *30* (8), 1202-1211. DOI: 10.1002/jcc.21148.
23. Marrero-Ponce, Y.; Marrero, R. M.; Torrens, F.; Martinez, Y.; Bernal, M. G.; Zaldivar, V. R.; Castro, E. A.; Abalo, R. G. Non-stochastic and stochastic linear indices of the molecular pseudograph's atom-adjacency matrix: a novel approach for computational in silico screening and "rational" selection of new lead antibacterial agents. *J Mol Model.* **2006**, *12* (3), 255-271.
24. Alafeef, M.; Srivastava, I.; Pan, D. Machine Learning for Precision Breast Cancer Diagnosis and Prediction of the Nanoparticle Cellular Internalization. *ACS sensors.* **2020**, *5* (6), 1689-1698. DOI: 10.1021/acssensors.0c00329.

25. Barnard, A. S.; Opletal, G. Predicting structure/property relationships in multi-dimensional nanoparticle data using t-distributed stochastic neighbour embedding and machine learning. *Nanoscale*. **2019**, *11* (48), 23165-23172. DOI: 10.1039/c9nr03940f.
26. He, J.; He, C.; Zheng, C.; Wang, Q.; Ye, J. Plasmonic nanoparticle simulations and inverse design using machine learning. *Nanoscale*. **2019**, *11* (37), 17444-17459. DOI: 10.1039/c9nr03450a.
27. Mikolajczyk, A.; Sizochenko, N.; Mulkiewicz, E.; Malankowska, A.; Rasulev, B.; Puzyn, T. A chemoinformatics approach for the characterization of hybrid nanomaterials: safer and efficient design perspective. *Nanoscale*. **2019**, *11* (24), 11808-11818. DOI: 10.1039/c9nr01162e.
28. Palizhati, A.; Zhong, W.; Tran, K.; Back, S.; Ulissi, Z. W. Toward Predicting Intermetallics Surface Properties with High-Throughput DFT and Convolutional Neural Networks. *J Chem Inf Model*. **2019**, *59* (11), 4742-4749. DOI: 10.1021/acs.jcim.9b00550.
29. Sun, B.; Fernandez, M.; Barnard, A. S. Machine Learning for Silver Nanoparticle Electron Transfer Property Prediction. *J Chem Inf Model*. **2017**, *57* (10), 2413-2423. DOI: 10.1021/acs.jcim.7b00272.
30. Yan, T.; Sun, B.; Barnard, A. S. Predicting archetypal nanoparticle shapes using a combination of thermodynamic theory and machine learning. *Nanoscale*. **2018**, *10* (46), 21818-21826. DOI: 10.1039/c8nr07341d.
31. Ferreira da Costa, J.; Silva, D.; Caamaño, O.; Brea, J. M.; Loza, M. I.; Munteanu, C. R.; Pazos, A.; García-Mera, X.; González-Díaz, H. Perturbation Theory/Machine Learning Model of ChEMBL Data for Dopamine Targets: Docking, Synthesis, and Assay of New l-Prolyl-l-leucyl-glycinamide Peptidomimetics. *ACS Chemical Neuroscience*. **2018**, *9* (11), 2572-2587. DOI: 10.1021/acscemneuro.8b00083.
32. Munteanu, C. R. Técnicas de ingeniería informática e inteligencia artificial para clasificación: aplicaciones para el descubrimiento de fármacos y dianas moleculares. PhD, Universidade da Coruña, La Coruña, España, 2013.
33. Escobar Cubiella, M. Q. Modelos bioinformáticos y estudio de receptores de proteínas mediante el uso de redes complejas para el desarrollo y diseño de fármacos eficaces en patologías del sistema nervioso central. PhD, Universidad de Santiago de Compostela, Santiago de Compostela, España, 2012.
34. Nocedo-Mena, D.; Cornelio, C.; Camacho-Corona, M. D. R.; Garza-Gonzalez, E.; Waksman de Torres, N.; Arrasate, S.; Sotomayor, N.; Lete, E.; Gonzalez-Diaz, H. Modeling Antibacterial Activity with Machine Learning and Fusion of Chemical Structure Information with Microorganism Metabolic Networks. *J Chem Inf Model*. **2019**, *59* (3), 1109-1120. DOI: 10.1021/acs.jcim.9b00034.
35. Ortega-Tenezaca, B.; Gonzalez-Diaz, H. IFPTML mapping of nanoparticle antibacterial activity vs. pathogen metabolic networks. *Nanoscale*. **2021**, *13* (2), 1318-1330. DOI: 10.1039/d0nr07588d.
36. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. Computational modeling in nanomedicine: prediction of multiple antibacterial profiles of nanoparticles using a quantitative structure-activity relationship perturbation model. *Nanomedicine (Lond)*. **2015**, *10* (2), 193-204. DOI: 10.2217/nnm.14.96.

CHAPTER 3. PREDICTING METABOLIC REACTION NETWORKS WITH PERTURBATION-THEORY MACHINE LEARNING (PTML) MODELS

Checking the connectivity (structure) of complex Metabolic Reaction Networks (MRNs) models proposed for new microorganisms with promising properties is an important goal for chemical biology. In principle, we can perform a hand-on checking (Manual Curation). However, this is a hard task due to the high number of combinations of pairs of nodes (possible metabolic reactions). In this work, we used Combinatorial, Perturbation Theory, and Machine Learning, techniques to seek a CPTML model for MRNs >40 organisms compiled by Barabásis' group. First, we quantified the local structure of a very large set of nodes in each MRN using a new class of node index called Markov linear indices f_k . Next, we calculated CPT operators for 150000 combinations of query and reference nodes of MRNs. Last, we used these CPT operators as inputs of different ML algorithms. The CPTML linear model obtained using LDA algorithm is able to discriminate nodes (metabolites) with correct assignation of reactions from not correct nodes with values of accuracy, specificity, and sensitivity in the range of 85-100% in both training and external validation data series. Meanwhile, PTML models based on Bayesian network, J48-Decision Tree and Random Forest algorithms were identified as the three best non-linear models with accuracy greater than 97.5%. The present work opens a door to the study of MRNs of multiple organisms using PTML models.

1. INTRODUCTION

The application of computational tools in Chemical Biology is a challenging goal; which becomes notably difficult if we consider the study of very large and complex biological networks.² In this context, Metabolic Reaction Networks (MRNs) of organisms are among the more important systems to be studied by Chemical Biology.³⁻⁶ MRNs are complex networks formed by combinations of thousands of chemical reactions or transformations (links) of metabolites (nodes) in a living organism. Computational chemists have excellent tools to manage single chemical reactions, but difficulties emerge when one must consider complex MRNs. Despite the similarity in metabolic pathways (conservation), the structure of MRNs is often very different in organisms of differing species and researchers have to propose specific MRNs models for each organism.⁷ The construction of MRNs for different organisms use different experimental techniques and many times also rely upon sequence alignment computational techniques. It means that we can use computational techniques to find an enzyme in a query organism (o) with high sequence similarity to another enzyme present in the Metabolome of another organism of a different species (s). Then, we can presuppose that the metabolic reaction catalyzed by this enzyme exists on both organisms. This combination of different experimental and alignment-based computational techniques leads to a vast amount of information. After that, researchers use the different method to process all this information and propose consensus MRNs models for the new organisms.^{8,9} In this context, it is straightforward to realize that we need to carry out a checking of the connectivity (structure) of the alternative complex MRNs models proposed for new microorganisms. In principle, we can perform a hand-on checking (Manual Curation) but this is a hard task due to the high number of combinations of pairs of nodes (possible metabolic reactions).

In fact, the numerous challenges that arise on the assembly of complex network models are not unique to MRNs. Complex network analysis provides an effective approach to diverse

problems in the bio-molecular, technological, and social sciences.¹⁰⁻¹³ In all these cases, the search for models able to predict and/or evaluate the structure or properties of the networks proposed is an exciting field within the complex network sciences. For instance, one of the simplest models is a linear equation which predicts the lethality of proteins in yeast, based on simple structural parameters like the node degree of the protein within the interaction network.¹⁴ However, in order to solve more complex problems, we eventually need to search for not so simple structural parameters and/or rely on more complicated multivariate models.

Specifically, the verification or checking (curation) of new models of MRN for new organisms with promising properties is useful for the biotechnology industry. In principle, we can calculate numerical parameters of the connectivity of alternative MRNs and use them as input for a Machine Learning (ML) algorithm able to discriminate viable from unviable MRNs. On the other hand, Perturbation Theory (PT) models allow us to predict the solutions to a query problem (q) based on a previously known solution for a similar problem or problem of reference (r). Specifically, we classify a model as Combinatorial PT (CPT) model when we can use it to study all possible combinations of q and r pairs. In a recent work, we outlined a new type of ML method called CPTML = CPT + ML which leverages ideas from both PT and ML models.¹⁵ The CPTML method uses a different kind of CPT operators to predict the properties of one system based on the properties of a system of reference. For instance, Moving Average (MA) operators used in Box-Jenkins's ARIMA models in time series analysis.¹⁶ The MA operators of structural descriptors are useful to quantify variations in network interconnection patterns due to multiple conditions or parameters (e.g. date of an assay in a time-course experiment) in models of complex datasets in Organic Chemistry, Medicinal Chemistry, Nanotechnology, *etc.*¹⁷⁻²² Very recently, Speck-Planche and Cordeiro²³⁻²⁶ have used this kind of models in chemical combinatorial sciences. Recently,²⁷ we applied CPTML ideas to predict the effect of inter-species perturbations in combinations of nodes in MRNs using Markov Chain (MC) indices. However, there are no other reports of CPTML models of inter-species perturbations in MRNs, to the best of our knowledge. Therefore, the search for novel numerical descriptors useful in the development of CPTML models of complex networks in general, and MRNs, in particular, is an area of increasing interest. The structure of a network-like system is a function of the system components (nodes) and the relationships between them (interactions). Therefore, most network structure descriptors are Topological Indices (TIs) that codify information about the connectivity (topology) of the network. Specifically, MC theory has been applied in a number of studies of complex networks to calculate network invariants.²⁷⁻³³ On the other hand, the atom-based linear indices, developed by Marrero *et al.*,^{34, 35} have been successfully applied to studies of small molecules. However, there is no previous work extending these indices to the study of complex networks. Here, we describe the conceptual extension of these linear indices to create novel descriptors of MRN topological structure. We go on to demonstrate the effectiveness of these new descriptors in the task of recognizing MRN structure against a null model derived from randomized networks.

We can fit a PTML model using different types of ML algorithms (linear or non-linear). For example, Bediaga *et al.*³⁶ use non-linear Artificial Neural Network (ANN) algorithms to fit PTML models with a ChEMBL dataset of preclinical assays of anti-cancer compounds. Also, González-Durruthy *et al.*³⁷ used ANN algorithms to find PTML models of the potential ability of carbon nanotubes to induce mitochondrial toxicity-based inhibition of the mitochondrial H-FOF1-ATPase from in vitro assays. On the other hand, Ferreira da Costa *et al.*³⁸ have compared linear and non-linear ML methods like ANN, Random Forest, and Deep Learning algorithms, in the search of PTML models supporting the organic synthesis,

chemical characterization, and pharmacological assay of a new series of PLG Peptide-mimetic compounds.

In the present study, we report a new model for the prediction of the connectivity (local structure) of a query metabolite (node q) in the MRN of a query organism (o). We attack this problem considering that with know the local structure of a similar metabolite of reference (node r) in the MRN of another species (s). Firstly, we quantify the local structure of the nodes in the MRNs with a new class of node index called Markov linear indices f_k of order k . These indices are a generalization of linear indices of molecular graphs,^{34,35} here adapted for complex networks. Next, we calculate the values of different perturbation theory operators $\Delta f_k(q, r)$, $\Delta f_k(q, o)$, $\Delta f_k(r, s)$, and $\Delta \Delta f_k(q, o, r, s)$. These operators are able to quantify the deviations (perturbations) on the local structure of one node in the MRN with respect to the structure of a similar node and/or all nodes in the MRN of a potentially differing species. We calculated these perturbation operators for 150 000 pairs of query and reference nodes from the MRNs of >40 organisms compiled by Barabási's group.⁷ Last, we use these PT operators as inputs of the linear and no linear ML methods. The PTML linear and no linear models obtained achieves high values of accuracy, specificity, and sensitivity in both training and validation series, and compared with previous reports.²⁷ This model may become a useful tool to predict the interconnections for newly characterized metabolites in the MRNs of engineered organisms. This proposed approach could help to recognize which metabolites are the most determinant within novel MRNs. Besides, this type of characterization helps to identify if the specific metabolite composition of the new synthetic biology systems remains robust to variation of internal and external conditions (resistance to pathogens, *etc.*). In **Figure 3.1**, we show the workflow used in this work to develop the new CPTML model for MRNs checking.

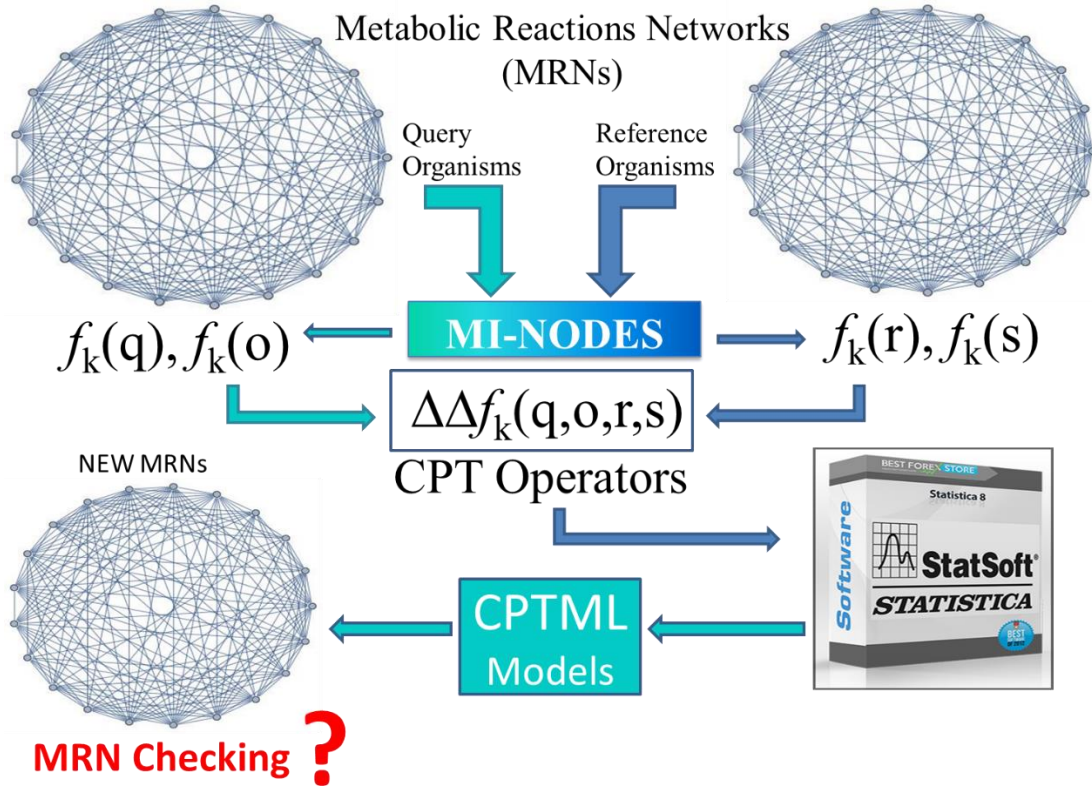


Figure 3.1. Workflow of the PTML method applied to MRNs checking.

2. METHODS

2.1 Dataset of complex networks.

We obtained the data of the MRNs from Barabási's group website. In these data, a unique number identifies each substrate in the metabolite network for each organism. The data are formatted as directed links: from \rightarrow to. The data were originally compiled in,⁷ taken from the 'intermediate metabolism and bioenergetics' portions of the WIT database, and were previously used to understand the large-scale organization of metabolic networks.⁷ **Table 3.1** shows the number of metabolites and metabolic reactions for all organisms studied in this work.

Table 3.1. Details of the metabolic networks of >40 organisms.

Organism Name	Symbol	N ^a	L _{in} ^b	L _{out} ^b	Organism Name	Symbol	N ^a	L _{in} ^b	L _{out} ^b
<i>Aeropyrum pernix</i>	AP	204	588	575	<i>Chlorobium tepidum</i>	CL	389	1097	1062
<i>Archaeoglobus fulgidus</i>	AG	496	1527	1484	<i>Rhodobacter capsulatus</i>	RC	670	2174	2122
<i>Methanobacterium thermoautotrophicum</i>	TH	430	1374	1331	<i>Rickettsia prowazekii</i>	RP	214	510	504
<i>Methanococcus jannaschii</i>	MJ	424	1317	1272	<i>Neisseria gonorrhoeae</i>	NG	406	1298	1270
<i>Pyrococcus furiosus</i>	PF	316	901	867	<i>Neisseria meningitidis</i>	NM	381	1212	1181
<i>Pyrococcus horikoshii</i>	PH	323	914	882	<i>Campylobacter jejuni</i>	CJ	380	1142	1115
<i>Aquifex aeolicus</i>	AA	419	1278	1249	<i>Helicobacter pylori</i>	HP	375	1181	1144
<i>Chlamydia pneumoniae</i>	CQ	194	401	391	<i>Escherichia coli</i>	EC	778	2904	2859
<i>Chlamydia trachomatis</i>	CT	215	479	462	<i>Salmonella typhi</i>	TY	819	3008	2951
<i>Synechocystis sp.</i>	CY	546	1782	1156	<i>Actinobacillus actinomycetemcomitans</i>	AB	395	1202	1166
<i>Porphyromonas gingivalis</i>	PG	424	1192	1221	<i>Haemophilus influenzae</i>	HI	526	1773	1746
<i>Mycobacterium bovis</i>	MB	429	1247	1244	<i>Pseudomonas aeruginosa</i>	PA	734	2453	2398
<i>Mycobacterium leprae</i>	ML	422	1271	1823	<i>Treponema pallidum</i>	TP	207	562	555
<i>Mycobacterium tuberculosis</i>	MT	587	1862	2741	<i>Borrelia burgdorferi</i>	BB	187	442	438
<i>Bacillus subtilis</i>	BS	785	2794	1218	<i>Thermotoga maritima</i>	TM	338	1004	976
<i>Enterococcus faecalis</i>	EF	386	1244	1578	<i>Deinococcus radiodurans</i>	DR	815	2870	2811
<i>Clostridium acetobutylicum</i>	CA	494	1624	525	<i>Emericella nidulans</i>	EN	383	1095	1081
<i>Mycoplasma genitalium</i>	MG	209	535	466	<i>Saccharomyces cerevisiae</i>	SC	561	1934	1889
<i>Mycoplasma pneumoniae</i>	MP	178	470	1298	<i>Caenorhabditis elegans</i>	CE	462	1446	1418
<i>Streptococcus pneumoniae</i>	PN	416	1331	1277	<i>Oryza sativa</i>	OS	292	763	751
<i>Streptococcus pyogenes</i>	ST	403	1300		<i>Arabidopsis thaliana</i>	AT	302	804	789

Notes. ^aN = number of nodes (metabolites), ^bL = input-output links (metabolic reactions), according to Jeong *et al.*⁷

2.2 Markov linear indices for complex networks.

The linear indices (f_k) proposed by Marrero *et al.*^{34, 35} are interesting due to their efficient encoding of molecular structure in QSAR studies. Previous application of these indices include the prediction of antiprotozoal inhibitory activity for novel quinoxalinones,³⁹ classification models for tyrosinase inhibitory activity discrimination,⁴⁰ *etc.* Nevertheless, the majority of models developed until now only predict outputs for one target, including those based on classic linear indices. In a recent work,⁴¹ we developed a multi-output model for inhibitors of the ubiquitin-proteasome pathway with potential anti-cancer applications.⁴²

These classic atom-based linear indices measure the connectivity between pairs of atoms (nodes) placed at different atom-atom topological distances (d_{ij}) in a chemical graph.³⁴ Using Markov chain theory, higher-order analogues, M_k , of classic atom-based linear indices are calculated. In so doing, the more general concept of node is applied, such that the indices may be used for networks of any node type, with different levels of matter complexity (i.e. atoms, drugs, proteins, organisms, etc.). Here, we present for the first time the generalization of atom-based linear indices of molecular graphs to the study of complex networks using Markov chain theory. We combine the concept behind Marrero's molecular descriptors, specifically the atom-based linear indices,⁴³ and the MARCH-INSIDE (MI) software algorithm of González-Díaz *et al.*⁴⁴⁻⁴⁷ This generalization was done with the aim to extend the atom-based linear indices for the study of different types of complex networks.

The MRN comprises n nodes which represent metabolites (vector of \mathcal{R}^n). In this sense, the adjacency matrix \mathbf{M} (which represents the deterministic bonding and atom connectivity) is modified by the stochastic Markov matrix or order k , ${}^k\Pi$, which indicates the probability, ${}^k p_{ij}$, of metabolite transformation (i.e. that a directed edge is traversed between two nodes). The k^{th} order Markov linear index f_k is defined as:

$$f_k(i) = \sum_{j=1}^n {}^k p_{ij} \cdot \delta_i \quad (1)$$

Where, ${}^k p_{ij} = {}^k p_{ji}$ (${}^k\Pi$ is an asymmetric square matrix), n is the number of metabolites in the network, and δ_i is the degree of node (metabolite) i . In addition to the MRN data from Barabási's group, we also generated MRN with incorrect random patterns using the software CentiBin.⁴⁸ Specifically, we created the Erdos Renyi Random Network ERRN (1000, 0.1) and the Kleinsberg Small World Network KSWN (1000). Both random networks have 1000 nodes each. These random networks are denoted in the tables as ER and SK, respectively. We used the software MI-NODES to calculate the Markov linear indices f_k with order $k = [0, 5]$ for all the true and random MRNs.²⁷⁻³³

2.3 PT operators for MRNs of different organisms.

A new methodology that uses PT in multiple-condition QSPR/QSAR problems was introduced by Gonzalez-Díaz *et al.*⁴⁹ In the present report, the prediction of the effect of perturbations in the connectivity of MRN is studied based on an adaptation of this theory. In a metabolic network, each metabolite may act (in principle) as substrate or product of a metabolic reaction. This means that we may have a series of chemical reactions (a metabolic pathway) that leads to the production of this metabolite and/or to the consumption of this metabolite. We can represent the sub-network (local metabolic pathway) of all the reactions that lead to the production or consumption of the metabolite q inside an organism o as a sub-graph of M . This sub-graph, denoted as L_{qo} , is rooted on the metabolite q within organism o . L_{qo} may contain many links (chemical reactions) inside the network. This is the sub-graph of immediate incoming/outgoing edges with respect to node q . Consequently, this sub-graph has a diameter $d = 2$ (longest topological distance between one substrate of the reactions with the metabolite q); give that there is a sub-graph (local metabolic pathway) centered on q . These sub-graphs are star graphs centered on q . Simple visual inspection of these sub-graphs allows us to detect specific connectivity patterns for each metabolite. This local connectivity pattern clearly changes when we compare pattern L_{qo} of one metabolite, q , with L_{ro} of another metabolite, r , in the same organism. More importantly, the local metabolic patterns also have variations or perturbations when we compare the local metabolic pathway L_{qo} of metabolite, q , in the organism, o , with the pattern L_{qs} of the same metabolite q in another organism from another species s .²⁷

In this sense, we require a new CPTML model able to predict the connectivity pattern (L_{qo}) of a metabolite, q , in the MRN of the organism, o , given that we already know the connectivity L_{rs} for a metabolite of reference, r , in the MRN of another organism of species, s . Our model takes into consideration perturbations in the local connectivity (metabolite connectivity) or global connectivity (full organism metabolic changes) in the new MRN with respect to the MRN of reference. We can measure these perturbations numerically using the following functions of linear indices.⁴⁹

$$\Delta f_k(q, r) = f_k(q) - f_k(r) \quad (2)$$

$$\Delta f_k(q, o) = f_k(q) - f_k(o)_{avg} \quad (3)$$

$$\Delta f_k(r, s) = f_k(r) - f_k(s)_{avg} \quad (4)$$

$$\Delta \Delta f_k(q, o, r, s) = [f_k(q) - f_k(o)] - [f_k(r) - f_k(s)] \quad (5)$$

Here, $f_k(q)$ is the k^{th} order descriptor of the sub-graph centered on node q , while $f_k(o)_{avg}$ is the average of the k^{th} order descriptor computed for each node in the MRN for organism o . The perturbation operators Δf_k and $\Delta \Delta f_k$, are inspired by the idea of MA operators used in Box-Jenkins models in time series analysis.^{16, 42} We explored four types of operators in our study. The first type of operator, $\Delta f_k(q, r) = f_k(q) - f_k(r)$, is the difference between the local linear indices of the new metabolite q and the reference metabolite r . We designed this operator to measure the changes in the local connectivity of the new node (metabolite) with respect to the metabolite of reference. The second type of operator $\Delta f_k(q, o) = f_k(q) - f_k(o)_{avg}$ and $\Delta f_k(r, s) = f_k(r) - f_k(s)_{avg}$ was created to measure the deviation in the connectivity of a given metabolite (q or r) with respect to all the metabolites within the same organism (o or s). We constructed this operator to measure perturbations in the reactions of one metabolite with respect to the overall metabolism of the entire organism. The last type of operator $\Delta \Delta f_k(q, o, r, s) = [f_k(q) - f_k(o)_{avg}] - [f_k(r) - f_k(s)_{avg}]$ combines both local and overall metabolic perturbations. It is important to note that the terms $f_k(o)_{avg}$ and $f_k(s)_{avg}$ cancel out when the query organism, o , and the organism of reference, s , are the same.

2.4 PTML linear model for MRNs of different organisms.

Within this framework, it is possible to propose and test different relationship between perturbations of input/output conditions with L_{qo} . This is a discrete-value function (Boolean) ideal for classification techniques. The following equation is a general form of the model that includes only additive perturbations of linear functions.

$$\lambda(L_{qo})_{new} = a \cdot \lambda(L_{rs})_{ref} + \sum_{k=0}^{k=5} b_k \cdot \Delta f_k(q, o) + \sum_{k=0}^{k=5} c_k \cdot \Delta f_k(r, s) + \sum_{k=0}^{k=5} d_k \cdot \Delta \Delta f_k(q, o, r, s) + e_0 \quad (6)$$

The first input term is the scoring function $\lambda(L_{rs})_{ref}$ for the connectivity pattern of the MRN of reference r . In this work, we used the identity function, $\lambda = I$, for $\lambda(L_{rs})_{ref}$. Consequently, $\lambda(L_{rs})_{ref} = L_{rs}$ the connectivity pattern of the metabolite of reference r . The values of the sum operators run from $k = 0$ to $k = 5$ as we calculated only the first terms of the family of linear indices with $k = 0$ to 5 . In this work, we are going to use the LDA module of implemented on the software STATISTICA to seek the model. Then, the output $\lambda(L_{qo})_{new}$ is one scoring function λ (get real values) for the connectivity pattern L_{qo} of metabolite q in the MRN of the organism o (output). Linear forward stepwise strategy for variable selection is going to be used to select the input variables on the classification equation. Fisher ratio test determines which attributes (variables) enter the model. Chi-square (χ^2) and error level $p < 0.01$ is used to detect significant separation between the two classes ($L_{qo} = 1 / L_{qo} = 0$). The canonical correlation coefficient (R_{can}) is going to be used to measure the strength of the linear relationship.

This model uses the operators of linear indices, f_k , of MRN to predict the effects of inter-species variations in metabolic connectivity patterns L_{qo} . In general, the connectivity to be predicted (output of the model) $\lambda(L_{qo})_{new} > 0 \Rightarrow L_{qo} = 1$ when both metabolites q and r have all the correct connections on the MRNs their respective organisms o and s . Conversely, $L_{qo} = 0$ when q is another metabolite different from r or has an incorrect connectivity pattern (the metabolic reactions assigned are not correct for this metabolite). Correct here means that the connections coincide with those determined experimentally are accepted to be correct according to the dataset of Barabási's group. In addition, the connectivity of reference (input of the model) $L_{rs} = 1$ when the metabolite of reference r has all the correct connections (metabolic reactions) on the MRN of reference r . On the other hand, $L_{rs} = 0$ when r is a metabolite with incorrect connectivity patterns.

2.5 PTML non-linear models.

The PTML non-linear models were developed using Waikato Environment for Knowledge Analysis (WEKA), version 3.8.0. Five classification algorithms were applied: Bayesian network (BN), multinomial logistic ridge regression (LRR), J48 decision tree (J48), Multilayer perceptron (MLP), and Random Forest (RF).

Bayesian Network (BN). A Bayesian network is a Directed Acyclic Graph (DAG) $G = (X; A)$, where each node $X_i \in X$ represents a random variable in a domain, and each arc $a_{i,j} \in A$ describes a direct dependence relationship between two variables X_i and X_j . Associated with each node X_i , there is a conditional probability distribution represented by $\theta_i = P(X_i | \Pi(X_i))$, which quantifies how much the node X_i depends on its parents $\Pi(X_i)$. $Val(X_1, \dots, X_n)$ is the set of possible values of variables (X_1, \dots, X_n) . As the graph structure G qualitatively characterizes the independence relationship among random variables, the conditional probability distribution quantifies the strength of dependencies between a node and its parent nodes. It can be proved that a Bayesian network $(X; A)$ uniquely encodes the joint probability distribution of the domain variables $X = (X_1, \dots, X_n)^{50}$

$$P(X) = \prod_{i=1}^n \theta_i \quad (7)$$

Multi-Layer Perceptron (MLP). A network consists of interconnected nodes that form three kinds of layers: input, hidden and output. There can be more than one hidden layer. In the case of the "hidden" layer, a single layer of hidden units was used for the classification models. The activation function for this ANN model was the sigmoidal function.⁵¹

Binary Logistic Regression (BLR). A binomial logistic regression (often referred to simply as logistic regression), predicts the probability P that an observation falls into one of two categories of a dichotomous $Y =$ dependent variable based on one or more independent variables $X = \{X_1, X_2, \dots, X_n\}$ that can be either continuous or categorical. The regression logistic model is described by equation (8), where $LR(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$ represent the classification function with β_i coefficients. The value of dependent variable Y is 1, if $P(X) > 0.5$ and in other case Y is 0.⁵⁰

$$P(X) = \frac{1}{1 + e^{-(LR(X))}} \quad (8)$$

Random Forest (RF). RF, is a very sophisticated algorithm to handle a large number of classification problems, where numerous decision trees are assembled by voting mechanism.⁵² For each tree, a different training set is generated by randomly re-sampling the data set with

replacement resulting in a training set (bootstrap learning set) that contains approximately two-thirds of the samples in the data set original.

J48 decision tree (J48). The C 4.5 is an algorithm developed by Ross Quinlan⁵³ that employ the basic methodology of divide-and-conquer described in CART.⁵⁴ The model displays a flowchart-like tree structure where each internal node represents a test on selected variable, branch denote test outcomes, and each terminal (leaf) node assigns a class label. The differences are in the tree structure, the splitting criteria, the pruning method, and the way missing values are handled.^{53, 55}

3. RESULTS AND DISCUSSION

3.1 PTML linear models.

In this study, we constructed a new classification model to predict the local connectivity pattern L_{qo} for the query metabolites q in the MRN of the new organism (o) using as reference local connectivity L_{rs} of other metabolites (r) on MRNs of different species (s). Firstly, we calculated the values of the Markov linear indices $f_k(q)$ and their perturbation operators Δf_k and $\Delta \Delta f_k$ for all the metabolites (nodes) in the MRN of 40 organisms. In **Table 3.2**, we depict some of the average values of the linear indices for many organisms. We also created a supplementary material with the full list of 155 000 perturbations predicted (material provided as online supplementary material). In the next step, we calculate the values of the PT operators for 155000 combinations of query vs. reference metabolites selected the correct MRNs and randomized models of MRNs networks (examples of incorrect connectivity). For doing this, our database has 109 979 combinations of nodes with correct patterns and 45 021 combinations of nodes with incorrect patterns. After that, the dataset was randomly divided in training and prediction subsets in a 3:1 rate, respectively. The training set remains with 116 052 cases and the test set have 38 948 cases.

Table 3.2. Average values of f_k for the metabolic networks of >40 organisms.

Org. ^a	Node degrees ^b			Markov Linear Indices ^c					
	δ	δ_{in}	δ_{out}	f_0	f_1	f_2	f_3	f_4	f_5
AA	0.10	0.10	0.10	0.10	-0.65	-0.59	-0.60	-0.52	-0.50
AB	0.04	0.04	0.04	0.04	0.20	0.08	0.09	0.03	-0.01
AG	-0.05	-0.05	-0.05	-0.05	-0.66	-0.68	-0.69	-0.67	-0.65
AP	-0.06	-0.06	-0.06	-0.06	0.15	0.04	0.12	0.03	0.03
AT	-0.79	-0.79	-0.79	-0.79	-0.64	-0.72	-0.74	-0.80	-0.82
BB	-2.13	-2.13	-2.13	-2.13	-0.51	-0.64	-0.65	-0.71	-0.71
BS	1.09	1.09	1.09	1.09	3.23	3.03	2.83	2.70	2.55
CA	0.69	0.69	0.69	0.69	-0.60	-0.42	-0.42	-0.27	-0.24
CE	0.58	0.58	0.58	0.58	-0.53	-0.46	-0.51	-0.45	-0.43
CJ	0.06	0.06	0.06	0.06	-0.47	-0.35	-0.40	-0.32	-0.31
CL	-0.26	-0.26	-0.26	-0.26	0.85	0.95	1.13	1.14	1.12
CQ	-3.11	-3.11	-3.11	-3.11	0.25	0.25	0.17	0.06	-0.01
CT	-2.56	-2.56	-2.56	-2.56	-0.69	-0.98	-1.00	-1.17	-1.18
CY	0.81	0.81	0.81	0.81	1.06	1.08	1.23	1.15	1.18
DR	1.04	1.04	1.04	1.04	0.44	0.29	0.47	0.37	0.39
EC	1.43	1.43	1.43	1.43	0.57	0.78	0.78	0.94	0.94
EF	0.67	0.67	0.67	0.67	-0.64	-0.49	-0.47	-0.32	-0.27
EN	-0.04	-0.04	-0.04	-0.04	-0.51	-0.59	-0.63	-0.70	-0.72
HI	0.85	0.85	0.85	0.85	1.29	0.96	0.98	0.81	0.78
HP	0.65	0.65	0.65	0.65	0.24	0.50	0.60	0.82	0.84
MB	-0.13	-0.13	-0.13	-0.13	-0.63	-0.80	-0.82	-0.90	-0.91
MG	-1.29	-1.29	-1.29	-1.29	-0.66	-0.90	-0.93	-1.08	-1.10
MJ	0.12	0.12	0.12	0.12	-0.65	-0.62	-0.63	-0.58	-0.57
ML	0.04	0.04	0.04	0.04	-0.56	-0.70	-0.75	-0.83	-0.85
MP	-0.96	-0.96	-0.96	-0.96	-0.59	-0.86	-0.92	-1.10	-1.14
MT	0.41	0.41	0.41	0.41	1.22	1.24	1.38	1.39	1.43
NG	0.51	0.51	0.51	0.51	-0.58	-0.46	-0.49	-0.41	-0.41
NM	0.55	0.55	0.55	0.55	-0.56	-0.24	-0.22	0.04	0.10
OS	-0.96	-0.96	-0.96	-0.96	-0.64	-0.69	-0.69	-0.69	-0.68
PA	0.76	0.76	0.76	0.76	3.51	3.42	2.97	2.78	2.73
PF	-0.24	-0.24	-0.24	-0.24	-0.67	-0.88	-0.90	-1.02	-1.05
PG	-0.51	-0.51	-0.51	-0.51	-0.66	-0.47	-0.44	-0.25	-0.20
PH	-0.36	-0.36	-0.36	-0.36	-0.60	-0.70	-0.75	-0.80	-0.81
PN	0.57	0.57	0.57	0.57	-0.63	-0.41	-0.39	-0.20	-0.15
RC	0.54	0.54	0.54	0.54	1.97	1.93	2.43	2.37	2.51
RP	-1.46	-1.46	-1.46	-1.46	-0.60	-0.86	-0.89	-1.05	-1.07
SC	1.46	1.46	1.46	1.46	0.16	0.38	0.40	0.51	0.57

ST	0.67	0.67	0.67	0.67	-0.34	-0.38	-0.38	-0.38	-0.37
TH	0.49	0.49	0.49	0.49	-0.62	-0.46	-0.47	-0.36	-0.35
TM	0.05	0.05	0.05	0.05	-0.51	-0.47	-0.49	-0.46	-0.47
TP	-0.72	-0.72	-0.72	-0.72	-0.27	-0.42	-0.47	-0.58	-0.59
TY	1.34	1.34	1.34	1.34	0.52	0.56	0.39	0.31	0.28
YP	0.07	0.07	0.07	0.07	-0.00	0.76	0.75	1.16	1.14

Notes. ^a Org = organism, Organisms names are in two letters code and metabolite names in numeric code, according to Jeong *et al.*⁷ ^b Average node degrees, δ . ^c All indices have been standardized to z-scaled values to avoid scale errors in visual comparison.

As can be observed in **Table 3.3**, the obtained PTML model classified correctly the 90.5% of the cases in the training set, with only 11 079 misclassified cases out of 116 052 cases. In the case of the prediction set, the same behavior was observed with an accuracy (Ac) value of 89.8%, and only 3 967 misclassified cases. Both series have specificity (Sp) values of 100%, this means that none of the negative cases (incorrect patterns) is misclassify. An adequate value of sensitivity (Sn) is observed for the training and prediction series with an 86.5% and 85.7%, respectively. In general, the PTML model has a good performance for describing the correct/incorrect connectivity pattern as showed in the performance of the statistical parameters of the current classification equation. The following equation (7) was the best PTML model found includes only additive perturbations of linear functions:

$$\lambda(L_{qo})_{\text{new}} = -0.33127 \cdot \lambda(L_{rs})_{\text{ref}} - 0.73350 \cdot \Delta\Delta f_0(q, o, r, s) + 0.12583 \cdot \Delta\Delta f_1(q, o, r, s) - 0.18410 \cdot \Delta\Delta f_2(q, o, r, s) - 1.96212 \quad (7)$$

$$N = 116052, Rc = 0.83, \chi^2 = 135844.3, \quad p < 0.01$$

The output $\lambda(L_{qo})_{\text{new}}$ is the scoring function λ for the connectivity pattern L_{qo} of metabolite q in the MRN the organism o (output). The first input term is the scoring function $\lambda(L_{rs})_{\text{ref}}$ for the connectivity pattern of the MRN of reference. In this work, we used the identity function, $\lambda = I$, for $\lambda(L_{rs})_{\text{ref}}$. Consequently, $\lambda(L_{rs})_{\text{ref}} = L_{rs}$ the connectivity pattern of the metabolite of reference r , see details on the previous section. Notably, the only one type of operator in the final CPTML model is $\Delta\Delta f_k(q, o, r, s)$, which reflects both local and global effects. This coincides with the excellent results obtained with Box-Jenkins MA operators⁵⁶⁻⁵⁹ and perturbation models⁴⁹ for related problems.

Table 3.3. Results of CPTML model for metabolic networks of >40 organisms.

Data sub-set	Stat. Param.	Pred. %	Predicted MRN perturbations	
			$L_{qo} = 1$	$L_{qo} = 0$
$L_{qo} = 1$	Sp	100	71207	11079
$L_{qo} = 0$	Sn	86.5	0	33766
Train total	Ac	90.5		
$L_{qo} = 1$	Sp	100	23726	3967
$L_{qo} = 0$	Sn	85.7	0	11255
Validation total	Ac	89.8		

3.2 PTML non-linear models.

Finally, a comparative study of our linear model with non-linear models obtained using ML classification algorithms was carried out. The ML methods implemented Waikato Environment for Knowledge Analysis (WEKA) was used to process our dataset, results are shown in **Table 3.4**. All PTML non-linear models displayed better performance than PTML linear model, especially for the training set with a global Accuracy of 91.25–99%. Analysing each model in particular, one can reveal that ANN algorithm of type MLP, and ensemble decision trees (RF) are the most powerful learning algorithms (Ac~97.5%). However, the best Accuracy results are from Bayesian network (BN) (global Ac of 99% for training, 98.78% test and 98.96% for cross-validation).

Table 3.4. Results of CPTML-non linear models for metabolic networks of >40 organisms.

ML Model	Parameters ^a				
	Data set	Ac(%)	Sn(%)	Sp(%)	MCC
BN	Training	99.00	99.16	97.45	0.98
	Test	98.92	99.16	97.16	0.97
	CV	98.98	99.12	97.42	0.98
MLP	Training	92.99	97.04	82.13	0.85
	Test	92.69	97.01	81.29	0.84
	CV	91.02	91.91	80.14	0.8
BLR	Training	91.25	97.31	78.04	0.81
	Test	91.04	99.12	76.69	0.81
	CV	91.25	97.31	78.05	0.81
RF	Training	98.59	98.98	96.28	0.97
	Test	98.46	98.89	95.90	0.96
	CV	98.55	99.18	95.97	0.97
J48	Training	98.99	99.18	97.39	0.98
	Test	97.89	95.52	97.13	0.95
	CV	98.97	99.16	97.35	0.98

^a Parameters, Ac(%) = Accuracy, Sn(%) = Sensitivity, Sp(%) = Specificity, and MCC = Mathew's Correlation Coefficient

In **Figure 3.2**, we summarized the results obtained graphically. While BLR model only presents a similar-to-lower goodness-of-fit in comparison with the PTML linear model and most other models. In conclusion, BN, J48 and RF were identified as the three best PTML non-linear models based on the consensus analysis of MCC, and overall accuracy. However, the improvement from PTML linear models to PTML non-linear models was not cost-benefit efficiently. It means, PTML non-linear models obtained are notably more complicated with a high number of parameters as compared to PTML linear model.

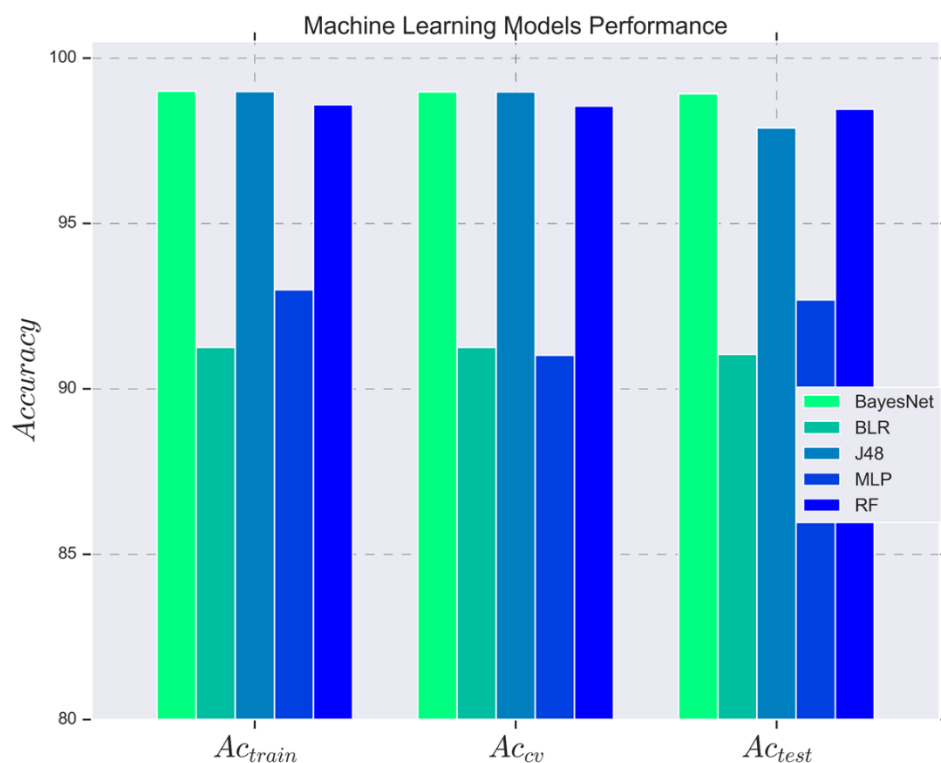


Figure 3.2. PTML non-linear model performance for metabolic networks of >40 organisms.

4. CONCLUSIONS

The results showed that the extension and generalization of atom-based linear indices to the Markov linear indices of complex networks is straightforward to realize. We also demonstrated that these indices are very useful to predict the effect of structural and inter-species variations (perturbations) in MRNs. Regarding the methodological objectives, the linear model only included one subsystem (metabolic networks of different microorganisms) and showed a good fit ($Sp=100\%$, $Sn=86.5\%$ and $Ac=90.5\%$). In this chapter other subsystem information blocks were not considered to analyze the problem presented in the thesis.

Regarding the practical objectives, the CPTML model obtained shows promising results with an accuracy (Ac), specificity (Sp), and sensitivity (Sn) between 85-100% for perturbations in a total of 155000 combinations of nodes in MRN of >40 organisms, overcoming previous studies in the same issue. Meanwhile, the performance of models was found to be improved by including different non-linear models. Leading to models with excellent internal accuracy and predictively on external data to classify correctly structural and inter-species variations (perturbations) in MRNs. The $\Delta\Delta f_k(q,o,r,s)$ are probably the most important (with respect to is $\Delta\Delta f_k(q,o)$ and is $\Delta\Delta f_k(r,s)$) because they quantify local (node connectivity) and global (organism) information at the same time. The new CPTML may become a useful tool to check out the structure of MRN of new organisms in biotechnology.

5. REFERENCES

1. Diéguez-Santana, K.; Casañola-Martin, G. M.; Green, J. R.; Rasulev, B.; González-Díaz, H. Predicting Metabolic Reaction Networks with Perturbation-Theory Machine Learning (PTML) Models. *Current Topics in Medicinal Chemistry*. **2021**, *21* (9), 819-827. DOI: 10.2174/1568026621666210331161144.
2. Kamps, D.; Dehmelt, L. Deblurring Signal Network Dynamics. *ACS chemical biology*. **2017**, *12* (9), 2231-2239. DOI: 10.1021/acscchembio.7b00451.
3. Carbonell, P.; Parutto, P.; Baudier, C.; Junot, C.; Faulon, J.-L. RetroPath: Automated Pipeline for Embedded Metabolic Circuits. *ACS Synth. Biol.* **2014**, *3* (8), 565-577.
4. Stephanopoulos, G. Synthetic Biology and Metabolic Engineering. *ACS Synth. Biol.* **2012**, *1* (11), 514-525.
5. Libis, V.; Delépine, B.; Faulon, J.-L. Expanding Biosensing Abilities through Computer-Aided Design of Metabolic Pathways. *ACS Synth. Biol.* **2016**, *5* (10), 1076-1085.
6. Hadadi, N.; Hafner, J.; Shajkofci, A.; Zisaki, A.; Hatzimanikatis, V. ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. *ACS Synth. Biol.* **2016**, *5* (10), 1155-1166.
7. Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z. N.; Barabasi, A. L. The large-scale organization of metabolic networks. *Nature*. **2000**, *407* (6804), 651-654.
8. Ma, H.; Zeng, A.-P. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*. **2003**, *19* (2), 270-277.
9. Stanford, N. J.; Lubitz, T.; Smallbone, K.; Klipp, E.; Mendes, P.; Liebermeister, W. Systematic construction of kinetic models from genome-scale metabolic networks. *PLoS One*. **2013**, *8* (11), e79195. DOI: 10.1371/journal.pone.0079195.
10. Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D. U. Complex networks: Structure and dynamics. *Physics Reports*. **2006**, *424*, 175-308.
11. Bonchev, D. On the complexity of directed biological networks. *SAR and QSAR in Environmental Research*. **2003**, *14* (3), 199-214.
12. Bornholdt, S.; Schuster, H. G. *Handbook of Graphs and Complex Networks: From the Genome to the Internet*; WILEY-VCH GmbH & CO. KGa., 2003.
13. Breiger, R. The Analysis of Social Networks. In *Handbook of Data Analysis*, Hardy, M., Bryman, A. Eds.; Sage Publications, 2004; pp 505-526.
14. Jeong, H.; Mason, S. P.; Barabasi, A. L.; Oltvai, Z. N. Lethality and centrality in protein networks. *Nature*. **2001**, *411* (6833), 41-42.
15. Gonzalez-Diaz, H.; Arrasate, S.; Gomez-SanJuan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. M. General Theory for Multiple Input-Output Perturbations in Complex Molecular Systems. 1. Linear QSPR Electronegativity Models in Physical, Organic, and Medicinal Chemistry. *Current topics in medicinal chemistry*. **2013**, *13* (14), 1713-1741.
16. Box, G. E. P.; Jenkins, G. M. *Time series analysis*; Holden-Day, 1970.
17. Blazquez-Barbadillo, C.; Aranzamendi, E.; Coya, E.; Lete, E.; Sotomayor, N.; Gonzalez-Diaz, H. Perturbation theory model of reactivity and enantioselectivity of palladium-catalyzed Heck-Heck cascade reactions. *Rsc Advances*. **2016**, *6* (45), 38602-38610. DOI: 10.1039/c6ra08751e.
18. Casanola-Martin, G. M.; Le-Thi-Thu, H.; Perez-Gimenez, F.; Marrero-Ponce, Y.; Merino-Sanjuan, M.; Abad, C.; Gonzalez-Diaz, H. Multi-output Model with Box-Jenkins Operators of Quadratic Indices for Prediction of Malaria and Cancer Inhibitors Targeting Ubiquitin-Proteasome Pathway (UPP) Proteins. *Current Protein & Peptide Science*. **2016**, *17* (3), 220-227. DOI: 10.2174/1389203717999160226173500.

19. Romero-Duran, F. J.; Alonso, N.; Yanez, M.; Caamano, O.; Garcia-Mera, X.; Gonzalez-Diaz, H. Brain-inspired cheminformatics of drug-target brain interactome, synthesis, and assay of TVP1022 derivatives. *Neuropharmacology*. **2016**, *103*, 270-278. DOI: 10.1016/j.neuropharm.2015.12.019.
20. Kleandrova, V. V.; Luan, F.; Gonzalez-Diaz, H.; Ruso, J. M.; Speck-Planche, A.; Cordeiro, M. N. D. S. Computational Tool for Risk Assessment of Nanomaterials: Novel QSTR-Perturbation Model for Simultaneous Prediction of Ecotoxicity and Cytotoxicity of Uncoated and Coated Nanoparticles under Multiple Experimental Conditions. *Environmental Science & Technology*. **2014**, *48* (24), 14686-14694. DOI: 10.1021/es503861x.
21. Luan, F.; Kleandrova, V. V.; Gonzalez-Diaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, M. N. Computer-aided nanotoxicology: assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. *Nanoscale*. **2014**, *6* (18), 10623-10630. DOI: 10.1039/c4nr01285b.
22. Alonso, N.; Caamano, O.; Romero-Duran, F. J.; Luan, F.; Cordeiro, M. N. D. S.; Yanez, M.; Gonzalez-Diaz, H.; Garcia-Mera, X. Model for High-Throughput Screening of Multitarget Drugs in Chemical Neurosciences: Synthesis, Assay, and Theoretic Study of Rasagiline Carbamates. *Acs Chemical Neuroscience*. **2013**, *4* (10), 1393-1403. DOI: 10.1021/cn400111n.
23. Speck-Planche, A.; Dias Soeiro Cordeiro, M. N. Speeding up Early Drug Discovery in Antiviral Research: A Fragment-Based in Silico Approach for the Design of Virtual Anti-Hepatitis C Leads. *ACS Comb Sci*. **2017**, *19* (8), 501-512. DOI: 10.1021/acscombsci.7b00039.
24. Kleandrova, V. V.; Ruso, J. M.; Speck-Planche, A.; Dias Soeiro Cordeiro, M. N. Enabling the Discovery and Virtual Screening of Potent and Safe Antimicrobial Peptides. Simultaneous Prediction of Antibacterial Activity and Cytotoxicity. *ACS combinatorial science*. **2016**, *18* (8), 490-498. DOI: 10.1021/acscombsci.6b00063.
25. Speck-Planche, A.; Cordeiro, M. N. Computer-aided discovery in antimicrobial research: In silico model for virtual screening of potent and safe anti-pseudomonas agents. *Combinatorial chemistry & high throughput screening*. **2015**, *18* (3), 305-314.
26. Speck-Planche, A.; Cordeiro, M. N. Simultaneous virtual prediction of anti-Escherichia coli activities and ADMET profiles: A chemoinformatic complementary approach for high-throughput screening. *ACS Comb Sci*. **2014**, *16* (2), 78-84. DOI: 10.1021/co400115s.
27. Vergara-Galicia, J.; Prado-Prado, F. J.; Gonzalez-Diaz, H. Galvez-Markov Network Transferability Indices: Review of Classic Theory and New Model for Perturbations in Metabolic Reactions. *Current Drug Metabolism*. **2014**, *15* (5), 557-564.
28. Duardo-Sanchez, A.; Gonzalez-Diaz, H.; Pazos, A. MIANN Models of Networks of Biochemical Reactions, Ecosystems, and US Supreme Court with Balaban-Markov Indices. *Current Bioinformatics*. **2015**, *10* (5), 658-671. DOI: 10.2174/1574893610666151008012752.
29. Duardo-Sanchez, A.; Gonzalez-Diaz, H.; Pazos, A. MI-NODES Multiscale Models of Metabolic Reactions, Brain Connectome, Ecological, Epidemic, World Trade, and Legal-Social Networks. *Current Bioinformatics*. **2015**, *10* (5), 692-713.
30. Duardo-Sanchez, A.; Munteanu, C. R.; Riera-Fernandez, P.; Lopez-Diaz, A.; Pazos, A.; Gonzalez-Diaz, H. Modeling Complex Metabolic Reactions, Ecological Systems, and Financial and Legal Networks with MIANN Models Based on Markov-Wiener Node

- Descriptors. *Journal of Chemical Information and Modeling*. **2014**, *54* (1), 16-29. DOI: 10.1021/ci400280n.
31. Gonzalez-Diaz, H.; Arrasate, S.; Gomez-San Juan, A.; Sotomayor, N.; Lete, E.; Speck-Planche, A.; Ruso, J. M.; Luan, F.; Dias Soeiro Cordeiro, M. N. Matrix Trace Operators: From Spectral Moments of Molecular Graphs and Complex Networks to Perturbations in Synthetic Reactions, Micelle Nanoparticles, and Drug ADME Processes. *Current Drug Metabolism*. **2014**, *15* (4), 470-488.
 32. Gonzalez-Diaz, H.; Riera-Fernandez, P.; Pazos, A.; Munteanu, C. R. The Rucker-Markov invariants of complex Bio-Systems: Applications in Parasitology and Neuroinformatics. *Bio Systems*. **2013**, *111* (3), 199-207. DOI: 10.1016/j.biosystems.2013.02.006.
 33. Riera-Fernandez, P.; Munteanu, C. R.; Martin-Romalde, R.; Duardo-Sanchez, A.; Gonzalez-Diaz, H. Markov-Randic Indices for QSPR Re-Evaluation of Metabolic, Parasite-Host, Fasciolosis Spreading, Brain Cortex and Legal-Social Complex Networks. *Current Bioinformatics*. **2013**, *8* (4), 401-415.
 34. Marrero Ponce, Y.; Castillo Garit, J. A.; Torrens, F.; Romero Zaldivar, V.; Castro, E. A. Atom, atom-type, and total linear indices of the "molecular pseudograph's atom adjacency matrix": Application to QSPR/QSAR studies of organic compounds. *Molecules*. **2004**, *9* (12), 1100-1123. Scopus.
 35. Marrero-Ponce, Y. Linear indices of the "molecular pseudograph's atom adjacency matrix": Definition, significance-interpretation, and application to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors. *Journal of Chemical Information and Computer Sciences*. **2004**, *44* (6), 2010-2026. Scopus.
 36. Bediaga, H.; Arrasate, S.; González-Díaz, H. PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS combinatorial science*. **2018**, *20* (11), 621-632. DOI: 10.1021/acscmbosci.8b00090.
 37. González-Durruthy, M.; Manske Nunes, S.; Ventura-Lima, J.; Gelesky, M. A.; González-Díaz, H.; Monserrat, J. M.; Concu, R.; Cordeiro, M. N. D. S. MitoTarget Modeling Using ANN-Classification Models Based on Fractal SEM Nano-Descriptors: Carbon Nanotubes as Mitochondrial F0F1-ATPase Inhibitors. *Journal of Chemical Information and Modeling*. **2019**, *59* (1), 86-97. DOI: 10.1021/acs.jcim.8b00631.
 38. Ferreira da Costa, J.; Silva, D.; Caamaño, O.; Brea, J. M.; Loza, M. I.; Munteanu, C. R.; Pazos, A.; García-Mera, X.; González-Díaz, H. Perturbation Theory/Machine Learning Model of ChEMBL Data for Dopamine Targets: Docking, Synthesis, and Assay of New l-Prolyl-l-leucyl-glycinamide Peptidomimetics. *ACS Chemical Neuroscience*. **2018**, *9* (11), 2572-2587. DOI: 10.1021/acscchemneuro.8b00083.
 39. Martins Alho, M. A.; Marrero-Ponce, Y.; Barigye, S. J.; Meneses-Marcel, A.; Machado Tugores, Y.; Montero-Torres, A.; Gómez-Barrio, A.; Nogal, J. J.; García-Sánchez, R. N.; Vega, M. C.; et al. Antiprotozoan lead discovery by aligning dry and wet screening: Prediction, synthesis, and biological assay of novel quinoxalinones. *Bioorganic and Medicinal Chemistry*. **2014**, *22* (5), 1568-1585. DOI: 10.1016/j.bmc.2014.01.036 Scopus.
 40. Rescigno, A.; Casanola-Martin, G. M.; Sanjust, E.; Zucca, P.; Marrero-Ponce, Y. Vanilloid derivatives as tyrosinase inhibitors driven by virtual screening-based QSAR models. *Drug Test Anal.* **2011**, *3* (3), 176-181, Research Support, Non-U.S. Gov't. DOI: 10.1002/dta.187.
 41. Casañola-Martin, G. M.; Le-Thi-Thu, H.; Pérez-Giménez, F.; Marrero-Ponce, Y.; Merino-Sanjuán, M.; Abad, C.; González-Díaz, H. Multi-output model with Bob-

- Jenkins operators of linear indices to predict multi-target networks of ubiquitin-proteasome system inhibitors. *Mol. Divers.* **2015**, *19*, 347-356.
42. Hill, T.; Lewicki, P. *STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining*; StatSoft, 2006
 43. Casañola-Martín, G. M.; Khan, M. T. H.; Marrero-Ponce, Y.; Ather, A.; Sultankhodzhaev, M. N.; Torrens, F. New tyrosinase inhibitors selected by atomic linear indices-based classification models. *Bioorganic and Medicinal Chemistry Letters.* **2006**, *16* (2), 324-330. DOI: 10.1016/j.bmcl.2005.09.085 Scopus.
 44. Gonzalez-Diaz, H.; Duardo-Sanchez, A.; Ubeira, F. M.; Prado-Prado, F.; Perez-Montoto, L. G.; Concu, R.; Podda, G.; Shen, B. Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, anti-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers. *Curr Drug Metab.* **2010**, *11* (4), 379-406.
 45. Gonzalez Diaz, H.; Olazabal, E.; Castanedo, N.; Sanchez, I. H.; Morales, A.; Serrano, H. S.; Gonzalez, J.; de Armas, R. R. Markovian chemicals "in silico" design (MARCH-INSIDE), a promising approach for computer aided molecular design II: experimental and theoretical assessment of a novel method for virtual screening of fasciolicides. *J Mol Model.* **2002**, *8* (8), 237-245. DOI: 10.1007/s00894-002-0088-7.
 46. Gonzalez-Diaz, H.; Torres-Gomez, L. A.; Guevara, Y.; Almeida, M. S.; Molina, R.; Castanedo, N.; Santana, L.; Uriarte, E. Markovian chemicals "in silico" design (MARCH-INSIDE), a promising approach for computer-aided molecular design III: 2.5D indices for the discovery of antibacterials. *J Mol Model.* **2005**, *11* (2), 116-123. DOI: 10.1007/s00894-004-0228-3.
 47. Gonzalez-Diaz, H.; Prado-Prado, F.; Ubeira, F. M. Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. *Current topics in medicinal chemistry.* **2008**, *8* (18), 1676-1690.
 48. Junker, B. H.; Koschützki, D.; Schreiber, F. Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics.* **2006**, *7*, 219.
 49. Gonzalez-Diaz, H.; Arrasate, S.; Gomez-San Juan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. M. New Theory for Multiple Input-Output Perturbations in Complex Molecular Systems. 1. Linear QSPR Electronegativity Models in Physical, Organic, and Medicinal Chemistry. *Curr Top Med Chem.* **2013**, *13*, 1713-1741. From Nlm.
 50. Dieguez-Santana, K.; Pham-The, H.; Rivera-Borroto, O. M.; Puris, A.; Le-Thi-Thu, H.; Casanola-Martin, G. M. A Two QSAR Way for Antidiabetic Agents Targeting Using α -Amylase and α -Glucosidase Inhibitors: Model Parameters Settings in Artificial Intelligence Techniques. *Letters in Drug Design & Discovery.* **2017**, *14* (8), 862-868. DOI: 10.2174/1570180814666161128121142.
 51. Witten, H. I.; Frank, E. *Data Mining: Practical machine learning tools and techniques*; Morgan Kaufmann, 2005.
 52. Breiman, L. Random Forests. *Machine Learning.* **2001**, *45* (1), 5-32, journal article. DOI: 10.1023/a:1010933404324.
 53. Quinlan, R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers, 1993.
 54. Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; CRC press., 1984.
 55. Diéguez-Santana, K.; Rivera-Borroto, O. M.; Puris, A.; Pham-The, H.; Le-Thi-Thu, H.; Rasulev, B.; Casañola-Martín, G. M. Beyond Model Interpretability using LDA and Decision Trees for α -Amylase and α -Glucosidase Inhibitor Classification Studies. *Chemical Biology & Drug Design.* **2019**. DOI: 10.1111/cbdd.13518.

56. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. Chemoinformatics in anti-cancer chemotherapy: multi-target QSAR model for the in silico discovery of anti-breast cancer agents. *European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences*. **2012**, *47* (1), 273-279. DOI: 10.1016/j.ejps.2012.04.012 From Nlm.
57. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. Unified multi-target approach for the rational in silico design of anti-bladder cancer agents. *Anti-cancer agents in medicinal chemistry*. **2013**, *13* (5), 791-800.
58. Speck-Planche, A.; Kleandrova, V. V.; Cordeiro, M. N. New insights toward the discovery of antibacterial agents: multi-tasking QSBER model for the simultaneous prediction of anti-tuberculosis activity and toxicological profiles of drugs. *European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences*. **2013**, *48* (4-5), 812-818. DOI: 10.1016/j.ejps.2013.01.011 From Nlm.
59. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. Multi-target inhibitors for proteins associated with Alzheimer: in silico discovery using fragment-based descriptors. *Current Alzheimer research*. **2013**, *10* (2), 117-124. From Nlm.

CHAPTER 4. MACHINE LEARNING MAPPING OF METABOLIC NETWORKS VS. CHEMBL DATA OF ANTIBACTERIAL COMPOUNDS

Antibacterial drugs (AD) change the metabolic status of bacteria, contributing to bacterial death. However, antibiotic resistance and the emergence of Multi-drug-resistant bacterial strains increase interest in understanding metabolic network (MN) mutations and the interaction of AD *vs.* MN. In this work, we used the IFPTML = Information Fusion (IF) + Perturbation-Theory (PT) + Machine Learning (ML) algorithm for the study of a large dataset of ChEMBL database that contains >165 000 AD assays *vs.* > 40 MNs of multiple bacteria species. We built a Linear Discriminant Analysis (LDA) and 17 ML models based on the linear index based on atoms to predict antibacterial compounds. IFPTML-LDA model presented the following results for the training subset: specificity (Sp) = 76.1%, sensitivity (Sn) = 72.3%, and Accuracy (Acc) = 74.3%. Among IFPTML-non-Linear, the k Nearest Neighbors (KNN) show the best results with Sn =99.2%, Sp=95.5%, Acc=97.4% and AUROC=0.998 in training sets. IFPTML linear and non-linear models of the AD *vs.* MN have good statistical parameters, and they could contribute to finding new metabolic mutations in antibiotic resistance and reducing time/costs in antibacterial drug research.

1. INTRODUCTION

Antibiotics have become the foundation of modern medicine. However, at the beginning of 2017, the World Health Organization published a list of global priorities on antibiotic-resistant bacteria.¹ Continued efficacy is threatened by the global dissemination of antibiotic-resistance determinants, driven in large part by improper use of antibiotics in clinical, community, and agricultural settings.² To develop effective next-generation antibacterial therapies, we must better understand how bacteria respond to antibiotics.³ Molecular screens have identified compounds that limit bacterial growth *in vitro*. Despite the abundance of bioactive chemicals, only a few biological functions are targeted.⁴ Antibiotics that disrupt these energy-consuming pathways disrupt the metabolic balance.

In previous decades, Levy *et al.*⁵ argued about the limited period of clinical- utility that antibiotics have before being compensated for the inevitable emergence of resistance. Meanwhile, new antibiotics are desperately needed to combat bacterial resistance.⁶ An alternative chemical space for the abundant compounds of affected antibiotics has not yet been identified. In addition, the development of antibiotics has a low yield on the multiple diseases caused by microbes, and the antibiotic line has been operating at an alarmingly slow pace in recent decades.⁷ Most recently approved antibiotics are chemically modified derivatives of the existing drug classes; many are naturally occurring.^{8,9} Therefore, resistant strains may readily mutate to resist these analogs if their existing resistance mechanisms do not already exhibit partial cross-effectiveness.¹⁰

Furthermore, this bacterial resistance to conventional antibiotics has also been attributed to excessive broad-spectrum antibiotics,¹¹ which requires scientists to find fast, accessible, and cheap methods to discover new drugs and target molecules against infectious microorganisms. In this sense, understanding the metabolism of pathogens plays an important role. Metabolic

networks (MN) are represented by the set of metabolic pathways that are a series of biochemical reactions in which the product (output) of a reaction serves as a substrate (input) to another reaction.¹² Novel applications of MN reconstructions of human pathogens have recently been described. These studies have focused on elucidating resistance metabolic dependencies and identifying potential drug targets and antibacterials.¹³⁻¹⁵ The influence of the changes in MNs over the capacity for survival of different microorganisms has been demonstrated by Barabási's group and other authors.¹⁶

On the other hand, the importance of metabolic mutations in antibiotic resistance is frequently underestimated.¹⁷ Recently, Lopatkin *et al.*¹⁸ demonstrated that metabolic mutations arise in clinically relevant bacteria in response to antibiotic therapy. They are using a variety of in vitro evolution procedures and comprehensive sequencing data analysis. *E. coli* as a model pathogen provided proof that metabolic mutations can develop in response to antibiotic treatment.¹⁸ This research has provided a new perspective on the development of antibiotic resistance by shedding light on the complexities of metabolic alterations.³ Their findings may assist in explaining the prevalence of (multi-) drug-resistant bacterial strains isolated in areas with little or no antibiotic exposure, as well as the documented increase in antibiotic resistance following extensive herbicide or other environmentally hazardous material application.¹⁸ Antibacterial drugs (AD) change the metabolic status of bacteria, contributing to bacterial death, *e.g.*, via oxidative damage or stasis through translation inhibition, resulting in lower cellular respiration.³ The metabolic state of bacteria influences antibiotic sensitivity; hence, modifying the metabolic state of bacteria can improve antibiotic efficacy.^{3, 17} In this sense, the interaction of AD and MN can contribute to finding new metabolic mutations in antibiotic resistance, mainly toward (multi-) drug-resistant bacterial strains.

On the other hand, prediction from computer models has been widely used as an important alternative to obtain experimental evidence and save resources and research time in drug discovery and development.¹⁹ These methods allow establishing relationships between many datasets and structural molecular information that contributes to biological activity. to solve complex problems.²⁰ Additionally, machine learning (ML) allows us to process information as molecular descriptors. However, traditional techniques to extract metadata from complex preclinical assay databases are inadequate. This is the case for the ChEMBL database, which contains large datasets from various heterogeneous and autonomous sources that attempt to explore complex and evolving relationships between data²¹ from preclinical trials.²²

Numerous applications of cheminformatics and other computational approaches have been developed to aid in discovering AD against various bacteria. However, they are limited to predicting their biological activity in a certain strain under specified conditions.²³ González-Díaz *et al.* created IFPTML, a multi-output, input-coded multi-label machine learning technique, to address this type of challenge. Perturbation Theory (PT) + Machine Learning (ML) + Information Fusion (IF) is the acronym for the IFPTML algorithm.²⁴ The scoring function $f(s_{ij})$ calc is produced by the IFPTML model. IFPTML has been applied to complicated data analysis jobs in molecular sciences,^{24, 25} infectious disease,²⁶ nanotechnology,^{27, 28} *etc.*

These problems have different drugs, drug cocktails, proteins, vaccines, MN, epidemiological networks, *etc.*^{25, 29-33}

In the present work, we propose the combination of the fundamentals of Information Fusion (IF) Perturbation Theory (PT) and Machine Learning (ML) methods to build an IFPTML (PTML + IF) model as a solution for this type of data.^{27, 34-36} This model is especially suitable for databases with similar large data features and combinatorial information. This paper analyzed a large dataset (>155000 preclinical assays) against different bacterial strains downloaded from the ChEMBL database. We merged this dataset with structural information for >40 MNs from different microorganisms reported by Barabási's group.¹⁶ In all these cases, those without biological values, measurements, or assay conditions were removed.³⁷ For this purpose, we used Moving Average (MAs) operators to express the perturbations in the assays and PT Multiplier Operators (PTOs) to perform data fusion and dimension reduction. Last, we applied linear discriminant analysis (LDA) and non-linear ML algorithms to find the best IFPTML predictive model. The general workflow used for the IFPTML model for AD vs. MN is shown in **Fig 4.1**.

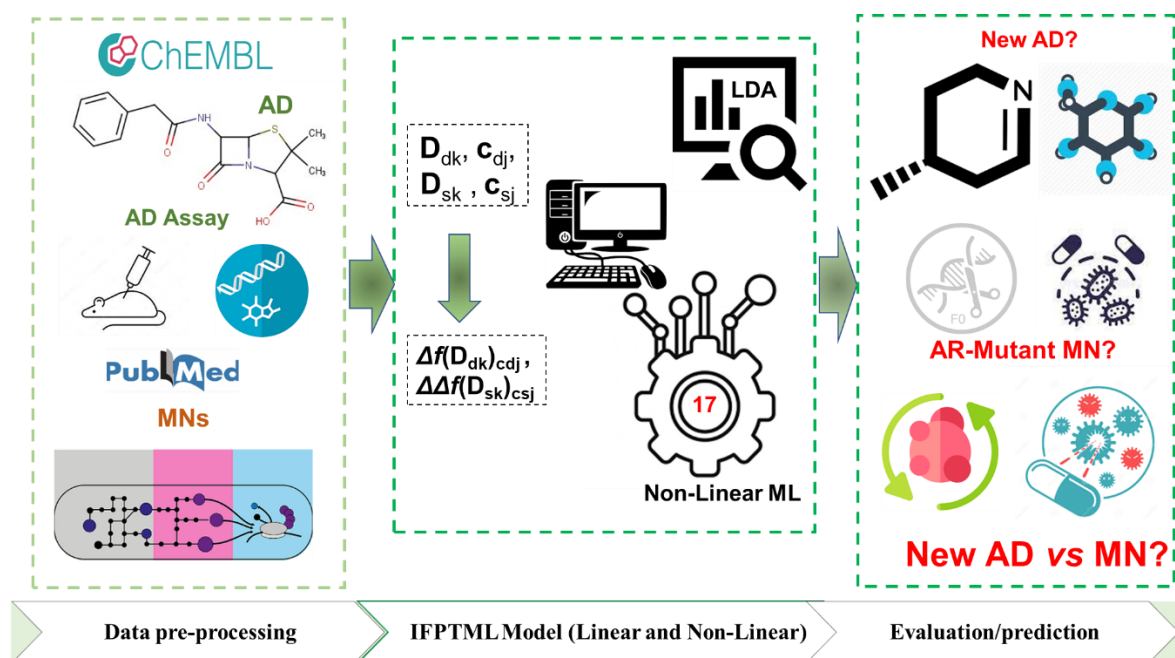


Figure 4.1. IFPTML model for AD vs. MN development workflow.

2. MATERIALS AND METHODS

2.1 ChEMBL data set of antibacterial compounds

We downloaded a large dataset of preclinical assays of ADs from the ChEMBL database. The dataset was created through a data fusion process between the ChEMBL dataset and Barabási's group MNs released by Jeong *et al.*¹⁶ In this sense, we only searched in the ChEMBL database biological activity assays of ADs against organisms present in the MNs dataset. The steps carried out were the following:

In the ChEMBL, the different organisms were searched by targets and assays and saved in an Excel file. Subsequently, we merge the datasets obtained with each keyword into a single file. Later, we performed the data curation, eliminating all duplicate cases and reporting no biological activity value. The data of the organisms *Methanococcus jannaschii* and *Treponema pallidum* are excluded since the two compounds reported in the ChEMBL have no biological activity measured by what they are not considered. After data curation, we found that the ChEMBL AD activity dataset contains the values of >300 parameters (MIC, IC₅₀, etc.) for >155000 biological assays of >50000 compounds vs. >25 bacteria species with >90 strains.

2.2 IFPTML analysis steps

IFPTML analysis has different steps that can be separated into three phases (IF + PT + ML). **Fig. 4.2** depicts the IFPTML method workflow for AD vs. MN analysis, including the general procedures described in this paper. The first step of the IF phase is to obtain values v_i , and v_j for the different biological properties c_{d0} and c_{s0} of the two subsystems (AD and MN). Next, we need preprocess all the observed values with different units, scales, degrees of uncertainty, etc. to obtain dimensionless functions characterizing the system as a whole, AD vs. MN cases. Barabási's group released the MN dataset as gzipped ASCII files.¹⁶ The numbers of nodes (metabolites), input-output links (metabolic reactions), node degree, topological indices, full names, and codes of >40 bacteria species studied here appear in **Table S02** (Supporting Information **S00**). In the IF approach, the chemical compounds' structures of ADs ($f_k(D_i)$ values) were fused with structural information included in the MNs datasets of the various species.

The output $f(v_{ij})_{\text{calc}}$ was calculated as a linear combination of scores for various c_i . c_j is a generic term that refers to a variety of multi-output assay circumstances, such as targets, assays, organisms, strains, and MN. In this sense, c_0 is the biological activity v_{ij} Minimal Inhibitory Concentration (MIC ($\mu\text{g}\cdot\text{mL}^{-1}$)) or Minimal Bactericide Concentration (MBC ($\mu\text{g}\cdot\text{mL}^{-1}$)), etc, c_1 is the specific protein (ChEMBL database), c_2 is the assay organism in experiment, c_3 is the specific strain of assay organism, c_4 is MN microorganism specie, c_5 is the target type, and c_7 is mappings to the ChEMBL targets. Then, the parameters f_k , $\Delta f_k(c_q)$, and $\Delta\Delta f_k(c_q)$ are the independent input variables and $f(v_{ij})=1$ is the input dependent variable. The molecular descriptors D_{ik} , of linear indexes based on atoms, include $f_q(N, M, w)_g$, for each chemical q_{th} . **Eq. 1** shows the general definition of linear indexes based on atoms (Eq. 1).

$$f_{qk}(G, N_1, M, w)_g = f_{qk}(w)_g = \sum_{i=1}^{n_g} |f_i|_g \quad (1)$$

Where N_1 is the selected matrix norm (Manhattan distance), M is the graphic-theoretical electronic density matrix. While (w) is physicochemical weight used. In this case, Ghose-Crippen LogP, electronegativity and volume of van der Waals. Finally, the different groups of atoms calculated for the compounds were: H (A) bond acceptors, C atoms in the aliphatic chain (C), donors of the H link (D), C atoms in the aromatic portion (P) and heteroatoms (X).³⁸

Next, we must define and obtain/calculate the values of all vectors corresponding to the structural descriptors D_{dk} and D_{sk} for the two subsystems. Additionally, we must define and obtain/calculate the vector elements c_{dj} and c_{sj} with all AD and MN bacteria labels/assay

conditions. Following that, we transformed the estimated molecular descriptors \mathbf{D}_{dk} and \mathbf{D}_{sk} to Box–Jenkins MA operators. The PTOs estimated in this work include the chemical structure and/or physicochemical properties of the AD subsystem $\Delta f(\mathbf{D}_{dk})$, as well as structural information about the bacteria's MN $\Delta \Delta f(\mathbf{D}_{sk})$. They were written in the form of deviation terms for each subsystem $f(\mathbf{D}_{dk})$ and $f(\mathbf{D}_{sk})$ with respect to the average value for the respective subsystems of reference $\langle f(\mathbf{D}_{dk})_{cdj} \rangle$ and $\langle f(\mathbf{D}_{sk})_{csj} \rangle$. As a result, the initial terms $f(\mathbf{D}_{dk})$ and $f(\mathbf{D}_{sk})$ in these formulas denote the subsystem, while the averages denote the assay. The following equations were utilized (**Eq. 2-3**)

$$\Delta f(\mathbf{D}_{dk}) = f(\mathbf{D}_{dk}) - \langle f(\mathbf{D}_{dk}) \rangle_{cdj} \quad (2)$$

$$\Delta \Delta f(\mathbf{D}_{sk}) = f(\mathbf{D}_{sk}) - \langle f(\mathbf{D}_{sk}) \rangle_{csj} \quad (3)$$

2.3 IFPTML linear model

The IFPTML model was obtained from the merger of several cheminformatics methods. The output of the IFPTML model is the scoring function values $f(v_{ij})_{calc}$ for the biological activity of the i^{th} compound assayed in the j^{th} preclinical assay with conditions $c_j = (c_0, c_1, \dots, c_7)$ against the s^{th} bacteria species with MNs. The model starts with a value of reference $f(v_{ij})_{ref}$ and adds the effect of perturbations (PT operators) in the conditions of assay, or the bacteria strain used, etc. The PT operators Δf_k based on Box–Jenkins moving average (MA) operators has been used in previous works to solve different problems.^{32, 39, 40} The linear classification models were developed using Linear Discriminant Analysis (LDA). **Eq.4** shows the general form of the IFPTML linear models.

$$f(v_{ij})_{calc} = a + b \cdot f(v_{ij})_{ref} + \sum_{k=0}^{k=5} c_k \cdot \Delta f(\mathbf{D}_{dk}) + \sum_{k=0}^{k=5} d_k \cdot \Delta \Delta f(\mathbf{D}_{sk}) \quad (4)$$

The following statistical parameters were utilized to validate the model: the number of training examples (N), and the overall values of Model quality was determined using parameters such as Sensitivity (Sn), Specificity (Sp), Chi-square (χ^2), and the p-level. LDA algorithms was run using the STATISTICA 6.0 program⁴¹ **Fig. 4.2.** shows the IFPTML information processing detailed workflow.

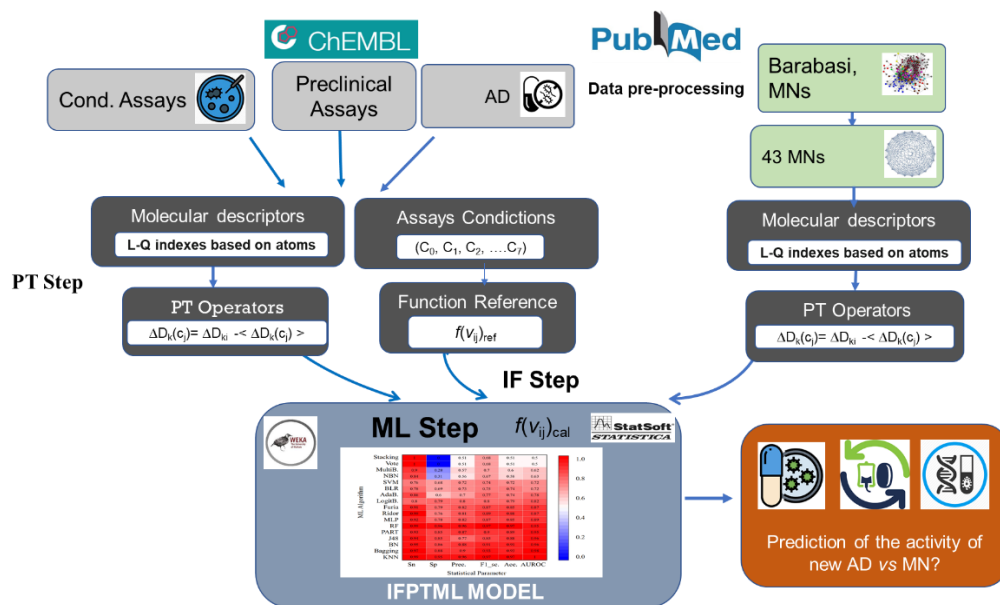


Figure 4.2. IFPTML information processing detailed workflow.

2.4 IFPTML non-linear models

Next, we decided to run several non-linear ML algorithms developed using the Waikato Environment for Knowledge Analysis (WEKA) software package, version 3.8.5.⁴² In total, we used 17 ML algorithms to build these alternative non-linear IFPTML classification models from the present dataset. These included classifiers such as Bayesian networks, decision trees, ensemble methods, rule-based classifiers, neural networks, and functions. Each technique adopts a learning algorithm to identify the model that best fits the relationship between the input data set and the class. The Bayesian Network K2/B (BN) and Nave Bayes (NBN) classifiers were based on Bayes' theorem. The classification trees applied were the pruned or unpruned C4.5 decision tree (J48) developed by Ross Quinlan⁴³ and the Random Forest (RF) classifier.⁴⁴ This technique is an extension of Bagging, with the addition of randomized feature selection. RF first selects a subset of features at random, then performs the traditional split selection technique inside the selected feature subset.⁴⁵

Different ensemble methods were used. They include meta-algorithms that aim at combining weak learners' skills such as bagging, boosting, voting, and stacking. In the first case, bagging methods are used to lower the variance of a base estimator (*e.g.*, decision tree) before constructing an ensemble from it. They are a quick and easy technique to improve a single model without changing the fundamental base algorithm.⁴⁵ An implementation of CART (SimpleCart) was applied based on classifiers trees in the Weka package.⁴⁶ The second group is the boosting algorithms that are capable of transforming weak learners into strong ones. Intuitively, a weak learner does little better than a random guess, whereas a strong learner performs quite near to perfect.⁴⁵ In this work, we used Adaboost, LogitBoost, and MultiBoosting, which are three representative algorithms of this family of algorithms.⁴⁷ These models were built in conjunction with classifiers trees based on entropy (DecisionStump). Voting is a straightforward ensemble procedure that generates two or more sub-models. Each

sub-model provides predictions that are merged in some way, such as by taking the mean or the mode of the predictions, allowing each sub-model to vote on what the outcome should be.⁴⁸ The last, Stacking is a general method considered as a simple extension to voting ensembles, where an individual learner is combined. Individuals are considered first-level learners, while combiners are called second-level or meta-learners.⁴⁵ In this work, the meta classifier ZeroR was used as the base model.

Artificial neural network (ANN) is a non-linear classification approach inspired by biological neural networks. Objects (compounds) are represented by feature vectors. Each feature is sent to an input neuron with a weight. Input is routed to the output layer via hidden layers based on these weights.⁴⁹ The output layer mixes these signals (*e.g.*, activity or class prediction). Weights are initially set at random. The weights are changed as the network is fed data, so that the overall output approximates the observed endpoint values for the chemicals.⁵⁰ In our work the “hidden” layer was proved from 2-13 ranging in 1-unit steps (a single layer of hidden) to predict of the antibacterial compounds.

Other functions such as Support vector machine (SVM), k Nearest Neighbors (KNN), and Binary Logistic Regression (BLR) were implemented. SVM is a method that works well with noisy data.⁵¹ Identifying a stiff choice hyperplane that leads to the greatest possible margins across activity classes leads to models. For non-linear data, use kernels to transpose the original feature space to higher dimensions. In this work, we used the polynomial kernel function. KNN is a lazy learning classification method, that allocates new compounds to the most prevalent class of known compounds in their near neighborhood.⁵² Several combinations of parameters were proven. One of them, the number of nearest neighbors (k), was varied from 1-20 (in 1-unit step). In addition, we employed the four distances (Chebyshev, Edit, Euclidean and Manhattan) of the LinearNNSearch in a feature space. Finally, BLR is an algorithm that can be used for predicting a categorical variable (*e.g.*, Yes/No, Pass/Fail) using a set of independent variables (*s*).^{53, 54}

In the case of the Rule-based classifiers, three methods were applied. PART is a decision list that builds a partial C4.5 decision tree in each iteration and transforms the best leaf into a rule,⁵⁵ Ripple-Down Rule (Ridor) learner generates a default rule and then the exceptions to the default with the least (weighted) error rate. The exceptions are a set of rules that predict classes other than those chosen by the default,⁵⁶ and the Fuzzy Unordered Rules Induction Algorithm (FURIA) is a novel fuzzy rule-based classification method introduced by Hühn and Hüllermeier.⁵⁷

The performance metrics used were Area Under Receiver Operating Characteristic (AUROC), Accuracy (Acc), Sn, Sp, Precision, and F1 score.

Domain of Applicability (DoA). Producing reliable forecasts necessitates an understanding of the model's limitations and applicability. The Domain of Applicability (DoA) can be defined either using similarity measures based on Euclidean distances between all training and test composites or with the leverage approach.^{58, 59} We employ the leveraging technique. After calculating the hat matrix for the structural domain, the residuals and LOO residuals of the

response variables were plotted against the leverages (the diagonal values of the hat matrix (h) in order to visually define the DoA. (Williams plot).⁶⁰ Chemicals that exceeded specified threshold values were identified as outliers in terms of reactivity and leverage. Three residuals were used as response thresholds. Leverage was set to the critical hat value ($h^* = 3(p+1)/n$, where p denotes the number of model descriptors and n is the number of training compounds.⁶⁰ Gramatica⁶¹, classified ($h > h^*$) as a structurally significant chemical. In addition to test series, the DoA was performed for an external series composed of 224719 compounds (without antibacterial activity).

3. RESULTS AND DISCUSSION

3.1 IFPTML linear model.

The IFPTML model projected is the combination of PTML modeling and Information Fusion (IF) procedures. The model starts with the expected value of biological activity and incorporates the effect of different perturbations in the system. The model has two input variables: the expected-value function $f(v_{ij})_{ref}$ and the Δf , $\Delta\Delta f$ PT operators. In **Table 4.1**, we show selected variables of the IFPTML-LDA model for the different conditions used in the model. The criteria selected are those expected to be more relevant in biological activity (AD vs. MR) terms.

Table 4.1. IFPTML workflow variables model.

Phase	Step	Name	Symbol	Information	Formula / Description
IF	0	Value	v_{ij}	Biological Activity	Value v_{ij} (MIC, MBC, <i>etc.</i>) of the parameter (labeled c_0) determined for the i^{th} compound under assay conditions $c_j = [c_0, c_1, c_2 \dots c_{\text{max}}]$
	1	Objective function	$f(v_{ij})_{obs}$	Biological Activity	$f(v_{ij})_{obs} = 1$ IF ($v_{ij} > \text{cutoff}_j$ AND $d(c_0) = 1$) OR ($v_{ij} < \text{cutoff}_j$ AND $d(c_0) = -1$) ELSE $f(v_{ij})_{obs} = 0$ Boolean variable obtained from the original biological activity value v_{ij}
	2	Reference Function	$f(v_{ij})_{ref}$	Drugs Chemical Structure	f_{14q} Expected value of Linear indices (C atoms in aliphatic chain/Non-Stochastic Matrix Order 2)
PT	3		Δf_i	Drug structure vs. Protein accession	$[d_{14q} - <d_{14q}(c_{1q})>]$ Account for variability on Linear indices (C atoms in aliphatic chain/Non-Stochastic Matrix Order 2) of the Drug structure of metabolite q in the MN, under same conditions c_1 (specific protein of the ChEMBL database)

		[d _{14q} - <d _{14q} (c _{4q})>]
<i>4f₂</i>	Drug structure vs. MN Microorganism	Account for variability on Linear indices (C atoms in aliphatic chain/Non-Stochastic Matrix Order 2) of the Drug structure of metabolite q in the MN, concerning MN Microorganism (c ₄)
		[d _{15q} - <d _{15q} (c _{7q})>]
<i>4f₃</i>	Drug structure vs. Target mapping ChEMBL	Account for variability on Linear indices (C atoms in aliphatic chain/Non-Stochastic Matrix Order 3) of the Drug structure of metabolite q in the MN, under conditions c ₇ (Mappings to ChEMBL targets). It included different Target Mapping ChEMBL such as Non-molecular, Protein Unassigned, Homologous protein, Multiple proteins, Multiple homologous proteins, Homologous protein complex, Molecular (non-protein), Protein complex.
		[d _{14q} - <d _{14q} (c _{5q})>]
<i>4f₄</i>	Drug structure vs. Target type	Account for variability on Linear indices (C atoms in aliphatic chain/Non-Stochastic Matrix Order 2) of the Drug structure of metabolite q in the MN, under conditions c ₅ (Different target types). It included different types of ChEMBL targets as Organism, Single protein, Unchecked, Cell-line, Nucleic-Acid, Protein complex, ADMET, Protein family, No target, Tissue, Protein complex group, Protein-protein interaction.
		[d _{15q} - <d _{15q} (c _{1q})>]
<i>4f₅</i>	Drug structure vs. Protein accession	Account for variability on Linear indices (C atoms in aliphatic chain/Non-Stochastic Matrix Order 3) of the Drug structure of metabolite q in the MN, respect to a specific protein in a ChEMBL database (c ₁).
<i>4f₆</i>		[d _{15q} - <d _{15q} (c _{5q})>]

	Drug structure vs. Target type	Account for variability on Linear indices (C atoms in aliphatic chain/Non-Stochastic Matrix Order 3) of the Drug structure of metabolite q in the MN, under same conditions c_5 (Different target types).
		$[d_{01o} - \langle d_{01o}(c_{1o}) \rangle] - [d_{01s} - \langle d_{01s}(c_{5s}) \rangle]$
	ΔAf_1 Metabolic Network structure vs. Protein Accession	Account for variability on Linear indices (Global indices /Non-Stochastic Matrix Order 1) of the query organism o and the organism of reference s in the MN, for the same specific protein in a ChEMBL database (c_1).
		$[d_{00o} - \langle d_{00o}(c_{3o}) \rangle] - [d_{00s} - \langle d_{00s}(c_{3s}) \rangle]$
	ΔAf_2 Metabolic Network structure vs. Assay Strain	Account for variability on Linear indices (Global indices /Non-Stochastic Matrix Order 0) of the query organism o and the organism of reference s in the MN, with respect to the structure of the drugs assayed against the same strain of assay organism (c_3).
		$[d_{02o} - \langle d_{02o}(c_{4o}) \rangle] - [d_{02s} - \langle d_{02s}(c_{4s}) \rangle]$
4	ΔAf_3 Metabolic Network structure vs. MN Microorganism	Account for variability on Linear indices (Global indices /Non-Stochastic Matrix Order 2) of the query organism o and the organism of reference s in the MN, with respect to the same MN Microorganism (c_4)
		$[d_{03o} - \langle d_{03o}(c_{4o}) \rangle] - [d_{03s} - \langle d_{03s}(c_{4s}) \rangle]$
	ΔAf_4 Metabolic Network vs. MN Microorganism	Account for variability on Linear indices (Global indices /Non-Stochastic Matrix Order 3) of the query organism o and the organism of reference s in the MN, with respect to same MN Microorganism (c_4)
		$[d_{03o} - \langle d_{03o}(c_{5o}) \rangle] - [d_{03s} - \langle d_{03s}(c_{5s}) \rangle]$
	ΔAf_5 Metabolic Network structure vs. Target Type	Account for variability on Linear indices (Global indices /Non-Stochastic Matrix Order 3) of the query organism o and the organism of reference s in the MN, with the same types of ChEMBL targets.

	5	Output Function	$f(v_{ij})_{calc}$	Score of Biological Activity	$f(v_{ij})_{calc} = a + b_k f(v_{ij})_{ref} + c_k \cdot \Delta f(D_{dk}) + d_k \cdot \Delta \Delta f(D_{sk})$ Real valued output of the model
ML	6	Predicted Probability	$p(f(v_{ij})_{obs} = 1)$	Score of Biological Activity	$p(f(v_{ij})_{obs} = 1) = 1 / (1 - (\pi_0) / (\pi_1)) \cdot \exp(-f(v_{ij})_{calc})$ Predicted probability of $f(v_{ij})_{obs} = 1$
	7	Predicted Class	$f(v_{ij})_{obs}$	Predicted Class	$f(v_{ij})_{obs} = 1$ IF $p(f(v_{ij})_{obs} = 1) > 0.5$ ELSE $f(v_{ij})_{obs} = 0$ Predicted Biological Activity Class

The probabilities used a priori to fit the model were set $\pi_0(f(v_{ij}=0)) = \pi_1(f(v_{ij}=1)) = 0.5$. The molecular descriptors were transformed to Box–Jenkins moving averages. Two Duplex Linear Indices Atom based Level descriptors were used (with C atoms in aliphatic chain and Total (Global) indices). In the first, Non-Stochastic Matrix Order 2 and 3 were included in the model. In the second, the Non-Stochastic Matrix Order varied from 0 to 3. The output of the model v_{ij} is a scoring function of the biological activity value of the i^{th} drug in the different combinations of conditions of assay c_{sj} and c_{dj} . The classification of one compound as active in this work is based on the desirability $d(c_0)$ of the biological property $v_{ij}(c_0)$ and the predefined value of cutoff. The threshold value of the biological activity $v_{ij}(c_0)$ (MIC) to consider one compound as active or not was set as less than $4213 \mu\text{g}\cdot\text{mL}^{-1}$ or less than the average for properties non measured. The drugs were considered to be active ($f(v_{ij})_{obs}=1$) when $v_{ij} >$ cutoff and priori desirability function $d(c_0) = 1$; then $f(v_{ij})_{obs}=1$. Furthermore, if $v_{ij} <$ cutoff and $d(c_0) = -1$ then $f(v_{ij})_{obs}=1$; otherwise, $f(v_{ij})_{obs}=0$. When we want to maximize the value of biological activity $s_{ij}(c_0)$, for example, inhibition (%), the desirability $d(c_0) = 1$. On the contrary, $d(c_0) = -1$ when we want to minimize the value of biological activity $v_{ij}(c_0)$; for example, potency (nM), IC_{50} (nM), K_i (nM), or EC_{50} (nM). Otherwise, when the necessity of maximizing or minimizing $v_{ij}(c_0)$ is unclear, the value of desirability was assumed to be $d(c_0) = 0$. In any case, the values of $d(c_0)$ for the same property may be customized (switched) for a specific situation.⁶²

Eq 5 show detailed explanation about all the input variables analyze, and the equation of the best model found is the following:

$$f(v_{ij}) = -5.475 + 0.023 \cdot f(v_{ij})_{ref} - 0.047 \cdot \Delta f_1 - 0.008 \cdot \Delta f_2 - 0.002 \cdot \Delta f_3 + 0.030 \cdot \Delta f_4 + 0.014 \cdot \Delta f_5 - 0.011 \cdot \Delta f_6 + 0.829 \cdot \Delta \Delta f_1 - 0.178 \cdot \Delta \Delta f_2 - 3.349 \cdot \Delta \Delta f_3 + 2.229 \cdot \Delta \Delta f_4 + 0.498 \cdot \Delta \Delta f_5 \quad (5)$$

$$N=115662, \chi^2=25774.24, p<0.01$$

Statistical parameters of the model are N is the number of cases applied to train the model, χ^2 is the Chi-square statistics, and p is the p-level.

As shown in **Eq. 9**, the parameters Δf_1 , Δf_2 , Δf_3 , Δf_6 , $\Delta \Delta f_2$, and $\Delta \Delta f_3$ all have a negative effect on the numerical score of the biological activity; these parameters correspond to the boundary conditions for the measure, target, and data curation, respectively. On the other hand, the

variables $f(v_{ij})_{\text{ref}}$, Δf_4 , Δf_5 , $\Delta \Delta f_1$, $\Delta \Delta f_4$, and $\Delta \Delta f_5$ (protein, MN organism, and target type) all influence the activity positively. Additionally, we may obtain the parameters that contribute the most to the activity using this equation. In the instance of $\Delta \Delta f_4$, the coefficient is 2.229, which is a very realistic value considering that the most significant variations in activity, even among identical compounds, are explained by the diverse techniques employed to assess the activity. The same holds true for the $\Delta \Delta f_3$ parameter, which has a coefficient of 3.349 in the equation and contributes significantly negatively to activity.

Table 4.2 shows the classification matrices and summarizes the results in terms of Sn = sensitivity (%), Sp = specificity (%), and Acc = accuracy (%) for training and validation series. The IFPTML-LDA model presented very high-performance parameters in both training and validation series. The cases in training and validation series were selected with a random, stratified, and representative sampling. The obtained IFPTML model classified correctly the ~74.3% of the cases in the training and validation set. Both series have adequate values of sensitivity (Sn) and specificity (Sp) ~76%, and 72%, respectively. In general, the IFPTML model has a good performance for describing the correct/incorrect connectivity pattern as showed in the performance of the statistical parameters of the current classification equation.

Table 4.2. IFPTML linear model results for ChEMBL AD vs. MNs.

Series	Set	Stat. Param ^a	%	$f(v_{ij})_{\text{pred}}=0$	$f(v_{ij})_{\text{pred}}=1$
Training	$f(v_{ij})_{\text{pred}}=0$	Sp	76.1	45254	14227
	$f(v_{ij})_{\text{pred}}=1$	Sn	72.3	15548	40633
	Total	Acc	74.3		
Validation	$f(v_{ij})_{\text{pred}}=0$	Sp	76.2	15107	4719
	$f(v_{ij})_{\text{pred}}=1$	Sn	72.1	5219	13507
	Total	Acc	74.2		
Screening	$f(v_{ij})_{\text{pred}}=0$	Sp	-	0	0
	$f(s_{ij})_{\text{pred}}=1$	Sn	72.3	62243	162476
	Total	Acc	72.3		

^aSn=sensitivity (%), Sp=specificity (%), and Acc= accuracy (%). The positive (1) and negative control cases (0) were assigned as follows: if a priori desirability function $d(c_0)=-1$, then $f(v_{ij})_{\text{obs}}=1$ when $s_{ij}<\text{cutoff}$. In addition, if $d(c_0)=1, 0$, then $f(v_{ij})_{\text{obs}}=1$ when $v_{ij}>\text{cutoff}$; otherwise, $f(v_{ij})_{\text{obs}}=0$.

3.2 IFPTML Non-Linear models.

We also trained another type of IFPTML model using a different class of ML algorithms. Specifically, we used 17 ML classifiers. The performance of these models is summarized in **Table 4.3**, and the graphical representation of the results can be visualized in **Figures 4.3 and 4.4**.

Table 4.3. IFPTML-Non-linear AD vs. MN systems models.

Models ^a	Sub-set ^b	Stat. ^c	Val. (%)	Class	Observed		AURO C ^d
				Pred.	1	0	
KNN	t	Sn	99.18	1	58991	2549	0.998

		Sp	95.46	0	490	53632	
	v	Sn	91.92	1	18224	2446	0.924
		Sp	86.94	0	1602	16280	
RF	t	Sn	98.63	1	58669	2229	0.953
		Sp	96.03	0	812	53952	
	v	Sn	93.96	1	18628	2430	0.945
		Sp	87.02	0	1198	16296	
Bagging	t	Sn	97.46	1	57969	6722	0.982
		Sp	88.04	0	1512	49459	
	v	Sn	95.86	1	19005	2823	0.96
		Sp	84.92	0	821	15903	
BN	t	Sn	95.48	1	56791	7870	0.964
		Sp	85.99	0	2690	48311	
	v	Sn	93.91	1	18619	2970	0.947
		Sp	84.14	0	1207	15756	
J48-DT	t	Sn	93.90	1	27684	8160	0.958
		Sp	85.48	0	1797	48021	
	v	Sn	96.00	1	2976	22009	0.944
		Sp	84.11	0	15750	16543	
Part	t	Sn	93.06	1	55352	8508	0.955
		Sp	84.86	0	4129	47673	
	v	Sn	92.41	1	18321	2972	0.946
		Sp	84.13	0	1505	15754	
MLP	t	Sn	92.10	1	54783	12241	0.888
		Sp	78.21	0	4698	43940	
	v	Sn	92.02	1	18243	4138	0.885
		Sp	77.90	0	1583	14588	
FURIA	t	Sn	91.31	1	54315	11705	0.871
		Sp	79.17	0	5166	44476	
	v	Sn	91.45	1	18131	3967	0.869
		Sp	78.82	0	1695	14759	
Ridor	t	Sn	98.67	1	58687	13452	0.874
		Sp	76.06	0	794	42729	
	v	Sn	98.31	1	19490	4615	0.868
		Sp	75.36	0	336	14111	

LogitBoost	t	Sn	79.87	1	47506	12025	0.819
		Sp	78.60	0	11975	44156	
	v	Sn	79.84	1	15830	4078	0.817
		Sp	78.22	0	3996	14648	
AdaBoost	t	Sn	86.14	1	51234	22219	0.783
		Sp	60.45	0	8247	33962	
	v	Sn	85.98	1	17047	7527	0.782
		Sp	59.80	0	2779	11199	
BLR	t	Sn	77.91	1	46343	17405	0.722
		Sp	69.02	0	13138	38776	
	v	Sn	77.99	1	15463	5867	0.769
		Sp	68.67	0	4363	12859	
SVM	t	Sn	76.09	1	45257	17959	0.721
		Sp	68.03	0	14224	38222	
	v	Sn	76.02	1	15072	6050	0.719
		Sp	67.69	0	4754	12676	
MultiBoostA B	t	Sn	89.56	1	53274	40450	0.623
		Sp	28.00	0	6207	15731	
	v	Sn	89.57	1	17759	13482	0.622
		Sp	28.00	0	2067	5244	
NBN	t	Sn	84.04	1	49988	38683	0.628
		Sp	31.15	0	9493	17498	
	v	Sn	84.02	1	16657	12900	0.629
		Sp	31.11	0	3169	5826	
Stacking (ZeroR)	t	Sn	100.00	1	59481	56181	0.5
		Sp	0.00	0	0	0	
	v	Sn	100.00	1	19826	18726	0.5
		Sp	0.00	0	0	0	
Vote	t	Sn	100.00	1	59481	56181	0.5
		Sp	0.00	0	0	0	
	v	Sn	100.00	1	19826	18726	0.5
		Sp	0.00	0	0	0	

^aML-Classification Models. kNN= k Nearest Neighbors, RF= Random Forest, Bagging, BN= Bayes network, J48-DT=J48 decision tree, Part, MLP= Multi-Layer Perceptron. FURIA= Fuzzy Unordered Rules Induction Algorithm, Ridor= Ripple-Down Rule, LogitBoost, AdaBoost, BLR= Binary Logistic Regression, SVM= Support Vector Machines, MultiBoostAB, NBN= Naïve Bayes, Stacking (ZeroR), and Vote. ^b Sub-set. t=: Training set, v= Validation set. ^c Stat. Statistical performance. Sn=Sensibility, Sp= Specificity. ^dAUROC: Area under ROC value.

As expected, almost 10 of the 17 ML models displayed better Sn and Sp values than the IFPTML-LDA model. They are KNN, RF, Bagging, BN, J48-DT, Part, MLP, FURIA, Ridor, and LogitBoost. However, AdaBoost, BLR, SVM, MultiBoostAB, NBN, Stacking (ZeroR), and Vote show a lower values of Sp than the IFPTML-LDA model. In the case of the Stacking (ZeroR), and Vote (Sn=0%) and AUROC=0.5, it indicates that classification is no better than random guessing. Thus, these techniques are not suitable for AD vs.MN data processing. In terms of accuracy, the first ten algorithms mentioned also presented good performance, with a global Ac= 80 – 97.4%, suggesting that this dataset herein is predominated by non-linear classification.

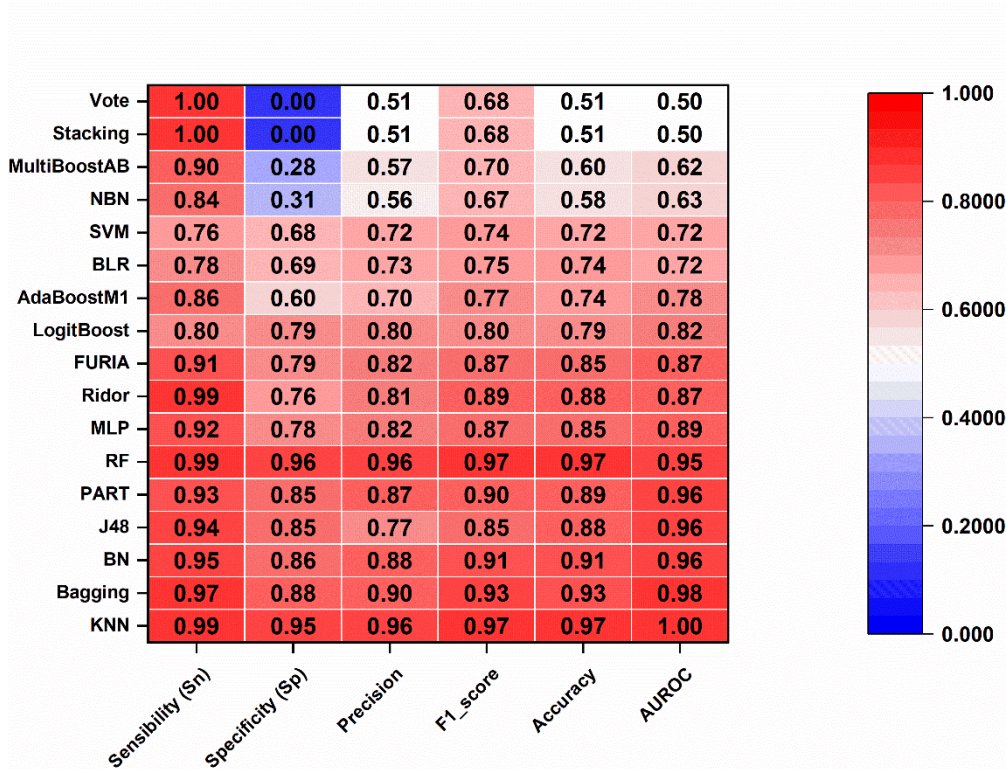


Figure. 4.3. Detailed score for the Training Set considering 17 ML techniques applied.

Sn: Sensibility, Sp: Specificity, Prec.: Precision, F1_sc.: F1 score, Acc.: Accuracy, and AUROC: Area under ROC value.

Otherwise, in the validation set, the same algorithms KNN, RF, Bagging, BN, J48-DT, Part, MLP, FURIA, Ridor, LogitBoost are superior to the IFPTML-LDA model. In addition, these techniques display adequate goodness-of-fit and goodness-of-prediction. They are consistently performing well on both the training and test sets (see **Table 4.3**). In particular, the Sn rates for active and inactive classes are >91%, suggesting high discriminant ability for further virtual screening applications.

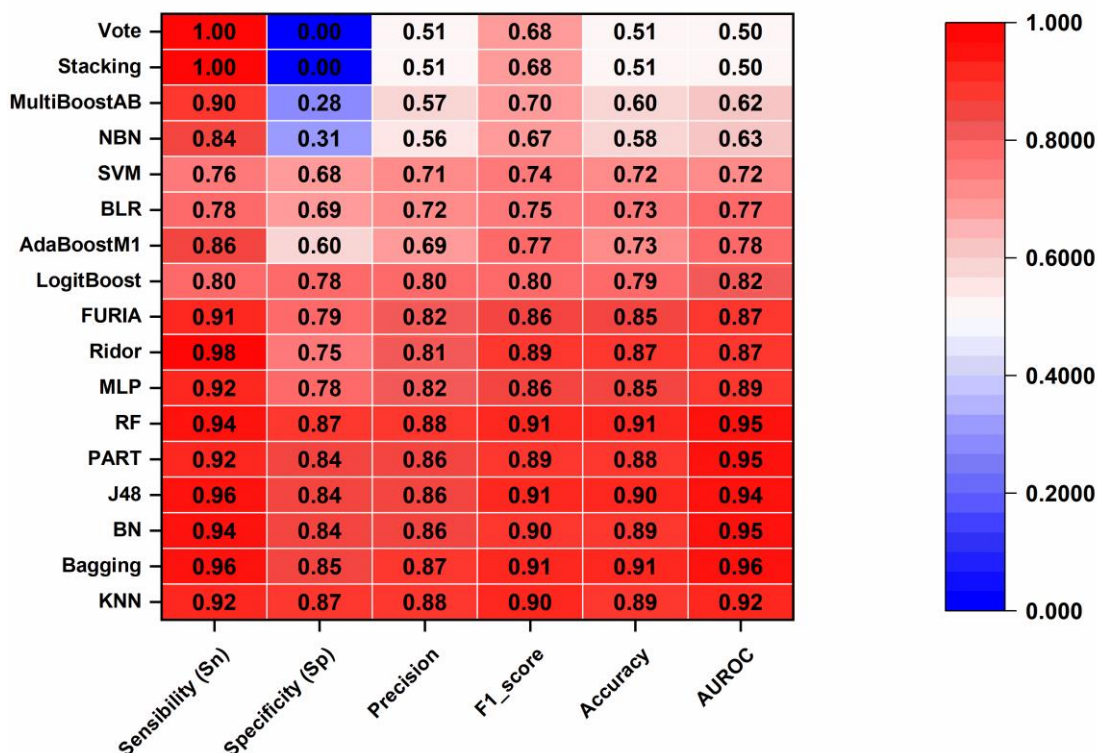


Figure. 4.4. Detailed score for the Test Set considering 17 ML technique applied.

Sn: Sensibility, Sp: Specificity, Prec.: Precision, F1_sc.: F1 score, Acc.: Accuracy, and AUROC: Area under ROC value.

In the training/validation set, the KNN, Bagging, BN, J48, PART, and RF show AUROC > 0.95. The ROC curve is created by plotting the true positive rate against the false-positive rate at different thresholds. Values close to 1 indicate that classification is almost perfect across all thresholds; thus, these six techniques are considered good classifiers for a dataset. They are the most accurate models as determined by a consensus examination of their overall accuracy and AUROC parameters. However, the improvement from LDA to ML models was not considerable and selecting a model suitable for virtual screening assays is challenging.

Domain of Applicability (DoA).

The DoA of the IFPTML-LDA model is illustrated in Figure 4.5, as a double ordinate plot of residuals test sets (first ordinate) and plot of residuals external validation (second ordinate) vs. leverages (abscissa) (William Plot). Within the domain, the examples fall within a rectangular area defined by a band of two residuals and a leverage threshold of $h = 0.00033$.^{19, 63, 64} As can be observed, the majority of test and validation examples fall inside this range. There are, however, a significant number of examples with leverage greater than the threshold but with LOO and standard residuals under the limits. In these instances, where the leverage value is greater than h^* , the prediction should be regarded as untrustworthy. Greater than warning leverage (h^*) indicates that the composite's expected reaction can be extrapolated from the model, and hence the predicted value should be used with extreme caution. As a result, there are no instances in either the training or prediction series where the residual values are greater

than the range defined for residuals and residual LOO. As a result, there are no outliers reported. As a result, our model is capable of accurately predicting new chemicals in this DoA.

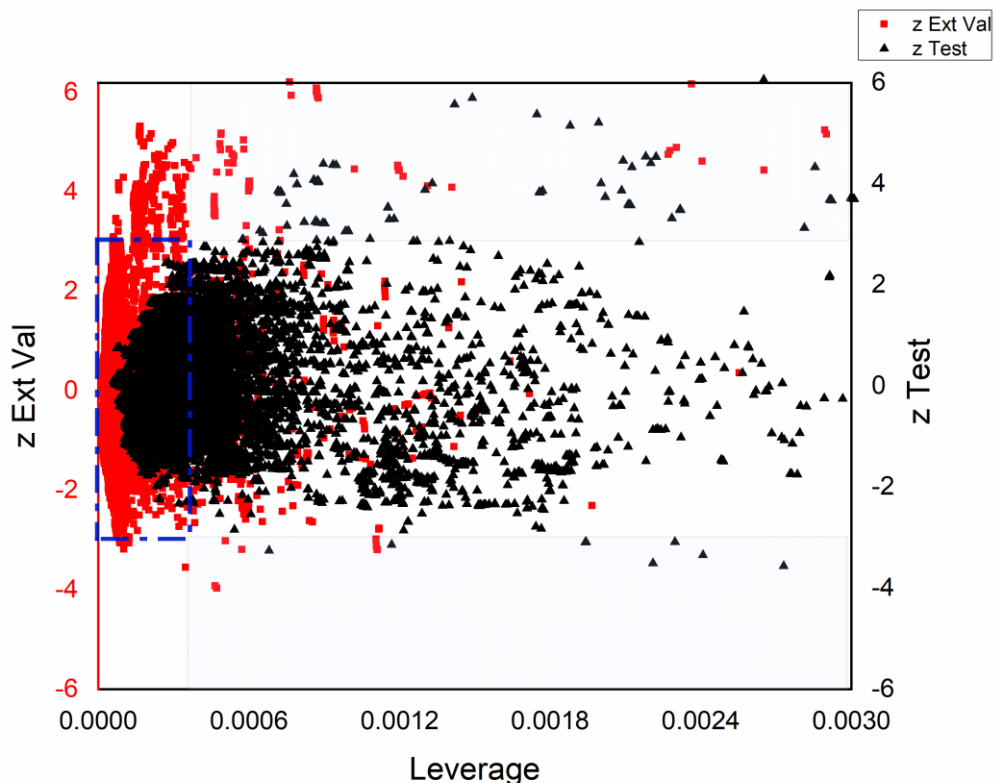


Figure 4.5. William's plot of residuals versus leverages for AD vs MN in the test and external validation sets.

3.3 Comparison with other heterogeneous series of compounds approaches

The linear and non-linear IFPTML of the AD vs. MN were compared with other reports based on a heterogeneous series of compounds previously described in the literature of discovering antibacterial compounds. **Table 4.4** shows a comparison between the present model and some of these models is shown (Heterogeneous Series of compounds, Drug family >10). An analysis of **Table 4.4** reveals that current work has the greatest dataset (very complex and notably larger data set in the number of compounds). Only six previous models have more than 10000 compounds. The model presented in this report has a large number of parameters (12) when compared with other models that have 6-8 as the number of parameters. However, **models 3, 5, and 6** show a greater number of variables 62⁶⁵ and 21⁶⁶, respectively.

The LDA predominates among the techniques used to realize the models (6 of the 13). Two model include KNN (**Model 3 and 5**)^{65, 66} and ANN (**Model 4 and 10**)^{23, 67} even though SVM is analyzed in one model (**Model 6**)⁶⁶ the Iterative stochastic elimination (ISE)⁶⁸ and Self-organizing map (SOM) (Kohonen).⁶⁹ In the case of accuracy, it should be noted that all compared models have precision values higher than 75 %. However, the accuracy values of the RF and KNN techniques in this study (97.4%) are higher than those of other studies carried out

with similar data sets, such as Nocedo *et al.*⁷⁰ (88.6%). The predominant validation technique was the external predicting series used in 12 out of 13 models, including this one. This shows that we used a proven validation technique. As shown in **Table 4.4** (**Model 1-3, 7, and 8**), the models are notable to predict multiple species, they only predict a single type of microorganism.

Table 4.4. Chemoinformatic approaches for the development of novel antibacterial compounds (Heterogeneous Series of compounds, Drug family >10).

Model ^a	n ^b	Act. ^b	Var. ^b	Tech. ^c	Acc (%)	Val ^d	Multi Species ^e	MO ^f	Net ^g	Ref. ^h
1	667	363	7	LDA	92.9	i	No	No	No	71
2	2030	1006	8	LDA	90.4	i	No	No	No	72
3	4346	520	62	kNN	95	ii	No	No	No	65
4	11576	4208	4	ANN	97	i	<i>ST</i>	Yes	No	23
5	7517	2066	21	kNN	99.3	i	<i>MRSA</i>	Yes	No	66
6	7517	2066	21	SVM	92.9	i	<i>MRSA</i>	Yes	No	66
7	37834	13203	5	LDA	95	i	No	Yes	No	73
8	2230	1051	3	LDA	86.3	i	No		No	74
9	30181	12474	6	LDA	90	i	<i>FN/PI</i>	Yes	No	67
10	54682	19912	6	ANN	90	i	<i>PS</i>	Yes	No	67
11	3500	628	4	ISE	94.6	i	MBS	Yes	No	68
12	74567	8724	6	SOM	75.5	i	<i>EC</i>	Yes	No	69
13	83605	10030	6	LDA	88.6	i	MBS	Yes	Yes	70
14	115662	42209	12	LDA	74.3	i	MBS	Yes	Yes	This work
15	115662	42209	12	kNN	97.4	i	MBS	Yes	Yes	
16	115662	42209	12	RF	97.4	i	MBS	Yes	Yes	

^a Number of the Model. ^b n=Total number of cases in training and/or validation series, Act=Active drugs, and Vars. = Variables in the model. ^c Technique: LDA = Linear discriminant analysis, ANN= artificial neural network, BLR=binary logistic regression, BN=Bayesian Network, DT=decision tree, ISE=Iterative stochastic elimination, SOM=self-organizing map (Kohonen), RF=Random Forest., KNN=K-Nearest-Neighbor. ^d Val: Validation Methods. (i) external predicting series, test set, (ii) 100-times-averaged resubstitution technique. ^e Multi Species: MBS=Multiple bacterial strain, *MRSA*=Methicillin-resistant *Staphylococcus aureus*, *FN*=*Fusobacterium necrophorum*, *PI*=*Prevotella intermedia*, *EC*=*Escherichia coli*, *PS*=*Pseudomonas spp*, *SS*=*Streptococcus spp*. ^f MO = Multi Output: multi-output models can predict more than one type of biological activity (MIC, IC50, MBC, etc.). ^g Net. =MN_s; Models able to account for changes in the MN_s of different microorganisms. ^h Reference.

Recently, multispecies models have been developed; some of them predict biological activity exclusively for members of the same genus or subgroup of bacteria (**Models 4 to 13**). The current IFPTML model can predict any compound's antibacterial activity against a various bacteria strain, including their MNs. This makes it possible to vary a certain reaction inside a bacterium and identify the changes in its metabolic pathway. As a result, binding sites that the

drug's activity can target are revealed. Furthermore, the drug search is handled. The model's adoption can reduce the number of candidates, resulting in time and resource savings.

4. CONCLUSIONS

Bacterial resistance to conventional antibiotics has been attributed to the use of broad-spectrum antibiotics. Understanding the metabolism of pathogens plays an important role in discovering new drugs and targets for antibacterial treatment. The influence of changes in metabolic networks on the capacity for survival of different microorganisms has been demonstrated. In this chapter, we developed an NIFPTML-LDA model for predicting the antibacterial activity, taking into account the structure of MN. Regarding the methodological objectives, the linear model included two subsystems (preclinical antibacterial drugs and metabolic networks of different microorganisms) and showed a good fit. The information from the two subsystems did not significantly influence the robustness of the models to analyze the problem presented in the thesis.

Regarding the practical objectives, NIFPTML-LDA models allowed us to predict the antibacterial activity and suitability of >160 000 biological assays of >50000 compounds vs. >25 different types of bacteria species with >90 strains. The model showed good predictive power (Sn, Sp, and Acc = 74%) compared to other ML linear and non-linear models (e.g., SOM models) reported in this work and from literature. Among the 17 ML algorithms used to create non-linear IFPTML classification models, the KNN, Bagging, BN, J48, PART, and RF models show the highest AUROC, Accuracy, F1 score, Sn, and Sp values (>85% in training/validation sets). We can conclude that the IFPTML model reported could be a simple, useful, and adaptable instrument, reducing time and costs in antibacterial drug research.

5. REFERENCES

1. Tacconelli, E.; Magrini, N. Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. *World Health Organization*. **2017**, 1-7.
2. Bush, K.; Courvalin, P.; Dantas, G.; Davies, J.; Eisenstein, B.; Huovinen, P.; Jacoby, G. A.; Kishony, R.; Kreiswirth, B. N.; Kutter, E.; et al. Tackling antibiotic resistance. *Nat Rev Microbiol*. **2011**, *9* (12), 894-896. DOI: 10.1038/nrmicro2693 PubMed.
3. Stokes, J. M.; Lopatkin, A. J.; Lobritz, M. A.; Collins, J. J. Bacterial Metabolism and Antibiotic Efficacy. *Cell metabolism*. **2019**, *30* (2), 251-259. DOI: 10.1016/j.cmet.2019.06.009 From NLM.
4. Kohanski, M. A.; Dwyer, D. J.; Collins, J. J. How antibiotics kill bacteria: from targets to networks. *Nat Rev Microbiol*. **2010**, *8* (6), 423-435. DOI: 10.1038/nrmicro2333 From NLM.
5. Levy, S. B.; Bonnie, M. Antibacterial resistance worldwide: Causes, challenges and responses. *Nature Medicine*. **2004**, *10* (12S), S122-S129. DOI: 10.1038/nm1145.
6. Brown, E. D.; Wright, G. D. Antibacterial drug discovery in the resistance era. *Nature*. **2016**, *529* (7586), 336-343, Review. DOI: 10.1038/nature17042 Scopus.
7. Butler, M. S.; Blaskovich, M. A.; Cooper, M. A. Antibiotics in the clinical pipeline at the end of 2015. *Journal of Antibiotics*. **2017**, *70* (1), 3-24. DOI: 10.1038/ja.2016.72.

8. Coates, A. R.; Halls, G.; Hu, Y. Novel classes of antibiotics or more of the same? *British Journal of Pharmacology*. **2011**, *163* (1), 184-194. DOI: 10.1111/j.1476-5381.2011.01250.x.
9. Lehar, S. M.; Pillow, T.; Xu, M.; Staben, L.; Kajihara, K. K.; Vandlen, R.; DePalatis, L.; Raab, H.; Hazenbos, W. L.; Hiroshi Morisaki, J.; et al. Novel antibody-antibiotic conjugate eliminates intracellular *S. aureus*. *Nature*. **2015**, *527* (7578), 323-328, Article. DOI: 10.1038/nature16057 Scopus.
10. Luo, X.; Qian, L.; Xiao, Y.; Tang, Y.; Zhao, Y.; Wang, X.; Gu, L.; Lei, Z.; Bao, J.; Wu, J.; et al. A diversity-oriented rhodamine library for wide-spectrum bactericidal agents with low inducible resistance against resistant pathogens. *Nature Communications*. **2019**, *10* (1), Article. DOI: 10.1038/s41467-018-08241-3 Scopus.
11. Zaengle-Barone, J. M.; Jackson, A. C.; Besse, D. M.; Becken, B.; Arshad, M.; Seed, P. C.; Franz, K. J. Copper Influences the Antibacterial Outcomes of a β -Lactamase-Activated Prochelator against Drug-Resistant Bacteria. *ACS Infectious Diseases*. **2018**, *4* (6), 1019-1029. DOI: 10.1021/acscinfecdis.8b00037.
12. Roche-Lima, A.; Domaratzki, M.; Fristensky, B. Metabolic network prediction through pairwise rational kernels. *BMC Bioinformatics*. **2014**, *15* (1), 318, journal article. DOI: 10.1186/1471-2105-15-318.
13. Dunphy, L. J.; Papin, J. A. Biomedical applications of genome-scale metabolic network reconstructions of human pathogens. *Current Opinion in Biotechnology*. **2018**, *51*, 70-79. DOI: 10.1016/j.copbio.2017.11.014.
14. Levin-Reisman, I.; Ronin, I.; Gefen, O.; Braniss, I.; Shores, N.; Balaban, N. Q. Antibiotic tolerance facilitates the evolution of resistance. *Science*. **2017**, *355* (6327), 826-830. DOI: 10.1126/science.aaj2191.
15. Lupoli, T. J.; Vaubourgeix, J.; Burns-Huang, K.; Gold, B. Targeting the Proteostasis Network for Mycobacterial Drug Discovery. *ACS Infectious Diseases*. **2018**, *4* (4), 478-498. DOI: 10.1021/acscinfecdis.7b00231.
16. Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z. N.; Barabasi, A. L. The large-scale organization of metabolic networks. *Nature*. **2000**, *407* (6804), 651-654.
17. Wareth, G.; Neubauer, H.; Sprague, L. D. A silent network's resounding success: how mutations of core metabolic genes confer antibiotic resistance. *Signal Transduction and Targeted Therapy*. **2021**, *6* (1), 301. DOI: 10.1038/s41392-021-00717-x.
18. Lopatkin, A. J.; Bening, S. C.; Manson, A. L.; Stokes, J. M.; Kohanski, M. A.; Badran, A. H.; Earl, A. M.; Cheney, N. J.; Yang, J. H.; Collins, J. J. Clinically relevant mutations in core metabolic genes confer antibiotic resistance. *Science (New York, N.Y.)*. **2021**, *371* (6531). DOI: 10.1126/science.aba0862 From NLM.
19. Dieguez-Santana, K.; Pham-The, H.; Villegas-Aguilar, P. J.; Le-Thi-Thu, H.; Castillo-Garit, J. A.; Casañola-Martin, G. M. Prediction of acute toxicity of phenol derivatives using multiple linear regression approach for *Tetrahymena pyriformis* contaminant identification in a median-size database. *Chemosphere*. **2016**, *165*, 434-441. DOI: <https://doi.org/10.1016/j.chemosphere.2016.09.041>.
20. Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*. **2018**, *23* (8), 1538-1546. DOI: 10.1016/j.drudis.2018.05.010.
21. Wu, X.; Zhu, X.; Wu, G.; Ding, W. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*. **2014**, *26* (1), 97-107. DOI: 10.1109/TKDE.2013.109.
22. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: a large-scale

- bioactivity database for drug discovery. *Nucleic acids research*. **2012**, *40* (Database issue), D1100-1107. DOI: 10.1093/nar/gkr777 From NLM.
23. Speck-Planche, A.; Kleandrova, V. V.; Cordeiro, M. N. D. S. Chemoinformatics for rational discovery of safe antibacterial drugs: Simultaneous predictions of biological activity against streptococci and toxicological profiles in laboratory animals. *Bioorg. Med. Chem.* **2013**, *21* (10), 2727-2732. DOI: 10.1016/j.bmc.2013.03.015.
 24. Gonzalez-Diaz, H.; Arrasate, S.; Gomez-SanJuan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. M. General theory for multiple input-output perturbations in complex molecular systems. 1. Linear QSPR electronegativity models in physical, organic, and medicinal chemistry. *Current topics in medicinal chemistry*. **2013**, *13* (14), 1713-1741. DOI: 10.2174/1568026611313140011.
 25. Quevedo-Tumaili, V. F.; Ortega-Tenezaca, B.; Gonzalez-Diaz, H. Chromosome Gene Orientation Inversion Networks (GOINs) of Plasmodium Proteome. *J Proteome Res.* **2018**, *17* (3), 1258-1268. DOI: 10.1021/acs.jproteome.7b00861.
 26. Gonzalez-Diaz, H.; Riera-Fernandez, P.; Pazos, A.; Munteanu, C. R. The Rucker-Markov invariants of complex Bio-Systems: applications in Parasitology and Neuroinformatics. *Bio Systems*. **2013**, *111* (3), 199-207. DOI: 10.1016/j.biosystems.2013.02.006.
 27. Santana, R.; Zuluaga, R.; Gañan, P.; Arrasate, S.; Onieva, E.; Gonzalez-Diaz, H. Designing Nanoparticle Release Systems for Drug-Vitamin Cancer Co-Therapy with Multiplicative Perturbation-Theory Machine Learning (PTML) Models. *Nanoscale*. **2019**, 10.1039/C9NR05070A. DOI: 10.1039/C9NR05070A.
 28. Diéguez-Santana, K.; González-Díaz, H. Towards Machine Learning Discovery of Dual Antibacterial Drug-Nanoparticle Systems. *Nanoscale*. **2021**, *13*, 17854-17870. DOI: 10.1039/D1NR04178A.
 29. Diez-Alarcia, R.; Yanez-Perez, V.; Muneta-Arrate, I.; Arrasate, S.; Lete, E.; Meana, J. J.; Gonzalez-Diaz, H. Big Data Challenges Targeting Proteins in GPCR Signaling Pathways; Combining PTML-ChEMBL Models and [(35)S]GTPgammaS Binding Assays. *ACS Chem Neurosci*. **2019**, *10* (11), 4476-4491. DOI: 10.1021/acchemneuro.9b00302.
 30. Gonzalez-Diaz, H.; Herrera-Ibata, D. M.; Duardo-Sanchez, A.; Munteanu, C. R.; Orbegozo-Medina, R. A.; Pazos, A. ANN multiscale model of anti-HIV drugs activity vs AIDS prevalence in the US at county level based on information indices of molecular graphs and social networks. *J Chem Inf Model*. **2014**, *54* (3), 744-755. DOI: 10.1021/ci400716y.
 31. Gonzalez-Diaz, H.; Riera-Fernandez, P. New Markov-autocorrelation indices for re-evaluation of links in chemical and biological complex networks used in metabolomics, parasitology, neurosciences, and epidemiology. *J Chem Inf Model*. **2012**, *52* (12), 3331-3340. DOI: 10.1021/ci300321f.
 32. Martinez-Arzate, S. G.; Tenorio-Borroto, E.; Barbabosa Pliego, A.; Diaz-Albiter, H. M.; Vazquez-Chagoyan, J. C.; Gonzalez-Diaz, H. PTML Model for Proteome Mining of B-Cell Epitopes and Theoretical-Experimental Study of Bm86 Protein Sequences from Colima, Mexico. *J Proteome Res.* **2017**, *16* (11), 4093-4103. DOI: 10.1021/acs.jproteome.7b00477.
 33. Diéguez-Santana, K.; Casañola-Martin, G. M.; Green, J. R.; Rasulev, B.; González-Díaz, H. Predicting Metabolic Reaction Networks with Perturbation-Theory Machine Learning (PTML) Models. *Current Topics in Medicinal Chemistry*. **2021**, *21* (9), 819-827. DOI: 10.2174/1568026621666210331161144.

34. Blay, V.; Yokoi, T.; González-Díaz, H. Perturbation Theory–Machine Learning Study of Zeolite Materials Desilication. *Journal of Chemical Information and Modeling*. **2018**, *58* (12), 2414-2419. DOI: 10.1021/acs.jcim.8b00383.
35. Ferreira da Costa, J.; Silva, D.; Caamaño, O.; Brea, J. M.; Loza, M. I.; Munteanu, C. R.; Pazos, A.; García-Mera, X.; González-Díaz, H. Perturbation Theory/Machine Learning Model of ChEMBL Data for Dopamine Targets: Docking, Synthesis, and Assay of New l-Prolyl-l-leucyl-glycinamide Peptidomimetics. *ACS Chemical Neuroscience*. **2018**, *9* (11), 2572-2587. DOI: 10.1021/acschemneuro.8b00083.
36. Simón-Vidal, L.; García-Calvo, O.; Oteo, U.; Arrasate, S.; Lete, E.; Sotomayor, N.; González-Díaz, H. Perturbation-Theory and Machine Learning (PTML) Model for High-Throughput Screening of Parham Reactions: Experimental and Theoretical Studies. *Journal of Chemical Information and Modeling*. **2018**, *58* (7), 1384-1396. DOI: 10.1021/acs.jcim.8b00286.
37. Duardo-Sanchez, A.; Munteanu, C. R.; Riera-Fernandez, P.; Lopez-Diaz, A.; Pazos, A.; Gonzalez-Diaz, H. Modeling Complex Metabolic Reactions, Ecological Systems, and Financial and Legal Networks with MIANN Models Based on Markov-Wiener Node Descriptors. *Journal of Chemical Information and Modeling*. **2014**, *54* (1), 16-29. DOI: 10.1021/ci400280n.
38. Valdés-Martín, J. R.; Marrero-Ponce, Y.; García-Jacas, C. R.; Martínez-Mayorga, K.; Barigye, S. J.; Vaz d'Almeida, Y. S.; Pham-The, H.; Pérez-Giménez, F.; Morell, C. A. QuBiLS-MAS, open source multi-platform software for atom- and bond-based topological (2D) and chiral (2.5D) algebraic molecular descriptors computations. *Journal of Cheminformatics*. **2017**, *9* (1), 35, journal article. DOI: 10.1186/s13321-017-0211-5.
39. Durán, F. J. R.; Alonso, N.; Caamaño, O.; García-Mera, X.; Yañez, M.; Prado-Prado, F. J.; González-Díaz, H. Prediction of Multi-Target Networks of Neuroprotective Compounds with Entropy Indices and Synthesis, Assay, and Theoretical Study of New Asymmetric 1,2-Rasagiline Carbamates. *Int. J. Mol. Sci.* **2014**, *15* (9), 17035-17064. DOI: 10.3390/ijms150917035.
40. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. D. S. Multi-target drug discovery in anti-cancer therapy: Fragment-based approach toward the design of potent and versatile anti-prostate cancer agents. *Bioorganic & Medicinal Chemistry*. **2011**, *19* (21), 6239-6244. DOI: 10.1016/j.bmc.2011.09.015.
41. Hill, T.; Lewicki, P. *STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining*; StatSoft, 2006
42. Frank, E.; Hall, M. A.; Witten, I. H. *The WEKA workbench*; Morgan Kaufmann, 2016.
43. Quinlan, R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers, 1993.
44. Breiman, L. Random Forests. *Machine Learning*. **2001**, *45* (1), 5-32, journal article. DOI: 10.1023/a:1010933404324.
45. Zhou, Z.-H. *Ensemble methods: foundations and algorithms*; Chapman and Hall/CRC Press, 2012.
46. Breiman, L. Bagging predictors. *Machine Learning*. **1996**, *24* (2), 123-140. DOI: 10.1007/BF00058655.
47. Hastie, T.; Tibshirani, R.; Friedman, J. H. *The elements of statistical learning: Data mining, inference, and prediction*; Springer open, 2008.
48. Kuncheva, L. I. *Combining pattern classifiers: methods and algorithms*; John Wiley & Sons, 2014.

49. Yosipof, A.; Guedes, R. C.; García-Sosa, A. T. Data Mining and Machine Learning Models for Predicting Drug Likeness and Their Disease or Organ Category. *Front Chem.* **2018**, *6*, 162-162. DOI: 10.3389/fchem.2018.00162 PubMed.
50. Witten, H. I.; Frank, E. *Data Mining: Practical machine learning tools and techniques*; Morgan Kaufmann, 2005.
51. Vapnik, V. *The nature of statistical learning theory*; Springer science & business media, 1999.
52. Aha, D. W.; Kibler, D.; Albert, M. K. Instance-based learning algorithms. *Machine Learning.* **1991**, *6* (1), 37-66. DOI: 10.1007/BF00153759.
53. Dieguez-Santana, K.; Pham-The, H.; Rivera-Borroto, O. M.; Puris, A.; Le-Thi-Thu, H.; Casanola-Martin, G. M. A Two QSAR Way for Antidiabetic Agents Targeting Using α -Amylase and α -Glucosidase Inhibitors: Model Parameters Settings in Artificial Intelligence Techniques. *Letters in Drug Design & Discovery.* **2017**, *14* (8), 862-868. DOI: 10.2174/1570180814666161128121142.
54. Le Cessie, S.; Van Houwelingen, J. C. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C.* **1992**, *41* (1), 191-201. DOI: 10.2307/2347628.
55. Frank, E.; Witten, I. H. Generating accurate rule sets without global optimization. In *Fifteenth International Conference on Machine Learning*, San Francisco, CA, 1998; Morgan Kaufmann Publishers Inc: pp 144-151.
56. Gaines, B. R.; Compton, P. Induction of ripple-down rules applied to modeling large databases. *Journal of Intelligent Information Systems.* **1995**, *5* (3), 211-228.
57. Hühn, J.; Hüllermeier, E. FURIA: an algorithm for unordered fuzzy rule induction. *Data Mining Knowledge Discovery.* **2009**, *19* (3), 293-319, journal article. DOI: 10.1007/s10618-009-0131-8.
58. Afantitis, A.; Melagraki, G.; Tsoumanis, A.; Valsami-Jones, E.; Lynch, I. A nanoinformatics decision support tool for the virtual screening of gold nanoparticle cellular association using protein corona fingerprints. *Nanotoxicology.* **2018**, *12* (10), 1148-1165, Article. DOI: 10.1080/17435390.2018.1504998 Scopus.
59. Papadiamantis, A. G.; Jänes, J.; Voyiatzis, E.; Sikk, L.; Burk, J.; Burk, P.; Tsoumanis, A.; Ha, M. K.; Yoon, T. H.; Valsami-Jones, E.; et al. Predicting Cytotoxicity of Metal Oxide Nanoparticles Using Isalos Analytics Platform. *Nanomaterials.* **2020**, *10* (10), 2017. DOI: 10.3390/nano10102017.
60. Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. *Alternatives to Laboratory Animals.* **2005**, *33* (2), 1-19.
61. Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR & Combinatorial Science.* **2007**, *26* (5), 694-701. DOI: 10.1002/qsar.200610151.
62. Bediaga, H.; Arrasate, S.; González-Díaz, H. PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS Combinatorial Science.* **2018**, *20* (11), 621-632. DOI: 10.1021/acscmbosci.8b00090.
63. Zhang, S.; Golbraikh, A.; Oloff, S.; Kohn, H.; Tropsha, A. A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *Journal of chemical information and modeling.* **2006**, *46* (5), 1984-1995. DOI: 10.1021/ci060132x From NLM.
64. Diéguez-Santana, K.; Rasulev, B.; González-Díaz, H. Towards rational nanomaterial design by predicting drug-nanoparticle system interaction vs. bacterial metabolic networks. *Environmental Science: Nano.* **2022**, 10.1039/D1EN00967B. DOI: 10.1039/D1EN00967B.

65. Karakoc, E.; Cherkasov, A.; Sahinalp, S. C. Distance based algorithms for small biomolecule classification and structural similarity search. *Bioinformatics*. **2006**, *22* (14), e243-e251, Conference Paper. DOI: 10.1093/bioinformatics/btl259 Scopus.
66. Wang, L.; Le, X.; Li, L.; Ju, Y.; Lin, Z.; Gu, Q.; Xu, J. Discovering new agents active against methicillin-resistant *Staphylococcus aureus* with ligand-based approaches. *Journal of Chemical Information and Modeling*. **2014**, *54* (11), 3186-3197, Article. DOI: 10.1021/ci500253q Scopus.
67. Speck-Planche, A.; Cordeiro, M. N. D. S. Enabling virtual screening of potent and safer antimicrobial agents against noma: Mtk-QSBER model for simultaneous prediction of antibacterial activities and ADMET properties. *Mini-Reviews in Medicinal Chemistry*. **2015**, *15* (3), 194-202. DOI: 10.2174/138955751503150312120519.
68. Masalha, M.; Rayan, M.; Adawi, A.; Abdallah, Z.; Rayan, A. Capturing antibacterial natural products with in silico techniques. *Molecular Medicine Reports*. **2018**, *18* (1), 763-770, Article. DOI: 10.3892/mmr.2018.9027 Scopus.
69. Ivanenkov, Y. A.; Zhavoronkov, A.; Yamidanov, R. S.; Osterman, I. A.; Sergiev, P. V.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Veselov, M. S.; Ayginin, A. A.; et al. Identification of novel antibacterials using machine-learning techniques. *Frontiers in Pharmacology*. **2019**, *10* (JULY), Article. DOI: 10.3389/fphar.2019.00913 Scopus.
70. Nocedo-Mena, D.; Cornelio, C.; Camacho-Corona, M. d. R.; Garza-González, E.; Waksman de Torres, N.; Arrasate, S.; Sotomayor, N.; Lete, E.; González-Díaz, H. Modeling Antibacterial Activity with Machine Learning and Fusion of Chemical Structure Information with Microorganism Metabolic Networks. *Journal of Chemical Information and Modeling*. **2019**, *59* (3), 1109-1120. DOI: 10.1021/acs.jcim.9b00034.
71. González-Díaz, H.; Torres-Gómez, L. A.; Guevara, Y.; Almeida, M. S.; Molina, R.; Castañedo, N.; Santana, L.; Uriarte, E. Markovian chemicals "in silico" design (MARCH-INSIDE), a promising approach for computer-aided molecular design III: 2.5D indices for the discovery of antibacterials. *Journal of molecular modeling*. **2005**, *11* (2), 116-123. DOI: 10.1007/s00894-004-0228-3.
72. Marrero-Ponce, Y.; Medina-Marrero, R.; Torrens, F.; Martinez, Y.; Romero-Zaldivar, V.; Castro, E. A. Atom, atom-type, and total nonstochastic and stochastic quadratic fingerprints: A promising approach for modeling of antibacterial activity. *Bioorg. Med. Chem*. **2005**, *13* (8), 2881-2899, Article. DOI: 10.1016/j.bmc.2005.02.015 Scopus.
73. Speck-Planche, A.; Cordeiro, M. N. D. S. Simultaneous virtual prediction of anti-*Escherichia coli* activities and admet profiles: A chemoinformatic complementary approach for high-throughput screening. *ACS combinatorial science*. **2014**, *16* (2), 78-84. DOI: 10.1021/co400115s.
74. Castillo-Garit, J. A.; Marrero-Ponce, Y.; Barigye, S. J.; Medina-Marrero, R.; Bernal, M. G.; De La Vega, J. M. G.; Torrens, F.; Arán, V. J.; Pérez-Giménez, F.; García-Domenech, R.; et al. In silico antibacterial activity modeling based on the TOMOCOMD-CARDD approach. *Journal of the Brazilian Chemical Society*. **2015**, *26* (6), 1218-1226, Article. DOI: 10.5935/0103-5053.20150087 Scopus.

**CHAPTER 5. TOWARDS MACHINE LEARNING DISCOVERY
OF DUAL ANTIBACTERIAL DRUG-NANOPARTICLE SYSTEMS**

Artificial Intelligence/Machine Learning (AI/ML) algorithms may speed up the design of DADNP systems formed by Antibacterial Drugs (AD) and Nanoparticles (NP). In this work, we used IFPTML = Information Fusion (IF) + Perturbation-Theory (PT) + Machine Learning (ML) algorithm for the first time to study of a large dataset of putative DADNP systems composed by >165000 ChEMBL AD assays and 300 NP assays vs. multiple bacteria species. We trained alternative models with Linear Discriminant Analysis (LDA), Artificial Neural Networks (ANN), Bayesian Networks (BNN), K-Nearest Neighbor (KNN) and other algorithms. IFPTML-LDA model was simpler with values of $Sp \approx 90\%$ and $Sn \approx 74\%$ in both training (>124K cases) and validation (>41K cases) series. IFPTML-ANN and KNN models are notably more complicated even when they are more balanced $Sn \approx Sp \approx 88.5\% - 99.0\%$ and $AUROC \approx 0.94 - 0.99$ in both series. We also carried out a simulation (>1900 calculations) of the expected behavior for putative DADNPs in 72 different biological assays. The putative DADNPs studied are formed by 27 different drugs with multiple classes of NP and types of coats. In addition, we tested the validity of our additive model with 80 DADNP complexes experimentally synthesized and biologically tested (reported in >45 papers). All these DADNPs show values of $MIC < 50 \mu\text{g}\cdot\text{mL}^{-1}$ (cutoff used) better than MIC of AD and NP alone (synergistic or additive effect). The assays involve DADNP complexes with 10 types of NP, 6 coating materials, NP size range 5-100 nm vs. 15 different antibiotics, and 12 bacteria species. The IFPTML-LDA model classified correctly 100% (80 out of 80) DADNP complexes as biologically active. IFPMTL additive strategy may become a useful tool to assist the design of DADNP systems for antibacterial therapy taking into consideration only information about AD and NP components by separate.

1. INTRODUCTION

The emergence of multidrug-resistant (MDR) strains, the high cost of Antibacterial Drug (AD) development, and other factors push researchers to look for alternatives to traditional antibiotic treatments.^{2,3} On the other hand, Nanoparticles (NP) are gaining importance as drug delivery systems in treating different infectious diseases.^{4,6} NP could also be modified in different ways to act as alternatives to antibiotics. For instance, NP may be coated with an extracellular vesicle membrane and loaded with AD compounds.⁷ These NP delivery systems are very interesting, but one classic AD is still the only active agent in the system. Alternatively, NP may be loaded with more than one active agent. For instance, a NP loaded with curcumin and miltefosine has synergistic anti-leishmanial antimicrobial activity.⁸ NP loaded with Ciprofloxacin (AD), papain (mucolytic), and dextran sulfate (polyelectrolyte) has shown to be promising for bronchiectasis therapy.⁹ In addition, the NP of Ag, Cu, Zn ions, and zinc and copper oxides have demonstrated antibacterial activity *per se*.¹⁰ For example, nanoscale gold particles have antibacterial properties against Gram-positive and Gram-negative bacteria.¹¹ Compared to standard antibiotics, they do not efficiently produce drug resistance because they target various molecules (DNA and protein) in bacteria, making it difficult for bacteria to establish a system that can defend against all damage.¹⁰ Antibacterial mechanisms primarily involve damaging the cytoderm and biofilms, producing reactive oxygen species, and releasing metal ions that cause bacterial cell damage¹².

Interestingly, some drug-free NP may also present an antibacterial activity *per se*.^{13, 14} This opens an area designing new NP systems with dual antibacterial functionalities. For simplicity, we are going to call them Dual Antibacterial Drug-Nanoparticles (DADNP) systems. Then, DADNPs are systems with dual AD activity due to AD and the NP core *per se*. DADNP has two or more active AD agents being one of them the NP core. For instance, Shahbandeh *et al.* designed a DADNP of AgNO₃NP loaded with Imipenem and demonstrated the antibacterial activity against MDR *P. aeruginosa* isolates of the DADANP and the free components (AD and NP) by separate.¹⁵ Unfortunately, DADNP discovery may be time and resources consuming due to the very high number of AD and NP candidates to be tested and the high number of boundary conditions c_j to be defined for the experiments.

In this context, we should consider that Artificial Intelligence (AI) and/or Machine Learning (ML) algorithms have been used in nanoscience's to reduce time and costs. For instance, Mu *et al.* present a model that investigates the link between 26 physicochemical parameters of 51 Metal oxide Nanoparticles (MeONPs) and cytotoxicity in *E. coli*.¹⁶ The parameters of enthalpy of gaseous cation production (4H) and polarization force (Z/r) were shown to be important in the poisonous effect of these MeONPs. The model also suggested that MeONPs and their released metal ions could jointly promote DNA damage and cell death in *E. coli*. Previous nanoparticles cytotoxicity models were developed to provide a scientific basis for creating safe nanomaterials. Puzyn *et al.*, developed a model to describe the cytotoxicity of 17 different types of MeONPs to *E. coli*.¹⁷ Pan *et al.*, identified the major factors responsible for MeONP cytotoxicity in various mechanisms of *E. coli* bacteria and HaCaT cells¹⁸. Fjodorova *et al.* used artificial neural network models to associate NP cytotoxicity with metal cation charge, metal electronegativity, and oxide metal atom count.¹⁹ Zhou *et al.* used MLR and SVM to predict NP cytotoxicity in *E. coli* using DFT-derived (B3LYP technique) quantum-chemical descriptors.²⁰ Kaweeterawat *et al.* modeled cytotoxicity using a different dataset of 24 MeONPs and a classification-based support SVM. The conduction band energy and hydration enthalpy were revealed to be relevant factors to predict *E. coli* toxicity.²¹ Recently, Kar *et al.* analyzed a diverse dataset of 25 MeONP using ML classification methods for understanding the mechanisms of *E. coli* nanotoxicity.²²

In fact, AI/ML models have been used to solve different problems in the interface of chemistry and infectious diseases research.²³⁻²⁶ Consequently, AI/ML are expected to be useful also for selecting NP and AD candidates for experimental testing of DADNP systems.²⁷ However, a critical drawback for using AI/ML methods to design new DADNP is the insufficient number of DADNP systems that have been experimentally tested to date. According to Gajewicz *et al.* this factor may be affecting the development of AI/ML models in all NP research areas.²⁸ A temporal solution may be the selection of the components (AD and NP) of the new DADNP based on the individual properties of each component. In general, we can call this the additive approach to the design of NP drug delivery systems. This approach has been used to predict NP-Anticancer drugs and NP-Antimalarial drugs delivery systems before. The main weakness of the additive approach is that it may underscore positive synergies. It could predict a positive DADNP with a lower (but still high) probability than expected. However, the main advantage is the critical reduction of time, resources, and use of laboratory animals.^{29, 30} One crucial opportunity for applying the additive approach is the very high number of AD already tested.

For instance, the public database ChEMBL has thousands of preclinical assays of candidates to AD hits.³¹⁻³⁵ According to the additive approach, the more active ChEMBL compounds may be good AD candidates for DANP systems. In addition, there are no public databases of antibacterial NP, but an increasing number of examples have been reported in scientific literature.³⁶⁻⁵³

Another important drawback of developing DADNP systems with AI/ML methods is the high complexity of the data to be analyzed. Vectors of experimental conditions may represent the AD preclinical assays \mathbf{c}_{dj} different from those of the NP experiments \mathbf{c}_{nj} . It includes, on one side different c_{d0} = AD activity parameters to be measured (IC₅₀, MIC, *etc.*), c_{d1} = bacteria species, c_{d2} = bacteria strains for AD assays. On the other side, we have multiple c_{n0} = NP activity parameters to be measured (IC₅₀, MIC, *etc.*), c_{n1} = bacteria species, c_{n2} = NP shape, c_{n3} = coating agents, *etc.* Consequently, DADNP discovery by the additive method needs an AI/ML technique to carry out multi-output and multi-label classification.⁵⁴⁻⁵⁷ The AI/ML method used also includes a pre-processing stage to carry out Information Fusion (IF) of the two datasets (AD and NP). Unfortunately, almost all AI/ML models reported using only the structural/molecular descriptors of the AD or NP system as an input. Consequently, they omit all other non-structural variables, experimental conditions, or AD or NP labels, respectively.⁵⁸⁻⁶² As a result, they are not multi-output models and/or do not predict outputs for multiple labels, different organisms, cell lines.^{17, 63-74} Sizochenko *et al.* published one of the few methods to predict NP toxicity for multiple species.⁷⁵ Predicting NP toxicity and not the antibacterial activity is also the main goal of many of these studies.^{17 76}

In order to solve this kind of problem, González-Díaz *et al.* developed IFPTML, a multi-output, and input-coded multi-label ML technique. IFPTML acronym is self-explanatory and stands for Perturbation-Theory (PT) + Machine Learning (ML) + Information Fusion (IF) algorithm.⁷⁷ IFPTML model output is the scoring function $f(v_{ij})_{\text{calc}}$. IFPTML has been used in molecular sciences and also infectious disease research to complex data analysis tasks. It includes mapping drug, target protein, or parasite vaccine epitopes *vs.* information about cell lines, assay organisms, host organisms, bacteria metabolic networks, parasite spreading networks, or even the social network of HIV/AIDS epidemiology in the USA at the county level.⁷⁸⁻⁸⁵ IFPTML has also been applied to NP systems considering NP structure and coating agents, NP synthesis conditions, loaded drug structure, co-therapy loaded drugs, assay conditions, *etc.*^{29, 30, 86-88} Accordingly, in this work, we developed the first IFPTML model for DADNP systems design, including AD and NP components at the same time.

2. MATERIALS AND METHODS

2.1 IFPTML DADNP data analysis phases

Firstly, we obtained the outcomes of many preclinical assays of NP and AD from two datasets already published.^{89, 90} This involved getting for each NP and AD the vectors \mathbf{D}_{nk} and \mathbf{D}_{dk} of molecular descriptors. We also obtained the respective vectors \mathbf{c}_{nj} and \mathbf{c}_{dj} of labels/assay conditions. Next, we transformed all the elements of the vectors \mathbf{D}_{dk} and \mathbf{D}_{nk} into the values $\text{Sh}(\mathbf{D}_{dk})$ and $\text{Sh}(\mathbf{D}_{nk})$ of Shannon's information measures. After that, we calculated the values $\Delta\text{Sh}(\mathbf{D}_{dk})_{cdj}$ and $\Delta\text{Sh}(\mathbf{D}_{nk})_{enj}$ of the respective PTOs. Subsequently, we performed an IF procedure with the NP and AD data to obtain the DADNP working dataset. Last, we trained/validated alternative IFPTML models using different ML techniques. In **Figure 5.1**,

we depict graphically the general workflow used to obtain the IFPTML DADNP predictive model.

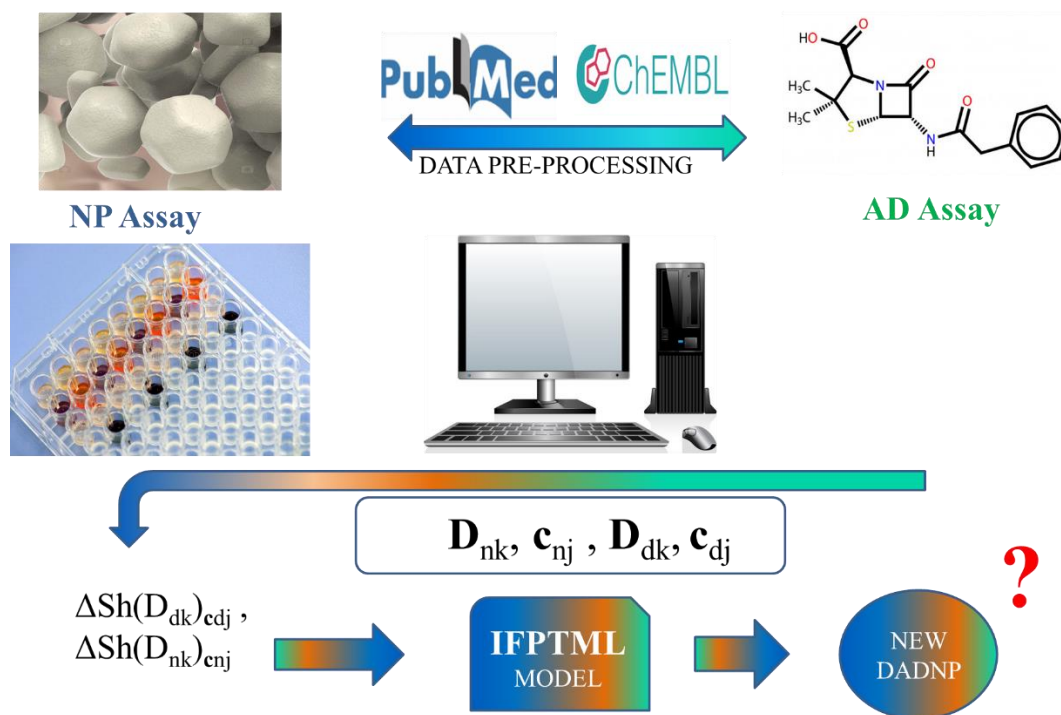


Figure 5.1. IFPTML algorithm workflow for DADNP systems.

Step 1 Data collection. Step 2. Data pre-processing. Step 3. Information Fusion (NP and AD assay). Step 4. Shannon-entropy scaling and PTO calculation. Step 5. Models' construction and evaluation. Step 6. DADNP systems prediction.

2.2 ChEMBL and NP datasets

The ChEMBL dataset use here includes >160 000 outcomes of AD preclinical assays for 55 931 compounds. Each compound has the outcome of at least 1 out of >300 biological activity parameters (MIC, IC₅₀, *etc.*). Each compound was assayed against at least 1 out of >90 bacteria strains of >25 bacterial species. The chemical structure of each AD candidate compound was encoded into a vector of molecular descriptors $D_{dk} = [D_{d1}, D_{d2}, D_{d3}]$. The elements of this vector are the molecular descriptors of the i^{th} compound: D_{d1} = Logarithm of the n-Octanol/Water Partition coefficient (LOGP_{*i*}), D_{d2} = Topological Polar Surface Area (PSA_{*i*}), D_{d3} = Number of Violations of Lipinski's Rule (NVL_{*i*}). The specific labels or conditions of each assay were encoded into the vectors $c_{dj} = [c_{d0}, c_{d1}, c_{d2}]$. The elements of these vectors are c_{d0} = name of the biological parameter (MIC, IC₅₀, *etc.*), c_{d1} = name of the bacteria species, c_{d2} = label or code of the bacteria strain. Please do not confuse the numeric value of the biological activity parameter $v_{ij}(c_{d0})$ with the name of the biological activity parameter c_{d0} . This dataset was obtained from a previous dataset reported before by our group after a new verification and pre-processing.⁹⁰ We also used a previously reported dataset with the outcomes of $N_n = 300$ preclinical assays of metal, metal salt, and metal oxide NPs against different bacteria species (s).⁸⁹ The NPs have a core made of metal, metal oxide, or metal salt. The NP assays have multiple experimental

variables conditioning the nature of the assay c_{nj} . We listed all the specific conditions of one assay as a vector $\mathbf{c}_{nj} = [c_{n1}, c_{n2}, c_{n3}, \dots, c_{nmax}]$. It includes the report of 1 out of 4 possible NP action parameters for 34 possible bacteria/strains. The data also contains NP shapes, NP physicochemical properties, NP coating agents, and time of assay.⁸⁹

2.3 IF step for observed biological parameters

The first step to obtaining the IFPTML model for DADNP systems was defining and obtaining the values of the objective function. The objective function is the function we want to fit with a ML model using the vectors of descriptors for each case \mathbf{D}_k . The objective function is often obtained after a mathematical transformation of the original theoretical or observed property of the system under study.⁹¹⁻⁹³ In the present IFPTML model, we have two sets of observed values ($v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$) and two sets of input vectors (\mathbf{D}_{dk} and \mathbf{D}_{nk}) for the AD and NP subsystems (S_d and S_n), respectively. In addition, we found many different biological parameters c_{d0} and c_{n0} . For instance, we find properties like Minimal Inhibitory Concentration (MIC ($\mu\text{g}\cdot\text{mL}^{-1}$)) or Minimal Bactericide Concentration (MBC ($\mu\text{g}\cdot\text{mL}^{-1}$)), *etc.* Do not help to solve the problem that the $v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$ values compiled are not exact numbers in many cases. Many reports in both datasets are of the type of MIC ($\mu\text{g}\cdot\text{mL}^{-1}$) < 100. In addition, we have to consider that to obtain optimal DADNP systems; we want to maximize some properties and minimize others. We conceptualize this fact with the parameter desirability.

The parameter desirability was set $d(c_{d0}) = 1$ or $d(c_{n0}) = 1$ when we want to maximize the value $v_{ij}(c_{d0})$ or $v_{nj}(c_{n0})$ respectively. Remember, the different AD and NP parameter have names or labels c_{d0} and c_{n0} , respectively. Examples of biological activity parameters (c_{d0}) with $d(c_{d0}) = 1$ are the Selectivity ratio, Inhibition (%), *etc.* Conversely, negative desirability $d(c_{d0}) = -1$ parameters are for instance MIC($\mu\text{g}\cdot\text{mL}^{-1}$), IC₅₀($\mu\text{g}\cdot\text{mL}^{-1}$), *etc.* These facts increase the uncertainty of the data and make it difficult to develop a regression model. To summarize, it is a common practice in drug discovery to use a cutoff value to split AD or even NP assays into promising and not promising. Consequently, to obtain our final objective function, we must pre-process all observed $v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$ values to eliminate or minimize inaccuracies. In addition, we need to re-scale $v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$ values to obtain a dimensionless variable not affected by scales. Last, the IF processing step for both parameters $v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$ allows obtaining an objective function of the putative DADNP system. In **Figure 5.2**, we depict a workflow summarizing all the steps of information flow (variable scaling, fusion, processing, *etc.*) of the IFPTML algorithm used here.

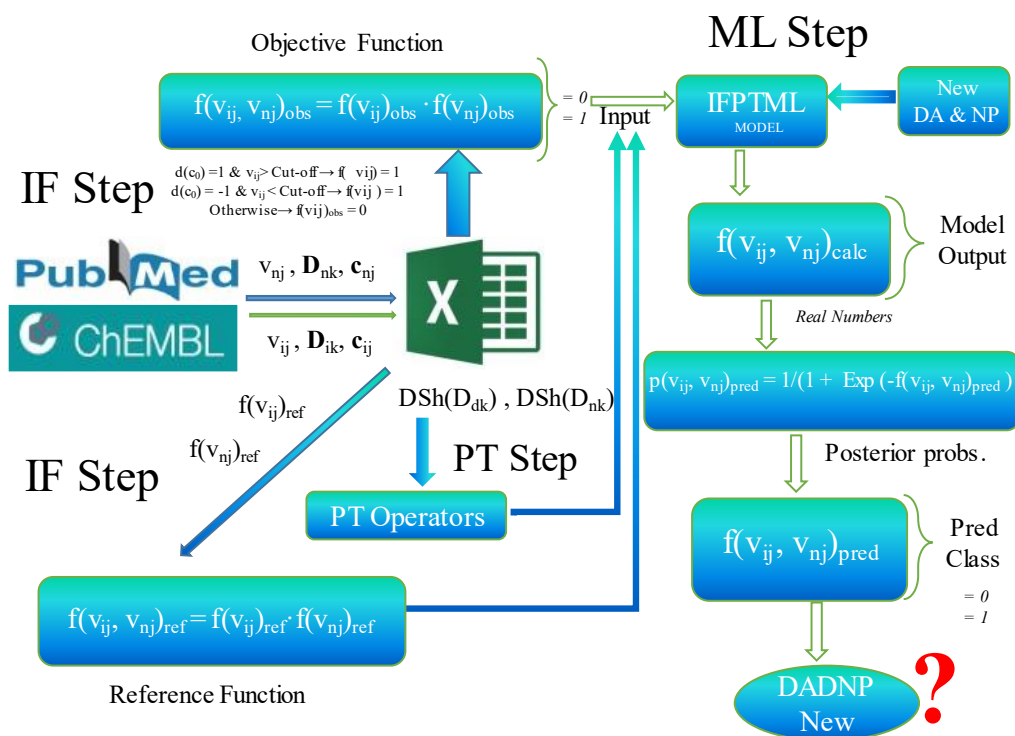


Figure 5.2. IFPTML detailed information processing workflow.

Firstly, we re-scaled the original parameters $v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$ to obtain the corresponding Boolean (dummy) functions $f(v_{ij}(c_{d0}))_{obs}$ and $f(v_{nj}(c_{n0}))_{obs}$. The scaling of $v_{ij}(c_{d0})$ was as follow: $f(v_{ij}(c_{d0}))_{obs} = 1$ when $v_{ij}(c_{d0}) > \text{cutoff}$ and $d(c_{d0}) = 1$ or $v_{ij}(c_{d0}) < \text{cutoff}$ and desirability $d(c_{d0}) = -1$, $f(v_{ij}(c_{d0})) = 0$ otherwise. By analogy, $v_{nj}(c_{n0})$ scaling was: $f(v_{nj}(c_{n0}))_{obs} = 1$ when $v_{nj}(c_{n0}) > \text{cutoff}$ and $d(c_{n0}) = 1$ or $v_{nj}(c_{n0}) < \text{cutoff}$ and $d(c_{n0}) = -1$, $f(v_{ij}(c_{d0}), v_{nj}(c_{n0})) = 0$ otherwise. The values $f(v_{ij}(c_{d0}))_{obs} = 1$ and $f(v_{nj}(c_{n0}))_{obs} = 1$ points to an strong desired effect of both the AD and the NP over the target bacteria.¹⁰ Accordingly, the objective function was defined as follow $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{obs} = f(v_{ij}(c_{d0}))_{obs} \cdot f(v_{nj}(c_{n0}))_{obs}$. Then as result of the IF-scaling $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{obs}$ depends on the i^{th} AD compound, the n^{th} NP system, the c^{th} CA used as coat, the s^{th} specie of assay, and the j^{th} sets of assay conditions. Otherwise, $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{obs} = 0$, meaning that at least one of the previous conditions fail.

2.4 IF step for function of reference

Once we defined the objective function, we must define the input variables of the IFPTML model. The first and unique of his kind input variable of this model is the function of reference $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{ref}$. In IFPTML models $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{ref}$ place an special role because this function represent the expected probability $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{ref} = p(f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{ref} = 1)$ of obtaining the desired level of activity for a property obtained from already known systems. The model starts with the value this function for an already known system or sub-set of systems used as reference. Later the IFPTML model adds the effect of deviations (perturbations) of the query system from the systems of reference (PT ideas, see next section). Consequently, $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{ref}$ is also a function based on observed (not predicted) outcomes. In this work, the reference function for putative DADNP systems was obtained by IF-scaling of the original

$v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$ values as well. In the previous section, we explained how to transform these values into the $f(v_{ij}(c_{d0}))_{obs}$ and $f(v_{nj}(c_{n0}))_{obs}$ functions. Once we get the values of these functions for all cases on the AD and NP datasets, we are in the position to count the number of positive outcomes $n(f(v_{ij}(c_{d0})) = 1)$ and $n(f(v_{nj}(c_{n0})) = 1)$. Next, we can divide these values by the total number of cases obtaining the functions of reference (expected probabilities) for the AD and NP systems alone. These values are $f(v_{ij}(c_{d0}))_{ref} = p(f(v_{ij}(c_{d0}))_{ref} = 1) = n(f(v_{ij}(c_{d0}))_{ref} = 1)/n(c_{n0})_j$ and $f(v_{nj}(c_{n0}))_{ref} = p(f(v_{nj}(c_{n0}))_{ref} = 1) = n(f(v_{nj}(c_{n0}))_{ref} = 1)/n(c_{n0})_j$. From this, the calculation of the function of reference is straightforward to realize as the product of the probabilities for each subsystem $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{ref} = p(f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{ref} = 1) = p(f(v_{ij}(c_{d0}))_{ref} = 1) \cdot p(f(v_{nj}(c_{n0}))_{ref} = 1)$. The function of reference used here is then another expression of the IF step (union) of both AD and NP datasets.

2.5 Shannon-entropy scaling of physicochemical information

This IFPTML analysis considers the vectors \mathbf{D}_{dk} and \mathbf{D}_{nk} having as components the physicochemical parameters or molecular descriptors of the AD and NP. The vector \mathbf{D}_{nk} also includes the descriptors D_{c1k} and D_{c2k} of the two possible NP coat agents. They are very diverse but almost are physicochemical properties. The AD vector lists the elements: $\mathbf{D}_{dk} = [D_{d1}, D_{d2}, D_{d3}, D_{d4}]$. These elements are the AD descriptors $D_{d1} = \text{Logarithm of the n-Octanol/Water Partition coefficients (LOGP}_i)$, $D_{d2} = \text{Topological Polar Surface Area (PSA}_i)$, $D_{d3} = \text{Number of Violations to Lipinski's Rule (NVLRI}_i)$, and $D_{d4} = \text{Molecular Weight (Mw}_i)$. The NP vector lists the elements: $\mathbf{D}_{nk} = [D_{n1}, D_{n2}, D_{n3}, D_{n4}, D_{n5}, D_{n6}, D_{n7}, D_{n8}]$. They are: $D_{n1} = \text{NP Molar Volume (AMV)}$, $D_{n2} = \text{Average Atomic Electronegativity (AAE)}$, $D_{n3} = \text{Average Atomic Polarizability (AAP)}$, and $D_{n4} = \text{Average Particle Size (APS) of the NP core in nanometers (nm)}$. In addition, the vector includes the elements: $D_{n5} = \text{LOGP}_{ca1}$, $D_{n6} = \text{PSA}_{ca1}$, $D_{n5} = \text{LOGP}_{ca2}$ and $D_{n6} = \text{PSA}_{ca2}$ of the first (c_{a1}) and second (c_{a2}) NP coating agents (ca). They have different units and scales, making it necessary to the re-scaling and/or standardization of all the information into the same scale towards the subsequent IF and ML processing. As one IF process is involved, we selected Shannon's entropy information measure as the scaling transformation. All the AD, NP, and NP coat variables have been transformed using the following equations.

$$p(D_k) = \frac{1}{(1 + \text{Exp}(-D_k/1000))} \quad (1)$$

$$\text{Sh}(D_k) = -p(D_k) \cdot \log_2(p(D_k)) \quad (2)$$

2.6 PT data preprocessing

In addition to \mathbf{D}_{dk} and \mathbf{D}_{nk} vectors, this IFPTML analysis also considers the vectors \mathbf{c}_{dj} and \mathbf{c}_{nj} as components of the non-numerical experimental conditions and/or labels for AD and NP assays. Using the $\text{Sh}(D_{dk})$ and $\text{Sh}(D_{nk})$ values explained before, we can calculate the PTOs of the AD and NP assays to account for this additional information. We used here two kinds of PTOs. The first is the AD and NP MA PTOs (**Equation 3** and **Equation 4**). They are used to account for AD and NP structural and assay information. The PTOs $\Delta\text{Sh}(D_{dk})$ and $\Delta\text{Sh}(D_{nk})$ codify AD and NP structural and/or physicochemical information on the parameters $\text{Sh}(D_{dk})$ and $\Delta\text{Sh}(D_{nk})$, respectively. The PTOs $\Delta\text{Sh}(D_{dk})$ and $\Delta\text{Sh}(D_{nk})$ codify AD and NP biological assay information with the parameter $\langle\text{Sh}(D_{dk})_{cdj}\rangle$ and $\langle\text{Sh}(D_{nk})_{cnj}\rangle$, respectively. They are the

values of the are average operator $\langle \rangle$ for $\text{Sh}(D_{dk})$ and $\Delta\text{Sh}(D_{nk})$ running overall cases with the same sub-set of experimental conditions c_{dj} and c_{nj} , respectively. Consequently, they should give specific values for each assay with at least one different element (experimental condition) of the vector c_{dj} or c_{nj} . In consequence, they can be used to indicate which assay we are using.^{29, 30, 86-88} Please, see values of $\langle \text{Sh}(D_{dk})_{c_{dj}} \rangle$ and $\langle \text{Sh}(D_{nk})_{c_{nj}} \rangle$ values in **Table S1** of Supporting Information file SI00.doc. The second type of PTOs used is the AD-NP coat MA Balance (MAB) PTO $\Delta\Delta\text{Sh}(D_{ca1}, D_{ca2}, D_{dk})$ (**Equation 5**). The MAB PTO accounts for the similarities in the information of AD vs. the NP coating agent. PTOs based directly on MA and/or linear and non-linear transformations of MA have been used for AD and NP discovery before.^{30, 88, 94} However, the MAB is reported here for the first time (see Results and Discussion). The MAS is another expression of the combined IF+PT additive processing of both AD and NP datasets.

$$\Delta\text{Sh}(D_{dk}) = \Delta\text{Sh}(D_{dk}) - \langle \text{Sh}(D_{dk})_{c_{dj}} \rangle \quad (3)$$

$$\Delta\text{Sh}(D_{nk}) = \Delta\text{Sh}(D_{nk}) - \langle \text{Sh}(D_{nk})_{c_{nj}} \rangle \quad (4)$$

$$\Delta\Delta\text{Sh}(D_{ca1}, D_{ca2}, D_{dk}) = \Delta\text{Sh}(D_{dk}) - [\Delta\text{Sh}(D_{ca1}) + \Delta\text{Sh}(D_{ca2})] \quad (5)$$

2.7 IF step and design of training and validation subsets

The dataset cases should be assigned to the training (set = t) or validation (set = v) series. The procedure of cases sampling used should be random, representative, and stratified.⁹⁵ As an additional condition, our sampling should take into consideration the IF-scaling process. Firstly, we downloaded the AD activity dataset from ChEMBL, which has random uploads from many sources worldwide and randomly selected journal papers dealing with NP antibacterial activity. Next, we organized all the cases based on the following labels c_{d0} , c_{d1} , c_{d2} , c_{n0} , c_{n1} , and c_{n2} . All cases were ordered by sorting the labels from A to Z (remember, these are non-numeric variables in nature). The order of priority of the labels on the process of ordering was $c_{d0} \Rightarrow c_{n0} \Rightarrow c_{d1} \Rightarrow c_{n1} \Rightarrow c_{d2} \Rightarrow c_{n2}$. It means that first, we ordered the cases by c_{d0} , next by c_{n0} , and so on. This priority order takes into account the IF process by alternating labels from both AD and NP datasets. After that, 3 out of each 4 cases were assigned to set = t and 1 out of 4 sets = v from top to down of the list. This increases the probability that almost all the levels of each label are represented in set = t and set = v (stratified sampling). This also increases the probability that almost all levels of each label are in a proportion 3/4 in set = t and 1/3 in set = v (representative sampling). The 75% vs. 25% proportion between set = t and set = v is not the only but is very commonly used.⁹⁵

2.8 IFPTML additive cross-over linear model

IFPTML DADNP model uses as input the PTOs described above to encode information of the putative DADNP system and the respective subsystems AD and NP. Joint objective function $f(v_{ij}, v_{nj})_{obs}$, reference function $f(v_{ij}, v_{nj})_{ref}$, post IF PTOs $\Delta\Delta\text{Sh}(D_{1c}, D_{2c}, D_{dk})$, and resulting output function $f(v_{ij}, v_{nj})_{calc}$ performs dataset cross-over codification of AD and NP information. IFPTML linear models tested for this system have the following general equation:

$$\begin{aligned}
 f(v_{ij}, v_{nj})_{calc} &= a_0 + a_1 \cdot f(v_{ij}, v_{nj})_{ref} \\
 &+ \sum_{k=1, j=1}^{k=kmax, j=jmax} a_{k,j} \cdot \Delta Sh(D_{ki})_{cd_j} \quad (6) \\
 &+ \sum_{k=1, j=1}^{k=kmax, j=jmax} a_{k,j} \cdot \Delta Sh(D_{kn})_{cn_j} + \sum_{k=1, j=1}^{k=kmax, j=jmax} a_{k,j} \cdot \Delta \Delta Sh(D_{ki}, D_{kn})_{cd_j, cn_j}
 \end{aligned}$$

2.9 IFPTML models training and validation

The LDA algorithm was used to find the preliminary model in the first instance. We used Forward Step-Wise (FSW) procedure as a variable selection strategy to select the input features automatically. The program used was STATISTICA 6.0.⁹⁵ After that, an Expert-Guided Selection (EGS) heuristic was used to retrain the LDA model with the more important features selected by FSW and other missing features. The quality of all the IFPTML models found was assed calculating Sensitivity (Sn), Specificity (Sp), Accuracy (Ac), Chi-square (χ^2), and the *p*-level.^{96, 97}

3. RESULTS AND DISCUSSION

3.1 IFPTML DADNP additive linear model

As we mentioned in the introduction, ML techniques are being applied to solve multiple practical problems in Nanotechnology.⁹⁸⁻¹⁰³ In this work, we focused on using the IFPTML algorithm to map AD *vs.* NP preclinical assays. Very recently, Speck-Planche *et al.* reported many IFPTML models of the toxicity and antimicrobial activity of NPs *vs.* multiple species in different conditions but did not consider the AD as part of the system.^{87, 89, 104} Curiously, Nocedo *et al.*, published an IFPTML model that predicts AD activity *vs.* multiple species, conditions of the assay, *etc.* but does not consider NP as part of the system.⁹⁰ Consequently, both models fail to consider both components of the DADNP system (AD and NP) altogether. Accordingly, in this work, we developed the first IFPTML model for DADNP systems design, including AD and NP components at the same time. Consequently, it has multiple AD assay conditions, assay strains, and biological properties but NP types, coating agents, *etc.* In so doing, as part of the IF-scaling process, we created the objective function $f(v_{ij}, v_{nj})_{obs} = f(v_{ij})_{obs} \cdot f(v_{nj})_{obs}$. These functions minimize the effect of uncertainty and eliminate the heterogeneity of scales. After calculating the PTOs (input variables), we used the ML techniques to fit the $f(v_{ij}, v_{nj})_{obs}$ function an obtain the IFPTML models. The best IFPTML-LDA model obtained for the design of DADNP systems was the following.

$$\begin{aligned}
f(v_{ij}, v_{nj})_{calc} = & 71.148 + 4.353 \cdot f(v_{ij}, v_{nj})_{ref} - 1351.098 \cdot \Delta\text{Sh}(\text{LOG}P_i)_{cd_j} \quad (7) \\
& + 440.516 \cdot \Delta\text{Sh}(\text{AM}V_n)_{cn_j} - 36.049 \cdot \Delta\text{Sh}(\text{AP}_n)_{cn_j} \\
& + 120.600 \cdot \Delta\text{Sh}(\text{AP}S_n)_{cn_j} + 1562.732 \cdot \Delta\text{Sh}(t)_{cn_j} \\
& + 31.828 \cdot \Delta\Delta\text{Sh}(\text{PSA}_{ca1}, \text{PSA}_{ca2}, \text{PSA}_i)_{cd_j, cn_j}
\end{aligned}$$

$$N_{\text{train}} = 124318 \quad \chi^2 = 30385.73 \quad p\text{-level} < 0.05$$

The output function $f(v_{ij}, v_{nj})_{calc}$ is the real-valued numeric function that may be used to scoring DADNP systems. The $f(v_{ij}, v_{nj})_{calc}$ function was obtained by fitting the objective function $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{obs}$ with the ML algorithm using the PTOs as input variables. The quality of all the IFPTML models found was assessed by calculating Sn, Sp, Ac, χ^2 , and the p -level.⁹⁵ The Sn, Sp, and Ac are >75%; in fact, they are in the range of 79-92% overall (including training and validation series). These parameters were in the correct ranges reported in the literature for ML classification techniques (see **Table 5.1**).^{96, 97} This model includes all the essential variables AD structure and assay conditions, NP properties, CA structure, NP assay conditions, *etc.*

Table 5.1. IFPTML DADNP model results summary.

$f(v_{ij}, v_{nj})_{obs}$	Stat.	(%)	$f(v_{ij}, v_{nj})_{pred}$	
Train			1	0
0	Sp	90.6	104394	10891
1	Sn	74.3	2331	6742
Total	Ac	89.4		
Validation	Stat.	(%)	0	1
0	Sp	90.5	34792	3637
1	Sn	73.5	802	2222
Total	Ac	89.3		

Notably, the MA PTOs were not able to account for all the relevant information. In the particular case of CA structure, the FSW was unable to include the simple MA PTOs $\Delta\text{Sh}(D_{ca1})$ and $\Delta\text{Sh}(D_{ca2})$. The failure of FSW to get $\Delta\text{Sh}(D_{ca1})$ and $\Delta\text{Sh}(D_{ca2})$ of the CAs is due to the low variance of these parameters and the low variability of the experimental data reported. However, based on Occam's principle, we should use the minimum necessary features to solve the problem (no more, no less).¹⁰⁵ Consequently, as part of the IF-scaling process, we calculated the PTO input variables of the type MABs represented by $\Delta\Delta\text{Sh}(D_{ca1}, D_{ca2}, D_i)$. MABs account for the similarities/dissimilarities (perturbations) on the information of AD *vs.* the CA used as NP coating. PTOs based directly on MA and/or linear and non-linear transformations of MA have been used for AD and NP discovery before.^{30, 88, 94} However, the MAB is reported here for the first time. The MA PTOs used here to account for information of the NP or AD datasets. They have been calculated previous to the IF process and run over one single dataset. Consequently, they cannot account for information on the AD-NP pairs that are candidates to form the DADNP. Conversely, MAB is a post-IF PTOs accounting for information on both the AD and the NP. More precisely, MABs quantify the physicochemical information of AD and the NP coating system using the $\text{Sh}(D_{dk})$ and $\text{Sh}(D_{nk})$ descriptors. MABs also quantify the

experimental conditions and/or non-numerical labels \mathbf{c}_{dj} and \mathbf{c}_{nj} of the AD and NP assays through the parameters $\langle \text{Sh}(D_{dk})_{\mathbf{c}_{dj}} \rangle$ and $\langle \text{Sh}(D_{nk})_{\mathbf{c}_{nj}} \rangle$, see **Equation 8**.

$$\begin{aligned} \Delta\Delta\text{Sh}(D_{ca1}, D_{ca2}, D_{dk}) &= \Delta\text{Sh}(D_{dk}) - [\Delta\text{Sh}(D_{ca1}) + \Delta\text{Sh}(D_{ca2})] \quad (8) \\ \Delta\Delta\text{Sh}(D_{ca1}, D_{ca2}, D_{dk}) &= \left[\text{Sh}(D_{dk}) - \langle \text{Sh}(D_{dk})_{\mathbf{c}_{dj}} \rangle \right] - \\ &\quad \left\{ \left[\text{Sh}(D_{ca1k}) - \langle \text{Sh}(D_{ca1k})_{\mathbf{c}_{nj}} \rangle \right] - \left[\text{Sh}(D_{ca2k}) - \langle \text{Sh}(D_{ca2k})_{\mathbf{c}_{nj}} \rangle \right] \right\} \end{aligned}$$

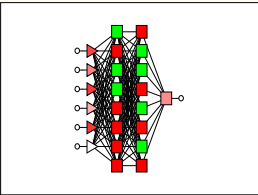
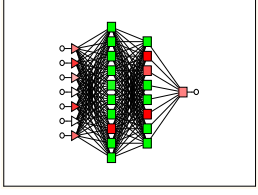
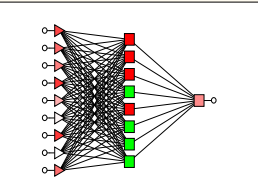
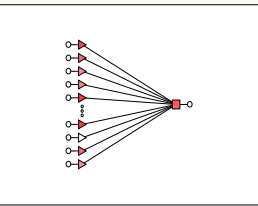
In this sense, MABs belong to an essentially new class of crossover operators involving information of more than one data set simultaneously. It is important to note that the MABs are linear PTOs. Consequently, in this work, we can only calculate homogeneous MABs with all the parameters D_{dk} and D_{nk} of the exact nature (same property and units). Specifically, in this work, we tested only the MAB PTOs $\Delta\Delta\text{Sh}(\text{LOGP}_{1c}, \text{LOGP}_{2c}, \text{LOGP}_i)$ and $\Delta\Delta\text{Sh}(\text{PSA}_{ca1}, \text{PSA}_{ca2}, \text{PSA}_i)$ based on the LOGP ($\text{LOGP}_{ca1}, \text{LOGP}_{ca2}, \text{LOGP}_i$) and PSA ($\text{PSA}_{ca1}, \text{PSA}_{ca2}, \text{PSA}_i$) values of the i^{th} AD and the first and second CAs respectively. In any case, Shannon's entropy eliminates the original units transforming all the parameters to the same scale. It opens the door to the use of heterogeneous MABs based on descriptors with different scales and units.

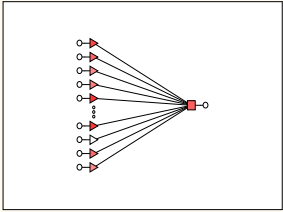
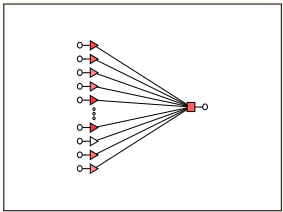
3.2 IFPTML-ANN linear vs. non-linear models

We also used ANN to test the linear hypothesis validity and propose alternative non-linear models. ANN linear models or Linear Neural Networks (LNN) are similar to LDA models because both are linear equations. Consequently, we used the IFPTML-LNN model to test the strength of the linear additive relationship among PTOs and the DADNP objective function. IFPTML-LNN models reported here presented high Sn and Sp $\approx 79 - 80\%$ values in the training and validation series, see **Table 5.2**. Similar to the IFPTML-LDA model, the Sp and Sn values are slightly unbalanced concerning each other but near to constant if we compare training vs. validation series. The Area Under Receiver Operating Characteristic (AUROC)⁹⁵ curve values are 0.86-0.87 for these models' training and validation series. Specifically, AUROC values of IFPTML-LNN models are significantly different from random (RND) behavior with AUROC = 0.5,⁹⁵ see **Figure 5.3**.

All this confirms the strength of the linear hypothesis used here. However, the values of Sn and Sp obtained still have a margin from improvement. Consequently, we increased the number of variables in the IFPTML-LNN models from 9 to 10 and 11. In this study, no significant change was detected. As a result, we also considered the non-linear hypothesis here to increase Sp and Sn values. The IFPTML-MLP 9:9-8-1:1 model with nine neurons in the input layer (input variables) and eight neurons in the hidden layer showed more balanced SN and Sp $\approx 88\%$ values. See the summary of results in **Table 5.2**. See detailed results for all cases in Supporting Information file SI01.xlsx. More complicated IFPTML-MLP2 models with two hidden layers do not show significant improvement.

Table 5.2. IFPTML-ANN DADNP systems models.

IFPTML-ANN Models ^a	Sub set	Stat.	Val. (%)	$f(v_{ij}(C_{d0}), v_{nj}(C_{n0}))$	Observed		AUROC
				Predicted	1	0	
MLP 6:6-8-8-1:1  BP100,CG20,CG1b	t	Sn	88.7	1	8049	12216	0.94
		Sp	89.4	0	1024	103070	
	v	Sn	88.0	1	2662	4015	0.95
		Sp	89.6	0	362	34413	
MLP 7:7-10-8-1:1  BP100,CG20,CG1b	t	Sn	86.8	1	7872	15450	0.93
		Sp	86.6	0	1201	99836	
	v	Sn	86.0	1	2602	5090	0.92
		Sp	86.8	0	422	33338	
MLP 9:9-8-1:1  BP100,CG20,CG1b	t	Sn	88.5	1	8030	13413	0.94
		Sp	88.4	0	1043	101873	
	v	Sn	88.5	1	2676	4460	0.94
		Sp	88.4	0	348	33968	
LNN 9:9-1:1  PI	t	Sn	80.3	1	7282	23035	0.87
		Sp	80.0	0	1791	92251	
	v	Sn	79.7	1	2411	7723	0.86
		Sp	79.9	0	613	30705	
LNN 10:10-1:1	t	Sn	80.3	1	7285	23052	0.87

 <p>PI</p>	v	Sp	80.0	0	1788	92234	0.86
		Sn	79.8	1	2412	7726	
<p>LNN 11:11-1:1</p>  <p>PI</p>	t	Sn	80.4	1	7294	22849	0.87
		Sp	80.2	0	1779	92437	
	v	Sn	79.7	1	2410	7664	0.86
		Sp	80.1	0	614	30764	

Considering all the previous factors, we were pivoting between IFPTML-LDA or IFPTML-LNN model and the IFPTML-MLP model. A critical point in favor of the IFPTML-MLP model is his notably higher value of AUROC = 0.94 and the notably better behavior (shape) of the ROC curve concerning the IFPTML-LNN linear models and RND classifier behavior, see **Figure 5.3**. Once again, Occam’s razor comes to the rescue herein by checking if minimal necessary features (no more, no less) are being considered.¹⁰⁵ We carried out a feature sensitivity analysis on the input variables.

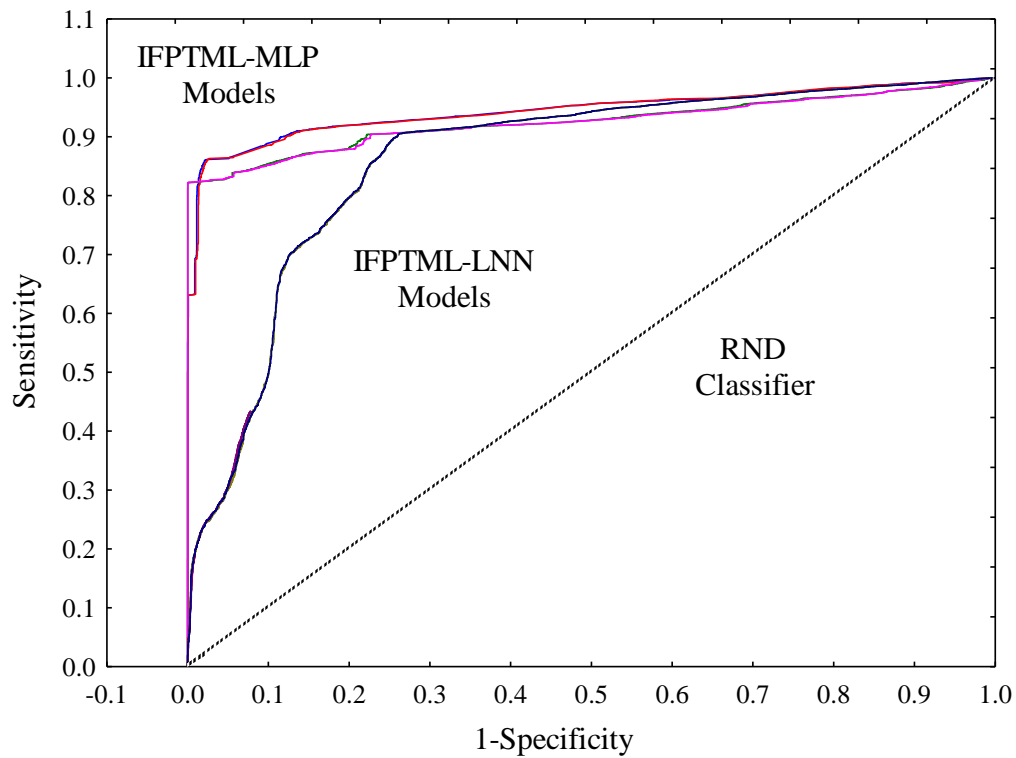


Figure 5.3. AUROC analysis of IFPTML-MLP and IFPTML-LNN models.

In **Figure 5.4**, we can see that the IFPMTL-LNN models include important parameters from the EGS point of view. Almost all parameters have a significant contribution with a Sensitivity ≥ 1 .⁹⁵ However, in most cases, it is only marginally higher with a Sensitivity $\approx 1.00 - 1.08$. On the other hand, the IFPTML-MLP model also includes the important parameters according to the EGS criteria, but they have notably higher sensitivity values $\approx 1.00 - 2.52$. MLP2 has even higher values of feature Sensitivity $\approx 1.00 - 3.31$. However, as we mentioned before, there is no gain on Sp and Sn values to justify the notably higher complexity of the model, see **Figure 5.4**.

Subsystem	Variables	LNN						MLP		MLP2			
		t	v	t	v	t	v	t	v	t	v	t	v
AD & NP	$f(c_{a0}, c_{n0})_{ref}$	1.01	1.01	1.01	1.01	1.01	1.01	1.04	1.04	1.17	1.16	0	0
AD	$\Delta Sh(ALOGP_i)_{cj}$	0	0	0	0	1.00	1.00	1.00	1.00	0	0	0	0
	$\Delta Sh(PSA_i)_{cj}$	0	0	1.00	1.00	1.00	1.00	0	0	0	0	0	0
	$\Delta Sh(NVLR_i)_{cj}$	1.00	1.00	1.00	1.00	1.00	1.00	1.03	1.03	1.02	1.02	1.01	0.99
NP	$\Delta Sh(AMV_n)_{cn}$	1.01	1.01	1.01	1.01	1.01	1.01	1.43	1.42	1.54	1.53	1.96	1.94
	$\Delta Sh(AAE_n)_{cn}$	1.00	1.00	1.00	1.00	1.00	1.00	2.52	2.49	2.35	2.30	3.31	3.25
	$\Delta Sh(AAP_n)_{cn}$	1.01	1.01	1.01	1.01	1.01	1.00	2.10	2.16	0	0	3.01	3.04
	$\Delta Sh(APS_n)_{cn}$	1.08	1.08	1.08	1.08	1.08	1.08	1.94	1.98	2.70	2.62	2.93	2.95
	$\Delta Sh(t_n)_{cn}$	1.01	1.01	1.01	1.01	1.01	1.01	1.91	1.94	1.21	1.18	2.51	2.47
AD & NP	$\Delta \Delta Sh(LOGP_i, LOGP_{ca1}, LOGP_{ca2})$	1.00	1.00	1.00	1.00	1.00	1.00	0	0	0	0	1.20	1.22
	$\Delta \Delta Sh(PSA_i, PSA_{ca1}, PSA_{ca2})$	1.00	1.00	1.00	1.00	1.00	1.00	1.20	1.23	0	0	0	0

Figure 5.4. IFPTML-ANN model input variable sensitivity analysis for AD&NP, AD, and NP subsystems.

3.3 IFPTML-WEKA Non-Linear models

Next, we decided to run several non-linear ML algorithms implemented on other software. They can be seen as an alternative to the algorithms implemented in STATISTICA. In total, we used six ML algorithms to build these alternative non-linear IFPTML classification models from the present dataset. These models were developed using the Waikato Environment for Knowledge Analysis (WEKA) software package, version 3.8.0.¹⁰⁶ These included decision tree classifiers, rule-based classifiers, neural networks, Bayesian networks, and functions. Each technique adopts a learning algorithm to identify the model that best fits the relationship between the input data set and the class. The classification algorithms applied were: Bayesian Network of K2 and B (BN), and Naïve Bayes (NBN) classifiers based on Bayes' theorem, J48 decision tree, developed by Ross Quinlan¹⁰⁷ (J48), and Random Forest (RF)¹⁰⁸, k Nearest Neighbors (KNN)¹⁰⁹, Support Vector Machine (SVM), Binary Logistic Regression (BLR) and Rule-based classifiers as PART¹¹⁰, JRip¹¹¹ and Furia-C.¹¹²

Table 5.3 shows the typical statistical values of the IFPTML models based on these algorithms. The analysis of the values of all the IFPTML models (Training/Validation Series) present good performances (Accuracy global 88.8-98.3%), and five of them showed a better performance than the PTML DADNP model (89.4%) (except BN with 88.8%). Similarly, AUROC values are high (92-99%) in most cases. SVM and Jrip show low values (0.5 and 0.7, respectively). In the analysis and comparison of the ten algorithms used, RF and KNN stand

out as having the highest precision, sensitivity, specificity, and AUROC, good binary classification models for those data under study. On the other hand, RF, KNN, BN, and NBN models have high Sn and Sp values, and their differences are slight. They can be considered models with an excellent predictive capacity for positive and negative data. However, SVM, J48, PART, JRip, and BLR have very low Sn results, Sn (0-41.6%) and differ from Sp (~98%), which denotes that they are not good techniques to classify this data set.

Table 5.3. IFPTML- WEKA Non-Linear models.

Models ^a	Sub set	Stat.	Val. (%)	Class	Observed		AUROC
				Pred.	1	0	
RF	t	Sn	77.9	1	7070	1424	0.99
		Sp	98.8	0	2003	113862	
	v	Sn	85.7	1	2591	293	0.99
		Sp	93	0	433	38135	
KNN	t	Sn	84.8	1	7694	2047	0.99
		Sp	98.2	0	1379	113239	
	v	Sn	92.2	1	2788	490	0.99
		Sp	98.7	0	236	37938	
BLR	t	Sn	19	1	1721	2525	0.94
		Sp	98	0	7352	112761	
	v	Sn	19.6	1	592	827	0.94
		Sp	98	0	2432	37601	
BN	t	Sn	89.2	1	8091	12999	0.95
		Sp	89	0	982	102287	
	v	Sn	88.6	1	2678	4326	0.95
		Sp	89	0	346	34102	
NBN	t	Sn	81.7	1	7413	11145	0.92
		Sp	90	0	1660	104141	
	v	Sn	80.9	1	2446	3701	0.92
		Sp	90	0	578	34727	
J48-DT	t	Sn	39.1	1	3547	1664	0.96
		Sp	98.6	0	5526	113622	
	v	Sn	38	1	1148	598	0.95
		Sp	98.4	0	1876	37830	
SVM	t	Sn	0.0	1	0	0	0.5
		Sp	100	0	9073	115286	

	v	Sn	0	1	0	0	0.5
		Sp	100	0	3024	38428	
PART	t	Sn	35.7	1	3235	1511	0.97
		Sp	98.7	0	5838	113775	
	v	Sn	35.1	1	1061	509	0.96
		Sp	98.7	0	1963	37919	
JRip	t	Sn	41.6	1	3777	1973	0.7
		Sp	98.3	0	5296	113313	
	v	Sn	41.4	1	1251	661	0.7
		Sp	98.3	0	1773	37767	
Furia-C	t	Sn	63	1	5715	3358	0.94
		Sp	94.3	0	6575	108711	
	v	Sn	62.5	1	1889	1135	0.93
		Sp	94.3	0	2175	36253	

3.4 Comparison to previous models

Other works published previously study different problems involving NPs and/or AD components. In fact, most of these works study the antimicrobial activity of NPs or AD activity vs. multiple species using IFPTML models. However, none of these models studied both the biological activity of both NP and AD at the same time or the possibility of forming DADNP systems. For instance, Concu *et al.* developed a PTML-ANN model for jointly predicting multiple toxicological profiles of NPs under diverse experimental conditions.¹⁰⁴ The model is derived from 54,371 NP-NP pair cases generated by applying the perturbation theory to a set of 260 unique NPs. However, this model of Concu *et al.* does not consider the information about AD biological activity. In the case of Luan *et al.*, the model contains 1681 cases (NP-NP pairs) derived from a raw dataset of 41 nanoparticles.⁸⁷ In other hand, Speck-Planche *et al.* created the model from 69,231 nanoparticle-nanoparticle (NP-NP) pairs built from the dataset of 300 NPs with varying chemical compositions, sizes, shapes, and surface coatings.⁸⁹ On the other hand, Nocedo-Mena *et al.* developed a IFPTML model for antibacterial compound prediction that considers the structure of the compound, assay conditions (different activity parameters or bacterial strains), and variations in the MRN of the bacteria.⁹⁰

In terms of the use of AI/ML algorithms in our work, we used several techniques such as LDA and ANN that have been used in previous works, as well as alternative non-linear IFPTML models (BN, RF,... *etc*) that increase the variability of processing algorithms that can be used to predict the output values of new data in AD-NP. The IFPTML-LDA and IFPTML-ANN models' values are adequate Sp \approx 88.5-90% and Sn \approx 74-88.5% training (>124K cases) and validation (>41K cases) series. Furthermore, the IFPTML-KNN model shows Sn and Sp \approx 85-99%, and AUROC = 0.99 in training/validation series outperformed. These findings are consistent with previous research. The IFPTML-LDA linear model in⁹⁰ presented values of Sp = 90.31/90.40 and Sn = 88.14/88.07 in training/validation series. On the other hand,⁸⁷ and¹⁰⁴

obtained 93% and $\approx 98\%$ sensitivity values for the LDA and ANN perturbation models, respectively. In any case, the statistical comparison of our present model and the previous one is not very informative because they deal with different problems. In conclusion, the present model seems to be the first reported in the literature to predict DADNP systems by selecting both AD and NP components using AI/ML models.

3.5 IFPTML DADNP simulation experiment

We used the IFPTML model to simulate the values of probability of several DADNP. The linear IFPTML-LDA model was selected for its simplicity, and the value of probability $p(\text{DADNP}_{in})_{cdj, cnj}$ was calculated with which the DADNP_{in} system formed by the i^{th} AD_i and the n^{th} NP_n is expected to have the desired level of biological activity on the assay conditions c_{dj} and c_{nj} . The study included $N_{AD} = 27$ compounds with AD activity, approved by the FDA and/or demonstrated active in various assays. We also included $N_{NP} = 72$ assays of NP vs. different bacteria species in the study, including multiple MDR strains. We carried out a total $N_{tot} = N_{AD} \cdot N_{NP} = 27 \cdot 72 = 1944$ calculations of the probability of success of the putative DADNP in the assays selected. The model identified some DADNP systems as promising for further assays. Only the 1% of the DADNP calculated were predicted with very high $p(\text{DADNP}_{in})_{cdj, cnj} > 0.9$. Please see details on Supporting Information file S00.doc. In **Figure 5.5**, we can see a selection of DADNP assays predicted.

	B	C	D	E	F	G	H	I	J	K	L	M
5					CEF	IMIDAZOL	MACRO	PEN	QUIN	TETRA	TETRA	TRIAZ
6	cn1 = Organism	cn2 = Strain	NP	Coat	Ciprofloxacin	Metronidazole	Erythromycin	Penicillin V	Nalidic acid	Tetracycline	Minocycline	Voriconazole
7	Escherichia coli	ATCC25922	Ag	PVP	0.010	0.004	0.006	0.007	0.005	0.008	0.010	0.006
8	Escherichia coli	ATCC25922	Ag	Lactose	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	Klebsiella pneumoniae	MDR	Au	PDT/CQ	0.809	0.338	0.594	0.616	0.472	0.676	0.796	0.515
10	Klebsiella pneumoniae	MDR	Au	PDT/CPB	0.853	0.356	0.636	0.651	0.499	0.712	0.837	0.543
11	Klebsiella pneumoniae	MDR	Au	PDT	0.651	0.272	0.489	0.497	0.381	0.543	0.638	0.414
12	Pseudomonas aeruginosa	ATCC27853	Au	PDT/CQ	0.853	0.356	0.636	0.651	0.499	0.712	0.837	0.543
13	Pseudomonas aeruginosa	ATCC27853	Au	PDT	0.809	0.338	0.594	0.615	0.472	0.676	0.796	0.515
14	Pseudomonas aeruginosa	ATCC27853	Au	PDT/Ach	0.835	0.348	0.619	0.636	0.488	0.697	0.820	0.531
15	Pseudomonas aeruginosa	MDR	Au	PDT/CQ	0.809	0.338	0.594	0.616	0.472	0.676	0.796	0.515
16	Pseudomonas aeruginosa	MDR	Au	PDT/G	0.316	0.132	0.199	0.234	0.177	0.267	0.321	0.201
17	Pseudomonas aeruginosa	MDR	Au	PDT/Mel	0.819	0.342	0.604	0.623	0.478	0.684	0.805	0.521
18	Pseudomonas aeruginosa	MDR	Au	PDT/DMB	0.260	0.108	0.161	0.192	0.145	0.220	0.265	0.165
19	Pseudomonas aeruginosa	MDR	Au	PDT	0.786	0.328	0.573	0.597	0.457	0.657	0.775	0.500
20	Pseudomonas aeruginosa	MDR	Au	PDT/Ach	0.838	0.350	0.622	0.638	0.490	0.700	0.823	0.533
21	Staphylococcus aureus	ATCC6538P	Au	PDT/Mel	0.647	0.270	0.483	0.494	0.379	0.540	0.635	0.412
22	Staphylococcus aureus	ATCC6538P	Au	PDT/CQ	0.437	0.183	0.329	0.334	0.256	0.365	0.428	0.278
23	Escherichia coli	ATCC8739	Fe3O4	PGA	0.586	0.244	0.400	0.441	0.336	0.493	0.586	0.373
24	Staphylococcus aureus	ATCC10832	Fe3O4	PGA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
25	Staphylococcus aureus	ATCC6538	SiO2	DMA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Figure 5.5. IFPTML-LDA DADNP systems simulation (selected results).

The predicted DADNP assays contain ten classes of AD, including cephalosporins (CEF), quinolones (QUIN), tetracyclines (TETRA), macrolides (MACRO), Penicillin (PEN), Triazoles (TRIAZ), Imidazole's (IMIDAZOL), etc. Among the results, the DADNP systems formed by Ciprofloxacin and Au NP coated with PDT/CQ, PDT/Mel, or PDT/Ach seem to be promising for further assays vs. MDR *P. aeruginosa* strains ($p=0.809-0.838$). However, the DADNP systems formed by Ciprofloxacin and Au NP coated with PDT/DMB could not halt the infection of the same strain. Slightly lower results on MDR *P. aeruginosa* strains can be obtained with DADNP systems formed by the same NP (Au) y coated (PDT/CQ, PDT/Mel, or PDT/Ach) but with Minocycline as AD ($p=0.796-0.823$). Other DADNP systems formed by

Ciprofloxacin and Au NP coated with PDT/CQ y PDT/CPB also show good results against MDR *K. pneumoniae* ($p=0.809-0.853$), although lower with PDT ($p=0.651$). However, this is only a punctual example, and all predictions made with this method should be taken with caution and corroborated experimentally. The great advantage of this IFPTML method is not the possibility of making a good prediction with a few tests. The actual use of the IFPTML model is to make fast and inexpensive preliminary *in silico* screening of large numbers of DADNP systems. After that, we can shortlist the more promising DADNP systems for experimental assay, taking into account $p(\text{DADNP}_{\text{in}})_{\text{cdj, cnj}}$ values and expert opinion, similar cases from the literature if any, etc. This could be a valuable tool to direct the experimental search instead of costly and slow trial and error tests.

3.6 DADNP experimental cases simulation

In addition, we used the IFPTML model to predict the values of probability of several DADNP experimentally synthesized, biologically tested, and reported in the literature previously. The inclusion criteria for the study were the following. We included cases reporting 1) AD antibacterial activity, 2) NP antibacterial activity, and 3) the DADNP complex antibacterial activity. We included cases with both additive and synergistic activity of the DADNP. The cases selected have at least one report of one biological activity parameter. The revision included 45 papers in total finding positive DADNP cases in a total of 15 papers.¹¹³⁻¹²⁷ From these papers, we extracted 80 reports of tests of DADNP complexes with at least one positive experimental outcome. The NP used to assemble the DADNP complexes have a size range from 5 nm to 100 nm. Some DADNP constructed after addition of the coat and the AD may present a size >100 nm. The AD used to assemble the DADNPs have also a wide range of hydrophobicity ranging from hydrophilic drugs ($\text{LOGP} < 0$) to lipophilic drugs ($\text{LOGP} > 0$). In all cases, the parameter determined experimentally for the AD, NP, and DADNP complex was the MIC ($\mu\text{g}\cdot\text{mL}^{-1}$). In all cases, $\text{MIC}(\mu\text{g}\cdot\text{mL}^{-1}) < 50$ (cutoff used in the model) for the DADNP complexes. The assay times reported were within the range 12 to 24 hours. Note that the design of the DADNP involve the use of coating agents that may help to increase the stability and/or bioavailability of the complexes over time.¹¹³⁻¹²⁷

In the **Figure 5.6**, we depict the surface scatterplot of experimental $\text{MIC}(\mu\text{g}\cdot\text{mL}^{-1})$ values vs. distribution histograms of NP size and AD hydrophobicity. The types of NP present were metallic NP (Ag, Au, and Zn), double metal NP (ZnCu), metal oxide NP (Fe_3O_4 , CuO, ZnO, etc.), and metal salt (MoS_2). The coating material used were mainly polymers as Triethylene Glycol (TEG), Polyethylene glycol (PEG), Thioglycolic acid (TGA), Polydopamine (PDA), Alginate and Chitosan. The cases cover a wide range of microorganisms including different strains of *S. aureus*, *P. Aeruginosa*, *E. Faecium*, *E. Coli*, *E. faecalis*, *S. epidermidis*, *B. subtilis*, *A. Baumannii*, *S. enterica*, *S. mutans*, *E. faecium*, *M. luteus*, and *K. pneumoniae*. Several strains of these microorganisms are Multi-Drug Resistant (MDR) strains; e.g., *P. aeruginosa* strains. Anyhow, the DADNP complexes found included a diverse group of AD such as QUIN (Ofloxacin, Ciprofloxacin), TETRA (Tetracycline), MACRO (Gentamicin, Vancomycin, Rifampicin), PEN (Ampicillin, Meropenem), Lipopeptide (Daptomycin), Aminoglycoside (Tobramycin, Kanamycin, Streptomycin), Polypeptide (Polymyxin B), Glycylcycline (Tigecycline), and Amphenicols (Chloramphenicol).¹¹³⁻¹²⁷

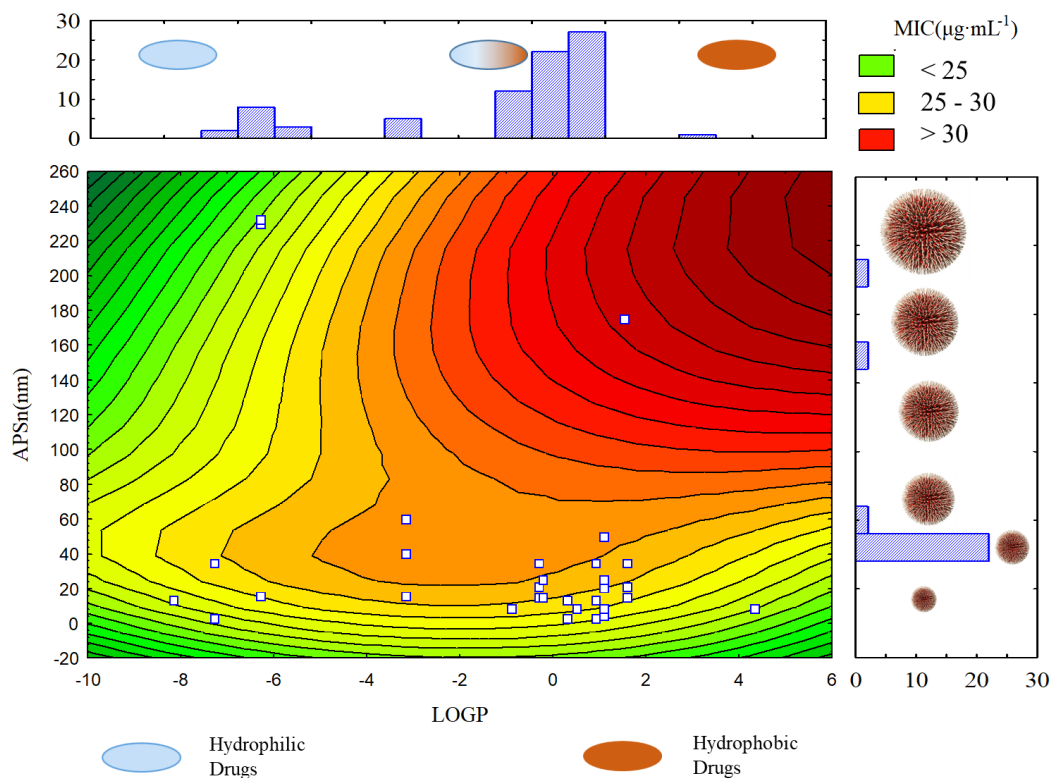


Figure 5.6. MIC($\mu\text{g}\cdot\text{mL}^{-1}$) Surface scatterplot vs. Histograms of NP size and AD Hydrophobicity distribution.

We can conclude that this experimental set of preclinical assays of DADNP complexes has a high structural and biological diversity. Intentionally, our original set of AD assays and NP assays use to assemble putative DADNP complexes a train our model has also a very high structural and biological diversity. It could help our additive model to learning how to discriminate active form non-active DADNP complexes from an additive approach. In fact, our IFPTML model was able to predict as positive all the cases (80 out of 80) with a high probability $p(\text{DADNP}_{\text{in}})_{\text{cdj, enj}} > 0.99$ in all cases (see **Table 5.4**). The result is very interesting because it supports the idea that our IFPTML additive model is able to identify properly DADNP complexes experimentally studied with high structural and biological diversity including both additive and synergic cases. Please see details on Supporting Information file S02.xlsx.

Table 5.4. IFPTML study of experimentally tested DANP complexes.

Drug	NP	APSn (nm)	$c_{dl} =$ Specie	MI C ($\mu\text{g} \cdot \text{mL}^{-1}$)	Class. Obs. ^a	Class. Pre d. ^b	p^c	$t(h)_d$	Ref e
Metal / Shape n.d.									
Vancomycin	Au	5	<i>E. faecium</i>	4	1	1	0.99	12	119
Vancomycin	Au	5	<i>E. coli</i>	8	1	1	0.99	12	119
Daptomycin	Au	50	<i>S. aureus</i>	2	1	1	0.99	24	123
Gentamicin	Au	16	<i>E. coli</i>	7.4	1	1	0.99	24	114
Ampicillin	Au	13.54	<i>E. coli</i>	45	1	1	0.99	24	125
Streptomycin	Au	13.54	<i>E. coli</i>	7	1	1	0.99	24	125
Kanamycin	Au	13.54	<i>E. coli</i>	12	1	1	0.99	24	125
Ampicillin	Au	13.54	<i>M. luteus</i>	0.45	1	1	0.99	24	125
Streptomycin	Au	13.54	<i>M. luteus</i>	17	1	1	0.99	24	125
Kanamycin	Au	13.54	<i>M. luteus</i>	23	1	1	0.99	24	125
Ampicillin	Au	13.54	<i>S. aureus</i>	0.37	1	1	0.99	24	125
Kanamycin	Au	13.54	<i>S. aureus</i>	5.8	1	1	0.99	24	125
Metal / Quasi-spherical									
Tetracycline	Au	25	<i>E. coli</i>	6	1	1	0.99	20	115
Tetracycline	Au	25	<i>S. aureus</i>	16	1	1	0.99	20	115
Tetracycline	Ag	15	<i>E. coli</i>	16	1	1	0.99	20	115
Tetracycline	Ag	15	<i>E. coli</i>	25	1	1	0.99	20	115
Tetracycline	Ag	15	<i>S. aureus</i>	16	1	1	0.99	20	115
Tetracycline	Ag	15	<i>E. coli</i>	32	1	1	0.99	20	115
Tetracycline	Ag	15	<i>S. aureus</i>	32	1	1	0.99	20	115
Metal/ Spherical									
Meropenem	CuZn	21	<i>P. aeruginosa</i>	25.2	1	1	0.99	24	116
Ciprofloxacin	CuZn	21	<i>P. aeruginosa</i>	3.2	1	1	0.99	24	116
Vancomycin	Ag	20.5	<i>S. aureus</i>	0.1	1	1	0.99	24	118
Vancomycin	Ag	20.5	<i>E. faecalis</i>	0.1	1	1	0.99	24	118
Vancomycin	Ag	20.5	<i>S. epidermitis</i>	0.02	1	1	0.99	24	118
Polymyxin B	Ag	8.4	<i>A. baumannii</i>	0.00 4	1	1	0.99	18	127
Rifampicin	Ag	8.4	<i>A. baumannii</i>	0.5	1	1	0.99	18	127
Tigecycline	Ag	8.4	<i>A. baumannii</i>	1	1	1	0.99	18	127
Chloramphenicol	Ag	35	<i>E. coli</i>	10.1	1	1	0.99	24	126
Chloramphenicol	Ag	35	<i>S. enterica</i>	10.5	1	1	0.99	24	126
Chloramphenicol	Ag	35	<i>S. aureus</i>	17.5	1	1	0.99	24	126
Kanamycin	Ag	35	<i>E. coli</i>	22	1	1	0.99	24	126
Kanamycin	Ag	35	<i>S. enterica</i>	21.8	1	1	0.99	24	126
Ampicillin	Ag	3	<i>E. faecium</i>	0.02	1	1	0.99	24	121
Ampicillin	Ag	3	<i>S. aureus</i>	0.2	1	1	0.99	24	121
Ampicillin	Ag	3	<i>E. coli</i>	0.4	1	1	0.99	24	121
Ampicillin	Ag	3	<i>E. coli</i>	0.2	1	1	0.99	24	121
Ampicillin	Ag	3	<i>P. aeruginosa</i>	0.6	1	1	0.99	24	121
Ampicillin	Ag	3	<i>S. mutans</i>	0.2	1	1	0.99	24	121
Chloramphenicol	Ag	3	<i>E. faecium</i>	0.1	1	1	0.99	24	121
Chloramphenicol	Ag	3	<i>S. aureus</i>	0.3	1	1	0.99	24	121
Chloramphenicol	Ag	3	<i>E. coli</i>	0.7	1	1	0.99	24	121
Chloramphenicol	Ag	3	<i>E. coli</i>	0.2	1	1	0.99	24	121
Chloramphenicol	Ag	3	<i>S. epidermitis</i>	0.7	1	1	0.99	24	121
Chloramphenicol	Ag	3	<i>S. mutans</i>	0.1	1	1	0.99	24	121
Kanamycin	Ag	3	<i>E. faecium</i>	0.2	1	1	0.99	24	121

Kanamycin	Ag	3	<i>S. aureus</i>	0.2	1	1	0.99	24	121
Kanamycin	Ag	3	<i>E. coli</i>	0.5	1	1	0.99	24	121
Kanamycin	Ag	3	<i>E. coli</i>	0.2	1	1	0.99	24	121
Kanamycin	Ag	3	<i>P. aeruginosa</i>	0.5	1	1	0.99	24	121
Kanamycin	Ag	3	<i>S. mutans</i>	0.15	1	1	0.99	24	121
Vancomycin	Au	8.4	<i>S. aureus</i>	8	1	1	0.99	14	122
Vancomycin	Au	8.4	<i>E. coli</i>	8	1	1	0.99	14	122
Vancomycin	Au	8.4	<i>A. baumannii</i>	8	1	1	0.99	14	122
Vancomycin	Au	8.4	<i>E. faecalis</i>	16	1	1	0.99	14	122
Vancomycin	Au	8.4	<i>E. faecium</i>	16	1	1	0.99	14	122
Vancomycin	Au	8.4	<i>P. aeruginosa</i>	8	1	1	0.99	14	122
Vancomycin	Au	8.4	<i>E. faecalis</i>	16	1	1	0.99	14	122
Vancomycin	Au	8.4	<i>E. faecium</i>	32	1	1	0.99	14	122
Gentamicin	Ag	40	<i>S. aureus</i>	1.1	1	1	0.99	14	124
Gentamicin	Ag	40	<i>K. pneumoniae</i>	0.4	1	1	0.99	14	124
Gentamicin	Ag	40	<i>P. aeruginosa</i>	0.9	1	1	0.99	14	124
Gentamicin	Zn	60	<i>S. aureus</i>	1.1	1	1	0.99	14	124
Metal Oxide / Spherical									
Tobramycin	Fe ₃ O ₄	16	<i>P. aeruginosa</i>	0.2	1	1	0.99	24	113
Tobramycin	Fe ₃ O ₄	230	<i>P. aeruginosa</i>	0.2	1	1	0.99	24	113
Tobramycin	Fe ₃ O ₄	232	<i>P. aeruginosa</i>	0.2	1	1	0.99	24	113
Meropenem	CuO	15	<i>P. aeruginosa</i>	13.9	1	1	0.99	24	116
Ciprofloxacin	CuO	15	<i>P. aeruginosa</i>	4.8	1	1	0.99	24	116
Meropenem	ZnO	35	<i>P. aeruginosa</i>	34.7	1	1	0.99	24	116
Ciprofloxacin	ZnO	35	<i>P. aeruginosa</i>	4	1	1	0.99	24	116
Vancomycin	Mn ₂ F e ₂ O ₄	25	<i>S. epidermitis</i>	0.6	1	1	0.99	24	117
Vancomycin	Mn ₂ F e ₂ O ₄	25	<i>S. aureus</i>	0.8	1	1	0.99	24	117
Vancomycin	Mn ₂ F e ₂ O ₄	25	<i>B. subtilis</i>	0.8	1	1	0.99	24	117
Vancomycin	Mn ₂ F e ₂ O ₄	25	<i>S. aureus</i>	1	1	1	0.99	24	117
Vancomycin	Mn ₂ F e ₂ O ₄	25	<i>E. coli</i>	39.1	1	1	0.99	24	117
Metal Salt / Nanoflakes									
Ofloxacin	MoS ₂	175	<i>S. aureus</i>	25	1	1	0.99	24	120

^a Class. Obs: $f(v_{ij}, v_{nj})_{\text{robs}}$, ^b Class. Pred: $f(v_{ij}, v_{nj})_{\text{pred}}$, ^c p: probability, calculated as $p = \frac{1}{1 + \text{Exp}(-f(v_{ij}, v_{nj})_{\text{calc}})}$, ^d t(h): Time of assay, ^e Ref: Reference.

4. CONCLUSIONS

DADNP systems are a promising alternative to classic AD therapy in the current context of MDR bacteria emergencies. DADNP systems discovery is difficult due to the high number of AD and NP systems combinations and conditions to be tested. ML models may help, but the low number of real DADNP experimentally characterized complex applications. Additive IFPTML models may become a pragmatic solution, for the moment, by taking into consideration the higher abundance of experimental tests for DADNP components AD and NP alone. Regarding the methodological objectives, the linear model included two subsystems (preclinical antibacterial drugs and nanoparticles) and showed a good fit ($S_n=74.3\%$, $S_p=90.6\%$, and $A_c=89.4\%$). The information from the two subsystems did not significantly influence the robustness of the models to analyze the problem presented in the thesis.

Regarding the practical objectives, the IFPTML-LDA and IFPTML-ANN models' values are adequate $Sp \approx 88.5-90\%$ and $Sn \approx 74-88.5\%$ in training (>124K cases) and validation (>41K cases) series. In addition, IFPTML-LDA and IFPTML-LNN models are more straightforward and still accurate options to predict putative DADNP systems. IFPTML-ANN non-linear models of type IFPTML-MLP and other ML algorithms are more complicated but have better statistical parameters (Sn and $Sp \approx 85-99\%$, and AUROC = 0.99 in training/validation series outperformed). The IFPTML DADNP simulation experiment shows that the DADNP systems formed by Ciprofloxacin and Au NP coated with PDT/CQ, PDT/Mel, or PDT/Ach seem to be promising for further assays vs. MDR *P. aeruginosa* strains ($p=0.809-0.838$). The IFPTML linear and additive model was able to predict 80 experimental cases of DADNPs complexes reported in the literature with high structural and biological diversity. In conclusion, the IFPTML models in general may offer a fast and inexpensive solution to the pre-screening of putative DADNP systems in order to reduce costs and time in posterior experimental screening.

5. REFERENCES

1. Diéguez-Santana, K.; González-Díaz, H. Towards machine learning discovery of dual antibacterial drug–nanoparticle systems. *Nanoscale*. **2021**, *13* (42), 17854-17870, 10.1039/D1NR04178A. DOI: 10.1039/D1NR04178A.
2. Fischbach, M. A.; Walsh, C. T. Antibiotics for emerging pathogens. *Science*. **2009**, *325* (5944), 1089-1093, Review. DOI: 10.1126/science.1176667 Scopus.
3. McKenna, M. The antibiotic paradox: why companies can't afford to create life-saving drugs. *Nature*. **2020**, *584* (7821), 338-341. DOI: 10.1038/d41586-020-02418-x.
4. Najer, A.; Wu, D.; Nussbaumer, M. G.; Schwertz, G.; Schwab, A.; Witschel, M. C.; Schäfer, A.; Diederich, F.; Rottmann, M.; Palivan, C. G.; et al. An amphiphilic graft copolymer-based nanoparticle platform for reduction-responsive anticancer and antimalarial drug delivery. *Nanoscale*. **2016**, *8* (31), 14858-14869, Article. DOI: 10.1039/c6nr04290b Scopus.
5. Anwar, A.; Mungroo, M. R.; Anwar, A.; Sullivan, W. J., Jr.; Khan, N. A.; Siddiqui, R. Repositioning of Guanabenz in Conjugation with Gold and Silver Nanoparticles against Pathogenic Amoebae *Acanthamoeba castellanii* and *Naegleria fowleri*. *ACS Infect Dis*. **2019**, *5* (12), 2039-2046. DOI: 10.1021/acsinfecdis.9b00263.
6. Clemens, D. L.; Lee, B. Y.; Plamthottam, S.; Tullius, M. V.; Wang, R.; Yu, C. J.; Li, Z.; Dillon, B. J.; Zink, J. I.; Horwitz, M. A. Nanoparticle Formulation of Moxifloxacin and Intramuscular Route of Delivery Improve Antibiotic Pharmacokinetics and Treatment of Pneumonic Tularemia in a Mouse Model. *ACS Infect Dis*. **2019**, *5* (2), 281-291. DOI: 10.1021/acsinfecdis.8b00268.
7. Gao, F.; Xu, L.; Yang, B.; Fan, F.; Yang, L. Kill the Real with the Fake: Eliminate Intracellular *Staphylococcus aureus* Using Nanoparticle Coated with Its Extracellular Vesicle Membrane as Active-Targeting Drug Carrier. *ACS Infect Dis*. **2019**, *5* (2), 218-227. DOI: 10.1021/acsinfecdis.8b00212.
8. Tiwari, B.; Pahuja, R.; Kumar, P.; Rath, S. K.; Gupta, K. C.; Goyal, N. Nanotized Curcumin and Miltefosine, a Potential Combination for Treatment of Experimental Visceral Leishmaniasis. *Antimicrobial agents and chemotherapy*. **2017**, *61* (3). DOI: 10.1128/AAC.01169-16.
9. Tran, T. T.; Hadinoto, K. Ternary nanoparticle complex of antibiotic, polyelectrolyte, and mucolytic enzyme as a potential antibiotic delivery system in bronchiectasis therapy.

- Colloids and surfaces. B, Biointerfaces.* **2020**, *193*, 111095. DOI: 10.1016/j.colsurfb.2020.111095.
10. Zhao, X.; Jia, Y.; Li, J.; Dong, R.; Zhang, J.; Ma, C.; Wang, H.; Rui, Y.; Jiang, X. Indole Derivative-Capped Gold Nanoparticles as an Effective Bactericide in Vivo. *ACS Applied Materials & Interfaces.* **2018**, *10* (35), 29398-29406. DOI: 10.1021/acsami.8b11980.
 11. Slavin, Y. N.; Asnis, J.; Häfeli, U. O.; Bach, H. Metal nanoparticles: understanding the mechanisms behind antibacterial activity. *Journal of Nanobiotechnology.* **2017**, *15* (1), 65. DOI: 10.1186/s12951-017-0308-z.
 12. Gu, X.; Xu, Z.; Gu, L.; Xu, H.; Han, F.; Chen, B.; Pan, X. Preparation and antibacterial properties of gold nanoparticles: a review. *Environmental Chemistry Letters.* **2021**, *19* (1), 167-187. DOI: 10.1007/s10311-020-01071-0.
 13. Fatima, F.; Siddiqui, S.; Khan, W. A. Nanoparticles as Novel Emerging Therapeutic Antibacterial Agents in the Antibiotics Resistant Era. *Biological trace element research.* **2020**. DOI: 10.1007/s12011-020-02394-3.
 14. Chernousova, S.; Epple, M. Silver as antibacterial agent: ion, nanoparticle, and metal. *Angewandte Chemie.* **2013**, *52* (6), 1636-1653. DOI: 10.1002/anie.201205923.
 15. Shahbandeh, M.; Taati Moghadam, M.; Mirnejad, R.; Mirkalantari, S.; Mirzaei, M. The Efficacy of AgNO₃ Nanoparticles Alone and Conjugated with Imipenem for Combating Extensively Drug-Resistant *Pseudomonas aeruginosa*. *International journal of nanomedicine.* **2020**, *15*, 6905-6916. DOI: 10.2147/IJN.S260520.
 16. Mu, Y.; Wu, F.; Zhao, Q.; Ji, R.; Qie, Y.; Zhou, Y.; Hu, Y.; Pang, C.; Hristozov, D.; Giesy, J. P.; et al. Predicting toxic potencies of metal oxide nanoparticles by means of nano-QSARs. *Nanotoxicology.* **2016**, *10* (9), 1207-1214. DOI: 10.1080/17435390.2016.1202352.
 17. Puzyn, T.; Rasulev, B.; Gajewicz, A.; Hu, X.; Dasari, T. P.; Michalkova, A.; Hwang, H. M.; Toropov, A.; Leszczynska, D.; Leszczynski, J. Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nat Nanotechnol.* **2011**, *6* (3), 175-178. DOI: 10.1038/nnano.2011.10.
 18. Pan, Y.; Li, T.; Cheng, J.; Telesca, D.; Zink, J. I.; Jiang, J. Nano-QSAR modeling for predicting the cytotoxicity of metal oxide nanoparticles using novel descriptors. *RSC Advances.* **2016**, *6* (31), 25766-25775, 10.1039/C6RA01298A. DOI: 10.1039/C6RA01298A.
 19. Fjodorova, N.; Novic, M.; Gajewicz, A.; Rasulev, B. The way to cover prediction for cytotoxicity for all existing nano-sized metal oxides by using neural network method. *Nanotoxicology.* **2017**, *11* (4), 475-483. DOI: 10.1080/17435390.2017.1310949.
 20. Zhou, Z.; Tang, X.; Dai, W.; Shi, J.; Chen, H. Nano-QSAR models for predicting cytotoxicity of metal oxide nanoparticles (MONPs) to *E. coli*. **2017**, *95* (8), 863-866. DOI: 10.1139/cjc-2017-0172.
 21. Kaweeterawat, C.; Ivask, A.; Liu, R.; Zhang, H.; Chang, C. H.; Low-Kam, C.; Fischer, H.; Ji, Z.; Pokhrel, S.; Cohen, Y.; et al. Toxicity of Metal Oxide Nanoparticles in *Escherichia coli* Correlates with Conduction Band and Hydration Energies. *Environmental Science & Technology.* **2015**, *49* (2), 1105-1112. DOI: 10.1021/es504259s.
 22. Kar, S.; Pathakoti, K.; Tchounwou, P. B.; Leszczynska, D.; Leszczynski, J. Evaluating the cytotoxicity of a large pool of metal oxide nanoparticles to *Escherichia coli*: Mechanistic understanding through In Vitro and In Silico studies. *Chemosphere.* **2021**, *264*, 128428. DOI: 10.1016/j.chemosphere.2020.128428.
 23. Yu, Y.; O'Rourke, A.; Lin, Y. H.; Singh, H.; Eguez, R. V.; Beyhan, S.; Nelson, K. E. Predictive Signatures of 19 Antibiotic-Induced *Escherichia coli* Proteomes. *ACS Infect Dis.* **2020**, *6* (8), 2120-2129. DOI: 10.1021/acsinfectdis.0c00196.

24. Pribut, N.; Kaiser, T. M.; Wilson, R. J.; Jecs, E.; Dentmon, Z. W.; Pelly, S. C.; Sharma, S.; Bartsch, P. W., 3rd; Burger, P. B.; Hwang, S. S.; et al. Accelerated Discovery of Potent Fusion Inhibitors for Respiratory Syncytial Virus. *ACS Infect Dis.* **2020**, *6* (5), 922-929. DOI: 10.1021/acsinfecdis.9b00524.
25. Wang, X.; Perryman, A. L.; Li, S. G.; Paget, S. D.; Stratton, T. P.; Lemenze, A.; Olson, A. J.; Ekins, S.; Kumar, P.; Freundlich, J. S. Intrabacterial Metabolism Obscures the Successful Prediction of an InhA Inhibitor of *Mycobacterium tuberculosis*. *ACS Infect Dis.* **2019**, *5* (12), 2148-2163. DOI: 10.1021/acsinfecdis.9b00295.
26. Cooper, S. J.; Krishnamoorthy, G.; Wolloscheck, D.; Walker, J. K.; Rybenkov, V. V.; Parks, J. M.; Zgurskaya, H. I. Molecular Properties That Define the Activities of Antibiotics in *Escherichia coli* and *Pseudomonas aeruginosa*. *ACS Infect Dis.* **2018**, *4* (8), 1223-1234. DOI: 10.1021/acsinfecdis.8b00036.
27. Duncan, G. A.; Bevan, M. A. Computational design of nanoparticle drug delivery systems for selective targeting. *Nanoscale.* **2015**, *7* (37), 15332-15340, Article. DOI: 10.1039/c5nr03691g Scopus.
28. Gajewicz, A. What if the number of nanotoxicity data is too small for developing predictive Nano-QSAR models? An alternative read-across based approach for filling data gaps. *Nanoscale.* **2017**, *9* (24), 8435-8448, Article. DOI: 10.1039/c7nr02211e Scopus.
29. Urista, D. V.; Carrue, D. B.; Otero, I.; Arrasate, S.; Quevedo-Tumaili, V. F.; Gestal, M.; Gonzalez-Diaz, H.; Munteanu, C. R. Prediction of Antimalarial Drug-Decorated Nanoparticle Delivery Systems with Random Forest Models. *Biology.* **2020**, *9* (8). DOI: 10.3390/biology9080198.
30. Santana, R.; Zuluaga, R.; Gañán, P.; Arrasate, S.; Onieva, E.; González-Díaz, H. Designing nanoparticle release systems for drug-vitamin cancer co-therapy with multiplicative perturbation-theory machine learning (PTML) models. *Nanoscale.* **2019**, *11* (45), 21811-21823, Article. DOI: 10.1039/c9nr05070a Scopus.
31. Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research.* **2019**, *47* (D1), D930-D940. DOI: 10.1093/nar/gky1075.
32. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; et al. The ChEMBL database in 2017. *Nucleic acids research.* **2017**, *45* (D1), D945-D954. DOI: 10.1093/nar/gkw1074.
33. Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Kruger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The ChEMBL bioactivity database: an update. *Nucleic acids research.* **2014**, *42* (Database issue), D1083-1090. DOI: 10.1093/nar/gkt1031.
34. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research.* **2012**, *40* (Database issue), D1100-1107. DOI: 10.1093/nar/gkr777 From NLM.
35. Ebejer, J. P.; Charlton, M. H.; Finn, P. W. Are the physicochemical properties of antibacterial compounds really different from other drugs? *Journal of cheminformatics.* **2016**, *8*, 30. DOI: 10.1186/s13321-016-0143-5.
36. Nabil, A.; Elshemy, M. M.; Asem, M.; Abdel-Motaal, M.; Gomaa, H. F.; Zahran, F.; Uto, K.; Ebara, M. Zinc Oxide Nanoparticle Synergizes Sorafenib Anticancer Efficacy with Minimizing Its Cytotoxicity. *Oxidative medicine and cellular longevity.* **2020**, *2020*, 1362104. DOI: 10.1155/2020/1362104.

37. Caron, W. P.; Morgan, K. P.; Zamboni, B. A.; Zamboni, W. C. A review of study designs and outcomes of phase I clinical studies of nanoparticle agents compared with small-molecule anticancer agents. *Clinical cancer research : an official journal of the American Association for Cancer Research*. **2013**, *19* (12), 3309-3315. DOI: 10.1158/1078-0432.CCR-12-3649.
38. Ruparelia, J. P.; Chatterjee, A. K.; Duttagupta, S. P.; Mukherji, S. Strain specificity in antimicrobial activity of silver and copper nanoparticles. *Acta Biomater*. **2008**, *4* (3), 707-716. DOI: 10.1016/j.actbio.2007.11.006.
39. Pramanik, A.; Laha, D.; Bhattacharya, D.; Pramanik, P.; Karmakar, P. A novel study of antibacterial activity of copper iodide nanoparticle mediated by DNA and membrane damage. *Colloids and surfaces. B, Biointerfaces*. **2012**, *96*, 50-55. DOI: 10.1016/j.colsurfb.2012.03.021.
40. Azam, A.; Ahmed, A. S.; Oves, M.; Khan, M. S.; Habib, S. S.; Memic, A. Antimicrobial activity of metal oxide nanoparticles against Gram-positive and Gram-negative bacteria: a comparative study. *International journal of nanomedicine*. **2012**, *7*, 6003-6009. DOI: 10.2147/IJN.S35347.
41. Azam, A.; Ahmed, A. S.; Oves, M.; Khan, M. S.; Memic, A. Size-dependent antimicrobial properties of CuO nanoparticles against Gram-positive and -negative bacterial strains. *International journal of nanomedicine*. **2012**, *7*, 3527-3535. DOI: 10.2147/IJN.S29020.
42. Hossain, S. T.; Mukherjee, S. K. Toxicity of cadmium sulfide (CdS) nanoparticles against Escherichia coli and HeLa cells. *Journal of hazardous materials*. **2013**, *260*, 1073-1082. DOI: 10.1016/j.jhazmat.2013.07.005.
43. Botequim, D.; Maia, J.; Lino, M. M.; Lopes, L. M.; Simoes, P. N.; Ilharco, L. M.; Ferreira, L. Nanoparticles and surfaces presenting antifungal, antibacterial and antiviral properties. *Langmuir*. **2012**, *28* (20), 7646-7656. DOI: 10.1021/la300948n.
44. Taglietti, A.; Diaz Fernandez, Y. A.; Amato, E.; Cucca, L.; Dacarro, G.; Grisoli, P.; Necchi, V.; Pallavicini, P.; Pasotti, L.; Patrini, M. Antibacterial activity of glutathione-coated silver nanoparticles against gram positive and gram negative bacteria. *Langmuir*. **2012**, *28* (21), 8140-8148, Article. DOI: 10.1021/la3003838 Scopus.
45. Hossain, S. T.; Mukherjee, S. K. CdO nanoparticle toxicity on growth, morphology, and cell division in Escherichia coli. *Langmuir*. **2012**, *28* (48), 16614-16622, Article. DOI: 10.1021/la302872y Scopus.
46. Premanathan, M.; Karthikeyan, K.; Jeyasubramanian, K.; Manivannan, G. Selective toxicity of ZnO nanoparticles toward Gram-positive bacteria and cancer cells by apoptosis through lipid peroxidation. *Nanomedicine*. **2011**, *7* (2), 184-192. DOI: 10.1016/j.nano.2010.10.001.
47. Inbaraj, B. S.; Kao, T. H.; Tsai, T. Y.; Chiu, C. P.; Kumar, R.; Chen, B. H. The synthesis and characterization of poly(gamma-glutamic acid)-coated magnetite nanoparticles and their effects on antibacterial activity and cytotoxicity. *Nanotechnology*. **2011**, *22* (7), 075101. DOI: 10.1088/0957-4484/22/7/075101.
48. Hu, X.; Cook, S.; Wang, P.; Hwang, H. M. In vitro evaluation of cytotoxicity of engineered metal oxide nanoparticles. *Sci Total Environ*. **2009**, *407* (8), 3070-3072. DOI: 10.1016/j.scitotenv.2009.01.033.
49. Zhao, Y.; Chen, Z.; Chen, Y.; Xu, J.; Li, J.; Jiang, X. Synergy of non-antibiotic drugs and pyrimidinethiol on gold nanoparticles against superbugs. *Journal of the American Chemical Society*. **2013**, *135* (35), 12940-12943. DOI: 10.1021/ja4058635.
50. Zhen, J. B.; Kang, P. W.; Zhao, M. H.; Yang, K. W. Silver Nanoparticle Conjugated Star PCL-b-AMPs Copolymer as Nanocomposite Exhibits Efficient Antibacterial Properties. *Bioconjugate chemistry*. **2020**, *31* (1), 51-63. DOI: 10.1021/acs.bioconjchem.9b00739.

51. Arasoglu, T.; Derman, S.; Mansuroglu, B. Comparative evaluation of antibacterial activity of caffeic acid phenethyl ester and PLGA nanoparticle formulation by different methods. *Nanotechnology*. **2016**, *27* (2), 025103. DOI: 10.1088/0957-4484/27/2/025103.
52. Elizabeth, E.; Baranwal, G.; Krishnan, A. G.; Menon, D.; Nair, M. ZnO nanoparticle incorporated nanostructured metallic titanium for increased mesenchymal stem cell response and antibacterial activity. *Nanotechnology*. **2014**, *25* (11), 115101. DOI: 10.1088/0957-4484/25/11/115101.
53. Wong, M. S.; Chen, C. W.; Hsieh, C. C.; Hung, S. C.; Sun, D. S.; Chang, H. H. Antibacterial property of Ag nanoparticle-impregnated N-doped titania films under visible light. *Scientific reports*. **2015**, *5*, 11978. DOI: 10.1038/srep11978.
54. Zhou, H.; Cao, H.; Matyunina, L.; Shelby, M.; Cassels, L.; McDonald, J. F.; Skolnick, J. MEDICASCY: A Machine Learning Approach for Predicting Small-Molecule Drug Side Effects, Indications, Efficacy, and Modes of Action. *Molecular pharmaceutics*. **2020**, *17* (5), 1558-1574. DOI: 10.1021/acs.molpharmaceut.9b01248.
55. Sun, L.; Yang, H.; Cai, Y.; Li, W.; Liu, G.; Tang, Y. In Silico Prediction of Endocrine Disrupting Chemicals Using Single-Label and Multilabel Models. *J Chem Inf Model*. **2019**, *59* (3), 973-982. DOI: 10.1021/acs.jcim.8b00551.
56. Kolesov, A.; Kamyshenkov, D.; Litovchenko, M.; Smekalova, E.; Golovizin, A.; Zhavoronkov, A. On multilabel classification methods of incompletely labeled biomedical text data. *Computational and mathematical methods in medicine*. **2014**, *2014*, 781807. DOI: 10.1155/2014/781807.
57. Heider, D.; Senge, R.; Cheng, W.; Hullermeier, E. Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. *Bioinformatics*. **2013**, *29* (16), 1946-1952. DOI: 10.1093/bioinformatics/btt331.
58. Serafim, M. S. M.; Kronenberger, T.; Oliveira, P. R.; Poso, A.; Honorio, K. M.; Mota, B. E. F.; Maltarollo, V. G. The application of machine learning techniques to innovative antibacterial discovery and development. *Expert Opin Drug Discov*. **2020**, *15* (10), 1165-1180. DOI: 10.1080/17460441.2020.1776696.
59. Durrant, J. D.; Amaro, R. E. Machine-learning techniques applied to antibacterial drug discovery. *Chem Biol Drug Des*. **2015**, *85* (1), 14-21. DOI: 10.1111/cbdd.12423.
60. Khosravian, M.; Faramarzi, F. K.; Beigi, M. M.; Behbahani, M.; Mohabatkar, H. Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. *Protein and peptide letters*. **2013**, *20* (2), 180-186. DOI: 10.2174/092986613804725307.
61. Fjell, C. D.; Jenssen, H.; Hilpert, K.; Cheung, W. A.; Pante, N.; Hancock, R. E.; Cherkasov, A. Identification of novel antibacterial peptides by chemoinformatics and machine learning. *Journal of medicinal chemistry*. **2009**, *52* (7), 2006-2015. DOI: 10.1021/jm8015365.
62. Yang, X. G.; Chen, D.; Wang, M.; Xue, Y.; Chen, Y. Z. Prediction of antibacterial compounds by machine learning approaches. *J Comput Chem*. **2009**, *30* (8), 1202-1211. DOI: 10.1002/jcc.21148.
63. Manganeli, S.; Leone, C.; Toropov, A. A.; Toropova, A. P.; Benfenati, E. QSAR model for predicting cell viability of human embryonic kidney cells exposed to SiO₂ nanoparticles. *Chemosphere*. **2016**, *144*, 995-1001. DOI: 10.1016/j.chemosphere.2015.09.086.
64. Toropova, A. P.; Toropov, A. A.; Rallo, R.; Leszczynska, D.; Leszczynski, J. Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions. *Ecotoxicology and environmental safety*. **2015**, *112*, 39-45. DOI: 10.1016/j.ecoenv.2014.10.003.

65. Toropova, A. P.; Toropov, A. A.; Veselinovic, A. M.; Veselinovic, J. B.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. Nano-QSAR: Model of mutagenicity of fullerene as a mathematical function of different conditions. *Ecotoxicology and environmental safety*. **2016**, *124*, 32-36. DOI: 10.1016/j.ecoenv.2015.09.038.
66. Rybinska-Fryca, A.; Mikolajczyk, A.; Puzyn, T. Structure-activity prediction networks (SAPNets): a step beyond Nano-QSAR for effective implementation of the safe-by-design concept. *Nanoscale*. **2020**. DOI: 10.1039/d0nr05220e.
67. Le, T. C.; Yin, H.; Chen, R.; Chen, Y.; Zhao, L.; Casey, P. S.; Chen, C.; Winkler, D. A. An Experimental and Computational Approach to the Development of ZnO Nanoparticles that are Safe by Design. *Small*. **2016**, *12* (26), 3568-3577. DOI: 10.1002/sml.201600597.
68. Ahmadi, S.; Toropova, A. P.; Toropov, A. A. Correlation intensity index: mathematical modeling of cytotoxicity of metal oxide nanoparticles. *Nanotoxicology*. **2020**, 1-9. DOI: 10.1080/17435390.2020.1808252.
69. Ojha, P. K.; Kar, S.; Roy, K.; Leszczynski, J. Toward comprehension of multiple human cells uptake of engineered nano metal oxides: quantitative inter cell line uptake specificity (QICLUS) modeling. *Nanotoxicology*. **2019**, *13* (1), 14-34, Article. DOI: 10.1080/17435390.2018.1529836 Scopus.
70. Sizochenko, N.; Gajewicz, A.; Leszczynski, J.; Puzyn, T. Reply to the comment on "Causation or only correlation? Application of causal inference graphs for evaluating causality in nano-QSAR models" by D. A. Tasi, J. Csontos, B. Nagy, Z. Konya and G. Tasi, *Nanoscale*, 2018, 10, C8NR02377H. *Nanoscale*. **2018**, *10* (44), 20867-20868. DOI: 10.1039/c8nr07975g.
71. Tasi, D. A.; Csontos, J.; Nagy, B.; Konya, Z.; Tasi, G. Comment on "Causation or only correlation? Application of causal inference graphs for evaluating causality in nano-QSAR models" by N. Sizochenko, A. Gajewicz, J. Leszczynski and T. Puzyn, *Nanoscale*, 2016, 8, 7203. *Nanoscale*. **2018**, *10* (44), 20863-20866. DOI: 10.1039/c8nr02377h.
72. Villaverde, J. J.; Sevilla-Moran, B.; Lopez-Goti, C.; Alonso-Prados, J. L.; Sandin-Espana, P. Considerations of nano-QSAR/QSPR models for nanopesticide risk assessment within the European legislative framework. *Sci Total Environ*. **2018**, *634*, 1530-1539. DOI: 10.1016/j.scitotenv.2018.04.033.
73. Sizochenko, N.; Leszczynska, D.; Leszczynski, J. Modeling of Interactions between the Zebrafish Hatching Enzyme ZHE1 and A Series of Metal Oxide Nanoparticles: Nano-QSAR and Causal Analysis of Inactivation Mechanisms. *Nanomaterials*. **2017**, *7* (10). DOI: 10.3390/nano7100330.
74. Manganelli, S.; Benfenati, E. Nano-QSAR Model for Predicting Cell Viability of Human Embryonic Kidney Cells. *Methods in molecular biology*. **2017**, *1601*, 275-290. DOI: 10.1007/978-1-4939-6960-9_22.
75. Sizochenko, N.; Mikolajczyk, A.; Jagiello, K.; Puzyn, T.; Leszczynski, J.; Rasulev, B. How the toxicity of nanomaterials towards different species could be simultaneously evaluated: A novel multi-nano-read-across approach. *Nanoscale*. **2018**, *10* (2), 582-591, Article. DOI: 10.1039/c7nr05618d Scopus.
76. Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Gini, G.; Puzyn, T.; Leszczynska, D.; Leszczynski, J. Novel application of the CORAL software to model cytotoxicity of metal oxide nanoparticles to bacteria *Escherichia coli*. *Chemosphere*. **2012**, *89* (9), 1098-1102, Article. DOI: 10.1016/j.chemosphere.2012.05.077 Scopus.
77. Gonzalez-Diaz, H.; Arrasate, S.; Gomez-SanJuan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. M. General theory for multiple input-output perturbations in complex molecular systems. 1. Linear QSPR electronegativity models in physical, organic, and

- medicinal chemistry. *Current topics in medicinal chemistry*. **2013**, *13* (14), 1713-1741. DOI: 10.2174/1568026611313140011.
78. Alonso, N.; Caamano, O.; Romero-Duran, F. J.; Luan, F.; MN, D. S. C.; Yanez, M.; Gonzalez-Diaz, H.; Garcia-Mera, X. Model for high-throughput screening of multitarget drugs in chemical neurosciences: synthesis, assay, and theoretic study of rasagiline carbamates. *ACS Chem Neurosci*. **2013**, *4* (10), 1393-1403. DOI: 10.1021/cn400111n.
79. Diez-Alarcia, R.; Yanez-Perez, V.; Muneta-Arrate, I.; Arrasate, S.; Lete, E.; Meana, J. J.; Gonzalez-Diaz, H. Big Data Challenges Targeting Proteins in GPCR Signaling Pathways; Combining PTML-ChEMBL Models and [(35)S]GTPgammaS Binding Assays. *ACS Chem Neurosci*. **2019**, *10* (11), 4476-4491. DOI: 10.1021/acchemneuro.9b00302.
80. Gonzalez-Diaz, H.; Riera-Fernandez, P.; Pazos, A.; Munteanu, C. R. The Rucker-Markov invariants of complex Bio-Systems: applications in Parasitology and Neuroinformatics. *Bio Systems*. **2013**, *111* (3), 199-207. DOI: 10.1016/j.biosystems.2013.02.006.
81. Duardo-Sanchez, A.; Munteanu, C. R.; Riera-Fernandez, P.; Lopez-Diaz, A.; Pazos, A.; Gonzalez-Diaz, H. Modeling Complex Metabolic Reactions, Ecological Systems, and Financial and Legal Networks with MIANN Models Based on Markov-Wiener Node Descriptors. *Journal of Chemical Information and Modeling*. **2014**, *54* (1), 16-29. DOI: 10.1021/ci400280n.
82. Gonzalez-Diaz, H.; Riera-Fernandez, P. New Markov-autocorrelation indices for re-evaluation of links in chemical and biological complex networks used in metabolomics, parasitology, neurosciences, and epidemiology. *J Chem Inf Model*. **2012**, *52* (12), 3331-3340. DOI: 10.1021/ci300321f.
83. Concu, R.; MN, D. S. C.; Munteanu, C. R.; Gonzalez-Diaz, H. PTML Model of Enzyme Subclasses for Mining the Proteome of Biofuel Producing Microorganisms. *J Proteome Res*. **2019**, *18* (7), 2735-2746. DOI: 10.1021/acs.jproteome.8b00949.
84. Martinez-Arzate, S. G.; Tenorio-Borroto, E.; Barbabosa Pliego, A.; Diaz-Albiter, H. M.; Vazquez-Chagoyan, J. C.; Gonzalez-Diaz, H. PTML Model for Proteome Mining of B-Cell Epitopes and Theoretical-Experimental Study of Bm86 Protein Sequences from Colima, Mexico. *J Proteome Res*. **2017**, *16* (11), 4093-4103. DOI: 10.1021/acs.jproteome.7b00477.
85. Quevedo-Tumaili, V. F.; Ortega-Tenezaca, B.; Gonzalez-Diaz, H. Chromosome Gene Orientation Inversion Networks (GOINs) of Plasmodium Proteome. *J Proteome Res*. **2018**, *17* (3), 1258-1268. DOI: 10.1021/acs.jproteome.7b00861.
86. Kleandrova, V. V.; Luan, F.; Gonzalez-Diaz, H.; Ruso, J. M.; Speck-Planche, A.; Cordeiro, M. N. Computational tool for risk assessment of nanomaterials: novel QSTR-perturbation model for simultaneous prediction of ecotoxicity and cytotoxicity of uncoated and coated nanoparticles under multiple experimental conditions. *Environ Sci Technol*. **2014**, *48* (24), 14686-14694. DOI: 10.1021/es503861x.
87. Luan, F.; Kleandrova, V. V.; González-Díaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, M. N. D. S. Computer-aided nanotoxicology: Assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. *Nanoscale*. **2014**, *6* (18), 10623-10630, Article. DOI: 10.1039/c4nr01285b Scopus.
88. Santana, R.; Zuluaga, R.; Ganan, P.; Arrasate, S.; Onieva, E.; Montemore, M. M.; Gonzalez-Diaz, H. PTML Model for Selection of Nanoparticles, Anticancer Drugs, and Vitamins in the Design of Drug-Vitamin Nanoparticle Release Systems for Cancer Cotherapy. *Mol Pharm*. **2020**, *17* (7), 2612-2627. DOI: 10.1021/acs.molpharmaceut.0c00308.

89. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. Computational modeling in nanomedicine: prediction of multiple antibacterial profiles of nanoparticles using a quantitative structure-activity relationship perturbation model. *Nanomedicine (Lond)*. **2015**, *10* (2), 193-204. DOI: 10.2217/nnm.14.96.
90. Nocado-Mena, D.; Cornelio, C.; Camacho-Corona, M. D. R.; Garza-Gonzalez, E.; Waksman de Torres, N.; Arrasate, S.; Sotomayor, N.; Lete, E.; Gonzalez-Diaz, H. Modeling Antibacterial Activity with Machine Learning and Fusion of Chemical Structure Information with Microorganism Metabolic Networks. *J Chem Inf Model*. **2019**, *59* (3), 1109-1120. DOI: 10.1021/acs.jcim.9b00034.
91. Li, Y.; Li, H.; Pickard, F. C. t.; Narayanan, B.; Sen, F. G.; Chan, M. K. Y.; Sankaranarayanan, S.; Brooks, B. R.; Roux, B. Machine Learning Force Field Parameters from Ab Initio Data. *Journal of chemical theory and computation*. **2017**, *13* (9), 4492-4503. DOI: 10.1021/acs.jctc.7b00521.
92. Xia, R.; Kais, S. Quantum machine learning for electronic structure calculations. *Nat Commun*. **2018**, *9* (1), 4195. DOI: 10.1038/s41467-018-06598-z.
93. Na, G. S.; Chang, H.; Kim, H. W. Machine-guided representation for accurate graph-based molecular machine learning. *Physical chemistry chemical physics : PCCP*. **2020**, *22* (33), 18526-18535. DOI: 10.1039/d0cp02709j.
94. Santana, R.; Zuluaga, R.; Gañán, P.; Arrasate, S.; Onieva, E.; González-Díaz, H. Predicting coated-nanoparticle drug release systems with perturbation-theory machine learning (PTML) models. *Nanoscale*. **2020**, *12* (25), 13471-13483, Article. DOI: 10.1039/d0nr01849j Scopus.
95. Hill, T.; Lewicki, P. *Statistics: Methods and Applications*; StatSoft, Inc., 2005.
96. Huberty, C. J.; Olejnik, S. *Applied MANOVA and discriminant analysis*; John Wiley & Sons, Inc., 2006.
97. Hanczar, B.; Hua, J.; Sima, C.; Weinstein, J.; Bittner, M.; Dougherty, E. R. Small-sample precision of ROC-related estimates. *Bioinformatics*. **2010**, *26* (6), 822-830, Article. DOI: 10.1093/bioinformatics/btq037 Scopus.
98. Bian, L.; Sorescu, D. C.; Chen, L.; White, D. L.; Burkert, S. C.; Khalifa, Y.; Zhang, Z.; Sejdic, E.; Star, A. Machine-Learning Identification of the Sensing Descriptors Relevant in Molecular Interactions with Metal Nanoparticle-Decorated Nanotube Field-Effect Transistors. *ACS Appl Mater Interfaces*. **2019**, *11* (1), 1219-1227. DOI: 10.1021/acsami.8b15785.
99. Alafeef, M.; Srivastava, I.; Pan, D. Machine Learning for Precision Breast Cancer Diagnosis and Prediction of the Nanoparticle Cellular Internalization. *ACS sensors*. **2020**, *5* (6), 1689-1698. DOI: 10.1021/acssensors.0c00329.
100. Sun, B.; Fernandez, M.; Barnard, A. S. Machine Learning for Silver Nanoparticle Electron Transfer Property Prediction. *J Chem Inf Model*. **2017**, *57* (10), 2413-2423. DOI: 10.1021/acs.jcim.7b00272.
101. Barnard, A. S.; Opletal, G. Predicting structure/property relationships in multi-dimensional nanoparticle data using t-distributed stochastic neighbour embedding and machine learning. *Nanoscale*. **2019**, *11* (48), 23165-23172. DOI: 10.1039/c9nr03940f.
102. He, J.; He, C.; Zheng, C.; Wang, Q.; Ye, J. Plasmonic nanoparticle simulations and inverse design using machine learning. *Nanoscale*. **2019**, *11* (37), 17444-17459. DOI: 10.1039/c9nr03450a.
103. Yan, T.; Sun, B.; Barnard, A. S. Predicting archetypal nanoparticle shapes using a combination of thermodynamic theory and machine learning. *Nanoscale*. **2018**, *10* (46), 21818-21826. DOI: 10.1039/c8nr07341d.

104. Concu, R.; Kleandrova, V. V.; Speck-Planche, A.; Cordeiro, M. Probing the toxicity of nanoparticles: a unified in silico machine learning model based on perturbation theory. *Nanotoxicology*. **2017**, *11* (7), 891-906. DOI: 10.1080/17435390.2017.1379567.
105. Van Den Berg, H. A. Occam's razor: from Ockham's via moderna to modern data science. *Science progress*. **2018**, *101* (3), 261-272. DOI: 10.3184/003685018X15295002645082.
106. Frank, E.; Hall, M. A.; Witten, I. H. *The WEKA workbench*; Morgan Kaufmann, 2016.
107. Quinlan, R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers, 1993.
108. Breiman, L. Random Forests. *Machine Learning*. **2001**, *45* (1), 5-32, journal article. DOI: 10.1023/a:1010933404324.
109. Aha, D. W.; Kibler, D.; Albert, M. K. Instance-based learning algorithms. *Machine Learning*. **1991**, *6* (1), 37-66. DOI: 10.1007/BF00153759.
110. Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*. **2018**, *23* (8), 1538-1546. DOI: 10.1016/j.drudis.2018.05.010.
111. Lehar, S. M.; Pillow, T.; Xu, M.; Staben, L.; Kajihara, K. K.; Vandlen, R.; DePalatis, L.; Raab, H.; Hazenbos, W. L.; Hiroshi Morisaki, J.; et al. Novel antibody-antibiotic conjugate eliminates intracellular *S. aureus*. *Nature*. **2015**, *527* (7578), 323-328, Article. DOI: 10.1038/nature16057 Scopus.
112. Hühn, J.; Hüllermeier, E. J. D. M.; Discovery, K. FURIA: an algorithm for unordered fuzzy rule induction. **2009**, *19* (3), 293-319, journal article. DOI: 10.1007/s10618-009-0131-8.
113. Armijo, L. M.; Wawrzyniec, S. J.; Kopciuch, M.; Brandt, Y. I.; Rivera, A. C.; Withers, N. J.; Cook, N. C.; Huber, D. L.; Monson, T. C.; Smyth, H. D. C.; et al. Antibacterial activity of iron oxide, iron nitride, and tobramycin conjugated nanoparticles against *Pseudomonas aeruginosa* biofilms. *Journal of Nanobiotechnology*. **2020**, *18* (1), 35. DOI: 10.1186/s12951-020-0588-6.
114. Burygin, G. L.; Khlebtsov, B. N.; Shantrokha, A. N.; Dykman, L. A.; Bogatyrev, V. A.; Khlebtsov, N. G. On the Enhanced Antibacterial Activity of Antibiotics Mixed with Gold Nanoparticles. *Nanoscale Res Lett*. **2009**, *4* (8), 794-801. DOI: 10.1007/s11671-009-9316-8 PubMed.
115. Djafari, J.; Marinho, C.; Santos, T.; Igrejas, G.; Torres, C.; Capelo, J. L.; Poeta, P.; Lodeiro, C.; Fernández-Lodeiro, J. New Synthesis of Gold- and Silver-Based Nano-Tetracycline Composites. *ChemistryOpen*. **2016**, *5* (3), 206-212. DOI: <https://doi.org/10.1002/open.201600016>.
116. Eleftheriadou, I.; Giannousi, K.; Protonotariou, E.; Skoura, L.; Arsenakis, M.; Dendrinou-Samara, C.; Sivropoulou, A. Cocktail of CuO, ZnO, or CuZn Nanoparticles and Antibiotics for Combating Multidrug-Resistant *Pseudomonas aeruginosa* via Efflux Pump Inhibition. *ACS Applied Nano Materials*. **2021**, *4* (9), 9799-9810. DOI: 10.1021/acsanm.1c02208.
117. Esmaeili, A.; Ghobadianpour, S. Vancomycin loaded superparamagnetic MnFe₂O₄ nanoparticles coated with PEGylated chitosan to enhance antibacterial activity. *International Journal of Pharmaceutics*. **2016**, *501* (1), 326-330. DOI: 10.1016/j.ijpharm.2016.02.013.
118. Esmaeillou, M.; Zarrini, G.; Ahangarzadeh Rezaee, M.; Shahbazi Mojarrad, J.; Bahadori, A. Vancomycin Capped with Silver Nanoparticles as an Antibacterial Agent against Multi-Drug Resistance Bacteria. *Adv Pharm Bull*. **2017**, *7* (3), 479-483. DOI: 10.15171/apb.2017.058 PubMed.

119. Gu, H.; Ho, P. L.; Tong, E.; Wang, L.; Xu, B. Presenting Vancomycin on Nanoparticles to Enhance Antimicrobial Activities. *Nano Letters*. **2003**, *3* (9), 1261-1263. DOI: 10.1021/nl034396z.
120. Huang, Y.; Gao, Q.; Li, X.; Gao, Y.; Han, H.; Jin, Q.; Yao, K.; Ji, J. Ofloxacin loaded MoS₂ nanoflakes for synergistic mild-temperature photothermal/antibiotic therapy with reduced drug resistance of bacteria. *Nano Research*. **2020**, *13* (9), 2340-2350. DOI: 10.1007/s12274-020-2853-2.
121. Hwang, I. S.; Hwang, J. H.; Choi, H.; Kim, K. J.; Lee, D. G. Synergistic effects between silver nanoparticles and antibiotics and the mechanisms involved. *Journal of medical microbiology*. **2012**, *61* (Pt 12), 1719-1726. DOI: 10.1099/jmm.0.047100-0 From NLM.
122. Lai, H.-Z.; Chen, W.-Y.; Wu, C.-Y.; Chen, Y.-C. Potent Antibacterial Nanoparticles for Pathogenic Bacteria. *ACS Applied Materials & Interfaces*. **2015**, *7* (3), 2046-2054. DOI: 10.1021/am507919m.
123. Meeker, D. G.; Jenkins, S. V.; Miller, E. K.; Beenken, K. E.; Loughran, A. J.; Powless, A.; Muldoon, T. J.; Galanzha, E. I.; Zharov, V. P.; Smeltzer, M. S.; et al. Synergistic Photothermal and Antibiotic Killing of Biofilm-Associated Staphylococcus aureus Using Targeted Antibiotic-Loaded Gold Nanoconstructs. *ACS Infectious Diseases*. **2016**, *2* (4), 241-250. DOI: 10.1021/acsinfecdis.5b00117.
124. Punjabi, K.; Mehta, S.; Chavan, R.; Chitalia, V.; Deogharkar, D.; Deshpande, S. Efficiency of Biosynthesized Silver and Zinc Nanoparticles Against Multi-Drug Resistant Pathogens. **2018**, *9* (2207), Original Research. DOI: 10.3389/fmicb.2018.02207.
125. Saha, B.; Bhattacharya, J.; Mukherjee, A.; Ghosh, A.; Santra, C.; Dasgupta, A. K.; Karmakar, P. In Vitro Structural and Functional Evaluation of Gold Nanoparticles Conjugated Antibiotics. *Nanoscale Res Lett*. **2007**, *2* (12), 614-622. DOI: 10.1007/s11671-007-9104-2 PMC.
126. Vazquez-Muñoz, R.; Meza-Villezcás, A.; Fournier, P. G. J.; Soria-Castro, E.; Juárez-Moreno, K.; Gallego-Hernández, A. L.; Bogdanchikova, N.; Vazquez-Duhalt, R.; Huerta-Saquero, A. Enhancement of antibiotics antimicrobial activity due to the silver nanoparticles impact on the cell membrane. *PloS one*. **2019**, *14* (11), e0224904-e0224904. DOI: 10.1371/journal.pone.0224904 PubMed.
127. Wan, G.; Ruan, L.; Yin, Y.; Yang, T.; Ge, M.; Cheng, X. Effects of silver nanoparticles in combination with antibiotics on the resistant bacteria *Acinetobacter baumannii*. *International journal of nanomedicine*. **2016**, *11*, 3789-3800. DOI: 10.2147/IJN.S104166 PubMed.

**CHAPTER 6. TOWARDS RATIONAL NANOMATERIAL DESIGN
BY PREDICTION OF DRUG-NANOPARTICLE SYSTEMS
INTERACTION VS. BACTERIA METABOLIC NETWORKS**

Appearance of Multidrug-Resistant (MDR) strains with perturbed Metabolic Networks (MN) push researcher to improve Antibacterial Drugs (AD). Some Nanoparticles (NP) may present antibacterial activity apart from acting as delivery systems. Then, developing Dual Antibacterial Drug-Nanoparticles (DADNP) systems becomes an option. However, testing DADNP *vs.* strains with different MN is a hard and costly task. Artificial Intelligence (AI) or Machine Learning (ML) could accelerate it by predicting bacteria sensibility. In this work, we used a Perturbation-Theory Machine Learning Information Fusion (IFPTML) analysis and mapping of DADNP (AD + NP) systems *vs.* MN of pathogenic bacteria species as a new application of AI/ML methods. Furthermore, most existing AI/ML models do not use as input vectors for \mathbf{c}_j of experimental conditions of assays (*i.e.*, bacteria specie, strain, NP shape, *etc.*). A working solution may be the use of an AI/ML method with an Information Fusion (IF) additive approach. Additive IF use as input the sets of vectors \mathbf{D}_{dk} , \mathbf{D}_{nk} , \mathbf{D}_{mk} and \mathbf{c}_{dk} , \mathbf{c}_{nk} , \mathbf{c}_{sk} with information about AD, NP, and MN structure and assays by separate. Accordingly, the IFPTML algorithm was selected to seek predictive models based on a ChEMBL dataset of >160000 AD assays enriched with 300 NP assays and >25 MN of different bacteria species. IFPTML use IF process to join the three datasets, PT Operators (PTOs) to codify \mathbf{D}_{dk} , \mathbf{D}_{nk} , \mathbf{D}_{sk} and \mathbf{c}_{dk} , \mathbf{c}_{nk} , \mathbf{c}_{sk} vectors information, and ML algorithms to train the model. IFPTML Linear Discriminant Analysis (LDA) model with $Sp \approx 90\%$ and $Sn \approx 80\%$ and best Artificial Neural Networks (ANN) model found with $Sp \approx Sn \approx 95\%$ in training/validation series presented good results. This kind of model could be useful for DADNP systems discovery. We also run a simulation >140000 points of putative DADNP systems *vs.* wild type and Knockout (KO) computationally-generated bacteria strains. The linear and additive IFPTML model was able to predict 102 experimental cases of complex DADNPs with a high degree of structural and biological variety. This led us to introduce the concept of MDR computational surveillance that could help to detect new strains of MDR bacteria.

1. INTRODUCTION

Eroom's law is an empirical observation stating that drug discovery processes are getting slower and costly every year consuming high amounts of time and resources with the subsequent environmental impact. In this sense, Eroom's law (Moore's law backwards) is the dark-side sister in chemistry of Moore's law of computational sciences. In fact, failure rates in drug clinical trials have scaled up to 90% after model organisms testing with costs rising to \$2.6 billion per drug. In this context, Artificial Intelligence (AI) and/or Machine Learning (ML) algorithms may help to speed up this process.¹ In fact, AI/ML methods have been used to solve different problems in pharmaceutical industry and nanotechnology.²⁻⁴ In the specific field of Antibacterial Drugs (AD) discovery the situation is not different. Multidrug-resistant (MDR) bacteria strains present mutations in specific gen due to environmental stressing conditions (antibiotic abuse, temperature variation, *etc.*) decreasing AD capability to halt bacterial infections.⁵ This in turn may trigger a domino effect promoting changes on their metabolism equivalent to rearrangements on the topological structure of their Metabolic Networks (MNs).⁶

7

Several authors have treated the human health and environmental risk assessment of engineered nanomaterials (ENM). For instance, Mikolajczyk, *et al.* developed nano-QSAR models for describing the cytotoxicity of 34 TiO₂-based NPs modified by (poly)metallic clusters (Au, Ag, Pt) to the Chinese hamster ovary cell line.⁸ On the other hand, Sizochenko *et al.* analyzed the genotoxicity of NPs (silicon and metal oxide).⁹ NPs could also affect community health, and how increased trophic complexity affects the interactions between organisms and nanomaterials. For instance, Wu *et al.* investigated the effect of exposures to AgNPs on simple microcosms (algae and bacteria) and increasingly complex microcosms containing predatory invertebrates and developing vertebrates.¹⁰ Thus, studies including multi-species scenarios could be more realistic and like natural conditions.

Nanosafety is the assessment of the dangers to human health and the environment posed by the use of ENMs, as well as their toxicity.¹¹ The rapid growth in the number of nanomaterials has raised concerns about possible toxic effects on human health and the environment.¹² The greatest concerns stem from contact with biological systems and those that are part of the constituents of medical devices, as well as pharmaceutical and cosmetic products.¹³ The results of AD-NP vs MN interactions can contribute to extending the scope of the faithful use of nanomaterials for contemporary pharmaceutical design based on nanobiotechnology to fight infections and other diseases.¹⁴ The safety of nanosystems has been recently treated by das Neves *et al.* They reviewed the key nanomaterial properties that govern the interplay between NP–mucosa interactions, and the importance of mathematical and computational models to characterize these interactions in nanomedicine and nanotoxicology.¹⁵

Rational Nanomaterial design could reduce cost and environmental impact. For instance, the design of DADNP systems is the understanding of the interaction of physical properties of NPs and ADs in different biological phenotypes that could allow the future rational design of these systems with desirable pharmacological properties. Analysis of property-distribution relationships is essential to inform the design of NPs with optimal pharmacokinetics and improved drug distribution.¹⁶ This picture points to the importance of MNs of pathogenic bacteria to research on new AD design.

In this context, we can use AI/ML models to simulate DADNP activity over bacteria with different MN for example. MNs, can be represented as graphs composed by sets of nodes (gene, protein, enzymes, metabolites, etc.) linked by arrows (metabolic reactions, transport process, signaling, *etc.*).¹⁷ Consequently, we can quantify their structure with numerical parameters useful as input of AI/ML algorithms. The public databases ChEMBL contains thousands of reports of preclinical assays of potential AD hits that it may be used to design new DADNP systems.^{18, 19} In addition, there is a growing number of experimental reports of NP with antibacterial action; see references in Supporting Information (SI) file SI00.doc. and the previous report of consensus MNs for multiple pathogen bacteria by Jeong *et al.*¹⁷ is a useful tool to understand the interrelationship of substrate-metabolism of diverse microorganisms. In another hand, the DADNP discovery with ML analysis is the very low number of experimentally cases of study useful to train the models. Gajewicz *et al.* analyzed this difficulty in a previous report.²⁰ Finally, this additive approach has the risk of neglecting the possibility of emergency of synergies among the subsystems.^{21, 22}

This work could contribute open an important window of opportunity for DADNP discovery. In addition, it could imply an important reduction of costs and time by using AI/ML and Networks analysis techniques. We propose a combination of the fundamentals of Information Fusion (IF), Perturbation Theory (PT), and Machine Learning (ML) methods to build an IFPTML (PT + ML + IF) model as a solution for this type of data. This model is especially suitable for databases with similar large data features and combinatorial information. IFPTML approach to the present allow us to treat the initial information about each sub-system (AD, NP, and MN) by separated. We can call this as the additive approach to the study of complex Bio-molecular Systems and/or Synthetic Biology systems. In this analysis we can see that the additive approach has some difficulty per se. One to be considered is the high complexity of the Data Analysis challenge posed by this problem. The system being analyzed (S_i) has three complex subsystems. These subsystems are: $S_1 = i^{\text{th}}$ AD hit, $S_2 = n^{\text{th}}$ NP system, and $S_3 = MN$ of the s^{th} bacteria specie. Each one of these subsystems can be quantitatively described using three sets of vectors \mathbf{D}_{dk} , \mathbf{D}_{nk} , and \mathbf{D}_{mk} . The elements of these vectors are descriptors of the chemical structure of each sub-system D_{dk} , D_{nk} , and D_{mk} . In addition, we can assign vectors of non-numeric labels or experimental conditions for each subsystem \mathbf{c}_{dj} , \mathbf{c}_{nj} , and \mathbf{c}_{sj} . Be aware that AD assays have multiple conditions \mathbf{c}_{dj} not necessarily identical to NP experiments \mathbf{c}_{nj} . Examples of these conditions are $c_{d0} = \text{AD assay outcome (MIC, etc.)}$, $c_{d1} = \text{AD assay bacteria species}$, $c_{d2} = \text{AD assays strains}$, $c_{n0} = \text{NP assay outcome (IC}_{50}, \text{etc.)}$, $c_{n1} = \text{bacteria species}$, $c_{n2} = \text{NP shape, etc.}$. The MN has also a vector of labels for the bacteria specie \mathbf{c}_{sj} with labels like $c_{s1} = \text{Pathogenicity}$, $c_{s2} = \text{Gram staining, etc.}$

Regarding the present problem, IFPTML have been applied to closed related problems. IFPTML has been used to predict NP alone or NP-Drug systems as well considering multiple conditions of assay.²¹⁻²⁵ Speck-Planche *et al.* developed IFPTML models of AD and NP antibacterial activity by separate, but never considered them together.^{24, 26, 27} Nocedo *et al.*, carried out an IFPTML mapping of AD activity *vs.* MN of multiple species, but not included NP as part of the equation.²⁸ In addition, Duardo *et al.*, Fernández-Riera *et al.*, and others reported IFPTML models of MN but not included AD or NP components.²⁹⁻³¹ Consequently, IFPTML have been used before to solve parts of the present problem. However, there are no reports of IFPTML models including the three components AD, NP, and MN of this problem at the same time. In view of that, this work reports the first IFPTML analysis and mapping of DADNP systems *vs.* MN of pathogenic bacteria species. The work has three main parts. Firstly, the paper develops the IFPMTL models for DADNP*vs.*MN mapping using alternative ML techniques. Next, the work shows the result of the IFPTML simulation of the effect of DADNP systems over real and computationally generated KO strains with changes on their MN topological structure. Then, the IFPTML model was used to predict the values of probability of several DADNP experimentally synthesized, biologically tested, and reported in the literature previously. At the end, IFPTML models were used to compared with other previous ML models that involving AD assays, NP antibacterial assays, and/or MN of bacteria.

2. MATERIALS AND METHODS

2.1 IFPTML analysis steps

. IFPTML analysis has different steps that can be separated into three phases (IF + PT + ML). The first step of IF phase is to obtain values v_{ij} , v_{nj} , and v_{sj} for the different biological properties c_{d0} , c_{n0} , and c_{s0} of the three subsystems AD, NP, and MN. We obtained them from datasets already published.^{27, 28} Next, we need preprocess all the observed values with different units, scales, degrees of uncertainty, etc. to obtain dimensionless functions characterizing the system as a whole DADNP vs. MN cases. The two main functions obtained in one IFPTML analysis are the objective function $f(v_{ij}, v_{nj}, v_{sj})_{obs}$ and the function of reference $f(v_{ij}, v_{nj}, v_{sj})_{ref}$. Next, we need to define and get/calculate the values of all vectors of structural descriptors \mathbf{D}_{dk} , \mathbf{D}_{nk} , and \mathbf{D}_{sk} for the three subsystems. We also need to define and get/calculate the elements of the vectors \mathbf{c}_{dj} , \mathbf{c}_{nj} , and \mathbf{c}_{sj} with all AD assay, NP assay, and MN bacteria labels/assay conditions. After that, we scaled all the elements of the vectors \mathbf{D}_{dk} , \mathbf{D}_{nk} , and \mathbf{D}_{sk} into Shannon's information measures $Sh(\mathbf{D}_{dk})$, $Sh(\mathbf{D}_{nk})$, and $Sh(\mathbf{D}_{sk})$. At this point, we enter into the PT phase of the analysis. In PT phase we zip all structural/labeling information into PTOs of each subsystem: $\Delta Sh(\mathbf{D}_{dk})_{cdj}$, $\Delta Sh(\mathbf{D}_{nk})_{cnj}$, and $\Delta Sh(\mathbf{D}_{sk})_{csj}$. Last, we enter the ML phase. In the first step of ML phase one can proceed to training/validating alternative IFPTML models using different ML techniques. In last step we use the selected IFPTML models to run simulations and carry out predictions. The following paragraphs offer a more detailed explanation of these phases. In **Figure 6.1** we show IFPTML algorithm workflow for DADNP vs. MN mapping and analysis, including the general steps given on the present work.

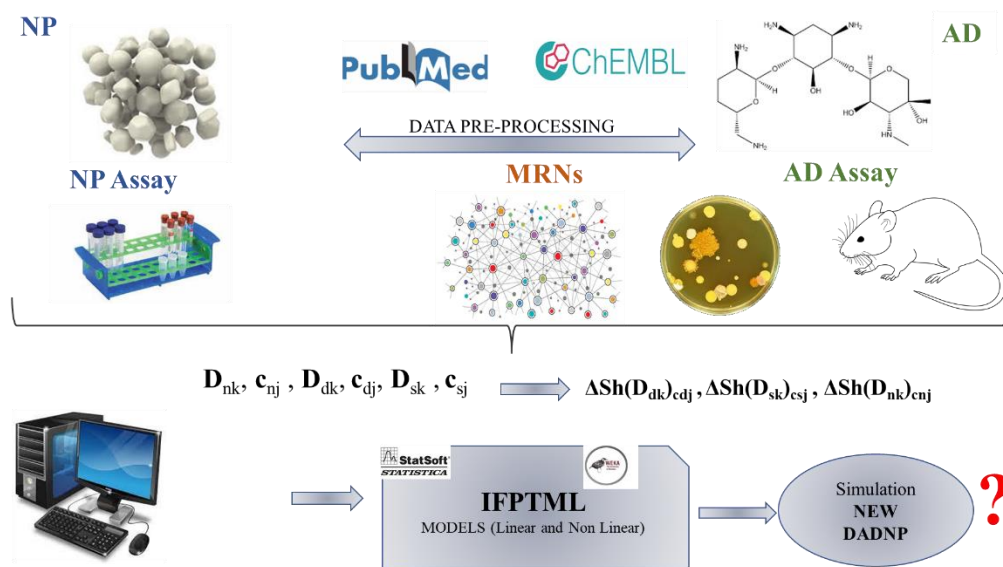


Figure 6.1. IFPTML workflow for DADNP vs. MNs mapping.

IF-Step 1, ChEMBL AD, NP, and MN datasets. The ChEMBL AD activity dataset contains the values of >300 parameters (MIC, IC₅₀, etc.) for >160 000 biological assays of >50000 compounds vs. >25 bacteria species with >90 strains. The NP dataset includes 1 out of 4 parameters of activity for 300 pre-clinical assays of NP vs. 34 bacteria species/strains (s).²⁷ The MN dataset released by Jeong *et al.*¹⁷ have the MN of >20 bacteria species. We encoded information about the structure of AD compounds, NP cores and coat, and MN structure into

vectors $\mathbf{D}_{dk} = [D_{d1}, D_{d2}, D_{d3}, \dots, D_{dmax}]$, $\mathbf{D}_{nk} = [D_{n1}, D_{n2}, D_{n3}, \dots, D_{nmax}]$, and $\mathbf{D}_{sk} = [D_{s1}, D_{s2}, D_{s3}, \dots, D_{smax}]$, respectively. We used the vectors $\mathbf{c}_{dj} = [c_{d0}, c_{d1}, c_{d2}, \dots, c_{dmax}]$, $\mathbf{c}_{nj} = [c_{n0}, c_{n1}, c_{n2}, \dots, c_{nmax}]$, and $\mathbf{c}_{mj} = [c_{s0}, c_{s1}, c_{s2}, \dots, c_{smax}]$ to codify information about the AD, NP, and MN conditions of assay or labels. See detailed information about the three datasets on File SI00.doc.²⁷

IF-Step 2, objective and reference functions. Firstly, we re-scaled the experimental values v_{ij} and v_{nj} of biological activity of AD and NP to obtain binary functions $f(v_{ij})_{obs}$ and $f(v_{nj})_{obs}$. The values $f(v_{ij})_{obs} = 1$ and $f(v_{nj})_{obs} = 1$ points to an strong desired effect of both the AD and the NP over the target bacteria.¹⁰ Otherwise, $f(v_{ij}) = f(v_{nj})_{obs} = 0$. By definition, the biological activity of the s^{th} bacteria specie with MNs is also a binary function $f(v_{nj})_{obs}$. The function $f(v_{nj})_{obs} = 1$ when the bacteria is pathogenic or $f(v_{nj})_{obs} = 0$ otherwise. The IF additive approach presupposes the best option is a system made by the best subsystems. The IF also tries to ensure the higher coherence possible among the information about the subsystems. Consequently, the objective function $f(v_{ij}, v_{nj}, v_{sj})_{obs}$ was defined as follows, see **Figure 6.2**.

$$f(v_{ij}, v_{nj}, v_{sj})_{obs} = f(v_{ij})_{obs} \cdot f(v_{nj})_{obs} \cdot f(v_{sj})_{obs} \cdot f(v_{jsn})_{obs} \quad (1)$$

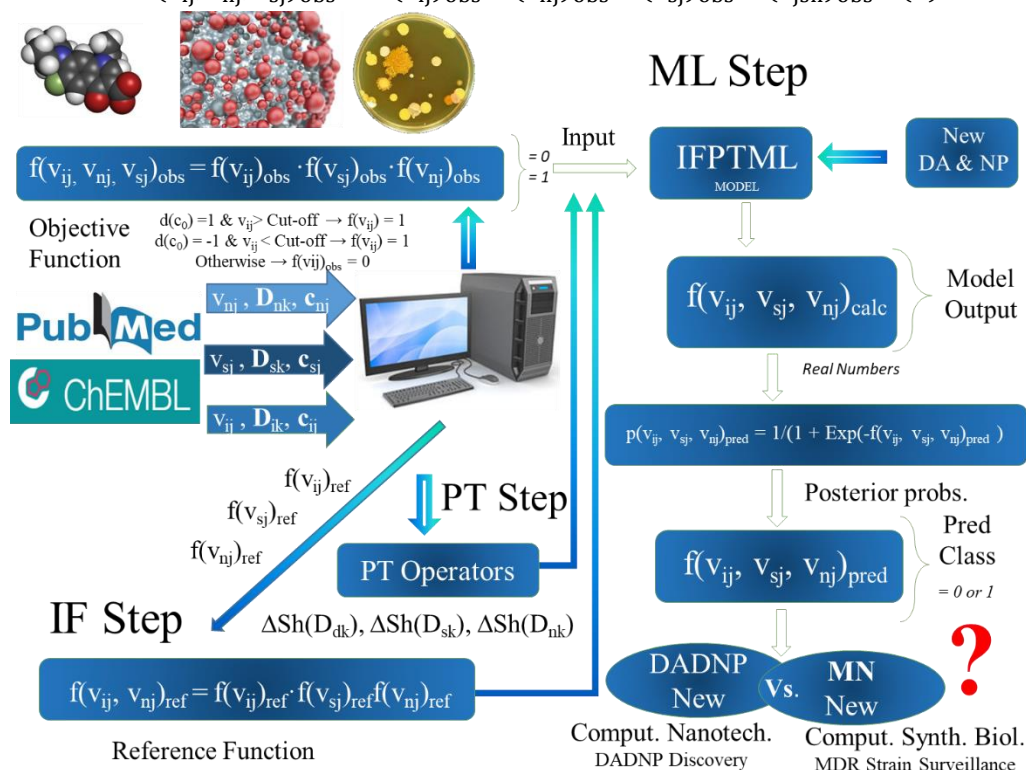


Figure 6.2. IFPTML information processing detailed workflow.

This function $f(v_{ij}, v_{nj}, v_{sj})_{obs} = 1$ when all the subsystems have the desired levels of their own properties. It means that, $f(v_{ij}, v_{nj}, v_{sj})_{obs} = 1$ if all AD, NP, and MN has individually the desired level of the biological property ($f(v_{ij})_{obs} = 1$, $f(v_{nj})_{obs} = 1$, and $f(v_{sj})_{obs} = 1$). In order to decide whether AD has the desired level of biological property or not we used the following expression. When $(v_{ij} > \text{cutoff}_j \ \& \ d_j(c_0) = 1)$ OR $(v_{ij} < \text{cutoff}_j \ \& \ d_j(c_0) = -1)$ then $f(v_{ij}) = 1$, else $f(v_{ij}) = 0$. The parameters cutoff_j and cutoff_{nj} are the threshold values (cutoff)

used to decide if the AD or NP biological effects are strong or weak. The values of cutoff depend on the nature of the biological parameter c_0 measured under conditions c_j for the AD or the parameter c_{n0} under conditions c_{nj} for the NP. The desirability functions $d_j(c_0)$ and $d_{nj}(c_0)$ indicate if we desire to maximize ($d_j(c_0) = 1$ or $d_{nj}(c_0) = 1$) or minimize ($d_j(c_0) = -1$ or $d_{nj}(c_0) = -1$) the values v_{ij} or v_{nj} of these biological parameters c_{j0} and c_{n0} . In the case of $f(v_{sj}) = 1$ if the s^{th} bacteria specie is pathogenic for humans or $f(v_{sj}) = 0$ otherwise. Please, see the steps in **Figure 6.2** and values of cutoff, desirability, *etc.* in the file SI01.xlsx. The same rule applies to NP. The following equations define more clearly these conditional functions.

$$f(v_{ij})_{\text{obs}} = \begin{cases} = 1 \text{ IF } v_{ij} > \text{cutoff}_j \text{ AND } d(c_0) = 1 \text{ OR} \\ = 1 \text{ IF } v_{ij} < \text{cutoff}_j \text{ AND } d(c_0) = -1 \text{ OR} \\ = 0 & \text{Otherwise} \end{cases} \quad (2)$$

$$f(v_{nj})_{\text{obs}} = \begin{cases} = 1 \text{ IF } v_{nj} > \text{cutoff}_{nj} \text{ AND } d(c_{n0}) = 1 \text{ OR} \\ = 1 \text{ IF } v_{nj} < \text{cutoff}_{nj} \text{ AND } d(c_{n0}) = -1 \text{ OR} \\ = 0 & \text{Otherwise} \end{cases} \quad (3)$$

After obtaining the objective function we should define the input variables of the IFPTML model. The first input variable is the function of reference $f(v_{ij}, v_{nj}, v_{sj})_{\text{ref}}$. In this work, the function of reference is the probability $f(v_{ij}, v_{nj}, v_{sj})_{\text{ref}} = p(f(v_{ij}, v_{nj}, v_{sj})_{\text{ref}} = 1)$. This is the probability with which the systems of reference have a desired level of activity $f(v_{ij}, v_{nj}, v_{sj})_{\text{ref}} = 1$ for the same parameters of activity. As this is an additive approach, we obtained this probability as the product of each individual probability with which all the subsystems have a desired level of activity by separate. Each individual probability $p(f(v_{ij})_{\text{ref}} = 1) = n(f(v_{ij})_{\text{obs}} = 1)/n_{c0}$ and $p(f(v_{nj})_{\text{ref}} = 1) = n(f(v_{nj})_{\text{obs}} = 1)/n_{c0}$, are equal to the number of positive subsystems $n(f(v_{ij})_{\text{obs}} = 1)$ or $n(f(v_{nj})_{\text{obs}} = 1)$ between the total number of subsystems n_{c0} or n_{c0} with the same biological parameter. The term $p(f(v_{sj})_{\text{ref}} = 1) = 1$ by definition for all pathogenic bacteria.

$$f(v_{ij}, v_{nj}, v_{sj})_{\text{ref}} = f(v_{ij})_{\text{ref}} \cdot f(v_{nj})_{\text{ref}} \cdot f(v_{sj})_{\text{ref}} \quad (4)$$

$$f(v_{ij}, v_{nj}, v_{sj})_{\text{ref}} = p(f(v_{ij})_{\text{ref}} = 1) \cdot p(f(v_{nj})_{\text{ref}} = 1) \cdot p(f(v_{sj})_{\text{ref}} = 1) \quad (5)$$

IF-Step 3, Shannon's information scaling of input variables. This IFPTML model considers that the system under study (S) is composed of various subsystems ($S = S_d + S_n + S_s$) with $S_d = \text{AD}$, $S_n = \text{NP}$, $S_s = \text{MN}$. The structure of each subsystem is encoded with the vectors of molecular/structural descriptors \mathbf{D}_{dk} , \mathbf{D}_{nk} , and \mathbf{D}_{sk} , respectively. The vectors of the subsystem S_d have the elements $\mathbf{D}_{dk} = [D_{d1}, D_{d2}, D_{d3}, D_{d4}]$. These elements are the descriptors the i^{th} AD. Last, the vectors of the subsystem S_s have the elements: $\mathbf{D}_{sk} = [D_{s1}, D_{s2}, D_{s3}]$. All the D_{dk} , D_{nk} , and D_{sk} values have different scales/units. Consequently, we used Shannon's entropy information measure to quantify all the information in the same scale.³² Please, see the values of $\text{Sh}(D_{nk})$ and $\text{Sh}(D_{sk})$ for different NP and MN in **Table S1** and **Table S2** of file SI00.doc. The same transformation was used for all D_{dk} , D_{nk} , and D_{sk} variables.

$$p(D_k) = \frac{1}{(1 + \text{Exp}(-D_k/1000))} \quad (6)$$

$$\text{Sh}(D_k) = -p(D_k) \cdot \log(p(D_k)) \quad (7)$$

IF-Step 4, data fusion vs. subset sampling. We assigned all cases to either training (subset = t) or validation (subset = v) series. Sampling is desired to random, representative, and stratified, as much as possible.³³ Additionally, in this work sampling should take into consideration the AD, NP, and MN, IF process. We selected the original data from the three datasets randomly to create triads. These triads are formed by one AD, one NP, and one MN cases (representing putative DADNP vs. MN interactions). However, we need to impose some constrains in some labels due to the IF process. The cases forming one triad have the same value of the labels c_{0d} and c_{0n} (same biological property) of AD and NP whenever it was possible. The cases of the triads have also the same c_{d1} , c_{n1} , and c_{s1} (bacteria specie) whenever it was possible. All triads have been ordered according to these main labels (stratified sampling). Subsequently, cases were assigned to set = t and set = v (representative sampling) in a proportion 75% vs. 25%.³³ See details in File SI00.doc

PT-Step 1, PTO additive calculation. As we mentioned before in the PT phase, we zip all structural/labeling information into PTOs of each subsystem, see **Figure 2**. The additive PTOs calculated in this work are: $\Delta\text{Sh}(D_{dk})$, $\Delta\text{Sh}(D_{nk})$, and $\Delta\text{Sh}(D_{sk})$. The PTOs of type $\Delta\text{Sh}(D_{dk})$ and $\Delta\text{Sh}(D_{nk})$ codify chemical structure and/or physicochemical properties of AD and NP subsystems. The PTOs of type $\Delta\text{Sh}(D_{sk})$ quantify structural information of the MN of the bacteria. We calculate the PTOs as the deviation of the information of the subsystems $\text{Sh}(D_{dk})$, $\text{Sh}(D_{nk})$, and $\text{Sh}(D_{sk})$ with respect to the average value for the respective subsystems of reference $\langle\text{Sh}(D_{dk})_{cdj}\rangle$, $\langle\text{Sh}(D_{nk})_{cnj}\rangle$, and $\langle\text{Sh}(D_{sk})_{csj}\rangle$. The average is calculated for all cases with the same vectors of labels/conditions \mathbf{c}_{dj} , \mathbf{c}_{nj} , \mathbf{c}_{sj} , respectively. Consequently, in these expressions the first terms $\text{Sh}(D_{dk})$, $\text{Sh}(D_{nk})$, and $\text{Sh}(D_{sk})$ identify the subsystem and the averages identify the assay. Please, see values in SI00.doc. The equations used are the following.

$$\Delta\text{Sh}(D_{dk}) = \text{Sh}(D_{dk}) - \langle\text{Sh}(D_{dk})\rangle_{\mathbf{c}_{dj}} \quad (8)$$

$$\Delta\text{Sh}(D_{nk}) = \text{Sh}(D_{nk}) - \langle\text{Sh}(D_{nk})\rangle_{\mathbf{c}_{nj}} \quad (9)$$

$$\Delta\text{Sh}(D_{sk}) = \text{Sh}(D_{sk}) - \langle\text{Sh}(D_{sk})\rangle_{\mathbf{c}_{sj}} \quad (10)$$

PT-Step 2, PTO cross-over calculation. The previous PTOs have been calculated previous to the IF process. This is because they codify information for one specific subsystem. We also calculated PTOs posterior to the IF process. We call them cross-over PTOs because they quantify information of two or more subsystems at time. The cross-over PTOs calculated here quantify the difference of the information about the AD from the information about the NP coating system and their particular assay conditions. In these operators $\Delta\text{Sh}(D_{dk})$ represent the AD and the drug assay conditions. The operators of the type $\Delta\text{Sh}(D_{ca1k})$ and $\Delta\text{Sh}(D_{ca2k})$ represent the first (c_{a1}) and second (c_{a2}) coating agents of the NP and the NP assay conditions \mathbf{c}_{nj} . They have the following formula.

$$\Delta\text{Sh}(D_{dk}, D_{nk}) = \Delta\text{Sh}(D_{dk}) + [\Delta\text{Sh}(D_{ca1k}) + \Delta\text{Sh}(D_{ca2k})] \quad (11)$$

ML-Step 1, additive cross-over linear model. IFPTML DADNP proposed here is, in first instance, a linear model. The output is the scoring function $f(v_{ij}, v_{nj}, v_{sj})_{\text{calc}}$ used to calculate the posterior probabilities which the DADNP is short listed for experimental biological assay, see **Figure 2**. We can obtain the model by using a ML to fit the objective function $f(v_{ij}, v_{nj}, v_{sj})_{\text{obs}}$. After, we obtain in first instance a linear IFPTML model. This model uses as input the function of reference $f(v_{ij}, v_{nj}, v_{sj})_{\text{ref}}$ and the PTOs of type $\Delta\text{Sh}(D_{dk})$, $\Delta\text{Sh}(D_{nk})$, $\Delta\text{Sh}(D_{sk})$, and $\Delta\Delta\text{Sh}(D_{nk})$,

D_{dk}). The PTOs $\Delta\text{Sh}(D_{dk})$, $\Delta\text{Sh}(D_{nk})$, and $\Delta\text{Sh}(D_{sk})$ quantify information about the AD, NP, MN and the biological assays of the triad. The PTO $\Delta\Delta\text{Sh}(D_{1c}, D_{2c}, D_{dk})$ is a cross-over operator because it quantifies information about the AD and the NP coating agents. The operator also quantifies information about the AD and NP assays at the same time. This operator is another expression of the IF process. Firstly, we ran a Linear Discriminant Analysis (LDA) technique as a first approach. Forward Step-Wise (FSW) feature selection strategy allowed automatic selection of input variables. Next, Expert-Guided Selection (EGS) was used incorporated missing features. Parameters like Sensitivity (Sn), Specificity (Sp), Chi-square (χ^2), and the p -level were used to check model quality. STATISTICA 6.0 software was used to ran all the algorithms.³³ The general form of the IFPTML-LDA linear models proposed is the following:

$$\begin{aligned}
 f(v_{ij}, v_{nj}, v_{sj})_{calc} &= a_0 + a_1 \cdot f(v_{ij}, v_{nj}, v_{sj})_{ref} \\
 &+ \sum_{k=1, j=1}^{k=kmax, j=jmax} a_{k,j} \cdot \Delta\text{Sh}(D_{ki})_{cdj} \quad (12) \\
 &+ \sum_{k=1, j=1}^{k=kmax, j=jmax} a_{k,j} \cdot \Delta\text{Sh}(D_{kn})_{cnj} + \sum_{k=1, j=1}^{k=kmax, j=jmax} a_{k,j} \cdot \Delta\text{Sh}(D_{sn})_{csj} \\
 &+ \sum_{k=1, j=1}^{k=kmax, j=jmax} a_{k,j} \cdot \Delta\Delta\text{Sh}(D_{ki}, D_{kn})_{cdj, cnj}
 \end{aligned}$$

ML-Step 2, non-linear models. We also trained/validated various models using Artificial Neural Networks (ANN). Different ANN topologies/techniques were tested including Linear Neural Networks (LNN), Multi-Layer Perceptrons (MLP), and Radial Basis Functions (RBF). The training algorithms used to optimize the IFPTML-ANN models were Back Propagation (BP), Conjugated Gradient (CG), K-Means Center Assignment (KM), K-Nearest Neighbor Deviation Assignment (KN), and Pseudo-Invert Linear Least Squares Optimization (PI). Specifically, we used BP100 and CG20b for MLP, PI for LNN, and KM, KN, and PI for RBF. STATISTICA 6.0 software was used to ran all the algorithms.³³ Area Under Receiver Operating Characteristic (AUROC) curve together with Sn, Sp, Matthew's correlation coefficient (MCC),³⁴(**Eq13**), F1 score (**Eq14**), χ^2 , and the p -level were used to check model quality. In addition, we applied the estimation metrics of the random correlation model of classification proposed by Lucic et al.^{35, 36}. These parameters for the validation of the classification model quality difference are based on the Ac Real (Q_2) and the corresponding random model ($Q_{2, md}$). See **Equation 15-18**. In addition, to test the robustness of the IFPTML-LDA model, the Y-randomization test was performed.^{37, 38} The training set function was randomized at 5%, 10%, 15%, 20%, 25%, 30%, 35%, and 40% of the total training set (active and inactive).^{39, 40}

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (13)$$

$$F_1 score = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (14)$$

$$Q_2 = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

$$Q_{2,rand} = \frac{(TP + FP) \cdot (TP + FN) + (TN + FN) \cdot (FN + FP)}{(TP + TN + FP + FN)^2} \quad (16)$$

$$\Delta Q_2 = 100 \cdot (Q_2 - Q_{2,rand}) \quad (\%) \quad (17)$$

$$Q_{2,rand-bal} = \frac{(TP + FP)^2 + (TN + FN)^2}{(TP + TN + FP + FN)^2} \quad (18)$$

where: MCC: Matthew's correlation coefficient, TP: true positive, TN: true negative, FP: false positive, and FN: false negative, Q_2 : Real accuracy, $Q_{2,rand}$: Random accuracy, ΔQ_2 : difference between the real model accuracy and the corresponding random accuracy, and $Q_{2,rand-bal}$: Most probable random accuracy for balanced model.

Generating reliable predictions requires knowledge of model limitations and applicability. The Domain of Applicability (DoA) can be defined using similarity measures based on Euclidean distances between all training and test composites or by using leverage method.^{41, 42} We apply the leverage method. In this approach, after the calculation of the hat matrix for the structural domain, the residuals and LOO residuals of the response variables were mapped against the leverages (the diagonal values of the hat matrix (h)) to visually characterize the DoA (Williams plot).⁴³ Chemicals exceeding certain threshold values were identified as response and leverage outliers. Response threshold values were set at ± 2 residuals and LOO residuals. The leverage threshold was set to the critical hat value ($h^* = 3(p+1)/n$, where p is the number of model descriptors and n is the number of training compounds.⁴³ According to Gramatica⁴⁴, ($h > h^*$) was considered a structurally influential chemical.

ML-Step 3, IFPTML mapping of DADNP vs. MN of strains. In this work we also used the IFPTML model created to study the possible susceptibility of new mutant bacteria to putative AD systems. In so doing, we simulated the new strains as mutants with changes (perturbations) on the parameters of their MN with respect to the existing wild type species. In particular, we focused on the parameters $L_{in}(\pm\delta_{in})$ or $L_{out}(\pm\delta_{out})$. These are the average values of in-degree and out-degree for all the nodes (metabolites) on the MN of the new strain. The parameters of the new mutant strain were calculated as follows:

$$\begin{aligned} L_{in}(\pm\delta_{in}) &= L_{in}(0) \pm \delta_{in} \cdot L_{in}(0) \quad (19) \\ L_{out}(\pm\delta_{out}) &= L_{out}(0) \pm \delta_{out} \cdot L_{out}(0) \quad (20) \end{aligned}$$

These equations indicate that the parameters $L_{in}(\pm\delta_{in}\%)$ and $L_{out}(\pm\delta_{out}\%)$ of the new MN are equal to the original values $L_{in}(0)$ and $L_{out}(0) \pm$ a certain fraction δ_{in} or δ_{out} of the same original values. These fractions can be interpreted as the relative change with respect to the original value, by rearranging the previous equations. We can use a 100-scaling factor if we want to express the change in (%).

$$\delta_{in}(\%) = 100 \cdot \frac{[L_{in}(\pm\delta_{in}) - L_{in}(0)]}{L_{in}(0\%)} = 100 \cdot \frac{\Delta L_{in}(\pm\delta_{in})}{L_{in}(0\%)} \quad (21)$$

$$\delta_{out}(\%) = 100 \cdot \frac{[L_{out}(\pm\delta_{out}) - L_{out}(0)]}{L_{out}(0)} = 100 \cdot \frac{\Delta L_{out}(\pm\delta_{out})}{L_{out}(0)} \quad (22)$$

For instance, let be a wild type of mutant with average $L_{in}(0) = 10$, we can obtain a new mutant with $L_{in}(\pm\delta_{in}) = L_{in}(0) \pm \delta_{in} \cdot L_{in}(0) = 10 + 1 \cdot 10 = 20$. It means that the new

mutant has a relative increment of $\delta_{in}(\%) = 100 \cdot (20 - 10)/10 = 100\%$ in L_{in} with respect to the wild type. We have taken advantage of these equations to create a new code in order to easily identify all mutants in this work. From now on, we are going to label all mutants and wild type bacteria with the following code $BB(\pm\delta_{in}, \pm\delta_{out})$. In this code, BB is the two-letter code of the different bacteria species. As we mentioned before $\pm\delta_{in}$ and $\pm\delta_{out}$ are the relative changes of the mutant strain with respect to the original value. Then, in this notation, δ_{in} and δ_{out} indicate of magnitude of the changes (perturbation) and (\pm) indicates the sign of the change with respect to the wild type. For instance, $EC(-0.1, +0.22)$ is the code for a mutant strain of *Escherichia coli* (EC) with $\delta_{in} = -0.1$ and $\delta_{out} = +0.2$. This represents a decrease of $\delta_{in}(\%) = -10\%$ and one increase of $\delta_{out}\% = +20\%$ with respect to wild type EC . In turns, in could be interpreted as a decrease of $\delta_{in}(\%) = -10\%$ in Anabolism and one increase of $\delta_{out}\% = +20\%$ in Catabolism with respect to wild type EC . Accordingly, the wild type EC is denoted by $EC(0, 0)$ and has $\delta_{in} = 0$ and $\delta_{in} = 0$. In general, we can scale the parameters $D_{ks}(\pm\delta_k) = L_{in}(\pm\delta_{in})$, $L_{out}(\pm\delta_{out})$, etc. of the new MN using the same Shannon's entropy procedure. See the equation of probability and entropy of the new MN as a function of the original parameter or wild type MN. The new parameters $Sh(D_{ks}(\pm\delta))$ can be substituted into the equation of one IFPTML model. This should allow us calculating the values of $f(v_{ij}, v_{nj}, v_{sj})_{calc}$ and subsequently the values of probability $p(f(v_{ij}, v_{nj}) = 1)_{calc}$ of activity for the AD systems with respect to the new mutant strains.

$$p(D_k(\pm\delta_k)) = \frac{1}{(1 + \text{Exp}(-D_k(\pm\delta_k)/1000))} \quad (23)$$

$$p(D_k(\pm\delta_k)) = \frac{1}{(1 + \text{Exp}(-(D_{ks}(0) \pm \delta_k \cdot D_{ks}(0))/1000))}$$

$$\text{Sh}(D_k(\pm\delta_k)) = -p(D_k(\pm\delta_k)) \cdot \log(D_k(\pm\delta_k)) \quad (24)$$

3. RESULTS AND DISCUSSION

3.1 IFPTML DADNP additive linear model.

AI/ML algorithms are gaining momentum with multiple applications in Nanotechnology.⁴⁵⁻⁵⁰ However, there is still models fail to account for complex problems, that involve multiple subsystems at the same time. The present work introduces the first IFPTML models able to map putative DADNP (AD + NP) systems vs. MN of bacteria. The best IFPTML-LDA linear model found here was the following.

$$\begin{aligned} f(v_{ij}, v_{nj}, v_{sj})_{calc} &= 89.776 + 5.600 \cdot f(v_{ij}, v_{nj}, v_{sj})_{ref} - 85.466 \cdot \Delta\text{Sh}(\text{LOGP}_i)_{cdj} \quad (25) \\ &- 1126.814 \cdot \Delta\text{Sh}(L_{ins})_{csj} + 1074.061 \cdot \Delta\text{Sh}(L_{outs})_{cnj} + 527.286 \cdot \Delta\text{Sh}(AMV_n)_{cnj} \\ &+ 131.830 \cdot \Delta\text{Sh}(APS_n)_{cnj} + 1225.895 \cdot \Delta\text{Sh}(t)_{cnj} \\ &+ 25.973 \cdot \Delta\Delta\text{Sh}(PSA_{ca1}, PSA_{ca2}, PSA_i)_{cdj, cnj} \\ N_{train} &= 124366 \quad \chi^2 = 41396.82 \quad p\text{-level} < 0.05 \quad \lambda\text{-Wilks} = 0.7168 \end{aligned}$$

The performance of this IFPTML model was assessed with S_n , S_p , χ^2 , and the p -level.³³ The value of the p -level < 0.05 for the χ^2 test indicates that the model is able to separate both classes

significantly. In fact, Sn, Sp, and Ac are in the range ≈ 80 -90% for both training and validation series, see **Table 6.1**. These values are very good for this kind of models, taking into consideration the high complexity of the data analyzed.⁵¹ This model includes all the important variables AD structure and assay conditions, NP properties, CA structure, NP assay conditions, MN structural parameters, *etc.* The output $f(v_{ij}, v_{nj}, v_{sj})_{\text{calc}}$ variable is real but dimensionless numeric parameter without upper or bottom boundaries. This difficult its use for comparison with observed values or outputs of alternative models. Consequently, using a sigmoid function one can transform this output into a probability function. The resulting function is the probability $p(f(v_{ij}, v_{nj}, v_{sj})=1)_{\text{calc}}$ given the prior probabilities $\pi_0 = 1 - \pi_1$ of both groups. This is the probability with which the DADNP introduced in the IFPTML model should be short listed for experimental assay *vs.* the bacteria with the MN used. The selected DADNP are those with $p(f(v_{ij}, v_{nj}, v_{sj})=1)_{\text{calc}} > 0.5$ which implies that $f(v_{ij}, v_{nj}, v_{sj})_{\text{pred}} = 1$, see **Figure 6.2**.³³ The file SI01.xls contains the values of $f(v_{ij}, v_{nj}, v_{sj})_{\text{obs}}$, $f(v_{ij}, v_{nj}, v_{sj})_{\text{pred}}$, and $p(f(v_{ij}, v_{nj}, v_{sj})=1)_{\text{calc}}$ for >160000 DADNP assays studied in training and validation.

$$p(f(v_{ij}, v_{nj}, v_{sj}) = 1) = \frac{1}{1 + \left(\frac{\pi_0}{\pi_1}\right) \cdot \text{Exp}\left(-f(v_{ij}, v_{nj}, v_{sj})_{\text{calc}}\right)} \quad (26)$$

Table 6.1. IFPTML DADNP *vs.* MN model results summary.

Data Set	Stat.	(%)	Classification	$f(v_{ij}, v_{nj}, v_{ij})_{\text{pred}}$	
			$f(v_{ij}, v_{nj}, v_{ij})_{\text{obs}}$	0	1
Train	Sp	90.5	0	104771	11045
	Sn	80.2	1	1617	6933
	Ac	89.8	Total		
Validation	Sp	90.7	0	35045	3604
	Sn	81.7	1	513	2283
	Ac	90.1	Total		

We confirmed our proposition to model DADNP activity *vs.* bacteria with known MN using ML techniques. IFPTML is a multi-output and input-coded multi-label ML technique developed precisely to target this kind of problems.^{52, 53} IFPTML use IF process to join the three datasets, PT Operators (PTOs) to codify \mathbf{D}_{dk} , \mathbf{D}_{nk} , \mathbf{D}_{mk} and \mathbf{c}_{dk} , \mathbf{c}_{nk} , \mathbf{c}_{mk} vectors information, and ML algorithms to train the model. PTOs are functions of Moving Averages (MA) denoted by $\Delta V(\mathbf{D}_{\text{dk}})_{\text{ej}}$ previously demonstrated to be useful in the study of NP systems.^{22, 25, 54} IFPTML has been employed to solve many multi-output and multi-label problems in Medicinal Chemistry, Epidemiology, Proteomics, Systems Biology, *etc.* These problems have different combinations of drugs, drug cocktails, proteins, vaccines, MN, PINs, US Epidemiological networks, *etc.*^{30, 55-58}

3.2 IFPTML-ANN linear *vs.* non-linear models.

In first instance, we used ANN models to test the strength of our linear hypothesis. Interestingly, the Linear Neural Networks (LNN) presented high values of $\text{Sn} \approx \text{Sp} \approx 87\%$ and $\text{AUROC} \approx 0.94$. This seems to confirm that the hypothesis of a linear relationship among the

PTOs and the output is not a random or chance finding, see **Table 6.2** and **Figure 6.3**. However, there is still some margin to improve the performance of the linear models found. As a result, we used non-linear ANN algorithms to improve the performance of our IFPTML linear models. The IFPTML-RBF model with $S_n \approx S_p \approx 75\%$ showed a lower performance than the IFPTML-LDA and IFPTML-LNN linear models. However, the IFPTML-MLP non-linear models reach significantly higher levels $S_n \approx S_p \approx 93-95\%$, see **Table 6.2**. These models are the better alternative we found to IFPTML-LDA model. The model also presented very high AUROC ≈ 0.97 . It indicates that our IFPTML-ANN models differ significantly from a random (RND) classifier AUROC = 0.5,³³ see **Figure 6.3**. These values are excellent for this kind of models according to literature.³³ The relationship among the ANN topology and model performance was another question to answer. We found no significant improvement by increasing the number of hidden layer topologies having 7-10 neurons.

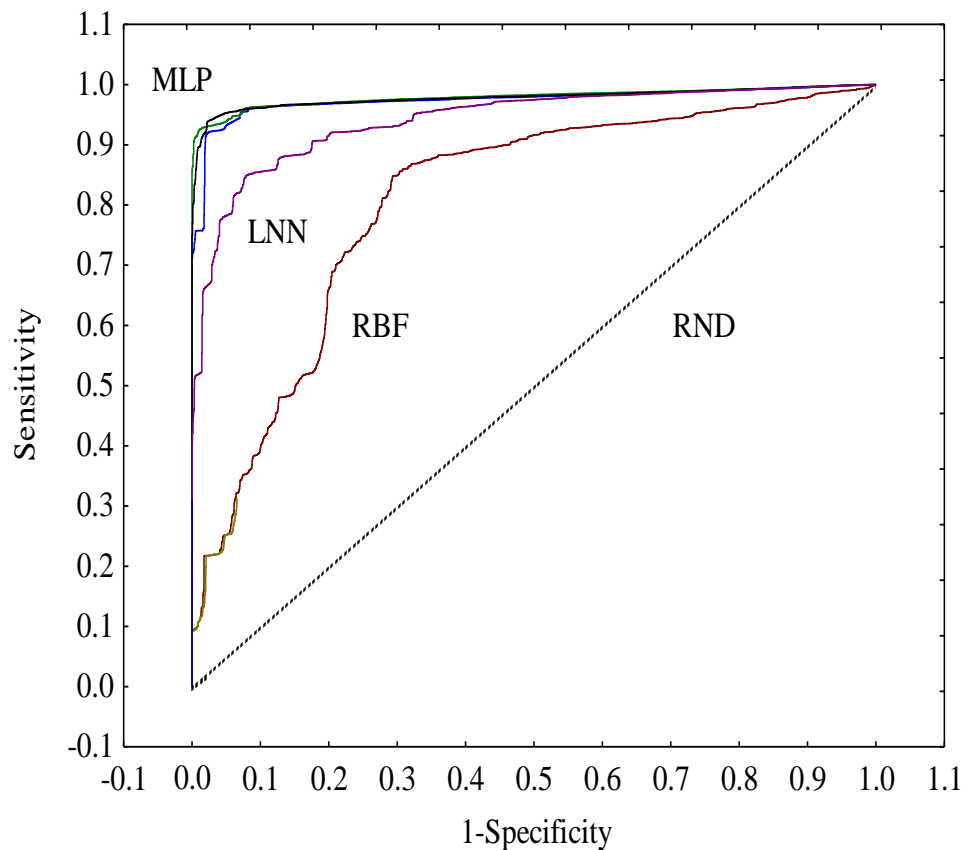
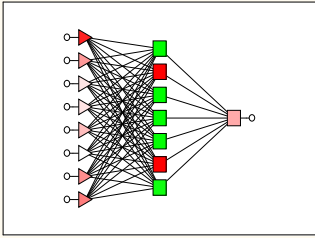
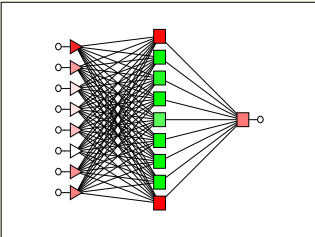
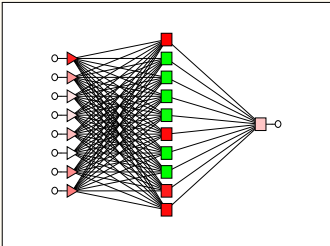
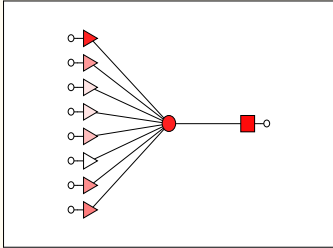
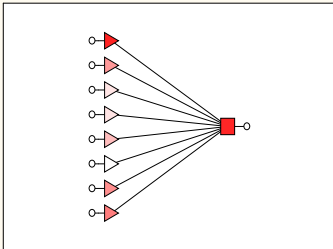


Figure 6.3. IFPTML ROC curve analysis.

Table 6.2. IFPTML-ANN DADNP vs. MN models.

IFPTML-ANN Model	Data Set	Stat. Par.	$f(i,j,s,n,c)$ (%)	Pred. Obs.	Pred.		AUROC
					1	0	
MLP 8:8-7-1:1 	t	Sn	93.6	1	7999	551	0.973
		Sp	93.6	0	7448	108368	
	v	Sn	94	1	2628	168	0.974
		Sp	93.9	0	2373	36276	
MLP 8:8-9-1:1 	t	Sn	94.1	1	8047	503	0.978
		Sp	94.2	0	6751	109065	
	v	Sn	94.7	1	2649	147	0.979
		Sp	94.2	0	2233	36416	
MLP 8:8-10-1:1 	t	Sn	94.5	1	8083	467	0.976
		Sp	94.9	0	5930	109886	
	v	Sn	95.5	1	2670	126	0.978
		Sp	95.1	0	1913	36736	
RBF 8:8-1-1:1 	t	Sn	75.1	1	6421	2129	0.804
		Sp	74.5	0	29539	86277	
	v	Sn	74.8	1	2092	704	0.806
		Sp	74.9	0	9714	28935	
LNN8:8-1:1 	t	Sn	86.7	1	7414	1136	0.939
		Sp	87.4	0	14582	101234	
	v	Sn	87.3	1	2441	355	0.941
		Sp	87.6	0	4775	33874	

3.3 IFPTML-WEKA AI/ML models.

Next, we decided to run several non-linear ML algorithms developed using the Waikato Environment for Knowledge Analysis (WEKA) software package, version 3.8.5.⁵⁹ We employed twelve ML algorithms in total to create these different non-linear IFPTML classification models using the current dataset. These included decision tree classifiers, neural networks, Bayesian networks, boosting algorithms and deep learning. Each technique adopts a learning algorithm to identify the model that best fits the relationship between the input data set and the class. Two classification algorithms based on Bayes' theorem were applied, Bayesian Network (BN), and Naïve Bayes (NBN). The boosting algorithms used were Adaboost, LogitBoost, and MultiBoosting, which are three representative algorithms of this family of algorithms.⁶⁰ They are well-known methods to build ensembles of classifiers with very good performance in medicinal chemistry.⁶¹ In the case of Deep Learning (DL), this technique was implemented with WekaDeeplearning4j (deep learning package for the Weka workbench). The DL Model was built with two Neural Network Layers architectures: DenseLayer and OutputLayer.⁶² Moreover, the J48 decision tree, developed by Ross Quinlan⁶³ (J48), and RF⁶⁴ were applied as representations of DT classifiers. Other functions such as SVMs (Linear and Non-Linear Functions),^{65, 66} k Nearest Neighbors (KNN),⁶⁷ and Binary Logistic Regression (BLR)⁶⁸ were implemented.

The typical statistical values of the IFPTML models based on these techniques are shown in **Table 6.3**. The values of all the IFPTML models (Training/Validation Series) demonstrate good results (Accuracy global 88.4-97.01 %), with 11 of them outperforming the IFPTML DADNP vs. MN model (89.8 %) (except NBN with 88.4 %). Similarly, AUROC scores are typically high (93-99 percent). SVMs (both linear and non-linear) have low values (0.5 and 0.78, respectively). In the case of non-linear SVM, the kernel used was Radial Basis Function (RBF). RF and KNN stand out as having the greatest precision, Sn, Sp, and AUROC, good binary classification models for the data under consideration, in the analysis and comparison of the ten algorithms used. In the case of kNN, the number of nearest neighbors (k) was 1 and types of nearest neighbor search algorithm was LinearNNSearch with EuclideanDistance. On the other hand, the RF, KNN, BN, and NBN models all have high Sn and Sp values, with only a tiny difference. They can be considered models having a high capability for positive and negative data prediction. However, SVMs, BLR, DL, Multi Boost, and Adaboost show very low Sn values, Sn (0-59 %), which contrasts with Sp (>97 %), indicating that they are ineffective at classifying this data set.

Table 6.3. IFPTML-WEKA AI/ML models.

Models ^a	Set ^b	Stat. ^c	Val. (%)	Class	Observed		AUROC ^d
				Pred.	1	0	
KNN	t	Sn	91.27	1	7804	2963	0.991
		Sp	97.44	0	746	112853	
	v	Sn	85.62	1	2394	1156	0.981
		Sp	97.01	0	402	37493	
BN	t	Sn	90.18	1	7710	5529	0.98
		Sp	95.23	0	840	110287	
	v	Sn	90.56	1	2532	1839	0.98
		Sp	95.24	0	264	36810	
RF	t	Sn	89.60	1	7661	2820	0.99
		Sp	97.57	0	889	112996	
	v	Sn	84.87	1	2373	1090	0.984
		Sp	97.18	0	423	37559	
NBN	t	Sn	85.78	1	7334	13226	0.962
		Sp	88.58	0	1216	102590	
	v	Sn	86.70	1	2424	4389	0.964
		Sp	88.64	0	372	34260	
J48-DT	t	Sn	84.91	1	7260	2976	0.986
		Sp	97.43	0	1290	112840	
	v	Sn	84.12	1	2352	1027	0.984
		Sp	97.34	0	444	37622	
Logit-Boost	t	Sn	75.95	1	6494	2901	0.976
		Sp	97.50	0	2056	112915	
	v	Sn	77.43	1	2165	979	0.977
		Sp	97.47	0	631	37670	
DL	t	Sn	58.18	1	4974	2863	0.956
		Sp	97.53	0	3576	112953	
	v	Sn	58.44	1	1634	904	0.95
		Sp	97.66	0	1162	37745	
SVM	t	Sn	58.18	1	4974	2966	0.778
		Sp	97.44	0	3576	112850	
	v	Sn	58.66	1	1640	932	0.781

		Sp	97.59	0	1156	37717	
BLR	t	Sn	51.30	1	4386	2308	0.959
		Sp	98.01	0	4164	113508	
	v	Sn	51.43	1	1438	717	0.961
		Sp	98.14	0	1358	37932	
Ada-Boost	t	Sn	46.36	1	3964	1854	0.967
		Sp	98.40	0	4586	113962	
	v	Sn	46.57	1	1302	597	0.968
		Sp	98.46	0	1494	38052	
Multi-BoostAB	t	Sn	0.00	1	0	0	0.937
		Sp	100.00	0	8550	115816	
	v	Sn	0.00	1	0	0	0.939
		Sp	100.00	0	2796	38649	
LibSVM	t	Sn	0.00	1	0	0	0.5
		Sp	100.00	0	8550	115816	
	v	Sn	0.00	1	0	0	0.5
		Sp	100.00	0	2796	38649	

^a ML-Classification Models. RF: Random Forest, kNN: k Nearest Neighbors, BLR: Binary Logistic Regression, BN: Bayes network, NBN: Naïve Bayes, J48-DT: J48 decision tree, Ada Boost, Logit Boost, Multi Boost, DL: Deep Learning, Lib SVM and SVM: Support Vector Machines. ^b Sub-set. T: Training set, v: Validation set. ^c Stat. Statistical performance. Sn: Sensibility, Sp: Specificity. ^d AUROC: Area under ROC value.

The Matthew correlation coefficient (MCC) shows a very strong positive relationship for MLP, KNN, RF, BN J48, and LogitBoost (See **Table 6.4**). The remaining techniques (including LDA) show strong positive relationships except for MultiBoostAB, and LibSVM, which, as mentioned before, were not good classifiers for the case under study. On the other hand, the highest F1 score values are shown in KNN and RF (>80%). This metric, according to,^{36, 69} is the most convenient for estimating the quality of models on imbalanced sets. Real accuracies (Q_2) ranged from 88.4% to 97.1%, and $Q_{2, \text{rnd}}$ ranged from 78.9 to 93.1. The value of the difference between the actual model accuracy and the corresponding random accuracy ranged from 5.7-11.4. This value, according to ³⁵ is considered significantly lower in comparison with the maximal possible (i.e., $\Delta Q_2 = 50\%$), which can be obtained only for the most difficult two-state classification model. However, the analysis presented by the cited authors was for balanced data (50:50), and in our case, the data structure contains ~ 93% inactive vs. 7% active cases. Hence, the values of the MLP (8: 8-10-1: 1, 8: 8-9-1: 1, 8: 8-7-1: 1), KNN, RF, BN, and J48 models are close to it and may have an adequate contribution to the actual accuracy of the estimation or prediction at the most likely random accuracy level.

Table 6.4. IFPTML DADNP vs. MN parameter for validation models (expressed in %).

Models ^a	Set ^b	Q ₂ , Ac ^c	F ₁ score ^d	MCC ^e	Q _{2,rd} ^f	ΔQ ₂ ^g	Q _{2,rd-bal} ^h
MLP 8:8-10-1:1	t	94.9	71.6	71.6	83.4	11.4	80.0
	v	95.1	72.4	72.4	83.7	11.4	80.3
MLP 8:8-9-1:1	t	94.2	68.9	69.0	82.9	11.3	79.0
	v	94.3	69.0	69.2	83.1	11.2	79.2
MLP 8:8-7-1:1	t	93.6	66.7	66.8	82.4	11.2	78.2
	v	93.9	67.4	67.6	82.8	11.1	78.8
KNN	t	97.0	80.8	79.8	85.7	11.4	84.2
	v	96.2	75.4	74.1	85.8	10.4	84.3
RF	t	97.0	80.5	79.4	85.9	11.2	84.6
	v	96.3	75.8	74.4	86.0	10.3	84.7
BN	t	94.9	70.8	70.1	83.9	10.9	81.0
	v	94.9	70.7	70.1	84.1	10.8	81.1
J48	t	96.6	77.3	75.8	86.0	10.5	84.9
	v	96.5	76.2	74.7	86.2	10.2	85.0
NBN	t	88.4	50.4	50.6	78.9	9.5	72.4
	v	88.5	50.5	51.0	79.0	9.5	72.5
LogitBoost	t	96.0	72.4	70.3	86.6	9.4	86.0
	v	96.1	72.9	71.0	86.7	9.4	86.0
LDA	t	89.8	52.3	51.5	80.6	9.2	75.3
	v	90.1	52.6	52.0	81.0	9.1	75.6
DL	t	94.8	60.7	58.0	87.7	7.1	88.8
	v	95.0	61.3	58.7	88.0	7.1	88.8
SMO	t	94.7	60.3	57.6	87.6	7.1	88.0
	v	95.0	61.1	58.5	87.9	7.1	88.4
BLR	t	94.8	57.5	55.3	88.5	6.3	89.8
	v	95.0	58.1	56.0	88.8	6.2	90.1
AdaBoost	t	94.8	55.2	53.6	89.1	5.7	91.1
	v	95.0	55.5	54.0	89.3	5.7	91.3
MultiBoostAB	t	93.1	-	-	93.1	0.0	100.0
	v	93.3	-	-	93.3	0.0	100.0
LibSVM	t	93.1	-	-	93.1	0.0	100.0
	v	93.3	-	-	93.3	0.0	100.0

^a ML-Classification Models. MLP: MultiLayer Perceptron, RF: Random Forest, kNN: k Nearest Neighbors, BLR: Binary Logistic Regression, BN: Bayes network, NBN: Naïve Bayes, J48-DT: J48 decision tree, Ada Boost, Logit Boost, LDA: Linear discriminant analysis, Multi Boost, DL: Deep Learning, Lib SVM and SVM: Support Vector Machines. ^b Sub-set. T: Training set, v: Validation set. ^c Q₂, Ac: Real accuracy, ^d F1 score, ^e MCC: Matthew's correlation coefficient, ^f Q_{2,rd}: Random

accuracy, ${}^s\Delta Q_2$: difference between the real and random accuracy, ${}^hQ_{2, \text{rnd-bal}}$: Most probable random accuracy for balanced model.

Figure 6.4 shows the values of Ac, Sn, and Sp obtained from randomization test Y (three randomizations of 5%, 10%, 15%, 20%, 25%, 30%, 35%, and 40% of the total training set (active and inactive). The figure reflects that the Ac values in percentage of randomized Y of these models were decreasing, from 89.8% to 39.5%, respectively. This result demonstrates that the overall good classification values were not due to chance correlations or structural redundancy in the training set.

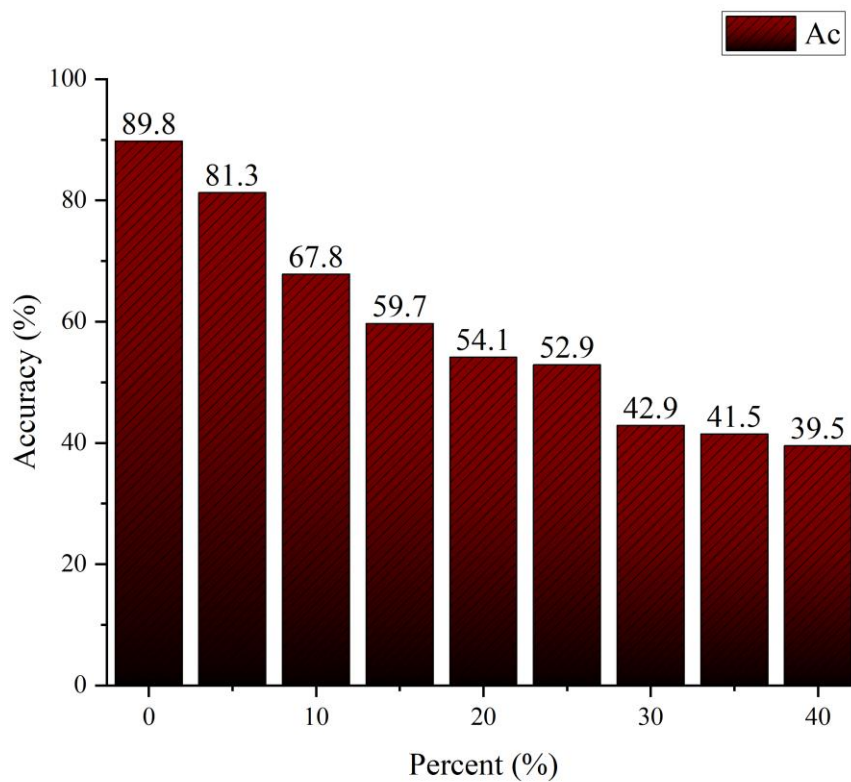


Figure 6.4. Values of the Accuracy in the Y-randomization test, from the different training data divisions.

The DoA of the IFPTML-LDA model is shown in **Figure 6.5**, the double ordinate plot of the residuals (first ordinate), and Leave-One-Out (LOO), (second ordinate) vs. the leverages (abscissa) (William Plot). The cases within the domain lie in the rectangular area within a band of ± 2 for the residuals and a leverage threshold of $h = 0.0002$.^{39, 70} As can be seen, most of the cases used in training and validation fall within this zone. However, there are a large number of cases that have leverage higher than the threshold but show LOO residuals and standard residuals within the limits. In these cases, with a high leverage value ($h > h^*$), the prediction should be considered unreliable. Leverage greater than warning leverage (h^*) means that the predicted response of the composite can be extrapolated from the model, and therefore the predicted value should be used with great care. Consequently, there are no cases in either the

training or prediction series with residual values outside the range ± 2 to that established for residuals and residual LOO. Therefore, no outliers are reported. Therefore, this model can be used with adequate accuracy for the prediction of new compounds in this DoA.⁴⁰

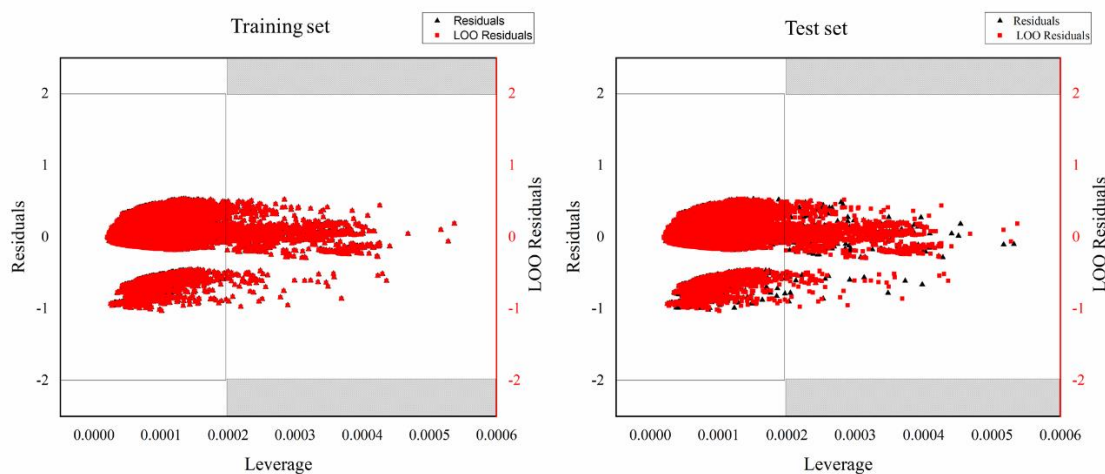


Figure 6.5. William's plot of residuals versus leverages for DADNP vs MN in the training and test sets.

3.4 IFPTML compared previous models.

After the internal comparison of the IFPTML-LDA vs. IFPTML-ANN models we decided to go ahead with comparison to other models reported before. We performed a review of the literature to carry out a comparison with other ML models. Only models involving AD assays, NP antibacterial assays, and/or MN of bacteria were included. Models including other drugs, other NP activities, or MN of other organisms were not included. We selected a total of 26 ML models reported in the literature. Many ML models involving AD have been reported, only those with large heterogeneous series of AD compounds have been included. Notably, most of the ML models focus on the study of AD and related compounds ignoring other subsystems of the present problem; see **Table 6.5**. Main part of these models focus on predict the probability of presenting AD activity in general without specifying the target bacteria species.⁷¹⁻⁷⁸ Only more recent models using PTML (PT + ML) methodology considers multiple biological parameters (multi-output) and multiple species.⁷⁹⁻⁸⁴ However, these PTML models also ignore the NP or MN components. One exception found was the PTML model by Speck-Planche *et al.* (includes NP but ignores MN and AD).^{27, 28} Another series of PTML models include MN but ignore NP and AD components. Nocado *et al.* published the first IFPTML *per se* on this topic because they carried out the IF of the AD and MN datasets.²⁸ However, this IFPTML model includes AD and MN but ignore NP subsystem. As result of this search, we can conclude the IFPTML models reported in this work are the only one considering the three subsystems at time AD, NP, and MD. On the other hand, Ortega *et al.*⁸⁵ and Diéguez *et al.*⁸⁶ also published IFPTML approaches to the present problem. Nevertheless, both models fail to account for one of the parts of the system. The first omit the AD component and the second omit the MN

component of the DADNP-MN interaction. An alternative to model developed in the present work is the combination some of the pairs of the previous models. For instance, if combine Speck-Planche *et al.* model (NP component) with Nocedo *et al.* model (AD + MN components) at the same time we can obtain a DADNP-MN interaction model. We can also combine Ortega *et al.* model (NP + MN) with Diéguez *et al.* model (AD + NP). In this last case, the NP introduced on both models have to be the same. However, all these combinations involve the use of two equations with different kind of parameters, errors, *etc.* Conversely, the model present here is a single linear and additive model involving all the components of the DADNP-MN (AD, NP, and MN) at the same time.

Table 6.5. ML models of AD compounds, MN of bacteria, and/or NP antibacterial systems.

Model Type	m ^a	AD ^a	NP ^b	MN ^c	MO ^d	MS ^e	MUL ^e	ML ^f	n ^g	Ac. ^h (%)	Val. ^d	Ref.
IFPTML	1	Yes	Yes	Yes	Yes	Yes	Yes	RF	>165K	97.0	i	This work
	2	Yes	Yes	Yes	Yes	Yes	Yes	ANN	>165K	95.0	i	Thiswork
	3	Yes	Yes	Yes	Yes	Yes	Yes	LDA	>165K	89.8	i	Thiswork
	4	No	Yes	Yes	Yes	Yes	Yes	LDA			i	85
	5	Yes	Yes	No	Yes	Yes	Yes	LDA			i	86
	6	Yes	Yes	No	Yes	Yes	Yes	KNN			i	86
	7	Yes	No	Yes	Yes	Yes	Yes	LDA	83605	88.6	i	28
PTML	8	No	No	Yes	Yes	Yes	No	LDA	>100K	72.3	i	30
	9	No	No	Yes	Yes	Yes	No	LDA	>100K	78.0	i	29
	10	No	No	Yes	Yes	Yes	No	LDA	>300K	85.0	i	87
	11	No	Yes	No	Yes	Yes	Yes	LDA	300	77.7	i	27
	12	Yes	No	No	Yes	Yes	Yes	LDA	2488	90.0	i	84
	13	Yes	No	No	Yes	Yes	Yes	LDA	30181	90.0	i	83
	14	Yes	No	No	Yes	Yes	Yes	ANN	54000	90.0	i	83
	15	Yes	No	No	Yes	Yes	Yes	LDA	3592	96.0	i	82
	16	Yes	No	No	Yes	No	Yes	LDA	37800	95.0	i	81
	17	Yes	No	No	Yes	Yes	Yes	ANN	11576	97.0	i	80
	18	Yes	No	No	Yes	Yes	Yes	LDA	12000	90.0	i	79
ML	19	Yes	No	No	No	No	No	LDA	667	92.9	i	78
	20	Yes	No	No	No	No	No	LDA	661	92.6	ii	77
	21	Yes	No	No	No	No	No	BLR	661	94.7	ii	77
	22	Yes	No	No	No	No	No	ANN	661	-	iii	77
	23	Yes	No	No	No	No	No	LDA	352	91.0	i	76
	24	Yes	No	No	No	No	No	LDA	111	94.0	i	75
	25	Yes	No	No	No	No	No	ANN	111	89.0	i	75
	26	Yes	No	No	No	No	No	LDA	-	> 90	i	74
	27	Yes	No	No	No	No	No	LDA	972	86.8	i	73
	28	Yes	No	No	No	No	No	LDA	458	~ 85	i	72
	29	Yes	No	No	No	No	No	LDA	433	~ 85	i	71

^a m = model number, AD = The model includes Antibacterial Drug (AD) compounds. ^b NP = The model includes Nanoparticles. ^c MN = The model includes metabolic network. ^d MO = Multi Output: multi-output models (MIC, IC₅₀, MBC, etc.). ^e MUL/MS: MUL = Multi-label model include in the inputs multiple labels of experimental condition like organism, target protein, cell lines, if necessary; MS = Multi-species model (one special case of MUL model). ^f ML= ML Technique: LDA = Linear Discriminant Analysis, RF=Random Forest, ANN= Artificial Neural Network, BLR= Binary Logistic Regression. n =Total number of cases (AD compounds, MN links, and/or NP assays, etc.) in training and/or validation series. ^hAc(%) = Accuracy of the model. ^d Val. =Validation methods: i) external validation series, ii) leave-30%-out cross validation, and iii) 100-times-averaged re-substitution technique. Furthermore, note that methods ii and iii are cross-validation methods.

3.5 IFPTML mapping of DADNP vs. MN of strains.

In the introduction, we mentioned the importance of predicting which bacteria strains with different MN may lead to MDR strains.⁸⁸ MDR strain surveillance is defined as the task of making a follow up of the new MDR strains as soon as them appear.^{89, 90} Once the IFPTML-LDA model was trained we can use it to map perturbations on MN of mutant strains vs. DADNP systems. We can define this kind of simulation as a computer-based MDR surveillance experiment. In the real world, techniques like Knock Out (KO) may be used to experimentally generate new strains with modifications on gen and protein expression. KO strains may be useful to identify genome-proteome vs. phenotype relationships.⁹¹ In the present simulation we are going to use both MNs of real wild type bacteria and computationally-generated MNs of KO strains. These KO mutant strains are mutants of the real bacterial (wild type) with a computationally perturbed metabolism. We call here a real bacteria those with the exact parameters of the MNs used training the model. Then, these mutant strains are not real but plausible synthetic organisms that have been computationally (*in silico*) generated. The simulation involved putative DADNP vs. real and computationally-generated synthetic bacteria strains in >20000 assays. These assays involve >3000 putative DADNP systems formed by >1600 compounds (mostly FDA approved drugs and some experimentally assayed compounds) with 16 NP and 17 CA options. The assays include different biological activity parameters (IC₅₀, MIC, etc.) vs. 22 Bacteria species. For the sake of simplicity, we used the IFPTML-LDA model (linear model) to run the simulation. In so doing, we changed the $\Delta Sh(D_{sk})$ values of the MNs of each real bacteria by the values of $\Delta Sh(D_{sk})$ calculated for their artificial synthetic mutant strains. We created these strains by increasing/decreasing *at random* the L_{in} and/or L_{out} degrees of the nodes of their MNs in a 1%. In first instance, 1% was considered a degree of perturbation large enough to cause important changes in the output of the model but still keeping the main features of the original MN in certain extend. Consequently, we obtained six types of mutants according to their incremental/subtractive perturbations in L_{in} , L_{out} , or both. It makes a total of $N_{run} > N_{assay} \cdot (1 + N_{mutant}) = 20000 \cdot (6 + 1) > 140000$ simulation runs. The number 1 accounts for the wild type bacteria species calculated for comparison purposes. Incremental perturbations imply increasing the number of chemical reactions (higher L_{in} and/or L_{out}) in the MN and consequently increasing overall metabolism. Conversely, subtractive perturbations imply decreasing the number of chemical reactions (lower L_{in} and/or L_{out}) in the MN and consequently decreasing overall metabolism. Specifically, changes in L_{in} imply changes on the number of substrates (educts) *per* product

(Anabolism unbalance with respect to wild type). Changes in L_{out} imply changes on the number of product (adducts) *per* substrate (Anabolism unbalance with respect to wild type).

In this discussion we are going to focus mainly on strains with subtractive perturbations. This includes subtractive perturbations in both directions $BB(-\delta_{in}, -\delta_{out})$ of metabolism. The group also include mutants with subtractive perturbations in only one direction $BB(-\delta_{in}, \delta_{out})$ or $BB(\delta_{in}, -\delta_{out})$. These perturbations in the MN are realizable in synthetic biology by experimental Knockout (KO), silencing with siRNAs, *etc.*, of the gene encoding for the respective enzymes.⁹² Firstly we focused on KO strains with code $BB(0, -0.1)$. These are KO strains with unchanged Anabolism ($\delta_{in}(\%) = 0\%$) and decreased Catabolism $\delta_{out}(\%) = -10\%$. The computational generation of the MNs of these KO strains is not trivial. Very often the elimination of one protein (due KO of the gen) results in the domino-effect (- sign) elimination of one or more reactions (links) of the MN and/or neighbor nodes (metabolites) and/or changes in topology as well. Interestingly, KO of one gene may result in subtractive but also on incremental perturbations. In other cases, appear “silent mutants” with changes in topology of the MN but overall constant average L_{in} and L_{out} values, see **Figure 6.6**.

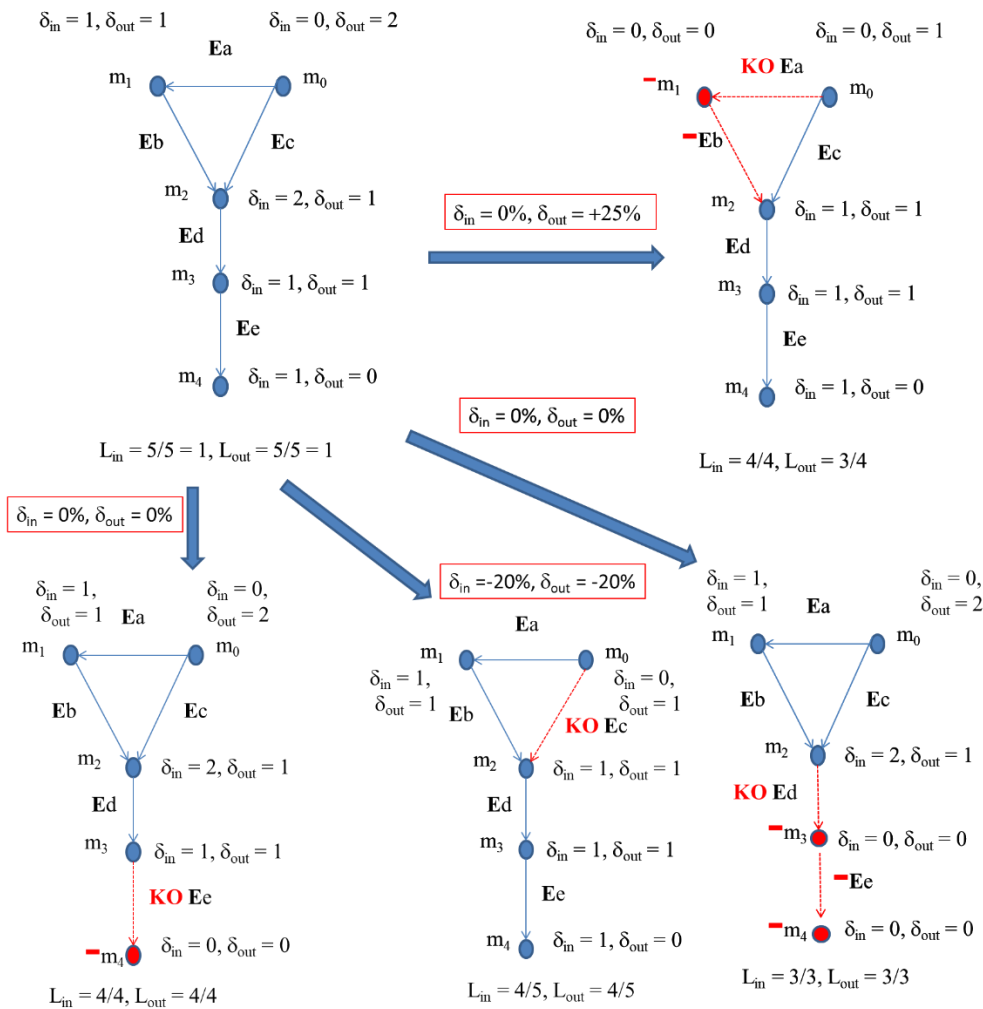


Figure 6.6. Effect of KO of gene over MNs topology.

For instance, it is straightforward to realize that KO of *Ea gene* deletes Enzyme **Ea** from the MN of the wild mutant. This in turn leads to the disappearance of metabolite m_1 and the reaction catalyzed by enzyme **Eb** due to lack of educts (substrates). As result we have the mutant **BB(0, +0.25)** with 25% decrease in Catabolism but not average changes in Anabolism as measured by L_{in} and L_{out} (average parameters). However, KO of gen **Ec** results in subtractive perturbation with reduction of both L_{in} and L_{out} creating mutants of type **BB(-0.2, -0.2)**. Last but not least, KO of gene **Ed** or **Ee** results in silent mutants **BB(0, 0)** with important changes on MN topology which are undetected by average L_{in} and L_{out} values because a proportional elimination of nodes. Solution to silent mutants' problem is easy by accounting for the perturbations (δ_k) on other MN topological descriptors. The simplest and meaningful alternatives are the number of nodes/metabolites (m) and/or number of links/reactions (r). Resulting in an extended notation of the mutants as follows **BB(δ_{in} , δ_{out} , δ_m , δ_r)**. This notation clearly differentiates wild type mutant **BB(0, 0, 0, 0)** from KO **Ed** with extended notation **BB(0, 0, -0.4, -0.4)** and **Ee** mutants with notation **BB(0, 0, -0.2, -0.2)**. We decided to continue using here **BB(- δ_{in} , - δ_{out})** notation because they are the only MN variables in the IFPTML model and to keep it simple.

After clarifying the notation we used the IFPTML-LDA model to predict the probabilities $p(f(v_{ij}, v_{sj}) = 1)$ with which all mutants generated are susceptible to AD. It means the probabilities with which $f(v_{ij}, v_{sj})_{pred} = 1$ for this pair AD and MN. In order to adapt the notation to this specific simulation experiment from now own we symbolized this probabilities as $p(f(v_{ij}, v_{sj}) = 1) = p(\text{BB}(\delta_{in}, \delta_{out}))_{ij}$. The **Table 6.6** summarizes the $p(\text{BB}(\delta_{in}, \delta_{out}))_{ijn}$ probability values predicted by the IFPTML-LDA model for different AD vs. KO strains with code **BB(0, -0.1)**.

IFPTML model predicts that many of these synthetic mutants could be susceptible to AD systems activity $p(\text{BB}(\delta_{in}, \delta_{out}))_{ijn} > 0.5$ but other may become MDR strains with $p(\text{BB}(\delta_{in}, \delta_{out}))_{ijn} < 0.5$. The drugs forming these AD are **MOX** = Moxifloxacin, **IMI** = Imipenem, **AZT** = Azithromycin, **GAT** = Gatifloxacin, **CLT** = Clarithromycin, **TET** = Tetracycline, **AMP** = Ampicillin, **KAN** = Kanamycin, **CFT** = Ceftriaxone, **PEN** = Penicillin. For instance, the first block of AD systems depicted correspond to the uncoated NP (Coat = N/A) complexes of Ag, Cu, CuI, CuO, Fe₂O₃, and ZnO with multiple drugs. The IFPTML predicts that the ASM of the strain **BS (0, -0.1)** should be susceptible to the AD systems of Ag, CuI, CuO, Fe₂O₃, and ZnO with $p(\text{BB}(\delta_{in}, \delta_{out}))_{ijn} > 0.7$ for different drugs (Avg. = 0.7 – 0.9 range). These ASM strains should be less susceptible to AD formed by Cu NP (Avg. = 0.48). To cite few examples, the IFPTML also predicts that the ASM strain **EC (0, -0.1)** should be susceptible to the AD systems (Avg. = 0.6 – 1 range) but **EF(0, -0.1)** strains should be highly resistant (Avg. = 0.04 – 0.23 range). Overall, all the strains studied should be MDR strains vs. AD systems of **CFT** with other NP (Avg. = 0.49).

Table 6.6. IFPMTL probabilities for AD vs. BB(0, -0.1) mutants (selected examples).

Bact	AD	Dru	MO	IM	AZ	GA	CL	TE	AM	KA	CF	PE	
MN	NP	Coat	Avg	0.65	0.6	0.5	0.6	0.5	0.4	0.50	0.70	0.4	0.5
BS	Ag	N/A	0.90		0.9				0.9	0.88	0.86		0.9
BS	Cu	N/A	0.48		0.5		0.1		0.4	0.50	0.45	0.4	0.5
BS	CuI	N/A	0.90		0.9				0.9	0.89	0.87		0.9
BS	CuO	N/A	0.74			0.7		0.7	0.3	0.78	0.74	0.7	0.7
BS	Fe ₂ O	N/A	0.71						0.7	0.78	0.74	0.7	0.5
BS	ZnO	N/A	0.75		0.8		0.7			0.78	0.74	0.7	0.7
EC	Ag	N/A	0.92	0.96	0.9		0.8	0.9			0.95		
EC	Au	PDT/CQ	1.00	1.00	1.0		1.0				1.00		
EC	Cu	N/A	0.69	0.76	0.7		0.6				0.72		
EC	CuI	N/A	0.89	0.89	0.8		0.8				0.86		
EC	CuO	N/A	0.83	0.76	0.8		0.7				0.87		
EF	Ag	N/A	0.23		0.2	0.3	0.3	0.1	0.2	0.21	0.17		
EF	Cu	N/A	0.04		0.0	0.0	0.0	0.0	0.0	0.05	0.03		
EF	CuI	N/A	0.26	0.38	0.2	0.2	0.2		0.2	0.08	0.26	0.3	
EF	CuO	N/A	0.12	0.12	0.1	0.1	0.1	0.1	0.1	0.12	0.09	0.1	0.2
EF	Fe ₂ O	N/A	0.11	0.12	0.1		0.1		0.1	0.12	0.08	0.1	0.1
EF	ZnO	N/A	0.11	0.12	0.1		0.1		0.1	0.10		0.1	
HI	Ag	N/A	0.86	0.85	0.8	0.8	0.9	0.8		0.84		0.8	0.9
HI	Cu	N/A	0.46	0.43	0.4	0.4	0.6	0.4		0.42		0.4	
HI	CuI	N/A	0.86	0.86	0.8	0.7	0.9	0.8		0.87		0.8	0.9
HI	CuO	N/A	0.68	0.77	0.7	0.7	0.7	0.7	0.4	0.71		0.7	0.7
HI	Fe ₂ O	N/A	0.71	0.77	0.7	0.7	0.7	0.7		0.74		0.7	0.7
HI	ZnO	N/A	0.69	0.72	0.7	0.5	0.7	0.7		0.74		0.7	
MT	Ag	N/A	0.90	0.72			0.9				0.91		
MT	CuI	N/A	0.83	0.82			0.9				0.83		
MT	CuO	N/A	0.71	0.87			0.6	0.6			0.83		
NG	Ag	N/A	0.26	0.42		0.2			0.1	0.29		0.2	0.1
NG	CuI	N/A	0.23	0.32		0.1			0.1	0.39		0.1	0.1
NG	CuO	N/A	0.15			0.1	0.1		0.0	0.25		0.1	0.0
PA	Au	PDT/Me	0.09	1.00	1.0		1.0				1.00		1.0
PA	Au	PDT/CQ	1.00	1.00	1.0		1.0				1.00		
PA	Au	PDT/CP	1.00	1.00	1.0		1.0				1.00		1.0
PA	Au	PDT/G	1.00	1.00	1.0						1.00		
PA	Au	PDT/AC	1.00		1.0	1.0	1.0				1.00		
PA	Au	PDT/D	1.00	1.00	1.0		0.9				1.00		
PA	Au	PDT	1.00	0.99	0.9		0.9				0.99		0.9
PA	CuO	N/A	0.99	0.89	0.6	0.8	0.8				0.83		0.8

^aMOX = Moxifloxacin, IMI = Imipenem, AZT = Azithromycin, GAT = Gatifloxacin, CLT = Clarithromycin,

TET = Tetracycline, AMP = Ampicillin, KAN = Kanamycin, CFT = Ceftriaxone, PEN = Penicillin.

In addition to BB (0, -0.1) mutants we also simulated the behavior of BB (-0.1,0). This second group includes KO mutants with a decreased Anabolism with respect to the wide type $\delta_{out} < 0$ and unaltered Catabolism $\delta_{in} = 0$. The **Figure 6.7** plots the probabilities predicted by the IFPTML model for both groups of KO mutants. It is interesting that we found three different series of results. These seem to be formed by assays of AD with the same AD and specific set of NP vs. different KO mutants. The linear series have predicted probabilities $p(-0.1,0) \approx p(0,-0.1)$ for both classes of KO mutants BB(-0.1, 0) and BB(0, -0.1) of the same specie BB. The linear series presuppose a gradual linear change in the AD activity regardless the MN of mutant strains has a decrease of Anabolism or Catabolism. It may indicate that if a BB has this behavior, it could be a species with the certain resilience to mutations that deprived both directions of the metabolism. However, there are other series with clear non-linear convex or concave behaviors. The study of the behavior of series of assays could be interesting but a classification by quadrants could be more systematic. More precisely we can classify BB species in four different quadrants of this chart. BB within quadrant I are predicted to be susceptible to AD activity and then less dangerous. Species within quadrants II, III, and IV could generate MDR strains due to mutations. The KO mutants of species in quadrant II seem to become MDR strains only if they have mutations that deprive they Catabolism ($\delta_{out} < 0 \Rightarrow p(-0.1, 0) < 0.5$).

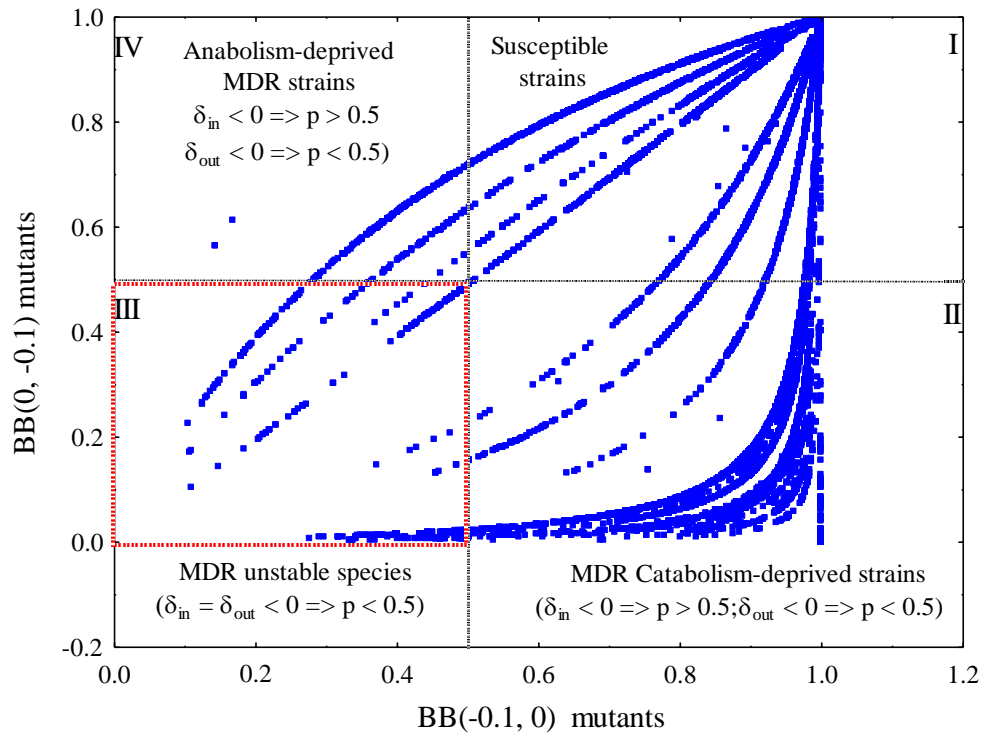


Figure 6.7. IFPTML mapping of KO mutants.

Conversely, the KO mutants of species in quadrant III seem to become MDR strains only if they have mutations that deprive they Anabolism ($\delta_{in} < 0 \Rightarrow p(-0.1, 0) < 0.5$). Species in the quadrant IV are probably the more unstable because they KO mutants seem to become MDR strains irrespective of the kind of metabolism deprived. These species should be specially subjected to MDR surveillance. In the file S01.xlsx we show the results of all these simulation

runs. See average values in sheet SAVG and full simulation results in sheet SIMUL. This excel book is an interactive example with active cells (sheet SAVG) allowing the reader to change the values in order to run his own simulations (we set $\delta_{in} = \delta_{out}$ for simplicity). The sheet SIMUL of the same book allows $\delta_{in} \neq \delta_{out}$ simulations as well. All in all, these studies should be taken with caution. All conclusions withdrawn from this kind of studies should be corroborated experimentally. The real utility expected from the model should be the fast and inexpensive calculation of very large series of AD vs. very large series of wild type and mutant strains. This could help to short list some AD and/or bacteria strains for experimental assays. Prediction of few examples to obtain 100% accurate experimental reproduction of the results is discarded due to the probabilistic nature of the model.

3.6 DADNP experimental cases simulation.

Furthermore, we utilized the IFPTML model to estimate the probability values of numerous DADNPs that have already been synthesized, biologically evaluated, and reported in the literature. The following were the study's inclusion criteria. We considered examples that reported 1) AD antibacterial activity, 2) NP antibacterial activity, 3) DADNP complex antibacterial activity, and 4) the microorganism's known metabolic network. We considered examples with DADNP activity that was both additive and synergistic. At least one report of one biological activity metric was found in the cases chosen. The revision includes a total of 65 studies that found positive DADNP instances in a total of 21 papers.⁹³⁻¹¹³ We collected 102 reports of DADNP complex tests with at least one favorable experimental outcome from these journals. The size of the NP used to build the DADNP complexes ranges from 5 nm to 100 nm. Some DADNP formed following the inclusion of the coat and the AD may have a size >100 nm. Additionally, the AD utilized to build the DADNPs exhibit a wide range of hydrophobicity, from hydrophilic (LOGP < 0) to lipophilic (LOGP > 0). The MIC ($\mu\text{g}\cdot\text{mL}^{-1}$) was determined experimentally for AD, NP, and DADNP complex. In all cases, MIC ($\mu\text{g}\cdot\text{mL}^{-1}$) < 50 (cutoff used in the model) was used for the DADNP complexes simulation. The reported assay times ranged from 12 to 24 hours. The DADNP design incorporates coating agents that may aid boost the complexes' stability and/or bioavailability over time.⁹³⁻¹¹³

The surface scatterplot of AD Hydrophobicity vs. the histograms of NP size and MN reactions number are shown in **Figure 6.8**. Most of the DADNP assays present values of metabolic networks between 2200-2450 reactions, LOGP, 0-2, and NP size less than 50 nm. Metallic nanoparticles (Ag, Au, and Zn), double metal nanoparticles (ZnCu), metal oxide nanoparticles (Fe_3O_4 , CuO, ZnO, and others), and metal salts (MoS_2 and AgNO_3) were all detected. Polyvinylpyrrolidone (PVP), Polyethylene glycol (PEG), Thioglycolic acid (TGA), Polydopamine (PDA), Triethylene Glycol (TEG), Alginate, and Chitosan were the most frequently utilized coating materials. The cases involve a diverse array of bacteria, including several strains of *P. Aeruginosa*, *E. Coli*, *S. aureus*, *E. faecalis*, *E. Faecium*, *S. epidermidis*, *B. subtilis*, *A. Baumannii*, *S. enterica*, *Y. pestis*, and *K. pneumoniae*. In any case, the DADNP complexes discovered contained a wide set of ADs, including PEN (Ampicillin, Meropenem, Imepenem), TETRA (Tetracycline), MACRO (Gentamicin, Vancomycin, Rifampicin), QUIN

(Ofloxacin, Ciprofloxacin), Aminoglycoside (Tobramycin, Kanamycin, Kanamycin, Streptomycin) Amphenicols (Chloramphenicol), Lipopeptide (Daptomycin), and Polypeptide (Polymyxin B).^{34, 78-97}

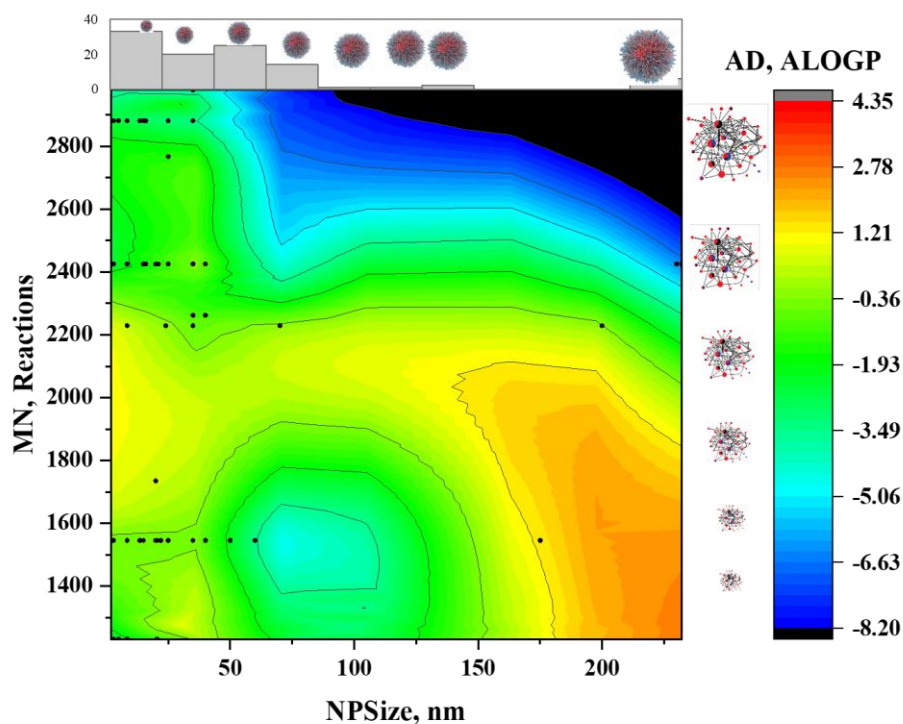


Figure 6.8. AD Hydrophobicity Surface scatterplot vs. Histograms of NP size and MN Reactions distribution.

We may infer that this experimental collection of DADNP complexes preclinical experiments contains a significant degree of structural and biological variety. Intentionally, our initial collection of AD and NP assays vs MN used to assemble putative DADNP complexes a train with a significant structural and biological variety. It may aid our additive model in learning how to differentiate between active and inactive DADNP complexes using an additive approach. Indeed, our IFPTML model was able to predict as positive all 102 cases ($p(\text{DADNP}_{in})_{c_{dj}, c_{nj}} > 0.99$ in all cases. **Table 6.7** summarizes selected IFPTML studies of experimentally validated DADNP complexes. The discovery is significant because it demonstrates that our IFPTML additive model is capable of correctly identifying experimentally tested DADNP complexes with a high degree of structural and biological diversity, including both additive and synergic examples. The study of the interaction of the physical properties of conjugated nanoparticles with antibacterial drugs vs. the various metabolic networks of microorganisms could have an impact on the reduction of costs and environmental impacts. The evaluation of these systems to obtain DADNP systems with desirable pharmacological properties (mainly towards MDR strains) is an area of great interest in nanomedicine and nanotoxicology. Thus, these complex systems are important in the rational design of nanomaterials and Nanosafety computational surveillance.

Table 6.7. Examples of IFPTML study of experimentally tested DADNP complexes.

NP Type	DADNP ^a	APSn(n m) ^b	Shape Obs. ^c	Specie ^d	MIC(μ gml ⁻¹) ^e	p(v _{ij} , v _{nj} , v _{sj}) _{obs} ^f	t(h) ^g	Re f. ^h
Doubled Metal	MER-CuZnNP-TEG	21	Sph	<i>P. aeruginosa</i>	25.197	0.999	24	⁹⁶
	CIP-CuZnNP-TEG	21	Sph	<i>P. aeruginosa</i>	3.184	0.999	24	⁹⁶
Metal	AMP-AgNP	3	Sph	<i>P. aeruginosa</i>	0.600	0.999	24	¹⁰¹
	KAN-AgNP	3	Sph	<i>P. aeruginosa</i>	0.500	0.999	24	¹⁰¹
	VAN-AgNP-TGA	20.5	Sph	<i>S. epidermitis</i>	0.020	0.999	24	⁹⁸
	AMP-AgNP	3	Sph	<i>E. coli</i>	0.375	0.999	24	¹⁰¹
	CHL-AgNP	3	Sph	<i>E. coli</i>	0.667	0.999	24	¹⁰¹
	KAN-AgNP	3	Sph	<i>E. coli</i>	0.500	0.999	24	¹⁰¹
	KAN-AuNP	20	Sph	<i>Y. pestis</i>	1.700	0.999	24	¹¹³
	IMI-AgNP	24.	Sph	<i>A. baumannii</i>	2.000	0.999	24	¹⁰⁸
	VAN-AgNP-TGA	20.5	Sph	<i>E. faecalis</i>	0.100	0.999	24	⁹⁸
	POL-AgNP	8.4	Sph	<i>A. baumannii</i>	0.004	0.999	18	¹⁰⁷
	RIF-AgNP	8.4	Sph	<i>A. baumannii</i>	0.527	0.999	18	¹⁰⁷
	TIG-AgNP	8.4	Sph	<i>A. baumannii</i>	1.041	0.999	18	¹⁰⁷
	AMP-AgNP	3	Sph	<i>E. coli</i>	0.222	0.999	24	¹⁰¹
	CHL-AgNP	3	Sph	<i>E. coli</i>	0.200	0.999	24	¹⁰¹
	KAN-AgNP	3	Sph	<i>E. coli</i>	0.222	0.999	24	¹⁰¹
	GEN-AgNP	40	Sph	<i>K. pneumoniae</i>	0.375	0.999	14	¹⁰¹
	AMP-AgNP	3	Sph	<i>E. faecium</i>	0.021	0.999	24	¹⁰¹
	AMP-AgNP	3	Sph	<i>S. aureus</i>	0.222	0.999	24	¹⁰¹
	CHL-AgNP	3	Sph	<i>E. faecium</i>	0.088	0.999	24	¹⁰¹
	CHL-AgNP	3	Sph	<i>S. aureus</i>	0.300	0.999	24	¹⁰¹
	KAN-AgNP	3	Sph	<i>S. aureus</i>	0.167	0.999	24	¹⁰¹
	AMP-AuNP-PEG	1.86	Sph	<i>S. aureus</i>	0.580	0.999	20	¹¹²
	VAN-AgNP-TGA	20.5	Sph	<i>S. aureus</i>	0.050	0.999	24	⁹⁸
Metal Oxide	CIP-CuONP-PEG	15	Sph	<i>P. aeruginosa</i>	4.752	0.999	24	⁹⁶
	CIP-ZnONP-PEG	35	Sph	<i>P. aeruginosa</i>	3.960	0.999	24	⁹⁶
	TOB-Fe ₃ O ₄ NP-PEG/Alg	16	Sph	<i>P. aeruginosa</i>	0.220	0.999	24	⁹³
	VAN-Mn ₂ Fe ₂ O ₄ NP-PEG/Ch	25	Sph	<i>B. subtilis</i>	0.780	0.999	24	⁹⁷
	VAN-Mn ₂ Fe ₂ O ₄ NP-PEG/Ch	25	Sph	<i>S. epidermitis</i>	0.610	0.999	24	⁹⁷
	IMI-AgNO ₃ NP	25	Cil	<i>P. aeruginosa</i>	4.000	0.999	24	¹¹¹

Metal Salt	OFL-MoS ₂ NP-Ch	175	Nf	<i>S. aureus</i>	25.000	0.999	24	¹⁰⁰
------------	----------------------------	-----	----	------------------	--------	-------	----	----------------

Notes: ^aDADNP: Dual Antibacterial Drug-Nanoparticles. (Include material coated). Antibacterial Drug. MER: Meropenem, CIP: Ciprofloxacin, AMP: Ampicillin, KAN: Kanamycin, VAN: Vancomycin, CHL: Chloramphenicol, IMI: Imipenem, RIF: Rifampicin, GEN: Gentamycin, POL: Polymyxin B, TIG: Tigecycline, TOB: Tobramycin, OFL: Ofloxacin. Coated material. TEG: Triethylene Glycol, TGA: Thioglycolic acid, PEG: Polyethylene glycol, Alg: Alginate, Ch: Chitosan. ^b: APSn(nm): Average of Size of Nanoparticle. ^c Shape Obs: Shape observed. Sph: Spherical, Cil: Cilindrical, Nf: Nanoflakes. ^d Specie: Microorganism species and metabolic network. ^e MIC ($\mu\text{g ml}^{-1}$): Minimum inhibitory concentration. ^f $p(v_{ij}, v_{nj}, v_{sj})_{\text{obs}}$: probability, calculated as $p(\text{DADNP}_{\text{in}} \text{ vs } \text{MN}/\mathbf{c}_{\text{dj}}, \mathbf{c}_{\text{nj}}, \mathbf{c}_{\text{sj}})_{\text{pred}} = 1/(1+\text{Exp}(-f(v_{ij}, v_{nj}, v_{sj})_{\text{calc}}))$. ^g t(h) Time of assay. ^h Ref. Reference.

4. CONCLUSIONS

Developing Dual Antibacterial Drug-Nanoparticles (DADNP) systems may become a new weapon on the arsenal to fight AD resistant MDR strains with different MNs. However, testing DADNP vs. strains with different MN is a hard and costly task. ML models may help to speed up the process. The IFPTML algorithm with an additive approach may be a practical solution to the DADNP discovery. This approach may be useful until researchers can accumulate larger experimental datasets of DADNP systems. Regarding the methodological objectives, The linear model included three subsystems (preclinical antibacterial drugs, metabolic network and nanoparticles with coating agents) and showed a good fit (Sn= 80.2%, Sp= 90.5% and Ac=89.8%). The information from the three subsystems did not significantly influence the robustness of the models to analyze the problem presented in the thesis.

Regarding the practical objectives, IFPTML-LDA was the simpler and still accurate model found. IFPTML-MLP models are more complicated but outperformed the linear models. Among the twelve ML algorithms used to create nonlinear IFPTML classification models, the RF, KNN, BN and NBN models had the highest Sn and Sp values. The IFPTML linear and additive model was able to predict 102 experimental cases of DADNPs complexes with a high degree of structural and biological variety reported in the literature. IFPTML models proposed could be useful for computational screening by short listing for experimental assays the more promising candidates. We introduce here the concept of MDR&Nanosafety computational surveillance. IFPTML models may be used also to run a large simulation of sensibility Knockout (KO) synthetic strains to DADNP systems in different assay conditions. These simulations may predict the answer to DADNP of bacteria with perturbations in MN structure. This in turn may help to detect new MDR bacteria strains.

5. REFERENCES

1. Zhavoronkov, A. Artificial Intelligence for Drug Discovery, Biomarker Development, and Generation of Novel Chemistry. *Molecular pharmaceutics*. **2018**, *15* (10), 4311-4313. DOI: 10.1021/acs.molpharmaceut.8b00930.
2. Feldmann, C.; Yonchev, D.; Stumpfe, D.; Bajorath, J. Systematic Data Analysis and Diagnostic Machine Learning Reveal Differences between Compounds with Single- and Multitarget Activity. *Molecular pharmaceutics*. **2020**. DOI: 10.1021/acs.molpharmaceut.0c00901.

3. Kosugi, Y.; Hosea, N. Direct Comparison of Total Clearance Prediction: Computational Machine Learning Model versus Bottom-Up Approach Using In Vitro Assay. *Molecular pharmaceuticals*. **2020**, *17* (7), 2299-2309. DOI: 10.1021/acs.molpharmaceut.9b01294.
4. Minerali, E.; Foil, D. H.; Zorn, K. M.; Lane, T. R.; Ekins, S. Comparing Machine Learning Algorithms for Predicting Drug-Induced Liver Injury (DILI). *Molecular pharmaceuticals*. **2020**, *17* (7), 2628-2637. DOI: 10.1021/acs.molpharmaceut.0c00326.
5. Fischbach, M. A.; Walsh, C. T. Antibiotics for emerging pathogens. *Science*. **2009**, *325* (5944), 1089-1093, Review. DOI: 10.1126/science.1176667 Scopus.
6. Nagar, S. D.; Aggarwal, B.; Joon, S.; Bhatnagar, R.; Bhatnagar, S. A Network Biology Approach to Decipher Stress Response in Bacteria Using Escherichia coli As a Model. *Omic : a journal of integrative biology*. **2016**, *20* (5), 310-324. DOI: 10.1089/omi.2016.0028.
7. Larocque, M.; Chenard, T.; Najmanovich, R. A curated C. difficile strain 630 metabolic network: prediction of essential targets and inhibitors. *BMC systems biology*. **2014**, *8*, 117. DOI: 10.1186/s12918-014-0117-z.
8. Mikolajczyk, A.; Gajewicz, A.; Mulkiewicz, E.; Rasulev, B.; Marchelek, M.; Diak, M.; Hirano, S.; Zaleska-Medynska, A.; Puzyn, T. Nano-QSAR modeling for ecosafe design of heterogeneous TiO₂-based nano-photocatalysts. *Environmental Science: Nano*. **2018**, *5* (5), 1150-1160, 10.1039/C8EN00085A. DOI: 10.1039/C8EN00085A.
9. Sizochenko, N.; Syzochenko, M.; Fjodorova, N.; Rasulev, B.; Leszczynski, J. Evaluating genotoxicity of metal oxide nanoparticles: Application of advanced supervised and unsupervised machine learning techniques. *Ecotoxicology and Environmental Safety*. **2019**, *185*, 109733. DOI: <https://doi.org/10.1016/j.ecoenv.2019.109733>.
10. Wu, F.; Harper, B. J.; Harper, S. L. Differential dissolution and toxicity of surface functionalized silver nanoparticles in small-scale microcosms: impacts of community complexity. *Environmental Science: Nano*. **2017**, *4* (2), 359-372, 10.1039/C6EN00324A. DOI: 10.1039/C6EN00324A.
11. Costa, P. M.; Fadeel, B. Emerging systems biology approaches in nanotoxicology: Towards a mechanism-based understanding of nanomaterial hazard and risk. *Toxicology and Applied Pharmacology*. **2016**, *299*, 101-111. DOI: 10.1016/j.taap.2015.12.014.
12. Ray, P. C.; Yu, H.; Fu, P. P. Toxicity and environmental risks of nanomaterials: challenges and future needs. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev*. **2009**, *27* (1), 1-35. DOI: 10.1080/10590500802708267 PubMed.
13. Zielińska, A.; Costa, B.; Ferreira, M. V.; Miguéis, D.; Louros, J. M. S.; Durazzo, A.; Lucarini, M.; Eder, P.; Chaud, M. V.; Morsink, M.; et al. Nanotoxicology and Nanosafety: Safety-By-Design and Testing at a Glance. *Int J Environ Res Public Health*. **2020**, *17* (13), 4657. DOI: 10.3390/ijerph17134657 PubMed.
14. Singh, A. V.; Laux, P.; Luch, A.; Sudrik, C.; Wiehr, S.; Wild, A.-M.; Santomauro, G.; Bill, J.; Sitti, M. Review of emerging concepts in nanotoxicology: opportunities and challenges for safer nanomaterial design. *Toxicology Mechanisms and Methods*. **2019**, *29* (5), 378-387. DOI: 10.1080/15376516.2019.1566425.
15. das Neves, J.; Sverdlov Arzi, R.; Sosnik, A. Molecular and cellular cues governing nanomaterial–mucosae interactions: from nanomedicine to nanotoxicology. *Chemical Society Reviews*. **2020**, *49* (14), 5058-5100, 10.1039/C8CS00948A. DOI: 10.1039/C8CS00948A.
16. Mitchell, M. J.; Billingsley, M. M.; Haley, R. M.; Wechsler, M. E.; Peppas, N. A.; Langer, R. Engineering precision nanoparticles for drug delivery. *Nature Reviews Drug Discovery*. **2021**, *20* (2), 101-124. DOI: 10.1038/s41573-020-0090-8.

17. Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z. N.; Barabasi, A. L. The large-scale organization of metabolic networks. *Nature*. **2000**, *407* (6804), 651-654.
18. Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research*. **2019**, *47* (D1), D930-D940. DOI: 10.1093/nar/gky1075.
19. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; et al. The ChEMBL database in 2017. *Nucleic acids research*. **2017**, *45* (D1), D945-D954. DOI: 10.1093/nar/gkw1074.
20. Gajewicz, A. What if the number of nanotoxicity data is too small for developing predictive Nano-QSAR models? An alternative read-across based approach for filling data gaps. *Nanoscale*. **2017**, *9* (24), 8435-8448, Article. DOI: 10.1039/c7nr02211e Scopus.
21. Urista, D. V.; Carrue, D. B.; Otero, I.; Arrasate, S.; Quevedo-Tumaili, V. F.; Gestal, M.; Gonzalez-Diaz, H.; Munteanu, C. R. Prediction of Antimalarial Drug-Decorated Nanoparticle Delivery Systems with Random Forest Models. *Biology*. **2020**, *9* (8). DOI: 10.3390/biology9080198.
22. Santana, R.; Zuluaga, R.; Gañán, P.; Arrasate, S.; Onieva, E.; González-Díaz, H. Designing nanoparticle release systems for drug-vitamin cancer co-therapy with multiplicative perturbation-theory machine learning (PTML) models. *Nanoscale*. **2019**, *11* (45), 21811-21823, Article. DOI: 10.1039/c9nr05070a Scopus.
23. Kleandrova, V. V.; Luan, F.; Gonzalez-Diaz, H.; Ruso, J. M.; Speck-Planche, A.; Cordeiro, M. N. Computational tool for risk assessment of nanomaterials: novel QSTR-perturbation model for simultaneous prediction of ecotoxicity and cytotoxicity of uncoated and coated nanoparticles under multiple experimental conditions. *Environ Sci Technol*. **2014**, *48* (24), 14686-14694. DOI: 10.1021/es503861x.
24. Luan, F.; Kleandrova, V. V.; Gonzalez-Diaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, M. N. Computer-aided nanotoxicology: assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. *Nanoscale*. **2014**, *6* (18), 10623-10630. DOI: 10.1039/c4nr01285b.
25. Santana, R.; Zuluaga, R.; Ganan, P.; Arrasate, S.; Onieva, E.; Montemore, M. M.; Gonzalez-Diaz, H. PTML Model for Selection of Nanoparticles, Anticancer Drugs, and Vitamins in the Design of Drug-Vitamin Nanoparticle Release Systems for Cancer Cotherapy. *Mol Pharm*. **2020**, *17* (7), 2612-2627. DOI: 10.1021/acs.molpharmaceut.0c00308.
26. Concu, R.; Kleandrova, V. V.; Speck-Planche, A.; Cordeiro, M. Probing the toxicity of nanoparticles: a unified in silico machine learning model based on perturbation theory. *Nanotoxicology*. **2017**, *11* (7), 891-906. DOI: 10.1080/17435390.2017.1379567.
27. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. Computational modeling in nanomedicine: prediction of multiple antibacterial profiles of nanoparticles using a quantitative structure-activity relationship perturbation model. *Nanomedicine (Lond)*. **2015**, *10* (2), 193-204. DOI: 10.2217/nmm.14.96.
28. Nocedo-Mena, D.; Cornelio, C.; Camacho-Corona, M. D. R.; Garza-Gonzalez, E.; Waksman de Torres, N.; Arrasate, S.; Sotomayor, N.; Lete, E.; Gonzalez-Diaz, H. Modeling Antibacterial Activity with Machine Learning and Fusion of Chemical Structure Information with Microorganism Metabolic Networks. *J Chem Inf Model*. **2019**, *59* (3), 1109-1120. DOI: 10.1021/acs.jcim.9b00034.

29. Duardo-Sanchez, A.; Munteanu, C. R.; Riera-Fernandez, P.; Lopez-Diaz, A.; Pazos, A.; Gonzalez-Diaz, H. Modeling complex metabolic reactions, ecological systems, and financial and legal networks with MIANN models based on Markov-Wiener node descriptors. *Journal of chemical information and modeling*. **2014**, *54* (1), 16-29. DOI: 10.1021/ci400280n.
30. Gonzalez-Diaz, H.; Riera-Fernandez, P. New Markov-autocorrelation indices for re-evaluation of links in chemical and biological complex networks used in metabolomics, parasitology, neurosciences, and epidemiology. *J Chem Inf Model*. **2012**, *52* (12), 3331-3340. DOI: 10.1021/ci300321f.
31. Diéguez-Santana, K.; Casañola-Martin, G. M.; Green, J. R.; Rasulev, B.; González-Díaz, H. Predicting Metabolic Reaction Networks with Perturbation-Theory Machine Learning (PTML) Models. *Current topics in medicinal chemistry*. **2021**, *21* (9), 819-827. DOI: 10.2174/1568026621666210331161144 From NLM.
32. Shannon, C. E. A Mathematical Theory of Communication. *The Bell System Technical Journal*. **1948**, *27*, 379-423.
33. Hill, T.; Lewicki, P. *Statistics: Methods and Applications*; StatSoft, Inc., 2005.
34. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. **2020**, *21* (1), 6. DOI: 10.1186/s12864-019-6413-7.
35. Batista, J.; Vikić-Topić, D.; Lučić, B. The Difference Between the Accuracy of Real and the Corresponding Random Model is a Useful Parameter for Validation of Two-State Classification Model Quality. *Croatica Chemica Acta*. **2016**, *89* (4), 527-534. DOI: 10.5562/cca3117.
36. Lučić, B.; Batista, J.; Bojović, V.; Lovrić, M.; Kržić, A. S.; Bešlo, D.; Nadramija, D.; Vikić-Topić, D. Estimation of random accuracy and its use in validation of predictive quality of classification models within predictive challenges. *Croatica Chemica Acta*. **2019**, *92* (3), 379-391. DOI: <https://doi.org/10.5562/cca3551>.
37. Papadiamantis, A. G.; Afantitis, A.; Tsoumanis, A.; Valsami-Jones, E.; Lynch, I.; Melagraki, G. Computational enrichment of physicochemical data for the development of a ζ -potential read-across predictive model with Isalos Analytics Platform. *NanoImpact*. **2021**, *22*, 100308. DOI: <https://doi.org/10.1016/j.impact.2021.100308>.
38. Golbraikh, A.; Tropsha, A. Beware of q²! *Journal of Molecular Graphics and Modelling*. **2002**, *20* (4), 269-276. DOI: [https://doi.org/10.1016/S1093-3263\(01\)00123-1](https://doi.org/10.1016/S1093-3263(01)00123-1).
39. Zhang, S.; Golbraikh, A.; Oloff, S.; Kohn, H.; Tropsha, A. A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *J Chem Inf Model*. **2006**, *46* (5), 1984-1995. DOI: 10.1021/ci060132x From NLM.
40. Casañola-Martín, G. M.; Marrero-Ponce, Y.; Tareq Hassan Khan, M.; Torrens, F.; Pérez-Giménez, F.; Rescigno, A. Atom- and Bond-Based 2D TOMOCOMD-CARDD Approach and Ligand-Based Virtual Screening for the Drug Discovery of New Tyrosinase Inhibitors. *Journal of Biomolecular Screening*. **2008**, *13* (10), 1014-1024. DOI: 10.1177/1087057108326078.
41. Afantitis, A.; Melagraki, G.; Tsoumanis, A.; Valsami-Jones, E.; Lynch, I. A nanoinformatics decision support tool for the virtual screening of gold nanoparticle cellular association using protein corona fingerprints. *Nanotoxicology*. **2018**, *12* (10), 1148-1165, Article. DOI: 10.1080/17435390.2018.1504998 Scopus.
42. Papadiamantis, A. G.; Jänes, J.; Voyiatzis, E.; Sikk, L.; Burk, J.; Burk, P.; Tsoumanis, A.; Ha, M. K.; Yoon, T. H.; Valsami-Jones, E.; et al. Predicting Cytotoxicity of Metal

- Oxide Nanoparticles Using Isalos Analytics Platform. *Nanomaterials*. **2020**, *10* (10), 2017. DOI: 10.3390/nano10102017.
43. Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. *Alternatives to Laboratory Animals*. **2005**, *33* (2), 1-19.
 44. Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR & Combinatorial Science*. **2007**, *26* (5), 694-701. DOI: 10.1002/qsar.200610151.
 45. Bian, L.; Sorescu, D. C.; Chen, L.; White, D. L.; Burkert, S. C.; Khalifa, Y.; Zhang, Z.; Sejdic, E.; Star, A. Machine-Learning Identification of the Sensing Descriptors Relevant in Molecular Interactions with Metal Nanoparticle-Decorated Nanotube Field-Effect Transistors. *ACS Appl Mater Interfaces*. **2019**, *11* (1), 1219-1227. DOI: 10.1021/acsami.8b15785.
 46. Alafeef, M.; Srivastava, I.; Pan, D. Machine Learning for Precision Breast Cancer Diagnosis and Prediction of the Nanoparticle Cellular Internalization. *ACS sensors*. **2020**, *5* (6), 1689-1698. DOI: 10.1021/acssensors.0c00329.
 47. Sun, B.; Fernandez, M.; Barnard, A. S. Machine Learning for Silver Nanoparticle Electron Transfer Property Prediction. *J Chem Inf Model*. **2017**, *57* (10), 2413-2423. DOI: 10.1021/acs.jcim.7b00272.
 48. Barnard, A. S.; Opletal, G. Predicting structure/property relationships in multi-dimensional nanoparticle data using t-distributed stochastic neighbour embedding and machine learning. *Nanoscale*. **2019**, *11* (48), 23165-23172. DOI: 10.1039/c9nr03940f.
 49. He, J.; He, C.; Zheng, C.; Wang, Q.; Ye, J. Plasmonic nanoparticle simulations and inverse design using machine learning. *Nanoscale*. **2019**, *11* (37), 17444-17459. DOI: 10.1039/c9nr03450a.
 50. Yan, T.; Sun, B.; Barnard, A. S. Predicting archetypal nanoparticle shapes using a combination of thermodynamic theory and machine learning. *Nanoscale*. **2018**, *10* (46), 21818-21826. DOI: 10.1039/c8nr07341d.
 51. Huberty, C. J.; Olejnik, S. *Applied MANOVA and discriminant analysis*; John Wiley & Sons, Inc., 2006.
 52. Sun, L.; Yang, H.; Cai, Y.; Li, W.; Liu, G.; Tang, Y. In Silico Prediction of Endocrine Disrupting Chemicals Using Single-Label and Multilabel Models. *J Chem Inf Model*. **2019**, *59* (3), 973-982. DOI: 10.1021/acs.jcim.8b00551.
 53. Heider, D.; Senge, R.; Cheng, W.; Hullermeier, E. Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. *Bioinformatics*. **2013**, *29* (16), 1946-1952. DOI: 10.1093/bioinformatics/btt331.
 54. Santana, R.; Zuluaga, R.; Gañán, P.; Arrasate, S.; Onieva, E.; González-Díaz, H. Predicting coated-nanoparticle drug release systems with perturbation-theory machine learning (PTML) models. *Nanoscale*. **2020**, *12* (25), 13471-13483, Article. DOI: 10.1039/d0nr01849j Scopus.
 55. Diez-Alarcia, R.; Yanez-Perez, V.; Muneta-Arrate, I.; Arrasate, S.; Lete, E.; Meana, J. J.; Gonzalez-Diaz, H. Big Data Challenges Targeting Proteins in GPCR Signaling Pathways; Combining PTML-ChEMBL Models and [(35S)]GTPgammaS Binding Assays. *ACS Chem Neurosci*. **2019**, *10* (11), 4476-4491. DOI: 10.1021/acchemneuro.9b00302.
 56. Duardo-Sanchez, A.; Munteanu, C. R.; Riera-Fernandez, P.; Lopez-Diaz, A.; Pazos, A.; Gonzalez-Diaz, H. Modeling Complex Metabolic Reactions, Ecological Systems, and Financial and Legal Networks with MIANN Models Based on Markov-Wiener Node Descriptors. *Journal of Chemical Information and Modeling*. **2014**, *54* (1), 16-29. DOI: 10.1021/ci400280n.

57. Martinez-Arzate, S. G.; Tenorio-Borroto, E.; Barbabosa Pliego, A.; Diaz-Albiter, H. M.; Vazquez-Chagoyan, J. C.; Gonzalez-Diaz, H. PTML Model for Proteome Mining of B-Cell Epitopes and Theoretical-Experimental Study of Bm86 Protein Sequences from Colima, Mexico. *J Proteome Res.* **2017**, *16* (11), 4093-4103. DOI: 10.1021/acs.jproteome.7b00477.
58. Quevedo-Tumaili, V. F.; Ortega-Tenezaca, B.; Gonzalez-Diaz, H. Chromosome Gene Orientation Inversion Networks (GOINs) of Plasmodium Proteome. *J Proteome Res.* **2018**, *17* (3), 1258-1268. DOI: 10.1021/acs.jproteome.7b00861.
59. Frank, E.; Hall, M. A.; Witten, I. H. *The WEKA workbench*; Morgan Kaufmann, 2016.
60. Hastie, T.; Tibshirani, R.; Friedman, J. H. *The elements of statistical learning: Data mining, inference, and prediction*; Springer open, 2008.
61. Chen, P.; Pan, C. Diabetes classification model based on boosting algorithms. *BMC Bioinformatics.* **2018**, *19* (1), 109. DOI: 10.1186/s12859-018-2090-9.
62. Lang, S.; Bravo-Marquez, F.; Beckham, C.; Hall, M.; Frank, E. Wekadeeplearning4j: A deep learning package for weka based on deeplearning4j. *Knowledge-Based Systems.* **2019**, *178*, 48-50. DOI: 10.1016/j.knosys.2019.04.013.
63. Quinlan, R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers, 1993.
64. Breiman, L. Random Forests. *Machine Learning.* **2001**, *45* (1), 5-32, journal article. DOI: 10.1023/a:1010933404324.
65. Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems technology.* **2011**, *2* (3), 1-27. DOI: 10.1145/1961189.1961199.
66. Keerthi, S. S.; Shevade, S. K.; Bhattacharyya, C.; Murthy, K. R. K. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation.* **2001**, *13* (3), 637-649. DOI: 10.1162/089976601300014493 %J Neural Computation (accessed 10/14/2021).
67. Aha, D. W.; Kibler, D.; Albert, M. K. Instance-based learning algorithms. *Machine Learning.* **1991**, *6* (1), 37-66. DOI: 10.1007/BF00153759.
68. Le Cessie, S.; Van Houwelingen, J. C. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C.* **1992**, *41* (1), 191-201. DOI: 10.2307/2347628.
69. Cooper, C. I.; Yao, D.; Sendorek, D. H.; Yamaguchi, T. N.; P'ng, C.; Houlahan, K. E.; Caloian, C.; Fraser, M.; Ellrott, K.; Margolin, A. A.; et al. Valection: design optimization for validation and verification studies. *BMC Bioinformatics.* **2018**, *19* (1), 339. DOI: 10.1186/s12859-018-2391-z.
70. Dieguez-Santana, K.; Pham-The, H.; Villegas-Aguilar, P. J.; Le-Thi-Thu, H.; Castillo-Garit, J. A.; Casañola-Martin, G. M. Prediction of acute toxicity of phenol derivatives using multiple linear regression approach for Tetrahymena pyriformis contaminant identification in a median-size database. *Chemosphere.* **2016**, *165*, 434-441. DOI: 10.1016/j.chemosphere.2016.09.041.
71. Mishra, R. K.; Garcia-Domenech, R.; Galvez, J. Getting discriminant functions of antibacterial activity from physicochemical and topological parameters. *J Chem Inf Comput Sci.* **2001**, *41* (2), 387-393. DOI: 10.1021/ci000303c.
72. Murcia-Soler, M.; Perez-Gimenez, F.; Garcia-March, F. J.; Salabert-Salvador, M. T.; Diaz-Villanueva, W.; Castro-Bleda, M. J.; Villanueva-Pareja, A. Artificial neural networks and linear discriminant analysis: a valuable combination in the selection of new antibacterial compounds. *J Chem Inf Comput Sci.* **2004**, *44* (3), 1031-1041. DOI: 10.1021/ci030340e.
73. Murcia-Soler, M.; Perez-Gimenez, F.; Garcia-March, F. J.; Salabert-Salvador, M. T.; Diaz-Villanueva, W.; Medina-Casamayor, P. Discrimination and selection of new

- potential antibacterial compounds using simple topological descriptors. *Journal of molecular graphics & modelling*. **2003**, 21 (5), 375-390. DOI: 10.1016/s1093-3263(02)00184-5.
74. Mut-Ronda, S.; Salabert-Salvador, M. T.; Duarte, M. J.; Anton-Fos, G. M. Search compounds with antimicrobial activity by applying molecular topology to selected quinolones. *Bioorganic & medicinal chemistry letters*. **2003**, 13 (16), 2699-2702. DOI: 10.1016/s0960-894x(03)00544-4.
 75. Garcia-Domenech, R.; de Julian-Ortiz, J. V. Antimicrobial activity characterization in a heterogeneous group of compounds. *J Chem Inf Comput Sci*. **1998**, 38 (3), 445-449. DOI: 10.1021/ci9702454.
 76. Molina, E.; Diaz, H. G.; Gonzalez, M. P.; Rodriguez, E.; Uriarte, E. Designing antibacterial compounds through a topological substructural approach. *J Chem Inf Comput Sci*. **2004**, 44 (2), 515-521. DOI: 10.1021/ci0342019.
 77. Cronin, M. T.; Aptula, A. O.; Dearden, J. C.; Duffy, J. C.; Netzeva, T. I.; Patel, H.; Rowe, P. H.; Schultz, T. W.; Worth, A. P.; Voutzoulidis, K.; et al. Structure-based classification of antibacterial activity. *J Chem Inf Comput Sci*. **2002**, 42 (4), 869-878. DOI: 10.1021/ci025501d.
 78. Gonzalez-Diaz, H.; Torres-Gomez, L. A.; Guevara, Y.; Almeida, M. S.; Molina, R.; Castanedo, N.; Santana, L.; Uriarte, E. Markovian chemicals "in silico" design (MARCH-INSIDE), a promising approach for computer-aided molecular design III: 2.5D indices for the discovery of antibacterials. *J Mol Model*. **2005**, 11 (2), 116-123. DOI: 10.1007/s00894-004-0228-3.
 79. Speck-Planche, A.; Kleandrova, V. V.; Cordeiro, M. N. New insights toward the discovery of antibacterial agents: multi-tasking QSBER model for the simultaneous prediction of anti-tuberculosis activity and toxicological profiles of drugs. *European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences*. **2013**, 48 (4-5), 812-818. DOI: 10.1016/j.ejps.2013.01.011 From Nlm.
 80. Speck-Planche, A.; Kleandrova, V. V.; Cordeiro, M. N. Chemoinformatics for rational discovery of safe antibacterial drugs: simultaneous predictions of biological activity against streptococci and toxicological profiles in laboratory animals. *Bioorg Med Chem*. **2013**, 21 (10), 2727-2732. DOI: 10.1016/j.bmc.2013.03.015.
 81. Speck-Planche, A.; Cordeiro, M. N. D. S. Simultaneous virtual prediction of anti-escherichia coli activities and admet profiles: A chemoinformatic complementary approach for high-throughput screening. *ACS combinatorial science*. **2014**, 16 (2), 78-84. DOI: 10.1021/co400115s.
 82. Kleandrova, V. V.; Ruso, J. M.; Speck-Planche, A.; Dias Soeiro Cordeiro, M. N. Enabling the Discovery and Virtual Screening of Potent and Safe Antimicrobial Peptides. Simultaneous Prediction of Antibacterial Activity and Cytotoxicity. *ACS combinatorial science*. **2016**, 18 (8), 490-498. DOI: 10.1021/acscombsci.6b00063.
 83. Speck-Planche, A.; Cordeiro, M. N. D. S. Enabling virtual screening of potent and safer antimicrobial agents against noma: Mtk-QSBER model for simultaneous prediction of antibacterial activities and ADMET properties. *Mini-Reviews in Medicinal Chemistry*. **2015**, 15 (3), 194-202. DOI: 10.2174/138955751503150312120519.
 84. Speck-Planche, A.; Kleandrova, V. V.; Ruso, J. M.; Cordeiro, M. N. First Multitarget Chemo-Bioinformatic Model To Enable the Discovery of Antibacterial Peptides against Multiple Gram-Positive Pathogens. *J Chem Inf Model*. **2016**, 56 (3), 588-598. DOI: 10.1021/acs.jcim.5b00630.

85. Ortega-Tenezaca, B.; Gonzalez-Diaz, H. IFPTML mapping of nanoparticle antibacterial activity vs. pathogen metabolic networks. *Nanoscale*. **2021**, *13* (2), 1318-1330. DOI: 10.1039/d0nr07588d.
86. Diéguez-Santana, K.; González-Díaz, H. Towards Machine Learning Discovery of Dual Antibacterial Drug-Nanoparticle Systems. *Nanoscale*. **2021**, *13*, 17854-17870. DOI: 10.1039/D1NR04178A.
87. Vergara-Galicia, J.; Prado-Prado, F. J.; Gonzalez-Diaz, H. Galvez-Markov network transferability indices: review of classic theory and new model for perturbations in metabolic reactions. *Current drug metabolism*. **2014**, *15* (5), 557-564. DOI: 10.2174/1389200215666140605125827.
88. Cho, H.; Kim, K. S. Escherichia coli OxyS RNA triggers cephalothin resistance by modulating the expression of CRP-associated genes. *Biochemical and biophysical research communications*. **2018**, *506* (1), 66-72. DOI: 10.1016/j.bbrc.2018.10.084.
89. Sader, H. S.; Rhomborg, P. R.; Fuhrmeister, A. S.; Mendes, R. E.; Flamm, R. K.; Jones, R. N. Antimicrobial Resistance Surveillance and New Drug Development. *Open forum infectious diseases*. **2019**, *6* (Suppl 1), S5-S13. DOI: 10.1093/ofid/ofy345.
90. Vernet, G.; Mary, C.; Altmann, D. M.; Doumbo, O.; Morpeth, S.; Bhutta, Z. A.; Klugman, K. P. Surveillance for antimicrobial drug resistance in under-resourced countries. *Emerging infectious diseases*. **2014**, *20* (3), 434-441. DOI: 10.3201/EID2003.121157.
91. Orth, J. D.; Conrad, T. M.; Na, J.; Lerman, J. A.; Nam, H.; Feist, A. M.; Palsson, B. O. A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Molecular systems biology*. **2011**, *7*, 535. DOI: 10.1038/msb.2011.65.
92. Nakashima, N.; Miyazaki, K. Bacterial cellular engineering by genome editing and gene silencing. *International journal of molecular sciences*. **2014**, *15* (2), 2773-2793. DOI: 10.3390/ijms15022773.
93. Armijo, L. M.; Wawrzyniec, S. J.; Kopciuch, M.; Brandt, Y. I.; Rivera, A. C.; Withers, N. J.; Cook, N. C.; Huber, D. L.; Monson, T. C.; Smyth, H. D. C.; et al. Antibacterial activity of iron oxide, iron nitride, and tobramycin conjugated nanoparticles against Pseudomonas aeruginosa biofilms. *Journal of Nanobiotechnology*. **2020**, *18* (1), 35. DOI: 10.1186/s12951-020-0588-6.
94. Burygin, G. L.; Khlebtsov, B. N.; Shantrokha, A. N.; Dykman, L. A.; Bogatyrev, V. A.; Khlebtsov, N. G. On the Enhanced Antibacterial Activity of Antibiotics Mixed with Gold Nanoparticles. *Nanoscale Res Lett*. **2009**, *4* (8), 794-801. DOI: 10.1007/s11671-009-9316-8 PubMed.
95. Djafari, J.; Marinho, C.; Santos, T.; Igrejas, G.; Torres, C.; Capelo, J. L.; Poeta, P.; Lodeiro, C.; Fernández-Lodeiro, J. New Synthesis of Gold- and Silver-Based Nano-Tetracycline Composites. *ChemistryOpen*. **2016**, *5* (3), 206-212. DOI: <https://doi.org/10.1002/open.201600016>.
96. Eleftheriadou, I.; Giannousi, K.; Protonotariou, E.; Skoura, L.; Arsenakis, M.; Dendrinou-Samara, C.; Sivropoulou, A. Cocktail of CuO, ZnO, or CuZn Nanoparticles and Antibiotics for Combating Multidrug-Resistant Pseudomonas aeruginosa via Efflux Pump Inhibition. *ACS Applied Nano Materials*. **2021**, *4* (9), 9799-9810. DOI: 10.1021/acsanm.1c02208.
97. Esmaili, A.; Ghobadianpour, S. Vancomycin loaded superparamagnetic MnFe₂O₄ nanoparticles coated with PEGylated chitosan to enhance antibacterial activity. *International Journal of Pharmaceutics*. **2016**, *501* (1), 326-330. DOI: 10.1016/j.ijpharm.2016.02.013.
98. Esmailou, M.; Zarrini, G.; Ahangarzadeh Rezaee, M.; Shahbazi Mojarrad, J.; Bahadori, A. Vancomycin Capped with Silver Nanoparticles as an Antibacterial Agent

- against Multi-Drug Resistance Bacteria. *Adv Pharm Bull.* **2017**, 7 (3), 479-483. DOI: 10.15171/apb.2017.058 PubMed.
99. Gu, H.; Ho, P. L.; Tong, E.; Wang, L.; Xu, B. Presenting Vancomycin on Nanoparticles to Enhance Antimicrobial Activities. *Nano Letters.* **2003**, 3 (9), 1261-1263. DOI: 10.1021/nl034396z.
 100. Huang, Y.; Gao, Q.; Li, X.; Gao, Y.; Han, H.; Jin, Q.; Yao, K.; Ji, J. Ofloxacin loaded MoS₂ nanoflakes for synergistic mild-temperature photothermal/antibiotic therapy with reduced drug resistance of bacteria. *Nano Research.* **2020**, 13 (9), 2340-2350. DOI: 10.1007/s12274-020-2853-2.
 101. Hwang, I. S.; Hwang, J. H.; Choi, H.; Kim, K. J.; Lee, D. G. Synergistic effects between silver nanoparticles and antibiotics and the mechanisms involved. *Journal of medical microbiology.* **2012**, 61 (Pt 12), 1719-1726. DOI: 10.1099/jmm.0.047100-0 From NLM.
 102. Lai, H.-Z.; Chen, W.-Y.; Wu, C.-Y.; Chen, Y.-C. Potent Antibacterial Nanoparticles for Pathogenic Bacteria. *ACS Applied Materials & Interfaces.* **2015**, 7 (3), 2046-2054. DOI: 10.1021/am507919m.
 103. Meeker, D. G.; Jenkins, S. V.; Miller, E. K.; Beenken, K. E.; Loughran, A. J.; Powless, A.; Muldoon, T. J.; Galanzha, E. I.; Zharov, V. P.; Smeltzer, M. S.; et al. Synergistic Photothermal and Antibiotic Killing of Biofilm-Associated Staphylococcus aureus Using Targeted Antibiotic-Loaded Gold Nanoconstructs. *ACS Infectious Diseases.* **2016**, 2 (4), 241-250. DOI: 10.1021/acsinfecdis.5b00117.
 104. Punjabi, K.; Mehta, S.; Chavan, R.; Chitalia, V.; Deogharkar, D.; Deshpande, S. Efficiency of Biosynthesized Silver and Zinc Nanoparticles Against Multi-Drug Resistant Pathogens. *Front Microbiol.* **2018**, 9 (2207), Original Research. DOI: 10.3389/fmicb.2018.02207.
 105. Saha, B.; Bhattacharya, J.; Mukherjee, A.; Ghosh, A.; Santra, C.; Dasgupta, A. K.; Karmakar, P. In Vitro Structural and Functional Evaluation of Gold Nanoparticles Conjugated Antibiotics. *Nanoscale Res Lett.* **2007**, 2 (12), 614-622. DOI: 10.1007/s11671-007-9104-2 PMC.
 106. Vazquez-Muñoz, R.; Meza-Villezas, A.; Fournier, P. G. J.; Soria-Castro, E.; Juarez-Moreno, K.; Gallego-Hernández, A. L.; Bogdanchikova, N.; Vazquez-Duhalt, R.; Huerta-Saquero, A. Enhancement of antibiotics antimicrobial activity due to the silver nanoparticles impact on the cell membrane. *PLoS one.* **2019**, 14 (11), e0224904-e0224904. DOI: 10.1371/journal.pone.0224904 PubMed.
 107. Wan, G.; Ruan, L.; Yin, Y.; Yang, T.; Ge, M.; Cheng, X. Effects of silver nanoparticles in combination with antibiotics on the resistant bacteria *Acinetobacter baumannii*. *International journal of nanomedicine.* **2016**, 11, 3789-3800. DOI: 10.2147/IJN.S104166 PubMed.
 108. Zendegani, E.; Dolatabadi, S. The Efficacy of Imipenem Conjugated with Synthesized Silver Nanoparticles Against *Acinetobacter baumannii* Clinical Isolates, Iran. *Biological trace element research.* **2020**, 197 (1), 330-340. DOI: 10.1007/s12011-019-01962-6.
 109. Shaker, M. A.; Shaaban, M. I. Formulation of carbapenems loaded gold nanoparticles to combat multi-antibiotic bacterial resistance: In vitro antibacterial study. *International Journal of Pharmaceutics.* **2017**, 525 (1), 71-84. DOI: 10.1016/j.ijpharm.2017.04.019.
 110. Roshmi, T.; Soumya, K. R.; Jyothis, M.; Radhakrishnan, E. K. Effect of biofabricated gold nanoparticle-based antibiotic conjugates on minimum inhibitory concentration of bacterial isolates of clinical origin. *Gold Bulletin.* **2015**, 48 (1), 63-71. DOI: 10.1007/s13404-015-0162-4.
 111. Shahbandeh, M.; Taati Moghadam, M.; Mirnejad, R.; Mirkalantari, S.; Mirzaei, M. The Efficacy of AgNO₃ Nanoparticles Alone and Conjugated with Imipenem for Combating

- Extensively Drug-Resistant *Pseudomonas aeruginosa*. *International journal of nanomedicine*. **2020**, *15*, 6905-6916. DOI: 10.2147/IJN.S260520.
112. Fan, Y.; Pauer, A. C.; Gonzales, A. A.; Fenniri, H. Enhanced antibiotic activity of ampicillin conjugated to gold nanoparticles on PEGylated rosette nanotubes. *International journal of nanomedicine*. **2019**, *14*, 7281-7289. DOI: 10.2147/ijn.S209756
From NLM.
 113. Payne, J. N.; Waghvani, H. K.; Connor, M. G.; Hamilton, W.; Tockstein, S.; Moolani, H.; Chavda, F.; Badwaik, V.; Lawrenz, M. B.; Dakshinamurthy, R. Novel Synthesis of Kanamycin Conjugated Gold Nanoparticles with Potent Antibacterial Activity. *Front. Microbiol.* **2016**, *7* (607), Original Research. DOI: 10.3389/fmicb.2016.00607.

CHAPTER 7. CONCLUDING REMARKS

1. CONCLUSIONS

The main contribution of this doctoral thesis is the derivation of several predictive models of the antibacterial activity of drugs in the design phase by applying perturbation theory (PT) combined with machine learning methods (ML) and the fusion of information (IF) from preclinical assays, chemical structures, nanoparticles, and variations of metabolic reaction networks of multiple microorganisms.

The main findings of the work performed during this PhD thesis are summarized as follows:

- The state of the art on bacterial resistance, major antibiotics, protein targets, mechanisms of action, databases of preclinical, clinical trials, and other sources of information useful for computational modeling were explored. In this critical analysis, machine learning techniques and performance evaluation metrics algorithms applied in the field of antibacterial drugs were compiled.
- A PTML computational model was built to study the connectivity (structure) of a metabolite in the metabolic reaction networks of a query organism. Analysis of the dataset included the number of nodes (metabolites), input-output links (metabolic reactions), node degree, topological indices, and the full names and codes of > 40 bacterial species.
- There is a low number of experimentally tested DADNP systems, but a high number of experimentally tested ADs and NPs, so additive IFPTML models may become a pragmatic solution for the time being when taking into account the greater abundance of experimental evidence for DADNP components in ADs and NPs alone.
- A chemo-computational methodology based on machine learning techniques with perturbation theory and information fusion was proposed that quantitatively related chemical and preclinical data from the ChEMBL database to metabolic network data. The linear and nonlinear IFPTML models of AD versus MN presented good statistical parameters. The IFPTML-LDA model presented specificity (S_p) of 76.1%, sensitivity (S_n) of 72.3%, and accuracy (A_c) of 74.3%. Among the IFPTML-nonlinear, the k Nearest Neighbors (KNN) showed the best results, with $S_n = 99.2\%$, $S_p = 95.5\%$, $A_c = 97.4\%$, and AUROC = 0.998 in the training sets.
- The first predictive model of the biological activity of antibacterial drugs functionalized with nanoparticle systems was constructed using the IFPTML method. This model included information on assay conditions and molecular descriptors. Different algorithms, such as Linear Discriminant Analysis (LDA), Artificial Neural Networks (ANN), Bayesian Networks (BNN), K -Nearest Neighbor (KNN), etc., were applied to find the model with the highest sensitivity, specificity, and accuracy, taking into account the complexity. We performed a simulation of the expected behavior of putative DADNPs in 72 different biological assays (> 1900 calculations) and tested the validity of the additive model with 80 experimentally synthesized and biologically tested DADNP complexes (reported in the literature). The IFPTML-LDA model correctly classified 100% of the DADNP complexes as biologically active. The IFPTML additive strategy may become a useful tool to aid in the design of DADNP systems for antibacterial therapy, taking into account only information about the AD and NP components separately.
- An analysis and mapping of DADNP (AD + NP) systems against MN of pathogenic bacterial species was developed by Information Fusion Machine Learning with Perturbation Theory (IFPTML) as a new application of AI/ML methods in the search for antibacterial drugs (AD) coping with the emergence of multidrug-resistant strains.

The Linear Discriminant Analysis (LDA) model of IFPTML with $Sp \approx 90\%$ and $Sn \approx 80\%$ and the best Artificial Neural Network (ANN) model found with $Sp \approx Sn \approx 95\%$ in the training and validation runs presented good results. This type of model could be useful for the discovery of new DADNP systems.

2. FUTURE WORKS

MLT contribute to drug discovery, are applied at different stages of development to accelerate the research process and reduce time and expenses. The various applications of ML in the antibacterial field can coordinate theoretical results such as chemical information, metabolic data of microorganisms and medical data to emerge as a tool for decision making. MLT (such as NN, SVM, DT, ensemble predictors, and Bayesian classifiers) could be used as predictive tools in chemoinformatic pipelines aimed at predicting the activity of unknown compounds and subsequently discovering new potential antibacterial agents. The predictive ability of models is determined by the inherent properties of each dataset, and the selection of the best MLT is related to its performance. ML approaches are increasingly opening new regions of chemical space for exploration. New techniques allow larger volumes of data to be processed and higher accuracy to be obtained. DL algorithms, RF and clustering methods have been gaining ground among ML techniques applied in AD discovery studies. However, in clinical trials, the application of ML is still limited even though there are varied sources of information that can generate absolute and methodological data to support decision making and deduction of risk failures in drug discovery. High numbers of antibacterial AD and NP experimentally tested are possibly the best opportunity at same time. In the face of the limitations of potential new antibiotic discovery and the global threat posed by antibiotic-resistant bacteria, strategies such as drug repurposing and combination antibacterial NP systems have emerged as promising approaches, although they present substantial obstacles to success. In the case of the former, to accelerate compounds in clinical studies, and in the latter, there remain many issues to be resolved in the safety of nanomaterials prior to clinical translation. Furthermore, in vivo toxicity and actual clinical effect must be attested at all times and with care. ML models may help, but the low number of real DADNP experimentally characterized complex applications. Additive IFPTML models may become a pragmatic solution, for the moment, by taking into consideration the higher abundance of experimental tests for DADNP components AD and NP alone.

Clearly, the remaining obstacles to the application of ML in antibacterial drug discovery must be removed or reduced. It is required to enhance research in new drug discovery and development processes. Access to robust and freely available data and increased collaboration between researchers and institutions.

Data availability, collaboration, and use of cloud-based web services: The prospects for ML-facilitated antibiotic discovery will depend in part on improved data. As broader data sources become publicly available, new ML questions can be raised and ongoing questions can be reviewed more rigorously. Although expanding public sources of experimental data will be crucial, collaboration between institutions can facilitate expansion of the empirical data set without private data sharing, as has been done in other areas of biomedical ML.¹ Increased data sharing of successful and

unsuccessful projects in the pharmaceutical industry has also been proposed as a means to accelerate research and development.²

The provision and sharing of data and models are being recognised as essential to improve the efficiency and transparency of research in various fields of drug discovery. Making them accessible using the FAIR (findable, accessible, interoperable, and reusable) principles is a cornerstone of open science.³ Making data and models available to all researchers increases the robustness of the science, enables model reuse, and expands the dataset of antibacterial drug data that can be used to train new models. One of the barriers to discovering new antibiotics is the lack of information sharing.⁴ For instance, The Pew Charitable Trusts launched SPARK: the Shared Platform for Antibiotic Research and Knowledge (<https://www.collaborativedrug.com/SPARK-data-downloads/>). SPARK is an online, publicly available, interactive database designed to help scientists build on previous research and generate new insights to advance the field's understanding of Gram-negative permeability. This viewpoint details how data is selected and integrated into the platform, how scientists can use SPARK to share their data, and the ways the scientific community can access and use this data to develop hypotheses.

Another case is the Global Antibiotic Research and Development Partnership (GARDP). It brings public and private partners together to speed the development and global availability of novel antibiotics to treat the most difficult drug-resistant bacterial diseases. GARDP's recently published strategy outlines its ambitious goal of delivering five novel medicines by 2025, focused on sexually transmitted infections, newborn sepsis, and infections in hospitalized adults and children.⁵ GARDP, in particular, has an outreach initiative called REVIVE (revive.gardp.org) to improve knowledge retention, and the antimicrobial R&D community is supported online to connect, share, and obtain information. Regular webinars, antimicrobial opinion pieces, and conference sessions are among REVIVE's initiatives.⁶

Drug re-purposing: Existing data can also be exploited for new purposes, as demonstrated by resources such as the Drug Repurposing Hub.⁷ In that sense, these resources such as drug repurposing have emerged as a drug discovery strategy in the face of difficult processes to find new antibacterial drugs in recent years. The review by Ananda Kumar et al.⁸ highlights the key role of drug repurposing in antibiotic development during 2016-2017 and addresses combination therapies with existing antibiotics. Additionally, they discuss the potential new implications of effectively combating multidrug-resistant (MDR) bacterial infections. A similar focus covers the review by Farha, et al.⁹ on repurposing existing drugs for antimicrobial purposes. In that paper, they discuss enabling screening platforms for antimicrobial discovery and present encouraging findings of new antimicrobial therapeutic strategies. The exploration of ¹⁰ with the use of DL methods to discover antibiotics from repurposing non-antibiotic pharmaceuticals was based on that approach.

New advances in nanomedicine: Recent advances in nanomedicine promise to be effective for pathogen treatments by enhancing the bactericidal capacity of antibiotics.¹¹ DADNP systems may be more cost-effective solutions due to their high tunability and broad adaptability to address different circumstances, including persistent cells in macrophages and biofilm infections.¹² These systems have been

investigated under controlled conditions with various nanoparticle complexes, mainly metallic (Ag and Au), and have been shown to have synergistic or additive effects compared to stand-alone AD or NPs, which makes them potentially promising for the treatment of MDR strains. Compared to traditional antibiotics, DADNP systems may be endowed with many functions, such as enhanced penetration, targeting, and absorption capabilities, change of the infectious microenvironment, and combination with other therapeutic strategies. However, several critical issues remain to be addressed in the application of strategies based on these systems, such as in vivo stability and long-term safety effects (in vivo toxicity and actual clinical effects should be attested throughout), a lack of standards for formulation delivery, scale-up feasibility, etc.¹² As a result, it will be necessary to increase long-term research and practice before large-scale application of nanoantibiotics for the treatment of resistant infections. It is expected that in the near future, nanoantibiotics will be able to fight resistant bacterial infections and save more lives.

Integration of the various approaches: Regarding future directions of employing MLT in antibiotic discovery, there is an urgent need to integrate the various approaches discussed. The use of available and shared drug data and information, computer-aided drug design, advanced MLT such as DL methods and other areas will play an important role in accelerating the urgent task of new antibiotic discovery. In addition, medicinal chemistry can use MLT to develop new antibacterial agents for numerous reasons/opportunities: the opportunity to exploit the ever-increasing available data; the increasing complexity of those data sets; the constant increase in computational power; and the advancement in algorithms and combined predictive pipelines. A drug design pipeline's first step could come from a vast undiscovered chemical space and must meet many ideal requirements. Generally, science and technology are continually renewing themselves. Regarding the first opportunity, multi-task or multi-objective QSAR research already exists and can be simply applied utilizing MLT.¹³⁻¹⁵ Multi-objective QSAR techniques predict many endpoints with the same model and are useful in multitarget drug design.^{16, 17}

3. REFERENCES

1. Sheller, M. J.; Edwards, B.; Reina, G. A.; Martin, J.; Pati, S.; Kotrotsou, A.; Milchenko, M.; Xu, W.; Marcus, D.; Colen, R. R.; et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*. **2020**, *10* (1), 12598. DOI: 10.1038/s41598-020-69250-1.
2. Kim, W.; Krause, K.; Zimmerman, Z.; Outtersson, K. Improving data sharing to increase the efficiency of antibiotic R&D. *Nature Reviews Drug Discovery*. **2021**, *20*, 1-2. DOI: 10.1038/d41573-020-00185-y.
3. Hodson, S.; Jones, S.; Collins, S.; Genova, F.; Harrower, N.; Laaksonen, L., ...; Wittenburg, P. *Turning FAIR into reality*; Final report and action plan from the EC expert group on FAIR data, 2018.
4. Thomas, J.; Navre, M.; Rubio, A.; Coukell, A. Shared Platform for Antibiotic Research and Knowledge: A Collaborative Tool to SPARK Antibiotic Discovery. *ACS Infectious Diseases*. **2018**, *4* (11), 1536-1539. DOI: 10.1021/acsinfecdis.8b00193.
5. Balasegaram, M.; Piddock, L. J. V. The Global Antibiotic Research and Development Partnership (GARDP) Not-for-Profit Model of Antibiotic Development. *ACS Infectious Diseases*. **2020**, *6* (6), 1295-1298. DOI: 10.1021/acsinfecdis.0c00101.

6. Pentz-Murr, A.; Piddock, L. J. V. Together towards a common goal: REVIVE, a community of antimicrobial researchers brought together by the Global Antibiotic Research & Development Partnership (GARDP). *Journal of Antimicrobial Chemotherapy*. **2019**, *74* (7), 1769-1770. DOI: 10.1093/jac/dkz077 %J Journal of Antimicrobial Chemotherapy (accessed 11/30/2021).
7. Corsello, S. M.; Bittker, J. A.; Liu, Z.; Gould, J.; McCarren, P.; Hirschman, J. E.; Johnston, S. E.; Vrcic, A.; Wong, B.; Khan, M.; et al. The Drug Repurposing Hub: A next-generation drug library and information resource. *Nature Medicine*. **2017**, *23* (4), 405-408, Letter. DOI: 10.1038/nm.4306 Scopus.
8. Ananda Kumar, K.; Grandhe Usha, R.; Kyeong, L.; Yongseok, C. Recent Drug-Repurposing-Driven Advances in the Discovery of Novel Antibiotics. *Current Medicinal Chemistry*. **2019**, *26* (28), 5363-5388. DOI: 10.2174/0929867325666180706101404.
9. Farha, M. A.; Brown, E. D. Drug repurposing for antimicrobial discovery. *Nature Microbiology*. **2019**, *4* (4), 565-577. DOI: 10.1038/s41564-019-0357-1.
10. Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackerman, Z.; et al. A Deep Learning Approach to Antibiotic Discovery. *Cell*. **2020**, *180* (4), 688-702.e613, Article. DOI: 10.1016/j.cell.2020.01.021 Scopus.
11. Eleraky, N. E.; Allam, A.; Hassan, S. B.; Omar, M. M. Nanomedicine Fight against Antibacterial Resistance: An Overview of the Recent Pharmaceutical Innovations. *Pharmaceutics*. **2020**, *12* (2), 142. DOI: 10.3390/pharmaceutics12020142 PubMed.
12. Wang, S.; Gao, Y.; Jin, Q.; Ji, J. Emerging antibacterial nanomedicine for enhanced antibiotic therapy. *Biomaterials Science*. **2020**, *8* (24), 6825-6839, 10.1039/D0BM00974A. DOI: 10.1039/D0BM00974A.
13. Cruz-Montegudo, M.; Cordeiro, M. N. D. S.; Tejera, E.; Dominguez, E. R.; Borges, F. Desirability-based multi-objective QSAR in drug discovery. *Mini-Reviews in Medicinal Chemistry*. **2012**, *12* (10), 920-935, Review. DOI: 10.2174/138955712802762329 Scopus.
14. Speck-Planche, A.; Kleandrova, V. V.; Cordeiro, M. N. D. S. New insights toward the discovery of antibacterial agents: Multi-tasking QSBER model for the simultaneous prediction of anti-tuberculosis activity and toxicological profiles of drugs. *European Journal of Pharmaceutical Sciences*. **2013**, *48* (4-5), 812-818. DOI: 10.1016/j.ejps.2013.01.011.
15. Speck-Planche, A.; Kleandrova, V. V.; Ruso, J. M.; Cordeiro, M. N. D. S. First Multitarget Chemo-Bioinformatic Model to Enable the Discovery of Antibacterial Peptides against Multiple Gram-Positive Pathogens. *Journal of Chemical Information and Modeling*. **2016**, *56* (3), 588-598. DOI: 10.1021/acs.jcim.5b00630.
16. González-Díaz, H.; Arrasate, S.; Gómez-San, A. J.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. M. General theory for multiple input-output perturbations in complex molecular systems. 1. linear QSPR electronegativity models in physical, organic, and medicinal chemistry. *Current Topics in Medicinal Chemistry*. **2013**, *13* (14), 1713-1741, Review. DOI: 10.2174/1568026611313140011 Scopus.
17. Liu, X.; Zhu, F.; Ma, X. H.; Shi, Z.; Yang, S. Y.; Wei, Y. Q.; Chen, Y. Z. Predicting targeted polypharmacology for drug repositioning and multi-target drug discovery. *Current Medicinal Chemistry*. **2013**, *20* (13), 1646-1661, Article. DOI: 10.2174/0929867311320130005 Scopus.

SCIENTIFIC OUTCOME

1. PUBLICATIONS

These research findings have been published in the following scientific articles:

1. **Paper I:** Diéguez-Santana, K., González-Díaz, H. Artificial Intelligence Discovery of new antibacterials: from classic drugs to dual nanoparticle-drug systems. Submitted to Chemical Society Reviews. ISSN 1460-4744. Q1 in Chemistry, Multidisciplinary. I.F:54.564
2. **Paper II:** Diéguez-Santana, K., Casañola-Martin, G. M., Green, J. R., Rasulev, B. & González-Díaz, H. Predicting Metabolic Reaction Networks with Perturbation-Theory Machine Learning (PTML) Models. Current Topics in Medicinal Chemistry. **2021**, 21(9), 819–827, Doi:10.2174/1568026621666210331161144. ISSN/eISSN: 1568-0266/1873-4294. Q2 in Medicinal Chemistry. I.F. 3.218.
3. **Paper III:** Diéguez-Santana, K., Casañola-Martin, G.M. Torres, R., Green, J. R., Rasulev, B. & González-Díaz, H. Machine Learning Mapping of Metabolic Networks vs. ChEMBL Data of Antibacterial Compounds. Submitted to ACS Molecular Pharmaceutics. **In process**. ISSN: 1543-8384. Q1 in Pharmacology & Pharmacy. IF: 4.939
4. **Paper IV:** Diéguez-Santana, K., González-Díaz, H. Towards Machine Learning Discovery of Dual Antibacterial Drug-Nanoparticle Systems. *Nanoscale*. **2021**, 13 (42), 17854-17870. Doi:10.1039/d1nr04178a. ISSN: 2040-3372. Q1 in Chemistry, Multidisciplinary/Materials Science, Multidisciplinary/ Nanoscience & Nanotechnology. I.F. 7.79.
5. **Paper V:** Diéguez-Santana, K., Rasulev, B. & González-Díaz, H. Towards Rational Nanomaterial Design by Prediction of Drug-Nanoparticle Systems Interaction vs. Bacteria Metabolic Networks. *Environmental Science: Nano*. **2022**. Doi: 10.1039/D1EN00967B. ISSN/eISSN: 2051-8153/2051-8161. Q1 in Chemistry, Multidisciplinary / Nanoscience & Nanotechnology / Environmental Sciences. I.F: 8.131

2. OTHER PUBLICATIONS

1. Castillo Garit, J. A., González Díaz, H., Cañizares Carmenate, Y., Torrens, F. Pham-The, H., Martínez López, Y., & Diéguez Santana, K. (2021). Aplicaciones y potencialidades de los métodos de diseño computacional en estudios ambientales y farmacocinéticos. *Anales de la Academia de Ciencias de Cuba*. **2021**. 11 (1). e811. ISSN 2304-0106. Disponible en: <http://www.revistaccuba.cu/index.php/revacc/article/view/811>
2. Diéguez-Santana, K., Puris, A., Rivera-Borroto, O. M., Pham-The, H., Le-Thi-Thu, H., Bakhtiyor Rasulev & Casañola-Martin, G. M. Beyond Model Interpretability using LDA and Decision Trees for α -Amylase and α -Glucosidase Inhibitor Classification Studies. *Chemical Biology & Drug Design*. **2019**. 94, 1414–1421. Doi:10.1111/cbdd.13518. ISSN:1747-0285. I.F: 2.328.
3. Dieguez-Santana, K., Puris, A., Rivera-Borroto, O. M., Casañola-Martin, G. M., González-Díaz. H., Fuzzy Approach to Dual inhibitors of α -amylase and α -glucosidase for the diabetes. Submitted to *Current Computer-Aided Drug Design*. ISSN: 1875-6697 (Online)/ISSN: 1573-4099 (Print). I. F: 1.606

4. Diéguez-Santana, K., Nachimba-Mayanchi, M. M., Puris, A., Torres Gutiérrez, R., González-Díaz, H. Prediction of acute toxicity of pesticides for *Americamysis bahia* using linear and nonlinear QSTR modelling approaches. Submitted to Environmental Research. ISSN: 0013-9351. I. F: 6.498.
5. Diéguez-Santana, K., Oviedo, B., Puris, A., Rivera-Borroto, O.M. Casañola-Martin, G.M., González-Díaz, H. ISIDA-ML Toxicity Prediction of phenol derivatives. Submitted to RSC Advances. Impact factor: 3.361. ISSN 2046-2069.

3. CONTRIBUTIONS TO CONFERENCES

1. Diéguez-Santana, K.; Puris Caceres, A. Y.; Rivera-Borroto, O. M.; Casañola-Martin, G. M. Dual inhibitors of α -amylase and α -glucosidase for the diabetes treatment: A fuzzy rules and machine learning approach. CHEMBIOMOL-06: Chem. Biol. & Med. Chem. Workshop, Bilbao-Rostock, Germany-Galveston, Texas, USA, **2020**.
2. Diéguez-Santana, K.; Casañola-Martin, G.M.; Green, J.R.; Rasulev, B. PTMLIF model of Metabolic Reaction Networks and ChEMBL Antibacterial Compounds. USEDAT-06: USA-Europe Data Analysis Training Program Workshop, Bilbao, Spain-Cambridge, UK-Miami, USA, **2020**.
3. Diéguez-Santana, K.; Puris Caceres, A.Y.; Rivera-Borroto, O.M.; Casañola-Martin, G.M. In silico toxicity prediction of phenol derivatives with ISIDA descriptors using multiple linear regression and machine learning approach. CHEMBIOINFO-06: Chem-Bioinformatics Congress, München, Germany-Chapel Hill, USA, **2020**.
4. Diéguez-Santana, K. Towards DADNP: Dual Antibacterial Drug-Nanoparticle Systems Machine learning Approach. NANOBIOIMAT-07: Nanotech. & Mat. Sci. Congress, Birmingham & Portsmouth, UK-Jackson & Fargo, USA, **2021**.
5. Diéguez-Santana, K.; Nachimba-Mayanchi, M. M. Machine learning-based prediction of toxicity of pesticide towards *Americamysis bahia*. MODECO-06: Molec. Diversity, Environ. Chem., and Economy Congress, Paris, France-Ohio, USA, **2021**.
6. Diéguez-Santana, K.; González-Díaz, H. IFPTML Study of Dual Antibacterial Drug-Nanoparticle (DADNP) Systems. NANOBIOIMAT-07: Nanotech. & Mat. Sci. Congress, Birmingham & Portsmouth, UK-Jackson & Fargo, USA, **2021**.
7. Diéguez-Santana, K.; Casañola-Martin, G. M.; Green, J. R.; Rasulev, B.; González-Díaz, H. Combinatorial Perturbation-Theory Machine Learning (CPTML) Models for Curation of Metabolic Reaction Networks. MODECO-06: Molec. Diversity, Environ. Chem., and Economy Congress, Paris, France-Ohio, USA, **2021**.
8. Rasulev, B.; Dieguez-Santana, K. Machine Learning Analysis of α -amylase Inhibitors. 07. NICXSM-07: North-Ibero-American Congress on Exp. & Simul. Methods, Valencia, Spain-Miami, USA, **2021**.

4. AWARDS

1. PREMIO NACIONAL DE LA ACADEMIA DE CIENCIAS DE CUBA. **2019**. Academia de Ciencias de Cuba. La Habana. Cuba. 15/04/2020. Title: Aplicaciones y potencialidades de los métodos de diseño computacional en estudios ambientales y farmacocinéticos. Authors: Castillo Garit, J. A.,

González Díaz, H., Cañizares Carmenate, Y., Torrens, F. Pham-The, H.,
Martínez López, Y., Diéguez Santana, K.