Combining Machine Learning models for improving their predictive quality and usefulness in biomedical applications

Tesis Doctoral

Doctorado en Biomedicina y Farmacia

Pablo Rodríguez Belenguer

Directores:

Dr. Manuel Pastor Maeso

Dr. Emilio Soria Olivas

Dr. Víctor Mangas Sanjuán

Universitat de València

Valencia - Noviembre 2023

VNIVERSITAT DE VALÈNCIA

Doctorado en Biomedicina y Farmacia

Dr. Manuel Pastor Maeso, Profesor Agregado del Departamento de Medicina y Ciencias de la Vida en la Universitat Pompeu Fabra, Dr. Emilio Soria Olivas, Catedrático del Departamento de Ingeniería Electrónica de la Universitat de Valencia, y Dr. Víctor Mangas Sanjuán, Profesor Titular del Departamento Farmacia y Tecnología Farmacéutica y Parasitología de la Universitat de Valencia,

CERTIFICAN:

Que el trabajo presentado por D. Pablo Rodríguez Belenguer, titulado "*Combining Machine Learning models for improving their predictive quality and usefulness in biomedical applications*", para obtener el grado de Doctor, ha sido realizado bajo nuestra dirección y asesoramiento.

Concluido el trabajo, autorizamos la presentación de la Tesis, para que sea juzgado por el tribunal correspondiente.

Lo que firmamos en Valencia a 2 de noviembre 2023

PASTOR MAESO MANUEL - 05394510K
Firmado digitalmente por PASTOR MAESO MANUEL - 05394510K
Fecha: 2023.11.02 19:15:42 +01'00'

Dr. Manuel Pastor Maeso

EMILIO |SORIA| OLIVAS
Firmado digitalmente por EMILIO|SORIA|OLIVAS
Fecha: 2023.11.02 19:27:21 +01'00'

Dr. Emilio Soria Olivas

MANGAS SANJUAN VICTOR - 48575410P
Firmado digitalmente por MANGAS SANJUAN VICTOR - 48575410P
Fecha: 2023.11.02 20:08:04 +01'00'

Dr. Víctor Mangas Sanjuán

Índice

Descripción General

Dentro de un cambio de paradigma hacia una toxicología mecanística, los modelos computacionales emergen como un poderoso complemento de los métodos experimentales. Sin embargo, es importante reconocer que los métodos experimentales, a pesar de su valor, enfrentan dificultades para abarcar todo el espectro de la biología. Debido a esta limitación, emplear un único modelo predictivo puede no ser suficiente. En su lugar, podría emplearse una combinación de modelos que representen fenómenos biológicos más simples.

La presente tesis se presenta mediante un compendio de publicaciones que pretenden desarrollar metodologías aplicables para la predicción de parámetros toxicológicos complejos a través de combinación de modelos, así como el estudio de la variabilidad asociada a dichas predicciones.

Rodríguez-Belenguer, P., March-Vila, E., Pastor, M., Mangas-Sanjuan, V., Soria-Olivas, E. (2023). Usage of Model Combination in Computational Toxicology. *Toxicology Letters*. https://doi.org/10.1016/j.toxlet.2023.10.013

En este trabajo revisamos las diversas formas en que se han combinado modelos computacionales en la literatura para abordar problemas toxicológicos prácticos. Consideramos que este enfoque (el uso de múltiples modelos combinados) es una estrategia interesante que podría generalizarse. Para ello proporcionamos una taxonomía de situaciones y directrices prácticas, ilustradas con numerosos ejemplos.

Rodríguez-Belenguer, P., Mangas-Sanjuan, V., Soria-Olivas, E., & Pastor, M. (2023). Integrating Mechanistic and Toxicokinetic Information in Predictive

Models of Cholestasis. ***Journal of Chemical Information and Modeling*****.** **https://doi.org/10.1021/acs.jcim.3c00945**

Este trabajo pretende proveer una metodología alternativa a los modelos directos QSAR para la predicción de colestasis, especialmente en los casos en los que se pretende predecir un nuevo compuesto que difiere significativamente de los de la serie de entrenamiento. Para ello es fundamental la incorporación de información mecanística, la cual se realiza a través de la combinación de múltiples modelos QSAR que representan fenómenos biológicos más simples. Además, se integra con información toxicocinética. Los resultados de esta metodología revelan un poder predictivo superior en comparación con el modelado directo QSAR en los escenarios de máxima disimilitud estructural entre nuevos compuestos con respecto a los de la serie de entrenamiento.

Rodríguez-Belenguer, P., Kopańska, K., Llopis-Lorente, J., Trenor, B., Saiz, J., & Pastor, M. (2023). Application of Machine Learning to improve the efficiency of electrophysiological simulations used for the prediction of drug-induced ventricular arrhythmia. ***Computer Methods and Programs in Biomedicine*****, 107345. https://doi.org/10.1016/j.cmpb.2023.107345**

En este trabajo se desarrollaron modelos multinivel como otra alternativa de modelado en situaciones de complejidad biológica. Para ello, se combinó el efecto de bloquear tres canales iónicos para producir arritmia ventricular, utilizando un complejo modelo electrofisiológico para predecir el biomarcador que representa la duración del potencial de acción al 90% de la repolarización (APD_{90}). Dado el alto coste computacional asociado con la obtención de las matrices electrofisiológicas, se evaluó a través de diferentes métricas, el resultado de predecir los valores de APD_{90} a partir del uso de modelos multinivel en diferentes muestreos regulares. El objetivo principal de este enfoque fue el de reducir el número de simulaciones necesarias para la

obtención de las matrices electrofisiológicas. Como resultado, esta metodología logró reducir 100 veces los tiempos de simulación.

Kopańska, K., Rodríguez-Belenguer, P., Llopis-Lorente, J., Trenor, B., Saiz, J., & Pastor, M. (2023). Uncertainty assessment of proarrhythmia predictions derived from multi-level in silico models. ***Archives of Toxicology*****,** ***97*****(10), 2721-2740. https://doi.org/10.1007/s00204-023-03557-6**

En el campo de la toxicología, las predicciones desempeñan un papel fundamental en la toma de decisiones y, por ende, es esencial garantizar la máxima confiabilidad de dichas predicciones. Por ello, en este trabajo se ofrece una representación más realista de las predicciones de los biomarcadores de proarritmia, derivadas del modelo multinivel del artículo anterior. Esto se consigue a través del estudio del impacto de las fuentes aleatorias de variabilidad (experimental e interindividual), tanto de manera individual como en un enfoque conjunto sobre las predicciones. Cabe destacar que el efecto de considerar simultáneamente ambos tipos de variabilidad no resultó ser aditivo y varió según el fármaco estudiado.

En definitiva, esta tesis contribuye a avanzar en la aplicación de la combinación de modelos para abordar la predicción de parámetros toxicológicos complejos y promueve una mayor comprensión de la variabilidad asociada a estas predicciones, para así dar lugar a una toma de decisiones más realista. Por todo ello, creemos que esta contribución puede ser útil para la comunidad toxicológica, proveyendo una metodología alternativa al modelado directo QSAR.

Agradecimientos

Probablemente esta sección sea la más injusta de todas ya que se corre el riesgo de no mencionar a personas que han contribuido a tu desarrollo profesional y personal, así que espero que nadie se sienta menospreciado, pero para no hacerlo más largo quiero incluir a las personas principales.

Primero de todo y centrándome en lo profesional, quiero agradecer a mi supervisor Emilio Soria Olivas por su apoyo incondicional desde el primer día que le conocí hasta el día de hoy. Siempre está dispuesto a escuchar a cualquiera con tal de ayudarle, y a mí lo hizo en el momento en el que más lo necesitaba en mi vida profesional. También agradecer a mi supervisor principal Manuel Pastor la paciencia que ha tenido cada día de su supervisión conmigo. Apostaste por mi sin dudarlo y eso es algo que jamás olvidaré. He disfrutado como un niño cada día de estos tres años trabajando en este magnífico equipo. Además, me has dado desde el principio confianza y tranquilidad para trabajar, lo que ha propiciado que estos tres años hayan sido los mejores de mi carrera profesional. También agradecer a mi supervisor Víctor Mangas Sanjuán por todo su apoyo y consejos, lo cierto es que siempre has estado ahí cuando los he necesitado. Quiero tener también palabras de agradecimiento hacia todos mis compañeros de laboratorio que desde el primer día me acogieron como a uno más: Giacomo, Eric, Tati, Adrián y, por último, Karo. ¡Cuánto nos hemos divertido y discutido juntos! Quiero decirte, que me has hecho ser un científico más completo y superarme día a día.

Respecto a familia y amigos, agradecer a mis padres y a mi hermana, que desde siempre han comprendido mi amor por la ciencia, aunque en ocasiones,

conllevara renunciar a momentos juntos. Gracias por escucharme cada día en los buenos y en los malos momentos. ¡Gracias por estar siempre ahí, no sé qué haría sin vosotros!

Agradecer a mi abuela Maruja (yayi) que me guía en el camino cada vez que me siento perdido. Agradecer a mi mejor amigo Pablo su apoyo incondicional, que tal y como decía mi abuela, eres la familia que se elige. Agradecer a toda la familia Ruso-Julve que desde el primer día me ha tratado como a uno más de la familia. Por último, agradecer a Candela, mi mujer, por ser esa persona que me lleva aguantando diecisiete años, la persona que siempre está ahí, la persona que también me ha compartido con la ciencia sin casi rechistar, la persona con la cual hemos creado lo más importante, a Leo. Hijo, muchas gracias por darnos tanto en tan poco tiempo, literalmente tú mueves nuestro mundo.

Gracias, y un millón de gracias a todos vosotros ya que hacéis que mi vida sea más fácil.

Listado de publicaciones

La presente tesis doctoral está basada en los siguientes manuscritos:

1. **Usage of model combination in computational toxicology**. Pablo Rodríguez-Belenguer, Eric March-Vila, Manuel Pastor, Victor Mangas-Sanjuan, Emilio Soria-Olivas. doi: https://doi.org/10.1016/j.toxlet.2023.10.013. Toxicology Letters. IF: 3.5, Q2 (Toxicology).
2. **Integrating Mechanistic and Toxicokinetic Information in Predictive Models of Cholestasis**. Pablo Rodríguez-Belenguer, Victor Mangas-Sanjuan, Emilio Soria-Olivas, Manuel Pastor. doi: 10.1021/acs.jcim.3c00945. Journal of Chemical Information and Modeling. IF 5.6, Q1 (Medicinal Chemistry, Computer Science).
3. **Application of machine learning to improve the efficiency of electrophysiological simulations used for the prediction of drug-induced ventricular arrhythmia**. Pablo Rodríguez-Belenguer, Karolina Kopańska, Jordi Llopis-Lorente, Beatriz Trenor, Javier Saiz, Manuel Pastor. doi: 10.1016/j.cmpb.2023.107345. Computer Methods and Programs in Biomedicine, IF 6.1, Q1 (Computer Science, Biomedical Engineering, Medical Informatics).
4. **Uncertainty assessment of proarrhythmia predictions derived from multi-level in silico models**. Karolina Kopańska, Pablo Rodríguez-Belenguer, Jordi Llopis-Lorente, Beatriz Trenor, Javier Saiz, Manuel Pastor. doi: 10.1007/s00204-023-03557-6. Archives of Toxicology. IF 6.1, Q1 (Toxicology).

Listado de abreviaturas

ADME: Absorption, Distribution, Metabolism, Excretion

AE: Adverse Event

AOP: Adverse Outcome Pathways

BCRP: Breast Cancer Resistance Protein

BSEP: Bile Salt Export Pump

CiPA: Comprehensive In Vitro Proarrhythmia Assay

CNN: Convolutional Neural Network

CPU: Central Processing Unit

DL: Deep Learning

DNN: Deep Neural Network

GNN: Graph Neural Network

IC_{50}: Concentración inhibitoria semimáxima

IST: *in silico* toxicology

KE: Key Event

KER: Key Event Relationship

LLM: Low-Level Models

LSTM: Long Short Term Memory

MIE: Molecular Initiating Event

ML: Machine Learning

MLP: Multi Layer Perceptron

MRE: Mean Relative Error

MRP: Multi Drug Resistence Protein

NAM: New Approach Methodology

NB: Naïve Bayes

NLDE: Non-Large Data-Points Error

OATP: Organinc Anion Transporting Polypeptides

PBPK: Physiologically Based Pharmacokinetic

P-gp: P-glycoprotein

PR: Polynomial transformation with Ridge regression

QIVIVE: Quantitative *In Vitro* to *In Vivo* Extrapolations

QSAR: Quantitative Structure-Activity Relationship

RA: Read-Across

RE: Relative Error

RF: Random Forest

SVM: Support Vector Machine

TK: Toxicokinetics

XGB: XGBoost

Introducción

La toxicología, históricamente, ha dependido en gran medida de experimentos en animales para evaluar la seguridad de compuestos químicos y fármacos. Esta aproximación observacional, si bien ha sido valiosa, presenta limitaciones éticas y científicas que han llevado a la búsqueda de enfoques alternativos para la evaluación de la toxicidad (Fischer et al., 2020).

En los últimos años se han hecho esfuerzos significativos para reducir, refinar y reemplazar (principio de las 3Rs) las pruebas en animales con metodologías de nuevo enfoque (New Approach Methodologies, NAMs) (Russell & Burch, 1960). Estas incluyen ensayos *in vitro*, que utilizan cultivos celulares o de tejidos para evaluar la toxicidad; pruebas *in chemico*, que se basan en análisis químicos y no requieren material biológico; y métodos *in silico*, que utilizan simulaciones computacionales y modelos matemáticos para predecir la toxicidad de un compuesto (https://www.epa.gov/). Este cambio de paradigma hacia los NAMs fue descrito en el informe "Toxicity Testing in the 21st Century: A Vision and a Strategy" por parte de la National Academy of Sciences y el National Research Council de los Estados Unidos en 2007 (Council, 2007). Puede considerarse que este informe estableció las bases para una nueva era en la evaluación de la toxicidad, basada en la comprensión de los mecanismos de toxicidad, que promueve el uso de métodos más éticos y eficaces para caracterizar la seguridad química en seres humanos.

Métodos en toxicología computacional

Los métodos *in silico* o *in silico* toxicology (IST) más comúnmente empleados se clasifican en cuatro grupos (Figura 1): extrapolación (Read-Across, RA), alertas estructurales o toxicóforos, modelos predictivos basados en la

caracterización de la relación cuantitativa estructura-actividad (Quantitative Structure-Activity Relationship, QSAR), y acoplamiento molecular o *docking*. Los tres primeros métodos se fundamentan en conocimiento y se basan en la teoría del bioisosterismo en la que "estructuras muy similares tienen bioactividades muy parecidas" (Johnson y Maggiora 1990). RA busca inferir las actividades biológicas de compuestos para los que se carece de información a partir de compuestos con una gran similitud estructural. A través de sistemas expertos, las alertas estructurales identifican grupos funcionales o subestructuras que han sido asociados con la aparición de efectos adversos (*Adverse Events, AEs*).

Los modelos QSAR son una de las técnicas más empleadas en IST. Estos modelos tienen la capacidad de predecir la actividad biológica (incluyendo AEs) de un compuesto a partir de su estructura química. Para ello, se emplean algoritmos de aprendizaje automático (Machine Learning, ML) y aprendizaje profundo (Deep Learning, DL), los cuales tienen la capacidad de predecir propiedades como Absorción, Distribución, Metabolismo, y Excreción (ADME) y/o toxicológicas. En el ámbito del ML, algunos de los algoritmos más comúnmente empleados son Random Forest (RF), XGBoost (XGB), Naïve Bayes (NB) y Support Vector Machine (SVM). En la literatura, hay multitud de ejemplos de aplicación de algoritmos de ML en toxicología, como es el caso del trabajo realizado por Ishfaq et al. (2022), en el que los autores construyeron modelos para predecir la actividad biológica de los inhibidores de la aromatasa. Del mismo modo, Trinh et al. (2022) emplearon modelos de *bagging* para predecir la toxicidad de nanomezclas de Ti_2O producidas en *Daphnia magna*.

Por su parte, en el DL destacan algoritmos como Deep Neural Network (DNN), Convolutional Neural Network (CNN), Graph Neural Network (GNN) y Long Short Term Memory (LSTM). Cada vez son más el número de trabajos que

utilizan DL en toxicología, como es el caso del trabajo publicado por Romano, Hao, y Moore (2022), en el cual utilizaron GNNs para predecir diferentes parámetros toxicológicos. Del mismo modo, Ulfa et al. (2021) emplearon otro tipo de redes neuronales, como es el caso de una combinación de convoluciones 1D con LSTM, para la predicción de la actividad biológica de unos compuestos químicos.

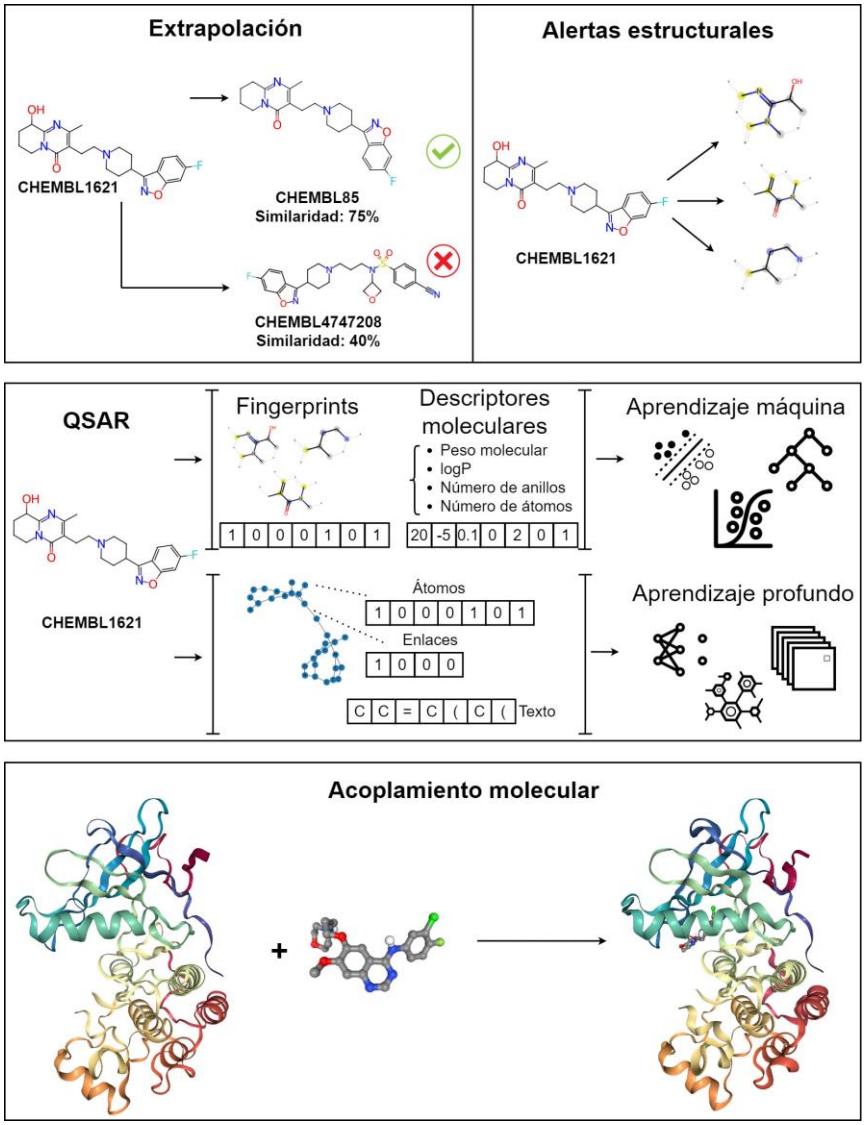


Figura 1: Esquema de los principales métodos empleados en toxicología computacional

En contraste con los algoritmos de ML, el DL presenta dos ventajas significativas. Primero, se destaca por la flexibilidad inherente de las estructuras basadas en redes neuronales, lo que le permite adaptarse a una amplia gama de problemas. Segundo, a diferencia de los algoritmos de ML, las redes neuronales tienen la capacidad de aprender y extraer automáticamente las características más relevantes de los datos, eliminando así la necesidad de realizar una selección manual de características. Por el contrario, los algoritmos de DL necesitan muchos más compuestos (varios miles) que los modelos de ML para así evitar el sobreajuste (Dargan et al. 2020).

Para construir los modelos QSAR, la estructura de los compuestos de la serie de entrenamiento debe ser descrita mediante un conjunto de variables (descriptores moleculares) que suelen incluir características fisicoquímicas (p. ej. peso molecular, solubilidad o número de átomos aceptores de hidrógeno) y/o características estructurales (por ejemplo, huellas digitales moleculares o *fingerprints*). La variable dependiente representa la actividad biológica, es decir, el efecto tóxico que se busca predecir. Dependiendo de la naturaleza de la variable dependiente el modelo puede ser de regresión, utilizado para predecir, por ejemplo, la concentración necesaria para inhibir el 50% de la actividad biológica de un compuesto (IC_{50}), o de clasificación, empleado para predecir, por ejemplo, si un compuesto es, o no, carcinógeno.

Finalmente, el acoplamiento molecular o *docking* es una técnica empleada para predecir cómo un ligando interacciona con un receptor en función de su estructura tridimensional. Esto se logra mediante la búsqueda de la conformación o posición relativa óptima del ligando dentro del receptor, con el objetivo de minimizar la energía libre de unión y maximizar la afinidad. En toxicología, esta técnica es importante cuando la toxicidad de un fármaco se origina debido a su interacción con una anti-diana (*antitarget*). Un ejemplo

clásico es el caso del receptor glucocorticoide, cuya inhibición puede ocasionar daños en el sistema inmune.

Principales limitaciones y complejidades presentes en la toxicología computacional

En los últimos tiempos, los métodos *in silico* han demostrado ser eficaces y convenientes para la evaluación de la toxicidad de compuestos químicos. Estos enfoques permiten realizar predicciones rápidas y precisas sin la necesidad de llevar a cabo ensayos en animales costosos y éticamente cuestionables (Council, 2007). Los métodos *in silico* son ideales para complementar otros tipos de ensayos en una estrategia secuencial, como métodos de cribado rápido, complementando métodos *in vitro* e *in vivo* (Raies & Bajic, 2016).

Sin embargo, es importante recordar que los métodos computacionales no pueden escapar de las complejidades inherentes a cualquier problema toxicológico. En general, estas complejidades se derivan de la complejidad de los procesos biológicos y químicos de los organismos, lo que obliga a los métodos experimentales y computacionales a introducir simplificaciones y suposiciones que limitan su potencial para representar la realidad. De hecho, los métodos *in vitro*, aunque éticos y efectivos, tienen una serie de limitaciones entre las cuales la más destacada es su dificultad para cubrir todos los fenómenos biológicos. Otra limitación de los métodos *in vitro* es la variabilidad asociada a sus experimentos, tales como las diferencias en las condiciones de laboratorio, la calidad de los reactivos, o las diferencias entre lotes de células o tejidos, lo que puede afectar la reproducibilidad y la precisión de los datos (Kernik et al., 2019). Estas limitaciones afectan directamente a los métodos computacionales, ya que consumen la información generada por los métodos experimentales. A raíz de todo lo expuesto, hemos identificado tres categorías

principales de complejidades, que están relacionadas con los siguientes aspectos: biológico, espacio químico, y metodológico.

Las complejidades biológicas suelen estar relacionadas con la existencia de múltiples mecanismos, muchos de ellos desconocidos, que pueden conducir a los mismos EAs. Las complejidades relacionadas con el espacio químico surgen cuando se intenta evaluar un compuesto químico que difiere significativamente en su estructura de los compuestos utilizados para construir el modelo, lo que a menudo resulta en predicciones poco precisas. Por último, las complejidades metodológicas se presentan cuando un solo algoritmo no puede establecer una relación adecuada entre los descriptores moleculares y la variable dependiente del modelo, cuando un solo tipo de variables no es suficiente para describir el problema toxicológico, o cuando el desequilibrio en la distribución de clases afecta negativamente el rendimiento final del modelo.

Una posible vía para mitigar algunos tipos de complejidades es la combinación de modelos QSAR de bajo nivel (Bringezu, Carlos Gómez-Tamayo, y Pastor 2021; Gadaleta et al. 2018; Heyndrickx et al. 2022; Kotsampasakou y Ecker 2017; March-Vila et al. 2023). A esta combinación de modelos predictivos se le conoce con el nombre de metamodelo. Esta aproximación tiene ventajas adicionales, porque ayuda a solventar la alta dependencia de los modelos directos QSAR con respecto a la estructura de los compuestos. Este problema está muy presente en el desarrollo de nuevos fármacos, donde en muchas ocasiones los candidatos a fármaco difieren significativamente de los compuestos utilizados para entrenar el modelo. Por tanto, en lugar de utilizar un modelo directo QSAR, se construyen modelos individuales que representen fenómenos más simples, con una mejor capacidad predictiva, y que, al ser combinados, pueden resolver de una manera más precisa el problema en cuestión. Esta estrategia busca superar las limitaciones inherentes de los modelos QSAR tradicionales, los cuales tienen dificultades para predecir

fenómenos biológicos complejos, para así permitir una consideración más amplia de la biología subyacente a los eventos toxicológicos.

Asimismo, esta metodología posee el potencial de ser empleada en la evaluación temprana de las propiedades toxicológicas de los candidatos a fármacos. No obstante, es importante destacar que su utilización en la toma de decisiones debe estar condicionada a la capacidad de estimar la incertidumbre de las predicciones. Esto se debe a que los resultados en toxicología son empleados para informar científicamente de decisiones (Gosling 2019; Maertens et al. 2022), ya sea dentro de las empresas o por parte de agencias reguladoras. Por ende, para poder tomar decisiones es esencial caracterizar la incertidumbre asociada a las predicciones.

Objetivos

El **objetivo general** de la tesis es:

Desarrollar, y validar una metodología general aplicable para la predicción de propiedades biológicas complejas, que presente ventajas en términos de calidad predictiva, así como en la estimación de la incertidumbre asociada a las predicciones con respecto a los métodos de referencia disponibles.

Los **objetivos específicos** para conseguir tal meta son:

1. Revisar las distintas estrategias de combinación de modelos de ML para abordar la predicción de parámetros biológicos complejos.
2. Evaluar el poder predictivo del modelado mecanístico en comparación con el modelado directo QSAR en condiciones de máxima disimilitud estructural en situaciones de aplicación relevante.
3. Evaluar la mejora en el poder predictivo del modelado mecanístico al incorporar información toxicocinética (Toxicokinetics, TK).
4. Desarrollar una metodología que permita identificar, caracterizar y cuantificar la variabilidad asociada a las predicciones obtenidas por la combinación de modelos.

Resultados y discusión

Complejidad biológica

La idea de combinar múltiples modelos, cada uno de los cuales representa mecanismos más simples, con el fin de mejorar la precisión de un modelo directo QSAR, cuenta con múltiples antecedentes que respaldan esta hipótesis. En el Capítulo 1 de nuestro trabajo de revisión (Rodríguez-Belenguer et al., 2023a), hemos identificado los principales tipos de metamodelos. La tabla 1 es la esencia del artículo dado que se resumen los diversos tipos de problemas encontrados en el ámbito de la toxicología, se analiza la complejidad que estos problemas generan, se explica por qué la construcción de un modelo directo QSAR podría no ser adecuada y se presenta el razonamiento detrás del uso de un metamodelo, junto con ejemplos ilustrativos. De entre los tres tipos de complejidades previamente mencionadas (biológica, espacio químico y metodológica), hemos decidido centrar esta tesis en la complejidad biológica, pese a que, en la revisión, hemos profundizado en todas ellas.

En un evento adverso hay presente una compleja red de fenómenos interconectados entre sí. En tales casos, un modelo directo QSAR puede tener dificultades para predecir el resultado conjunto de todos los fenómenos implicados, y la combinación de modelos de bajo nivel (Low-Level Models, LLM) se presenta como una estrategia prometedora para mejorar la calidad predictiva. Cada uno de estos LLM representa fenómenos biológicos más simples dentro de una red compleja, lo que contribuye a una representación más sencilla. En el contexto de los LLM, las rutas de eventos adversos (Adverse Outcome Pathways, AOP) emergen como una fuente de información mecanística idónea, de la cual puede extraerse información para la

identificación de las dianas a modelar y de sus interacciones. Aunque es importante destacar que los AOPs (Ankley et al., 2010) no fueron diseñados con este propósito específico, proporcionan un entorno transparente, accesible y estructurado que facilita la incorporación de información mecanística en los modelos gracias a AOPwiki (https://aopwiki.org/). Los AOPs conectan eventos moleculares iniciadores (Molecular Initiating Events, MIE) con eventos adversos (Adverse Outcome, AO) a través de una cadena causal de eventos clave (Key Events, KE), conectados mediante relaciones bien definidas (Key Event Relationships, KER).

Un metamodelo basado en una red de AOPs puede construirse integrando predicciones de multiples MIEs para predecir un parámetro toxicológico que describa el evento adverso. Para ello, se construyen modelos QSAR individuales correspondientes a cada MIE identificado, y se usa la predicción de estos eventos para construir un modelo de alto nivel que las relacione con las anotaciones biológicas del efecto adverso.

Todos los trabajos revisados coincidían en que la combinación de MIEs para la predicción de diferentes parámetros toxicológicos producía mejores resultados que los modelos directos QSAR (Gadaleta et al. 2018, 2022; Kleinstreuer et al. 2018; Kotsampasakou y Ecker 2017). Sin embargo, la mayoría de los estudios revisados empleaban datos *in vitro* para predecir resultados *in vivo*, ignorando el efecto de los procesos farmacocinéticos de las propiedades ADME en la predicción de la exposición. Esto introduce inconsistencias al intentar predecir datos *in vivo* desde variables puramente *in vitro*. Para abordar este problema una posible solución es utilizar modelos cuantitativos de extrapolación de datos *in vitro* a *in vivo* (Quantitative *In* Vitro to *In* Vivo Extrapolations, QIVIVE) (Punt et al. 2021), basados en un modelo farmacocinético basado en la fisiología (Physiologically Based Pharmacokinetic, PBPK). Este modelo simula eficazmente el comportamiento

longitudinal de una sustancia en un organismo, considerando fenómenos farmacocinéticos cruciales a lo largo del tiempo. El uso de modelos QIVIVE permite extrapolar las concentraciones *in vitro* a dosis *in vivo*. Por lo tanto, estos modelos proporcionan una herramienta valiosa para mejorar la precisión y relevancia de los metamodelos, avanzando en última instancia en nuestra comprensión del comportamiento de los compuestos en organismos vivos.

Integrando la TK sobre un metamodelo con múltiples MIEs para predecir colestasis

En virtud de lo anterior, en nuestro artículo (Rodríguez-Belenguer et al. 2023b) perteneciente al Capítulo 2, seleccionamos la colestasis inducida por fármacos como la adversidad a predecir. La colestasis es un evento adverso dosis dependiente caracterizado por una interrupción del flujo biliar, lo que conduce al aumento de las concentraciones de ácidos biliares hepáticos, pudiendo provocar necrosis y/o apoptosis hepática (Padda et al. 2011). El principal mecanismo es la inhibición de transportadores hepáticos encargados de facilitar el flujo de bilis desde el hígado hasta el intestino delgado. A pesar de que la bomba de exportación de sales biliares (Bile Salt Export Pump, BSEP) parece ser el principal MIE, no es el único transportador implicado. Por lo tanto, al construir un modelo *in silico* para predecir la colestasis, es fundamental considerar la contribución de otros transportadores que también podrían desempeñar un papel importante como MIEs. Entre ellos se encuentran las proteínas asociadas a la resistencia a múltiples fármacos (Multi Drug Resistence Protein, MRP2, MRP3 y MRP4), la proteína de resistencia al cáncer de mama (Breast Cancer Resistance Protein, BCRP), la glicoproteína-P (P-glycoprotein, P-gp) y los polipéptidos transportadores aniónicos (Organic Anion Transporting Polypeptides [OATP1B1 y OATP1B3]). Kotsampasakou y Ecker (2017) demostraron que la colestasis es un parámetro toxicológico lo

suficientemente complejo como para requerir enfoques diferentes al modelado directo QSAR. Esto se debe a que la colestasis, tal y como hemos visto, involucra numerosos mecanismos biológicos subyacentes, y los modelos directos QSAR probablemente tendrían una baja capacidad predictiva, siendo muy dependientes de las estructuras químicas, al no capturar adecuadamente la esencia de cada mecanismo. Por ello, el objetivo de este estudio fue desarrollar una metodología alternativa que mejore la calidad predictiva del modelado directo QSAR a través de la incorporación de información mecanística y TK, con el fin de superar su alta dependencia estructural.

La principal novedad de nuestro trabajo es la integración de información mecanística con toxicocinética, lo que permite la construcción de un metamodelo que compara las dosis *in vivo* obtenidas de los modelos QIVIVE con las dosis terapéuticas. De esta manera, este modelo aborda tanto el riesgo como la exposición, ofreciendo una perspectiva más completa y precisa de la colestasis inducida por fármacos.

Para determinar si esta metodología aporta ventajas en términos de calidad predictiva en comparación con los modelos directos QSAR, se llevó a cabo una evaluación utilizando diversas métricas. Se comparó el metamodelo que incorpora información TK con aquel que solo utiliza datos *in vitro* y con los modelos directos QSAR. Esta evaluación se realizó en situaciones de máxima disimilitud estructural, con el propósito de simular escenarios comunes en el descubrimiento de fármacos, donde se busca predecir la toxicidad de nuevos compuestos que difieren significativamente en estructura de los ya existentes en el mercado. Para ello, se utilizó un enfoque de validación cruzada (CV) con 20 repeticiones y 5 folds (20-Repeated 5-fold CV), cuyos resultados se compararon con los de un enfoque de validación cruzada basada en la semejanza estructural (Similarity 5-fold CV). Por lo que, si los resultados de cualquiera de los modelos evaluados mediante la CV basada en similitud son

menos robustos (Figura 2), esto indica una dependencia estructural de dicho modelo. Este mismo procedimiento se llevó a cabo también utilizando grupos de códigos anatomo-terapéuticos-químicos (Anatomical Therapeutic Chemical, ATC) (Figura 3) para evaluar si los modelos eran aplicables a compuestos con distintas propiedades farmacológicas.

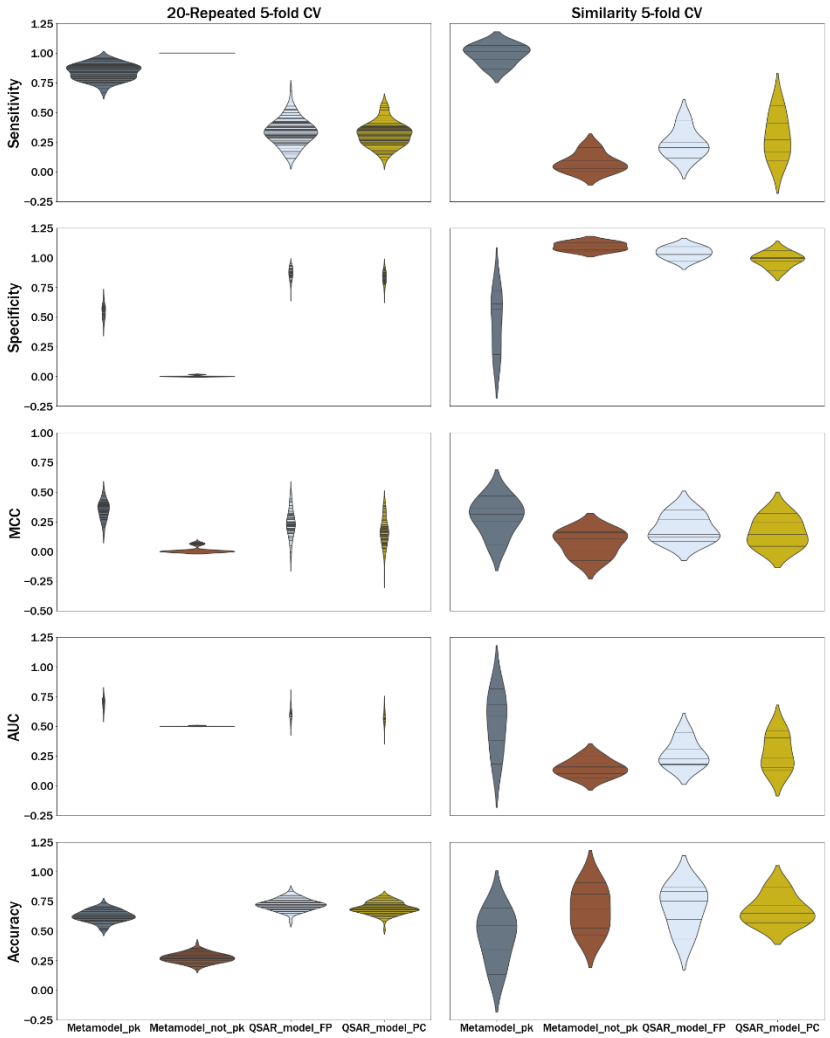


Figura 2: Diagrama de violín para diferentes métricas seleccionadas para evaluar la abstracción estructural de la metodología propuesta a través de una validación cruzada basada en semejanza estructural.

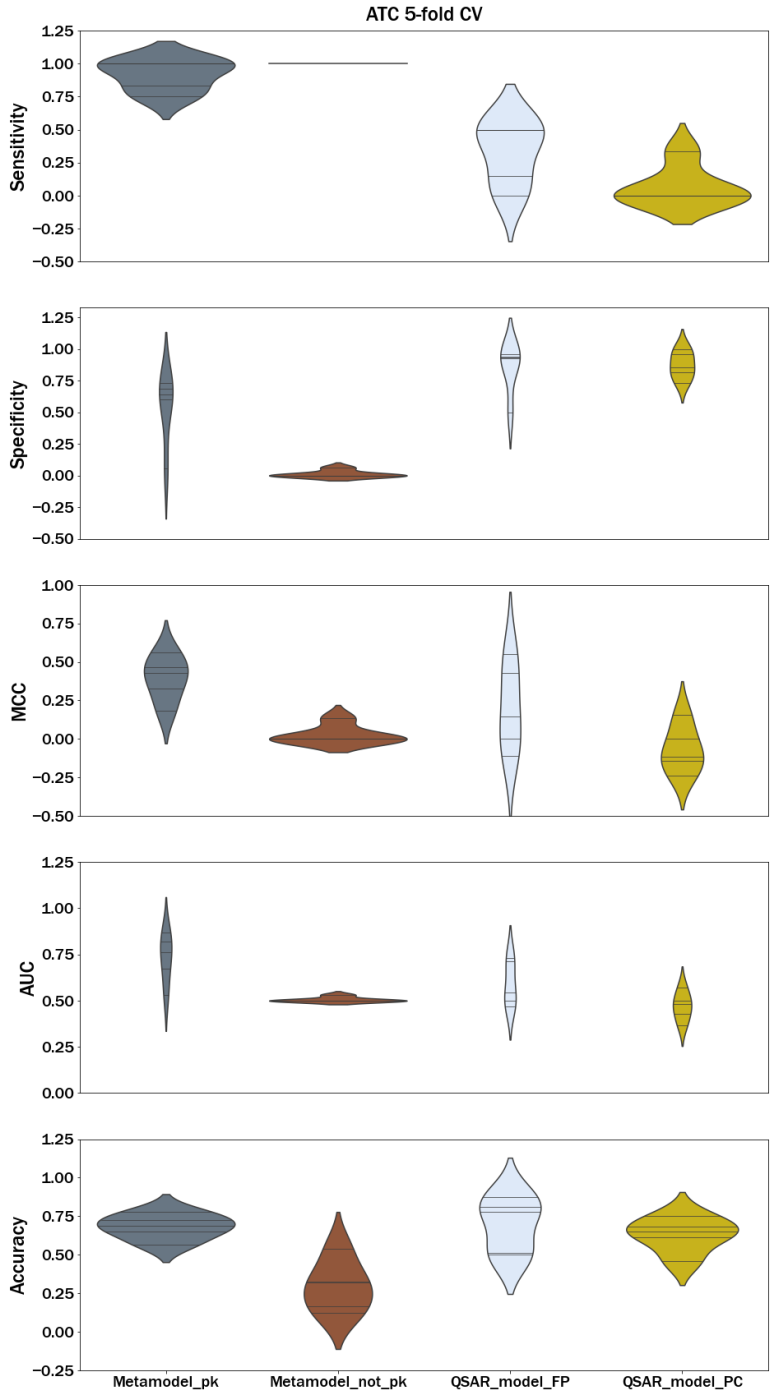


Figura 3: Diagramas de violín para diferentes métricas seleccionadas para evaluar la abstracción estructural de la metodología propuesta a través de una validación cruzada basada en códigos ATC.

El metamodelo incorporando información farmacocinética (Metamodel_pk) resultó ser el más sensible, con una sensibilidad superior al 80%, y con una especificidad superior al 50% (Figuras 2 y 3). Este resultado se mantuvo constante tanto en condiciones de evaluación normales a través de un 20-Repeated 5-fold Cross Validation, como en situaciones de máxima disimilitud estructural y farmacológica (Similarity 5-fold Cross Validation y 5 ATC-fold CV).

Asimismo, el Metamodel_pk resultó ser un modelo con un poder predictivo mucho mayor que el metamodelo que no incorpora información farmacocinética (Metamodel_not_pk), lo que resalta la importancia de considerar la farmacocinética en este tipo de estrategias. Los modelos directos QSAR, tanto los que utilizan fingerprints [fp] como descriptores fisicoquímicos [PC], resultaron ser altamente específicos pero muy poco sensibles. La sensibilidad de ambos modelos QSAR se vio reducida en condiciones de máxima disimilitud estructural y farmacológica, lo que reafirma su alta dependencia con respecto de las estructuras. Dado que, nos enfrentamos a un desequilibrio de clases en favor de la clase negativa, en estos casos, casi siempre suele aportar más valor un modelo con un equilibrio adecuado entre sensibilidad y especificidad, pero priorizando una sensibilidad más alta como es el caso del metamodelo que aporta información farmacocinética.

Por tanto, en este estudio la combinación de múltiples fenómenos biológicos más simples (MIEs) y la incorporación de información TK a través de modelos QIVIVE, produjo un rendimiento predictivo superior en comparación con el uso de modelos directos QSAR, especialmente en los casos de máxima disimilitud. Estos resultados sugieren que la metodología podría aplicarse en otros parámetros toxicológicos complejos, así como tener un potencial uso en la evaluación de riesgos al considerar la exposición y el riesgo en el metamodelo propuesto.

Modelos multinivel de arritmia ventricular

La combinación de múltiples MIEs no es la única opción de incorporar información mecanística en un metamodelo que trata de resolver la complejidad biológica. Es por ello que en el Capítulo 3, también hemos dirigido nuestra atención hacia una familia de modelos más complejos, llamados modelos multinivel (Rodríguez-Belenguer et al. 2023c).

En este trabajo, se desarrollaron modelos multinivel al combinar el efecto del bloqueo de tres canales iónicos para producir arritmia ventricular mediante un complejo modelo electrofisiológico. Esta aproximación se podría considerar más mecanística en comparación con la predicción a través de modelos directos QSAR, ya que se basa en el conocimiento del mecanismo por el cual los fármacos inducen arritmias ventriculares, afectando a la conductancia iónica que regula el potencial de membrana de los cardiomiocitos (Bartos, Grandi, y Ripplinger 2015).

El modelo electrofisiológico que se usa para predecir biomarcadores de arritmia ventricular a partir de las alteraciones de la conductancia requiere simulaciones computacionales muy complejas, lo que lo hace tedioso y no interactivo. Para abordar este problema, se pueden utilizar matrices de simulaciones precalculadas, lo que permite un cálculo instantáneo de biomarcadores como la duración del potencial de acción al 90% de la repolarización (APD_{90}). Sin embargo, la preparación de estas matrices (para ello usamos una versión modificada del modelo de O'Hara (O'Hara et al. 2011)) puede ser costosa en términos computacionales para los desarrolladores de métodos, lo que limita el alcance de las condiciones simuladas. Asimismo, es importante tener en cuenta que para proporcionar una descripción más completa de los mecanismos celulares de las arritmias inducidas por fármacos, la iniciativa Comprehensive In Vitro Proarrhythmia Assay (CiPA) propuso un

nuevo paradigma de pruebas en el cual la idea principal es utilizar los efectos de los fármacos medidos *in vitro* en múltiples canales iónicos (I_{Na}, I_{NaL}, I_{Kr}, I_{to}, I_{CaL}, I_{K1}, y I_{Ks}), en lugar de depender únicamente de I_{Kr}. En este sentido, se requieren estrategias para acortar los tiempos de simulación y, de esta forma, poder incluir un mayor número de canales iónicos.

Una simulación individual, que implica 500 latidos para un único conjunto de valores de entrada para tres canales iónicos, tarda alrededor de dos minutos y medio por CPU (Central Processing Unit). En nuestro caso, las matrices eran de 56*56*56 puntos (uno por cada canal iónico), lo que significa que, utilizando 32 CPUs, el tiempo necesario para obtener las matrices electrofisiológicas de nuestros modelos de arritmia ascendía a 56·56·56·2.5/32=13 720 minutos (228.7 horas). Por lo tanto, la implementación de estas matrices con una combinación de canales iónicos superior a tres (tal y como propone CiPA), se convierte en una limitación en sí misma, ya que el tiempo de cálculo se incrementa exponencialmente.

Por todo lo anterior, el objetivo principal de este trabajo fue reducir los tiempos necesarios para obtener las matrices electrofisiológicas, para ello se llevaron a cabo diferentes muestreos regulares en los que se evaluaron la calidad de los modelos, a través de métricas como el Error Relativo Medio en % (*Mean Relative Error,* MRE) y el porcentaje de datos con un Error Relativo (Relative Error*,* RE) inferior al 5% (*Non-Large Data-Points Error*, NLDE) para diversas aproximaciones de ML (transformación polinómica con regresión de Ridge [PR], SVM, y perceptrón multicapa [Multi-Layer Perceptron, MLP]). De esta manera, pudimos determinar la frecuencia con la que era necesario construir estas matrices con la certeza de que la información no utilizada no resultaba necesaria. Los modelos fueron validados en diferentes particiones, y a través de un conjunto de datos externo conteniendo 12 fármacos propuestos

por la iniciativa CiPA, los cuales presentan unas propiedades electrofisiológicas bien conocidas.

Los resultados obtenidos permitieron demostrar que, para este problema, el modelo que mejores resultados obtuvo fue SVM con un muestreo de uno cada cien puntos. En esta situación el MRE en test no superó el 0.20% (Figura 4) y no hubo ningún dato con un RE superior al 5% (Figura 5).

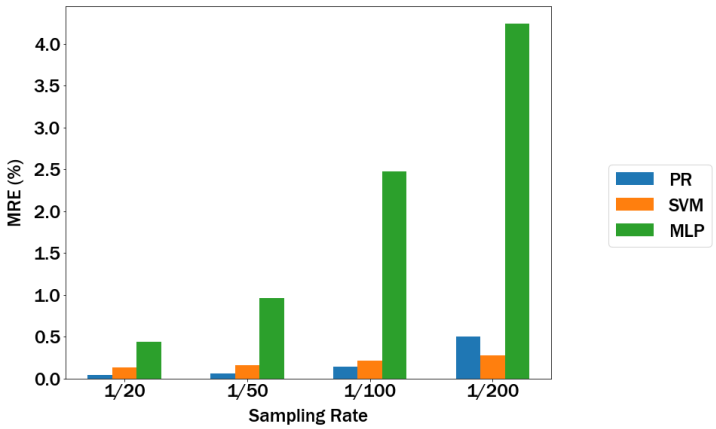


Figura 4: MRE(%) para los diferentes muestreos y modelos evaluados.

En la validación externa que utiliza los 12 fármacos propuestos por la iniciativa CiPA, el RE máximo fue prácticamente despreciable, de 1.5%, lo que supone un error de 4ms en la determinación del APD_{90}. En términos prácticos, esto implicaría, por ejemplo, un cambio en el valor del APD_{90} de 200 ms a 204 ms, sin que esto tenga ningún impacto en la consideración del riesgo arritmogénico.

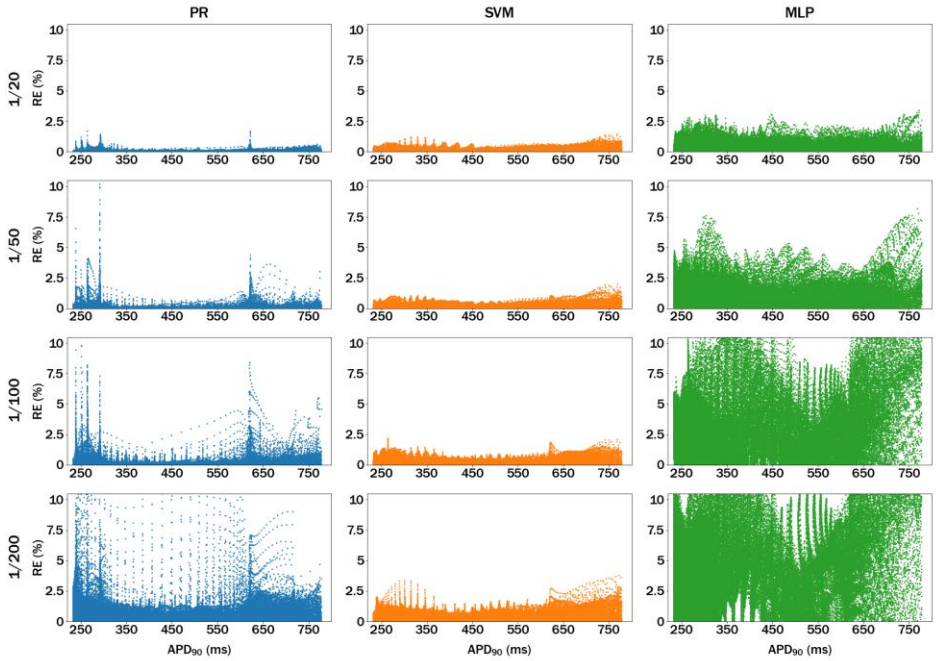


Figura 5: RE (%) en función de los valores experimentales de APD_{90}. Las columnas representan los tres modelos entrenados PR, SVM, y MLP. Las filas se corresponden con los diferentes ratios de muestreo evaluados.

Por lo tanto, se consiguió reducir de manera significativa la cantidad de simulaciones requeridas para efectuar predicciones precisas de biomarcadores de arritmia ventricular mediante la implementación de modelos multinivel con algoritmos de ML. Hemos evidenciado que la cantidad total de datos inicialmente simulados puede disminuir hasta un 1% de los utilizados hasta ahora, lo que implica una reducción sustancial en el tiempo de cálculo, pasando de las 228.7 horas originales a aproximadamente 2.29 horas. Este enfoque abre la posibilidad de modelar procesos biológicos más complejos, como aquellos que involucran cuatro o más canales iónicos.

Variabilidad asociada a las predicciones de modelos multinivel

A pesar de que los enfoques computacionales son una valiosa incorporación a los métodos puramente experimentales, es esencial realizar una exhaustiva evaluación de la variabilidad asociada a las predicciones con el fin de mejorar la confiabilidad de los métodos *in silico* (Gosling 2019).

En el pasado, se han propuesto diversas metodologías para caracterizar la variabilidad observada en los experimentos *in vitro* que miden el bloqueo de los canales iónicos debido a productos químicos (Elkins et al. 2013; Kramer et al. 2020; Li et al. 2017; Mirams et al. 2014). Por otro lado, la variabilidad interindividual asociada a los pacientes representa una fuente de variación relevante, consecuencia de múltiples factores vinculados con las características individuales de los pacientes. Al emplear enfoques *in silico*, los modelos electrofisiológicos que incorporan la IC_{50} específica de cada canal iónico en los biomarcadores de arritmia ventricular, utilizan numerosos parámetros que se ajustan para adaptarse a los resultados experimentales. Sin embargo, dado que los seres humanos no somos fisiológicamente idénticos, ningún modelo electrofisiológico puede producir resultados que representen adecuadamente a todos los pacientes ni explicar con precisión las diferencias observadas entre nosotros (Wisniowska, Tylutki, y Polak 2017). Por ello, los enfoques poblacionales se han descrito como una estrategia útil para considerar la variabilidad interindividual en los parámetros de los modelos *in silico*.

En el Capítulo 4, en nuestro trabajo (Kopańska y Rodríguez-Belenguer et al. 2023), hemos abordado la caracterización de la variabilidad asociada en los modelos multinivel. Para ello, se ha identificado la incertidumbre aleatoria. Así como, desarrollado métodos para caracterizar y propagar (via simulaciones de Monte-Carlo) este tipo de variabilidad seleccionada. Finalmente, se ha cuantificado la variabilidad presente en los resultados finales de los modelos multinivel mencionados anteriormente. Por todo ello, los objetivos de este

trabajo son, ofrecer una representación más realista de las predicciones de los biomarcadores de proarritmia, así como permitir el estudio del impacto de fuentes aleatorias de variabilidad, tanto individualmente como en conjunto, sobre las predicciones.

La incertidumbre aleatoria se debe a la variabilidad intrínseca y extrínseca, junto con errores de medición, que se utilizan para analizar las asociaciones con las entradas del modelo. Estos elementos se resumen como "variabilidad experimental" (Simulación A) y "variabilidad interindividual" (Simulación B), afectando los valores de IC_{50} y los parámetros predefinidos en los modelos de simulación de potencial de acción electrofisiológico (Simulación C es una combinación de ambas).

Al comparar las distribuciones de las tres simulaciones representadas en la Figura 6, que corresponden a los 12 compuestos de CiPA, se evidencian notables diferencias en términos de su amplitud y asimetría. En la Simulación A, se introdujeron valores aleatorios con una media de 0 y una desviación estándar de 0.5 en los valores de IC_{50} para generar las entradas del modelo. Por lo tanto, la forma y la anchura de estas distribuciones no están directamente influenciadas por las suposiciones utilizadas para caracterizar este tipo de variabilidad.

En la Simulación B, a diferencia de la Simulación A, la dispersión y la forma de las distribuciones, sí se deben a las suposiciones realizadas sobre la variabilidad interindividual. Por ello, la suma de números aleatorios distribuidos normalmente a los valores de salida en la Simulación B resulta en distribuciones de APD_{90} con un histograma normal y sin diferencias notables en el ancho.

Al combinar ambos tipos de variabilidad en la Simulación C, las distribuciones son bastante similares a las obtenidas en la Simulación B, pero con una

dispersión ligeramente mayor y cierta asimetría. Es crucial destacar que el efecto de considerar ambos tipos de variabilidad simultáneamente no es aditivo y varía según el fármaco en estudio.

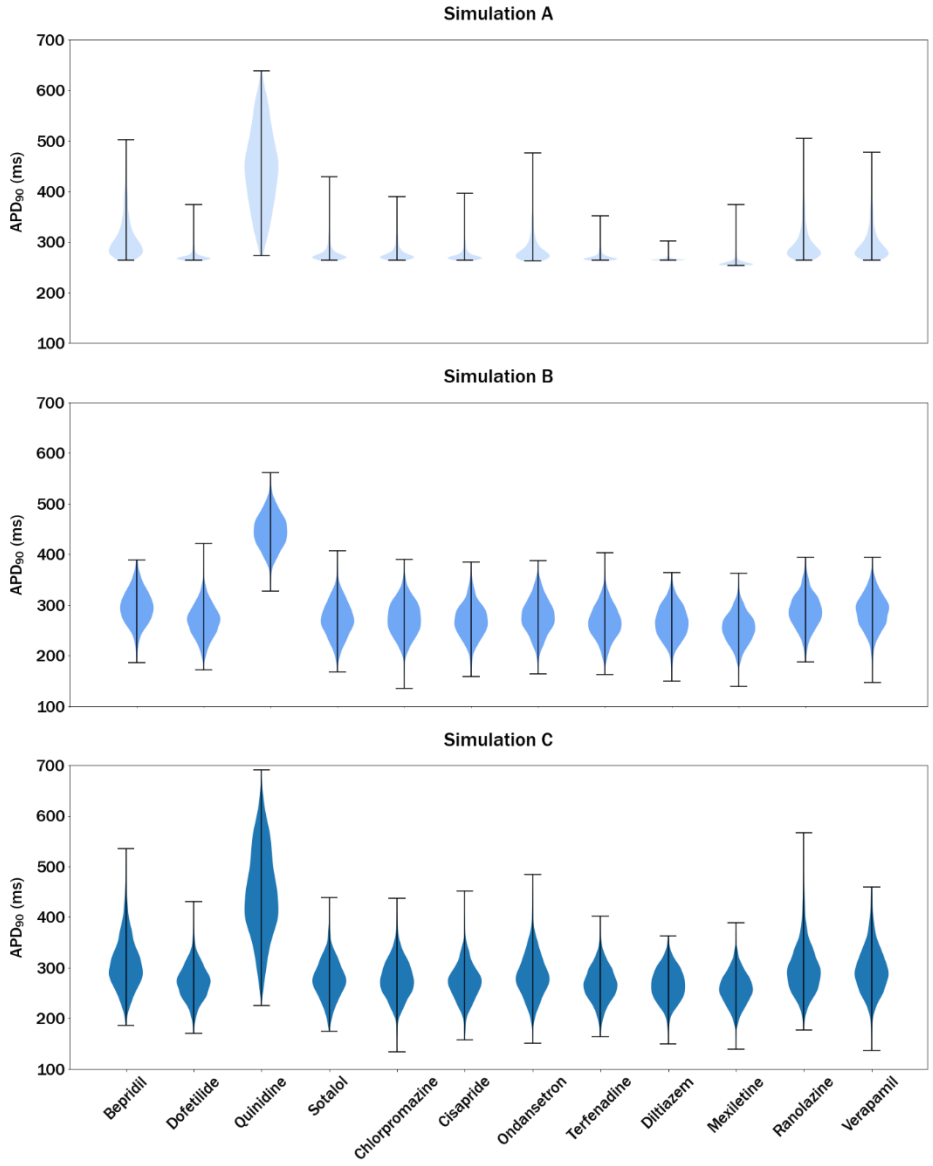


Figura 6: Gráficos de violín mostrando las distribuciones de los valores de APD_{90} obtenidos en diferentes simulaciones de Monte-Carlo introduciendo los siguientes tipos de variabilidad: Simulación A:Variabilidad experimental (Δ-pIC_{50}); Simulación B: Variabilidad interindividual (Δ-Parámetros); Simulación C: Combinación de variabilidad experimental e interindividual.

En cuanto a los gráficos de barras de la Figura 7 (utilizados para determinar los percentiles 10 y 90), no muestran grandes diferencias en las predicciones de APD_{90} generadas en las tres simulaciones realizadas para el mismo fármaco. Esto sugiere que la predicción real, calculada como el valor mediano del APD_{90}, apenas se ve afectada por el tipo de simulación y se mantiene constante.

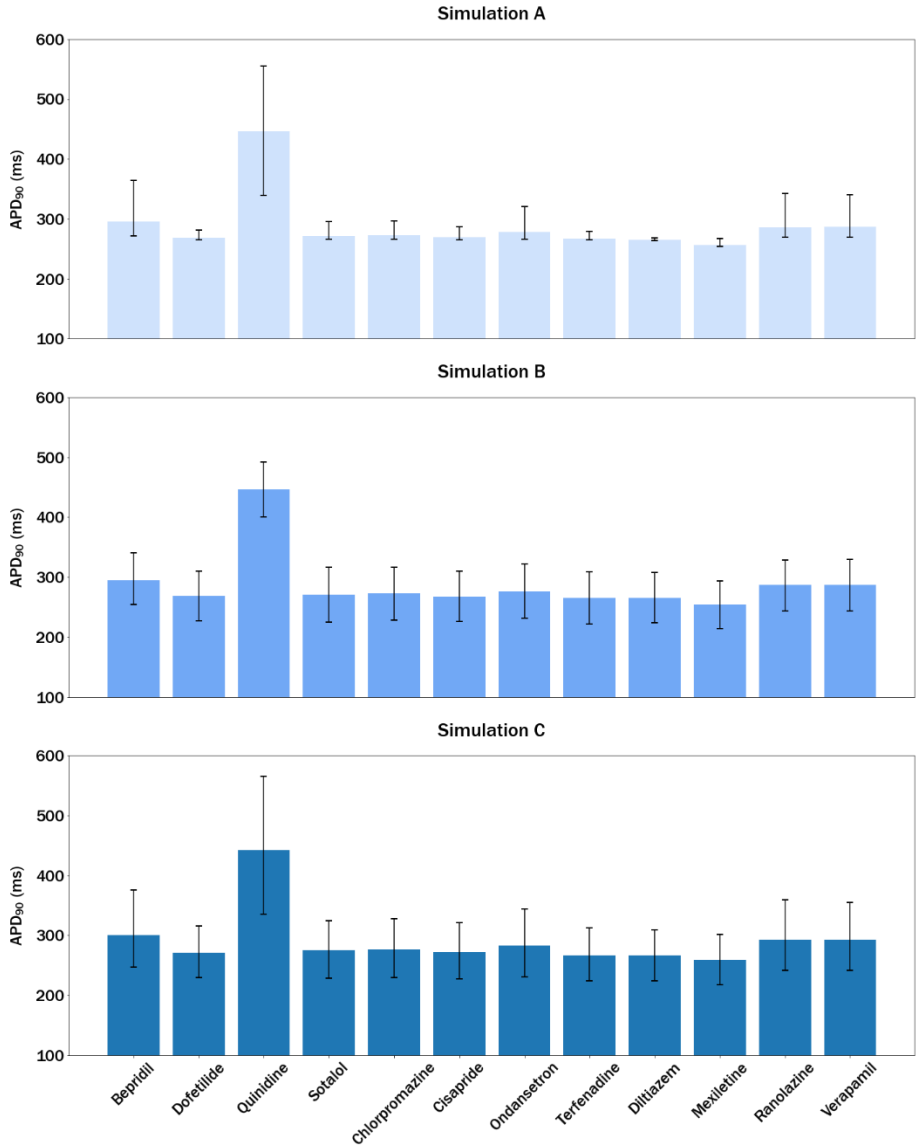


Figura 7: Gráficos de barras que muestran la mediana de las predicciones de APD_{90} obtenidas para los 12 compuestos CIPA, utilizando tres tipos de simulación.

Simulación A: Variabilidad experimental (Δ-pIC_{50}); Simulación B: Variabilidad interindividual (Δ-Parámetros); Simulación C: Combinación de variabilidad experimental e interindividual. Los intervalos representan los percentiles 10^{th} y 90^{th} obtenidos a partir de las distribuciones mostradas en la Figura 6.

En resumen, la inclusión de la variabilidad experimental en las entradas del modelo multinivel de seguridad cardíaca representó un avance significativo para aumentar la confiabilidad de las predicciones derivadas de estos modelos. Además, considerar las diferencias interindividuales en cuanto a los efectos de los medicamentos es especialmente importante cuando se trata de proteger a personas con mayor susceptibilidad a desarrollar arritmias cardíacas, ya que como se describe en Wisniowska, Tylutki, y Polak (2017): "Los seres humanos varían, por lo tanto, los modelos cardíacos deben tenerlo en cuenta...". Finalmente, la combinación de variabilidad u otros tipos de incertidumbre no implicó que los efectos de cada fuente se sumen en la predicción final.

Capítulo 1

Usage of Model Combination in Computational Toxicology

Pablo Rodríguez-Belenguer[1,2], Eric March-Vila[1], Manuel Pastor[1], Victor Mangas-Sanjuan[2,3], Emilio Soria-Olivas[4]

[1]Research Programme on Biomedical Informatics (GRIB), Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Hospital del Mar Medical Research Institute, 08003 Barcelona, Spain.

[2]Department of Pharmacy and Pharmaceutical Technology and Parasitology, Universitat de València, 46100 Valencia, Spain.

[3]Interuniversity Research Institute for Molecular Recognition and Technological Development, Universitat Politècnica de València, 46100 Valencia, Spain.

[4]IDAL, Intelligent Data Analysis Laboratory, ETSE, Universitat de València, 46100 Valencia, Spain.

Revista: Toxicology Letters

Editorial: Elsevier

Año: 2023

Cuartil: Q2

IF: 3.5

Introduction

Traditionally, toxicology has relied on animal experiments to assess the adverse effects of candidate drugs. Nonetheless, in recent years, significant efforts have been made to reduce, refine, and replace animal testing with New Approach Methodologies (NAMs) (Russell & Burch, 1960). This shift towards sustainable science was initiated by the publication of "Toxicity Testing in the 21st Century: A Vision and a Strategy" by the National Academy of Sciences and National Research Council of the USA (Council, 2007).

Despite these advances, inherent complexities in the phenomena under research persist in both experimental and computational methods (*in silico*) within the realm of NAMs, making it challenging to assess the toxicity of the chemicals under study. In this context, two fundamental questions emerge: what do we understand by "complexity", and what types of complexity can we encounter? According to the Cambridge Dictionary, complexity is defined as "*the state of having many parts and being difficult to understand or find an answer to a problem*". Furthermore, an added challenge lies in capturing the interaction among each of these components, resulting in emergent phenomena that are unpredictable at lower levels of representation. With respect to the types of complexities, we propose three major groups: mechanistic, chemical space and methodological.

On the one hand, mechanistic complexities encompass situations in which biological endpoints are constituted by several different processes interconnected in a network, and a reductionist approach may lead to a loss of information or a poor representation of the underlying mechanisms in a biological phenomenon. Chemical space complexities, on the other hand, arise when evaluating the toxicity of new compounds that differ significantly from those used to build a model. Finally, methodological complexities pertain to

the intricacies and challenges encountered in designing, implementing, and executing research methods and procedures. This complexity may occur due to the nature of the research problem, the need to account for multiple variable types, or because of the typical class imbalance problem.

Now, the following question that emerges is: what causes these complexities? Essentially, these complexities rely on the complexity of the biological and chemical processes in the organisms. These complexities frequently constrain experimental and computational methods, compelling them to make numerous assumptions that, in certain instances, may not align with reality. *In vitro* assays, one of the most widely used NAMs, have significant advantages such as speed, cost-effectiveness and ethical acceptability. However, they typically focus on individual cell types or tissues, thereby missing factors such as organ-to-organ communication, systemic effects, and intercellular interactions, affecting the overall relevance of the data collected (Hartung, 2018). Finally, it is essential to emphasize that computational methods, despite the significant benefits they offer, such as cost-effectiveness (Council, 2007), reduced reliance on animal testing, and high-throughput screening (Raies & Bajic, 2016), heavily depend on the data generated by experimental methods. This dependence exposes them to the limitations of experimental data, in addition to their own inherent limitations.

Hence, in this work, we will review how the inherent complexities in the field of computational toxicology have been addressed through the combination of multiple models and the integration of their results. Instead of following a systematic review approach we aim to provide readers with a practical guide on the effective utilization of model combination in the field of computational toxicology, offering insights on when, how, and for what purposes to employ this approach. To facilitate this, we will delve into the "metamodel" concept which represents the combination of multiple models, with each individual

model referred to as a LLM. As the primary focus of this review lies in predictive models (Raies & Bajic, 2016), it is noteworthy that each of the LLMs is constructed utilizing QSAR models. By integrating these LLMs, we can effectively tackle the complexity at hand and optimize our problem-solving capabilities. The metamodel framework enables us to benefit from a comprehensive and well-rounded approach that capitalizes on the unique attributes of each component (March-Vila et al., 2023; Bringezu et al., 2021; C.-H. Chen et al., 2020).

Type of metamodels

A metamodel is a supervised learning approach which involves the combination of several LLMs to achieve superior predictive performance compared to what a classical QSAR model could achieve (Polikar, 2006; Rokach, 2010). Figure 1 presents an overview of the process from problem formulation to its resolution using metamodels. The icons surrounding the head represent some of the day-to-day issues faced by a computational toxicologist, such as high dissimilarity between the test set and train set, complex biological phenomena, algorithmic limitations, or information accessibility issues which stem from the confidentiality of pharmaceutical companies' data, among others. These problems form the basis of the three types of complexities we analyse: mechanistic, chemical space, and methodological complexities. Therefore, once computational toxicologists identify the problem at hand and its associated complexity, they can attempt to solve it using metamodels. On the one hand, there may be a need to better describe the mechanisms of a complex endpoint, where each model represents a specific mechanism (mechanistic-based metamodel).

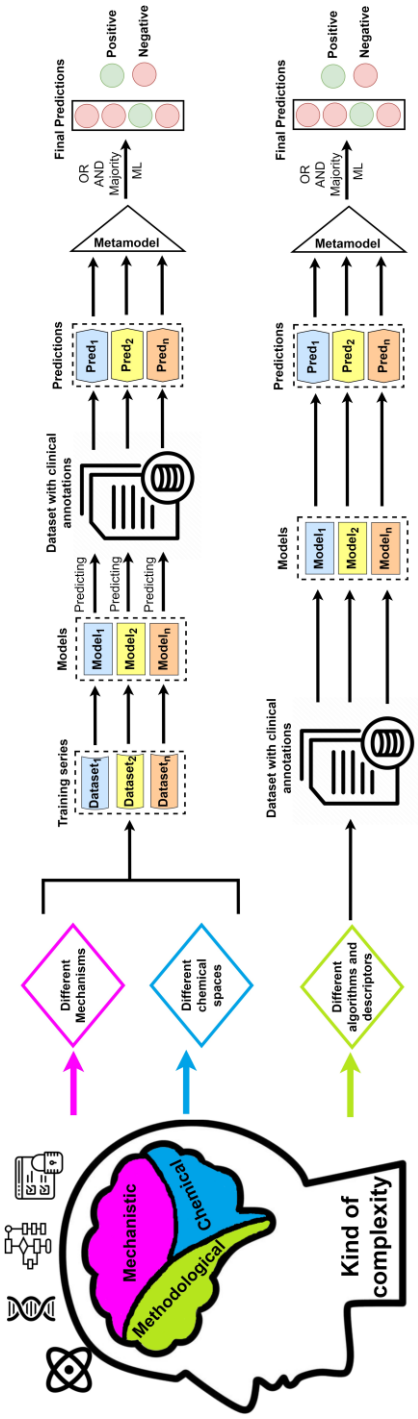


Figure 1: Overview of how three different types of complexities in toxicology are tackled through the combination of models.

On the other hand, there may be a requirement to improve the description of a complex chemical space, with *n* models representing different chemical spaces (fragment-based chemical spaces metamodel). Also, there may be a challenge in enhancing the prediction capability by combining models to better capture the complex association with an endpoint using different algorithms or descriptors (methodological-based metamodel). Regardless of the problem type, once the LLMs are constructed and used for predicting, their outputs are combined to obtain the desired outcome (Bringezu et al., 2021; Daghighi et al., 2022; Yu et al., 2022).

In the process of integrating predictions, another crucial step involves identifying the most suitable combinatorial strategy. This can be achieved through methods such as logical operations (OR, AND, Majority) or building a high-level machine learning model, which is trained using as input the predictions produced by the LLM (Pastor et al., 2021). On certain occasions, the choice of the method is not driven by the predictive performance, but rather by the most logical approach from a toxicological perspective. In such cases, considerations related to the safety and potential risks associated with the chemicals being analysed take precedence over the absolute performance of the model.

Table 1 presents a summary of the main problems associated to the complexities reviewed in this article, highlighting the underlying problem that motivates their use, as well as explaining the reasons why a classical QSAR model would not be a more suitable option, along with the kind of LLM and the main examples to be reviewed.

Table 1: Summary of the underlying problem that leads to proposing a metamodel, kind of metamodel, why a classical QSAR model is not convenient, rationale of metamodel use, low-level models and examples.

Problem	Kind of complexity	Why not a classical QSAR?	Rationale of metamodel use	Low-level models	Examples

Complex biological phenomenon	Mechanistic	Classical QSAR models cannot capture the intricacies of complex biological phenomena.	Need to simplify the problem into different models that represent the different processes of the phenomenon under study.	Molecular Initiating Events (MIEs) of an Adverse Outcome Pathway (AOP) or different levels of mechanistic information.	• Molecular Initiating event combination (Sapounidou et al., 2023; Gadaleta et al., 2022, 2018; Kleinstreuer et al., 2018; Kotsampasakou & Ecker, 2017). • Multi-level model (Mirams et al., 2014a; Rodríguez-Belenguer et al., 2023a).
Data accessibility due to structure confidentiality	Chemical space	It is impossible to combine all the datasets into one because you do not have access to the structures.	Each pharmaceutical company contains its own structures that can cover different positions in the chemical space.	Each dataset for each company represents a model.	• Combination of models without structure sharing (Bosc et al., 2021; Gedeck et al., 2017). • Federated learning (S. Chen et al., 2021; Heyndrickx et al., 2022b; Simm et al., 2021).
Lack of identifiability of patterns in the data	Chemical space	A classical QSAR may overlook or fail to account for the inherent heterogeneity within the dataset.	Models trained on individual clusters can capture the specific patterns within each subset. Combining the models enables leveraging the strengths of each cluster-specific model across the entire dataset.	Each low-level model is trained with the dataset belonging to a cluster obtained by a clustering technique.	• Cluster using unsupervised learning (H. Li et al., 2018; Samanipour et al., 2022).
Algorithm limitations	Methodological	A classical QSAR algorithm cannot excel in all scenarios.	The strengths and weaknesses of the different algorithms can be compensated by combining them with each other.	Different machine learning approximations for the same dataset.	• Algorithm combinations (Cerruela García et al., 2018; D'Souza et al., 2021; Grenet et al., 2019; Hanser et al., 2019; He et al., 2019; Liew et al., 2011; L. Wang et al., 2021; Yu et al., 2022).
Molecular descriptor limitations	Methodological	A classical QSAR model with so many types of variables may have multicollinearity problems.	The combination of models in which each one has a different type of variables allows you to attack the multicollinearity problem separately without the risk of losing information that is necessary.	Different kind of variables for each model.	• Descriptor combinations (Bugeac et al., 2021; Kwon et al., 2019; Smusz et al., 2013).
Class imbalance limitations	Methodological	A classical QSAR will tend to more effectively predicting the majority class.	The combination of appropriately balanced individual models will avoid biases in the prediction towards the majority class.	Replicating the minority class in each model and distribute the majority class evenly until a 50:50 ratio is reached.	• Creation of balanced subsets from original series (Bringezu et al., 2021; March-Vila et al., 2023).

Mechanistic-based metamodel

Understanding the mechanism of toxicity is crucial when evaluating a specific toxic response. Without knowledge of the underlying mechanism, it becomes

challenging to make informed decisions regarding the toxic effects of the drug under examination (Ross, 1989). In fact, Cronin & Richarz (2017) highlight a shift in the field of toxicology towards *"an assessment of the (perturbation of) normal biological pathways relating to toxicity allowing for a mechanistic basis to understanding the effects of chemicals"* (Cronin & Richarz, 2017). Hence, in computational toxicology, constructing models grounded on mechanistic knowledge offers the benefits of enhancing the model's predictive accuracy, its predictive performance when extrapolating, and increasing its interpretability by incorporating the biology that underlies the endpoint of interest (Benzekry, 2020). This has been a longstanding pursuit in the field of artificial intelligence (AI), aiming to move beyond "black box" models that lack transparency and fail to explain the reasons behind their predictions (Petch et al., 2022).

In the realm of *in silico* toxicology, classical QSAR models have shown efficacy in predicting simple biological phenomena (Chinen & Malloy, 2022; De et al., 2022). However, their performance tends to be inadequate when it comes to predicting more complex biological phenomena (e.g., cholestasis, steatosis, or neurotoxicity). Hence, in this review, a mechanistic-based metamodel is focused on the integration of the outputs of several LLMs focused on simpler biological phenomena, with each model representing a specific mechanism relevant to the desired endpoint. So, the combination of simpler biological phenomena at the receptor or organ level is investigated, thereby admitting that complex mechanisms cannot be directly modelled. This recognition arises from the understanding that a classical QSAR model would fail to capture the intricate information underlying each distinct biological phenomenon (Cherkasov et al., 2014).

To overcome the limitation that complex biological phenomena cannot be correctly modelled by a classical QSAR model, it would be useful to integrate

existing mechanistic knowledge into a rational framework, like the one provided by an Adverse Outcome Pathway (AOP). It is important to make clear that AOPs were not created with this purpose, however, they provide a structured and transparent framework which allows models to take advantage of the mechanistic information accessible on AOPwiki. AOPs link molecular initiating events (MIEs) to Adverse Outcomes (AO) through intermediate key events (KEs), providing a mechanistic understanding of the toxicity of chemical compounds (Ankley et al., 2010). One potential strategy to construct a metamodel based on AOP network (which can link different MIEs) consists of the integration of information from multiple MIEs to predict a selected toxicological endpoint. In this scenario, individual QSAR models are constructed corresponding to each identified MIE, enabling the prediction of these events for a dataset with clinical annotations. So, the prediction matrix obtained would serve as the input variables for a model, while the clinical annotations would act as the output variable. The final predictions can be derived by either employing a voting system among the different predicted MIEs or by retraining the prediction matrix using a classifier, as mentioned above. The choice on how to merge model outputs relies on the nature and severity of the adverse event under evaluation. For instance, in some cases, a logical OR could be the preferred approach as it would indicate that the presence of any of the studied mechanisms alone would classify the compound as toxic. However, in other cases, a majority or a greater number of involved mechanisms may need to provide a positive vote to determine whether a compound exhibits positive or negative activity.

Kotsampasakou & Ecker (2017) employed a combination of MIEs to predict cholestasis. In this work, the authors created an *in silico* method by constructing a metamodel that combined multiple transporters (acting as MIEs) related to cholestasis occurrence and compared it with a classical QSAR approach. In this case, the MIEs matrix was trained with different Machine

Learning (ML) classifiers (Kotsampasakou & Ecker, 2017). The findings indicated that the metamodel produced better predictive performance than the classical QSAR models. Also, Gadaleta et al. (2022) developed a MIE-based metamodel approach for predicting neurotoxicity. Fifteen QSAR models were created, each corresponding to a different MIE and were combined using a balanced random forest. The MIE predictions were used alongside chemical descriptors and structural fingerprints in various classifiers to compare their predictive performance. Overall, classifiers based on MIE predictions showed prediction accuracy similar to those based on chemical descriptors and structural fingerprints (Gadaleta et al., 2022). Other works that have been evaluated using similar methods have also demonstrated encouraging outcomes (Sapounidou et al., 2023; Gadaleta et al., 2018; Kleinstreuer et al., 2018).

However, although AOPs provide a valuable framework for this type of metamodeling, it is not the only way to combine biological information. In other studies, Mirams et al. (2014) developed a multi-level *in silico* tool combining information from various ion channels to predict the action potential duration at 90% of the repolarization (APD_{90}), which is an important measure of the cardiac cell's depolarization and repolarization time during an action potential (Mirams et al., 2014a). The authors were able to create a metamodel that incorporated this multi-level information, resulting in highly accurate predictions. This type of metamodel that combines ion channels for the prediction of a given biomarker was subsequently used in another work (Rodríguez-Belenguer et al., 2023a). Here, Rodríguez-Belenguer et al. (2023) employed a metamodel with various ion channels to predict APD_{90}, aiming to reduce the number of time-consuming electrophysiological simulations. The authors successfully demonstrated that decreasing the number of simulations led to an almost hundred-fold reduction in computation time.

This type of metamodel is limited by the difficulty of identifying the biological foundations of toxicological endpoints and a wide enough collection of involved mechanisms. For instance, when attempting to model an AOP network with multiple MIEs, the possibility of having poor knowledge about a specific MIE or the presence of missing data in any of the MIEs may occur. This can result in the metamodel lacking one or more crucial low-level models that could be relevant to the occurrence of the adverse event in question. In addition, the AOP itself is not always well-defined, and indeed, the AOP evolve with time, and they are arranged in networks. This is an additional, higher level of complexity.

Fragment-based chemical spaces metamodel

The drug-like chemical space has a potential size as vast as 10^{60} compounds (Hoffmann & Gastreich, 2019). Hence, in novel drug development, it is essential to have alternatives to classical QSAR models for accurately predicting the endpoint of interest for compounds under investigation, especially when dealing with those that may occupy different regions in the chemical space compared to the ones present in the training dataset. Combining models from different chemical spaces can enhance the predictivity of a classical QSAR model, with two scenarios motivating us to construct such metamodels. The first involves working with several chemical spaces in the pharmaceutical industry, which can be complicated since the intellectual property of the compounds belongs to each individual company, necessitating the search for strategies that allow working without the need to share compound structures. The second involves having access to a complete dataset from which we are unable to identify any patterns at first glance. Regarding the first case, different companies may possess molecules that occupy a specific position in the chemical space, while other companies may have other drugs with similar or different positions. In the second case, the use

of clustering algorithms (belonging to the branch of unsupervised ML algorithms) can help to identify data patterns. Regardless of the type of scenario, developing individual models for different chemical spaces and combining them could increase the predictivity of classical QSAR model that would otherwise miss the inherent heterogeneity in the dataset. One must estimate which approach -logical operations or machine learning - yields the most accurate predictions.

Fragment chemical spaces without sharing data

Martin & Zhu (2021) pointed out that collaboration between pharmaceutical companies can be hindered by the protection of intellectual property and trade secrets (Martin & Zhu, 2021). The specific biological structures and targets of interest to each company create further barriers to collaboration. Thus, it becomes crucial to adopt strategies that facilitate sharing of models without disclosing structures or activity data to enable secure collaboration among competitors. For this, different strategies can allow to work with confidential data, but the ones we have reviewed focus on sharing the individual predictions from each low-level model and the use of federated learning strategies (McMahan et al., 2016; Konečný et al., 2017). Both strategies are specifically designed to overcome the collaborative paradigm and enable data owners to jointly train a model without exposing their data to others. As an example, Gedeck et al. (2017) constructed a metamodel with a Bayesian ridge regression that practically reproduced the results of a classical QSAR model (Gedeck et al., 2017). In another work, Bosc et al. (2021) formed a consortium with different partners and trained multinomial naïve Bayes models (Manning et al., 2008) with eleven datasets for predicting malaria (Bosc et al., 2021). The metamodel exhibited good performance across various validation sets and had the significant advantage of being computationally efficient. They also developed a web application accessible through the link:

https://www.ebi.ac.uk/chembl/maip/. Regarding federated learning strategies, the MELLODY project utilized federated machine learning to train predictive models on data that remains on the owner's servers, without the need to transfer it to a central location. This approach ensures that the data and asset owners retain control of their information throughout the project. The federated model was trained on the platform by aggregating the gradients of all contributing partners in a cryptographic, secure way, which enabled the creation of a global federated model for drug discovery without sharing confidential datasets. The successful application of federated learning in MELLODY will lead to substantial efficiency gains in drug discovery and development, as it expands the data available to a broader set of stakeholders (Heyndrickx et al., 2022b). Other works that have been reviewed show that federated learning within computational toxicology is a powerful tool to work in a collaborative way between companies (S. Chen et al., 2021; Simm et al., 2021).

In essence, these kind of metamodels have the significant advantage of increasing the possibility of having similar compounds between both the training and test sets by working collaboratively between companies with different chemical spaces. However, it is worth noting that the model closest to the test set in the chemical space will perform better than the others and may not benefit from the combination of models. In contrast, the combination of all models should provide better results for companies that are further away from that chemical space. At the outset of the modelling process, there is no information available on the similarity of the test set to the dataset or to those of other pharmaceutical companies. While this technique may disadvantage a company, the collective benefits outweigh any potential drawbacks. One potential solution to overcome these disadvantages may be to eliminate individual company predictions that exceed a certain distance from the test set (this approach is being carried out in a work of our group). In this way, the best

LLM would not lose importance by the combination of models, and the worst models would still contribute to the combination.

Fragmenting chemical spaces with clustering approaches

Datasets often contain inherent heterogeneity, where different subsets of instances exhibit distinct patterns or behaviours. By clustering the data, you can identify and separate these subsets into individual clusters. Building separate models for each cluster enables targeted modelling, ensuring that each model captures the specific patterns within its assigned cluster.

One can either use expert knowledge to cluster the data, for instance using a specific chemical, molecular or pharmacokinetic descriptor and group the data that fall within certain thresholds, or unsupervised learning algorithms such as K-means (Selim & Ismail, 1984), DBScan (Ester et al., 1996) or hierarchical clustering (Johnson, Stephen C., 1967) to understand better how the data groups together and then create subsets based on the grouping. We can trace approaches to solve this issue back to 1977, when Svante Wold published SIMCA (Wold & Sjöström, 1977) as a way to identify clusters within chemical data.

The approach known as “clustering first, and then modelling” (H. Li et al., 2018; Yuan et al., 2007) has been applied in some works. Yuan et al. (2007) concluded in their work that the statistical results obtained by local models based on the subsets were much superior to those obtained by the global model based on the whole training set (Yuan et al., 2007) and Li et al. (2018) mentioned that by creating subsets of similar compounds the afterwards modelling shows better predictions because analogical chemicals are more likely to capture same category molecules precisely, suggesting that low-level models are superior to classical QSAR models (H. Li et al., 2018).

Golalipour et al. (2021) compiled a plethora of methods to cluster data and combined these clusters into an ensemble. They suggested that each clustering method has good efficiency on specific data (Golalipour et al., 2021), which means that one should review the data and the context in which is working to choose the clustering method that best suits its needs. Also, they mentioned that a clustering ensemble shows a better performance than the set of base clustering methods (Golalipour et al., 2021), a situation we have also observed in our ensemble models when compared to classical models.

An illustrative example of how to cluster the data either using clustering algorithms or expert knowledge, such as defined labels, can be found in the work of Samanipour et al. (2022). They collected acute fish toxicity data from different sources and derived the toxic classes of the compounds using on one side k-means clustering and on the other, Globally Harmonized System of Classification (GHS) (*UNECE, 2021*) labels referred to acute fish toxicity based on different thresholds, so the compounds can be classified from very low to high toxicity (Samanipour et al., 2022). Then, they compared the performance of a QSAR regression model against a descriptor-based direct classification model. Finally, they checked the categorization obtained from the k-means and the knowledge based on GHS, finding that there is a high level of similarity in the thresholds (Samanipour et al., 2022). We find these results illustrative enough to encourage clustering and subset selection using an unsupervised clustering algorithm to identify patterns within data and generate subsets that allow the creation of LLMs that overcome the limitations of the heterogeneity in the training series. Still, one must be cautious since each dataset must be well-defined and studied, alongside with the problem to solve.

In broad terms, by using clustering techniques, you can identify distinct subsets of data with similar characteristics (Samanipour et al., 2022). Each individual model can then specialize in learning patterns within a specific

cluster, potentially leading to improved performance. However, if the clusters are not well-defined or if there is significant overlap between clusters, the individual models may not effectively capture the desired patterns. Clustering errors or misclassifications can propagate to the metamodel, impacting its overall performance.

Methodological-based metamodel

In this category of metamodels, those that are built without depending on the examination of any biological mechanism, physicochemical attribute, or pharmacokinetic property are included. Their primary objective consists of improving model performance by addressing methodological challenges which are universal in the realm of computational toxicology. Some of the commonly reviewed in this work include the inadequacy of a single algorithm or variable type to effectively solve the problem and class imbalance. This approach is designed to tackle a specific aspect of the problem at hand, surpassing the limitations of a classical QSAR model. Here, the merging of outputs from individual models would prioritize achieving the optimal performance of a metamodel, as the resolution of the problem is not affected by biological or chemical complexities.

Algorithm limitations

While numerous ML algorithms have been employed to acquire knowledge on QSARs, there is not a universally recognized optimal algorithm for QSAR learning (Wu et al., 2021). Thus, understanding the working principles, advantages, and disadvantages of each algorithm is crucial to determine the potential benefits that a combination of algorithms can offer for a given problem or task (Table 2).

Table 2: Summary of the main ML algorithms in IST and their advantages and disadvantages.

Algorithms	Problem	Association	How does it work?	Advantages	Disadvantages
Linear regression (LR) (Hastie, Tibshirani, & Friedman, 2009)	Regression	Linear	A statistical model that finds the linear relationship between independent and dependent variables.	• Simple to implement. • Overfitting can be reduced by regularization.	• Subject to underfitting. • Large effect of outliers.
Partial least square (PLS) (H. O. A. Wold, 1968)	Regression	Linear	A statistical method that constructs a linear regression model by projecting both the predicted variables and observable variables onto a new space.	• Handling multicollinearity. • Dealing with high-Dimensional data.	• Subject to overfitting. • Large effect of outliers.
Ridge (L2) (Hoerl & Kennard, 1970a)	Regression[a]	Linear[b]	A linear regression technique introduces regularization to prevent overfitting by adding a penalty term to the loss function. Penalizes the sum of squares of the weights.	• Performs well with high-dimensional data. • Reduces the impact of irrelevant predictors.	• Biased estimates when there is substantial multicollinearity. • Shrinks all coefficients.
Lasso (L1) (Tibshirani, 1996)	Regression[a]	Linear[b]	A linear regression technique that introduces regularization and performs feature selection by adding an absolute penalty term to the loss function. Penalizes the sum of absolute values of the weights.	• Find relevant predictors. • Effective in situations where only a small number of predictors are truly important.	• May struggle with multicollinearity. • Sometimes can be very strict.
Naïve Bayes (Gareth James, 2013)	Classification	Linear	It is a simple probabilistic machine learning algorithm that makes predictions based on the application of Bayes' theorem with the assumption of independence between features.	• Fast and efficient. • Handles missing data gracefully.	• Assumes independence between features. • Assumption about the priors.
Polynomial (Hastie, Tibshirani, & Friedman, 2009)	Regression	Non-linear	A regression technique that fits the data to a polynomial function to capture nonlinear relationships.	• Works on any size of data. • Flexibility of shape.	• Can lead to overfitting if the degree of the polynomial is too high. • The higher the polynomial degree the higher of model complexity.
Support Vector Machines (SVM) (Cortes & Vapnik, 1995)	Regression and classification	Linear and non-linear	A supervised learning algorithm that finds an optimal hyperplane or boundary to separate data points or estimate continuous values after a nonlinear transformation of the input data (kernels).	• Effective in high-dimensional spaces. • Versatile kernel functions.	• Computationally Intensive. • Sensitive to noisy data or imbalance.
Decision Tree (DT) (Breiman, 1984)	Regression and classification	Non-linear	Creates a hierarchical structure of rules to make predictions by recursively splitting data based on feature values.	• Easy to interpretability. • Robust to outliers. • Independent of the input range.	• Lack of smoothness. • High Variance.

Random Forest (RF) (Breiman, 2001)	Regression and classification	Non-linear	An ensemble learning method that combines multiple decision trees, where the final result is obtained through voting or averaging.	• Improved generalization regarding individual decision trees. • Handling high-dimensional data. • The importance of the variables can be obtained.	• Less Interpretable than the decision trees. • Computational complexity.
K-Nearest Neighbors (KNN) (Fix & Hodges, 1951)	Regression and classification	Non-linear	Assigns a data point based on the majority vote or average of its k nearest neighbours.	• Simplicity. • No assumptions of data distribution.	• Feature Scaling. • Memory requirements.
XGBoost (XGB) (T. Chen & Guestrin, 2016)	Regression and classification	Non-linear	A gradient-boosting algorithm that uses a set of weak learners to build a powerful predictive model.	• Parallel and sequential processing. • Feature importance.	• Sensitive to outliers. • Memory usage. • Problems with high-dimensionality.
Neural Networks (NN) (Rumelhart et al., 1986a)	Regression and classification	Linear and Non-linear	A computational model that learns complex patterns and relationships between input and output data through interconnected layers of nodes (neurons).	• Ability to learn complex.patterns. • Flexibility.	• Large training data requirements. • Hyperparameter manual tuning.

[a] They can be incorporated into classification algorithms as hyperparameters
[b] They can be incorporated into algorithms suitable for non-linear associations

In this context, it is important to note that a classical QSAR modelling approach with the same model does not necessarily excel in all scenarios. Thus, considering the potential complementarity among different algorithms (Table 2) becomes crucial to develop more reliable models for toxicological predictions. For instance, in the field of Drug-induced liver injury (DILI) prediction, researchers such as Liew et al. (2011) were early adopters of model combinations to improve their predictions (Liew et al., 2011). The approach involved constructing a total of 617 base classifiers using a diverse set of 1,087 compounds. These base models incorporated K-NN and SVM, which were further stacked with a Naïve Bayes classifier. The performance of the ensemble was evaluated through internal validation using a five-fold cross-validation technique. The study revealed that the ensemble model exhibited proficient classification of positive compounds associated with hepatic effects. However, its performance was comparatively lower for negative compounds, especially when they possessed structural similarities. In the latter case, a classical QSAR model is probably more efficient.

In another work, Hanser et al. (2019) showed that using the same training series with a combination of statistical models like RF and SOHN (Hanser et al., 2014) along with an expert system such as Derek Nexus, produced a better outcome than the individual models by itself (Hanser et al., 2019). In another reviewed work, Ancuceanu et al. (2020) stacked a set of 78 ML models for predicting DILI achieving slightly superior results to other models published (Ancuceanu et al., 2020). The most common algorithms were Decision Tree, Random Forest, Support Vector machines and Neural Networks which were weighted using majority voting. The balance accuracy of this work (74%) was higher than the work published by He et al. (2019). Other works reviewed showed that the combination of different algorithms is a technique commonly used in the field of computational toxicology (Yu et al., 2022; Cerruela García et al., 2018; Grenet et al., 2019; D'Souza et al., 2021; L. Wang et al., 2021).

In summary, the combination of diverse mathematical approaches offers the advantage of compensating for the strengths and weaknesses of individual algorithms. However, it is important to note that such models tend to be treated as complete black boxes, focusing solely on improving performance without considering chemical, pharmacokinetic, or biological complexities.

Molecular descriptor limitations

The training of a model relies on accurately representing molecules using descriptors that effectively capture their properties and structural features. Literature offers numerous molecular descriptors, encompassing a wide range from basic molecule properties to intricate three-dimensional representations. These descriptors are often stored as vectors with hundreds or even thousands of elements. It is crucial to acknowledge that there is no single optimal choice for the best feature (Carracedo-Reboredo et al., 2021). Consequently, the selection and combination of features should be carefully

studied, considering the context and objectives of the modelling research (Carracedo-Reboredo et al., 2021).

Classical QSAR modelling using different descriptors, rather than employing a combination of models with these diverse descriptors, might show three major drawbacks:

1. High dimensionality: Combining multiple descriptors increases the overall dimensionality of the model. As the number of features grows, computational complexity and resource requirements can escalate significantly. This can lead to longer training times and challenges in optimizing the model's performance (W. Zhou et al., 2012).

2. Difficulty of interpretability: Incorporating numerous descriptors into a single model can make it more challenging to interpret and understand the contributions of individual features to the model's predictions. Extracting meaningful insights and interpreting feature importance becomes more complex when multiple features are combined (Matveieva & Polishchuk, 2021).

3. Multicollinearity: When using multiple descriptors that may be correlated or redundant, multicollinearity can appear. This can lead to instability in the model's performance and make it difficult to discern the true influence of each descriptor on the toxicological endpoint being predicted (Heo et al., 2019).

Hence, employing a combination of individual models, each containing one type of variable, can effectively address the challenges associated with classical QSAR modelling. In Table 3 there are listed the most used descriptors in computational toxicology.

Table 3: Main descriptors used in computational toxicology

Kind of descriptor	Brief description

Physico-chemical descriptor 1D	They allow the calculation of information based on specific fractions of a molecule. Examples: carbon atoms, cyanates, or nitriles (Carracedo-Reboredo et al., 2021).
Physico-chemical descriptor 2D	They rely on graphical representations of molecules, exhibiting theoretical structural properties which are preserved under isomorphism. Examples: Molecular weight, number of bonds, hydrogen bond acceptor (Carracedo-Reboredo et al., 2021).
Physico-chemical descriptor 3D	They consider the distances between bonds, bond angles, dihedral angles, and other measures. Examples: Asphericity, eccentricity, and inertial shape factor (Carracedo-Reboredo et al., 2021).
Molecular ACCess Systems keys fingerprint (MACCS)	They are constructed using SMART patterns and are optimized for substructure searching based on 2D molecular descriptors. There are two kinds: 166-bit keyset and a 960-bit keyset (Durant et al., 2002).
PubChem Fingerprints (PubChemFP)	This fingerprinting method encodes 881 structural key types, representing substructures found in a fraction of all compounds within the PubChem database. They are used by PubChem for similarity neighbour and similarity searching (Y. Wang et al., 2009).
Extended Connectivity Fingerprints (ECFP)	ECFP is a representation of a molecule's structure based on its connectivity pattern. It captures information about the presence or absence of specific chemical substructures and their connectivity to neighbouring atoms (Rogers & Hahn, 2010).
Atom pairs	AtomPairs2DFingerprint (APFP) captures information about the atomic environment and shortest path separations between pairs of atoms in a compound's topological representation. 780 distinct atom pairs at different topological distances are encoded (Schneider et al., 1999; Carhart et al., 1985). GraphOnlyFingerprint (GraphFP) encodes the 1024 unique paths of a fragment within the compound's structure (Steinbeck et al., 2003).
RDkit fingerprints	They are generated by considering all potential paths of specific lengths, originating from each heavy atom in the molecular graph (Landrum et al., 2020).

For instance, in the study conducted by Smusz et al. (2013), a multidimensional analysis of machine learning methods was employed to classify bioactive compounds (Smusz et al., 2013). Researchers constructed eleven learning algorithms, which included four meta-classifiers (with different input combinations). Various types of fingerprints, such as ECFP, MACCS, or PubChemFP, among others, were utilized in this analysis. The incorporation of meta-learning techniques resulted in an enhancement of the evaluation parameters. This suggests that the use of meta-learning approaches improved the performance of the classical models, leading to increased accuracy and predictive capability.

In another study performed by Kwon et al. (2019), an ensemble method was proposed and evaluated on nineteen bioassay datasets (Kwon et al., 2019). The results showed that the ensemble method consistently outperformed thirteen individual models. The researchers utilized three types of molecular fingerprints, namely PubChem, ECFP, MACCS, and SMILES. Regarding ML models, they employed SVM which achieved the highest average Area Under the Curve (AUC) value compared to other algorithms such as NN, RF, gradient boosting machines (GBM), and ordinary regression.

In the study conducted by Bugeac et al. (2021), the researchers constructed a metamodel consisting of 28 individual models, each utilizing a different set of

physicochemical descriptors and fingerprints (Bugeac et al., 2021). The performance of various classification algorithms was evaluated, including KNN, logistic regression, decision tree classifier, and ensemble methods. Among these algorithms, KNN, logistic regression, and decision tree classifier demonstrated the highest balanced accuracy. However, during nested cross-validation, the ensemble method exhibited slightly superior results. It suggests that the ensemble approach was able to harness the collective predictive power of the individual models, resulting in improved performance.

As previously mentioned, constructing a metamodel using individual models with different input variables can effectively address challenges related to high dimensionality, interpretability, and multicollinearity when dealing with each variable type separately. However, this approach overlooks the intricate complexities of biological systems that extend beyond these factors. Consequently, this oversight could lead to incomplete representations of biological responses, undermining the accuracy and robustness of predictions.

Balancing strategies

When dealing with real-world data we often see the problem of class imbalance, which consists in the overrepresentation of one class over the other one. This causes learning algorithms to bias towards the majority class (Krawczyk, 2016; Megahed et al., 2021). Using a metamodel to correct class imbalance can offer several advantages over classical QSAR models with over- or under-sampling strategies (Liu et al., 2022; March-Vila et al., 2023). One key advantage is that the metamodel allows for the creation of individual models that balance the classes, often leading to improved predictive power (Galar et al., 2012). The choice of how to balance datasets relies on both the size and the degree of imbalance in the original training dataset. To illustrate, when dealing with a large collection of compounds that exhibits a significant

imbalance, such as a ratio of 100:1, it is possible to create multiple subsets where each subset contains an equal number of instances from the underrepresented class while ensuring a fair representation of compounds from the overrepresented class. The way to combine these multiple subsets depends on the nature of the endpoint that is being studied, thus the choice of a logical OR, logical AND or majority voting should be decided according to how the endpoint is defined: is a compound positive for a certain endpoint no matter in which part of the chemical space of the subsets falls in, or should it be considered positive after several models have shown it is?

Galar et al. (2012) proposed a new taxonomy for ensemble-based techniques to deal with the imbalanced dataset, that consisted of cost-sensitive ensembles and data preprocessing followed by ensemble learning (Galar et al., 2012). They concluded that ensemble-based algorithms are worthwhile since they improve the results that are obtained by the usage of data preprocessing techniques and training a single classifier. They state that, despite the use of more classifiers making them more complex, the overall growth is justified by the better results that can be assessed (Galar et al., 2012).

Bringezu et al. (2021) solved their imbalance data problem by creating a series of low-level models with balanced datasets that stemmed from the main training series. This resulted in a classifier with high sensitivity and specificity (Bringezu et al., 2021). They also compared the performance of their ensemble models against a classical QSAR model that used the main training series without correcting the imbalance. The classical QSAR had good specificity but low sensitivity (0.95 and 0.47 respectively) which led to an accuracy of 0.71, whereas the ensemble model had a more balanced specificity/sensitivity ratio (0.87 and 0.92 respectively) which resulted in an accuracy of 0.87, thus improving the results of the classical QSAR (Bringezu et al., 2021).

Likewise, March-Vila et al. (2023) obtained a very unbalanced dataset that was performing badly as a model since a classical QSAR tends to predict more effectively the majority class. They decided to split the training series into multiple balanced sets, keeping the same negative annotated compounds in all the subsets, since those were the less represented class, and adding different positive compounds into each of the sets until they reached a balance between positive and negative annotations. They proceeded with the creation of one model for each dataset and a subsequent metamodel was created based on the previous low-level models. They found two main advantages in this approach: first, using an imbalance correction algorithm such as random oversampling (Menardi & Torelli, 2014) or random undersampling (Lemaître et al., 2017) would have affected the chemical space they were working on, and, in this way, the chemical space remained unaltered. And second, the performance of the metamodel surpassed the previously created models with the unbalanced datasets (March-Vila et al., 2023).

In this kind of strategy, it is important to consider that the metamodel approach assumes a static class distribution during the training phase. If there are significant changes in the class distribution during deployment or inference, the performance of the metamodel may deteriorate. In such cases, it becomes necessary to retrain or update the metamodel to adapt to the new distribution and ensure continued optimal performance.

Discussion

In this article, we have focused on reviewing three different types of model combinations: mechanistic-based metamodel, fragment-based chemical space metamodel, and methodological-based metamodel. These kind of metamodels have been developed to tackle the intrinsic complexities inherent in the problem under investigation within the field of computational

toxicology. In broad terms, we have observed that the predictivity of classical QSAR models might be constrained when describing complex biological phenomena, complex chemical spaces, or with different methodological challenges. These complexities could stem from the fact that classical QSAR models might face challenges in interpreting the specific mechanisms leading to the assessment criterion, the broad chemical space or the singularities of each kind of ML algorithm, training series and molecular descriptor, whereas a lower-level model may have an easier time capturing and interpreting these complexities due to its ability to delve into the underlying mechanisms and nuances of the data. Furthermore, there is an additional layer of complexity encompassing the three described above, which pertains to the intricacies of human beliefs and behaviour. This layer plays a significant role in determining the credibility of computational predictions, influencing others' willingness to use them in decision-making.

However, despite all the benefits shown in this review for metamodels, they typically combine individual models that use *in vitro* data to predict clinical annotations established in *in vivo* assays. As mentioned in the introduction, *in vitro* data may not directly correlate with *in vivo* data due to the absence of important factors such as absorption, distribution, metabolism, and excretion (ADME) processes, which is a limitation in itself. This limitation poses a significant obstacle to obtaining reliable results since we are trying to predict *in vivo* data from variables that contain only *in vitro* information. To address this issue, one possible solution may be to employ *in vitro* to *in vivo* extrapolation models (QIVIVE), based on a minimal physiologically based pharmacokinetic (mPBPK) model. This model effectively simulates the longitudinal behaviour of a substance in a living organism, considering crucial pharmacokinetic phenomena over time, such as ADME processes. The use of QIVIVE models would allow extrapolating *in vitro* concentrations to *in vivo* doses. Thus, QIVIVE models could provide a valuable tool to enhance the

accuracy and relevance of metamodels, ultimately advancing our understanding of drug behaviour and its effects in the context of the real world. In general, metamodels facilitate the integration of various types of data, including *in vitro* and *in vivo*, enabling a comprehensive assessment of chemical toxicity. This integration provides the foundation for a more accurate risk assessment by holistically considering hazard and exposure of substances across different levels of biological organization using metamodels.

If earlier we discussed how to enhance metamodels through QIVIVE models, another long-standing debate within the computational toxicology community aimed at improving statistical methods is the integration with expert systems. The combination of expert-based and statistical approaches exhibits substantial potential across various scientific domains, notably in the assessment of mutagenic impurities within the pharmaceutical industry. This collaborative synergy offers the prospect of producing more resilient and precise outcomes by harnessing human expertise alongside the objectivity of statistical models. Nevertheless, it is imperative to approach this amalgamation judiciously, particularly within the framework of regulatory standards, as exemplified by ICH M7. The primary concern pertains to the risk that an inadequately managed combination might compromise the transparency and scientific justification demanded during the evaluation of mutagenic impurities. To maintain the credibility and acceptance of such hybrid models, they must uphold transparency in their decision-making processes and ensure comprehensibility, aligning them with the principles enshrined in regulatory guidelines. Striking a delicate equilibrium that capitalizes on the strengths of both approaches while staying in compliance with regulatory standards is pivotal to the continued advancement and recognition of these models.

Despite the good performance shown by metamodels, it is still necessary to look at the data and understand the problem to be solved, since each context must be dealt with a specific solution rather than a general one that fits all. Even though we are in the era of AI and very powerful technologies appear, such as generative deep learning algorithms, we must keep in mind that properly handling the data, for instance with a rational curation process or hypothesis testing, is the first and key step that will allow us to develop a proper solution to a given problem.

Still, and considering the limitations shown in this review, it is of major importance that we collect the knowledge regarding metamodels since they have proven to be a useful and powerful approach for solving complex problems. At the end of the day, if we manage to develop tools that aid decision-making in a rational and justified manner, we will be able to better explain the reality surrounding us and we will have a better understanding of how the models work, avoiding black boxes thanks to a better interpretability.

Conclusions

Our review of combining different QSAR models for predicting toxicological endpoints has shed light on the effectiveness of this approach in addressing various complexities but also provided a guide on how, when, and for what purposes to utilize them. By categorizing these complexities as mechanistic, chemical space, and methodological, we aimed to provide a systematic understanding of the challenges faced in the field of computational toxicology.

The findings from the reviewed works overwhelmingly demonstrate that model combination yields superior performance compared to classical QSAR models. This notable improvement can be attributed to the ability of the combination approach to focus on individual sub-processes, allowing for a more targeted and accurate analysis of the specific toxicological endpoint of

interest. In contrast, the classical QSAR model, incorporating a heterogeneous mix of all sub-processes, may lead to confusion and less precise predictions. The versatility of the model combination approach lies in its ability to unravel the intricacies of mechanistic interactions, capture diverse chemical space representations, and overcome methodological limitations. By leveraging the strengths of different models and integrating their outputs, the metamodels offer a promising avenue for addressing the complexities present in toxicological studies.

The consolidation of information on model combination into a single article holds the potential to be a valuable resource for the computational toxicology community. However, we acknowledge that the subdivision of complexities giving rise to the different reviewed metamodels is our point of view which is essential for a comprehensive understanding of their applicability. Similarly, we recognize that further research is necessary to explore additional combinations of models and the integration of advanced computational techniques.

Funding

The authors received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 964537 (RISK-HUNT3R), which is part of the ASPIS cluster.

Conflicts of interest

The authors declare no conflict of interest.

References

Ancuceanu, R., Hovanet, M. V., Anghel, A. I., Furtunescu, F., Neagu, M., Constantin, C., & Dinu, M. (2020). Computational Models Using

Multiple Machine Learning Algorithms for Predicting Drug Hepatotoxicity with the DILIrank Dataset. *International Journal of Molecular Sciences*, *21*(6), Article 6. https://doi.org/10.3390/ijms21062114

Ankley, G. T., Bennett, R. S., Erickson, R. J., Hoff, D. J., Hornung, M. W., Johnson, R. D., Mount, D. R., Nichols, J. W., Russom, C. L., Schmieder, P. K., Serrrano, J. A., Tietge, J. E., & Villeneuve, D. L. (2010). Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environmental Toxicology and Chemistry*, *29*(3), 730-741. https://doi.org/10.1002/etc.34

Benzekry, S. (2020). Artificial Intelligence and Mechanistic Modeling for Clinical Decision Making in Oncology. *Clinical Pharmacology & Therapeutics*, *108*(3), 471-486. https://doi.org/10.1002/cpt.1951

Bosc, N., Felix, E., Arcila, R., Mendez, D., Saunders, M. R., Green, D. V. S., Ochoada, J., Shelat, A. A., Martin, E. J., Iyer, P., Engkvist, O., Verras, A., Duffy, J., Burrows, J., Gardner, J. M. F., & Leach, A. R. (2021). MAIP: a web service for predicting blood-stage malaria inhibitors. *Journal of Cheminformatics*, *13*(1), 13. https://doi.org/10.1186/s13321-021-00487-2

Breiman, L. (1984). *Classification and Regression Trees*. Wadsworth International Group.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Bringezu, F., Carlos Gómez-Tamayo, J., & Pastor, M. (2021). Ensemble prediction of mitochondrial toxicity using machine learning technology. *Computational Toxicology*, *20*, 100189. https://doi.org/10.1016/j.comtox.2021.100189

Bugeac, C. A., Ancuceanu, R., & Dinu, M. (2021). QSAR Models for Active Substances against Pseudomonas aeruginosa Using Disk-Diffusion Test Data. *Molecules*, *26*(6), 1734. https://doi.org/10.3390/molecules26061734

Carhart, R. E., Smith, D. H., & Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: Definition and applications. *Journal of Chemical Information and Computer Sciences*, *25*(2), 64-73. https://doi.org/10.1021/ci00046a002

Carracedo-Reboredo, P., Liñares-Blanco, J., Rodríguez-Fernández, N., Cedrón, F., Novoa, F. J., Carballal, A., Maojo, V., Pazos, A., & Fernandez-Lozano, C. (2021). A review on machine learning approaches and trends in drug discovery. *Computational and Structural Biotechnology Journal*, *19*, 4538-4558. https://doi.org/10.1016/j.csbj.2021.08.011

Cerruela García, G., García-Pedrajas, N., Luque Ruiz, I., & Gómez-Nieto, M. Á. (2018). An ensemble approach for in silico prediction of Ames mutagenicity. *Journal of Mathematical Chemistry*, *56*(7), 2085-2098. https://doi.org/10.1007/s10910-018-0855-z

Chen, C.-H., Tanaka, K., Kotera, M., & Funatsu, K. (2020). Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications. *Journal of Cheminformatics*, *12*(1), 19. https://doi.org/10.1186/s13321-020-0417-9

Chen, S., Xue, D., Chuai, G., Yang, Q., & Liu, Q. (2021). FL-QSAR: a federated learning-based QSAR prototype for collaborative drug discovery. *Bioinformatics (Oxford, England)*, *36*(22-23), 5492-5498. https://doi.org/10.1093/bioinformatics/btaa1006

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., Consonni, V., Kuz'min, V. E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., & Tropsha, A. (2014). QSAR Modeling: Where Have You Been? Where Are You Going To? *Journal of Medicinal Chemistry*, *57*(12), 4977-5010. https://doi.org/10.1021/jm4004285

Chinen, K., & Malloy, T. (2022). Multi-Strategy Assessment of Different Uses of QSAR under REACH Analysis of Alternatives to Advance Information Transparency. *International Journal of Environmental Research and Public Health*, *19*(7), 4338. https://doi.org/10.3390/ijerph19074338

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Council, N. R. (2007). *Toxicity Testing in the 21st Century: A Vision and a Strategy*. The National Academies Press. https://doi.org/10.17226/11970

Cronin, M. T. D., & Andrea-Nicole, R. (2017). *Relationship Between Adverse Outcome Pathways and Chemistry-Based in Silico Models to Predict Toxicity*.

Daghighi, A., Casanola-Martin, G. M., Timmerman, T., Milenković, D., Lučić, B., & Rasulev, B. (2022). In Silico Prediction of the Toxicity of Nitroaromatic Compounds: Application of Ensemble Learning QSAR Approach. *Toxics*, *10*(12). https://doi.org/10.3390/toxics10120746

De, P., Kar, S., Ambure, P., & Roy, K. (2022). Prediction reliability of QSAR models: An overview of various validation tools. *Archives of Toxicology*, *96*(5), 1279-1295. https://doi.org/10.1007/s00204-022-03252-y

D'Souza, S., Prema, K. V., & Balaji, S. (2021). Hierarchical Modeling of Binding Affinity Prediction Using Machine LearningTechniques. *2021 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, 61-65. https://doi.org/10.1109/DISCOVER52564.2021.9663690

Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences*, *42*(6), 1273-1280. https://doi.org/10.1021/ci010132r

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226-231.

Fix, E., & Hodges, J. L. (1951). *Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties*. USAF School of Aviation Medicine.

Gadaleta, D., Manganelli, S., Roncaglioni, A., Toma, C., Benfenati, E., & Mombelli, E. (2018). QSAR Modeling of ToxCast Assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. *Journal of Chemical Information and Modeling*, *58*(8), 1501-1517. https://doi.org/10.1021/acs.jcim.8b00297

Gadaleta, D., Spînu, N., Roncaglioni, A., Cronin, M. T. D., & Benfenati, E. (2022). Prediction of the Neurotoxic Potential of Chemicals Based on Modelling of Molecular Initiating Events Upstream of the Adverse Outcome Pathways of (Developmental) Neurotoxicity. *International Journal of Molecular Sciences*, *23*(6). https://doi.org/10.3390/ijms23063053

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-,

Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(4), 463-484. https://doi.org/10.1109/TSMCC.2011.2161285

Gareth James, D. W., Trevor Hastie, Robert Tibshirani. (2013). *An introduction to statistical learning: With applications in R*. New York : Springer, [2013] ©2013. https://search.library.wisc.edu/catalog/9910207152902121

Gedeck, P., Skolnik, S., & Rodde, S. (2017). Developing Collaborative QSAR Models Without Sharing Structures. *Journal of Chemical Information and Modeling*, *57*(8), 1847-1858. https://doi.org/10.1021/acs.jcim.7b00315

Globally Harmonized System of Classification and Labelling of Chemicals (GHS Rev. 9, 2021) | UNECE. (s. f.). Recuperado 8 de junio de 2023, de https://unece.org/transport/standards/transport/dangerous-goods/ghs-rev9-2021

Golalipour, K., Akbari, E., Hamidi, S. S., Lee, M., & Enayatifar, R. (2021). From clustering to clustering ensemble selection: A review. *Engineering Applications of Artificial Intelligence*, *104*, 104388. https://doi.org/10.1016/j.engappai.2021.104388

Grenet, I., Merlo, K., Comet, J.-P., Tertiaux, R., Rouquié, D., & Dayan, F. (2019). Stacked Generalization with Applicability Domain Outperforms Simple QSAR on in Vitro Toxicological Data. *Journal of Chemical Information and Modeling*, *59*(4), 1486-1496. https://doi.org/10.1021/acs.jcim.8b00553

Hanser, T., Barber, C., Rosser, E., Vessey, J. D., Webb, S. J., & Werner, S. (2014). Self organising hypothesis networks: A new approach for representing and structuring SAR knowledge. *Journal of Cheminformatics*, *6*(1), 21. https://doi.org/10.1186/1758-2946-6-21

Hanser, T., Steinmetz, F. P., Plante, J., Rippmann, F., & Krier, M. (2019). Avoiding hERG-liability in drug design via synergetic combinations of different (Q)SAR methodologies and data sources: A case study in an industrial setting. *Journal of Cheminformatics*, *11*(1), 9. https://doi.org/10.1186/s13321-019-0334-y

Hartung, T. (2018). Perspectives on In Vitro to In Vivo Extrapolations. *Applied in Vitro Toxicology*, *4*(4), 305-316. https://doi.org/10.1089/aivt.2016.0026

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

He, S., Ye, T., Wang, R., Zhang, C., Zhang, X., Sun, G., & Sun, X. (2019). An In Silico Model for Predicting Drug-Induced Hepatotoxicity. *International Journal of Molecular Sciences*, *20*(8), Article 8. https://doi.org/10.3390/ijms20081897

Heo, S., Safder, U., & Yoo, C. (2019). Deep learning driven QSAR model for environmental toxicology: Effects of endocrine disrupting chemicals on human health. *Environmental Pollution*, *253*, 29-38. https://doi.org/10.1016/j.envpol.2019.06.081

Heyndrickx, W., Mervin, L., Morawietz, T., Sturm, N., Friedrich, L., Zalewski, A., Pentina, A., Humbeck, L., Oldenhof, M., Niwayama, R., Schmidtke, P., Fechner, N., Simm, J., Arany, A., Drizard, N., Jabal, R., Afanasyeva, A., Loeb, R., Verma, S., … Ceulemans, H. (2022). MELLODDY: cross pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. *ChemRxiv*, *Cambridge*(Cambridge Open Engage). https://doi.org/10.26434/chemrxiv-2022-ntd3r

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55-67. https://doi.org/10.2307/1267351

Hoffmann, T., & Gastreich, M. (2019). The next level in chemical space navigation: Going far beyond enumerable compound libraries. *Drug Discovery Today*, *24*(5), 1148-1156. https://doi.org/10.1016/j.drudis.2019.02.013

Johnson, Stephen C. (1967). *Hierarchical clustering schemes | SpringerLink*. https://link.springer.com/article/10.1007/BF02289588

Kleinstreuer, N. C., Hoffmann, S., Alépée, N., Allen, D., Ashikaga, T., Casey, W., Clouet, E., Cluzel, M., Desprez, B., Gellatly, N., Göbel, C., Kern, P. S., Klaric, M., Kühnl, J., Martinozzi-Teissier, S., Mewes, K., Miyazawa, M., Strickland, J., van Vliet, E., … Petersohn, D. (2018). Non-animal methods to predict skin sensitization (II): An assessment of defined approaches. *Critical Reviews in Toxicology*, *48*(5), 359-374. https://doi.org/10.1080/10408444.2018.1429386

Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2017). *Federated Learning: Strategies for Improving Communication Efficiency*.

Kotsampasakou, E., & Ecker, G. F. (2017). Predicting Drug-Induced Cholestasis with the Help of Hepatic Transporters—An in Silico Modeling Approach. *Journal of Chemical Information and Modeling*, *57*(3), 608-615. https://doi.org/10.1021/acs.jcim.6b00518

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221-232. https://doi.org/10.1007/s13748-016-0094-0

Kwon, S., Bae, H., Jo, J., & Yoon, S. (2019). Comprehensive ensemble in QSAR prediction for drug discovery. *BMC Bioinformatics*, *20*, 521. https://doi.org/10.1186/s12859-019-3135-4

Landrum, G., Tosco, P., Kelley, B., sriniker, gedeck, NadineSchneider, Vianello, R., Ric, Dalke, A., Cole, B., AlexanderSavelyev, Swain, M., Turk, S., N, D., Vaucher, A., Kawashima, E., Wójcikowski, M., Probst, D., godin,

guillaume, … DoliathGavid. (2020). *RDKit* (Release_2020_03_1) [Software]. Zenodo. https://doi.org/10.5281/zenodo.3732262

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, *18*(1), 559-563.

Li, H., Cui, Y., Liu, Y., Li, W., Shi, Y., Fang, C., Li, H., Gao, T., Hu, L., & Lu, Y. (2018). Ensemble Learning for Overall Power Conversion Efficiency of the All-Organic Dye-Sensitized Solar Cells. *IEEE Access*, *6*, 34118-34126. https://doi.org/10.1109/ACCESS.2018.2850048

Liew, C. Y., Lim, Y. C., & Yap, C. W. (2011). Mixed learning algorithms and features ensemble in hepatotoxicity prediction. *Journal of Computer-Aided Molecular Design*, *25*(9), 855-871. https://doi.org/10.1007/s10822-011-9468-3

Liu, L., Wu, X., Li, S., Li, Y., Tan, S., & Bai, Y. (2022). Solving the class imbalance problem using ensemble algorithm: Application of screening for aortic dissection. *BMC Medical Informatics and Decision Making*, *22*, 82. https://doi.org/10.1186/s12911-022-01821-w

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

March-Vila, E., Ferretti, G., Terricabras, E., Ardao, I., Brea, J. M., Varela, M. J., Arana, Á., Rubiolo, J. A., Sanz, F., Loza, M. I., Sánchez, L., Alonso, H., & Pastor, M. (2023). A continuous in silico learning strategy to identify safety liabilities in compounds used in the leather and textile industry. *Archives of Toxicology*, *97*(4), 1091-1111. https://doi.org/10.1007/s00204-023-03459-7

Martin, E. J., & Zhu, X.-W. (2021). Collaborative Profile-QSAR: A Natural Platform for Building Collaborative Models among Competing Companies. *Journal of Chemical Information and Modeling*, *61*(4), 1603-1616. https://doi.org/10.1021/acs.jcim.0c01342

Matveieva, M., & Polishchuk, P. (2021). Benchmarks for interpretation of QSAR models. *Journal of Cheminformatics*, *13*(1), 41. https://doi.org/10.1186/s13321-021-00519-x

McMahan, H. B., Moore, E., Ramage, D., & y Arcas, B. A. (2016). Federated Learning of Deep Networks using Model Averaging. *ArXiv*, *abs/1602.0*.

Megahed, F. M., Chen, Y.-J., Megahed, A., Ong, Y., Altman, N., & Krzywinski, M. (2021). The class imbalance problem. *Nature Methods*, *18*(11), Article 11. https://doi.org/10.1038/s41592-021-01302-4

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*(1), 92-122. https://doi.org/10.1007/s10618-012-0295-5

Mirams, G. R., Davies, M. R., Brough, S. J., Bridgland-Taylor, M. H., Cui, Y., Gavaghan, D. J., & Abi-Gerges, N. (2014). Prediction of Thorough QT study results using action potential simulations based on ion channel screens. *Journal of Pharmacological and Toxicological Methods*, *70*(3), 246-254. https://doi.org/10.1016/j.vascn.2014.07.002

Pastor, M., Gómez-Tamayo, J. C., & Sanz, F. (2021). Flame: An open source framework for model development, hosting, and usage in production environments. *Journal of Cheminformatics*, *13*(1), 31. https://doi.org/10.1186/s13321-021-00509-z

Petch, J., Di, S., & Nelson, W. (2022). Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. *Canadian Journal of Cardiology*, *38*(2), 204-213. https://doi.org/10.1016/j.cjca.2021.09.004

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, *6*(3), 21-45. https://doi.org/10.1109/MCAS.2006.1688199

Raies, A. B., & Bajic, V. B. (2016). In silico toxicology: Computational methods for the prediction of chemical toxicity. *Wiley Interdisciplinary Reviews. Computational Molecular Science*, *6*(2), 147-172. https://doi.org/10.1002/wcms.1240

Rodríguez-Belenguer, P., Kopańska, K., Llopis-Lorente, J., Trenor, B., Saiz, J., & Pastor, M. (2023). Application of Machine Learning to improve the efficiency of electrophysiological simulations used for the prediction of drug-induced ventricular arrhythmia. *Computer Methods and Programs in Biomedicine*, 107345. https://doi.org/10.1016/j.cmpb.2023.107345

Rogers, D., & Hahn, M. (2010). Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, *50*(5), 742-754. https://doi.org/10.1021/ci100050t

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1), 1-39. https://doi.org/10.1007/s10462-009-9124-7

Ross, D. (1989). Mechanistic Toxicology: A Radical Perspective*. *Journal of Pharmacy and Pharmacology*, *41*(8), 505-511. https://doi.org/10.1111/j.2042-7158.1989.tb06516.x

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), Article 6088. https://doi.org/10.1038/323533a0

Samanipour, S., O'Brien, J. W., Reid, M. J., Thomas, K. V., & Praetorius, A. (2022). From Molecular Descriptors to Intrinsic Fish Toxicity of Chemicals: An Alternative Approach to Chemical Prioritization. *Environmental Science & Technology*. https://doi.org/10.1021/acs.est.2c07353

Sapounidou, M., Norinder, U., & Andersson, P. L. (2023). Predicting Endocrine Disruption Using Conformal Prediction – A Prioritization Strategy to Identify Hazardous Chemicals with Confidence. *Chemical Research in*

Toxicology, *36*(1), 53-65. https://doi.org/10.1021/acs.chemrestox.2c00267

Schneider, G., Neidhart, W., Giller, T., & Schmid, G. (1999). “Scaffold-Hopping” by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angewandte Chemie International Edition*, *38*(19), 2894-2896. https://doi.org/10.1002/(SICI)1521-3773(19991004)38:19<2894::AID-ANIE2894>3.0.CO;2-F

Selim, S. Z., & Ismail, M. A. (1984). K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(1), 81-87. https://doi.org/10.1109/TPAMI.1984.4767478

Simm, J., Humbeck, L., Zalewski, A., Sturm, N., Heyndrickx, W., Moreau, Y., Beck, B., & Schuffenhauer, A. (2021). Splitting chemical structure data sets for federated privacy-preserving machine learning. *Journal of Cheminformatics*, *13*(1), 96. https://doi.org/10.1186/s13321-021-00576-2

Smusz, S., Kurczab, R., & Bojarski, A. J. (2013). A multidimensional analysis of machine learning methods performance in the classification of bioactive compounds. *Chemometrics and Intelligent Laboratory Systems*, *128*, 89-100. https://doi.org/10.1016/j.chemolab.2013.08.003

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *Journal of Chemical Information and Computer Sciences*, *43*(2), 493-500. https://doi.org/10.1021/ci025584y

Russell, W.M.S, & Burch, R.L. (1960).The Principles of Humane Experimental Technique. *Medical Journal of Australia*, *1*(13), 500-500. https://doi.org/10.5694/j.1326-5377.1960.tb73127.x

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267-288.

Wang, L., Ding, J., Shi, P., Fu, L., Pan, L., Tian, J., Cao, D., Jiang, H., & Ding, X. (2021). Ensemble machine learning to evaluate the in vivo acute oral toxicity and in vitro human acetylcholinesterase inhibitory activity of organophosphates. *Archives of Toxicology*, *95*(7), 2443-2457. https://doi.org/10.1007/s00204-021-03056-6

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., & Bryant, S. H. (2009). PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, *37*(suppl_2), W623-W633. https://doi.org/10.1093/nar/gkp456

Wold, H. O. A. (1968). *Nonlinear Estimation by Iterative Least Square Procedures*.

Wold, S., & Sjöström, M. (1977). SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. En *Chemometrics: Theory and Application* (Vol. 52, pp. 243-282). AMERICAN CHEMICAL SOCIETY. https://doi.org/10.1021/bk-1977-0052.ch012

Wu, Z., Zhu, M., Kang, Y., Leung, E. L.-H., Lei, T., Shen, C., Jiang, D., Wang, Z., Cao, D., & Hou, T. (2021). Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. *Briefings in Bioinformatics*, *22*(4), bbaa321. https://doi.org/10.1093/bib/bbaa321

Yu, T.-H., Su, B.-H., Battalora, L. C., Liu, S., & Tseng, Y. J. (2022). Ensemble modeling with machine learning and deep learning to provide interpretable generalized rules for classifying CNS drugs with high prediction power. *Briefings in Bioinformatics*, *23*(1), bbab377. https://doi.org/10.1093/bib/bbab377

Yuan, H., Wang, Y., & Cheng, Y. (2007). Local and Global Quantitative Structure−Activity Relationship Modeling and Prediction for the Baseline Toxicity. *Journal of Chemical Information and Modeling*, *47*(1), 159-169. https://doi.org/10.1021/ci600299j

Zhou, W., Dai, Z., Chen, Y., Wang, H., & Yuan, Z. (2012). High-Dimensional Descriptor Selection and Computational QSAR Modeling for Antitumor Activity of ARC-111 Analogues Based on Support Vector Regression (SVR). *International Journal of Molecular Sciences*, *13*(1), 1161-1172. https://doi.org/10.3390/ijms13011161

Capítulo 2

Integrating Mechanistic and Toxicokinetic Information in Predictive Models of Cholestasis

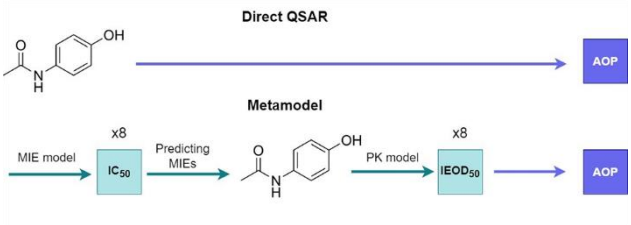
Pablo Rodríguez-Belenguer[1,2], Victor Mangas-Sanjuan[2,3], Emilio Soria-Olivas[4], Manuel Pastor[1]

[1]Research Programme on Biomedical Informatics (GRIB), Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Hospital del Mar Medical Research Institute, 08003 Barcelona, Spain.

[2]Department of Pharmacy and Pharmaceutical Technology and Parasitology, Universitat de València, 46100 Valencia, Spain.

[3]Interuniversity Research Institute for Molecular Recognition and Technological Development, Universitat Politècnica de València, 46100 Valencia, Spain.

[4]IDAL, Intelligent Data Analysis Laboratory, ETSE, Universitat de València, 46100 Valencia, Spain.



Revista: Journal of Chemical Information and Modeling

Editorial: American Chemical Society

Año: 2023

Cuartil: Q1

IF: 5.6

Introduction

There is an urgent need to replace, reduce, and refine (3Rs) animal experimentation. The knowledge obtained from past *in vivo* experiments can be reused to minimise the need to perform new assays, promoting sustainable science. New approach methodologies (NAMs) constitute an attractivealternative for assessing chemical hazards and estimating the effects of exposure, with the potential to support Toxicological Next Generation Risk Assessment (NGRA) and to promote the application of the 3R principles. Among the different approximations encompassed by the NAM term [1,2]*, in silico* methods are highly convenient on their own or as a complement to *in vitro* techniques.

While *in silico* toxicology (IST) offers benefits in terms of cost-effectiveness, high throughput, and ethical considerations, its ability to predict complex biological endpoints is still under debate [3]. Another difficulty is their integration with experimental data for risk assessment purposes, particularly in regulatory setups [4–7].

Quantitative structure-activity relationship (QSAR) is one of the most used methodologies in the IST field. It has been successfully used to predict *in vitro* results and simple toxicological endpoints [8,9]. However, the predictivity of QSAR models becomes limited when it comes to complex biological endpoints, such as organ toxicity. This is because complex biological endpoints result from multiple mechanisms and effects at different biological levels, making it more challenging to predict them accurately with QSAR. Additionally, QSAR models have only local validity, and the low structural similarity between the compounds in the validation and training sets can result in poor predictive performance [10]. Another drawback of the QSAR models is that, usually, they do not consider pharmacokinetic (PK) information, such as the absorption,

distribution, metabolism, and excretion (ADME) properties of compounds [11], and they might have difficulties to characterise the actual chemical risk of a compound since the toxicity of a compound is linked to the exposure [12]. QSAR methods can be important in transitioning to mechanism-based toxicology [13]. In this quest, Adverse Outcome Pathways (AOPs) have been developed to integrate existing mechanistic knowledge into a rational framework [14]. AOP connects known biological events linearly through a series of Key Events (KEs) from a Molecular Initiating Event (MIE) to the final Adverse Outcome (AO). The causal relationships between these KEs are defined by Key Event Relationships (KERs).

In 2013, the Organization for Economic Co-operation and Development (OECD) published the first version of the Guidance Document on Developing and Assessing Adverse Outcome Pathways with a conceptual background [15], followed by the publication of the User's Handbook Supplement in 2018 [16]. This supplement provides practical guidance and advice on applying AOPs in the context of risk assessment and highlights the benefits of using a mechanistic approach to comprehend adverse effects better. Moreover, this supplement contains practical instructions for AOP development and collaborative work on the databases AOP knowledgebase (AOP-KB)[17] and AOP-Wiki [18].

Computational methods could exploit the standardised knowledge representation that AOPs provide. Accordingly, *in silico* models with multiple molecular initiating events (MIEs) can be built to predict complex toxicological endpoints for which QSAR models do not provide quality results [19,20].

AOPs have also been incorporated in mechanistic-based toxicokinetic (TK)/toxicodynamic models that evaluate exposure-response relationships [21–23]. A common misconception is to consider that drugs with a very small IC_{50}

are "more toxic". However, this is not necessarily true as the likelihood and severity of adverse effects are more closely linked to the total amount of drug at the target site, rather than the drug's potency [24]. Even drugs with a high IC_{50} can cause toxicity if the dose administered in clinical use (the therapeutic dose [T_D]) is high enough. Therefore, to make decisions about the potential toxicity of drugs, IC_{50}s should be transformed to "point of departure doses" using quantitative *in vitro* to *in vivo* extrapolation (QIVIVE) models [25]. QIVIVE is derived from a minimal Physiological-based pharmacokinetic (PBPK) model, which reproduces the kinetic of a substance within a living organism over time, considering the main pharmacokinetic phenomena: absorption, distribution, metabolism, and excretion. The main objective of QIVIVE is to establish the *in vivo* dose which will produce a certain concentration in the blood (or tissues). This can correspond to *in vitro* concentrations like the half-maximal effective concentration (EC_{50}), IC_{50}, or half-maximal active concentration (AC_{50}). In this sense, QIVIVE can be considered a "reverse dosimetry" method, providing doses from concentrations.

In this work, we aim to develop a novel approach that integrates the contribution of multiple MIEs and the compound TK properties for the prediction of a complex toxicological endpoint. In this study, we will use hepatotoxicity as a representative example of a complex toxicological endpoint. Drug-induced liver injury (DILI) is one of the primary causes of attrition during clinical and preclinical studies and one of the main reasons for drug withdrawal from the market [26,27]. DILI can be categorised as either idiosyncratic or non-idiosyncratic based on its relationship with the drug dose. If DILI occurs independently of the dose, it is considered idiosyncratic, while it is considered non-idiosyncratic if DILI is dose-dependent. Non-idiosyncratic DILI can be classified into the following three categories [26] i) hepatocellular, ii)

cholestatic, and iii) mixed. Because DILI is a very broad endpoint, we will focus on cholestasis.

Cholestatic DILI is a dose-dependent adverse effect defined as a disruption of the bile flow, which increases hepatic bile acid concentrations, resulting in necrosis and/or apoptosis. Together with hepatocellular, it is one of the most severe manifestations of DILI [26,28]. Cholestasis is often produced by inhibiting the hepatic transporters responsible for facilitating bile flow from the liver to the small intestine [26]. Hepatic transporters are classified according to their location in the membranes: those belonging to the canalicular membrane and those belonging to the basolateral membrane. Canalicular membrane transporters regulate hepatic clearance, as well as the secretion of bile salts and conjugates into the bile. Basolateral membrane transporters regulate the uptake of drugs and transport endobiotics and xenobiotics from the blood to the hepatocyte [26].

Bile Salt Export Pump (BSEP), multidrug resistance-associated protein (MRP2), Breast cancer resistance protein (BCRP), and P-glycoprotein (P-gp) are canalicular membrane transporters [28–30], while MRP3, MRP4, and organic anion transporting polypeptides (OATP1B1 and OATP1B3) are basolateral membrane transporters. The role of BSEP inhibition is one of the most important mechanisms studied in cholestasis occurrence, being the main MIE described in the cholestasis AOP found in the AOP-wiki [31].

This study aims to add to existing QSAR methodologies a new approach which integrates mechanistic information for multiple MIE (using AOPs) and TK information (using QIVIVE models), providing a more complete and realistic description of the phenomenon studied. This approach will be illustrated by applying it to the prediction of the cholestatic properties of a series of compounds. The results of this case study will be used to discuss its

advantages compared to direct QSAR modelling, especially in the most common situations in drug development, where the candidates do not have much structural resemblance with the structures in the training series.

Material and methods

Cholestasis dataset

A series of chemical compounds with cholestasis annotations was obtained from Kotsampasakou and Ecker (2017), where the researchers extracted the annotations from PubMed (http://www.ncbi.nlm.nih.gov/pubmed), Google, Scopus (https://www.scopus.com/), and the SIDER database v2 searching the terms: "drug-induced cholestasis" or "cholestasis". The data was curated by removing inorganic compounds and compounds containing metallic elements. In the end, the series consisted of 577 compounds with 130 "positives" (cholestatic compounds) and 447 "negatives" (non-cholestatic compounds).

For our study, we applied additional curation, eliminating compounds whose administration route is not oral nor intravenous since only these routes provide relatively simple and well-understood absorption and elimination pathways. The filtered dataset contained 437 compounds (116 positives and 321 negatives).

The compounds in this series were characterised using unique IDs to facilitate the extraction of data from other sources: ChEMBL IDs were obtained using chembl-webresource-client 0.10.8 [32], DSSTox substance IDs or DTXSIDs were assigned using PubChemPy 1.0.4 [33], and Drugbank IDs were obtained from Drugbank version 5.1.9 [34]. In addition, for chemical comparisons by pharmacological groups, information on Anatomical Therapeutic Chemical (ATC) classification was added up to the second level of information

(Pharmacological or Therapeutic subgroup) using chembl-webresource-client and MedCode 1.3 [35].

Transporter QSAR models

Sets of compounds annotated with the pIC_{50} values were extracted from ChEMBL version 29 to build QSAR models for the main hepatic transporters involved in drug-induced cholestasis (low-level models, LLM). The process of selecting compounds that inhibit specific transporters involved two filters: the target organism (*homo sapiens*) and the target type (single protein). No filtering based on assay type was implemented to avoid compromising the number of compounds selected. This decision was made taking into account that a higher number of assays might introduce variability due to differences in experimental conditions and measurement techniques [36]. Compounds for which multiple experimental annotations were available were included as multiple data points. This procedure has the advantage of giving more weight to multiple-tested compounds and incorporating experimental variability, in contrast with alternative procedures in which a single mean or median is used to characterize their biological properties. The structures were standardised using a curation tool [37], removing inorganic compounds and compounds with metallic elements. Table 1 shows the hepatic transporters considered, with detailed information (transporter names with their target ChEMBL ID, acronym, number of compounds, and the mean and standard deviation [std] of the pIC_{50} distributions for each selected transporter) on the data extracted. To match the transporter inhibition data with the *in vivo* cholestasis data described above, we used the ChEMBL ID of the compounds.

Table 1: Information about compounds collected for each hepatic transporter.

Transporter	Acronym	N[a]	Mean_pIC$_{50}$	Std_pIC$_{50}$
P-glycoprotein (CHEMBL4302)	P-gp	1031	5.89	1.11
Breast cancer resistance protein (CHEMBL5393)	BCRP	1015	5.99	0.75
Organic anion transporting polypeptide 1 (CHEMBL1697668)	OATP1B1	63	5.49	0.61
Organic anion transporting polypeptide 3 (CHEMBL1743121)	OATP1B3	25	5.13	0.77
Multidrug resistance-associated protein 4 (CHEMBL1743128)	MRP4	106	4.70	0.47
Multidrug resistance-associated protein 2 (CHEMBL5748)	MRP2	57	4.69	0.42
Bile salt export pump (CHEMBL6020)	BSEP	361	4.68	0.51
Multidrug resistance-associated protein 3 (CHEMBL5918)	MRP3	43	4.52	0.45

[a]N is the number of compounds in the training series of each LLM.

Figure 1 displays the violin plots showing the distributions of pIC$_{50}$ values for the selected transporters. Consistent with the information in Table 1, the mean pIC$_{50}$ falls within the range of 4.5-6 for each transporter, with P-gp and BCRP exhibiting the highest means. Likewise, P-gp also exhibited the highest standard deviation likely due to the inclusion of a larger number of diverse assays conducted to this particular transporter.

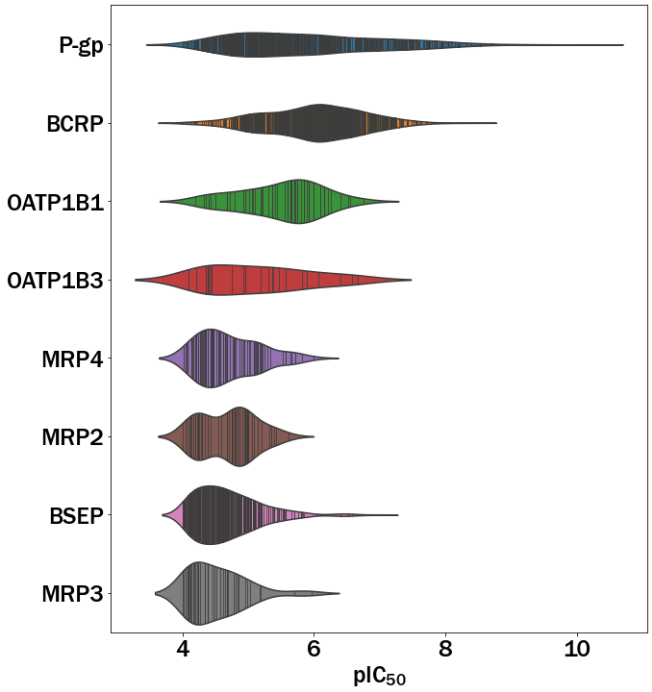


Figure 1: Violin plots of the pIC_{50} distributions of the eight hepatic selected transporters.

For each LLM, we obtained all compounds with IC_{50} annotations and developed a QSAR model using the pIC_{50} as the dependent variable. Morgan fingerprints (FP) (nbits=2048, radius=2, features=enabled) were computed using RDKit 2019.9.3 (Landrum 2016) and used as input variables for building four machine learning (ML) regression models for each LLM with scikit-learn version 0.24.1 [39]; XGBoost 1.4.2 (XGB) [40], Random Forest (RF) [41], K-nearest neighbours (KNN) [42], and Support Vector Machines (SVM) [43]. All models were trained using a grid search with 5-fold cross-validation (CV) to find the best hyperparameters based on the Mean Absolute Error (MAE) as the scoring metric. The model with the lowest MAE was selected for each of the LLMs. As part of our proposed hyperparameter grid for the SVM model, we included the linear kernel as an option in addition to the radial kernel, serving as an alternative to linear models. To ensure the robustness of the models, a 20-

Repeated 5-fold CV approach was employed for the model evaluation. The selection of twenty repetitions was made considering that a Repeated k-fold CV requires fewer replicates than the total number of compounds available. As there were only twenty-five compounds collected for inhibiting the OATP1B3 transporter, twenty replicates were selected for the analysis throughout the entire article to maintain consistency in the methodology when using Repeated k-fold CV [44]. The information corresponding to the settings of these models is provided in the supporting information Table S1.

These models (LLM) were used to predict eight transporter pIC_{50} values for the 437 compounds belonging to the cholestasis dataset. For compounds with known experimental activity, the mean of all available experimental values was used instead of the predictions. The final matrix contains 437 rows (compounds) and 8 columns (transporters).

In vitro to *in vivo* extrapolations

In vivo half maximal inhibitory equivalent oral doses ($IEOD_{50}$) were calculated from the IC_{50} values by applying QIVIVE methods, translating concentrations into *in vivo* doses. For calculating $IEOD_{50}$s, we used the High-Throughput Toxicokinetics (httk 2.1.0) library [45]. Monocompartmental (MC) models were built using the default parameters provided by the httk library. Order 1 kinetics assumes that the drug concentration in the body can be described by a single compartment, which is appropriate for drugs that distribute rapidly and evenly throughout the body, under the assumption that the effect of a peripheral distribution is negligible at a steady state. QSAR models usually assume that the compound is at a steady state without considering any time-dependent processes that may affect the drug concentration. Hence, after computing the steady-state concentrations (C_{ss}) using the MC model, the next step was to calculate the $IEOD_{50}$ for each drug that inhibits each transporter. The $IEOD_{50}$ is directly proportional to the in vitro IC_{50} and inversely proportional to C_{ss} [46].

The httk library [45] can compute the percentile of the specified $IEOD_{50}$ for the model. In our case, we obtained the 90th percentile as the larger the percentile predicted C_{ss} from the MC model, the lower $IEOD_{50}$, due to the inverse relationship between C_{ss} and $IEOD_{50}$. This approach is considered to be the most conservative as cholestasis is a dose-dependent adverse outcome, and any compound with a therapeutic dose (T_D) above the highest $IEOD_{50}$ among the selected transporters would be considered cholestatic. The information about the therapeutic doses was obtained by matching the Drugbank IDs in the cholestasis dataset with the corresponding entries in the Drugbank database (other sources of information consulted were: drugs [https://www.drugs.com/] and Medscape [https://reference.medscape.com/]).

The calculation of $IEOD_{50}$ requires obtaining physicochemical parameters such as molecular weight (MolWt), log P (octanol-water partition coefficient), and PK parameters such as intrinsic clearance (Cl_{int}) and plasma-unbound fraction (f_{ub}). For the compounds in the cholestasis dataset, MolWt and log P were computed using RDKit. Experimental $Cl_{int,}$ and f_{ub} values were extracted from the httk databases (only drugs that have been experimentally tested with human hepatocyte cells), using DTXSIDs to identify the compounds in both datasets whenever possible. For the rest of the compounds, these values were predicted using OPERA version 2.9 [47]. Compounds that were unable to have either f_{ub} or Cl_{int} values calculated by OPERA were eliminated from the cholestasis dataset. As a result, the dataset contained a total of 426 compounds, with 115 classified as positive and 311 as negative. Finally, compounds with $IEOD_{50}$ values larger than 10000 mg/kg/day (only 7 compounds were in this category) were removed from the cholestasis dataset since we consider that these doses are not realistic from a physiological point of view. After this elimination, the final dataset was reduced to 419

compounds (114 positives and 305 negatives). To further clarify the procedure, the filtering steps are summarized in Figure 2.

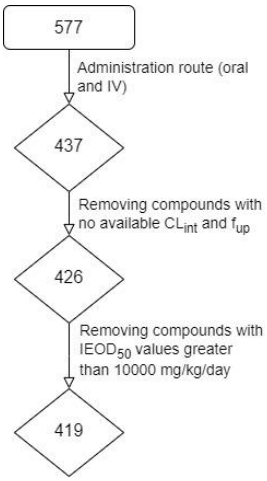


Figure 2: Cholestasis dataset filters**.**

Cholestasis model building

The cholestasis models were built using the series described above, using two different strategies: direct QSAR modelling and combining the predictions provided by the LLM (obtaining a metamodel). In the latter approach, we generated two different metamodels for assessing the advantages of incorporating PK information. In the metamodel incorporating PK information (Metamodel_pk), a compound was considered positive when its T_D was n times higher than the predicted $IEOD_{50}$ for any of the considered transporter, where n is a factor adjusted to balance the sensitivity and specificity of the metamodel. Regarding metamodel not incorporating PK information (Metamodel_not_pk), this model used IC_{50} information exclusively. A compound was classified as positive if the IC_{50} of any of the transporters was ≤300 μM, according to [48–51]. Both metamodels were constructed as scikit-learn estimators to fully utilize the functionalities of the scikit-learn library.

Direct QSAR models were built using physicochemical descriptors (PC) and FP as predictor variables, obtained using RDKit. The following algorithms were used: XGB, RF, Naïve Bayes approach (Multinomial Naïve Bayes [MNB] for FP and Gaussian Naïve Bayes [GNB] for PC descriptors) [52], and SVM. Like in the regression models, we added the linear kernel within the hyperparameter grid for SVM classification models.

In order to find the best hyperparameters, all models underwent a grid search with a 5-fold CV, utilizing the the Area Under the Receiver Operating Characteristic Curve (ROC AUC) score as the scoring metric. For the QSAR models, the algorithm achieving the highest ROC AUC among the four tested models was chosen as the optimal choice (Tables S2 and S3 provides further details).

Model evaluation

The model quality was evaluated using the model's sensitivity (S), specificity (SP), accuracy (A), Matthews correlation coefficient (MCC), and ROC AUC.

Comparison between Repeated k-fold and "similarity-based cross-validation" performances

To incorporate the similarity in the assessment of the model predictivity, we compared the results obtained with a standard 20-Repeated 5-fold CV [44] with a modified CV algorithm where the groups contain structurally dissimilar compounds. So, if the predictive power of a model is lower when using the "similarity-based CV", it would indicate that it is worse for predicting when the compounds in the test series are more structurally different from those in the training series. For this modified version of CV, we applied a hierarchical clustering to obtain five clusters (Cluster 1=51 compounds, Cluster 2=174 compounds, Cluster 3=61 compounds, Cluster 4= 79 compounds, and Cluster

5=54 compounds) using fingerprints as input variables and the Jaccard distance as the evaluation metric. The same number of folds was established for both types of CVs to allow a fair comparison. Each fold in the similarity 5-fold CV was trained using four clusters and validated with the remaining cluster, thus predicting compounds with low structural similarity to the training set of that fold. For detailed information on the search for optimal hyperparameter sets using both 20-Repeated 5-fold CV (Table S2.A in the supporting information) and similarity 5-fold CV (Table S2.B in the supporting information), refer to the supporting information. These tables provide further insights into the process of identifying the best hyperparameters for both types of cross-validation.

Likewise, intra- and inter-cluster similarities were evaluated using FP descriptors and the Tanimoto similarity metric. The mean similarity value of the three most similar compounds was computed intra- and inter-cluster. The supporting information's Figure S1 presents a heatmap displaying the Tanimoto similarity values for both intra- and inter-cluster comparisons. The similarity values intra-clusters showed minimal differences, ranging from 0.41 (Cluster 2) to 0.47 (Cluster 1). Regarding to the comparison inter-clusters, the similarity values ranged from 0.17 (Cluster 1-Cluster 5) to 0.31 (Cluster 2-Cluster 4). The observed results suggest that intra-cluster similarity outweighed inter-cluster similarity, indicating that this methodology has the potential to be worthy in evaluating the structural independence of the proposed approach.

Performances according to the "ATC-based cross-validation"

Such as mentioned above, our study intends to determine if the proposed methodology has advantages with respect to other approaches in terms of predictive quality when the compounds are different from those in the

models' training series. With this aim, complementing the "similarity-based cross-validation" described above, we applied a cross-validation procedure where drugs used in certain therapeutic areas (as identified by their ATC codes) are used to predict compounds used in different therapeutic areas. We started by compiling the ATCs for the compounds in our series for the five most represented ATCs: J01 (antibacterials for systemic use), N05 (psycholeptics), L01 (antineoplastic agents), C01 (cardiac therapy) and N02 (analgesics), as shown in Table 2. So, we conducted an ATC 5-fold CV, where each fold involved training on compounds from four of the five ATC and predicting the validation set of compounds from the remaining ATC. Additional information regarding the optimal hyperparameters for each evaluated model can be found in Table S3 of the supporting information. Also, we calculated the intra- and inter-ATC group similarities in the same way as before used for computing the similarities described above for the similarity 5-fold CV method. Within the same ATC code, molecular similarities ranged from 0.22 (L01) to 0.54 (J01), as shown in Figure S2 of the supporting information. When comparing compounds from different ATC codes, similarities ranged from 0.16 (J01-C01, N05-J01, L01-C01) to 0.27 (N05-N02). Similarly, to our previous CV strategy, intra-ATC similarity was found to be higher than inter-ATC similarity, justifying the use of this evaluation method, following the same approach as the previous one, for building models based on splits with high dissimilarity. This allows us to further verify that the performance of our proposed model is less dependent on the structural similarity between the training and test series than a direct QSAR model.

Table 2: Summary information of the top five ATC codes.

ATC	Number of compounds by class	Most common pharmacological groups

J01 (antibacterials for systemic use)	# Cholestatic compounds=20 # Non-cholestatic compounds=17	B-lactams and Penicillins
N05 (psycholeptics)	# Cholestatic compounds=12 # Non-cholestatic compounds=25	Psycholeptics and hypnotics
L01 (antineoplastic agents)	# Cholestatic compounds=6 # Non-cholestatic compounds=16	Alkylating agents ad plant alkaloids
C01 (cardiac therapy)	# Cholestatic compounds=3 # Non-cholestatic compounds=15	Cardiac stimulants and anthyarrhymics
N02 (analgesics)	# Cholestatic compounds=2 # Non-cholestatic compounds=15	Antimigraine and opioids

Statistical analyses

Student's t-tests at a 95% confidence level were used to determine whether there are statistically significant differences in the "Lipinski's rules of five" (Lipinski et al. 1997) variables between the positive and negative classes. This analysis was complemented with a two-way ANOVA to determine whether the effect of the transported type and the target class (as fixed factors) have a statistically significant effect on the $IEOD_{50}$ at a 95% confidence level.

Software

Table 3 shows a summary of the main software libraries and packages used in this study.

Table 3: Packages with their version used and main applicability.

Package	Version	Applicability	Language	References
scikit-learn	0.24.1	ML		[39]
numpy	1.19.5	Vector operations		[53]
statsmodels	0.12.2	Statistics		[54]
seaborn	0.11.1	Visualisation		[55]
matplotlib	3.3.4	Visualisation		[56]
RDKit	2019.9.3	Chemical	python 3.6.13	[38]
pandas	1.1.5	Dataframe operations		[57]
chembl-webresource-client	0.10.8	ChEMBL requests		[32]
PubChemPy	1.0.4	PubChem requests		[33]
MedCode	1.3	ATC codes		36
XGBoost	1.4.2	Boosting model		[40]
httk	2.1.0	Pharmacokinetic	R 4.2.1	[45]

Results and discussion

Overview

To detect potential differences between cholestatic and not cholestatic compounds due to physicochemical properties, we run a preliminary study using some of the variables used by "Lipinski's rules of five" [58].

The approach described here was based on physiological knowledge, where we constructed models for simpler phenomena (MIEs) that represent relevant components of the complex endpoint (AO) and combined the predictions incorporating toxicokinetic considerations. We started by gathering existing

information on the biological processes involved in this endpoint from the AOP wiki. Then, we developed QSAR models for each of the hepatic transporters identified as relevant MIEs: P-gp, BCRP, OATP1B1, OATP1B3, MRP4, MRP2, BSEP, and MRP3 (see Table 1), as described in the Methods Section. The predicted *in vitro* inhibitory information was exploited by applying logical rules. A simple logical OR on this prediction matrix was used to label compounds showing inhibitory activity for any of these transporters as a potential cholestatic compound (second part of Figure 3.A).

However, this approach has the limitation that the inhibitory *in vitro* concentrations obtained by the models can be non-representative of the ones reached in therapeutics due to differences in clearance, protein binding, bioavailability, and other pharmacokinetic parameters. For this reason, we incorporated toxicokinetic considerations to obtain $IEOD_{50}$s (representing *in vivo* doses) from the predicted IC_{50}s (representing *in vitro* data) using QIVIVE models. The proposed workflow (Figure 3.B) starts applying a PBPK model to obtain C_{ss} from the input pIC_{50}. Then, the QIVIVE approach allows obtaining $IEOD_{50}$ from the C_{ss}. Finally, the $IEOD_{50}$ are compared with T_D (obtained from public sources, as described in the Methods Section), and we used a logical OR rule to label as cholestatic the compounds for which any of the T_D is larger than the highest $IEOD_{50}$ among all transporters (first part of Figure 3.A).

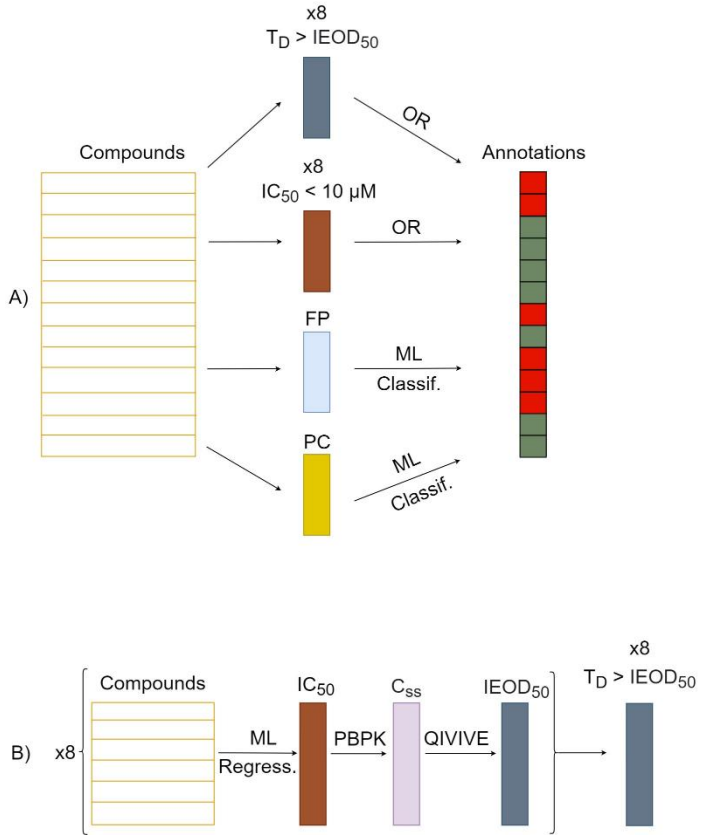


Figure 3: Scheme of the proposed methodology. A) High-level view of the four models being compared: Metamodel with PK information (grey), Metamodel without PK information (brown), direct QSAR with FP descriptors (light blue), and direct QSAR with PC (yellow). Red cells represent cholestatic compounds, and green cells represent non-cholestatic compounds. B) Scheme of the proposed workflow to introduce toxicokinetics in the modelling.

The results obtained using this approach were compared with a classical direct modelling method that uses compound structures to build QSAR models, using both FP and PC descriptors (summarised in the third and fourth sections of Figure 3.A), and predicted cholestasis directly without considering any mechanistic information.

The comparison of our approach with the direct QSAR models included an analysis of their performance using standard metrics but also an additional analysis for comparing their applicability for the prediction of dissimilar

compounds. This involves an evaluation of their predictive quality using "similarity-based" and "ATC-based" complementary CVs, carried out as described in the Methods Section.

Preliminary analyses

As a preliminary step, we studied possible differences in the physicochemical properties between the cholestatic and non-cholestatic compounds in the studied series of 419 compounds. Figure 4 shows the summary of density and scatter plots, separated by class for the compound's molecular weight (ExacMolWt), number of hydrogen bond acceptors (NumHAcceptors), number of hydrogen bond donors (NumHDonors), and log P (MolLogP). These are the properties represented by Lipinski's rules of five [58], which are known to describe important properties for the pharmacokinetic and pharmacodynamic characteristics of the compounds. The centre of the distribution is slightly higher for ExactMolWt, MolLogP, and NumHAcceptors in the set of cholestatic molecules compared to the non-cholestatic ones. However, according to the Student's t-test performed, the differences were not statistically significant for any of the properties studied at a 95% confidence level.

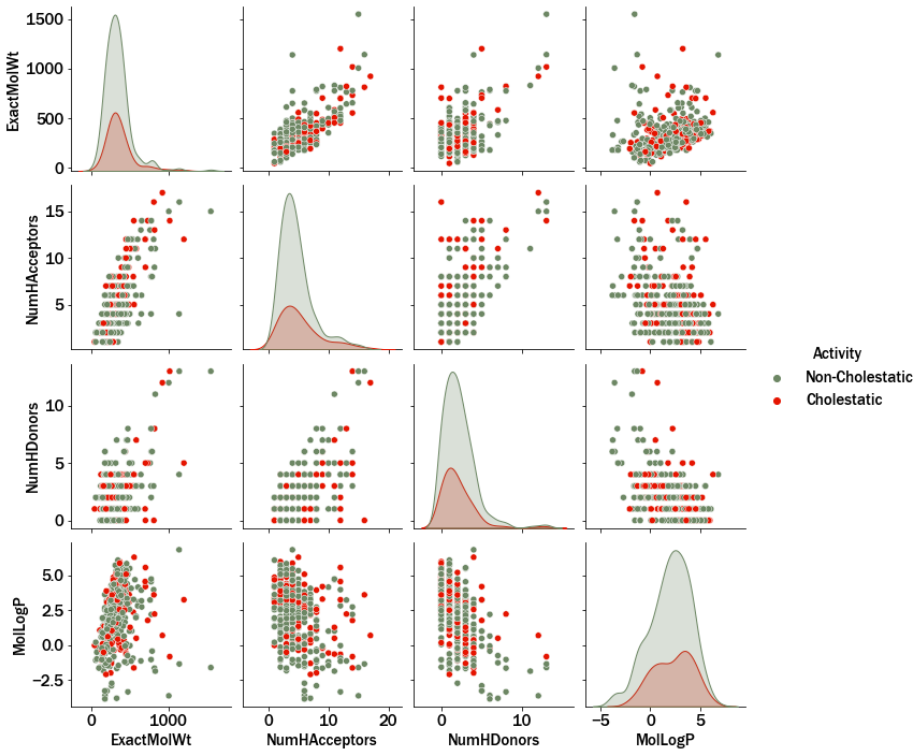


Figure 4: Distribution of Lipinski's rules of five: Molecular weight (ExactMolwt), Number of hydrogen bond donors (NumHDonors), Number of hydrogen bond acceptors (NumHAcceptors), and octanol-water partition coefficient log P (MolLogP).

Low-level models

Individual QSAR regression models for the eight transporters selected (Table 1) were built as described in the Methods Section. The violin plots in Figure 5 show the MAE distributions obtained from the 20-Repeated 5-CV for each of the eight low-level models. Particularly, the models for P-gp and OATP1B3 inhibition had the poorest performance. Regarding deviations between folds, P-gp's extensive data leads to minimal variations, while OATP1B3's limited data results in significant deviations. These findings emphasize the impact of variability between several assays and data availability on predictive performance, such as described above.

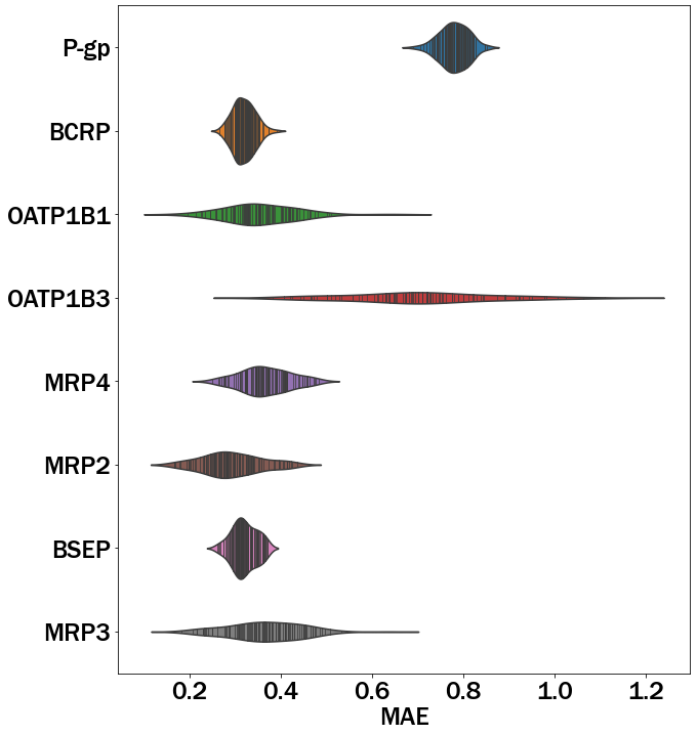


Figure 5: Violin plot with MAE obtained for each LLM.

Table 4 presents the mean and standard deviation of the twenty repetitions of the 5-fold CV. It reveals that P-gp (0.78) and OATP1B3 (0.68) had the highest mean MAEs, with a low standard deviation for P-gp (0.03) and a higher standard deviation for OATP1B3 (0.19). These observations align with the insights shared in the previous Figure 5. The remaining transporters exhibited similar mean MAE values. The models for BCRP and BSEP demonstrated less deviation between folds (akin to what was observed for P-gp), as these models had more training data compared to the others.

Table 4: Mean and std of MAEs obtained from the 20-Repeated 5-fold CV for the eight selected transporters.

Metrics	BCRP	MRP2	MRP3	MRP4	OATP1B1	OATP1B3	BSEP	P-gp
MAE_{mean}	0.32	0.29	0.36	0.36	0.36	0.68	0.33	0.78
MAE_{std}	0.02	0.06	0.08	0.05	0.07	0.19	0.03	0.03

Figure 6 depicts the box plot illustrating the predicted pIC_{50} distributions for each transporter. Notably, the worst performing models (P-gp and BCRP) exhibit similar values in their distributions. This finding could have the potential to impact the overall quality of the metamodels. It is important to note that these data impose an upper bound on the quality of the predictive models derived from them. While it could be tempting to push the model beyond this limit, doing so risks to produce model overfitting, compromising their predictive performance. The analysis showed statistically significant differences ($p<0.01$) in the pIC_{50} distributions between the different transporters and classes (two-way ANOVA, 95% confidence level, as described in the Methods Section).

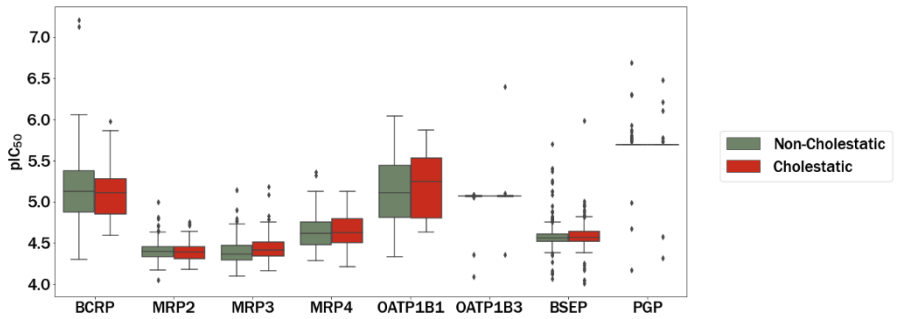


Figure 6: Box plots of the pIC_{50} distributions separated by class for each selected transporter.

Incorporating TK considerations

The predicted *in vitro* pIC_{50} cannot be expected to correlate directly with observed cholestatic outcomes without first transforming these to *in vivo* doses ($IEOD_{50}$) and then comparing these doses with the ones administered in clinical use. The first step, the computation of $IEOD_{50}$, was carried out using QIVIVE models, as described in the Methods Section. To evaluate the predictive power of the models built by OPERA for predicting f_{ub} and Cl_{int}, compounds of the cholestasis dataset with experimental values (from queries to the httk library) were predicted. Figure S3 in the supporting information

displays a scatter plot with the X-axis representing the experimental values extracted from httk, and the Y-axis showing the OPERA predictions for the same compounds for both f_{ub} (Figure S3.A in the supporting information) and Cl_{int} (Figure S3.B in the supporting information). In this Figure, minimal deviations between the actual values and the predictions can be observed. To further validate the predictive power, Table S4 in the supporting information presents the MAE values for both f_{ub} (MAE=0.07) and Cl_{int} (MAE=8.50), as well as the mean and standard deviation between the experimental values from httk and the predicted values from OPERA, which exhibit practically identical results. These results highlight the quality of the OPERA models in predicting pharmacokinetic parameters.

Figure 7 shows box plots with the $pIEOD_{50}$s ($-\log_{10}(IEOD_{50})$) of the eight selected transporters for cholestatic and non-cholestatic drugs, side by side. It can be seen that the median value of inhibitory potential for each transporter is nearly identical between cholestatic and non-cholestatic compounds. The analysis showed no statistically significant differences ($p<0.05$) in the $pIEOD_{50}$ distributions between the different transporters and classes (two-way ANOVA, 95% confidence level, as described in the Methods Section). The absence of statistical significance in the $pIEOD_{50}$ distributions between different transporter and classes, in contrast to the statistical significance observed in the pIC_{50} distributions, may be due to the fact that *in vitro* models can oversimplify or fail to fully capture the complexity of metabolic pathways that occur *in vivo*. Therefore, based on *in vivo* extrapolations the whole set of hepatic transporters could have the same relevance in predicting cholestasis.

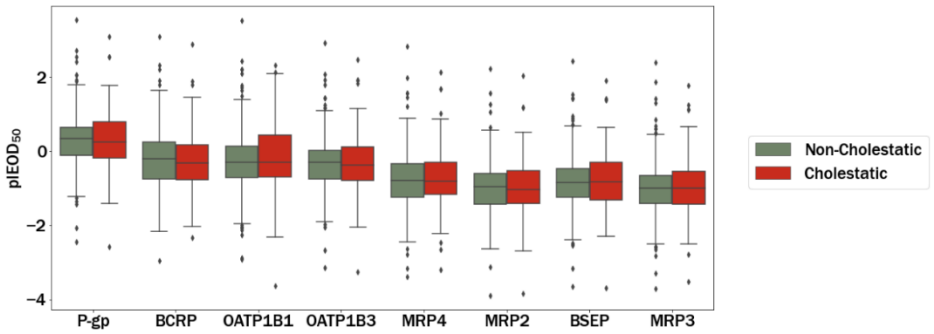


Figure 7: Box plot comparing the $pIEOD_{50}$ (potential of $IEOD_{50}$) predicted for the series studied and the eight transporters, separated by class. Transporters are shown in decreasing order, with respect to the median $pIEOD_{50}$.

The predictive quality of the metamodels and the direct QSAR models

The metamodel obtained using PK information (Metamodel_pk) was constructed by using a logical OR to combine the presence of T_D higher than the $IEOD_{50}$. In other words, a compound was predicted to be cholestatic if its T_D was higher than the predicted $IEOD_{50}$ for any of the selected hepatic transporters. For the metamodel without PK information (Metamodel_not_pk), a compound was predicted as cholestatic if any transporters had an IC_{50} below 300 µM. The direct QSAR models were built using fingerprints and physicochemical descriptors as predictor variables. The models and approaches utilized in this article have undergone a meticulous grid search, aiming to identify the optimal set of hyperparameters, such as described in Methods Section.

Comparison between Repeated k-fold and "similarity-based cross-validation" performances

The results of a direct QSAR model coming from a Repeated k-fold may be too optimistic and these results may not be representative of practical problems.

One of the main reasons is that, in real drug development applications, the structure of the new drug candidate is often very different from the structures of the training series. Therefore, to obtain a fairer comparison, we further evaluated the predictive quality of our models by assessing if they could accurately predict the properties of structurally diverse compounds. With this aim, we applied a "similarity-based CV" (described in detail in the Methods Section) where the series was split into five structurally dissimilar subgroups using hierarchical clustering. Then, we applied a similarity 5-fold CV where four subgroups were used to predict the remaining one, containing structurally dissimilar compounds. By comparing the predictive quality of this approach with those from a standard 20-Repeated 5-fold CV, where groups were assigned randomly, we can evaluate how dependent the structural similarity is on the prediction quality for all the studied models.

Table 5 presents the mean and standard deviation for both the 20-Repeated 5-fold CV and the similarity 5-fold CV.

Table 5: Mean and std of the Sensitivity (S), specificity (SP), AUC, MCC, and Accuracy (A) for each model in both the 20-Repeated 5-fold CV and Similarity 5-fold CV.

		S	S_std	SP	SP_std	AUC	AUC_std	MCC	MCC_std	A	A_std
20-Repeated 5-fold CV	**Metamodel_pk**	0.84	0,07	0.55	0,06	0.69	0,05	0.34	0,19	0.63	0,05
	Metamodel_not_pk	1.00	0,00	0.00	0,01	0.50	0,00	0.01	0,03	0.27	0,04
	QSAR_model_FP	0.34	0,11	0.87	0,05	0.60	0,05	0.23	0,12	0.72	0,05
	QSAR_model_PC	0.32	0,10	0.84	0,06	0.58	0,06	0.11	0,12	0.69	0,05
Similarity 5-fold CV	**Metamodel_pk**	0.81	0.06	0.54	0.15	0.67	0.09	0.29	0.14	0.62	0.10
	Metamodel_not_pk	0.06	0.06	0.98	0.02	0.52	0.03	0.07	0.10	0.74	0.09
	QSAR_model_FP	0.20	0.09	0.94	0.03	0.57	0.04	0.19	0.10	0.75	0.09
	QSAR_model_PC	0.24	0.14	0.90	0.04	0.57	0.05	0.16	0.11	0.74	0.06

Figure 8 displays violin plots for the sensitivity, specificity, MCC, AUC, and accuracy of each model based on the 20-Repeated 5-CV (left column) and in the similarity 5-fold CV (right column). Additionally, Table 5 provides a summary of the mean and standard deviation of the metrics presented in Figure 8. This Figure depicts how the average variability across folds was similar for both types of CV across different models (the std of the Metamodel_pk was lower than that of the QSAR models). The Metamodel_not_pk exhibited the least variability between folds for each metric.

The results observed in Figure 8 indicate that the Metamodel_pk was not affected by the decrease in structural similarity, as its sensitivity remained above 0.80 in both scenarios (Table 5). Concerning specificity, the Metamodel_pk did not exhibit any variations depending on the CV utilised, and its performance was the same regardless of the CV used (0.55 approximately). Here, it is important to clarify that the low performance in terms of the specificity of the metamodel with PK information could be due to that the lower predictivity of LLM with the worst performances (based on an analysis not included). For instance, the P-gp model only achieved correct predictions for the non-cholestatic activity of compounds in 7% of the cases, and the OATP1B3 model achieved 15% of successes. These results may have an impact on the final quality of the Metamodel_pk in terms of specificity. It should be noted that this observation is consistent with the description of the LLMs, where both transporter models exhibited the poorest performances.

In the first type of cross-validation (left column of Figure 8), the Metamodel_not_pk achieved a sensitivity of 1.00. However, in the second type of cross-validation (right column of Figure 8), the sensitivity dropped to approximately 0.00. Interestingly, despite these variations in sensitivity, both types of cross-validation resulted in ROC AUC scores that were very close to

0.5. When comparing the two metamodels, it was observed that the model incorporating PK information exhibited significantly a higher ROC AUC score (between 0.15-0.19 higher for both CV approaches) compared to the model without PK information.

In the Repeated k-fold CV, RF model showed the best performance for both QSAR models. However, when using Similarity k-fold CV, the XGB model outperformed the others in terms of sensitivity, MCC, and ROC AUC score. Thus, for QSAR models, we could say that the lower the structural similarity, the lower the sensitivity. Regarding the QSAR model utilizing FP descriptors, consistently exhibited slightly higher specificity compared to the model using PC descriptors. Both models showed similar values of sensitivity, with 0.34 for QSAR_model_FP and 0.32 for QSAR_model_PC in the Repeated k-fold CV, and 0.20 and 0.24, respectively, in the Similarity-based CV. Aggregated quality indexes like the AUC or the MCC show an improvement in the overall predictive quality of the Metamodel_pk. On the contrary, the accuracy is slightly better for QSAR_model_FP (about 0.7 for both kind of CV), the most specific model, since the proportion of positive annotations is low (approximately one positive compound for every three negative compounds), and a specific model has fewer false positives. Considering the low proportion of positive compounds, more sensitive models (such as Metamodel_pk) are far more valuable and useful. Furthermore, Table S5 in the supporting information displays the evaluation of each metric for every model in each fold of the similarity 5-fold CV. In broad terms, regardless of the similarity space used, the metamodel incorporating PK information outperformed both QSAR models, showing a higher sensitivity, MCC, and AUC.

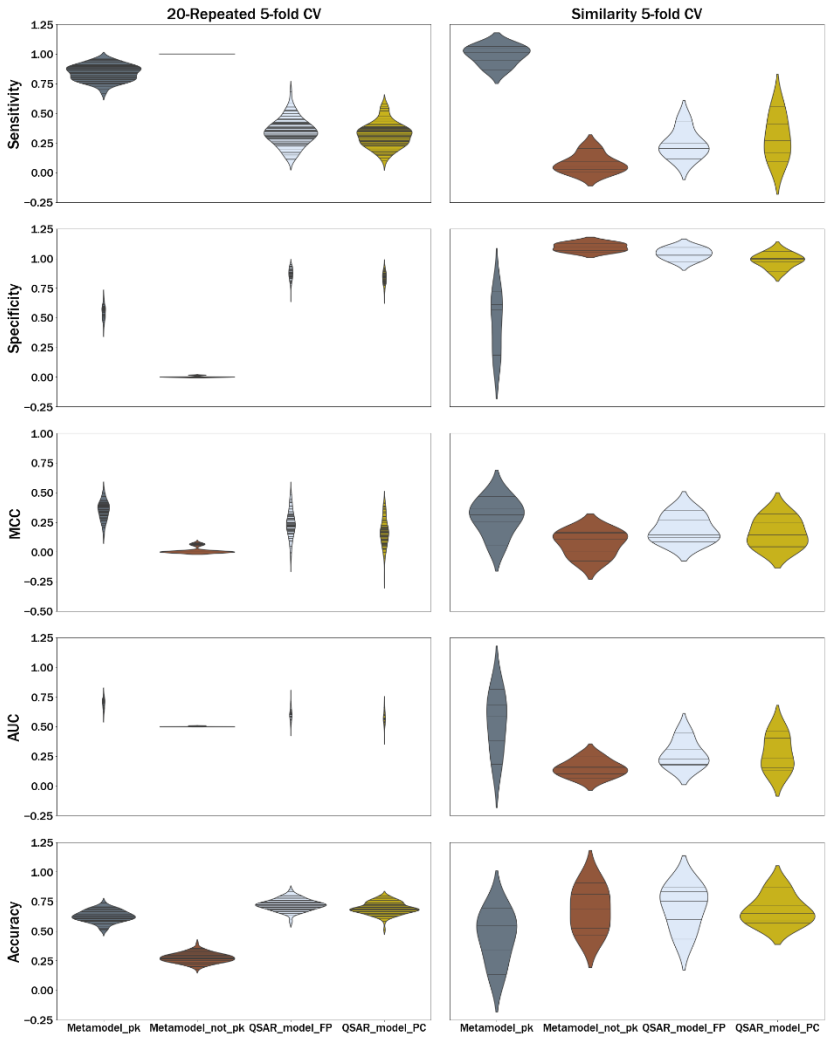


Figure 8: Sensitivity, specificity, MCC, AUC, accuracy according to the 20-Repeated 5-fold CV (left column) and the Similarity 5-fold CV (right column).

Performances according to the "ATC-based cross-validation."

The analysis by ATC codes allows the categorisation of drugs based on their therapeutic and pharmacological properties. In terms of structural independence, the ATC code can provide insight into the relationship between a drug's structure and its therapeutic properties. In this study, the predictive quality of the models was further tested by using a CV approach that closely resembles the previous point but where the folds were obtained by grouping

compounds with the same ATC code. This exercise aimed to check whether models trained with compounds from some therapeutic regions can accurately predict the toxicity of compounds belonging to different therapeutic areas, as characterised by their respective ATC codes.

Figure 9 illustrates the performance of the different models, providing further insights into their comparative robustness to predict structurally diverse compounds thanks to the use of the ATC-based k-fold approach. Comprehensive details for each model and metric can be found in Table S6. Additionally, Table S7 presents a comprehensive breakdown of the results for each fold, enabling a granular examination of the model performance across multiple metrics. The results of Figure 9 show the same variability across folds for different models and metrics that was discussed in previous plots, and the same trends with respect to the best models in terms of sensitivity and specificity. Regarding sensitivity, the best results were obtained for the similarity-based CV approach, with Metamodel_pk exhibiting much higher sensitivity (0.92) than other models. Overall, the metamodel outperformed other models in all evaluated metrics except for specificity and accuracy, as previously mentioned in the section on Similarity-based CV. Analyzing each fold based on Table S7 in the supporting information and comparing the MCC scores of different models, we observe that the Metamodel_pk achieved notably good MCC values (specifically for the ATC code of anticancer drugs (MCC_{L01}=0.47) and cardiovascular therapy (MCC_{C01}=0.56), whereas the QSAR models displayed MCC scores ranging from -0.24 to 0.15, depending on the selected QSAR model. Indeed, these two ATC groups (L01 and C01) frequently exhibit diverse molecular scaffolds, posing a greater challenge when attempting to predict them using models trained with other ATC groups. The inherent dissimilarity between these groups adds an additional layer of complexity to the prediction task. All this highlights the big drawback of QSAR

models and evidence that our methodology can address this issue by bridging the gap created by conventional models.

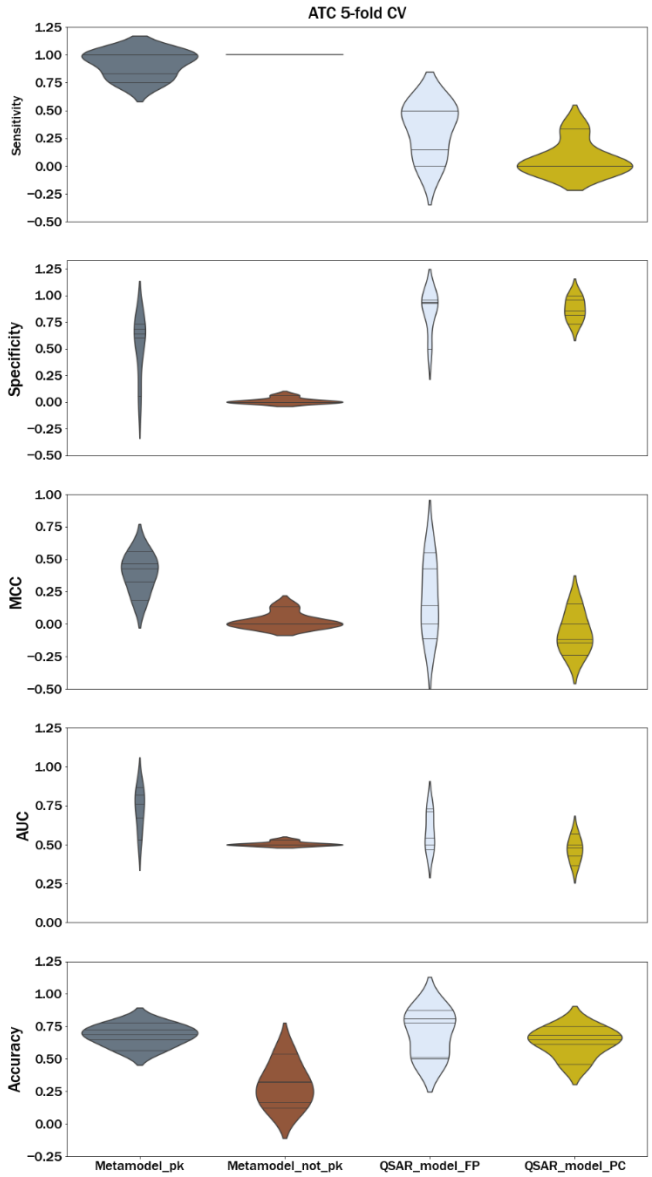


Figure 9: Sensitivity, specificity, MCC, AUC, accuracy according to the ATC 5-fold CV.

Discussion

The methodology presented here allows the prediction of cholestasis using an alternative approach to the direct QSAR models, which integrates mechanistic information and pharmacokinetic properties. To effectively execute this methodology, it is necessary to build low-level models that predict the IC_{50} of each inhibited transporter with utmost precision. These *in vitro* concentrations are subsequently extrapolated to $IEOD_{50}$ through QIVIVE models. Similarly, accurate models are essential for determining the experimental and physicochemical parameters that are used to feed the monocompartmental model underlying the calculation of C_{ss} in QIVIVE models.

Determining whether one model's predictive power surpasses another depends on the intended uses of the prediction results. Achieving the optimal balance between sensitivity and specificity is crucial in some scenarios. However, in discovery and early drug development, the main goal of *in silico* studies is the early detection of the potentially toxic compound. Therefore, a lower specificity is much preferable to a lower sensitivity. This is particularly true in the case of cholestasis since, as previously mentioned, it is a severe adverse event and a relevant mechanism of DILI, which is one of the primary causes of drug withdrawal or termination of clinical trials. Hence, in new drug development, sacrificing potential candidates may be preferable to avoid future financial losses worth millions of dollars.

It is worth noting that the IC_{50} values for many compounds were predicted as their experimental values were unknown. The LLMs exhibited similar performances, except for the P-gp and OATP1B3 transporters. In the case of the P-gp model, its significant assay variability, as described by other authors [36,59], directly impacts the measurement of compound biological activity, resulting in the lowest-performing low-level model. Similarly, the limited number of compounds available to train the OATP1B3 model contributes to

its consistently poor performance across different splits of the 20-Repeated 5-fold CV. Regarding the P-gp model, we could have applied stricter criteria to the assays to include more homogeneous data. However, to maintain methodological consistency, the same procedure would need to be applied to other transporters, potentially resulting in an extremely small number of compounds. Regarding to OATP1B3 transporter, for which it was not possible to build a good model, it could have been excluded from the metamodel. However, we preferred to keep it to maintain a more complete representation of all the targets involved in the biological mechanisms underlying cholestasis occurrence. Therefore, we opted for a controlled risk approach, monitoring subsequent evaluations where metamodel failures were observed. Therefore, the inaccuracy of the predicted values should be borne in mind as it limits the quality of the model. Even so, we consider that even a rough estimation of the pIC_{50} is likely to improve the overall predictive performance of the method, and it has value in exemplifying an approach which can be much improved when higher-quality estimations of these parameters (either experimental or predicted) can be generated.

In evaluating the predictive power of the selected strategies (Figures 8 and 9), it was found that Metamodel_pk exhibited substantially greater sensitivity than Metamodel_not_pk. The improved sensitivity is likely due to PK data providing information about a drug's behaviour in the body, including ADME processes and drug's exposure. The model that did not incorporate PK (Metamodel_not_pk) information only used *in vitro* data, which cannot accurately predict how a drug will behave *in vivo*. Similar findings were observed when comparing the Metamodel_pk to either of the two classical QSAR models, with the first one demonstrating higher sensitivity than the QSAR models.

The predictive performance of QSAR models and metamodels was further investigated by evaluating their quality in situations where the validation sets have a lower resemblance to the training sets (Figure 8). Our results confirmed the theory that QSAR models are highly dependent on the structural similarity between the test series compounds to the ones in the training series. It highlights the need for careful consideration of the selection of compounds used in the training set and the evaluation of the performance of the QSAR models for structurally diverse compounds. In contrast, our results showed that the metamodel based on PK information was not dependent on structural similarity, probably because it represents better the underlying biological mechanisms. Overall, our study provides further evidence for the differential performance of QSAR models and metamodel in predicting cholestasis and highlights the importance of considering the structural similarity of the validation compounds when evaluating the predictive quality of QSAR models. These findings suggest that the metamodel with PK information is much less dependent than the direct QSAR models of the structural similarity and, therefore can produce a better prediction for original structures, which is one of the common use scenarios in drug discovery.

This hypothesis was further confirmed using an additional validation analysis, in which we predicted compounds of a certain therapeutic area (as characterised by their ATC codes) using compounds from other therapeutic areas. In this type of analysis, we are aware that alternative strategies, such as clustering scaffolds, could be employed. However, this approach may introduce the challenge of having an abundance of scaffolds that may not be relevant to the analysis, requiring manual selection. Thus, by selecting the five most prevalent ATC groups within the dataset under investigation, we ensure a more objective analysis. The analysis revealed that the metamodel with PK information exhibited significantly higher sensitivity across all ATC groups than

the QSAR model. Concerning the groups J01, N05 or N02, the MCC was not much superior for the metamodel with respect to QSAR_model_FP (Table S7 of the supporting information). However, the metamodel with PK information demonstrated clear benefits in terms of predictive quality for the following L01 and C01 ATC subgroups. This could be due to that compounds belonging to ATC groups, such as antineoplastics or cardiac therapy, do not usually share common scaffolds with other ATC groups. Once more, this emphasises the recommended methodology for cases where the objective is to predict the cholestasis activity of a novel drug from a therapeutic category with many structural differences.

The results indicate that the metamodel incorporating PK information is better for typical applications in discovery or early development stages toxicity assessment than the direct models and Metamodel_not_pk. This could be explained by the fact that the metamodel with PK information considers both hazard and exposure, providing a more comprehensive representation of the underlying biological mechanisms of action.

Conclusion

Here, we present an innovative methodology integrating multiple biological phenomena (MIE) with pharmacokinetic properties (QIVIVE) to predict cholestasis as a more comprehensive approach to the phenomena of interest. The aim was to assess the predictive ability of the proposed methodology and direct QSAR modelling in three different ways: by evaluating the overall performance of the models (S, SP, MCC, ROC AUC, A) using classical methods as well as by using ad-hoc cross-validation approaches where the predicted compounds are selected to have low similarity (in terms of structural similarity or ATC codes) with the training series.

After comparing the predictive power of the proposed models, it was determined that, in broad terms, the metamodel with PK information outperformed both the metamodel without PK information and QSAR models in terms of sensitivity, MCC, and ROC AUC. Nevertheless, concerning accuracy, the outcomes were less favourable than those of the QSAR models. This outcome is understandable since there were significantly more negative compounds than positive ones, and ML models often predict the majority class more accurately due to higher information available, resulting in a slightly lower hit rate.

The Metamodel_pk showed structural independence as its sensitivity and specificity remained unchanged regardless of the similarity space tested (using the "similarity-based CV"). In contrast, the QSAR models showed decreasing sensitivity as similarity decreased. In the "ATC-based CV" the Metamodel_pk also showed a much higher sensitivity than the QSAR model, especially in cases where there are more diverse structures that keep a lower structural similarity, as in the case of antineoplastic agents (L01) or cardiac therapy (C01). Overall, the metamodel that included PK information demonstrated superior predictive performance for more diverse structures and cholestatic compounds.

In light of these results, we propose that this methodology can be applied to other complex toxicological endpoints, aiding experts in developing new frameworks to support NGRA as it considers both hazard and exposure for a more comprehensive toxicity assessment.

Data Availability

The code and datasets used in the study are publicly available from the GitHub repository: https://github.com/phi-grib/Cholestasis_paper.

Supporting Information

Low-level models optimization; 20 Repeated 5-fold CV, ATC 5-fold CV and Similarity 5-CV models optimization; Tanimoto similarity values intra and inter-clusters, and intra and inter-ATC; f_{ub} and CL_{int} performances; Similarity 5-fold CV and ATC 5-fold CV performances by fold; ATC 5-fold CV mean performances.

Author information

Corresponding author

Manuel Pastor - Research Programme on Biomedical Informatics (GRIB), Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Hospital del Mar Medical Research Institute, Barcelona, Spain; https://orcid.org/0000-0001-8850-1341; Email: manuel.pastor@upf.edu.

Authors

Pablo Rodríguez-Belenguer - Research Programme on Biomedical Informatics (GRIB), Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Hospital del Mar Medical Research Institute, Barcelona, Spain; Department of Pharmacy and Pharmaceutical Technology and Parasitology, Universitat de València, Valencia, Spain; https://orcid.org/0000-0001-5270-7452.

Víctor Mangas-Sanjuan - Department of Pharmacy and Pharmaceutical Technology and Parasitology, Universitat de València, Valencia, Spain; Interuniversity Research Institute for Molecular Recognition and Technological Development, Universitat Politècnica de València, Valencia, Spain; https://orcid.org/0000-0002-3388-5023.

Emilio Soria-Olivas - IDAL, Intelligent Data Analysis Laboratory, ETSE, Universitat de València, Valencia, Spain; https://orcid.org/0000-0002-9148-8405.

Author contributions

Manuel Pastor and Pablo Rodriguez-Belenguer conceived and designed the study. Material preparation, data collection, and *in silico* analysis were performed by Pablo Rodriguez-Belenguer. The results analysis was performed by Pablo Rodriguez-Belenguer and Manuel Pastor. All the authors analysed and discussed the results. The first draft of the manuscript was written by Pablo Rodriguez-Belenguer and Manuel Pastor, and all authors contributed to previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

The authors received funding from the eTRANSAFE project, Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777365, supported by European Union's Horizon 2020 and the EFPIA. Authors declare that this work reflects only the author's view and that IMI-JU is not responsible for any use that may be made of the information it contains. Also, this project received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 964537 (RISK-HUNT3R), which is part of the ASPIS cluster.

Conflicts of interest

The authors declare no conflict of interest.

Ethical standards

The manuscript does not contain clinical studies or patient data.

References

1. Chang X, Tan YM, Allen DG, Bell S, Brown PC, Browning L, Ceger P, Gearhart J, Hakkinen PJ, Kabadi SV, Kleinstreuer NC, Lumen A, Matheson J, Paini A, Pangburn HA, Petersen EJ, Reinke EN, Ribeiro AJS,

Sipes N, Sweeney LM, Wambaugh JF, Wange R, Wetmore BA, Mumtaz M. IVIVE: Facilitating the Use of In Vitro Toxicity Data in Risk Assessment and Decision Making. Toxics. 2022;10. doi:10.3390/toxics10050232

2. Rovida C, Escher SE, Herzler M, Bennekou SH, Kamp H, Kroese DE, Maslankiewicz L, Moné MJ, Patlewicz G, Sipes N, Van Aerts L, White A, Yamada T, Van de Water B. NAM-supported read-across: From case studies to regulatory guidance in safety assessment. ALTEX. 2021;38:140-150. doi:10.14573/altex.2010062

3. Belfield SJ, Firman JW, Enoch SJ, Madden JC, Erik Tollefsen K, Cronin MTD. A review of quantitative structure-activity relationship modelling approaches to predict the toxicity of mixtures. Computational Toxicology. 2023;25:100251. doi:https://doi.org/10.1016/j.comtox.2022.100251

4. Astuto MC, Di Nicola MR, Tarazona JV, Rortais A, Devos Y, Liem AKD, Kass GEN, Bastaki M, Schoonjans R, Maggiore A, Charles S, Ratier A, Lopes C, Gestin O, Robinson T, Williams A, Kramer N, Carnesecchi E, Dorne JCM. In Silico Methods for Environmental Risk Assessment: Principles, Tiered Approaches, Applications, and Future Perspectives. Methods Mol Biol. 2022;2425:589-636. doi:10.1007/978-1-0716-1960-5_23

5. Osman NA. Statistical methods for in silico tools used for risk assessment and toxicology. Physical Sciences Reviews. Published online January 7, 2022. doi:10.1515/PSR-2018-0166/MACHINEREADABLECITATION/RIS

6. Thomas PC, Bicherel P, Bauer FJ. How in silico and QSAR approaches can increase confidence in environmental hazard and risk assessment. Integr Environ Assess Manag. 2019;15:40-50. doi:10.1002/IEAM.4108

7. Myatt GJ, Bassan A, Bower D, Crofton KM, Cross KP, Graham JC, Hasselgren C, Jolly RA, Miller S, Pavan M, Tice RR. Increasing the acceptance of in silico toxicology through development of protocols and position papers. Computational Toxicology. 2022;21:100209. doi:10.1016/J.COMTOX.2021.100209

8. De P, Kar S, Ambure P, Roy K. Prediction reliability of QSAR models: an overview of various validation tools. Arch Toxicol. 2022;96:1279-1295. doi:10.1007/S00204-022-03252-Y/FIGURES/6

9. Chinen K, Malloy T. Multi-Strategy Assessment of Different Uses of QSAR under REACH Analysis of Alternatives to Advance Information

Transparency. International Journal of Environmental Research and Public Health 2022, Vol 19, Page 4338. 2022;19:4338. doi:10.3390/IJERPH19074338

10. Kolmar SS, Grulke CM. The effect of noise on the predictive limit of QSAR models. J Cheminform. 2021;13:92. doi:10.1186/s13321-021-00571-7

11. Madden JC, Thompson C V. Pharmacokinetic Tools and Applications. In: Benfenati E, ed. In Silico Methods for Predicting Drug Toxicity. Springer US; 2022:57-83. doi:10.1007/978-1-0716-1960-5_3

12. Calabrese EJ. Dose–Response Relationship. In: Wexler PBTE of T (Third E, ed. Academic Press; 2014:224-226. doi:https://doi.org/10.1016/B978-0-12-386454-3.00991-X

13. Krewski D, Acosta Jr D, Andersen M, Anderson H, Bailar III JC, Boekelheide K, Brent R, Charnley G, Cheung VG, Green Jr S, Kelsey KT. Toxicity testing in the 21st century: a vision and a strategy. J Toxicol Environ Health B Crit Rev. 2010;13:51-138. doi:10.1080/10937404.2010.483176

14. Vinken M, Knapen D, Vergauwen L, Hengstler JG, Angrish M, Whelan M. Adverse outcome pathways: a concise introduction for toxicologists. Arch Toxicol. 2017;91:3697-3707. doi:10.1007/s00204-017-2020-z

15. OECD. OECD Series on Testing and Assessment No. 184: Guideance Document on Developing and Assessing Adverse OutcomePathways. OECD, Paris, 2013. Published online 2013.

16. OECD. OECD Series on Testing and Assessment No. 233: Users' Handbook Supplement to the Guidance Document for Developing and Assessing AOPs. OECD, Paris, 2018. Published online 2018:1-60.

17. OECD. AOP knowledge base. Published 2014. Accessed September 15, 2022. https://aopkb.oecd.org/

18. Society for Advancement of AOPs. AOP-Wiki. Published 2014. Accessed September 15, 2022. https://aopwiki.org/aops

19. Matsuzaka Y, Totoki S, Handa K, Shiota T, Kurosaki K, Uesawa Y. Prediction Models for Agonists and Antagonists of Molecular Initiation Events for Toxicity Pathways Using an Improved Deep-Learning-Based Quantitative Structure–Activity Relationship System. International Journal of Molecular Sciences 2021, Vol 22, Page 10821. 2021;22:10821. doi:10.3390/IJMS221910821

20. Allen TEH, Goodman JM, Gutsell S, Russell PJ. Quantitative Predictions for Molecular Initiating Events Using Three-Dimensional Quantitative Structure-Activity Relationships. Chem Res Toxicol. 2020;33:324-332. doi:10.1021/ACS.CHEMRESTOX.9B00136/SUPPL_FILE/TX9B00136_SI_002.PDF

21. Testai E, Bechaux C, Buratti FM, Darney K, Di Consiglio E, Kasteel EE, Kramer NI, Lautz LS, Santori N, Skaperda ZV, Kouretas D. Modelling human variability in toxicokinetic and toxicodynamic processes using Bayesian meta-analysis, physiologically-based modelling and in vitro systems. EFSA Supporting Publications. 2021;18:6504E. doi:10.2903/SP.EFSA.2021.EN-6504

22. Gao Y, Kang L, Zhang Y, Feng J, Zhu L. Toxicokinetic and toxicodynamic (TK-TD) modeling to study oxidative stress-dependent toxicity of heavy metals in zebrafish. Chemosphere. 2019;220:774-782. doi:10.1016/J.CHEMOSPHERE.2018.12.197

23. Warner RM, Sweeney LM, Hayhurst BA, Mayo ML. Toxicokinetic Modeling of Per- and Polyfluoroalkyl Substance Concentrations within Developing Zebrafish (Danio rerio) Populations. Environ Sci Technol. Published online September 2, 2022. doi:10.1021/ACS.EST.2C02942

24. Lehman-McKeeman LD. Mechanisms of Toxicity. In: Klaassen CD, ed. Casarett & Doull's Toxicology: The Basic Science of Poisons, 9th Edition. McGraw-Hill Education; 2019.

25. Punt A, Pinckaers N, Peijnenburg A, Louisse J. Development of a Web-Based Toolbox to Support Quantitative In-Vitro-to-In-Vivo Extrapolations (QIVIVE) within Nonanimal Testing Strategies. Chem Res Toxicol. 2021;34:460-472. doi:10.1021/ACS.CHEMRESTOX.0C00307/ASSET/IMAGES/LARGE/TX0C00307_0008.JPEG

26. Kotsampasakou E, Ecker GF. Predicting Drug-Induced Cholestasis with the Help of Hepatic Transporters—An in Silico Modeling Approach. J Chem Inf Model. 2017;57:608-615. doi:10.1021/acs.jcim.6b00518

27. Norman BH. Drug Induced Liver Injury (DILI). Mechanisms and Medicinal Chemistry Avoidance/Mitigation Strategies. J Med Chem. 2020;63:11397-11419. doi:10.1021/acs.jmedchem.0c00524

28. Padda MS, Sanchez M, Akhtar AJ, Boyer JL. Drug-induced cholestasis. Hepatology. 2011;53:1377-1387. doi:10.1002/HEP.24229

29. Pauli-Magnus C, Meier PJ. Hepatobiliary transporters and drug-induced cholestasis. Hepatology. 2006;44:778-787. doi:10.1002/HEP.21359

30. Jazaeri F, Sheibani M, Nezamoleslami S, Moezi L, Dehpour AR. Current Models for Predicting Drug-induced Cholestasis: The Role of Hepatobiliary Transport System. Iran J Pharm Res. 2021;20:1. doi:10.22037/IJPR.2020.113362.14254

31. Vinken M, Landesmann B, Goumenou M, Vinken S, Shah I, Jaeschke H, Willett C, Whelan M, Rogiers V. Development of an Adverse Outcome Pathway From Drug-Mediated Bile Salt Export Pump Inhibition to Cholestatic Liver Injury. Toxicological Sciences. 2013;136:97-106. doi:10.1093/TOXSCI/KFT177

32. Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, Bellis L, Overington JP. ChEMBL web services: streamlining access to drug discovery data and utilities. Nucleic Acids Res. 2015;43:W612-20. doi:10.1093/nar/gkv352

33. Wang Y, Suzek T, Zhang J, Wang J, He S, Cheng T, Shoemaker BA, Gindulyte A, Bryant SH. PubChem BioAssay: 2014 update. Nucleic Acids Res. 2014;42:D1075-82. doi:10.1093/nar/gkt978

34. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34:D668-72. doi:10.1093/nar/gkj067

35. Cates Jill. A Python package for standardizing medical data. Published 2018. Accessed February 1, 2023. https://github.com/topspinj/medcodes

36. Elkins RC, Davies MR, Brough SJ, Gavaghan DJ, Cui Y, Abi-Gerges N, Mirams GR. Variability in high-throughput ion-channel screening data and consequences for cardiac safety assessment. J Pharmacol Toxicol Methods. 2013;68:112-122. Doi:10.1016/j.vascn.2013.04.007

37. March-Vila, Eric and Pastor M. Data curation. Published online 2020. Accessed February 1, 2023. https://github.com/phigrib/Data_curation

38. Landrum GA. RDKit: Cheminformatics and Machine Learning Software. Open-Source Cheminformatics Software, RDKit (2019). Published online 2016.

39. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-

learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-2830.

40. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016;13-17-August-2016:785-794. Doi:10.1145/2939672.2939785

41. Ho TK. Random decision forests. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. 1995;1:278-282. Doi:10.1109/ICDAR.1995.598994

42. Fix E, Hodges JL. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. Int Stat Rev. 1989;57:238-247. Doi:10.2307/1403797

43. Cortes C, Vapnik V, Saitta L. Support-vector networks. Machine Learning 1995 20:3. 1995;20:273-297. Doi:10.1007/BF00994018

44. Alpaydin E. Introduction to Machine Learning. 2nd ed. The MIT Press; 2010.

45. Pearce RG, Setzer RW, Strope CL, Sipes NS, Wambaugh JF. Httk: R Package for High-Throughput Toxicokinetics. J Stat Softw. 2017;79:1-26. Doi:10.18637/JSS.V079.I04

46. Wetmore BA. Quantitative in vitro-to-in vivo extrapolation in a high-throughput environment. Toxicology. 2015;332:94-101. Doi:10.1016/j.tox.2014.05.012

47. Mansouri K, Grulke CM, Judson RS, Williams AJ. OPERA models for predicting physicochemical properties and environmental fate endpoints. J Cheminform. 2018;10:10. Doi:10.1186/s13321-018-0263-1

48. Rodríguez-Pérez R, Gerebtzoff G. Identification of bile salt export pump inhibitors using machine learning: Predictive safety from an industry perspective. Artificial Intelligence in the Life Sciences. 2021;1:100027. Doi:https://doi.org/10.1016/j.ailsci.2021.100027

49. Bosc N, Atkinson F, Felix E, Gaulton A, Hersey A, Leach AR. Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. J Cheminform. 2019;11:4. Doi:10.1186/s13321-018-0325-4

50. Jain S, Grandits M, Richter L, Ecker GF. Structure based classification for bile salt export pump (BSEP) inhibitors using comparative structural

modeling of human BSEP. J Comput Aided Mol Des. 2017;31:507-521. Doi:10.1007/s10822-017-0021-x

51. Warner DJ, Chen H, Cantin LD, Kenna JG, Stahl S, Walker CL, Noeske T. Mitigating the Inhibition of Human Bile Salt Export Pump by Drugs: Opportunities Provided by Physicochemical Property Modulation, In Silico Modeling, and Structural Modification. Drug Metabolism and Disposition. 2012;40:2332-2341. Doi:10.1124/dmd.112.047068

52. Vikramkumar, B V, Trilochan. Bayes and naïve Bayes Classifier. Published online April 3, 2014. doi:10.48550/arxiv.1404.0933

53. Harris CR, Millman KJ, Van Der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R. Array programming with NumPy. Nature 2020 585:7825. 2020;585:357-362. doi:10.1038/s41586-020-2649-2

54. Seabold S, Perktold J. statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference. ; 2010.

55. Waskom ML. seaborn: statistical data visualization. J Open Source Softw. 2021;6:3021. doi:10.21105/joss.03021

56. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007;9:90-95. doi:10.1109/MCSE.2007.55

57. McKinney W. Data Structures for Statistical Computing in Python. In: van der Walt S, Millman J, eds. Proceedings of the 9th Python in Science Conference. ; 2010:51-56.

58. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 1997;23:3-25. doi:https://doi.org/10.1016/S0169-409X(96)00423-1

59. Landrum GA. The hazards of combining data from IC50 assays. Published 2023. Accessed February 1, 2023. Accessed June 17, 2023. https://greglandrum.github.io/rdkit-blog/posts/2023-06-12-overlapping-ic50-assays1.html

Capítulo 3

Application of machine learning to improve the efficiency of electrophysiological simulations used for the prediction of drug-induced ventricular arrhythmia

Pablo Rodríguez-Belenguer[1*], Karolina Kopańska[1*], Jordi Llopis-Lorente[2], Beatriz Trenor[2], Javier Saiz[2], Manuel Pastor[1]

[1]Research Programme on Biomedical Informatics (GRIB), Universitat Pompeu Fabra, Barcelona, Spain

[*]Both authors have contributed equally

[2]Centro de Investigación e Innovación en Bioingeniería (Ci2B), Universitat Politècnica de València, Valencia, Spain

Revista: Computer Methods and Programs in Biomedicine

Editorial: Elsevier

Año: 2023

Cuartil: Q1

IF: 6.1

Introduction

Assessing the arrhythmogenic risk of new drug candidates is an important step in safety studies. The mechanism by which drugs induce ventricular arrhythmias involves their binding to one or multiple ion channels, thereby altering the ionic conductance that controls cardiomyocyte membrane potential.[1] As a result, the form and duration of ventricular action potentials (APs) change, and the net effects can be observed at tissue and organ levels, such as the prolongation of the QT-interval on the surface ECG.[2] A significant prolongation of the QT-interval, is often linked to severe adversities such as early afterdepolarisations (EADs), which can quickly progress to one of the most severe effects of proarrhythmic drugs: the polymorphic ventricular tachycardia known as Torsade de Pointes (TdP).[3]

As the occurrence of TdP historically led to the withdrawal of several marketed drugs, the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) developed standardised guidelines for safety testing of novel medicines.[4] Resting upon the preclinical ICH S7b guideline,[5] the estimation of proarrhythmic risk is done through the integration of results from *in vitro* inhibition assay of the Rapid Delayed Rectifier Potassium Current (I_{Kr}) encoded by the Human Ether-a-go-go-related Gene (hERG) and an *in vivo* animal QT-prolongation study. Following the clinical guideline ICH E14,[6] the potential of a drug to delay ventricular repolarisation is assessed by measuring *in vivo* human QT/QTc interval prolongation.

Indeed, testing drugs in compliance with these regulatory requirements over the last two decades resulted in no further removal of marketed drugs due to ventricular arrhythmia. However, the consideration of *in vitro* effects of drugs on a single ion channel and the application of a conservative cut-off for QT-prolongation is the reason why several potentially useful drug candidates with

low toxicity risk are also discarded during the development stages. To provide a more complete description of the cellular mechanisms of drug proarrhythmia, a novel testing paradigm was proposed by the Comprehensive In Vitro Proarrhythmia Assay (CiPA) initiative.[7,8] The CiPA points out that the consideration of drug interactions with other currents along with the hERG is also important for the analysis of ventricular arrhythmia. The main aim behind the CiPA project is to combine *in vitro* measured drug effects on multiples ion channels (I_{Na}, I_{NaL}, I_{Kr}, I_{to}, I_{CaL}, I_{K1}, and I_{Ks}) with computational simulations, such as *in silico* reconstructions of cardiac myocyte electrophysiology, and to compare these results with *in vitro* human stem cell results and human ECG phase 1 clinical trials.[9]

Adding *in silico* elements to the cardiac safety testing pipeline has two main advantages, the first being the ability to fill data gaps when experimental results are not yet available at early stages of drug development and the second being an increased analytical accuracy due to the solid mechanistic foundation of the CiPA paradigm.[10]

Several works have been published on the implementation of the CiPA based *in silico* simulations for the prediction of ventricular arrhythmia and TdP biomarkers using predicted or experimentally determined drug-induced ion channel inhibition data.[11–19] Computational models of human and animal electrophysiology operate at different biological levels, ranging from a single channel to whole tissue simulations and vary in terms of the degree of complexity and abstraction, the underlying mathematical approaches, and physiological parameters.[11] Although the predictions generated by such models are considered valuable and relevant, they also have limitations related to their usability. Usually, computational safety models are designed based on the subjective scientific interests of the developers and the required efficiency to run on high-performance-computing platforms is seldom

reached. But most importantly, the simulation consists of multiple steps, making the prediction process rather tedious.[20] For example, Beattie *et al. (2013)*[21] presented a safety tool based on concentration-effect data for four cardiac ion channels (hERG, NaV1.5, CaV1.2, KCNQ1), in which drug-induced channel inhibition of selected compounds was predicted and used for the computation of QT interval changes in rabbit ventricles using computationally demanding one-dimensional tissue simulation.

To speed up the process, our group developed an *in silico* system that transforms multi-channel blockage into proarrhythmia biomarkers, such as action potential duration at 90% of the repolarisation (APD_{90}), in which the most computationally intensive steps are precomputed, allowing to produce results instantaneously.[22,23] In our system, input values are pre-processed by combining channel-specific half-maximal inhibitory concentration (IC_{50}) and the Hill coefficient for the currents I_{Kr}, I_{Ks}, and I_{CaL} with the concentration of the drug. The APD_{90} prolongation values are then predicted using isolated human ventricular myocyte models as a function of these three input values. Since the calculation can take a considerable time, the predictions are generated by making use of precomputed matrices comprising large sets of possible combinations of input values, each of which is associated with a particular value of the output biomarker. These technical features make the prediction system simple, practical, and rapid.[22,23]

Even if storing precomputed data matrices is a very convenient way to obtain predictions interactively, with minimal computational requirements for the end-user, the procedure has the drawback that the preparatory simulations that the method developer needs to run are extremely expensive in terms of computational power and time. This is because accurate predictions can only be produced when the input values cover a wide range of possibilities starting with safe and ending with very toxic representations of drug effects on each

considered ion channel. The number of combinations is calculated as X^n, being X the number of possible values considered for each input value and n the count of the input values considered (number of ion channels). This fact imposes a practical upper limit to the number of currents that can be considered since incorporating one more channel multiplies by X the number of simulations to run. Since incorporating additional currents could have substantial benefits, we studied how to overcome these limitations. A potential solution would be to train a machine learning (ML) model with part of the data array and use it to predict the rest of the data array, thereby reducing the number of required simulations. The use of ML in the field of arrhythmia and electrophysiology-oriented research is not new, and the spectrum of published ML applications in this area is very broad.[20,21] For example, classification and regression algorithms can be applied to build models describing the association between the molecular structure and the inhibitory potential of drugs on ion channels[24] or to produce high-level arrhythmogenic risk indicators.[25–27] Another example of the application of ML in combination with *in silico* simulations to improve the predictive results of the arrhythmogenic risk in post-infarction patients was described by Maleckar and colleagues (2020), who simulated the data for the analysis only partially and predicted the rest using ML methods.[28]

In this work, we describe an application of ML which aimed only to optimise the generation of precomputed matrices that link input ionic currents with output APD_{90} values. The basic idea was to train a model with a few of the array nodes and to use it to reconstruct the whole array. We show that even a tiny fraction of nodes (5% or less) can produce a very accurate estimation of the values obtained using simulations for the remaining part of the array (95% or more). Therefore, using an ML model can save up to 95% of the computation time and, more importantly, opens the possibility to precompute

matrices with more currents that can provide better, more useful predictions. In this work, we compare different machine learning approaches, optimise their parameters, and evaluate the quality of the predictions obtained using different sample sizes to make the most optimal choices for future simulations. Then, we present the best methodological settings and validate our selected model by predicting the APD_{90} for a series of compounds from the CiPA dataset. Lastly, we evaluate the value of our method by simulating a real production scenario where it was applied to a new electrophysiological simulation.

Methods

Data collection for model building

In silico action potential (AP) modelling of the healthy human endocardial cardiomyocyte and APD_{90} measurements were done using a modified version of the widely known model published by O'Hara and colleagues.[29] The modifications were designed to better reproduce the experimental data of drug effects. Briefly, the AP model modifications included: i) the scaling of the following conductances: I_{Kr} by 1.119, I_{NaL} by 2.274, I_{K1} by 1.414, I_{Ks} by 1.648, I_{CaL} by 1.018, and I_{Na} by 0.4; and ii) a reformulation of the activation and inactivation gates of I_{Na}. For further details about the electrophysiological model, see Llopis-Lorente et al. (2020).[16] Simulations were run with a basic cycle length of 1,000 ms, a stimulus of 1.5-fold the diastolic threshold of amplitude and a duration of 0.5 ms, at physiological temperature (37ºC) and the following extracellular concentrations: $[Na^+]$=140 nM, $[Ca^{2+}]$=1.8 nM and $[K^+]$=5.4 nM. Measurements of APD_{90} under drug effects were done after 500 beats starting from control -no drug- initial conditions.

In this work, we considered the effects of drug action on two combinations of cardiac ion channels. Primarily, aiming to improve the *in silico* modelling tool

described by Obiol-Pardo, we considered drug effects on I_{Kr}, I_{Ks} and I_{CaL} currents.[22,23] To evaluate the applicability of our new methodology to other combinations of ionic channels and validate the proposed machine learning methods, we selected the currents I_{Kr}, I_{CaL}, I_{NaL} that were recently described by Llopis-Lorente.[16] Drug effects on the AP were simulated using the simple pore block model.[30] Drug inhibition produced on each channel was simulated by scaling the channel's maximal conductance (g_i) using the standard Hill equation (**Equation 1**).

$$g_{i,drug} = g_i \left[1 + \left(\frac{D}{IC_{50,i}}\right)^h\right]^{-1} \quad \textbf{Equation 1}$$

where $g_{i,drug}$ is channel i's maximal conductance in the presence of the drug, D is the drug concentration, $IC_{50,i}$ is the half-maximal inhibitory concentration for that drug, and channel i and h is the Hill coefficient, which represents the number of molecules that are sufficient to block an ion channel.

A wide combination of input values representing the ratio $\left(\frac{\mathrm{D}}{\mathrm{IC_{50}}}\right)^{\mathrm{h}}$ for I_{Kr}, I_{Ks}, I_{CaL} was simulated and stored in an array. The array consisted of 3 channel input values: I_{Kr}, I_{Ks}, and I_{CaL}. Each of them represented the logarithm of the ratio $\left(\frac{\mathrm{D}}{\mathrm{IC_{50}}}\right)^{\mathrm{h}}$, as described in **Equation 2**. For each channel (I_{Kr}, I_{Ks}, I_{CaL}), the input value ranged from -3 to 2.5, with a step increment of 0.1. These values were chosen to cover the properties found in real molecules, avoiding the need to extrapolate the models. Therefore, the simulated array comprised 175,616 instances (56 data points for each current).

$$\text{Input value} = \log_{10}\left(\left[\frac{D}{IC_{50}}\right]^h\right) \quad \textbf{Equation 2}$$

The output value of the array was the APD_{90}, simulated as described above for each of the input values combinations. For each set of input values, an additional binary variable was included to indicate whether early

afterdepolarisations (EAD) occurred during the simulation of that drug (EAD=1) or not (EAD=0). An EAD was defined as any event with a positive voltage gradient (dV/dt > 0 mV/ms) after 100 ms from the beginning of the action potential or with a value of membrane voltage at the end of the beat being higher than resting membrane voltage (Vm > −40 mV).

The standard use of such array was as follows: for a given compound at a concentration D, **Equation 2** was applied for the three ionic channels (I_{Kr}, I_{Ks}, I_{CaL}). The results of Equation 2 were rounded to the first decimal and bounded between -3 and 2.5, i.e., if an input value was lower than -3 or higher than 2.5, the value was then transformed to -3 or 2.5. For each combination of the three calculated input values, the corresponding output (APD_{90}) was stored in a three-dimensional result array. For example, a drug with the following IC_{50}s: 1 nM for I_{Kr}, 1000 nM for I_{Ks} and 10 nM for I_{CaL} at a concentration of 1 nM yielded the data point [0, -3, -1], which led to an APD_{90} of 369.16 ms.

Electrophysiological simulations and generation of the APD_{90} array were carried out using MATLAB version R2021b. The table with the APD_{90} values for a wide combination of input values is available online, named “KrKsCaL.xlsx”, on the public repository of the Polytechnic University of Valencia (RIUNET, https://riunet.upv.es/handle/10251/183067).

Data pre-processing

We removed from the analysis all data points for which EADs were detected. Also, we applied filters to remove simulation results yielding APDs greater than 1000 ms. These conditions represent repolarisation abnormalities, and the numerical result is considered unreliable. Additionally, data points with an APD_{90} larger than the 3rd quartile plus 1.5 times the interquartile range were considered outliers and removed. This filter removed 1.4% of the data points, with values ranging between 777.59 and 865.47 ms. After the pre-processing, the number of simulation results was reduced to 140,269.

The data array was divided into training, validation, and test sets using four different sampling rates that were used in the models (**Table 1**). In each case, the training and validation series were extracted by picking the results at regular and pre-defined intervals to guarantee an even distribution of values for fitting and validation along with the explored range of input values. All remaining data were used as test series to evaluate the predictive performance of the models.

Table 1: Percentages of data from the original array sampled using four different rates to generate the training, validation, and test series for model building

Sampling rate	Training series	Validation series	Test series
1/20	5%	5%	90%
1/50	2%	2%	96%
1/100	1%	1%	98%
1/200	0.5%	0.5%	99%

Machine learning algorithms

Figure 1 shows a 3D representation of the APD_{90} values obtained for different combinations of two current pairs (I_{Kr} and I_{Ks}). The APD_{90} values are distributed on a non-linear 2D surface smoothly distributed. This observation suggests that by the application of ML algorithms suitable for processing non-linear data, we could obtain a good model fitting. In this work, we selected three different ML methods: Polynomial Transformation with Ridge regression (PR), Support Vector Machine (SVM), and Multilayer Perceptron (MLP). For each one, we optimised their hyperparameters and validated the models using three partitions and an external test set with selected CiPA compounds.

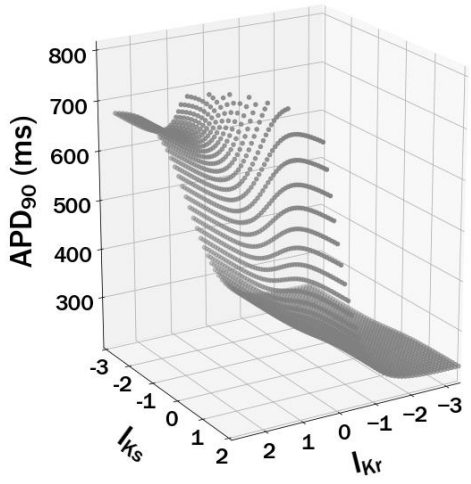


Figure 1: 3D plot showing the non-linear relationship between the APD_{90} and the input values (I_{Kr} and I_{Ks}) for the simulated data. In this plot, a fixed value of 0.3 was used for I_{CaL}.

Polynomial Regression

The PR model was built using polynomial regression (**Equation 3**), a form of linear regression in which the relationship between the independent and dependent variables is modelled as a polynomial of the n^{th} degree. In this algorithm,[31] the polynomial degree increases proportionally to the complexity of the data structure:

$$\hat{y} = b + w_1.x + w_2.x^2 \ldots + w_n.x^n \qquad \textbf{Equation 3}$$

Where ŷ is the target variable, n is the degree of the polynomial, x is the independent variable, w represents the model coefficients, and b is the offset.

To reduce the chance of overfitting the model by selecting a too high polynomial degree, Ridge regression[32] (**Equation 4**) was applied:

$$J(w,b) = \sum_{i=1}^{M} \left(y_i - b - \sum_{j=1}^{p} w_j.x_{ij}\right)^2 + \alpha \sum_{(j=1)}^{p} w_j^2 \qquad \textbf{Equation 4}$$

Ridge regression, which operates by performing L2 regularisation, penalises the model coefficients by adding the factor (α). The greater the factor α, the

greater the impact of the shrinkage penalty, resulting in a larger reduction of the magnitude of model coefficients. Therefore, finding an optimal value for α is particularly important to control model overfit.

Support vector machine

To build the SVM model, we used a non-linear support vector machine for regression (SVR) which can be explained by a line enclosed between two decision boundaries, where the width between is controlled by the parameter ε. [33] As the data points that lie within the boundaries get assigned a loss of 0, the best value of ε is the one that maximally increases the number of the data points included within. On the other hand, the error is computed using slack variables that quantify the distance from the decision boundaries Ɛ's to the points outside the margin. Support vector machine models strive towards a maximal error reduction as defined in **Equation 5**.

$$\text{Minimise } \frac{w^T w}{2} + C\sum_{i=1}^{N}(\xi_i + \xi_i{}^*) \text{ subject to } \begin{cases} y_i - w^T\phi(x_i) - b \le \varepsilon + \xi_i, \\ w^T\phi(x_i) + b - y_i \le \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \ge 0, i = 1, \dots, n \end{cases}$$

Equation 5

ξ_i and $\xi_i{}^*$ are the slack variables, $\| w \|$ represents the Euclidian normalisation of the weight (w) vector. C is a regularisation parameter where the strength of the regularisation is inversely proportional to this parameter. ϕ(x) is the transformation from input space into feature space, and b is the bias term.

To process non-linear data, support vector regressors perform the kernel trick,[34] a method that allows for a representation of the data only through a set of pairwise similarity comparisons between two instances in the input space. More precisely, a kernel function $K(x_i, x_j)$ takes as input the original low dimensional data points $(x_i,\ x_j)$ and computes a dot product of these data in the transformed high dimensional space, without explicitly determining their

coordinates in this feature space. In this work Radial Basis Function (RBF)[35] kernel (**Equation 6**) was used.

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2} \quad \textbf{Equation 6}$$

γ is the parameter of the gaussian kernel and $(x_i,\ x_j)$ are two selected input instances. In this work, scale mode γ (**Equation 7**) was selected because it is invariant against the scale of the inputs.

$$Y_{scale\ mode} = \frac{1}{n\ .\ x_{variance}} \quad \textbf{Equation 7}$$

Where n is the number of features and $x_{variance}$ corresponds to the variance in the input data.

Multilayer perceptron

Multilayer perceptron[36] is a feedforward artificial neural network class belonging to the family of supervised machine learning algorithms. The basic structure of an MLP consists of a dot product of the input data (x) with their weights (w) + the bias (b) and of an activation function which in most cases is non-linear (**Equation 8**). These inputs yield an output of a single neuron.

$$output = f(y) = f(\sum_{k=1}^{n} w_k\ .\ x_k + b) \quad \textbf{Equation 8}$$

The output obtained from the first neuron is transmitted to the next one through feedforward propagation. In order to reduce the error between the desired output and the predicted output, the weights are updated in a process of backpropagation.[37] The most important hyperparameters that impact the predictive performance of the neural network are hidden layers, activation function,[38] learning rate (lr), which controls the step-size in updating the weights, the L2 regularisation parameter penalty alpha (α), and the solver for weight optimisation.

Evaluation metrics

The three machine learning algorithms were applied to four training series generated with different sampling rates (as shown in **Table 1**) to build 12 models. The predictive performances of the models were compared using three evaluation metrics: Mean Absolute Error (MAE) (**Equation 9**), the Mean Relative Error in % (MRE) computed from Relative Error (**Equation 10**), and the percentage of data with Relative Error (RE) below 5% (non-large data-points error, NLDE). These metrics were used to quantify the differences between predicted and simulated APD_{90} values and to guarantee that the quantity of the sampled data from the original simulated data array is enough to build a robust ML model. We only consider acceptable the simulations with an RE below 5%.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \left|Y_i - \hat{Y}_i\right| \qquad \textbf{Equation 9}$$

$\hat{Y}_i$ corresponds to the predicted value, Y_i is the real value, and n is the number of data points.

$$RE(\%) = \frac{|Y_i - \hat{Y}_i|}{Y_i}.100 \qquad \textbf{Equation 10}$$

RE (%) values, computed as a function of $APD_{90,}$ were plotted for a visual evaluation

Hyperparameters of all described ML algorithms

Algorithm-specific hyperparameters selected for the optimisation of the ML models are listed in **Table 2**.

Table 2: Selected hyperparameters for the optimisation of selected ML models

Internal name	Algorithm	Hyperparameters
PR	Ridge regression with a polynomial transformation	Polynomial degree=[2-15], α=[1.10^{-6} – 10]
SVM	Support Vector Machine Regression	C=[0.1 – 30.10^{5}], kernel=RBF, Y=scale, Ɛ=0.1

MLP	Multilayer Perceptron	Hidden layers=[(50,50,50), (50,100,50), (100,)], learning rate=[constant, adaptative, sgd], solver= Adam, α=[0.05, 0.1,0.5], activation=ReLU

The hyperparameter tuning for the different models aimed to minimise the validation set MAE. We also tested whether the hyperparameters of the three selected algorithms can be optimised using only the training set or if it requires an additional validation set.

The scripts were developed using Python 3.8. Machine learning models were built and evaluated using standard libraries Scikit-learn,[39] NumPy,[40] Pandas[41] and Matplotlib.[42] The source code of the scripts used for building and validating the models, together with the datasets described and analysed in this manuscript, are available at GitHub (https://github.com/phi-grib/cardioML) and distributed as open source under GNU GLP-3.0 license.

Example case study using CiPA compounds

To obtain a more realistic evaluation, focused on the range of IC_{50} observed in commonly used drugs for the I_{Kr}, I_{Ks}, and I_{CaL} channels, as well as drug concentrations reached in their clinical use, we computed the input values as described in **Equation 2** for 12 CiPA drugs belonging to three different TdP risk classes (low, intermediate, high). For these compounds, we used the concentration corresponding to their Effective Free Therapeutic Plasma Concentration (EFTPC) values, the channel-specific half-maximal inhibitory concentrations (IC_{50}) and Hill coefficients (h) extracted from Llopis-Lorente et al. (2020).[16] D, IC_{50}s, h values and the corresponding input values used for the simulation of the 12 CiPA drugs are available at the file "12_CiPA_Drugs.D-IC50-h.xlsx" at GitHub (https://github.com/phi-grib/cardioML/blob/main/12_CiPA_Drugs.D-IC50-h.xlsx). The PR, SVM, and MLP models, trained with data sampled 1/100, were applied to these compounds to predict their APD_{90}.

The predicted results were eventually compared with the simulated APD_{90} read-out from the data array, and the differences were expressed as Relative Error (%).

Results

Overview

The starting point for this work was to generate a large number of APD_{90} values using electrophysiological simulations, as described in the Methods section. For these simulations, the input values represent the relation between the IC_{50}, Hill coefficient and the drug concentration for three ion channels I_{Kr}, I_{Ks}, and I_{CaL}. The output values are the APD_{90} we expect to obtain for cardiomyocytes exposed to a drug with the given I_{Kr}, I_{Ks}, and I_{CaL} input values. These values were collected in an array containing the APD_{90} values produced by the simulations for a wide combination of input values.

The next step was generating small samples of the original data, which were used to train ML models that were used to predict the remaining data as accurately as possible. The results were compared to identify the best ML methods and the lowest training series size producing acceptable results. Finally, the quality of the models was further compared, and the method was validated using 12 CiPA compounds.

Our study showed that a simulation of only 1-5% of data is sufficient to build an ML model able to produce accurate estimations of the remaining 99-95% of the APD_{90} values. Such a large reduction in the computation automatically translates into a substantial improvement of both the time and computing power required for the preceding data collection step. Consequently, this reduction opens the possibility of considering drug effects on more than three channels, thereby improving the mechanistic description of the *in silico* tool. From the model settings evaluated, the best results were obtained using SVM.

A sampling ratio 1/100 was considered a good trade-off between estimation quality and computation reduction, according to three quality evaluation metrics considered: MRE (%), MAE (%) and percentage of data points with RE below 5% computed for the training, validation, and test set. In the external validation using CiPA compounds, we showed that the maximum error obtained by the SVM model for the sampling ratio 1/100 barely exceeds 1.5% of RE, representing approximately 4 ms of deviation.

Compilation of the data array

As described above, a data array of APD_{90} obtained for different simulation input values (ratio of drug concentration over I_{Kr}, I_{Ks}, and I_{CaL} IC_{50}) was generated. This dataset consisted of simulated APD_{90} for 175,616 possible combinations of drug effects on channels I_{Kr}, I_{Ks}, and I_{CaL}. It covers a range of blockades from 0.1% to 99.7% for each channel. The pre-treatment applied removed values assigned the top cut-off value (1000 ms) and higher (see Methods section for details). **Figure 2** shows the final distributions of the APD_{90} values, where most of the values are concentrated around the physiological biomarker values (264 ms).

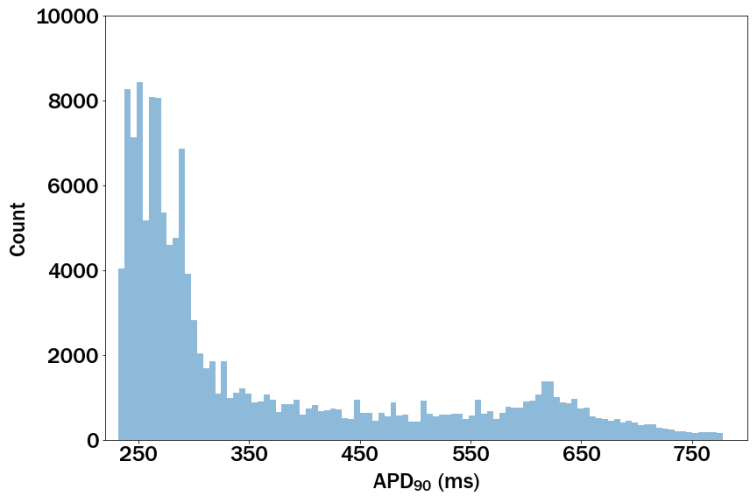


Figure 2: Distributions of APD_{90} values after data pre-processing

This data array resulted from the systematic application of electrophysiological simulations using a range of input values that start from practically safe scenarios (-3 indicates the ratio of 1:1000 between the effective free therapeutic plasma concentration and the IC_{50}). Larger APD_{90} values can only be observed for a few combinations of input values, with a slight concentration of around 610 ms.

Machine learning: fitting and quality

Before model building, the original data array was split into training, validation and test sets using regular and equal-sized patterns with four different sampling rates: 1/20, 1/50, 1/100, and 1/200. A first sample of data points was assigned as a training set and a second one as a validation set, whereby the rest of the data was devoted to the test set. Then, we used the training set to build PR, SVM and MLP models. A detailed description of the applied sampling rates and ML algorithms is provided in the Methods section. In this work, we optimised the hyperparameters of each algorithm by minimising the loss function on both the training (data known for the model) and the validation (independent data) series and compared the results to evaluate whether a separate validation set is necessary or somewhat redundant in the process of model optimisation. Furthermore, this allows assessing if the best modelling settings (hyperparameters determined for a specific algorithm and sampling) can be re-used to obtain a suitable model for another data set of similar nature without needing a validation set.

After building 12 models, their quality was evaluated using the MAE, MRE and NLDE, computed as explained in the Methods section. **Figure 3** summarises the results obtained in the calculation of MRE for each model and the four selected sampling ratios. In the general quality assessment of the models, the lowest MAE (results not shown) and MRE (%) were produced by the PR algorithm. Nevertheless, the differences between PR and SVM, considering

both evaluation metrics are minimal, of approximately 0.2%. Compared with the SVM and PR models, the MAE and MRE computed for the MLP model are generally higher and increase for low sampling ratios.

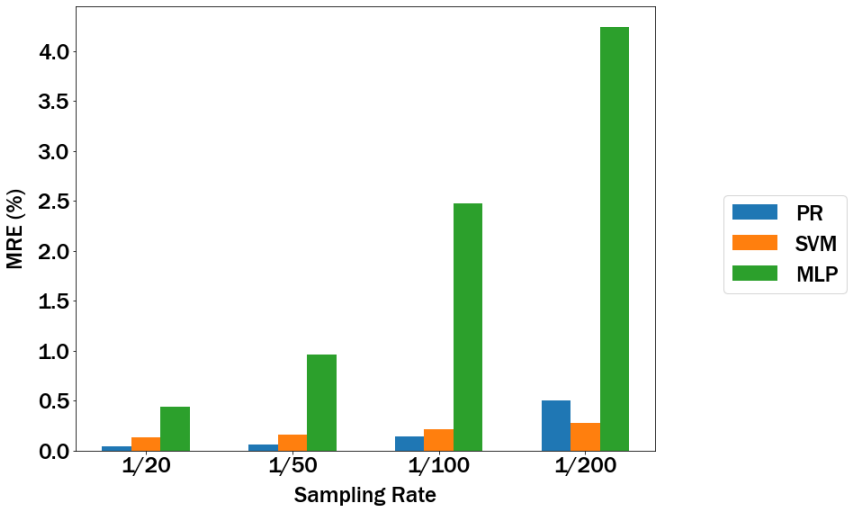


Figure 3: Selected evaluation metric for different ML models and partitions of data, MRE (%)

The plots in **Figure 4** illustrate the RE (%) calculated for the predicted APD_{90} values from the test series. We show the differences between the three tested models: PR (blue), SVM (orange), and MLP (green) and how the different sampling ratios impacted the evaluation metrics from the smallest to the highest. In models PR and SVM, the RE (%) range is smaller than for MLP. All the models have in common that the RE (%) is larger for APD_{90} below 300 ms and above 600 ms. In the graphical distribution of RE (%) along the APD_{90} axis, it is noticeable that the initial and end regions of the APD_{90} value range are the ones with the largest RE (%) increase. Nonetheless, out of the three model types, SVM is the only algorithm that does not make any prediction above the considered threshold of 5% of RE.

A closer observation of the differences between the APD_{90} obtained from the simulation and the predicted, expressed as RE (%), shows that the largest

errors have a periodic pattern. This can be observed, for example, in the region between 300 and 500 ms in the results of the PR with 1/100 sampling. These errors are produced by a border effect: the model does not fit well the data points located at the upper and lower limits of the input values. In these positions, there is an abrupt change of the surface, and some models struggle to fit the simulation results accurately. In particular, the use of equispaced sample points in PR can produce slight oscillations at the edges (Runge's phenomenon).[43]

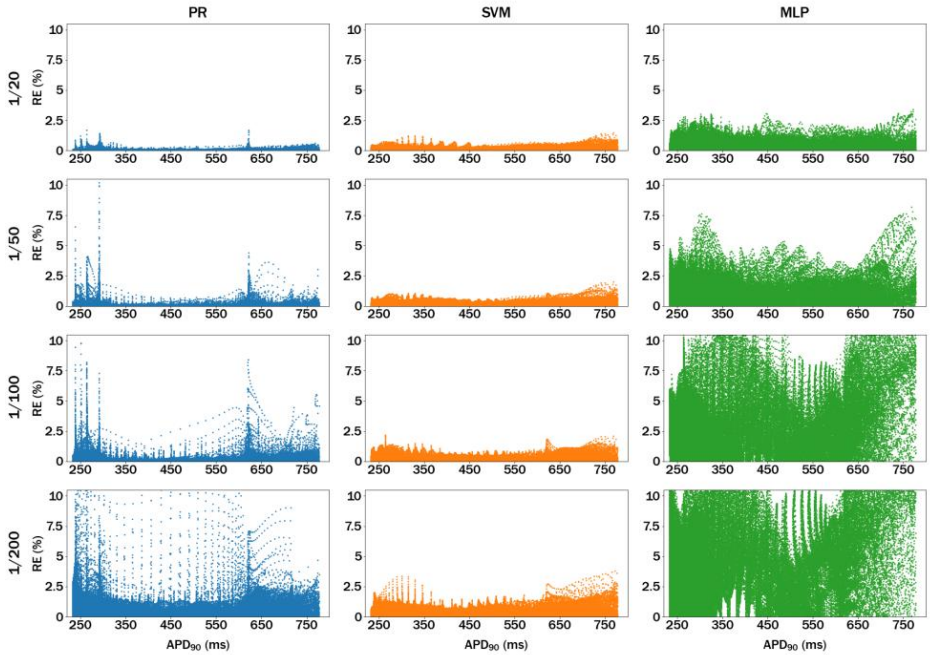


Figure 4: Each plot shows the RE (%) as a function of the experimental values of APD_{90}. Columns represent three trained models PR, SVM, and MLP. Rows correspond to the sampling ratios applied to the input data starting from 1/20 to 1/200.

Figure 5 represents a 3D plot, with APD_{90} in the Z (vertical axis) and I_{Kr} and I_{Ks} in the X and Y axes, respectively. A fixed value of 0.3 was used for I_{CaL} in all instances. For all models and sampling rates shown in the graphics, the predicted values correspond more precisely with the simulation results in the centre of the covered output ranges (APD_{90} between 300 and 600 ms). As

described above, the values predicted by the three different models are plotted using the following colours PR (blue), SVM (orange), and MLP (green), while grey was used to depict simulated values on each plot. Still, some models exhibit minor deviations in the borders for the reasons explained above. However, even in these areas, we obtain errors well below 5% for all SVM models.

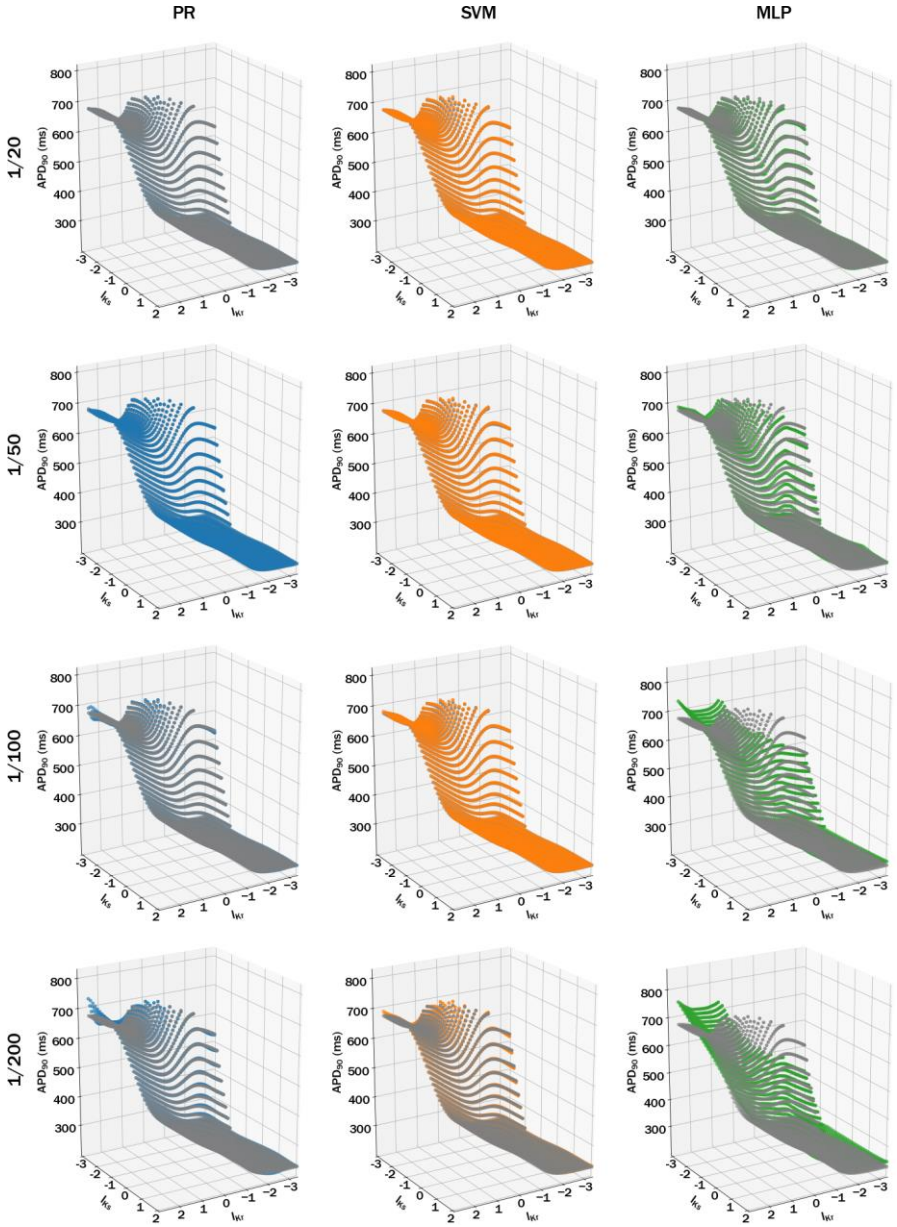


Figure 5: 3D plots representing I_{Kr}, I_{Ks} and APD_{90} for a fixed value of I_{CaL} equal to 0.3 to give an example. Columns represent three trained models PR, SVM, and MLP. Rows correspond to the sampling ratios applied to the input data starting from 1/20 to 1/200.

External validation using a set of CiPA compounds

A set of 12 CiPA compounds with well-defined cardiac electrophysiology, clinical response and known effective therapeutic concentration was used in our project to validate the predictive quality of the models.

Figure 6 (A) illustrates the APD_{90} simulated and predicted using the three ML models and the sampling rate of 1/100 for a set of 12 CiPA drugs. For all selected CiPA drugs except Quinidine, which poses a high risk of inducing TdP, the duration of the experimental APD_{90} interval lies below 300 ms. This trend remains unchanged for the APD_{90} values predicted by all three models. **Figure 6 (B)** illustrates the RE (%) for the CiPA dataset used for the external validation. The RE values are very low and below 1% in most cases. This external validation result confirms the results obtained in the validation and testing step of the model training, where the PR and SVM models perform comparatively well. In contrast, the predictions generated by the MLP model deviate more from the experimental values.

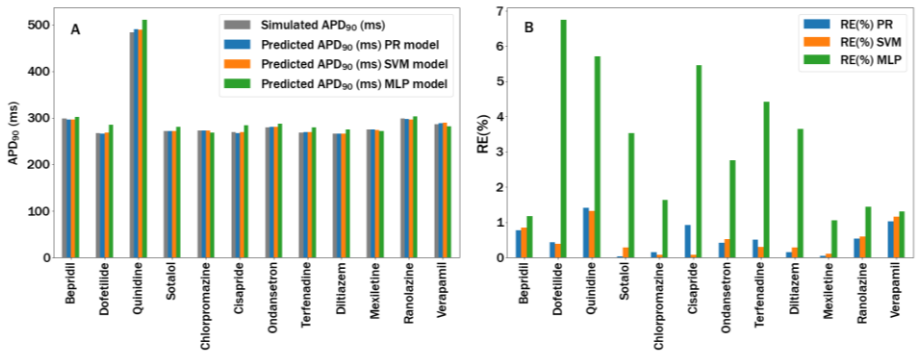


Figure 6: External validation of the three ML models built using the training set sampled 1/100 performed using a set of 12 CiPA drugs selected from three TdP risk classes. **A:** Simulated and predicted APD_{90} values. **B**: RE (%).

Simulation of future use by applying the developed methodology to another data array

Once a suitable sampling rate and algorithm were selected, and its hyperparameters were optimised, could these settings be used to fit biomarkers obtained from a different electrophysiological simulation? Should the hyperparameters be optimised again using a validation set? To answer these questions, a second pre-simulated data array was used. The simulations were carried out following the *in silico* action potential (AP) modelling protocol described in the Methods section but now using input values which reflect the degree of inhibition of three different ion currents (I_{Kr}, I_{NaL}, I_{CaL}).

Data simulation and sampling were done using methods equivalent to those described above. Further on, the assessment of the SVM using a sampling ratio of 1/100 was performed following two different approaches. The first option was identical to the methodology described for the array APD_{90} – (I_{Kr}, I_{Ks}, I_{CaL}), in which we used 1/100 data points for model training, 1/100 for validation, and 98/100 for testing. The hyperparameters for this model were determined based on the validation set. In the second scenario, we built an SVM model and optimised its hyperparameters as a function of the training set only, which was compiled by combining the training and validation sets (summing to 2 data points per 100).

Table 3: Performance metrics assessed for the model APD_{90} – (I_{Kr}, I_{Ks}, I_{CaL}) using (A) training, validation, and test set and (B) using the double amount of data for training and the rest for test set.

(A)

		SVM		
Sampling	Partition	MAE	MRE (%)	NLDE
1/100	Train	0.56	0.18	100.00
	Val	1.00	0.27	100.00
	Test	0.93	0.25	100.00

(B)

		SVM		
Sampling	Partition	MAE	MRE (%)	NLDE
1/100	Train	0.56	0.18	100.00
	Test	0.93	0.25	100.00

For the selected model and sampling rate, the results obtained using either two (**Table 3 (A)**) or three (**Table 3 (B)**) partitions are rather similar. Therefore, we found that in comparable situations, the same hyperparameters can be applied to train other models, making it unnecessary to include the validation partition.

Discussion

The methodology presented here allows the replacement of computationally costly simulations with estimations generated by a machine learning model. For the method to be profitable, the reduction must significantly impact the number of necessary simulations. In the Results section, it was shown that the number of data points available for training the model largely impacts the errors the model commits on average but selecting 1 of every 100 data points results in an excellent balance between the reduction of the calculations and the robustness and predictive accuracy of the simulation fitting.

Deciding on the necessary number of points required to capture the data structure is a problem-specific decision. In the current application, simulating 1/100 points would practically produce a one hundred-fold decrease in the number of required simulations and computation time, fulfilling our original objectives.

All in all, the described methodology led to the development of high-quality models able to produce APD_{90} values, which are a relatively accurate estimation of those produced by computationally intensive simulations. In this research, we obtained slight differences in the quality of the SVM as compared

to the PR model. The errors produced by PR at the borders can be justified by the use of regularly spaced sample points, and could be mitigated by the use of Chebyshev nodes .[43] However, for this particular work we considered that the use of an ad-hoc sampling for PR will not allow a fair comparison with other models. The advantage of applying polynomial transformation is the simplicity of the underlying mathematics, especially in contrast to the Neural Network or SVM models when large regularisation values are used for training. Therefore, PR would be the preferred algorithm if taking the lowest computational complexity as the criteria for choosing the model. But very often, fitting complex data requires the application of a high polynomial degree which goes in hand with a high probability of overfitting, which is the downside of PR. This issue can be resolved through the application of regularisation. The most common regularisation methods are Lasso (L1) and Ridge (L2). While Ridge regression introduces a penalty factor to shrink the magnitude of the model coefficients, Lasso eliminates some of the insignificant coefficients of the model. This difference was extremely important since all features in our input data were essential to model the biological problem correctly and therefore, L2 regularisation was selected instead of the more rigorous L1.

For this reason, if increasing the number of ion channels is the objective of future works, Polynomial Regression would not be the best choice. This is because incrementing the number of input values could yield less smooth surfaces, requiring an increase of the polynomial degree and more rigorous regularisation. On the other hand, the Support Vector Machine algorithm is characterised by a very high generalisation ability, even when the number of instances is less than the number of variables.[42] However, one of the downsides of SVMs for regression is its sensitivity to outliers, which highlights the importance of both data pre-processing and model optimisation. The

robustness of the SVM algorithm was confirmed in this work by obtaining high-quality models and precise predictions.

The third and last tested model, the MLP, did not generalise as well as the other two models. A possible explanation for this result may be the insufficient amount of data since Artificial Neural Networks generally require a lot of information to learn from and to predict well. Additionally, since the tuning of hyperparameters of MLP is comparatively expensive in terms of content and time, improving the performance of the neural network model would require testing a wider range of hyperparameters. Nevertheless, the scope of application of Multi-Layer Perceptron is wide and covers several modelling areas. To give a more related example, MLP algorithms were used with high accuracy in Arrhythmia Classification problems where the data was richer in specific information and valuable characteristics.[44]

With respect to the method limitations, the models described here were developed and optimised for a combination of three ion channels. When re-using this methodology for a different combination of channels or ventricular arrhythmia biomarkers, the model building and validation would need to be repeated to ensure high-quality results.

We used a specific model (a modified version of O'Hara and colleagues) to generate the APD_{90} array. There are many available models in the field for which the methodology is expected to work well. This, however, would need to be confirmed.

Conclusion

In this work, we have shown that it is possible to significantly reduce the number of simulations required to make accurate predictions of ventricular-arrhythmia biomarkers through the application of ML models. We demonstrated that the total amount of the originally simulated data points can

be reduced to just 1%. Such data reduction goes in hand with a significant reduction of the time necessary to produce an *in silico* prediction tool based on large pre-simulated datasets. The simple approach developed here opens up the possibility of modelling more complex biological processes, such as the alteration of ventricular-arrhythmia safety biomarkers as a response to an interaction of four and more ionic channels. Additionally, the methods described here are likely to be applicable to model other biomarkers than APD_{90} and even be applied to predict other computational simulation results in different fields of biomedical research. Lastly, the development of effective early-stage screening systems is aligned with the interests of pharmaceutical companies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors received funding from the eTRANSAFE project, Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777365. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. We also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016496. (SimCardioTest). J.L.L. is being funded by the Ministerio de Ciencia, Innovacion y Universidades for the "Formacion de Profesorado Universitario" (Grant Reference: FPU18/01659). The work was also partially supported by the Dirección General de Política Científica de la Generalitat Valenciana (PROMETEO/ 2020/043).

We wish to thank Prof. Emilio Soria-Olivas for the valuable discussion.

References

1. Bartos, D. C., Grandi, E. & Ripplinger, C. M. Ion channels in the heart. *Compr. Physiol.* **5**, 1423–1464 (2015).

2. Roden, D. M. Drug-induced prolongation of the QT interval. *N. Engl. J. Med.* **350**, 1013–1022 (2004).

3. Yap, Y. G. & Camm, A. J. Drug induced QT prolongation and torsades de pointes. *Heart* **89**, 1363 (2003).

4. Stockbridge, N., Morganroth, J., Shah, R. R. & Garnett, C. Dealing with global safety issues: Was the response to QT-liability of non-cardiac drugs well coordinated? *Drug Saf.* **36**, 167–182 (2013).

5. ICH S7B Non-clinical evaluation of the potential for delayed ventricular repolarization (QT interval prolongation) by human pharmaceuticals | European Medicines Agency. Available at: https://www.ema.europa.eu/en/ich-s7b-non-clinical-evaluation-potential-delayed-ventricular-repolarization-qt-interval. (Accessed: 25th May 2020)

6. ICH E14 Clinical evaluation of QT/QTc interval prolongation and proarrhythmic potential for non-antiarrhythmic drugs | European Medicines Agency. Available at: https://www.ema.europa.eu/en/ich-e14-clinical-evaluation-qtqtc-interval-prolongation-proarrhythmic-potential-non-antiarrhythmic. (Accessed: 25th May 2020)

7. Fermini, B. *et al.* A new perspective in the field of cardiac safety testing through the comprehensive in vitro proarrhythmia assay paradigm. *J. Biomol. Screen.* **21**, 1–11 (2016).

8. Sager, P. T., Gintant, G., Turner, J. R., Pettit, S. & Stockbridge, N. Rechanneling the cardiac proarrhythmia safety paradigm: A meeting report from the Cardiac Safety Research Consortium. *Am. Heart J.* **167**, 292–300 (2014).

9. Gintant, G., Sager, P. T. & Stockbridge, N. Evolution of strategies to improve preclinical cardiac safety testing. *Nat. Rev. Drug Discov. 2016 157* **15**, 457–471 (2016).

10. Li, Z. *et al.* Improving the in silico assessment of proarrhythmia risk by combining hERG (Human Ether-à-go-go-Related Gene) channel-drug binding kinetics and multichannel pharmacology. *Circ. Arrhythmia Electrophysiol.* **10**, (2017).

11. Hwang, M., Lim, C. H., Leem, C. H. & Shim, E. B. In silico models for

evaluating proarrhythmic risk of drugs. *APL Bioeng.* **4**, 021502 (2020).

12. Li, Z. *et al.* Assessment of an In Silico Mechanistic Model for Proarrhythmia Risk Prediction Under the CiPA Initiative. *Clin. Pharmacol. Ther.* **105**, 466–475 (2019).

13. Passini, E. *et al.* Human In Silico Drug Trials Demonstrate Higher Accuracy than Animal Models in Predicting Clinical Pro-Arrhythmic Cardiotoxicity. *Front. Physiol.* **8**, (2017).

14. Zhou, X. *et al.* Blinded in silico drug trial reveals the minimum set of ion channels for torsades de pointes risk assessment. *Front. Pharmacol.* **10**, 1643 (2020).

15. Lancaster, M. C. & Sobie, E. A. Improved Prediction of Drug-Induced Torsades de Pointes Through Simulations of Dynamics and Machine Learning Algorithms. *Clin. Pharmacol. Ther.* **100**, 371–379 (2016).

16. Llopis-Lorente, J. *et al.* In silico classifiers for the assessment of drug proarrhythmicity. *J. Chem. Inf. Model.* **60**, 5172–5187 (2020).

17. Christophe, B. Occurrence of early afterdepolarization under healthy or hypertrophic cardiomyopathy conditions in the human ventricular endocardial myocyte: In silico study using 109 torsadogenic or non-torsadogenic compounds. *Toxicol. Appl. Pharmacol.* **438**, 115914 (2022).

18. Yoo, Y., Marcellinus, A., Jeong, D. U., Kim, K. S. & Lim, K. M. Assessment of Drug Proarrhythmicity Using Artificial Neural Networks With in silico Deterministic Model Outputs. *Front. Physiol.* **12**, 2289 (2021).

19. Llopis-Lorente, J., Trenor, B. & Saiz, J. Considering population variability of electrophysiological models improves the in silico assessment of drug-induced torsadogenic risk. *Comput. Methods Programs Biomed.* **221**, 106934 (2022).

20. Cooper, F. R. *et al.* Chaste: Cancer, Heart and Soft Tissue Environment. *J. Open Source Softw.* **5**, 1848 (2020).

21. Beattie, K. A. *et al.* Evaluation of an in silico cardiac safety assay: using ion channel screening data to predict QT interval changes in the rabbit ventricular wedge. *J. Pharmacol. Toxicol. Methods* **68**, 88–96 (2013).

22. Romero, L. *et al.* In Silico QT and APD Prolongation Assay for Early Screening of Drug-Induced Proarrhythmic Risk. *J. Chem. Inf. Model.* **58**, 867–878 (2018).

23. Obiol-Pardo, C., Gomis-Tena, J., Sanz, F., Saiz, J. & Pastor, M. A multiscale simulation system for the prediction of drug-induced cardiotoxicity. *J. Chem. Inf. Model.* **51**, 483–492 (2011).

24. Khalifa, N., Kumar Konda, L. S. & Kristam, R. Machine learning-based QSAR models to predict sodium ion channel (Na v 1.5) blockers. *Future Med. Chem.* **12**, 1829–1843 (2020).

25. Varshneya, M., Mei, X. & Sobie, E. A. Prediction of arrhythmia susceptibility through mathematical modeling and machine learning. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

26. Aghasafari, P. *et al.* A deep learning algorithm to translate and classify cardiac electrophysiology. *Elife* **10**, (2021).

27. Cai, C. *et al.* Deep Learning-Based Prediction of Drug-Induced Cardiotoxicity. *J. Chem. Inf. Model.* **59**, 1073–1084 (2019).

28. Maleckar, M. M. *et al.* Combined In-silico and Machine Learning Approaches Toward Predicting Arrhythmic Risk in Post-infarction Patients. *Front. Physiol.* **12**, 1903 (2021).

29. O'Hara, T., Virág, L., Varró, A. & Rudy, Y. Simulation of the Undiseased Human Cardiac Ventricular Action Potential: Model Formulation and Experimental Validation. *PLOS Comput. Biol.* **7**, e1002061 (2011).

30. Mirams, G. R. *et al.* Simulation of multiple ion channel block provides improved early prediction of compounds' clinical torsadogenic risk. *Cardiovasc. Res.* **91**, 53–61 (2011).

31. Stigler, S. M. Gergonne's 1815 paper on the design and analysis of polynomial regression experiments. *Hist. Math.* **1**, 431–439 (1974).

32. Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55 (1970).

33. Drucker, H., Burges, C. J., Kaufman, L., Smola, A. & Vapnik, V. Support Vector Regression Machines. *Adv. Neural Inf. Process. Syst.* **9**, (1996).

34. M. A. Aizerman, È. M. Braverman, L. I. R. Theoretical foundation of potential functions method in pattern recognition. *Avtomat. i Telemekh* **25**, 917–936 (1964).

35. Broomnhead, D. S., Lowe DTIC SELECTE, D., Broomhead, D. & Lowe, D. Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks. (1988).

36. Rosenblatt, F. The perceptron: A probabilistic model for information

storage and organization in the brain. *Psychol. Rev.* **65**, 386–408 (1958).

37. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nat. 1986 3236088* **323**, 533–536 (1986).

38. Nair, V. & Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. (2010).

39. Pedregosa, F., Varoquaux, G. and Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

40. Harris, C. R. *et al.* Array programming with NumPy. *Nat. 2020 5857825* **585**, 357–362 (2020).

41. McKinney, W., & others. Data structures for statistical computing in python. *Proc. 9th Python Sci. Conf.* **445**, 51–56 (2010).

42. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

43. Trefethen, L. N. *Approximation Theory and Approximation Practice, Extended Edition*. (SIAM-Society for Industrial and Applied Mathematics, 2019).

44. Savalia, S. & Emamian, V. Cardiac Arrhythmia Classification by Multi-Layer Perceptron and Convolution Neural Networks. *Bioengineering* **5**, (2018).

Capítulo 4

Uncertainty assessment of proarrhythmia predictions derived from multilevel in silico models

Karolina Kopańska[1*], Pablo Rodríguez-Belenguer[1,2*], Jordi Llopis-Lorente[3], Beatriz Trenor[3], Javier Saiz[3], Manuel Pastor[1]

[1]*Research Programme on Biomedical Informatics (GRIB), Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Hospital del Mar Medical Research Institute, Barcelona, Spain*

[2]*Department of Pharmacy and Pharmaceutical Technology and Parasitology, Universitat de València, Valencia, Spain*

[3]*Centro de Investigación e Innovación en Bioingeniería (Ci2B), Universitat Politècnica de València, Valencia, Spain*

[*]*Both authors have contributed equally*

Revista: Archives of Toxicology

Editorial: Springer

Año: 2023

Cuartil: Q1

IF: 6.1

Introduction

Ventricular arrhythmias, especially polymorphic ventricular tachycardia known as Torsade de Pointes (TdP), are very serious and feared adverse drug effects. The early estimation of the potential of drug candidates to induce ventricular arrhythmias is therefore of the highest interest to all stakeholders in healthcare (Gintant et al., 2016b). The main mechanism of drug-induced ventricular arrhythmia involves the inhibition of one or multiple ion channels present in the membrane of ventricular myocytes. Such inhibitory effects prolong the action potential (AP) duration of ventricular cells triggering effects at the organ level. These prolongation effects can be observed at the patient level as changes in the duration and shape of QT-intervals on the surface ECG (Roden, 2004b; Yap & Camm, 2003b).

Since 2005, proarrhythmia assessment of pharmaceuticals for human use has been carried out according to the guidelines ICH S7b and ICH E14. In the non-clinical phase (ICH E14, 2005; ICH S7B, 2005), the risk is estimated by combining results from *in vitro* inhibition assays of the rapid delayed rectifier potassium current (I_{Kr}) encoded by the human ether-a-go-go-related gene (hERG) and an *in vivo* animal QT-prolongation studies, while in clinical phases (ICH E14, 2005; S7B, 2005), drug proarrhythmia is assessed by measuring *in vivo* human QT/QTc interval prolongation. A decade later, the comprehensive *in vitro* proarrhythmia assay (CiPA) initiative enriched the mechanistic description of proarrhythmia and complemented the assessment by incorporating *in silico* methodologies (Fermini et al., 2016b; Sager et al., 2014b). The four-stage CiPA paradigm highlights the value of considering drug effects on a set of ion currents (I_{Na}, I_{NaL}, I_{Kr}, I_{to}, I_{CaL}, I_{K1}, and I_{Ks}) as independent factors involved in arrhythmogenesis, instead of relying on I_{Kr}, only (Z. Li et al., 2017b; Sager et al., 2014b). The potency of drug-mediated inhibition of those ion channels, usually measured as the half maximal inhibitory concentration

(IC_{50}), serves as input for electrophysiological models, which translate this information into proarrhythmia biomarkers (Z. Li, Ridder, et al., 2019b; Park et al., 2019).

In the last decade, several efforts have been undertaken to enhance the assessment of proarrhythmia by introducing meta-models. Such meta-models are trained using larger series of simulation results, which allows for instantaneous predictions of selected proarrhythmia biomarkers. In particular, (Mirams et al., 2014b) described a meta-model built from simulated APD data for a series of different combinations variating the level of ion channel inhibition between 0% and 100% for five ionic transporters, including hERG, CaV1.2, NaV1.5, KCNQ1/MinK, and Kv4.3/KChIP2.2. Moreover, our groups also developed a multi-level *in silico* tool for the prediction of drug-induced action potential duration at 90% of repolarization (APD_{90}) and QT-interval prolongation (Obiol-Pardo et al., 2011b; Romero et al., 2018b). The core of this tool was a large 3D data array containing a large number of simulated APD_{90} prolongation effects generated by the inhibition of 3 relevant ion channels (I_{Kr}, I_{Ks}, and I_{CaL}). Since these values were pre-computed for a wide range of inhibition values, the method can provide an instantaneous estimate of the APD_{90} duration in ventricular cardiomyocytes, using as inputs the values of IC_{50}s for these channels and the plasma concentration of the drug. In a recent work, this approach was optimized by replacing the 3D data array with a machine learning (ML) model trained using only a small fraction of these costly computational simulations, leading to a significant reduction of the number of simulations required to obtain reliable APD_{90} estimates (Rodríguez-Belenguer et al., 2023b).

Although computational approaches are a valuable complement to purely experimental methods, a detailed assessment of the variability and uncertainty associated with the predictions is required to increase the

reliability of *in silico* methods (Gosling, 2019). Quantification of variability and uncertainty in computational modelling systems and their predictions has been the objective of previous works in the cardiac safety field (Mirams et al., 2016, 2020).

Several different methodologies have been described for the characterisation of variability observed when *in vitro* experiments are conducted to measure ion channel blockade produced by chemicals. Mirams et al. (2014) described the use of a meta-model for the characterisation of uncertainty in ion channel block and to further propagate these uncertainties considering a combination of channels. (Chang et al., 2017) analysed the uncertainty and variability in drug binding and drug ionic current block for TdP risk assessment using the non-parametric bootstrap method and a Bayesian inference approach. (Elkins et al., 2013) assessed the amount of between-experiment variability in drug-blockade of I_{Kr} (hERG), I_{Na} (NaV1.5), I_{CaL} (CaV1.2), I_{Ks} (KCNQ1/151ink), and I_{to} (Kv4.3/KchIP2.2) channels using concentration-effect curves fitted for positive control compounds from high-throughput-screening experiments performed at Glaxo Smith Kline and Astra Zeneca. (Kramer et al., 2020) performed an extensive analysis of variability in results obtained from automated patch-clamp measurements across analysis sites and experimental platforms, thereby pointing out the importance of following the principles of Good Laboratory Practice (GLP) to minimise variability.

Another important source of variability are inter-individual differences among patients receiving the same drug treatment. When applying *in silico* approaches, the electrophysiological models that integrate ion channel specific IC_{50} into ventricular arrhythmia biomarkers make use of a large number of parameters that were adjusted to fit experimental results. However, humans are not physiologically identical, and no single electrophysiological model can produce results suitable for representing all patients, nor accurately

explain the observed differences between patients (Wisniowska et al., 2017). Population-based approaches have been described as a useful strategy to consider the inter-individual variability in the parameters of *in silico* models. (Britton et al., 2013) analysed the inter-subject variability by generating a population of cellular AP models, each of which exerted small differences in parameters. These models were consequently filtered following physiologically based criteria and using acceptance-rejection criteria, as shown by (Llopis-Lorente et al., 2022b). Such populations of models can serve for the estimation of variability in the responses of a human population. Another approach for the analysis of biological variability was proposed by (Johnstone et al., 2016), who used Bayesian statistics to infer distributions of inputs and parameters, such as current maximal conductance. (Pathmanathan et al., 2015) performed an extensive analysis of uncertainty in the steady-state inactivation of the fast sodium current using an individual-based statistical method, the nonlinear mixed effects (NLME) modelling, to analyse voltage clamp data taken from a population of cells.

Once diverse sources of variability and uncertainty in model inputs and parameters are identified, Uncertainty Quantification (UQ) analysis should be conducted to characterise and quantify their impact on models' final outcomes. When input uncertainties are expressed using probabilistic terms, UQ is typically performed by applying sampling-based techniques to propagate them through the model, generating a distribution of model outputs. Monte Carlo (MC) simulations and Latin Hypercube Sampling (LHS) are the most popular methods for sampling-based uncertainty propagation (Clayton et al., 2020), but the application of other propagation approaches has been reported. For example, (Sobie, 2009) used multivariate regression for the assessment of the impact of variabilities in channel conductance, time constants, and steady state voltage offsets. In the second case study described by (Johnstone et al., 2016), they demonstrated the use of the Gaussian

Process (GP) emulator to assess the effects of the uncertainties in AP model parameters once they are propagated to the output ((Johnstone et al., 2016). Lately, (Hu et al., 2018) described the use of polynomial chaos for the propagation of uncertainties and global sensitivity analysis within a multi-level cardiac electrophysiology prediction framework. In most published works, the UQ was performed only on a subset of model parameters. (Pathmanathan et al., 2019) followed a different approach, suggesting that simpler models with a robust and complete UQ may be more useful than complex models without a full UQ. They performed the UQ on a canine cardiac cell model, which was reduced to relatively few parameters to which they assigned input distributions, controlled by a user-dependent hyperparameter.

In this work, we extend our multi-level *in silico* proarrhythmia model by integrating a comprehensive analysis of uncertainty. We start by identifying all sources of aleatory and epistemic uncertainty typically present in cardiac safety models. Focusing exclusively on aleatory uncertainty, we then investigate which of the identified sources affect the inputs of our model. We develop methods for the characterisation and propagation of the selected uncertainty types through the model, using applicable approaches and simple simulation methods, respectively. These methods aim to provide a more realistic representation of proarrhythmia biomarker predictions and allow for studying the individual and combined effect of different aleatory uncertainty sources on proarrhythmia biomarker predictions.

Methods

Multi-level *in silico* proarrhythmia model

In 2011 and 2018, we published two works (Obiol-Pardo et al., 2011c; Romero et al., 2018b) describing the development and refinement of a multi-level *in silico* method for predicting cardiac safety biomarkers (APD_{90} and QT-interval duration). This prediction method, shown in **Figure 1,** uses precomputed

simulations for estimating how compounds with different inhibitory effects on selected ionic currents can affect the ventricular tissue at certain concentrations. The inputs include IC_{50} values, obtained either in patch-clamp assays or predicted by *in silico* Quantitative Structure–Activity Relationship (QSAR) models, for three currents (here: I_{Kr}, I_{NaL}, I_{CaL}), the drug concentration, and a set of electrophysiological simulation parameters. Recently, we developed an optimised version of this method in which the high number of precomputed simulations was significantly reduced through the application of machine-learning (Rodríguez-Belenguer et al., 2023b).

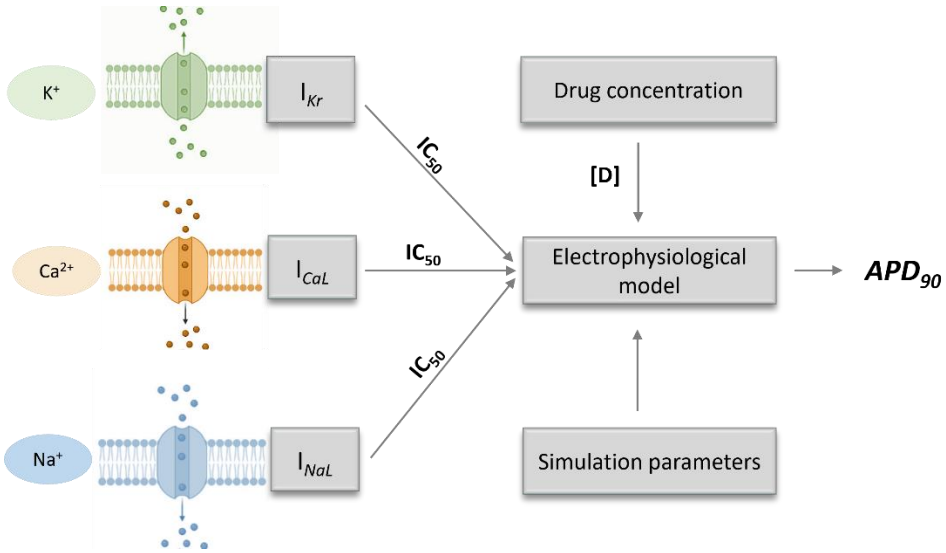


Fig. 1: A simplified schema of our multi-level *in silico* proarrhythmia model. For a single compound, the input comprises a set of IC_{50} values for the currents I_{Kr}, I_{NaL}, I_{CaL}, a drug concentration, and a set of electrophysiological simulation parameters. The model translates these inputs to an APD_{90} prediction.

Electrophysiological simulations

In silico action potential (AP) modelling of the healthy human endocardial cardiomyocyte and APD_{90} measurements were done using the widely known model published by (O'Hara et al., 2011b), modified as described by (Llopis-Lorente et al., 2020b). Here, we considered drug effects on APD_{90} as a function of the three selected currents; I_{Kr}, I_{NaL} and I_{CaL}, which are considered

particularly relevant for drug-induced occurrence of ventricular arrhythmias and are usually included in the pre-clinical ion channel screening panel at pharmaceutical companies (Chang et al., 2017).

Electrophysiological machine-learning model

We ran a set of electrophysiological simulations covering a wide range of combination of values for the ratio $\left(\frac{\mathrm{D}}{\mathrm{IC}_{50,\mathrm{i}}}\right)^{\mathrm{h}}$ for I_{Kr}, I_{NaL}, I_{CaL}. These ratios values were used to calculate channel inhibition via the simple pore block model (**Equation 1**).

$$g_{i,drug} = g_i \left[1 + \left(\frac{D}{IC_{50,i}}\right)^h\right]^{-1} \quad \textbf{Equation 1}$$

where $g_{i,\,drug}$ represents the maximal conductance of channel i in the presence of the drug, D is the drug concentration, $IC_{50,\,i}$ is the half-maximal inhibitory concentration for that drug, and channel i and h is the Hill coefficient.

The results obtained from the simulations (APD_{90}) were stored in an array, consisting of 3 input values (IV) corresponding to I_{Kr}, I_{NaL}, and I_{CaL} channels. Each IV was calculated by taking the logarithm of the ratio $\left(\frac{\mathrm{D}}{\mathrm{IC}_{50,\mathrm{i}}}\right)^{\mathrm{h}}$, as described in **Equation 2**. For each channel (I_{Kr}, I_{NaL}, I_{CaL}), the input value ranged from −3 to 2.5, with a step increment of 0.1.

$$IV = log_{10}\left(\left[\frac{D}{IC_{50}}\right]^h\right) \quad \textbf{Equation 2}$$

The standard utilisation of this array was as follows: for a given compound at a concentration D, **Equation 2** was applied independently for the three ionic channels (I_{Kr}, I_{NaL}, I_{CaL}). The resulting values were rounded to the first decimal and constrained between −3 and 2.5, i.e., if an input value was lower than −3 or higher than 2.5, the value was then transformed to −3 or 2.5, respectively. For each combination of the three calculated IV, the corresponding output

(APD_{90}) was retrieved from the array. For example, a drug with the following IC_{50}s: 1 nM for I_{Kr}, 1000 nM for I_{NaL} and 10 nM for I_{CaL} at a concentration of 1 nM yielded the data point [0, −3, −1], which led to an APD_{90} of 369.06 ms.

The results of these simulations (APD_{90}) were used to build an SVM model, as described in (Rodríguez-Belenguer et al., 2023b). This model can be effectively used to predict APD_{90} for any compound with an *IV* within the range covered by the model training series. Indeed, this range expands from -3 to 2.5 and is wide enough to represent the values found in most drugs and drug candidates. To limit the prediction space of this model, any *IV* minor than the minimum or superior to the maximum acceptable threshold is rounded accordingly. Hence, no values below -3 or above 2.5 are used to predict the APD_{90} values.

Uncertainty assessment protocol

According to EFSA's "Guidance on Uncertainty Analysis in Scientific Assessments" ((Benford, Halldorsson, Jeger, Knutsen, More, Naegeli, Noteborn, Ockleford, Ricci, Rychen, Schlatter, Silano, Solecki, Turck, Younes, Craig, Hart, Von Goetz, Koutsoumanis, Mortensen, Ossendorp, Martino, et al., 2018), UQ should commence with a comprehensive identification of all sources of uncertainty that have the potential to alter the assessment conclusion. In addition, the ECHA and the WHO recommend a complete and transparent characterisation of uncertainty in model inputs and the methodology by conducting a probabilistic analysis (European Chemicals Agency, 2012; Organization & on Chemical Safety, 2018).

In our protocol, the assessment question was defined as follows: "What is the APD_{90} that a certain drug will produce in an individual of a healthy population considering the compound's potency of inhibition of the considered ion channels at a specific concentration?" As a first step, we identified all aleatory and epistemic factors that contribute to the uncertainty in the output used to

answer the assessment question, when using the *in silico* proarrhythmia multi-level model. The next step was to investigate which sources of uncertainty affect the inputs of our model, thereby focusing specifically on the aleatory ones. Monte Carlo simulation was used to study how their effect on the input propagates through our model and is reflected on its output. Results of these simulations were expressed as values and intervals. The values can be interpreted as the most probable estimates of APD_{90} and the intervals as ranges of values within which the prediction could fall, given a certain level of credibility.

Identification of the main sources of variability and uncertainty in cardiac safety models

In order to correctly identify different sources of uncertainty, it is particularly important to distinguish between their aleatory or epistemic character (Benford, Halldorsson, Jeger, Knutsen, More, Naegeli, Noteborn, Ockleford, Ricci, Rychen, Schlatter, Silano, Solecki, Turck, Younes, Craig, Hart, Von Goetz, Koutsoumanis, Mortensen, Ossendorp, Germini, et al., 2018). To make a clear distinction, an overview of the most important sources of aleatory and epistemic uncertainties is presented in **Figure 2**, adapted from (Shamsi et al., 2020). The uncertainty types and the examples provided below apply to cardiac physiome models as previously described by (Mirams et al., 2016).

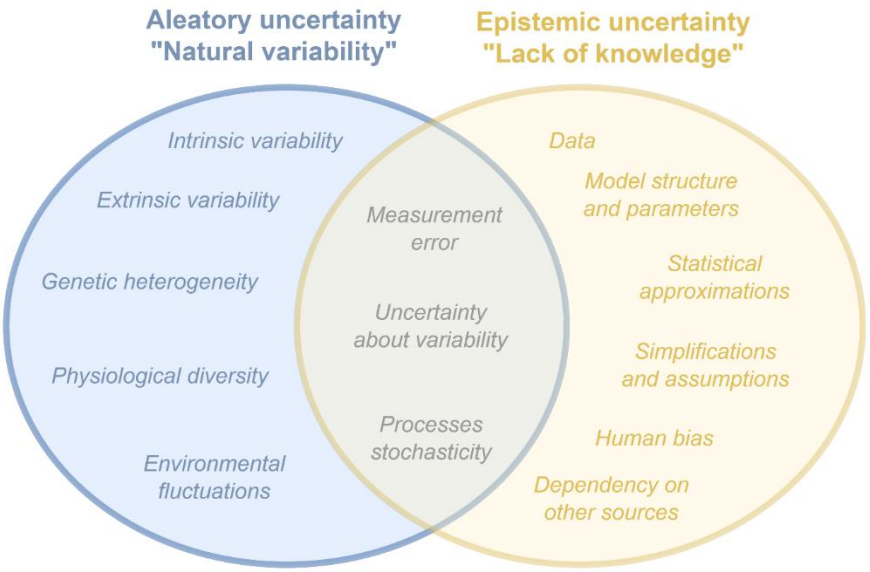


Fig. 2: Identified sources of aleatory and epistemic uncertainty affecting elements of *in silico* multi-level proarrhythmia models.

The term aleatory uncertainty, which is used interchangeably with variability, refers to the indispensable heterogeneity and diversity that occurs within biological populations, let them be biological samples or human individuals. Variability, which can be controlled and measured but never completely removed, is reflected in multiple values that a quantity of interest can take on. Generally, variability can be subdivided based on the criteria, whether the differences are observed within the same subject (e.g.: the same cell or the same person) or among different subjects (e.g.: a collection of cells or a specific human population). These types are referred to as intrinsic or extrinsic variability, respectively. Aleatory uncertainty can also be classified considering the biological levels of organisation at which differences can be observed. Both, intrinsic and extrinsic variability can have their onset at the genetic (DNA of an organism), physiological (an organism), the environmental (population of organisms) levels, as well as at all intermediate levels that connect them.

On the contrary to aleatory uncertainty, when a parameter can only have a single true value but the knowledge to define it is lacking, it is described as epistemic uncertainty, shortly called uncertainty (Johnstone et al., 2016). In the context of computational modelling, epistemic uncertainty can be attributed to the model either through its inputs or through the underlying methodology. As for epistemic uncertainty in the inputs, it results mainly from incomplete data-gathering steps or the sparseness of the collected information. Concerning the methodology, uncertainty can have its origin in the structure of the model, in the selected algorithms and parameters or the introduced interpolation or extrapolation factors. The overall methodological process, including steps that proceed or succeed in the actual prediction, is also subject to epistemic uncertainty. These encompass all assumptions, simplifications or statistical approximations made to develop the model or to interpret its results. Uncertainty can also arise as a result of coding errors or the failure to consider the dependency between sources of the required information.

Despite the theoretical differences, variability and uncertainty are tightly connected since the epistemic uncertainty about a quantity of interest is often expressed based on a summary of aleatory uncertainty. More specifically, when the knowledge to define parameters for the characterisation of variability is generally incomplete, or the assumptions made to do so are incorrect, there is uncertainty about variability (Benford, Halldorsson, Jeger, Knutsen, More, Naegeli, Noteborn, Ockleford, Ricci, Rychen, Schlatter, Silano, Solecki, Turck, Younes, Craig, Hart, Von Goetz, Koutsoumanis, Mortensen, Ossendorp, Germini, et al., 2018). There are further cases when the separation between aleatory and epistemic uncertainty is not clear. A very well-known example is the occurrence of measurement errors that combine both the imprecision resulting from inevitable fluctuations in the measurement process

and intrinsic and extrinsic variability between measurements of the same quantity (Johnstone et al., 2016).

Sources of variability considered in this work

Computational models can simultaneously be affected by more than one source of uncertainty. In this work, aleatory uncertainty, which as mentioned above is mainly referred to as variability, was the only characterised and quantified subtype of uncertainty. Particularly umbrella terms were used to group the variability sources that affect each specific input of our multi-level proarrhythmia model. The associations between model inputs and variability types were additionally marked within the basic structure of our model, as shown in **Figure 3**. There are several epistemic factors associated with the inputs and the methodology, each of which can be reduced or even removed by filling the knowledge gaps. However, even if we acknowledge its importance, the quantification of epistemic uncertainty is out of the scope of this publication.

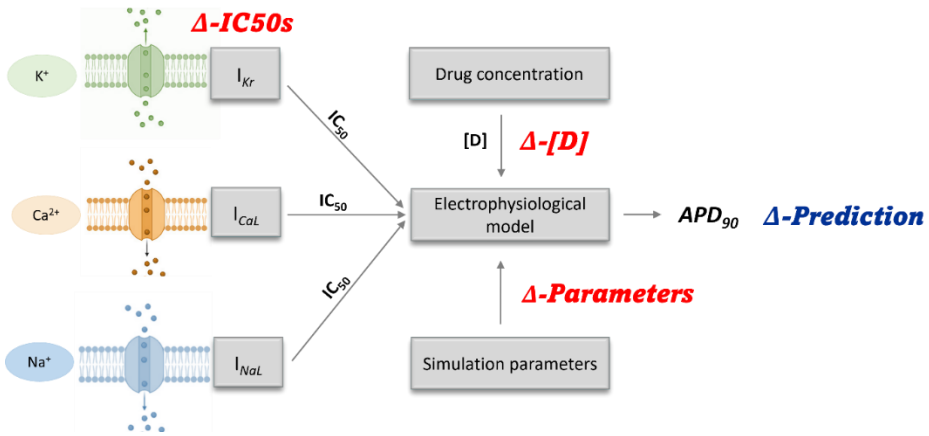


Fig. 3: Structure of our multi-level APD_{90} prediction model showing the sources of variability that affect model inputs addressed in this work. *Δ-IC_{50}s* represents the variability in the determined inhibitory drug effects on ion channels involved in physiological action potential generation. *Δ-Parameters* describes the variability in the electrophysiological model parameters due to inter-individual differences. *Δ-[D]* is the variability of the drug concentration obtained after the administration of the drug at therapeutic dosage due to inter-individual pharmacokinetic differences. Each

of the input variability sources contributes to the overall level of output variability, indicated as (*Δ-Prediction).*

In our model, the inhibitory effects of drugs targeting ion channels are introduced as IC_{50} values. These values can either be measured experimentally or predicted using QSAR models for each considered ion channel. For IC_{50}s measured experimentally, we assumed that the differences arising from intrinsic and extrinsic properties of analysed cellular systems can be summarised as *experimental variability (Δ-IC_{50}s)*. Here, we also account for the imprecision of repeated laboratory experiments since this factor cannot be separated from the measured values. Indeed, experimental variability could also be considered in the case of the Hill coefficient, which is a constant required to calculate the *IVs* for the model. Nevertheless, this constant is equal to one for many drugs, and even in a different case, the impact of a numeric change of h when computing *IV* (**Equation 2**) is rather small (Parikh et al., 2017; Romero et al., 2018b). Assuming that the consideration one more source of variability with a minimal impact on the predictions could introduce additional complexity and potentially complicate the interpretation of results for those variability sources whose impact on the prediction outcome is more significant, experimental variability associated with the Hill coefficient was not considered in this work.

The second model input affected by variability are the parameters defined to conduct electrophysiological cellular simulations. Here, we talk about the *inter-individual variability* (*Δ-Parameters*) that refers to the differences between individuals in the population. To be more specific, in the context of this publication the umbrella term *inter-individual variability* unites practically all sources of aleatory uncertainty shown in **Figure 2**. These include intrinsic and extrinsic differences between different cells within one human body and between several individuals, respectively. It also counts in genetic heterogeneity as well as environmental fluctuations, which together trigger

different epigenetic modifications and hence, physiological diversity between people and their hearts. Lastly, when cardiac activity is measured experimentally, random and measurement errors may also be taken into account.

Another important model input affected by the presence of variability is the *drug concentration (Δ-[D])*. When assessing the arrhythmogenic properties of a compound, it is common to use the Effective Free Therapeutic Plasma Concentration (EFTPC) to describe the protein unbound drug concentration present in the blood of patients treated with therapeutic doses. But the intrinsic and extrinsic variability also affects the pharmacokinetic (PK) processes of absorption, distribution, metabolism, and excretion, shortly ADME. Methods to address variability in drug concentration will be discussed later but will not be applied in our approach.

Quantitative characterisation of selected types of variability

Different guidelines recommend to derive measures of variability from representative observation data containing multiple instances of the quantities of interest that follow a certain distribution of frequencies and their spread (Hastie, Tibshirani, Friedman, et al., 2009; Shikano et al., 2012). Hence, the frequentist approach to probability was applied to characterise variability associated with the inputs of the multi-level proarrhythmia model. Incorporating pragmatic approximations based on different approaches described in detail below, it was assumed that experimental and inter-individual variability can be quantitatively described using normal probability distributions. The standard deviation (sd) was used to describe data spread.

Experimental variability in IC_{50}

Variability in experimentally measured pIC_{50} ($-\log_{10}(IC_{50})$) was characterised by (Elkins et al., 2013), who assumed that both, the pIC_{50} and sd parameter are

the same as, or very proximate to the one in control assays when the number of repeated measurements is high enough. The sd of the values measured in their study varied between ion channels, control compounds, and the number of repeats, reaching the minimum and maximum values of 0.08 and 0.2, respectively. Moreover, they showed that the pIC_{50} values collected in reiterated control assays on the same compound follow a logistic distribution.

We integrated these assumptions and represented the variability by considering that the experimental value is at the centre of a normal distribution, with a sd of 0.5. We chose a normal distribution for simplicity, due to its similarity with logistic distribution (similar in shape but with slightly higher kurtosis) (Hosmer Jr et al., 2013). The use of 0.5 is an approximation under the assumption that laboratory requirements stated in the GLP principles and stable testing conditions were not met during the measurement of IC_{50} values used in this work.

Inter-individual variability

To characterise the inter-individual variability, we applied the population-based approach previously described (Britton et al., 2013; Llopis-Lorente et al., 2022b; Muszkiewicz et al., 2016; Sobie, 2009). A modified version of the widely used AP endocardial model developed by O'Hara et al. (2011) (O'Hara et al., 2011b) was used as the baseline model. Assuming the baseline model represents the "averaged" model, an initial population of 1,000 models was generated by randomly and simultaneously applying a scaling factor to the 15 channel conductances of the AP model. These scale factors modifying the channel conductances were randomly sampled from a normal distribution with mean 1 and standard deviation 0.2, thus assuring most of the population (>99%) was in a range between ±60% with respect to the baseline model. This range covers the natural variability reported experimentally in human ventricular tissues (Fink et al., 2008; Romero et al., 2009; Volders et al., 2000).

The 1,000 models were simulated at 37ºC and at the following extracellular concentrations: $[Na^+]$ = 140 nM, $[Ca^{2+}]$ = 1.8 nM and $[K^+]$ = 5.4 nM. Then a calibration was performed. Plausible electrophysiological properties were defined according to experimental measurements for 15 biomarkers related to AP duration, amplitude of membrane potential, and calcium dynamics. Limits of acceptance for the 15 electrophysiological properties were taken from Table 1 in (Llopis-Lorente et al., 2022b). These ranges were obtained from a variety of experiments conducted on different hearts and cardiac regions (Britton et al., 2017; Coppini et al., 2013; Grandi et al., 2010; O'Hara et al., 2011b; Pieske et al., 2002; Sampedro, 2020; Schmidt et al., 1998). After calibration, 860 models presented a plausible electrophysiological behavior according to experimental data. Sacling factors of the final population are available in "ORdmD scaling factors.xlsx" at https://riunet.upv.es/handle/10251/182593.

Population of input value combinations

The population of 860 models was used to generate a distribution of APD_{90} predictions for a given set of 125 input value combinations, selected to represent properties similar to those of real compounds. These values spread regularly along all dimensions in the 3D array covering all possible combinations (5^3) of the five following values: -3, -1, -0.5, 0, 1 for the three channels I_{Kr}, I_{NaL}, I_{CaL}). Whether these distributions have the same shape and dispersion for diverse input values was first evaluated visually by plotting the value distributions as individual histograms.

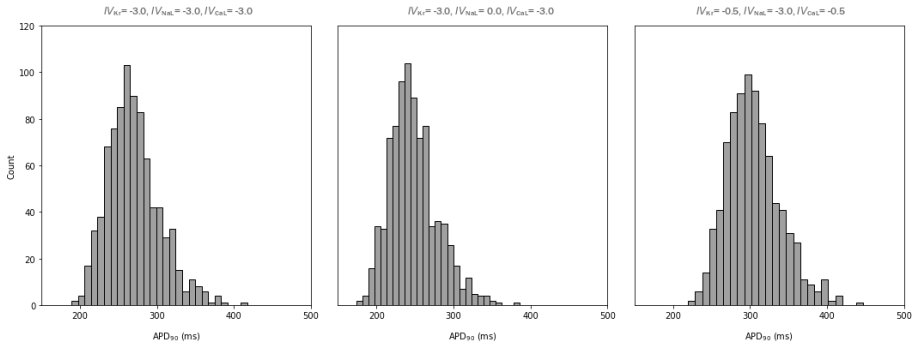


Fig. 4: Distributions of APD_{90} values generated by the population of 860 electrophysiological models for input values #0 (left), #3 (middle) and #60 (right), from the input value combinations shown in the graphics.

The example histograms in **Figure 4** represent the distribution of APD_{90} values obtained for three of these input value combinations. Left graphic, obtained using the input value combination (-3, -3, -3), shows an APD_{90} distribution generated assuming no inhibition of the selected channels. The remining two distributions illustrate distributions of output values produced for different input values combinations where inhibition was accounted for. The shape of the distributions is approximately normal (as checked using quantile-quantile plots) and for the 125 conditions tested, the average sd is of 35.4 ms, even if the dispersion is not homogeneous and different sd values were obtained for different input values.

The data table composed of 860 APD_{90} predictions generated for 125 input value combinations was used to build a model for predicting the dispersion (sd) of the distributions for a given set of input values. When generating predictions, this model produces an estimate of the dispersion of an APD_{90} distribution, for any drug with a combination of input values within the range covered by the models' training series. This predicted dispersion can be seen as an approximation of variability associated with APD_{90} prediction due to the inter-individual differences in the electrophysiological parameters. The models were built using a method similar to the one described extensively in our

previous work (Rodríguez-Belenguer et al., 2023b). SVM algorithm was used for the dispersion model, and the following hyperparameters were selected after optimizing the model: C = 1, kernel = Radial Basis Function (RBF), gamma= Scale. The goodness of fit was assessed as per mean absolute error (MAE = 0.35) computed for the test set.

Propagation and quantitative expression of variability in model outputs

Variability was propagated applying the forward Monte Carlo (MC) simulation approach (Kitagawa & Sato, 2001). The MC technique belongs to a broader group of stochastic simulation methods that allow for the generation of random numbers in order to solve problems of non-deterministic nature. The advantage of such a method is that no assumptions about the model must be made. Moreover, the simplicity and simultaneous correctness of the methodology are very convenient. In the context of variability assessment, MC requires the identification of all random components of a model and defining their interactions with other elements. It is important to consider the correlation between the level of randomness, or variability, and the number of samples needed to propagate such variability, thereby maintaining the reliability of the result. In other words, the greater the spread parameter describing the variability, the more samples must be drawn from the probability distribution. Moreover, as the result is highly dependent on the assumed distribution to be sampled with the MC method, the preparatory work to make correct assumptions with regard to the random variables is particularly important (Kroese & Rubinstein, 2012).

The simulations were run considering only experimental variability (Simulation A), only the variability due to inter-individual differences (Simulation B), or a combination of both variability types (Simulation C), as shown in **Figure 5**.

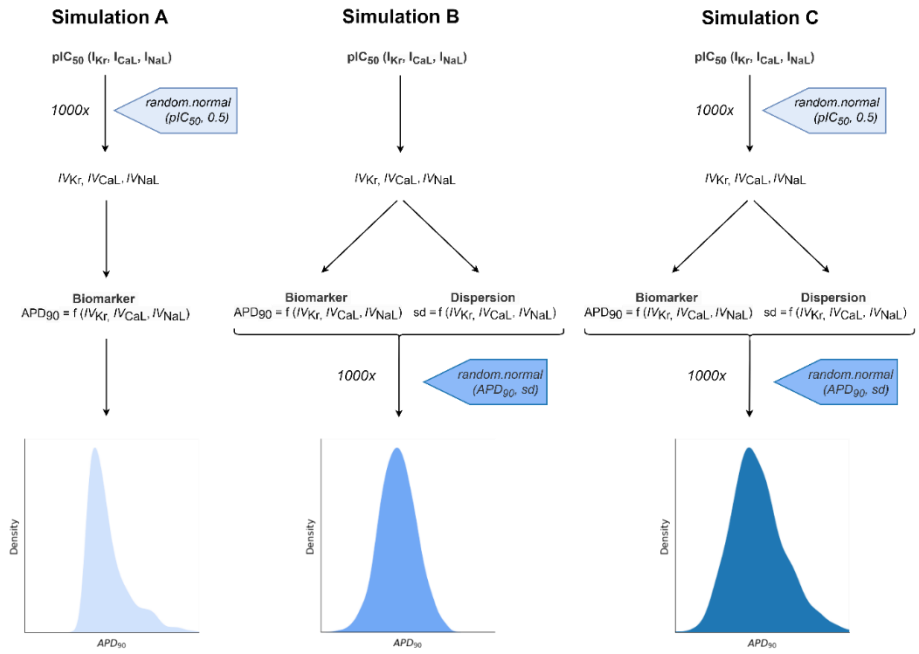


Fig. 5: Schema of the three simulation types carried out in this work. Simulation A – propagation of experimental variability associated with pIC_{50} values, Simulation B – propagation of inter-individual variability arising at the level of electrophysiological model parameters, Simulation C – propagation of combined experimental and inter-individual variability.

In all instances, the multi-level model described in **Figure 1** was applied 1000 times. In each simulation run, normally distributed random values were added up to specific elements of the model, using the random.normal(mu, sigma) function provided by the numpy library with a mu value of 0.0 and a sigma equal to the standard deviation of the variability represented, as described in the previous section.

In Simulation A, conducted to represent experimental variability in pIC_{50} values, the random value was added to the pIC_{50} used to compute the input values of the model. In Simulation B, aiming to represent the inter-individual variability, the model was run in the standard way and once the prediction was generated, the random values were added to the APD_{90} results using the sd computed by the dispersion model. In either case, the procedure is equivalent

to drawing the values from a normal distribution with the centre located in the original value and a standard deviation similar to the one obtained in the characterisation step. To analyse the combined effects of both types of variability, in Simulation C both approaches were merged; prior to the application of the model, the pIC_{50} values were modified with the random values as in Simulation A and, after generating each APD_{90} prediction, the output values were modified by adding the random values as in Simulation B.

In all three cases, the simulations generate output distributions of slightly different APD_{90} values. The centre of these distributions (median or 50^{th} percentile) was used as the point prediction, while the value range between the 10^{th} and the 90^{th} percentile was used as an interval representing the prediction variability, which can be interpreted as the 80% confidence interval.

An example case study using CiPA compounds

To evaluate the practical application of our methodology, we applied it on a set of 12 CiPA compounds. These compounds, officially selected as the CiPA training and calibration set, were chosen in this study because they belong to three risk classes (low, medium, high) and are well-characterised in terms of their arrhythmogenic mode of action. Moreover, these are real drugs, each of which inhibits the selected ion currents I_{Kr}, I_{NaL}, and I_{CaL} with a different potency at different therapeutic concentrations, resulting in a different combination of model input values. An overview of some important properties of the selected drugs extracted from (Colatsky et al., 2016a; Z. Li, Ridder, et al., 2019b; Llopis-Lorente et al., 2020a) is provided in **Table 1**.

Table 1: Compounds belonging to the CiPA training and calibration set and their main characteristics including the EFTPC in nM, IC_{50} values in nM, the h and the TdP and proarrhythmia risk class.

Name	EFTPC (nM)	I_{Kr}		I_{NaL}		I_{CaL}		Risk class
		IC_{50} (nM)	h	IC_{50} (nM)	h	IC_{50} (nM)	h	

Bepridil	33	144	1	339	1.9	638000	4.6	high
Dofetilide	2	75	1	837000	4.6	2300000	5.4	high
Quinidine	3237	971	1	2360	0.91	5100000	4.7	high
Sotalol	14690	290000	1	134000000	5.9	58000000	5.5	high
Chlorpromazine	38	650	1	673	1.8	6350	2	intermediate
Cisapride	2.6	72	1	421	2.2	4050000	5.6	intermediate
Ondansetron	139	1200	1	6870	1.2	9310000	0.2	intermediate
Terfenadine	4	129	1	98.3	1.1	1220000	5.2	intermediate
Diltiazem	122	7900	1	3040	1.1	31600	1.2	low
Mexiletine	4129	53000	1	4690	0.99	164000	0.96	low
Ranolazine	1948.2	8300	1	5950	0.99	6540000	3.8	low
Verapamil	81	460	1	982	1.2	11200	0.8	low

To obtain biomarker predictions that correspond with the arrhythmogenic potential of the drugs in clinical practice, the *IV*s were calculated using experimental IC_{50} values for I_{Kr}, I_{NaL}, and I_{CaL} channels and the EFTPC. As the starting point, a single APD_{90} biomarker prediction was generated using our default model for each of the 12 compounds. Then, experimental variability and inter-individual variability were characterised for these compounds and propagated through the model using the three different simulation types described above (**Figure 5).** For each drug, this procedure yielded a single biomarker prediction and an interval interpretable as an 80% confidence interval. These results were analysed in detail and critically discussed to evaluate the advantages of assessing the impact of input variability on the uncertainty in the output of the model, which contrasts with relying on single model predictions.

Software

The electrophysiological simulations and the generation of the APD_{90} array were carried out using MATLAB version R2021b. These results are available

online on the public repository of the Universitat Politècnica de València (https://riunet.upv.es/handle/10251/191820). The simulations were carried out using scripts written in Python 3.8. Machine learning models were built and evaluated using Scikit-learn version 0.24.2 (Pedregosa, F., Varoquaux, G. and Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, 2011), NumPy version 1.19.5 (Harris et al., 2020), Pandas version 1.1.5 (McKinney, W., 2010), Statsmodels version 0.12.2 (Seabold & Perktold, 2010). Graphics were generated with Matplotlib version 3.3.4 (Hunter, 2007). The source code of the python scripts, models and methods described here are freely accessible at GitHub (https://github.com/phi-grib/Cardiotox_uncertainty) and usable under GNU GPL v3 open source license.

Results

Overview

The main aim of this work was to develop methods for the assessment of uncertainty, mainly of aleatory type, in prediction results provided by the previously described *in silico* multi-level proarrhythmia model (**Figure 1)**. This model predicts the proarrhythmia biomarker APD_{90} of a certain compound from the experimentally measured or predicted inhibition potency of three ion currents (I_{Kr}, I_{NaL}, I_{CaL}) for a given drug concentration and channel-specific Hill coefficient.

The protocol for uncertainty assessment and quantification involved three steps:

1. Identification of the main sources of variability and uncertainty in cardiac safety models
2. Quantitative characterisation of selected types of variability

3. Propagation and quantitative expression of variability in model outputs

Independently of the type or source, we recognised that all uncertainty types identified (point 1) are interconnected and to some extent affect each other and the output of the model. Nevertheless, for this work we attempted to group them based on their association with the model inputs. Later, we characterised and quantified the individual and the combined effect of two selected variability types (points 2 & 3) on the predictions generated by our model.

This method was applied to a set of 12 CiPA drugs. The results of this use case were analysed, considering the benefits that such output could provide for drug developers and decision-makers.

Step 1: Identification of the main sources of variability and uncertainty in cardiac safety models

Figure 2 presented in the Methods section provides an overview of the most important sources of aleatory and epistemic uncertainty generally associated with cardiac safety models**.**

The origin of aleatory uncertainty was identified as intrinsic and extrinsic variability, as well as measurement errors. These aleatory elements were used to find associations with the inputs of our model. As a result, we summarised them under the umbrella terms “experimental variability” and “inter-individual variability”, affecting the input IC_{50} values and the parameters predefined in the electrophysiological action potential simulations models, respectively. The experimental variability of the Hill coefficient required to compute the input values of our model was not considered in this work, due to its minor impact (see Methods section for details). Additionally, the drug concentration is also subject to aleatory uncertainty, mainly due to intrinsic and extrinsic heterogeneity among subjects of the same population, leading

to differences in pharmacokinetic responses. Compared to the Hill coefficient, the impact of drug concentration on the numeric outcome of **Equation 2** computing the input values of the proarrhythmia model is larger. But, due to some limitations of this protocol, the impact of variability in drug concentration on the overall uncertainty levels in the prediction of the model was not quantified here.

With regard to epistemic uncertainty, the two main affected model components are the inputs from which the predictions are generated and the methodology underlying the prediction system. Experimental inputs are subject to epistemic uncertainty due to multiple unknown values and approximations introduced during laboratory measurements and in the consequent data processing. Some degree of epistemic uncertainty also accompanies all methodological steps, starting with the selection of models or algorithms, through the definition of their parameters and to the subjective expert judgements informing the model, to simplifications and assumptions accompanying the interpretation of the prediction results.

Step 2: Quantitative characterisation of selected types of variability

Experimental variability

Experimental variability was characterised based on assumptions and results previously published by (Elkins et al., 2013). Here, we assumed that IC_{50} values measured for different cardiac ion currents and different compounds are naturally associated with levels of deviation of similar magnitude as those of control compounds in published literature. In the simulations, this subtype of aleatory uncertainty was introduced by adding to the experimental pIC_{50} values a random value following a normal distribution with mean 0.0 and a sd of 0.5, as explained in the Methods section.

Inter-individual variability

Inter-individual variability was characterised following a multi-step approach based on a population of models. This model population, consisting of a total of 860 models, was generated by introducing variations into the default electrophysiological model used in this work as described in the Methods section. In particular, the parameters for every single model belonging to the population were equalled to those expected from a healthy population of patients. The population of models was then applied to predict APD_{90} values from a set of 125 input value combinations. The resulting 3D array served as training data to build a predictive model that can provides approximate estimates of the variability that can be attributed to the single APD_{90} prediction. This variability is expressed as predicted sd, as explained in the Methods section.

Step 3: Propagation and quantitative expression of variability in model outputs

The variability characterised in Step 2 was propagated through the model by running MC simulations, as shown in Methods in **Figure 5**. The MC simulations conducted in this work incorporate only the experimental variability into the input values (Simulation A), add up inter-individual variability into the APD_{90} predictions (Simulation B) or combine both types of simulations (Simulation C). See the Methods section for details. In all instances, these simulations turned single inputs into a collection of 1000 differently distributed output values. These distributions can be seen as a means to complement single predictions provided by our initial model by an informative value interval. Being a product of each prediction, the centre of such interval corresponds to the centre of the APD_{90} distribution (median or percentile 50^{th}) and ranges from the 10^{th} to the 90^{th} percentile. These intervals can informally be referred

to as the 80% confidence intervals and represent the central range of values which the model would produce 80% of the times.

An example case study using selected CiPA compounds

Value distributions resulting from variability propagation

To assess the practical value of the developed methodology, the above-described steps 1-3 were applied to a collection of 12 compounds with well-defined cardiac electrophysiology and proarrhythmia risk classes defining the severity of clinical effects, as previously characterised and published by the CiPA initiative (Colatsky et al., 2016a). The use of these drugs was further justified in the Methods section.

Application of the methodology on the example of the CiPA compound set yielded a collection of 1000 APD_{90} values for each CiPA drug and the considered simulation type. **Figure 6** shows three sections of violin plots, each representing results from the simulations A-C.

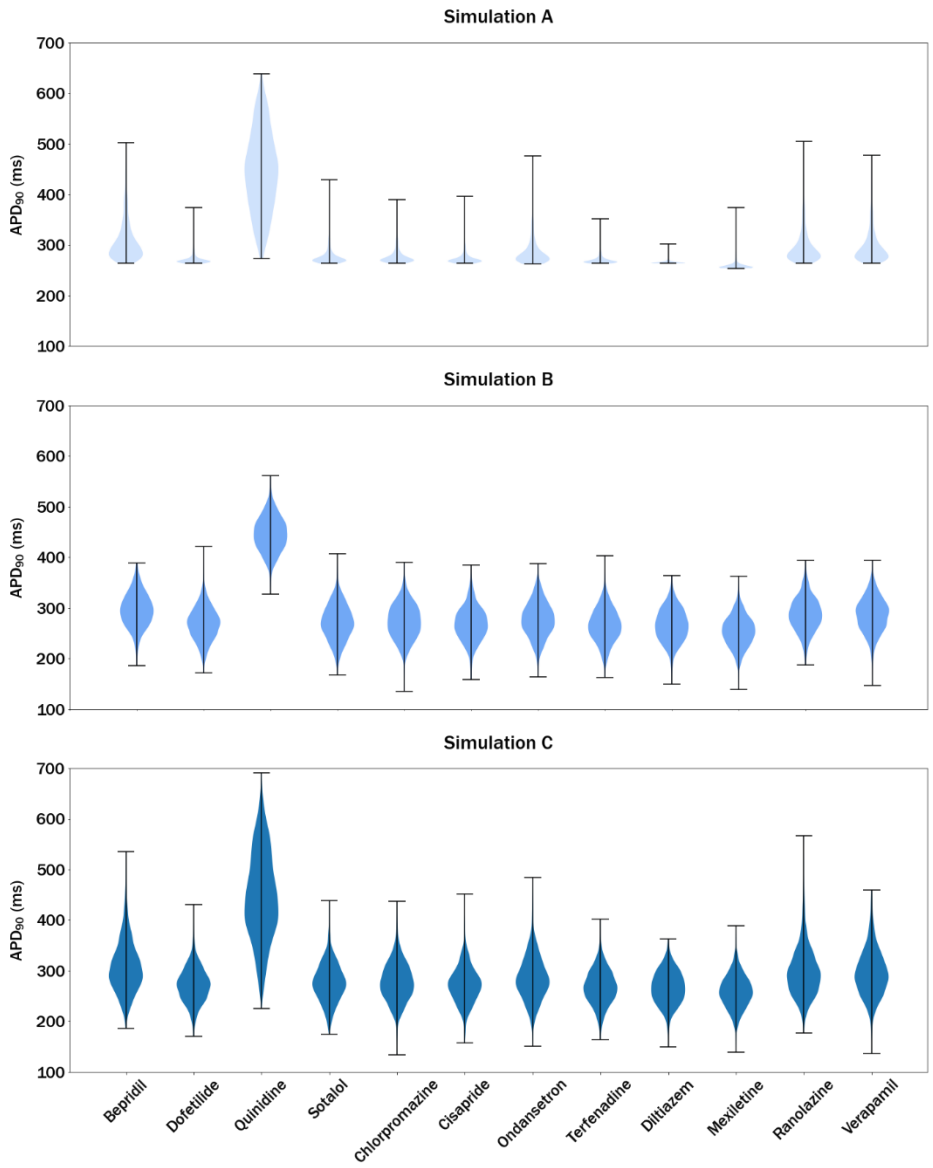


Fig. 6: Violin plots showing distributions of APD_{90} values obtained in different runs of Monte Carlo simulations introducing the following variability types. Simulation A: Experimental variability (Δ-pIC_{50}); Simulation B: Inter-individual variability (Δ-Parameters); Simulation C: Combination of experimental and inter-individual variability.

When comparing the distributions presented in **Figure 6**, obtained by propagating experimental variability (Simulation A) with those where inter-individual variability was considered (Simulations B and C), there are

remarkable differences in the width and skewness. As described in the Methods section, in Simulation A random numbers were added to the pIC_{50} values used to generate the model *IV*s. Hence, the shape and width of the output distributions do not depend directly on the assumptions made to characterise this variability type.

Conversely, the dispersion and the form of the output distributions essentially depend on how sensitive the output values are to small *IV*s changes in a certain region of the training series space. To understand this concept, the model describing the non-linear relationship between the APD_{90}s and the *IVs* can be visualised as a hyperplane. In some regions, this hyperplane is rather flat and therefore small changes on the *IV*s produce rather similar APD_{90} predictions. In other regions, this hyperplane is steeper wherefore small *IV* changes (e.g.: due to a pIC_{50} increase for a highly relevant channel) produce significant APD_{90} variations. For the analysed drugs, most of the distributions generated in Simulation A are right skewed, with the maximum value far from the distribution centre. This can be explained by the non-linear relationship between the *IV*s and the APD_{90}s: even if the *IV*s used in this simulation follow a normal distribution, the output values will not. Therefore, the propagation of experimental variability resulted in a condensation of APD_{90} predictions in a narrow area of around 275 ms and a great right skew of the distribution for the majority of the drugs included in this analysis. In the case of Bepridil, Ranolazine, and Verapamil introducing variability into the pIC_{50} values yielded *IV*s that fell within a sloped region of the prediction function, resulting in wider output distributions and minor right skew. The *IV*s computed for Quinidine, however, were spread differently producing a wide distribution of APD_{90} values with no notable skew.

As opposed to Simulation A, the dispersion and the form of distributions generated in Simulation B, shown in **Figure 6,** are a consequence of the

assumptions made about the inter-individual variability. Since they were generated by adding normally distributed random numbers to the output values, all APD_{90} distributions shown in Simulation B show a normal shape and exhibit no visible differences concerning the width. The minimal discrepancies in the width of the distributions can be justified with similarly minimal spread parameters predicted for these drugs as sd by the dispersion model (see Methods section). As the minimum and maximum sd values in the training series of the dispersion model were 26.93 and 55.18 ms, respectively, these values marked the possible prediction range for any kind of input combination. But since the *IV* combinations of the CiPA drugs did not reach these range limits, the predicted sd values to be considered as measures of the spread of each of these compounds varied between 31.64 and 37.21 ms. As this difference is quite a small relative to the predicted APD_{90} values, the observable differences between the width of the simulated distributions are minimal.

When combining both types of variability in one simulation run, we obtained the distributions shown in Figure 6C. In general, they are rather similar to the ones obtained in Simulation B, but with a slightly larger dispersion and a little skew. Importantly, the effect of considering both kinds of variability simultaneously is not additive, and the effect depends on the drug studied. For example, these effects were particularly noticeable for Bepridil, Quinidine, Ondansetron, Ranolazine, and Verapamil.

When comparing all three approaches, an additional difference between the plots is the sudden cut-off observed for the results of simulation A, where only experimental variability was considered. This cut-off is absent in distributions resulting from Simulations B and C. This difference can be explained by the limited range of *IV*s used in the model describing their associations with the pre-computed APD_{90}s (see Methods section). This means that any variation of

the *IV*s resulting in a decrease below the minimum value considered in the model (-3.0) generally does not result in any change of the output. As a consequence, many of the 1000 *IV*s generated during the simulation were simply converted into the cut-off values and produced exactly the same APD_{90} output. For many drugs, this effect was observed for the I_{CaL} channel, the inhibition of which usually requires the drug to be administered at higher concentrations. In comparison, the inhibition of the I_{Kr} channel at the EFTPCs of the CiPA drugs is more common, due to which the *IV*s computed for hERG channel had the greatest impact on the predicted APD_{90}. Conversely, the propagation of inter-individual variability in Simulation B added random numbers to the output values and was therefore not affected by these *IV* cut-offs.

In other words, in the case of Simulation A, after random values were added to the model inputs, these values were further processed by the model. In Simulation B, however, just one single value was predicted, and the distribution of values was simulated from the expected distribution parameters.

Value intervals as a quantitative expression of uncertainty in the output

The distributions of the predicted APD_{90} values were used to obtain intervals between the 10^{th} and 90^{th} percentiles for the 12 CiPA compounds, yielding the results shown in **Figure 7.**

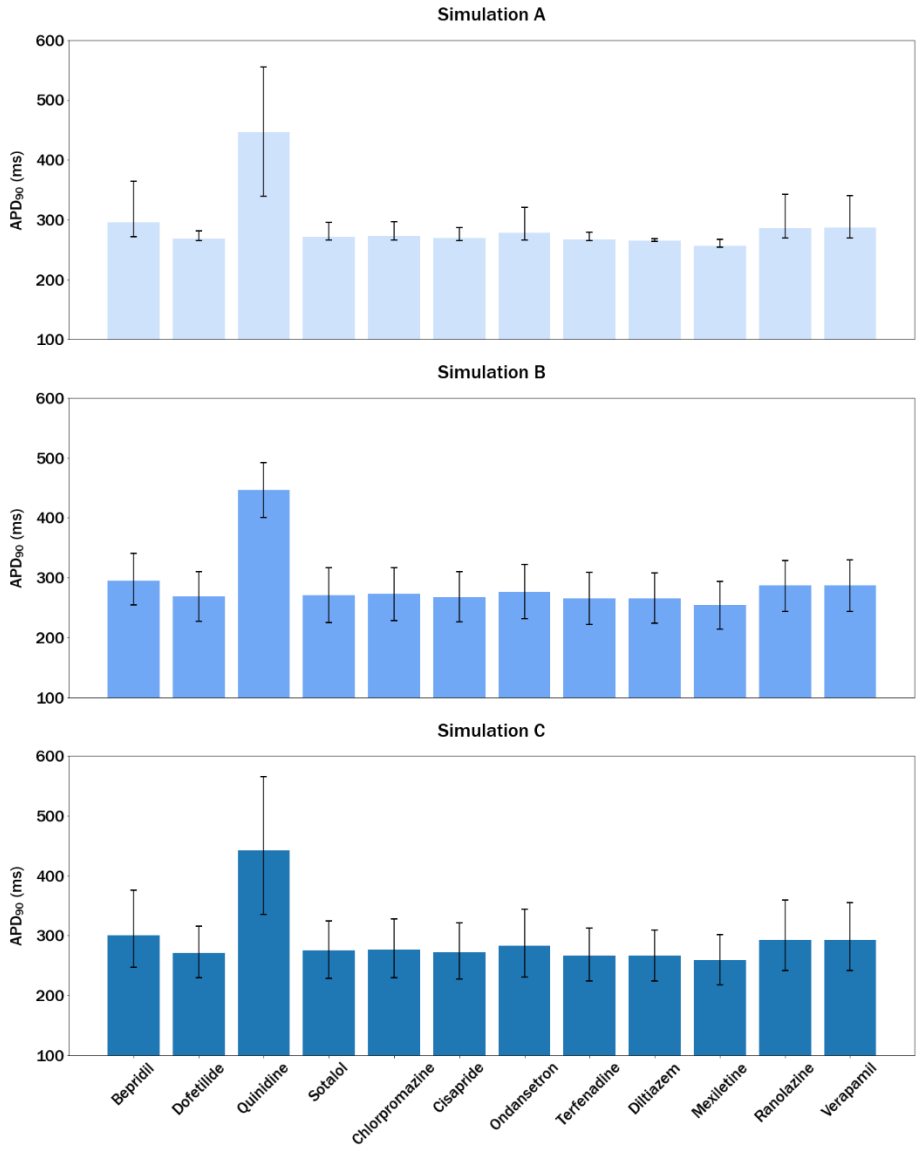


Fig. 7: Bar plots showing the median of the APD_{90} predictions obtained for the 12 CIPA compounds, using three simulation types. Simulation A: Experimental variability (Δ-pIC_{50}); Simulation B: Inter-individual variability (Δ-Parameters); Simulation C: Combination of experimental and inter-individual variability. The intervals represent the 10^{th} and 90^{th} percentiles obtained from the distributions shown in **Figure 6**.

The bar plots in **Figure 7** show no remarkable differences in the APD_{90} predictions generated in three different simulations conducted for the same drug. This observation allows concluding that the actual prediction, computed

as the median value of the APD_{90} distributions produced in simulation A-C, is barely affected by the simulation type and the biomarker prediction can be expected to remain unchanged. On the contrary, important differences can be observed in the widths of the intervals obtained by the different simulations. For Simulation A, these intervals vary between 4.3 and 216.6 ms, with a maximum difference exceeding 200 ms. In contrast, the intervals obtained for Simulation B are relatively similar for all tested compounds and range from 79.3 and 92.9 ms, approximately, thereby showing a maximum difference between two compounds of 13 ms. An overall increase in the intervals' width is noticeable when combining both types of variability. But combining two sources of variability does not lead to additive results, meaning that the combined result is not the sum of the two sources. Considering that the predicted numeric values could be potentially used to assign compounds into different risk classes (of TdP or arrythmia), it is possible that the interval ranges cross the boundaries of different classes, making then difficult to assign the compound to one of them.

In order to illustrate this situation we have shown in **Figure 8** the prediction intervals for Quinidine, Ondansetron, and Mexiletine belonging to the high, intermediate, and low-risk class of TdP, respectively as defined by the CiPA initiative (Colatsky et al., 2016b).

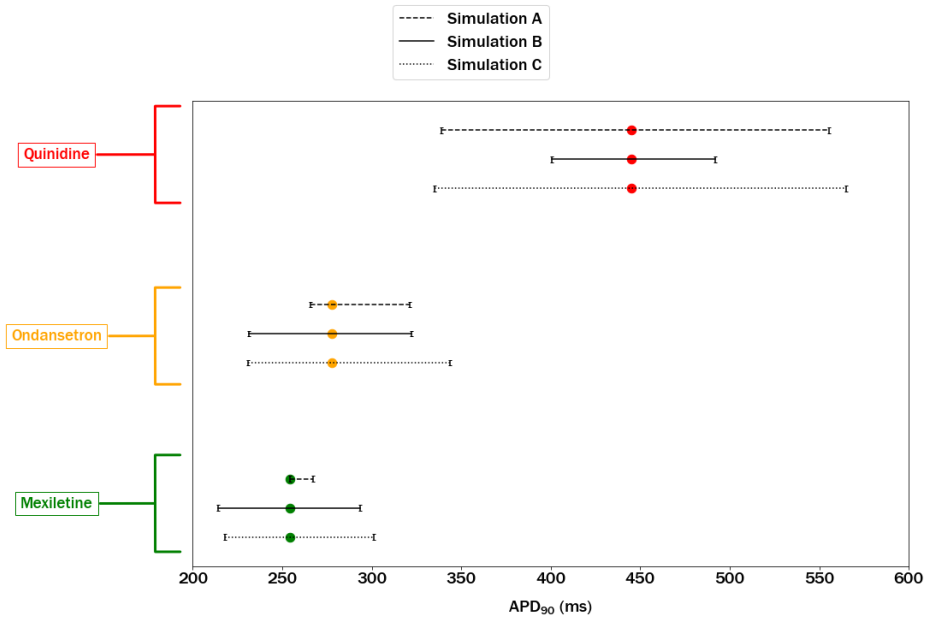


Fig. 8: Predicted APD_{90} values and their corresponding 80% intervals for three selected CiPA compounds assigned to the following arrhythmogenic risk classes: Quinidine as a high-risk drug (red); Ondansetron as an intermediate-risk drug (orange); Mexiletine as a low-risk drug (green). Intervals shown here were obtained in MC Simulations A-C as described in **Figure 4.**

It can be seen that the intervals computed for high-risk and low-risk drugs using any of the presented approaches do not overlap and would allow a clear class assignment. On the contrary, the APD_{90} interval computed for the intermediate-risk compound overlaps the interval of the low risk class using all three simulation scenarios as well as the high-risk class when the most conservative scenario is used. In general, the use of APD_{90} predictions intervals, compared with appropriately selected critical values, would allow for a more conservative classification approach, which incorporates into the prediction both the effects of the experimental and inter-individual variability.

Discussion

Obtaining a reliable risk evaluation for new drug candidates is one of the primary responsibilities of safety pharmacology. Regarding arrhythmogenic risk, the CiPA paradigm provided a standardised way for performing *in vitro*/*in*

silico-based cardiac safety assessment using proarrhythmia models (Hwang et al., 2020b). Despite the availability of a wide range of cardiac safety models stemming from the CiPA work, uncertainty analysis has been one of the last missing pieces to be addressed within this paradigm. Is in that context that this work proposes a protocol for the assessment of uncertainty and variability applicable to multi-level *in silico* proarrhythmia models.

A critical view on the developed methodology

Experimental variability

The central hypothesis behind this work is that there is a "true" pIC_{50} value when one specific ion channel is exposed to a certain concentration of a drug in one specific moment in time. However, the notion of a "true" pIC_{50} is relatively idealistic and therefore does not correspond to what can be expected in practical situations. This is because in the proposed "Uncertainty assessment protocol", the arrhythmogenic potential of drugs is assessed using a specific computational model and a combination of input values which are affected by multiple aleatory factors contributing to the overall levels of experimental variability. Hence, the consideration of experimental variability in cardiac safety model inputs is a step toward increased credibility of the predictions obtained from such models.

In this work, we assumed the same spread measure and the normality of the distributions describing the variability in the inhibition of each considered channel by each analysed drug. Even though the introduced assumptions were rather simple they allowed to test the effect of this variability in the final prediction, at a proof of concept level. In practice, since each pharmaceutical company has individual methods to perform the inhibition tests the standard deviation considered could be adjusted to match the characteristics of the analytical platform, as well as the structure and properties of the tested compounds. Importantly, in this study, we considered the overall variability

arising during the experiments, thereby combining the experimental errors with the biological properties of the samples. In the study performed by (Lei et al., 2020), the authors demonstrated that the extent to which the artefacts in patch-clamp experiments affect the overall levels of experimental variability is much greater than the cell-cell or between-cell differences. Indeed, adding this additional layer of detail to separate experimental errors from intrinsic/extrinsic variability would contribute to a better understanding of the toxicodynamic effects of drugs in the context of cardiac safety assessment.

Inter-individual variability

As for experimental variability, the consideration of inter-individual variability in cardiac model inputs can be seen as a step in the direction of realistic cardiac safety assessment. As described by (Wisniowska et al., 2017), “Humans vary, so cardiac models should account for that too…”. The importance of considering inter-individual differences with regard to drug effects is particularly important if it comes to the protection of individuals who are more prone to develop cardiac arrhythmias or TdP. The use of a population of models to account for such differences allows to obtain different AP responses under the same pharmacological intervention. As compared to classical simulation methods based on an averaged model producing unique values, another advantage of populational approaches is that they provide novel insights into physiological and pathophysiological variabilities (Ni et al., 2018). In addition, this approach has shown that TdP-risk assessment improves when taking into account the electrophysiological variability between cells (Llopis-Lorente et al., 2022a), therefore, increasing evidence points to the crucial role of variability in cardiac electrophysiological function.

Important to consider, however, are the characteristics of the population of interest. In this work, the electrophysiological model parameters, as well as the pre-processing of the simulated data, were based on criteria reflecting the

attributes of a healthy population. Hence, to predict outcomes for a population with any type of underlying conditions, the first calibration step of the population of models would need to be modified accordingly to account for specific characteristics of this population.

It is worth noting that the described approach for representing inter-individual variability was based on the assumption that variability equally impacts all the 15 channel conductances and that this variability is independent for each parameter of the electrophysiological model. These assumptions were based on a series of results presented in the available literature on this topic. Nevertheless, further modifications of the proposed methodology allowing to assign unequal measures representing the variability in the conductances of different ion channels and to consider possible dependency between these measures could add additional value.

In the context of this work, however, establishing identifiability of the true ion channel conductances values was not the aim. For interested readers, different strategies for the identifiability of the parameters of the AP model are presented in the review by Whittaker and colleagues (Whittaker et al., 2020).

Combination of variability

When combining experimental and inter-individual variability to produce a reasonable representation of proarrhythmia predictions, the emphasis should lie on appropriate interpretation of such results. From the theoretical perspective, the consideration of experimental variability is not necessary in clinical settings. Therefore, results obtained by combining these two variability sources do not intend to represent the variability in biomarker response that would be observed in a healthy human population. Nevertheless, when using computational proarrhythmia models which integrate some experimental values to produce estimates of human responses, the consideration of experimental variability is essential. In the latter case, the produced range of

values aims to represent the variability in predictions, given the limited ability to define the "true" pIC_{50} values together with inter-individual differences among subjects of a population.

As shown in this work, combining variability, or other types of types or uncertainty, does not mean that the effects of each source on the final prediction will sum up. Nevertheless, as the current methodology for combining different variability types affects the shape of the obtained distributions, the methodology could be adjusted to account for this dependency. To do so, an additional analysis of the dependencies between each of the input sources, as well as of their associated variabilities, could be included in future work.

Representation of results

Another important question is whether representing simulation results as a biomarker prediction with a corresponding 80% confidence interval has an advantage over standard methods yielding point estimates, only. As concluded by (Sahlin, 2015))"… a confidence interval is just an interval. It does not provide enough information to calculate an expected value or conservative value, which is important in rational decision making". However, a confidence interval provided together with the expected value is very useful for communicating uncertain results in a simple way. Such intervals allow for the inspection of values that would be produced in experiments or for individuals that do not represent the exact centre of the distribution from which they were drawn. Since variability is an innate element of all-natural and investigational processes, assuming that a fixed prediction is the exact centre of a specific distribution is rather ingenious. But, when intervals are provided together with single values to interpret the predictions, the scientific conclusion derived based on them automatically is considerate of the variation among biological samples or the physiology of patients. Another factor impacting the credibility

of confidence intervals is an adequate identification of all the sources of uncertainty and a correct characterisation and propagation of those, that indeed affect the prediction outcome. To know which sources should be prioritised for the UQ exercise, a prior sensitivity analysis is recommendable (Eck et al., 2016).

Suggestions for future work

Consideration of epistemic uncertainty

In this publication, although different sources of aleatory and epistemic uncertainty were identified, the described methods were mainly focused on the characterisation and propagation of two sources of variability. The protocol integrated only principles of the frequentist approach to probability. Indeed, when quantifying only variability reflecting the natural variability and randomness, the selection of normal distribution with standard deviation as the representation of sample spread was a reasonable decision. This is because real-valued random variables whose distributions are undefined are usually represented using normal distributions. As stated in the Central Limit Theorem, under some conditions, when a large series of random numbers are sampled from any population with a defined mean and sd, the initial distribution converges to a normal distribution as the number of samples increases (Devore & Berk, 2012).

However, epistemic uncertainty, also identified in this work, should not be expressed nor modelled using frequentist methods. Instead, the correct way to assess epistemic problems involves the application of the subjective probability theory, the most common application of which is the Bayesian theorem (van de Schoot et al., 2021). This involves starting with an initial belief, known as the prior probability, and updating it when new information becomes available yielding the posterior distribution. Nevertheless, applying Bayesian statistics to estimate the impact of purely epistemic factors (shown

in yellow in **Figure 2**) on the APD_{90} predictions would require major modifications of the developed uncertainty quantification protocol. But particularly important for this work and more feasible to implement would be the consideration of epistemic uncertainty about the aleatory uncertainties summarised as variability types. This would lead to a quantitative expression of the level of unknown in the metrics defined to characterise specific variability types, for instance, the constant sd value of 0.5 that was assumed to characterise experimental variability. Degrees of belief about the true parameters for this quantity could be derived using either objective measurements or subjective expert judgements. To propagate the uncertainty about variability in the quantity of interest, sampling of the resulting prior distributions could be integrated as part of a 2 dimensional Monte Carlo simulation. A result of such a simulation would not be a single distribution of values, but multiple distributions representing the uncertainty about variability (Benford, Halldorsson, Jeger, Knutsen, More, Naegeli, Noteborn, Ockleford, Ricci, Rychen, Schlatter, Silano, Solecki, Turck, Younes, Craig, Hart, Von Goetz, Koutsoumanis, Mortensen, Ossendorp, Germini, et al., 2018). Coming back to the previous example, the uncertainty about the level of experimental variability would be expressed as several distributions, each of which with a different centre (median or mean) and measure of spread.

Computational model inputs

There is a high interest in transforming the mixed-platform preclinical cardiac safety assessment of novel pharmaceutical products into purely *in silico* based methods without the need for extensive experimental testing. Therefore, the structure of our multi-level cardiotoxicity models allows both, experimental as well as predicted inputs. Since computational models, such as the PBPK or QSAR models, are built using experimental data, experimental variability, which was extensively described in this work, is also retained in the training

series used for building these models. But, when the plasma drug concentration or the channel specific IC_{50} are generated computationally, the level of epistemic uncertainty increases due to further limitations in the training data coverage or a high level of subjectivity impacting the parametrisation of the respective source models that predicts them.

For instance, if the intention is to predict proarrhythmic properties of a compound available in a public domain, such as the ChEMBL database (Gaulton et al., 2012), multiples datapoints would be available for the same compound, each of which is produced in a separate experiment following a specific protocol. These data points would first need to be extensively filtered to select the experimental parameters of interest and aggregated using statistical measures such as a mean or the median. This process, together with multiple unconsidered originating from differences in laboratory conditions, experimental design, and other factors, would contribute to the level of epistemic uncertainty. Despite of these factors, the predictive performance of purely computational proarrhythmia prediction systems highly depends on the selected biomarker. As shown by (Beattie et al., 2013b), the use of QSAR-derived data to simulate QT-interval shortening may yield nearly as good predictions as those produced using experimental data inputs. Conversely, utilising QSAR data to predict QT-interval prolongation significantly worsens the predictive performance. These two examples show the importance of comprehensive definition of the endpoint being modelled which should always precede the process of uncertainty analysis to ensure a correct determination of model limitations, variability sources and epistemic factors.

QSAR

The most widely accepted standard method for the quantification of reliability and uncertainty associated with QSAR model predictions are methods based on the concept of applicability domain (AD) (Sahlin et al., 2014). Predictions

generated for compounds having structurally or physio-chemically similar counterparts in the training set are generally considered reliable. Standard AD methods can be complemented by placing the model within a framework that can estimate the uncertainty levels in every single prediction. An example is the conformal prediction framework which guarantees the maximum allowed frequency of errors which will be committed by the conformal predictor (Alvarsson et al., 2021; Norinder et al., 2014; Svensson et al., 2018). Uncertainty resulting from lack of knowledge (e.g.: insufficient training data or anomalous samples in test data), that is predominant in model predictions is most commonly addressed by applying Bayesian inference, shortly introduced above (Sahlin, 2015).

PBPK

The arrhythmogenic potential of drug candidates is typically assessed at early stages of drug development when the compound can still be removed from the development pipeline without much economic harm. At these stages, the therapeutic concentration and other PK parameters required to compute the EFTPC are still unknown but the use of currently described methodologies to estimate point-of-departure concentrations is an interesting approach. These could be compared with experimental results produced at preclinical stages using physiologically based pharmacokinetic (PBPK) modelling to obtain plasma concentrations from the administered doses. PBPK models are mathematical algorithms based on ordinal differential equations (ODEs) describing physiological processes involved in the absorption, distribution, metabolism, and excretion of the drug (Piñero et al., 2018). Variability and uncertainty quantification in PBPK models is often initiated by a parametric sensitivity analysis to identify the PK parameters that are most susceptible. Since PK parameters are subject to inter-individual differences and PK simulations are often liable to lack of full information about the constants and

parameters in the ODEs, the UQ methods require combining the frequentist and conditional probabilistic approaches (Kuepfer et al., 2016). Consideration of uncertainty in PBPK simulations would allow to explore a range of clinically relevant drug concentrations, especially at the site of the pharmacological or toxicological action of the drug (e.g.: drug binding site at the ion channel protein in the membrane of ventricular myocytes) (Z. Li, Garnett, et al., 2019).

Conclusions

In this study, we developed and tested methods for the quantification of the impact of selected variability types on the uncertainty of APD_{90} predictions generated by an *in silico* multi-level proarrhythmia model. The aim was first to explore the effects of different types of variability, separately and in combination, by quantitatively characterising and propagating them throughout our complex model, and second to replace point predictions with value ranges that can be computed for predefined credibility levels (e.g.: 80%) and interpreted as confidence intervals.

The propagation of "experimental variability", associated with the input IC_{50} values, yielded distributions whose characteristics were defined by the location of the *IV*s within the hyperplane-like structure of model training data. This contrasts with the distributions resulting from the propagation of "inter-individual variability", linked with the parameters specified in the AP simulation models, whose shape and width were a direct consequence of the methodological assumptions and the predicted spread parameters, respectively. After a simultaneous propagation of both types, the distributions showed a combined effect of both, the non-linear relationship between the *IV*s and APD_{90} and the assumption of normality applied to model outputs. Importantly, combining two sources of variability did not lead to additive results, meaning that the combined result is not their sum.

Further, we showed how such distributions can be used to compute the proarrhythmia biomarker predictions together with value intervals of certain credibility. One of the main conclusions arising from this analysis was the that the actual biomarker prediction remains nearly unchanged when the simulations are performed, as compared to the initial method without UQ. Although we do not claim the undoubtful accuracy of these results, we consider that such representation of the predictions has excellent advantages over single-point estimates. These mainly include the possibility to inspect values that would be produced in experiments or for individuals that do not represent the exact centre of the distribution from which they were drawn. Hence, it allows to protect individuals who are more prone to develop cardiac arrhythmias or TdP, since interval ranges may cross the boundaries of different risk classes. Moreover, they provide a more realistic view on predictions in the context of drug candidate prioritisation and validation of clinical results, since the presence of uncertainty resulting from variability is usually neglected at these assessment stages.

Funding

The authors received funding from the eTRANSAFE project, Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777365, supported from European Union's Horizon 2020 and the EFPIA. Authors declare that this work reflects only the author's view, and that IMI-JU is not responsible for any use that may be made of the information it contains. Also, this project received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 964537 (RISK-HUNT3R), which is part of the ASPIS cluster. We also received funding from the SimCardioTest supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 101016496. J.L.L. is being funded by the Ministerio de Ciencia, Innovacion y Universidades

for the “Formacion de Profesorado Universitario” (Grant Reference: FPU18/01659). The work was also partially support by the Dirección General de Política Científica de la Generalitat Valenciana (PROMETEO/ 2020/043).

Author contribution

The ideas and methodologies described in this article evolved based on previous works of Obiol-Pardo *et al.* (2011) and Romero *et al.* (2018), directed by Manuel Pastor. Both first authors of this publication (Karolina Kopańska and Pablo Rodríguez-Belenguer) recently contributed to the development of the modelling framework, on top of which the uncertainty assessment protocol was developed, as described in Rodríguez-Belenguer, Kopańska *et al.* (2023). In this work, literature research, methods, and code development were equally divided between both authors, Karolina Kopańska and Pablo Rodriguez-Belenguer. Data collection and the development of methods for the conduct of electrophysiological biomarker simulations and the generation of electrophysiological model populations was done by Jordi Llopis-Lorente, Beatriz Trenor, Javier Saiz. The first draft of the manuscript was written by Karolina Kopańska and Pablo Rodríguez-Belenguer, under technical and academic supervision of Manuel Pastor. All authors were involved in the revision and approval of the last version of this manuscript.

Conflicts of interest

The authors declare no conflict of interest.

Ethical standards

The manuscript does not contain clinical studies nor patient data.

References

Alvarsson, Jonathan, Staffan Arvidsson McShane, Ulf Norinder, and Ola Spjuth. 2021. “Predicting With Confidence: Using Conformal Prediction

in Drug Discovery." *Journal of Pharmaceutical Sciences* 110(1): 42–49.

Beattie, Kylie A. et al. 2013. "Evaluation of an in Silico Cardiac Safety Assay: Using Ion Channel Screening Data to Predict QT Interval Changes in the Rabbit Ventricular Wedge." *Journal of pharmacological and toxicological methods* 68(1): 88–96. https://pubmed.ncbi.nlm.nih.gov/23624022/.

Benford, Diane, Thorhallur Halldorsson, Michael John Jeger, Helle Katrine Knutsen, Simon More, Hanspeter Naegeli, Hubert Noteborn, Colin Ockleford, Antonia Ricci, Guido Rychen, Josef R. Schlatter, Vittorio Silano, Roland Solecki, Dominique Turck, Maged Younes, Peter Craig, Andrew Hart, Natalie Von Goetz, Kostas Koutsoumanis, Alicja Mortensen, Bernadette Ossendorp, Laura Martino, et al. 2018. "Guidance on Uncertainty Analysis in Scientific Assessments." *EFSA Journal* 16(1).

Benford, Diane, Thorhallur Halldorsson, Michael John Jeger, Helle Katrine Knutsen, Simon More, Hanspeter Naegeli, Hubert Noteborn, Colin Ockleford, Antonia Ricci, Guido Rychen, Josef R. Schlatter, Vittorio Silano, Roland Solecki, Dominique Turck, Maged Younes, Peter Craig, Andrew Hart, Natalie Von Goetz, Kostas Koutsoumanis, Alicja Mortensen, Bernadette Ossendorp, Andrea Germini, et al. 2018. "The Principles and Methods behind EFSA's Guidance on Uncertainty Analysis in Scientific Assessment." *EFSA Journal* 16(1): e05122. https://onlinelibrary.wiley.com/doi/full/10.2903/j.efsa.2018.5122.

Britton, Oliver J. et al. 2013. "Experimentally Calibrated Population of Models Predicts and Explains Intersubject Variability in Cardiac Cellular Electrophysiology." *Proceedings of the National Academy of Sciences* 110(23): E2098–2105. https://www.pnas.org/content/110/23/E2098.

Britton, Oliver J et al. 2017. “The Electrogenic Na(+)/K(+) Pump Is a Key Determinant of Repolarization Abnormality Susceptibility in Human Ventricular Cardiomyocytes: A Population-Based Simulation Study.” *Frontiers in physiology* 8: 278.

Chang, Kelly C. et al. 2017. “Uncertainty Quantification Reveals the Importance of Data Variability and Experimental Design Considerations for in Silico Proarrhythmia Risk Assessment.” *Frontiers in Physiology* 8(NOV).

Clayton, Richard H. et al. 2020. “An Audit of Uncertainty in Multi-Scale Cardiac Electrophysiology Models.” *Philosophical Transactions of the Royal Society A* 378(2173). https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2019.0335.

Colatsky, Thomas et al. 2016. “The Comprehensive in Vitro Proarrhythmia Assay (CiPA) Initiative — Update on Progress.” *Journal of Pharmacological and Toxicological Methods* 81: 15–20.

Coppini, Raffaele et al. 2013. “Late Sodium Current Inhibition Reverses Electromechanical Dysfunction in Human Hypertrophic Cardiomyopathy.” *Circulation* 127(5): 575–84.

Devore, Jay L., and Kenneth N. Berk. 2012. *Modern Mathematical Statistics with Applications*. 2nd ed. Springer.

Eck, Vinzenz Gregor et al. 2016. “A Guide to Uncertainty Quantification and Sensitivity Analysis for Cardiovascular Applications.” *International Journal for Numerical Methods in Biomedical Engineering* 32(8): e02755. https://onlinelibrary.wiley.com/doi/full/10.1002/cnm.2755.

Elkins, Ryan C. et al. 2013. “Variability in High-Throughput Ion-Channel Screening Data and Consequences for Cardiac Safety Assessment.”

Journal of Pharmacological and Toxicological Methods 68(1): 112–22. https://pubmed.ncbi.nlm.nih.gov/23651875/.

European Chemicals Agency. 2012. "Guidance on Information Requirements and Chemical Safety Assessment Guidance for the Implementation of REACH. Chapter R.19: Uncertainty Analysis." http://echa.europa.eu.

Fermini, Bernard et al. 2016. "A New Perspective in the Field of Cardiac Safety Testing through the Comprehensive in Vitro Proarrhythmia Assay Paradigm." *Journal of Biomolecular Screening* 21(1): 1–11. https://journals.sagepub.com/doi/10.1177/1087057115594589.

Fink, Martin et al. 2008. "Contributions of HERG K+ Current to Repolarization of the Human Ventricular Action Potential." *Progress in biophysics and molecular biology* 96(1–3): 357–76.

Gaulton, Anna et al. 2012. "ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery." *Nucleic acids research* 40(D1): D1100–1107.

Gintant, Gary, Philip T. Sager, and Norman Stockbridge. 2016. "Evolution of Strategies to Improve Preclinical Cardiac Safety Testing." *Nature Reviews Drug Discovery 2016 15:7* 15(7): 457–71. https://www.nature.com/articles/nrd.2015.34.

Gosling, John Paul. 2019. "The Importance of Mathematical Modelling in Chemical Risk Assessment and the Associated Quantification of Uncertainty." *Computational Toxicology* 10: 44–50.

Grandi, Eleonora, Francesco S. Pasqualini, and Donald M. Bers. 2010. "A Novel Computational Model of the Human Ventricular Action Potential and Ca Transient." *Journal of Molecular and Cellular Cardiology* 48(1): 112–21.

Harris, Charles R. et al. 2020. "Array Programming with NumPy." *Nature 2020*

585:7825 585(7825): 357–62. https://www.nature.com/articles/s41586-020-2649-2.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. 2 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hosmer Jr, David W, Stanley Lemeshow, and Rodney X Sturdivant. 2013. 398 *Applied Logistic Regression*. John Wiley \& Sons.

Hu, Zhiyong, Dongping Du, and Yuncheng Du. 2018. "Generalized Polynomial Chaos-Based Uncertainty Quantification and Propagation in Multi-Scale Modeling of Cardiac Electrophysiology." *Computers in Biology and Medicine* 102: 57–74.

Hunter, John D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science and Engineering* 9(3): 90–95.

Hwang, Minki, Chul Hyun Lim, Chae Hun Leem, and Eun Bo Shim. 2020. "In Silico Models for Evaluating Proarrhythmic Risk of Drugs." *APL Bioengineering* 4(2): 021502. https://aip.scitation.org/doi/abs/10.1063/1.5132618 (February 21, 2022).

ICH E14. 2005. "Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs." https://www.ema.europa.eu/en/ich-e14-clinical-evaluation-qtqtc-interval-prolongation-proarrhythmic-potential-non-antiarrhythmic.

ICH S7B. 2005. *Non-Clinical Evaluation of the Potential for Delayed Ventricular Repolarization (QT Interval Prolongation)*. https://www.ema.europa.eu/en/ich-s7b-non-clinical-evaluation-potential-delayed-ventricular-repolarization-qt-interval.

Johnstone, Ross H. et al. 2016. “Uncertainty and Variability in Models of the Cardiac Action Potential: Can We Build Trustworthy Models?” *Journal of Molecular and Cellular Cardiology* 96: 49–62.

Kitagawa, G, and S Sato. 2001. “Sequential Monte Carlo Methods in Practice.” In eds. A Doucet, N de Freitas, and N Gordon. Springer, 178–95.

Kramer, James et al. 2020. “Cross-Site and Cross-Platform Variability of Automated Patch Clamp Assessments of Drug Effects on Human Cardiac Currents in Recombinant Cells.” *Scientific Reports 2020 10:1* 10(1): 1–15.

Kroese, Dirk P., and Reuven Y. Rubinstein. 2012. “Monte Carlo Methods.” *Wiley Interdisciplinary Reviews: Computational Statistics* 4(1): 48–58.

Kuepfer, L. et al. 2016. “Applied Concepts in PBPK Modeling: How to Build a PBPK/PD Model.” *CPT: Pharmacometrics and Systems Pharmacology* 5(10): 516–31.

Lei, Chon Lok et al. 2020. “Accounting for Variability in Ion Current Recordings Using a Mathematical Model of Artefacts in Voltage-Clamp Experiments.” *Philosophical Transactions of the Royal Society A* 378(2173). https://royalsocietypublishing.org/doi/10.1098/rsta.2019.0348.

Li, Zhihua et al. 2017. “Improving the in Silico Assessment of Proarrhythmia Risk by Combining HERG (Human Ether-à-Go-Go-Related Gene) Channel-Drug Binding Kinetics and Multichannel Pharmacology.” *Circulation: Arrhythmia and Electrophysiology* 10(2). https://www.ahajournals.org/doi/abs/10.1161/CIRCEP.116.004628 (November 12, 2021).

Li, Zhihua et al. 2019. “Assessment of an In Silico Mechanistic Model for Proarrhythmia Risk Prediction Under the CiPA Initiative.” *Clinical Pharmacology and Therapeutics* 105(2): 466–75.

Li, Zhihua, Christine Garnett, and David G. Strauss. 2019. “Quantitative Systems Pharmacology Models for a New International Cardiac Safety Regulatory Paradigm: An Overview of the Comprehensive In Vitro Proarrhythmia Assay In Silico Modeling Approach.” *CPT: Pharmacometrics & Systems Pharmacology* 8(6): 371–79. https://onlinelibrary.wiley.com/doi/full/10.1002/psp4.12423.

Llopis-Lorente, Jordi et al. 2020. “In Silico Classifiers for the Assessment of Drug Proarrhythmicity.” *Journal of Chemical Information and Modeling* 60(10): 5172–87. https://pubs.acs.org/doi/abs/10.1021/acs.jcim.0c00201.

Llopis-Lorente, Jordi, Beatriz Trenor, and Javier Saiz. 2022. “Considering Population Variability of Electrophysiological Models Improves the in Silico Assessment of Drug-Induced Torsadogenic Risk.” *Computer Methods and Programs in Biomedicine* 221: 106934.

McKinney, W., & others. 2010. “Data Structures for Statistical Computing in Python.” *Proceedings of the 9th Python in Science Conference* 445: 51–56.

Mirams, Gary R. et al. 2016. “Uncertainty and Variability in Computational and Mathematical Models of Cardiac Physiology.” *Journal of Physiology* 594(23): 6833–47.

Mirams, Gary R., Steven A. Niederer, and Richard H. Clayton. 2020. “The Fickle Heart: Uncertainty Quantification in Cardiac and Cardiovascular Modelling and Simulation.” *Philosophical Transactions of the Royal Society A* 378(2173).

https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0119.

Mirams, Gary R et al. 2014. “Prediction of Thorough QT Study Results Using Action Potential Simulations Based on Ion Channel Screens.” *Journal of Pharmacological and Toxicological Methods* 70(3): 246–54. https://www.sciencedirect.com/science/article/pii/S1056871914002354.

Muszkiewicz, Anna et al. 2016. “Variability in Cardiac Electrophysiology: Using Experimentally-Calibrated Populations of Models to Move beyond the Single Virtual Physiological Human Paradigm.” *Progress in Biophysics and Molecular Biology* 120(1–3): 115–27.

Ni, Haibo, Stefano Morotti, and Eleonora Grandi. 2018. “A Heart for Diversity: Simulating Variability in Cardiac Arrhythmia Research.” *Frontiers in physiology* 9: 958.

Norinder, Ulf, Lars Carlsson, Scott Boyer, and Martin Eklund. 2014. “Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination.” *Journal of Chemical Information and Modeling* 54(6): 1596–1603.

O’Hara, Thomas, László Virág, András Varró, and Yoram Rudy. 2011. “Simulation of the Undiseased Human Cardiac Ventricular Action Potential: Model Formulation and Experimental Validation.” *PLOS Computational Biology* 7(5): e1002061. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002061.

Obiol-Pardo, Cristian et al. 2011. “A Multiscale Simulation System for the Prediction of Drug-Induced Cardiotoxicity.” *Journal of Chemical Information and Modeling* 51(2): 483–92.

Organization, World Health, and International Programme on Chemical Safety. 2018. *Guidance Document on Evaluating and Expressing Uncertainty in Hazard Characterization*. 2nd ed. World Health Organization.

Parikh, Jaimit, Viatcheslav Gurev, and John J. Rice. 2017. "Novel Two-Step Classifier for Torsades de Pointes Risk Stratification from Direct Features." *Frontiers in Pharmacology* 8(NOV): 816.

Park, Jin Sol, Ji Young Jeon, Ji Ho Yang, and Min Gul Kim. 2019. "Introduction to in Silico Model for Proarrhythmic Risk Assessment under the CiPA Initiative." *Translational and Clinical Pharmacology* 27(1): 12. /pmc/articles/PMC6989268/.

Pathmanathan, Pras et al. 2015. "Uncertainty Quantification of Fast Sodium Current Steady-State Inactivation for Multi-Scale Models of Cardiac Electrophysiology." *Progress in biophysics and molecular biology* 117(1): 4–18. https://pubmed.ncbi.nlm.nih.gov/25661325/.

Pathmanathan, Pras, Jonathan M. Cordeiro, and Richard A. Gray. 2019. "Comprehensive Uncertainty Quantification and Sensitivity Analysis for Cardiac Action Potential Models." *Frontiers in Physiology* 10(JUN): 721.

Pedregosa, F., Varoquaux, G. and Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12(85): 2825–30.

Pieske, Burkert et al. 2002. "Rate Dependence of $[Na^+]_i$ and Contractility in Nonfailing and Failing Human Myocardium." *Circulation* 106(4): 447–53. https://www.ahajournals.org/doi/abs/10.1161/01.CIR.0000023042.50192.F4.

Piñero, Janet, Laura I. Furlong, and Ferran Sanz. 2018. “In Silico Models in Drug Development: Where We Are.” *Current Opinion in Pharmacology* 42: 111–21.

Roden, Dan M. 2004. “Drug-Induced Prolongation of the QT Interval.” *The New England journal of medicine* 350(10): 1013–22. https://pubmed.ncbi.nlm.nih.gov/14999113/.

Rodríguez-Belenguer, Pablo et al. 2023. “Application of Machine Learning to Improve the Efficiency of Electrophysiological Simulations Used for the Prediction of Drug-Induced Ventricular Arrhythmia.” *Computer Methods and Programs in Biomedicine*: 107345. https://www.sciencedirect.com/science/article/pii/S0169260723000123.

Romero, Lucia et al. 2018. “In Silico QT and APD Prolongation Assay for Early Screening of Drug-Induced Proarrhythmic Risk.” *Journal of Chemical Information and Modeling* 58(4): 867–78. https://pubs.acs.org/doi/10.1021/acs.jcim.7b00440.

Romero, Lucía, Esther Pueyo, Martin Fink, and Blanca Rodríguez. 2009. “Impact of Ionic Current Variability on Human Ventricular Cellular Electrophysiology.” *Am J Physiol Heart Circ Physiol* 297(4).

Sager, Philip T. et al. 2014. “Rechanneling the Cardiac Proarrhythmia Safety Paradigm: A Meeting Report from the Cardiac Safety Research Consortium.” *American Heart Journal* 167(3): 292–300.

Sahlin, U., N. Jeliazkova, and T. Oberg. 2014. “Applicability Domain Dependent Predictive Uncertainty in QSAR Regressions.” *Molecular Informatics* 33(1): 26–35. https://onlinelibrary.wiley.com/doi/full/10.1002/minf.201200131.

Sahlin, Ullrika. 2015. “Assessment of Uncertainty in Chemical Models by Bayesian Probabilities: Why, When, How?” *Journal of Computer-Aided Molecular Design* 29(7): 583–94. https://link.springer.com/article/10.1007/s10822-014-9822-3.

Sampedro, D.A. 2020. “Theoretical Analysis of Autonomic Nervous System Effects on Cardiac Electrophysiology an Its Relationship with the Arrhythmias Risk.”

Schmidt, Ulrich et al. 1998. “Contribution of Abnormal Sarcoplasmic Reticulum ATPase Activity to Systolic and Diastolic Dysfunction in Human Heart Failure.” *Journal of Molecular and Cellular Cardiology* 30(10): 1929–37. https://www.sciencedirect.com/science/article/pii/S0022282898907489.

van de Schoot, Rens et al. 2021. “Bayesian Statistics and Modelling.” *Nature Reviews Methods Primers* 1(1): 1–26.

Seabold, Skipper, and Josef Perktold. 2010. “Statsmodels: Econometric and Statistical Modeling with Python.” In *9th Python in Science Conference*.

Shamsi, Mohammad Haris, Usman Ali, Eleni Mangina, and James O’Donnell. 2020. “A Framework for Uncertainty Quantification in Building Heat Demand Simulations Using Reduced-Order Grey-Box Energy Models.” *Applied Energy* 275: 115141.

Shikano, Susumu, Thomas Bräuninger, and Michael Stoffel. 2012. “Statistical Analysis of Experimental Data.” In *Experimental Political Science: Principles and Practices*, eds. Bernhard Kittel, Wolfgang J Luhan, and Rebecca B Morton. London: Palgrave Macmillan UK, 163–77. https://doi.org/10.1057/9781137016645_8.

Sobie, Eric A. 2009. “Parameter Sensitivity Analysis in Electrophysiological Models Using Multivariable Regression.” *Biophysical Journal* 96(4): 1264–74.

Svensson, Fredrik et al. 2018. “Conformal Regression for Quantitative Structure-Activity Relationship Modeling - Quantifying Prediction Uncertainty.” *Journal of Chemical Information and Modeling* 58(5): 1132–40.

Volders, Paul G A et al. 2000. “Progress in the Understanding of Cardiac Early Afterdepolarizations and Torsades de Pointes: Time to Revise Current Concepts.” *Cardiovascular Research* 46(3): 376–92. https://doi.org/10.1016/S0008-6363(00)00022-5.

Whittaker, Dominic G et al. 2020. “Calibration of Ionic and Cellular Cardiac Electrophysiology Models.” *WIREs Systems Biology and Medicine* 12(4): e1482. https://doi.org/10.1002/wsbm.1482.

Wisniowska, Barbara, Zofia Tylutki, and Sebastian Polak. 2017. “Humans Vary, so Cardiac Models Should Account for That Too!” *Frontiers in Physiology* 8(SEP): 700.

Yap, Yee Guan, and A. John Camm. 2003. “Drug Induced QT Prolongation and Torsades de Pointes.” *Heart* 89(11): 1363. /pmc/articles/PMC1767957/.

Conclusiones

1. Se revisaron los principales trabajos en los que se emplea la combinación de múltiples MIEs y modelos multinivel como dos aproximaciones diferentes para la predicción de parámetros toxicológicos complejos.
2. El modelado mecanístico, a través de la combinación de múltiples MIEs para la predicción de la colestasis condujo a la construcción de un metamodelo con un poder predictivo superior al modelado directo QSAR, ya que sus resultados fueron independientes del grado de similitud estructural, a diferencia de los modelos directos QSAR.
3. La incorporación de la toxicocinética al metamodelo anterior incrementó de forma sustancial su capacidad predictiva.
4. Los modelos multinivel de arritmia fueron empleados de forma exitosa reduciendo 100 veces los tiempos de obtención de las matrices electrofisiológicas frente a las aproximaciones de referencia.
5. Se desarrolló con éxito una metodología general para identificar, caracterizar, y cuantificar la variabilidad asociada en las predicciones del modelo multinivel, mejorando así la caracterización de su confiabilidad.
6. Se determinó que la combinación de la variabilidad experimental y la interindividual no tiene un efecto sumatorio en la variabilidad asociada sobre las predicciones del modelo multinivel.

Referencias

Ankley, G. T., Bennett, R. S., Erickson, R. J., Hoff, D. J., Hornung, M. W., Johnson, R. D., Mount, D. R., Nichols, J. W., Russom, C. L., Schmieder, P. K., Serrrano, J. A., Tietge, J. E., & Villeneuve, D. L. (2010). Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environmental Toxicology and Chemistry*, *29*(3), 730-741. https://doi.org/10.1002/etc.34

Bartos, D. C., Grandi, E., & Ripplinger, C. M. (2015). Ion channels in the heart. *Comprehensive Physiology*, *5*(3), 1423-1464. https://doi.org/10.1002/cphy.c140069

Bringezu, F., Carlos Gómez-Tamayo, J., & Pastor, M. (2021). Ensemble prediction of mitochondrial toxicity using machine learning technology. *Computational Toxicology*, *20*, 100189. https://doi.org/10.1016/j.comtox.2021.100189

Council, N. R. (2007). *Toxicity Testing in the 21st Century: A Vision and a Strategy*. The National Academies Press. https://doi.org/10.17226/11970

Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2020). A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning. *Archives of Computational Methods in Engineering*, *27*(4), 1071-1092. https://doi.org/10.1007/s11831-019-09344-w

Elkins, R. C., Davies, M. R., Brough, S. J., Gavaghan, D. J., Cui, Y., Abi-Gerges, N., & Mirams, G. R. (2013). Variability in high-throughput ion-channel screening data and consequences for cardiac safety assessment.

Journal of Pharmacological and Toxicological Methods, *68*(1), 112-122. https://doi.org/10.1016/j.vascn.2013.04.007

Fischer, I., Milton, C., & Wallace, H. (2020). Toxicity testing is evolving! *Toxicology Research*, *9*(2), 67-80. https://doi.org/10.1093/toxres/tfaa011

Gadaleta, D., Manganelli, S., Roncaglioni, A., Toma, C., Benfenati, E., & Mombelli, E. (2018). QSAR Modeling of ToxCast Assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. *Journal of Chemical Information and Modeling*, *58*(8), 1501-1517. https://doi.org/10.1021/acs.jcim.8b00297

Gadaleta, D., Spînu, N., Roncaglioni, A., Cronin, M. T. D., & Benfenati, E. (2022). Prediction of the Neurotoxic Potential of Chemicals Based on Modelling of Molecular Initiating Events Upstream of the Adverse Outcome Pathways of (Developmental) Neurotoxicity. *International Journal of Molecular Sciences*, *23*(6). https://doi.org/10.3390/ijms23063053

Gosling, J. P. (2019). The importance of mathematical modelling in chemical risk assessment and the associated quantification of uncertainty. *Computational Toxicology*, *10*, 44-50. https://doi.org/10.1016/j.comtox.2018.12.004

Heyndrickx, W., Mervin, L., Morawietz, T., Sturm, N., Friedrich, L., Zalewski, A., Pentina, A., Humbeck, L., Oldenhof, M., Niwayama, R., Schmidtke, P., Fechner, N., Simm, J., Arany, A., Drizard, N., Jabal, R., Afanasyeva, A., Loeb, R., Verma, S., … Ceulemans, H. (2022). MELLODDY: cross pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. *ChemRxiv*,

Cambridge(Cambridge Open Engage). https://doi.org/10.26434/chemrxiv-2022-ntd3r

Ishfaq, M., Aamir, M., Ahmad, F., M Mebed, A., & Elshahat, S. (2022). Machine Learning-Assisted Prediction of the Biological Activity of Aromatase Inhibitors and Data Mining to Explore Similar Compounds. *ACS Omega*, *7*(51), 48139-48149. https://doi.org/10.1021/acsomega.2c06174

Johnson, M., & Maggiora, G. M. (1990). *Concepts and applications of molecular similarity*. https://api.semanticscholar.org/CorpusID:117506064

Kernik, D. C., Morotti, S., Wu, H., Garg, P., Duff, H. J., Kurokawa, J., Jalife, J., Wu, J. C., Grandi, E., & Clancy, C. E. (2019). A computational model of induced pluripotent stem-cell derived cardiomyocytes incorporating experimental variability from multiple data sources. *The Journal of Physiology*, *597*(17), 4533-4564. https://doi.org/10.1113/JP277724

Kleinstreuer, N. C., Hoffmann, S., Alépée, N., Allen, D., Ashikaga, T., Casey, W., Clouet, E., Cluzel, M., Desprez, B., Gellatly, N., Göbel, C., Kern, P. S., Klaric, M., Kühnl, J., Martinozzi-Teissier, S., Mewes, K., Miyazawa, M., Strickland, J., van Vliet, E., … Petersohn, D. (2018). Non-animal methods to predict skin sensitization (II): An assessment of defined approaches. *Critical Reviews in Toxicology*, *48*(5), 359-374. https://doi.org/10.1080/10408444.2018.1429386

Kopańska, K., Rodríguez-Belenguer, P., Llopis-Lorente, J., Trenor, B., Saiz, J., & Pastor, M. (2023). Uncertainty assessment of proarrhythmia predictions derived from multi-level in silico models. *Archives of Toxicology*, *97*(10), 2721-2740. https://doi.org/10.1007/s00204-023-03557-6

Kotsampasakou, E., & Ecker, G. F. (2017). Predicting Drug-Induced Cholestasis with the Help of Hepatic Transporters—An in Silico Modeling Approach. *Journal of Chemical Information and Modeling*, *57*(3), 608-615. https://doi.org/10.1021/acs.jcim.6b00518

Kramer, J., Himmel, H. M., Lindqvist, A., Stoelzle-Feix, S., Chaudhary, K. W., Li, D., Bohme, G. A., Bridgland-Taylor, M., Hebeisen, S., Fan, J., Renganathan, M., Imredy, J., Humphries, E. S. A., Brinkwirth, N., Strassmaier, T., Ohtsuki, A., Danker, T., Vanoye, C., Polonchuk, L., … Gintant, G. (2020). Cross-site and cross-platform variability of automated patch clamp assessments of drug effects on human cardiac currents in recombinant cells. *Scientific Reports 2020 10:1*, *10*(1), 1-15. https://doi.org/10.1038/s41598-020-62344-w

Li, Z., Dutta, S., Sheng, J., Tran, P. N., Wu, W., Chang, K., Mdluli, T., Strauss, D. G., & Colatsky, T. (2017). Improving the in silico assessment of proarrhythmia risk by combining hERG (Human Ether-à-go-go-Related Gene) channel-drug binding kinetics and multichannel pharmacology. *Circulation: Arrhythmia and Electrophysiology*, *10*(2). https://doi.org/10.1161/CIRCEP.116.004628

Maertens, A., Golden, E., Luechtefeld, T. H., Hoffmann, S., Tsaioun, K., & Hartung, T. (2022). Probabilistic Risk Assessment – The Keystone for the Future of Toxicology. *ALTEX*, *39*(1), 3-29. https://doi.org/10.14573/altex.2201081

March-Vila, E., Ferretti, G., Terricabras, E., Ardao, I., Brea, J. M., Varela, M. J., Arana, Á., Rubiolo, J. A., Sanz, F., Loza, M. I., Sánchez, L., Alonso, H., & Pastor, M. (2023). A continuous in silico learning strategy to identify safety liabilities in compounds used in the leather and textile

industry. *Archives of Toxicology*, *97*(4), 1091-1111. https://doi.org/10.1007/s00204-023-03459-7

Mirams, G. R., Davies, M. R., Brough, S. J., Bridgland-Taylor, M. H., Cui, Y., Gavaghan, D. J., & Abi-Gerges, N. (2014). Prediction of Thorough QT study results using action potential simulations based on ion channel screens. *Journal of Pharmacological and Toxicological Methods*, *70*(3), 246-254. https://doi.org/10.1016/j.vascn.2014.07.002

O'Hara, T., Virág, L., Varró, A., & Rudy, Y. (2011). Simulation of the Undiseased Human Cardiac Ventricular Action Potential: Model Formulation and Experimental Validation. *PLOS Computational Biology*, *7*(5), e1002061. https://doi.org/10.1371/JOURNAL.PCBI.1002061

Padda, M. S., Sanchez, M., Akhtar, A. J., & Boyer, J. L. (2011). DRUG INDUCED CHOLESTASIS. *Hepatology (Baltimore, Md.)*, *53*(4), 1377-1387. https://doi.org/10.1002/hep.24229

Punt, A., Pinckaers, N., Peijnenburg, A., & Louisse, J. (2021). Development of a Web-Based Toolbox to Support Quantitative In-Vitro-to-In-Vivo Extrapolations (QIVIVE) within Nonanimal Testing Strategies. *Chemical Research in Toxicology*, *34*(2), 460-472. https://doi.org/10.1021/acs.chemrestox.0c00307

Raies, A. B., & Bajic, V. B. (2016). In silico toxicology: Computational methods for the prediction of chemical toxicity. *Wiley Interdisciplinary Reviews. Computational Molecular Science*, *6*(2), 147-172. https://doi.org/10.1002/wcms.1240

Rodríguez-Belenguer, P., March-Vila, E., Pastor, M., Mangas-Sanjuan, V., & Soria-Olivas, E. (2023a). Usage of model combination in computational toxicology. *Toxicology Letters*, *389*, 34-44. https://doi.org/10.1016/j.toxlet.2023.10.013

Rodríguez-Belenguer, P., Mangas-Sanjuan, V., Soria-Olivas, E., & Pastor, M. (2023b). Integrating Mechanistic and Toxicokinetic Information in Predictive Models of Cholestasis. *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/acs.jcim.3c00945

Rodríguez-Belenguer, P., Kopańska, K., Llopis-Lorente, J., Trenor, B., Saiz, J., & Pastor, M. (2023c). Application of Machine Learning to improve the efficiency of electrophysiological simulations used for the prediction of drug-induced ventricular arrhythmia. *Computer Methods and Programs in Biomedicine*, 107345. https://doi.org/10.1016/j.cmpb.2023.107345

Romano, J. D., Hao, Y., & Moore, J. H. (2022). Improving QSAR Modeling for Predictive Toxicology using Publicly Aggregated Semantic Graph Data and Graph Neural Networks. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, *27*, 187-198.

Russell, W.M.S, & Burch, R.L. (1960). The Principles of Humane Experimental Technique. *Medical Journal of Australia*, *1*(13), 500-500. https://doi.org/10.5694/j.1326-5377.1960.tb73127.x

Trinh, T. X., Seo, M., Yoon, T. H., & Kim, J. (2022). Developing random forest based QSAR models for predicting the mixture toxicity of TiO2 based nano-mixtures to Daphnia magna. *NanoImpact*, *25*, 100383. https://doi.org/10.1016/j.impact.2022.100383

Ulfa, A., Bustamam, A., Yanuar, A., Amalia, R., & Anki, P. (2021). Model QSAR Classification Using Conv1D-LSTM of Dipeptidyl Peptidase-4 Inhibitors. *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*, 1-6.

Wisniowska, B., Tylutki, Z., & Polak, S. (2017). Humans vary, so cardiac models should account for that too! En *Frontiers in Physiology* (Vol. 8,

Número SEP, p. 700). Frontiers Media S.A.
https://doi.org/10.3389/fphys.2017.00700