UNIVERSIDADE FEDERAL DE GOIÁS
FACULDADE DE FARMÁCIA

FRANCISCO LUCAS FEITOSA DE OLIVEIRA

# CYTO-SAFE: UMA FERRAMENTA DE APRENDIZADO DE MÁQUINA PARA IDENTIFICAÇÃO PRECOCE DE COMPOSTOS CITOTÓXICOS NA DESCOBERTA DE FÁRMACOS

GOIÂNIA/GO
2024

Rua 240, esquina com 5ª Avenida,
s/nº - Setor Leste Universitário
CEP 74605-170 - Goiânia - Goiás - Brasil.

Fone: (62) 3209-6044
Site: http://farmacia.ufg.br

UNIVERSIDADE FEDERAL DE GOIÁS
FACULDADE DE FARMÁCIA

# TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

## 1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): Francisco Lucas Feitosa de Oliveira

Título do trabalho: CYTO-SAFE: UMA FERRAMENTA DE APRENDIZADO DE MÁQUINA PARA IDENTIFICAÇÃO PRECOCE DE COMPOSTOS CITOTÓXICOS NA DESCOBERTA DE FÁRMACOS

## 2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [ X ] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

**Casos de embargo:**

- Solicitação de registro de patente;

- Submissão de artigo em revista científica;

- Publicação como capítulo de livro.

**Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.**

Documento assinado eletronicamente por **Carolina Horta Andrade**, **Professora do Magistério Superior**, em 04/12/2024, às 10:34, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Francisco Lucas Feitosa De Oliveira**, **Discente**, em 04/12/2024, às 10:51, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

FRANCISCO LUCAS FEITOSA DE OLIVEIRA

# CYTO-SAFE: UMA FERRAMENTA DE APRENDIZADO DE MÁQUINA PARA IDENTIFICAÇÃO PRECOCE DE COMPOSTOS CITOTÓXICOS NA DESCOBERTA DE FÁRMACOS

Trabalho de Conclusão de Curso apresentado ao curso de Farmácia da Universidade Federal de Goiás, como requisito parcial para a obtenção do título de Bacharel em Farmácia.

**Orientadora:** Profa. Dra. Carolina Horta Andrade

GOIÂNIA/GO
2024

Rua 240, esquina com 5ª Avenida,
s/nº - Setor Leste Universitário
CEP 74605-170 - Goiânia - Goiás - Brasil.

Fone: (62) 3209-6044
Site: http://farmacia.ufg.br

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Oliveira, Francisco Lucas Feitosa de
CYTO-SAFE: UMA FERRAMENTA DE APRENDIZADO DE MÁQUINA PARA IDENTIFICAÇÃO PRECOCE DE COMPOSTOS CITOTÓXICOS NA DESCOBERTA DE FÁRMACOS [manuscrito] / Francisco Lucas Feitosa de Oliveira. - 2024.
36 f.

Orientador: Prof. Dr. Carolina Horta Andrade.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Faculdade Farmácia (FF), Farmácia, Goiânia, 2024.
Bibliografia. Anexos. Apêndice.
Inclui siglas, gráfico, tabelas, lista de figuras, lista de tabelas.

1. Citotoxicidade. 2. Descoberta de Fármacos. 3. QSAR. 4. Aprendizado de Máquina. I. Andrade, Carolina Horta, orient. II. Título.

CDU 615.1

UNIVERSIDADE FEDERAL DE GOIÁS
FACULDADE DE FARMÁCIA

# ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Aos 29 dias do mês de novembro do ano de 2024 iniciou-se a sessão pública de defesa do Trabalho de Conclusão de Curso (TCC) intitulado "CYTO-SAFE: UMA FERRAMENTA DE APRENDIZADO DE MÁQUINA PARA IDENTIFICAÇÃO PRECOCE DE COMPOSTOS CITOTÓXICOS NA DESCOBERTA DE FÁRMACOS", de autoria de **Francisco Lucas Feitosa de Oliveira**, do curso de Farmácia da Faculdade de Farmácia da UFG. Os trabalhos foram instalados pela Profa. Dra. Carolina Horta Andrade – orientadora FF/UFG com a participação dos demais membros da Banca Examinadora: Prof. Dr. Artur Christian Garcia da Silva - FF/UFG e Me. Meryck Felipe Brito da Silva - FF/UFG. Após a apresentação, a banca examinadora realizou a arguição do(a) estudante. Posteriormente, de forma reservada, a Banca Examinadora atribuiu a nota final de 10,0 (Dez), tendo sido o TCC considerado aprovado.

Proclamados os resultados, os trabalhos foram encerrados e, para constar, lavrou-se a presente ata que segue assinada pelos Membros da Banca Examinadora.

Documento assinado eletronicamente por **Carolina Horta Andrade**, **Professora do Magistério Superior**, em 29/11/2024, às 16:20, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Artur Christian Garcia Da Silva**, **Professor do Magistério Superior**, em 29/11/2024, às 16:21, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Meryck Felipe Brito Da Silva**, **Usuário Externo**, em 04/12/2024, às 10:28, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4935163** e o código CRC **76DADC82**.

**Referência:** Processo nº 23070.056090/2024-41 SEI nº 4935163

## DEDICATÓRIA

Dedico este trabalho a pessoa mais especial da minha vida, minha mãe Carmina Maria, que lutou incansavelmente para que eu pudesse me dedicar aos estudos e conquistar meus sonhos. Seu amor e cuidado são fonte de inspiração para mim, esse diploma também é seu. Também dedico este trabalho a mim mesmo, por nunca desistir, mesmo durante os períodos mais obscuros.

## AGRADECIMENTOS

Inicialmente, gostaria de expressar minha profunda gratidão à minha família: minha mãe Carmina, meu pai Jânio, minha irmã Tainá e meu cunhado Philipe, pelo constante suporte e apoio ao longo de toda esta jornada. Agradeço também ao meu namorado, Luiz Otávio, que sempre me incentivou a perseguir meus objetivos.

Meu reconhecimento imensurável aos profissionais de saúde mental que me acompanharam ao longo desses anos. Sem o trabalho e a dedicação de vocês, eu não estaria aqui hoje celebrando esta conquista.

Agradeço às minhas amigas que me acompanham desde o início da graduação: Amanda, Bianca, Hozana e Isadora. Sem a companhia e a alegria de vocês, todo este percurso teria sido muito mais difícil.

Sou grato a todo o pessoal do 4º andar incluindo LabManas, NanoCIS e as servidoras do PPGCF, pelo apoio e pelos momentos de descontração compartilhados.

Agradeço também à minha orientadora, Profa. Carolina, por ter enxergado em mim potencial que eu mesmo não via e por me ajudar a crescer como cientista.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), meu agradecimento pelo incrível trabalho de incentivo aos estudantes de graduação através das bolsas de iniciação científica, bem como pelo reconhecimento do meu trabalho com o Prêmio Destaque em Iniciação Científica e Tecnológica.

À Universidade Federal de Goiás, que, mesmo diante de um desgoverno que ameaçou a educação brasileira, lutou para oferecer uma educação superior de qualidade e digna a todos os alunos.

Por fim, mas não menos importantes, agradeço a todos os colaboradores deste trabalho: Victoria, Igor, Sabrina, Joyce e Rodolpho. Obrigado por acreditarem nesta ideia comigo, vocês fizeram este projeto brilhar.

## RESUMO

**Introdução:** A citotoxicidade é a capacidade de uma substância causar dano irreversível a células vivas, levando à morte celular. A avaliação da citotoxicidade é crucial nos estágios iniciais do desenvolvimento de fármacos, permitindo identificar compostos tóxicos precocemente, mitigar riscos e reduzir testes em animais. Modelos de relações quantitativas entre estrutura e atividade (QSAR) que utilizam algoritmos de Inteligência Artificial (IA) podem prever a citotoxicidade a partir da estrutura química dos compostos. **Objetivo:** Desenvolver e validar modelos de QSAR para predizer a citotoxicidade de candidatos a fármacos e disponibilizá-los como uma aplicação web gratuita. **Métodos:** Utilizou-se um conjunto de dados da literatura com cerca de 90.000 compostos testados em fibroblastos embrionários de camundongos (3T3) e células renais embrionárias humanas (HEK 293). Após limpeza e curadoria dos dados, foram gerados descritores *Extended-Connectivity Fingerprints*. Os modelos foram gerados usando o algoritmo *Light Gradient Boosting* com 80% dos dados para treinamento e 20% para validação. **Resultados:** Os modelos gerados mostraram bom desempenho, com acurácia balanceada (BACC) de 0,91 após aplicação de técnica de balanceamento de dados. Os melhores modelos estão disponíveis no aplicativo web Cyto-Safe (http://cytosafe.labmol.com.br/), que incorpora elementos de IA explicável, permitindo a visualização de regiões moleculares associadas à citotoxicidade. **Conclusão:** Os modelos foram eficazes na classificação de compostos quanto à citotoxicidade nas linhagens 3T3 e HEK 293. O Cyto-Safe é um produto tecnológico que oferece à comunidade científica uma ferramenta rápida e confiável para avaliar a citotoxicidade de compostos químicos sem dados experimentais, acelerando a descoberta de fármacos.



**Palavras-chave:** Citotoxicidade; Descoberta de Fármacos; QSAR; Aprendizado de Máquina.

## ABSTRACT

**Introduction:** Cytotoxicity is the ability of a substance to cause irreversible damage to living cells, leading to cell death. Evaluating cytotoxicity is crucial in the early stages of drug development, allowing for the early identification of toxic compounds, mitigating risks, and reducing animal testing. Quantitative Structure-Activity Relationship (QSAR) models that use Artificial Intelligence (AI) algorithms can predict cytotoxicity based on the chemical structure of compounds. **Objective:** To develop and validate QSAR models to predict the cytotoxicity of drug candidates and make them available as a free web application. **Methods:** A dataset from the literature with approximately 90,000 compounds tested on mouse embryonic fibroblasts (3T3) and human embryonic kidney cells (HEK 293) was used. After data cleaning and curation, Extended-Connectivity Fingerprints descriptors were generated. The models were created using the Light Gradient Boosting algorithm with 80% of the data for training and 20% for validation. **Results:** The generated models showed good performance, with a balanced accuracy (BACC) of 0.91 after applying data balancing techniques. The best models are available in the web application Cyto-Safe (http://cytosafe.labmol.com.br/), which incorporates elements of explainable AI, allowing visualization of molecular regions associated with cytotoxicity. **Conclusion:** The models were effective in classifying compounds regarding cytotoxicity in the 3T3 and HEK 293 cell lines. Cyto-Safe is a technological product that offers the scientific community a fast and reliable tool to evaluate the cytotoxicity of chemical compounds without experimental data, accelerating drug discovery.



**Keywords:** Cytotoxicity; Drug Discovery; QSAR; Machine Learning.

## LISTA DE ABREVIAÇÕES E SIGLAS

3T3: Cultivo celular derivados de embriões de ratos BALB/C e Swiss.

ADME: Absorção, distribuição, metabolismo e exceção.

AI: do inglês, *Artificial Inteligence.*

ATP: Adenosina trifosfato.

AUC: do inglês, *Area under the curve.*

BACC: Acurácia balanceada

CSV: Formato de arquivo de valores separados por vírgulas.

DNA: Ácido desoxirribonucleico.

$EC_{50}$: Concentração que induz metade do efeito máximo.

F1: Média harmônica da Precisão e da taxa de verdadeiros positivos.

HEK 293: Cultivo celular de rins de embriões humanos.

hERG: do inglês, *human ether-a-go-go related gene.*

IA: Inteligência Artificial.

$IC_{50}$: Concentração que induz metade da inibição inibição máxima.

JCIM: do inglês, *Journal of Chemical Information and Modeling.*

$LD_{50}$: Concentração que que induz morte de metade dos animais testados

LDH: Lactato desidrogenase.

LGBM: do inglês, *Light Gradient Boosting Machine.*

MCC: Coeficiente de correlação de Matthews.

MTT: Brometo de 3-4,5-dimetil-tiazol-2-il-2,5-difeniltetrazólio.

OECD: Organização para a Cooperação e Desenvolvimento Econômico.

QSAR: Relações Quantitativas entre Estrutura Química e Atividade

SDF: Formato de arquivo, do inglês, *Structure Data File.*

Se: Sensibilidade.

SMILES: do inglês, *simplified molecular-input line-entry system.*

Sp: Especificidade.

XAI: do inglês, *Explainable AI.*

## LISTA DE FIGURAS

## LISTA DE EQUAÇÕES E TABELAS

## SUMÁRIO

# 1 INTRODUÇÃO

## 1.1 Citotoxicidade

### 1.1.1 Definição

A avaliação da citotoxicidade é uma etapa fundamental na pesquisa de novos fármacos, cosméticos e aditivos alimentares, sendo definida pela capacidade de um agente em causar danos celulares diretos, resultando em morte celular ou alterações no metabolismo celular (Çelik, 2018). Esse conceito abrange tanto a toxicidade desejada, como em agentes antitumorais, quanto a indesejada, que pode ocorrer no desenvolvimento de produtos que requerem comprovação de segurança. A citotoxicidade é avaliada *in vitro*, com testes de viabilidade que medem a integridade da membrana ou a capacidade metabólica, além de avaliar a sobrevivência celular e o potencial de crescimento celular. Ensaios que consideram interações celulares e sinalizações específicas, como reações inflamatórias, são cada vez mais investigados (Aslantürk, 2018; Freshney, 2010).

Esses efeitos podem ocorrer por meio de diversos mecanismos, por exemplo:

- Danos ao DNA: Podem levar a mutações, interferindo na replicação e transcrição genética (Menz *et al.*, 2023).
- Disfunção mitocondrial: Afeta a produção de ATP, levando a um déficit energético (Costa *et al.*, 2011).
- Estresse oxidativo: O acúmulo de espécies reativas de oxigênio pode danificar proteínas, lipídios e ácidos nucleicos (Sies, Berndt e Jones, 2024).
- Alterações na permeabilidade da membrana celular: Pode resultar em perda da homeostase iônica e equilíbrio osmótico (Li, Filice e Ding, 2014).

### 1.1.2 Avaliação experimental de respostas celulares à agressores externos

A avaliação experimental de citotoxicidade é fundamental para determinar os efeitos adversos de substâncias químicas, fármacos e outros compostos em células vivas. Os métodos tradicionais *in vitro* baseiam-se em culturas de células isoladas para avaliar a viabilidade celular, proliferação, apoptose ou necrose (Freshney, 2010).

A avaliação da citotoxicidade em células eucarióticas cultivadas *in vitro* frequentemente utiliza a integridade da membrana plasmática como principal parâmetro para distinguir células viáveis de não viáveis. Células que perderam essa integridade permitem a passagem de moléculas que, inicialmente, não eram permeáveis através da membrana danificada, caracterizando-as como mortas. Os métodos tradicionais para detecção de citotoxicidade envolvem duas abordagens principais: a utilização de marcadores que escapam do citoplasma

para o meio extracelular quando ocorre a morte celular, ou o uso de corantes vitais, que são pigmentos que, naturalmente, não permeiam células viáveis, mas podem penetrar seletivamente células mortas devido à membrana danificada (Riss *et al.*, 2019).

Entre os ensaios mais utilizados para avaliar a viabilidade celular estão o ensaio de Brometo de 3-4,5-dimetil-tiazol-2-il-2,5-difeniltetrazólio (MTT), Alamar Blue e Lactato Desidrogenase (LDH). Esses métodos analisam a atividade metabólica celular e a integridade da membrana plasmática, além de outros marcadores de citotoxicidade (Freshney, 2010; Kumar, Nagarajan e Uchil, 2018; Longhin *et al.*, 2022; Sylvester, 2011).

Um exemplo de marcador comumente utilizado é a adenosina trifosfato (ATP), que serve como molécula de energia química para as células. Quando a integridade da membrana celular é comprometida, as células liberam essa molécula para o meio extracelular. Kits comerciais como o CellTiter-Glo® geralmente utilizam enzimas que empregam esse ATP livre como cofator enzimático para transformar substratos inertes em marcadores luminescentes (Crouch *et al.*, 1993). Corantes como o azul de tripano têm a capacidade de penetrar seletivamente em células com a membrana danificada e se depositar nelas, resultando em um acúmulo de pigmento em células mortas (Avelar-Freitas *et al.*, 2014).

A busca por métodos alternativos na avaliação de citotoxicidade tem se intensificado, impulsionada por preocupações éticas, regulamentações mais rigorosas e a necessidade de modelos mais preditivos e eficientes. Esses métodos visam substituir, reduzir ou refinar o uso de animais em pesquisa, alinhando-se aos Princípios dos 3Rs (*Replacement, Reduction, Refinement*) (Tannenbaum e Bennett, 2015).

Uma abordagem promissora é o uso de modelos tridimensionais (3D) de cultura celular, como esferoides, organoides e bioimpressão 3D de tecidos (Edmondson *et al.*, 2014; Ravi *et al.*, 2015). Diferentemente das culturas bidimensionais tradicionais, esses modelos 3D replicam de forma mais fiel a arquitetura e as interações celulares presentes nos tecidos i*n vivo*. Isso permite uma avaliação mais precisa dos efeitos citotóxicos, incluindo a penetração do composto no tecido, gradientes de concentração e respostas celulares complexas.

Outra inovação significativa é o desenvolvimento de órgãos-em-chip (*organ-on-a-chip*) (Li *et al.*, 2017). Esses dispositivos micro fluídicos combinam células humanas vivas com engenharia de tecidos e microtecnologias para criar modelos que mimetizam a fisiologia e a mecânica de órgãos inteiros. Por exemplo, um chip que simula o fígado humano pode ser utilizado para avaliar a hepatotoxicidade de compostos, considerando fatores como metabolismo hepático e resposta imunológica local (Hassan *et al.*, 2020).

A modelagem computacional e as metodologias *in silico* também têm ganhado destaque como ferramentas complementares. Através de técnicas como QSAR (*Quantitative Structure-Activity Relationship*), é possível predizer a propriedade toxicológica de novos compostos com base em suas estruturas químicas. Essas abordagens aceleram o processo de triagem, reduzindo o número de compostos que necessitam de avaliação experimental detalhada (Bajorath, 2012).

Além disso, o uso de organismos-modelo alternativos, como o peixe-zebra (*Danio rerio*) em estágios embrionários, oferece vantagens como ciclo de vida curto, transparência corporal e facilidade de manipulação genética. Esses organismos permitem o estudo de efeitos tóxicos em um contexto multicelular e podem ser utilizados para triagem de alta capacidade (Lieschke e Currie, 2007).

A implementação de métodos alternativos é incentivada por iniciativas internacionais e regulamentações que promovem a validação de novas técnicas. A *Organisation for Economic Co-operation and Development* (OECD), por exemplo, desenvolveu a "Guidance Document 129", que utiliza valores de $IC_{50}$ obtidos *in vitro* para prever $LD_{50}$ em animais, reduzindo significativamente o uso de animais nesses testes. (OECD, 2010).

Apesar dos avanços, é importante reconhecer que esses métodos alternativos ainda enfrentam desafios, como a necessidade de validação extensiva e a limitação em replicar completamente a complexidade dos sistemas biológicos humanos. No entanto, a integração de abordagens alternativas com métodos tradicionais promete melhorar significativamente a eficiência, a ética e a relevância dos estudos de citotoxicidade, contribuindo para o desenvolvimento seguro e eficaz de novos produtos químicos e farmacêuticos (OECD, 2017).

## 1.2 Planejamento e desenvolvimento de novos fármacos

A indústria farmacêutica configura-se como uma das mais lucrativas globalmente. Em 2022, estimou-se que o valor de mercado agregado das 20 maiores fabricantes de medicamentos alcançou a cifra de 3,5 trilhões de dólares, enquanto a receita total das indústrias farmacêuticas somou 1,48 trilhão de dólares no mesmo período (Staciarini, 2023).

O desenvolvimento de novos fármacos é um processo caro e de alto risco, com uma duração média de aproximadamente 10 anos entre o início do projeto até a descoberta de um fármaco e sua disponibilização no mercado. Adicionalmente, estima-se que o custo médio desse processo gire em torno de 2 bilhões de dólares, sem garantia de sucesso (Hinkson, Madej e Stahlberg, 2020; Sun *et al.*, 2022). Embora seja uma atividade altamente dispendiosa, o mercado farmacêutico continua em expansão, sendo que, em janeiro de 2024, foram registrados

22.825 projetos de pesquisa e desenvolvimento de fármacos, abrangendo desde a fase pré-clínica até fases mais avançadas (Citeline, 2024).

O processo de descoberta de fármacos é dividido em cinco etapas principais, sendo elas: a fase de descoberta de um composto promissor, a fase pré-clínica, as fases clínicas (I, II e III), o registro nas agências regulatório-sanitárias competentes e a fase de farmacovigilância, também conhecida como fase clínica IV, conforme ilustrado na Figura 1 (Lombardino e Lowe, 2004; Singh *et al.*, 2023).

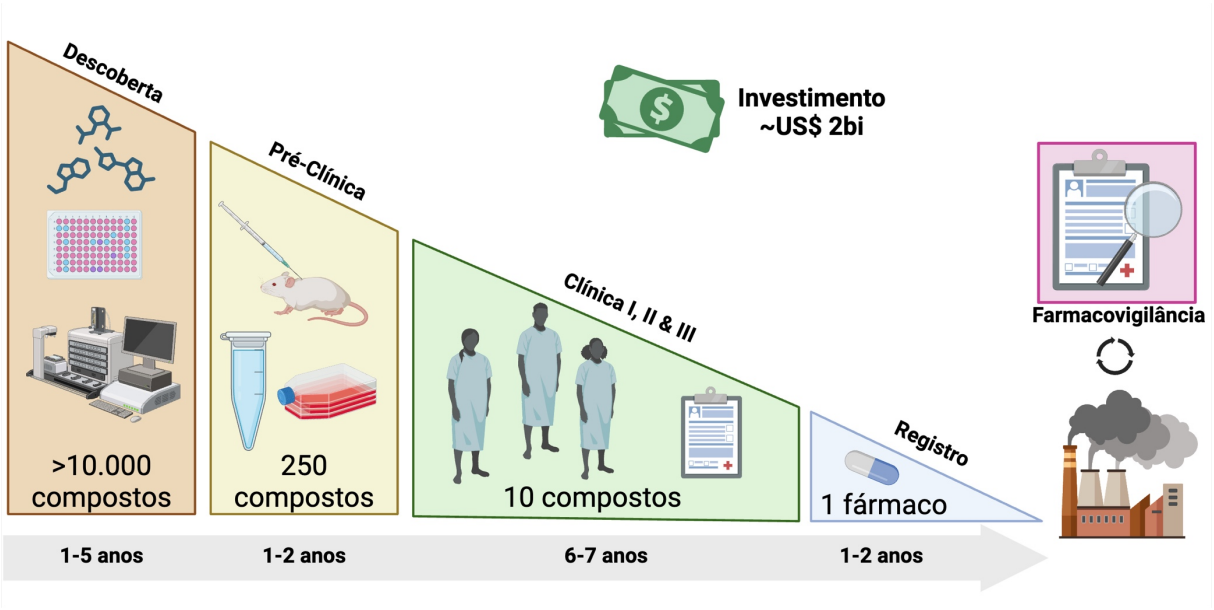


**Figura 1 -** Esquema geral do processo de descoberta de fármacos.

Fonte: O autor, utilizando software BioRender

Inicialmente, a fase de descoberta dedica-se à identificação de novos alvos biológicos e/ou novas moléculas candidatas que possam interagir de forma eficaz com esses alvos ou que possam ter atividade fenotípica. Neste estágio, empregam-se técnicas de biologia computacional e química medicinal, junto com a triagem de alto rendimento para explorar e avaliar milhares de compostos (Bleicher *et al.*, 2003; Raval, Kansagra e Ganatra, 2022; Trajanoska *et al.*, 2023). A triagem virtual desempenha um papel importante nessas etapas iniciais, permitindo a simulação computacional de novos compostos, suas interações moleculares, o que agiliza a priorização de candidatos promissores antes dos testes *in vitro* (Neves *et al.*, 2018). Esses compostos selecionados são então submetidos a testes *in vitro* para determinar sua atividade biológica, citotoxicidade e seletividade, visando selecionar as moléculas com maior potencial terapêutico.

Após esta triagem inicial, a fase pré-clínica é iniciada, onde os compostos selecionados são testados em modelos animais. Essa etapa é crucial para avaliar além da eficácia, a toxicidade, a farmacocinética e a farmacodinâmica dos candidatos, assegurando que sejam

eficazes e seguros para os subsequentes testes em seres humanos. Os estudos de ADME são realizados para entender como o fármaco é absorvido, distribuído, metabolizado e excretado pelo organismo (Shou, 2020).

Superada a fase pré-clínica, o processo evolui para as fases clínicas, subdivididas em três etapas: Fase I, onde a segurança do composto é testada em um pequeno grupo de voluntários saudáveis para definir a dosagem segura; Fase II, que expande o estudo para um grupo maior de pacientes para avaliar a eficácia do medicamento; e Fase III, onde o composto é testado em uma ampla população de pacientes, em diversos centros, para confirmar sua eficácia e monitorar possíveis efeitos colaterais, além de compará-lo com tratamentos já existentes (Van-Norman, 2016).

Concluídas as fases clínicas, o processo de registro do novo fármaco é iniciado. Nesta fase, todos os dados clínicos e pré-clínicos são compilados e submetidos às autoridades regulatórias para a revisão e aprovação do medicamento. Uma vez aprovado, o fármaco é finalmente comercializado, tornando-se disponível para venda e distribuição, oferecendo novas alternativas terapêuticas para o tratamento de diversas doenças (ANVISA, 2014).

Após o registro e comercialização, a fase de farmacovigilância ou fase clínica IV assegura o monitoramento contínuo da segurança e eficácia do medicamento. Essa etapa analisa dados de reações adversas, interações medicamentosas e eventos não previstos nos ensaios clínicos, permitindo ajustes regulatórios, como revisão de bulas, restrições de uso ou retirada do produto do mercado, quando necessário (ANVISA, 2020a; b).

## 1.3 Relações quantitativas entre estrutura química e atividade (QSAR)

### 1.3.1 História, evolução, aplicações e princípios

O conceito de QSAR (Relação Quantitativa Estrutura-Atividade) tem suas origens no século XIX, quando Brown e Fraser (1868) propuseram que a atividade farmacológica de uma substância tem correlação direta de sua constituição química. No início do século XX, Meyer (1899) e Overton (1901) observaram que a capacidade de uma substância induzir anestesia estava relacionada ao seu coeficiente de partição óleo/água, destacando a importância da lipofilicidade na atividade biológica.

Nas décadas seguintes, estudos de Hammett (1937) e Taft (1952) aprofundaram a compreensão dos efeitos eletrônicos e estéricos de substituintes químicos nas propriedades moleculares. Hansch e Fujita (1964) publicaram o primeiro estudo formal de QSAR, integrando parâmetros lipofílicos e eletrônicos para prever atividades biológicas em séries de compostos

relacionados. Esse trabalho pioneiro marcou o início da aplicação sistemática de modelos matemáticos na predição de atividades biológicas com base na estrutura química.

Com o avanço da tecnologia computacional e o desenvolvimento de métodos estatísticos mais sofisticados, o QSAR evoluiu para incorporar múltiplos descritores moleculares e modelar interações complexas usando técnicas como aprendizado de máquina e redes neurais profundas (Cherkasov *et al.*, 2013; Soares *et al.*, 2022). Esses modelos modernos permitem prever atividades biológicas com maior precisão, sem a necessidade de experimentação laboratorial extensiva, economizando tempo e recursos no desenvolvimento de novos fármacos.

Os modelos de QSAR baseiam-se no princípio de que moléculas com estruturas químicas semelhantes exibem atividades biológicas semelhantes, formalizado pela equação:

**Equação 1 -** Equação simplificada de QSAR

$$\boldsymbol{f(x) = \sum_{i=1}^{n} (C_i * D_i) + a}$$

Onde:

- f(x): representa a propriedade biológica ou resposta do sistema ao composto;
- $C_i$: são coeficientes ajustados por métodos estatísticos ou de aprendizado de máquina, que ponderam a influência de cada descritor na atividade biológica predita;
- $D_i$: são os descritores moleculares, que capturam aspectos relevantes da estrutura química;
- a: é uma constante que ajusta a predição geral.

Essa equação correlaciona características estruturais de uma molécula com suas propriedades biológicas ou físico-químicas. Os coeficientes $C_i$ são ajustados para minimizar erros de predição utilizando técnicas estatísticas ou de aprendizado de máquina. A metodologia geral de treinamento e validação de modelos de QSAR está ilustrada na Figura 2.

Nas últimas décadas, o QSAR tem sido amplamente utilizado no desenvolvimento de novos fármacos, na avaliação toxicológica de compostos e na descoberta de compostos com propriedades específicas. No contexto farmacêutico, esses modelos possibilitam prever a atividade de compostos antes de sua síntese e testes experimentais, otimizando o processo de descoberta e desenvolvimento de fármacos (Alves *et al.*, 2017; Neves *et al.*, 2022).

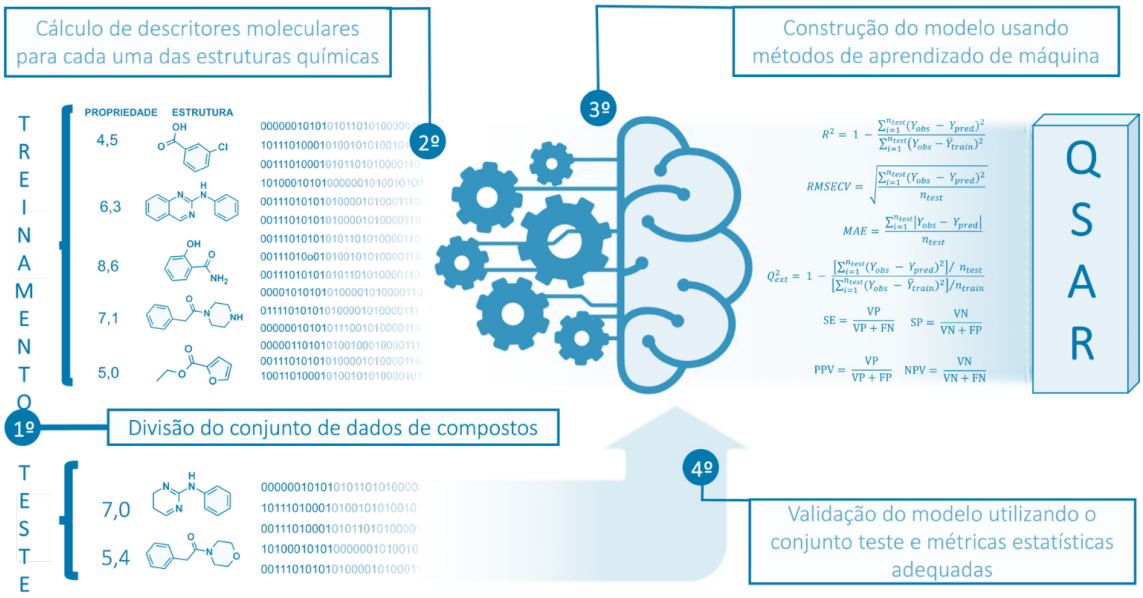


**Figura 2 -** Etapas do processo de treinamento e validação de modelos de QSAR

Fonte: Adaptado de Neves et al. (2022)

### 1.3.2 Descritores moleculares

Os descritores moleculares constituem elementos centrais nos modelos de QSAR, sendo variáveis quantitativas que traduzem características estruturais das moléculas em parâmetros que podem ser usados para prever suas propriedades biológicas. Em suma, descritores moleculares são o resultado de um cálculo matemático que converte a informação química em um número útil e interpretável por sistemas computacionais (Todeschini e Consonni, 2000).

Uma forma de classificar os descritores moleculares é segundo sua dimensionalidade: descritores 0D geralmente são propriedades experimentais e físico-químicas, como Log-P, massa molar, pKa, entre outros; descritores 1D utilizam-se de contagens de doadores e aceptores de hidrogênio, contagens de átomos ou de grupamentos funcionais específicos (Xue e Bajorath, 2000); descritores 2D representam estruturas bidimensionais das moléculas, como os descritores topológicos e os *fingerprints* moleculares (Rogers e Hahn, 2010); descritores 3D preocupam-se em representar características tridimensionais das moléculas, como, por exemplo, a área de superfície polar (Cramer, Patterson e Bunce, 1988; Klebe, Abraham e Mietzner, 1994). Há também descritores mais complexos, como 4D, 5D e 6D, que associam propriedades 3D com informações adicionais, como múltiplas conformações, modelo de encaixe induzido no receptor ou estados quânticos (Hopfinger *et al.*, 1997; Vedani e Dobler, 2002; Vedani, Dobler e Lill, 2005).

Em abordagens recentes, pesquisadores desenvolveram métodos livres de descritores (*descriptor-free*), que se beneficiam do conhecimento de modelos de linguagem natural como BERT e modelos baseados em redes *Transformers* para representar, em um espaço vetorial, as

características semânticas de uma palavra (Devlin *et al.*, 2018; Vaswani *et al.*, 2017). No caso da representação química, utilizam-se da fórmula molecular ou outras representações, como *simplified molecular-input line-entry system* (SMILES), para calcular tal vetorização (Ahmad *et al.*, 2022).

A seleção adequada dos descritores moleculares é um passo crucial no desenvolvimento de modelos de QSAR robustos e precisos, uma vez que eles devem capturar os aspectos mais relevantes da estrutura química que influenciam a resposta biológica (Seal *et al.*, 2021).

### 1.3.3 Boas práticas de desenvolvimento e validação de modelos de QSAR

A Organização para a Cooperação e Desenvolvimento Econômico, do inglês *Organisation for Economic Co-operation and Development* (OECD) estabeleceu diretrizes rigorosas para o desenvolvimento e validação de modelos QSAR para fins regulatórios, visando garantir a confiabilidade e a reprodutibilidade desses modelos (OECD, 2014). Entre as boas práticas recomendadas, destacam-se:

1. Propriedade biológica definida: A propriedade biológica que o modelo QSAR pretende prever deve ser bem definido e mensurável por experimentos.
2. Algoritmo claro: O método de desenvolvimento do modelo deve ser claro e não ambíguo, permitindo sua reprodução e validação por outros pesquisadores.
3. Domínio de aplicabilidade: O modelo QSAR deve ser aplicável a uma classe específica de compostos, com limitações claramente definidas sobre os tipos de moléculas para os quais as predições são válidas.
4. Validação robusta: A validação do modelo deve ser realizada por meio de técnicas estatísticas adequadas, como a validação cruzada, e utilizando conjuntos de dados externos. As métricas de desempenho do modelo, como acurácia e coeficiente de correlação, devem ser reportadas de forma detalhada.
5. Interpretação mecanística (quando aplicável): Sempre que possível, o modelo QSAR deve fornecer uma interpretação mecanística clara das relações entre a estrutura molecular e a atividade biológica ou toxicológica prevista. Isso envolve identificar quais propriedades químicas ou estruturais das moléculas estão correlacionadas com a resposta biológica.

Essas diretrizes asseguram que os modelos de QSAR possam ser aplicados com maior confiança, promovendo o uso de abordagens computacionais como métodos alternativos à experimentação animal.

## 1.4 JUSTIFICATIVA

A avaliação da citotoxicidade torna-se essencial no desenvolvimento de novos fármacos, garantindo que os compostos sejam seguros antes de avançarem para as fases clínicas (Gupta, 2016). Com o crescimento exponencial da indústria farmacêutica e o alto investimento necessário para a descoberta e desenvolvimento de fármacos, há uma demanda crescente por métodos mais eficientes, econômicos e éticos para a triagem de compostos (Singh *et al.*, 2023). Os métodos tradicionais de avaliação de toxicidade, embora eficazes, são frequentemente dispendiosos, demorados e dependentes do uso extensivo de animais em experimentação, o que levanta questões éticas e regulatórias (Tannenbaum e Bennett, 2015).

Nesse contexto, os modelos de QSAR emergem como uma ferramenta promissora. Ao correlacionar características estruturais de moléculas com suas atividades biológicas, os modelos QSAR permitem a previsão da citotoxicidade de novos compostos de forma rápida e sem a necessidade inicial de testes laboratoriais (Cherkasov *et al.*, 2013; Neves *et al.*, 2018). Isso não apenas acelera o processo de desenvolvimento de fármacos, mas também reduz os custos e aborda preocupações éticas associadas ao bem-estar animal.

Nosso grupo de pesquisa, ao longo dos últimos 15 anos, tem se dedicado ao desenvolvimento de modelos QSAR para diversos parâmetros toxicológicos, como o Pred-hERG que fornece modelos para predição de cardiotoxicidade pela inibição da subunidade do canal de potássio codificado pelo gene hERG (*human ether-a-go-go related gene*) (Sanches *et al.*, 2023, 2024), e o PredSkin que disponibiliza modelos para predição de sensibilização dérmica abrangendo todos os eventos-chave da via de sensibilização dérmica (Borba *et al.*, 2021). Ao aprimorar continuamente essas ferramentas preditivas, buscamos fornecer à comunidade científica e à indústria farmacêutica métodos confiáveis para a identificação precoce de potenciais riscos toxicológicos.

Portanto, este trabalho justifica-se pela necessidade de desenvolver e validar modelos de QSAR robustos para a avaliação da citotoxicidade, contribuindo para a otimização do processo de descoberta de fármacos, e disponibilizá-los gratuitamente como uma ferramenta intitulada Cyto-Safe.

# 2 ARTIGO CIENTÍFICO ACEITO PARA PUBLICAÇÃO NO JOURNAL OF CHEMICAL INFORMATION AND MODELING - Cyto-Safe: A Machine Learning Tool for Early Identification of Cytotoxic compounds in Drug Discovery

DOI: https://doi.org/10.1021/acs.jcim.4c01811

Francisco L. Feitosa[1-3]; Victoria F. Cabral[1-3]; Igor H. Sanches[1-3]; Sabrina Silva-Mendonca[1-3]; Joyce V. B. Borba[1-3]; Rodolpho C. Braga[4]; Carolina Horta Andrade[1-3*]

[1] Laboratory of Molecular Modeling and Drug Design (LabMol), Faculdade de Farmácia, Universidade Federal de Goiás, Goiânia, Goiás, Brazil.

[2] Center for the Research and Advancement in Fragments and molecular Targets (CRAFT), School of Pharmaceutical Sciences at Ribeirão Preto, University of São Paulo, Ribeirão Preto, SP, Brazil.

[3] Center for Excellence in Artificial Intelligence (CEIA), Institute of Informatics, Universidade Federal de Goiás, Goiânia, 74605-170, GO, Brazil.

[4] InsilicAll Inc., São Paulo, SP, Brazil.

Corresponding Author

* Address for correspondence: Laboratory for Molecular Modeling and Design, Faculdade de Fármacia, Universidade Federal de Goiás, Goiânia, GO, 74605-170, Brazil; Telephone: +55 62 3209-6451; E-mail: carolina@ufg.br

## 2.1 ABSTRACT

Cytotoxicity is essential in drug discovery, enabling early evaluation of toxic compounds during screenings to minimize toxicological risks. *In vitro* assays support high-throughput screening, allowing for efficient detection of toxic substances while considerably reducing the need for animal testing. Additionally, AI-based Quantitative Structure-Activity Relationship (AI-QSAR) models enhance early-stage predictions by assessing the cytotoxic potential of molecular structures, which helps prioritize low-risk compounds for further validation. We present a freely accessible web application designed for identifying potential cytotoxic compounds utilizing QSAR models. This application utilizes machine learning techniques and is built on a dataset of approximately 90,000 compounds, evaluated against two cell lines, 3T3 and HEK 293. Users can interact with the app by inputting a SMILES representation, uploading CSV or SDF files, or sketching molecules. The output includes a binary prediction for each cell line, a confidence percentage, and an explainable AI (XAI) analysis. The Cyto-Safe Web-App Version 1.0 is available at http://insightai.labmol.com.br/.

KEYWORDS: Cytotoxicity, Drug Discovery, QSAR Models, Machine Learning, In Vitro Assays, Explainable AI

## 2.2 INTRODUCTION

Cell viability and cytotoxicity assays are fundamental tools used in biomedical research to measure the cytotoxic effects of various substances on living cells. Cytotoxicity specifically refers to the detrimental impacts these substances have on cellular health, often leading to cell death[1]. These assays are critical for drug screening, identifying compounds that demonstrate cytotoxic effects which are then typically excluded in both target-based and cell-based phenotypic screenings. This is particularly vital in the context of oncology and neurodegenerative diseases, where understanding substance toxicity is crucial for drug development[2].

A significant advantage of *in vitro* assays is their ability to perform high-throughput screening, allowing researchers to efficiently identify toxic compounds or potential therapeutic agents from large numbers of samples. Additionally, these assays align with the growing emphasis on ethical research by reducing the need for animal testing, making them increasingly valuable in the scientific community[3,4]. They operate by measuring various cellular functions that indicate cytotoxicity, such as cell membrane permeability, enzyme activity, cell adherence, ATP production, co-enzyme production, and nucleotide uptake activity[5]. Accurately predicting chemical-induced cytotoxicity early in the drug development process (preferably before the compounds are even synthesized) is essential. Early detection of cytotoxicity can help prevent costly failures in later stages of development and ensures that only the most promising candidates advance.

The increasing costs associated with ADME/Tox (Absorption, Distribution, Metabolism, Excretion, and Toxicity) studies have spurred the development of *in silico* methods, particularly within large pharmaceutical companies that possess extensive and internally consistent datasets. While initial computational efforts outside the pharmaceutical sector faced limitations due to smaller datasets, the growing availability of larger datasets in public repositories has significantly improved the potential for successful model development[6]. The use of artificial intelligence (AI) to build cytotoxicity models shows great promise for enhancing early-stage cytotoxicity prediction. By predicting the cytotoxic potential of chemical compounds based on their molecular structures, these computational models can support virtual screening campaigns, helping to prioritize compounds with lower cytotoxic risks for further experimental validation. This approach streamlines the identification of viable drug candidates.

Building on insights from our previous research[7–10], we have developed QSAR models using a diverse dataset of compounds tested on 3T3 and HEK-293 cell lines. We present a new web-accessible application designed to predict the cytotoxicity potential of chemicals,

integrating a carefully curated database of 3T3 and HEK-293 cytotoxicity data. The web app features novel models for predicting cytotoxicity, developed exclusively with open-source tools. Available as a web version (version 1.0), it facilitates efficient virtual screening of chemical libraries, aiding in the identification of potential cytotoxic compounds for further investigation. The result includes a binary prediction for each cell line, a confidence percentage, and an explainable AI (XAI) analysis for visual interpretation of the results. This tool can be freely accessed at LabMol Insight AI portal http://insightai.labmol.com.br/.

## 2.3 CYTO-SAFE

### 2.3.1 Data collection

Cytotoxicity data was sourced from a dataset provided by the National Center for Advancing Translational Sciences (NCATS)[11]. This dataset includes the results of approximately 90,000 compounds tested for cytotoxicity using the luciferase assay, commercially known as CellTiter-Glo®, in a 48 hours of incubation time, across two different cell lines: 3T3 and HEK 293. The original data can be accessed via PubChem under AID 1345082 and AID 1345083.

### 2.3.2 Data cleaning and curation

Initially, both datasets comprised 93781 records. However, after eliminating entries with inconclusive outcomes and incomplete chemical structure information, the analysis yielded 67041 compounds from the 3T3 series and 64508 compounds from the HEK 293 series.

Subsequently, we implemented a rigorous data curation protocol as described by Fouches et al[12,13], resulting in 66,620 unique compounds for the 3T3 series. According to the threshold stablished of EC50 value of ≤10 μM [11], 62,613 compounds were labeled as non-cytotoxic, and 4,007 records were considered cytotoxic for the 3T3. In the HEK dataset, a total of 64,094 records were analyzed, with 6,141 compounds labeled as cytotoxic.

To tackle potential data unbalancing that might introduce bias into the classification models, we strategically applied the NearMiss v.3 under sampling method[14] executed on Imbalanced-learn package (https://imbalanced-learn.org), setting the sampling strategy as 0.2 and number of near neighbors to 50. This method enabled to get compounds in the majority class (non-cytotoxic) with the minimum distance from minority class examples (cytotoxic) maintaining a balanced proportion between cytotoxic and non-cytotoxic samples. By adopting this approach, our primary objective was to establish a more equitable and representative dataset, thereby enhancing the effectiveness of our classification model training process. The entire processed data is available in Supporting Information.

### 2.3.3 QSAR Modeling

Classification QSAR models were generated and validated in accordance with the established standards and principles of QSAR modeling[15,16]. The molecules were converted into a binary language, based on Extended Connectivity Fingerprints[17] with radius 2 (ECPF4) and 1024 bits, using the open-source library RDKit[18]. ECFP was chosen to capture detailed atomic environments without relying on predefined features. Light Gradient Boosting machine learning algorithm (LGBM)[19] was executed in Python 3.10. The dataset was stratified split into training (80%) and external (20%) sets. The external set was held out entirely during hyperparameter optimization to ensure unbiased evaluation. Bayesian optimization was conducted using the scikit-optimize, with 100 iterations and 10-fold cross-validation, optimized by balanced accuracy. Class weights were applied to handle class imbalance in the dataset. The selected hyperparameters for each model are provided in the Supporting Information. For all models, we calculated the following metrics on the external set: Balanced Accuracy (BACC), Matthew's correlation coefficient (MCC), Precision, Recall, F1 score, and plotted the confusion matrix.

### 2.3.4 Y-randomization

We performed fifty rounds of Y-randomization to assess the robustness and validity of our predictive models. Y-randomization involves shuffling the dependent variable (*Y*; cytotoxicity outcomes) while keeping the independent variables (*X*; molecular fingerprints) intact.

### 2.3.5 Deployment

The Cyto-Safe web application was deployed on a cloud-based platform, utilizing a Flask backend and a Jinja2 template-driven frontend to ensure scalability and responsive user interfaces. Machine learning models were then integrated within the Flask framework to facilitate both individual and batch predictions of up to 10 compounds via CSV or SDF file inputs. The application incorporates the Ketcher molecular editor (version 2.10.0; EPAM) to provide an intuitive interface for drawing and editing chemical structures, enhancing user experience and data accuracy. Prediction results can be exported as spreadsheets for detailed analysis or as web-based reports for immediate review.

### 2.3.6 Explainable AI (XAI)

In this study, we employed an Explainable AI (XAI) framework to interpret our model's binary classification predictions regarding the cytotoxicity of compounds in 3T3 and HEK-293 cell lines. We used the methodology proposed by Riniker and Landrum [20] that systematically removes bits in the molecular fingerprints that correspond to specific atoms or functional groups

and assess how these changes influenced the model's predictions. We normalized these contributions and visualized them using similarity maps and heatmaps analogous to topographical representations.

In these visualizations, structural fragments predicted to increase toxicity were highlighted in red, while those predicted to decrease toxicity were highlighted in green. This approach allowed us to identify key structural features affecting the model's decisions, providing deeper insights into the patterns and potential biases related to the predicted outcome.

## 2.4 RESULTS AND DISCUSSION

### 2.4.1 Modelling

The under-sampling technique was applied to the majority class (non-cytotoxic) using two different proportions: 1:1 and 1:5, relative to the minority class, cytotoxic. As a result, the balanced 3T3 dataset comprised 8,014 records for the 1:1 ratio and 24,042 for the 1:5 ratio. For the HEK 293 balanced datasets, the corresponding records were 12,282 and 36,846, respectively.

Further analysis was conducted using clustering to identify groups of structurally similar compounds that exhibit contrasting outcomes (cytotoxic *versus* non-cytotoxic), with the goal of simulating potential activity cliffs within the training set. The results revealed that only 11.8% of the clusters from the 3T3 dataset and 13.2% from the HEK293 dataset contained compounds with differing outcomes. These findings indicate a high level of data reliability while also highlighting the challenges that the algorithm must navigate to learn effectively from the training set (see Supporting Information - Supplementary Methods and Results).

Following training, all models were evaluated on the test set. The models exhibited satisfactory performance across both under-sampling proportion ratios, indicating their proficiency in distinguishing between cytotoxic and non-cytotoxic compounds. Notably, the models demonstrated robust generalization capabilities by accurately classifying samples not encountered during the training phase.

When comparing overall metrics, the models trained with the 1:5 ratio showed a slight improvement over those using the 1:1 ratio. This was particularly evident in the Matthews Correlation Coefficient (MCC), where the average values increased from 0.61 to 0.86, underscoring the reliability and informativeness of the predictions. Sensitivity (Se) also improved significantly, with average values rising from 0.65 to 0.83, indicating that the models retained their ability to correctly identify cytotoxic compounds despite the increased under-sampling. These results, detailed in Table 1, support the decision to adopt the 1:5 ratio in Cyto-Safe's back-end prediction algorithm.

**Table 1 -** Performance metrics of QSAR models predictions on the test sets for cytotoxicity classification in different cell lines and balancing proportions, using LGBM algorithm.

| | BACC | AUC | F1 | MCC | Precision | Se | Sp |
|---|---|---|---|---|---|---|---|
| 3T3 Unbalanced | 0.81 | 0.81 | 0.69 | 0.68 | 0.78 | 0.63 | 0.99 |
| 3T3 1:1 | 0.80 | 0.80 | 0.80 | 0.59 | 0.80 | 0.79 | 0.81 |
| 3T3 1:5 | 0.92 | 0.92 | 0.90 | 0.88 | 0.96 | 0.84 | 0.99 |
| HEK Unbalanced | 0.83 | 0.83 | 0.73 | 0.71 | 0.81 | 0.67 | 0.98 |
| HEK 1:1 | 0.81 | 0.81 | 0.81 | 0.63 | 0.81 | 0.81 | 0.81 |
| HEK 1:5 | 0.90 | 0.90 | 0.87 | 0.84 | 0.92 | 0.82 | 0.99 |

BACC: Balanced accuracy; AUC: Area under the curve; F1: F1 score; MCC: Matthew's correlation coefficient; Se: Sensibility; Sp: Specificity.

The t-SNE plots for each balanced approach (see Supporting Information) reflect the statistical improvements observed with a 1:5 ratio compared to both the 1:1 ratio and the unbalanced set. Unlike the 1:1 ratio, which excludes highly similar compounds and risks misleading predictions, the 1:5 ratio better captures the chemical space of the original unbalanced dataset.

Moreover, we applied the Y-randomization method to determine if the correlations identified by the model between molecular fragments and cytotoxic effects are genuine or simply artifacts of random associations. As a result of the fifty rounds on each dataset, we observed an average ACC of 0.5 for both datasets and MCC of 0.02 and 0.01 for 3T3 and HEK 293, respectively, confirming the robustness of the models.

### 2.4.2 Usability

As mentioned previously, users can easily draw a molecule for testing or upload a batch of molecules for prediction using a CSV or SDF file, with a limit of ten SMILES per request. The results are displayed in a list format for each model (3T3 and HEK 293), indicating whether each molecule is predicted to be toxic or non-toxic. Additionally, the predicted probabilities (representing the model's confidence) are provided as percentages, calculated using the *predict_proba* method from the LightGBM library. A dedicated button also allows users to initiate Explainable Artificial Intelligence (XAI) analysis (Figure 1).

Furthermore, in accordance with the best practices established by the Organization for Economic Co-operation and Development (OECD)[16], we have defined the applicability domain of the model's predictions to ensure their reliability. The threshold was set at 0.09, corresponding to the 5$^{th}$ percentile of the Tanimoto similarity distribution among compounds in the training set[21]. This relatively low threshold reflects the high structural diversity within the training set. A similarity distribution analysis is provided in Figure S3 (Supporting

Information). The information regarding whether the tested compound is within or outside the applicability domain is available on the prediction results page, helping users understand the limitations of the model's outputs.

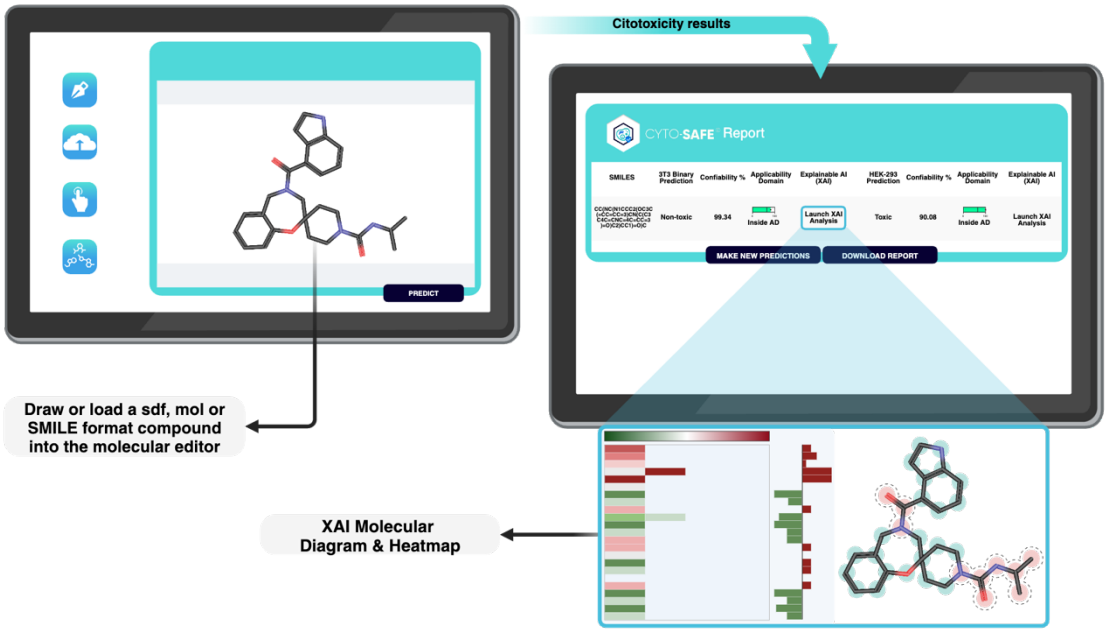


**Figure 1 -** General scheme of usage, outcome and XAI of Cyto-Safe web app.

### 2.4.3 Explainable AI (XAI) with molecular diagrams and heatmaps

Cyto-Safe 1.0 is equipped with explainable AI molecular diagrams and heatmaps to help users better understand the model's output. The molecular diagram displays the molecule contoured in either red or green, where red represents a strong influence on the model's prediction of "cytotoxic" and green indicates a strong influence on the prediction of "non-cytotoxic." As a case study, we evaluated the structures of two established drugs: Doxorubicin, a well-known chemotherapeutic agent with a cytotoxic mechanism of action, and Ibuprofen, a widely used nonsteroidal anti-inflammatory drug. Importantly, neither of these molecules was included in the training sets for the models.

Both models classified Doxorubicin as "cytotoxic," as anticipated, and the molecular diagram reinforced this classification by coloring almost the entire molecule in red in both predictions. In contrast, Ibuprofen was classified as "non-cytotoxic," with its molecular diagram predominantly highlighted in green, indicating the model's strong confidence in this prediction (Figure 2). This concordance between the molecular diagrams and the known pharmacological profiles of these compounds underscores the robustness and interpretability of the models.

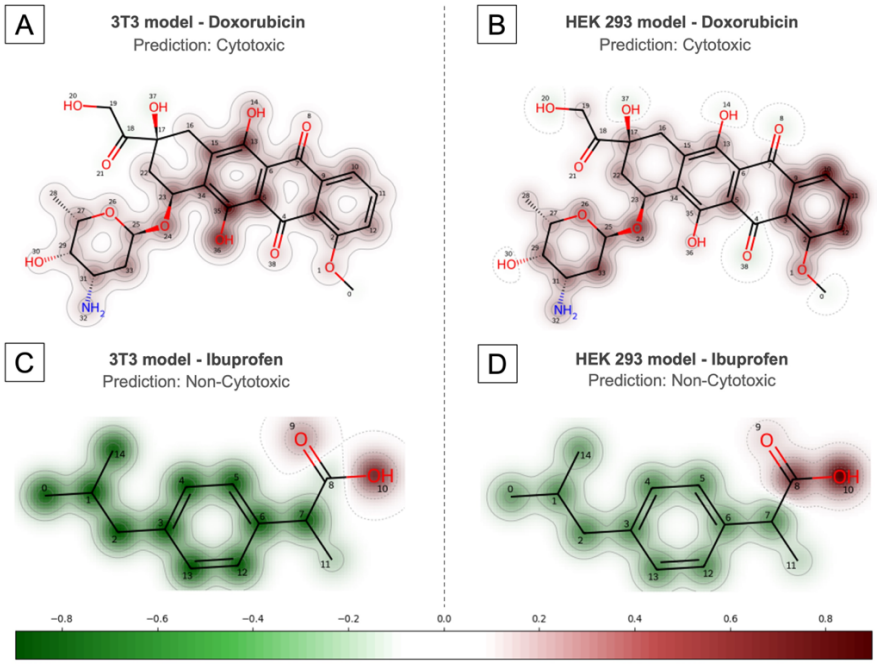


**Figure 2 -** Explainable AI (XAI) molecular diagrams illustrating the model's predictions for Doxorubicin on the 3T3 (A) and HEK-293 (B) models, and for Ibuprofen on the 3T3 (C) and HEK-293 (D) models. Red contoured regions highlight areas with a strong positive influence on predicted cytotoxicity, whereas green contoured regions indicate a strong positive influence on predicted non-toxicity. The intensity of the contour colors reflects the magnitude of their influence, with darker shades representing a greater impact on the model's predictions.

The atom influence is further illustrated in heatmaps provided in Figures S4-S7 (Supporting Information). These heatmaps employ the same color coding, with red contours indicating atoms that have a strong influence on the model's prediction of cytotoxicity and green contours representing atoms that have strong influence on the prediction of non-cytotoxicity. The intensity of the color contours reflects the strength of the influence, providing users with a detailed understanding of the factors affecting the model's predictions at both the fragment and atom levels.

The XAI feature of Cyto-Safe is essential for understanding the underlying mechanisms of both models' predictions. By offering molecular diagrams and atom-wise heatmaps, it identifies specific molecular regions that contribute to either "cytotoxic" or "non-cytotoxic" outcomes. This capability allows users to analyze structural fragments influencing toxicity, facilitating applications in drug design, safety assessments, and compound optimization.

### 2.4.4 Limitations

It is important to mention that the Cyto-Safe 1.0 was developed for predicting cytotoxicity using data from 3T3 and HEK 293 cell lines with the CellTiter-Glo® assay, faces limitations regarding dataset characteristics and generalizability. The reliance on only two cell lines - 3T3 (derived from mouse fibroblasts) and HEK 293 (from human embryonic kidney cells) - restricts the model's applicability to other biological contexts, as these cell lines may

not accurately reflect the behaviors of a broader range of cell types. Furthermore, variations in cytotoxicity responses among different cell types - including differences between permanent cell lines and primary cells - underscore the necessity of considering cell type-specific nuances when interpreting cytotoxicity data[22].

In addition to the cell line limitations, the specificity of the CellTiter-Glo® assay complicates the model's predictions. Different cytotoxicity assays, such as Lactate Dehydrogenase (LDH) release and MTT reduction, may detect varying aspects of cell viability, leading to divergent toxicological profiles. Additionally, variability in experimental conditions - such as cell density, culture medium, and incubation durations - further impacts reproducibility and the model's reliability across different settings[23]. The model's training data are confined to specific incubation times, limiting its applicability for different exposure durations. To enhance the model's robustness and generalizability, it would be necessary to incorporate a more diverse range of datasets, expanding beyond the current parameters under which it was developed[24]. Until such advancements are made, users should approach the model's predictions with a clear understanding of these constraints.

### 2.4.5 Comparative Analysis of QSAR Models for Cytotoxicity Prediction

Numerous QSAR models have been developed to predict cytotoxicity, each offering distinct strengths and facing particular challenges. Langdon et al. [25] employed Bayesian models to achieve cross-assay predictivity; however, the reliance on heterogeneous assay data introduces variability, potentially affecting the consistency and reliability of predictions, and their models lack an interpretability feature.

ProTox 3.0 [26] excels in providing multi-endpoint toxicity predictions by leveraging molecular similarity and machine learning. Its computational efficiency makes it suitable for large-scale screening. However, the model operates as a black box, limiting interpretability and hindering its application in guiding molecular modifications. Yin et al. [27] addressed the challenge of imbalanced datasets by implementing ensemble learning methods, which deliver strong predictive performance. Despite this, these models are not openly available to the community.

Other notable contributions include the work by Liu et al. [28], which focused on predicting microglial cytotoxicity using machine learning models integrated with feature selection and Shapley Additive Explanations. This method provided detailed substructure-level insights, enabling a deeper understanding of toxicological mechanisms. However, its requirement for local installation and computational skills limits its accessibility to users

without a strong technical background, and it is limited by only predicting microglial cytotoxicity.

Sun et al. [29] constructed predictive models based on multiple cell lines using support vector machines (SVMs). While these models achieved high predictive accuracy, they lack an explainable AI feature. Webel et al. [30] explored the use of deep learning to identify cytotoxic substructures, offering mechanistic insights via Deep Taylor Decomposition. Although their work is promising, it remains exploratory and does not provide a readily available tool for broader use.

Cyto-Safe offers a distinctive approach by integrating prediction accuracy with interpretability and accessibility. Its web-based interface eliminates the need for installations, allowing users from diverse backgrounds to easily conduct predictive analyses. Moreover, Cyto-Safe incorporates Explainable AI (XAI), which generates atom-level heatmaps that elucidate the structural features contributing to toxicity predictions. This transparency not only enhances user confidence but also provides valuable guidance for structural optimization. By supporting multiple data input formats, Cyto-Safe ensures a streamlined experience, catering to both expert and non-expert users.

In summary, Cyto-Safe bridges the gap between advanced predictive capabilities and practical usability, offering a comprehensive solution for cytotoxicity prediction while addressing the interpretability and accessibility limitations of existing models.

## 2.5 CONCLUSIONS

The Cyto-Safe web application demonstrates substantial efficacy in the binary classification of compounds based on cytotoxicity assessments in 3T3 and HEK 293 cells. Cyto-Safe is distinguished by its user-friendly interface, requiring no programming expertise, and its readiness for immediate deployment. Additionally, it offers transparent explanations of prediction outcomes, representing a significant advancement in the accessibility and usability of cytotoxicity assessment tools. The ongoing development of Cyto-Safe will include the expansion of predictive capabilities to encompass additional cell lines as new, high-quality data becomes available.

However, it's important to recognize the limitations of the model's predictive capabilities, as they are influenced by the specific experimental conditions used during training. The reliance on 3T3 and HEK 293 cell lines, along with defined incubation times, restricts the model's generalizability to other cell lines, assays, and exposure durations. To improve its robustness and generalizability, the model should be expanded to include a broader range of datasets, cell lines, assays, and incubation times.

In conclusion, Cyto-Safe is a valuable resource for both the scientific community and industry, facilitating the toxicity assessment of drug candidates. The tool is freely accessible at http://insightai.labmol.com.br/, enabling users to leverage its capabilities to optimize drug development processes.

## 2.6 DATA AVAILABILITY STATEMENT

All molecular structures used for each dataset modelled are provided in the supporting information (.xlsx file). The workflows used to calculate descriptors, split the data, train and validate the models are available on: https://github.com/LabMolUFG/cheminformatics_pipeline. The models are available on: https://github.com/LabMolUFG/cytosafe.

## 2.7 AUTHOR CONTRIBUTIONS

Each author has contributed significantly to this work. CHA acquired funding coordinated, designed, and supervised the project. FLF and VC provided the data curation, modeling, and validation. SSM and JVVBB analyzed the data and discussed the results. IHS performed the mechanistic interpretation. RCB implemented the tool in the server. FLF trained the models, analyzed the results, and wrote the first draft of the manuscript. All authors read, edited, and contributed to the final version of the manuscript.

## 2.8 ACKNOWLEDGMENTS

The authors would like to thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the recognition received through the 21st Edition of the award "Prêmio Destaque na Iniciação Científica e Tecnológica" granted to FLF. This work has been funded by CNPq (grants #440373/2022-0, #140631/2021-6 and #441038/2020-4), FAPEG (#202010267000272), and CAPES (finance code 001). CHA is a CNPq research fellow.

## 2.9 CONFLICT OF INTERESTS

RCB is founder and C.T.O. of InSilicAll, Inc. The remaining authors declare that there are no conflicts of interest.

## 2.10 REFERENCES

(1) Khalef, L.; Lydia, R.; Filicia, K.; Moussa, B. Cell Viability and Cytotoxicity Assays: Biochemical Elements and Cellular Compartments. *Cell Biochem Funct* **2024**, *42* (3). https://doi.org/10.1002/cbf.4007.

(2) Aslantürk, Ö. S. In Vitro Cytotoxicity and Cell Viability Assays: Principles, Advantages, and Disadvantages. In *Genotoxicity - A Predictable Risk to Our Actual World*; InTech, 2018. https://doi.org/10.5772/intechopen.71923.

(3) Sams-Dodd, F. Target-Based Drug Discovery: Is Something Wrong? *Drug Discov Today* **2005**, *10* (2), 139–147. https://doi.org/10.1016/S1359-6446(04)03316-1.

(4) Swinney, D. C. Phenotypic vs. Target-Based Drug Discovery for First-in-Class Medicines. *Clin Pharmacol Ther* **2013**, *93* (4), 299–301. https://doi.org/10.1038/clpt.2012.236.

(5) Riss, T.; Niles, A.; Moravec, R.; Karassina, N.; Vidugiriene, J. Cytotoxicity Assays: In Vitro Methods to Measure Dead Cells. In *Assay Guidance Manual [Internet]*; Bethesda (MD): Eli Lilly & Company and the National Center for Advancing Translational Sciences, 2019.

(6) Clark, A. M.; Dole, K.; Coulon-Spektor, A.; McNutt, A.; Grass, G.; Freundlich, J. S.; Reynolds, R. C.; Ekins, S. Open Source Bayesian Models. 1. Application to ADME/Tox and Drug Discovery Datasets. *J Chem Inf Model* **2015**, *55* (6), 1231–1245. https://doi.org/10.1021/acs.jcim.5b00143.

(7) Braga, R. C.; Alves, V. M.; Silva, M. F. B.; Muratov, E.; Fourches, D.; Lião, L. M.; Tropsha, A.; Andrade, C. H. Pred-hERG: A Novel Web-Accessible Computational Tool for Predicting Cardiac Toxicity. *Mol. Inf.* **2015**, *34* (10), 698–701. https://doi.org/10.1002/minf.201500040.

(8) Sanches, I. H.; Braga, R. C.; Alves, V. M.; Andrade, C. H. Enhancing HERG Risk Assessment with Interpretable Classificatory and Regression Models. *Chem. Res. Toxicol.* **2024**, *37* (6), 910–922. https://doi.org/10.1021/acs.chemrestox.3c00400.

(9) Borba, J. V. B.; Braga, R. C.; Alves, V. M.; Muratov, E. N.; Kleinstreuer, N.; Tropsha, A.; Andrade, C. H. Pred-Skin: A Web Portal for Accurate Prediction of Human Skin Sensitizers. *Chem. Res. Toxicol.* **2021**, *34* (2), 258–267. https://doi.org/10.1021/ACS.CHEMRESTOX.0C00186/SUPPL_FILE/TX0C00186_SI_002.PDF.

(10) Braga, R. C.; Alves, V. M.; Muratov, E. N.; Strickland, J.; Kleinstreuer, N.; Trospsha, A.; Andrade, C. H. Pred-Skin: A Fast and Reliable Web Application to Assess Skin Sensitization Effect of Chemicals. *J. Chem. Inf. Model.* **2017**, *57* (5), 1013–1017. https://doi.org/10.1021/ACS.JCIM.7B00194.

(11) Lee, O. W.; Austin, S.; Gamma, M.; Cheff, D. M.; Lee, T. D.; Wilson, K. M.; Johnson, J.; Travers, J.; Braisted, J. C.; Guha, R.; Klumpp-Thomas, C.; Shen, M.; Hall, M. D. Cytotoxic Profiling of Annotated and Diverse Chemical Libraries Using Quantitative High-Throughput Screening. *SLAS Discov* **2020**, *25* (1), 9–20. https://doi.org/10.1177/2472555219873068.

(12) Fourches, D.; Muratov, E.; Tropsha, A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J Chem Inf Model* **2010**, *50* (7), 1189–1204. https://doi.org/10.1021/ci100176x.

(13) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J Chem Inf Model* **2016**, *56* (7), 1243–1252. https://doi.org/10.1021/ACS.JCIM.6B00129.

(14) Jianping Zhang; Inderjeet Mani. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. In *Workshop on Learning from Imbalanced Data Sets II*; Washington, 2003.

(15) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol Inform* **2010**, *29* (6–7), 476–488. https://doi.org/10.1002/MINF.201000061.

(16) OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*; OECD, 2014. https://doi.org/10.1787/9789264085442-en.

(17) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J Chem Inf Model* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(18) RDKit: Open-Source Cheminformatics. https://www.rdkit.org/.

(19) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; NIPS'17; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp 3149–3157.

(20) Riniker, S.; Landrum, G. A. Similarity Maps - a Visualization Strategy for Molecular Fingerprints and Machine-Learning Methods. *J Cheminform* **2013**, *5* (1), 43. https://doi.org/10.1186/1758-2946-5-43.

(21) Gadaleta, D.; Mangiatordi, G. F.; Catto, M.; Carotti, A.; Nicolotti, O. Applicability Domain for QSAR Models. *IJQSPR* **2016**, *1* (1), 45–63. https://doi.org/10.4018/IJQSPR.2016010102.

(22) Ukelis, U.; Kramer, P.-J.; Olejniczak, K.; Mueller, S. O. Replacement of in Vivo Acute Oral Toxicity Studies by in Vitro Cytotoxicity Methods: Opportunities, Limits and Regulatory Status. *Regul Toxicol Pharmacol* **2008**, *51* (1), 108–118. https://doi.org/10.1016/j.yrtph.2008.02.002.

(23) Niepel, M.; Hafner, M.; Mills, C. E.; Subramanian, K.; Williams, E. H.; Chung, M.; Gaudio, B.; Barrette, A. M.; Stern, A. D.; Hu, B.; Korkola, J. E.; Gray, J. W.; Birtwistle, M. R.; Heiser, L. M.; Sorger, P. K.; Shamu, C. E.; Jayaraman, G.; Azeloglu, E. U.; Iyengar, R.; Sobie, E. A.; Mills, G. B.; Liby, T.; Jaffe, J. D.; Alimova, M.; Davison, D.; Lu, X.; Golub, T. R.; Subramanian, A.; Shelley, B.; Svendsen, C. N.; Ma'ayan, A.; Medvedovic, M.; Feiler, H. S.; Smith, R.; Devlin, K. A Multi-Center Study on the Reproducibility of Drug-Response Assays in Mammalian Cell Lines. *Cell Syst* **2019**, *9* (1), 35–48. https://doi.org/10.1016/j.cels.2019.06.005.

(24) Clothier, R. H. The FRAME Cytotoxicity Test (Kenacid Blue). In *In Vitro Toxicity Testing Protocols*; Humana Press: Totowa, NJ, 1995; pp 109–118. https://doi.org/10.1385/0-89603-282-5:109.

(25) Langdon, S. R.; Mulgrew, J.; Paolini, G. V.; Van Hoorn, W. P. Predicting Cytotoxicity from Heterogeneous Data Sources with Bayesian Learning. *J Cheminform* **2010**, *2* (1). https://doi.org/10.1186/1758-2946-2-11.

(26) Banerjee, P.; Kemmler, E.; Dunkel, M.; Preissner, R. ProTox 3.0: A Webserver for the Prediction of Toxicity of Chemicals. *Nucleic Acids Res* **2024**, *52* (W1), 513–520. https://doi.org/10.1093/nar/gkae303.

(27) Yin, Z.; Ai, H.; Zhang, L.; Ren, G.; Wang, Y.; Zhao, Q.; Liu, H. Predicting the Cytotoxicity of Chemicals Using Ensemble Learning Methods and Molecular Fingerprints. *J. Appl. Toxicol* **2019**, *39* (10), 1366–1377. https://doi.org/10.1002/jat.3785.

(28) Liu, Q.; He, D.; Fan, M.; Wang, J.; Cui, Z.; Wang, H.; Mi, Y.; Li, N.; Meng, Q.; Hou, Y. Prediction and Interpretation Microglia Cytotoxicity by Machine Learning. *J Chem Inf Model* **2024**. https://doi.org/10.1021/acs.jcim.4c00366.

(29) Sun, H.; Wang, Y.; Cheff, D. M.; Hall, M. D.; Shen, M. Predictive Models for Estimating Cytotoxicity on the Basis of Chemical Structures. *Bioorg Med Chem* **2020**, *28* (10). https://doi.org/10.1016/j.bmc.2020.115422.

(30) Webel, H. E.; Kimber, T. B.; Radetzki, S.; Neuenschwander, M.; Nazaré, M.; Volkamer, A. Revealing Cytotoxic Substructures in Molecules Using Deep Learning. *J Comput Aided Mol Des* **2020**, *34* (7), 731–746. https://doi.org/10.1007/s10822-020-00310-4.

# 3 REFERÊNCIAS BIBLIOGRÁFICAS

AHMAD, W. *et al.* ChemBERTa-2: Towards Chemical Foundation Models. 5 set. 2022.

ALVES, V. *et al.* QUIMIOINFORMÁTICA: UMA INTRODUÇÃO. **Química Nova**, 2017.

ANVISA. **RESOLUÇÃO DA DIRETORIA COLEGIADA - RDC Nº 31, DE 29 DE MAIO DE 2014**, 2014.

ANVISA. **INSTRUÇÃO NORMATIVA - IN Nº 63, DE 22 DE JULHO DE 2020**, 2020a.

ANVISA. **RESOLUÇÃO DE DIRETORIA COLEGIADA - RDC Nº 406, DE 22 DE JULHO DE 2020**, 2020b.

ASLANTÜRK, Ö. S. In Vitro Cytotoxicity and Cell Viability Assays: Principles, Advantages, and Disadvantages. *Em*: **Genotoxicity - A Predictable Risk to Our Actual World**. [s.l.] InTech, 2018. .

AVELAR-FREITAS, B. A. *et al.* Trypan blue exclusion assay by flow cytometry. **Brazilian Journal of Medical and Biological Research**, v. 47, n. 4, p. 307–315, 18 mar. 2014.

BAJORATH, J. Computational chemistry in pharmaceutical research: at the crossroads. **Journal of Computer-Aided Molecular Design**, v. 26, n. 1, p. 11–12, 15 jan. 2012.

BLEICHER, K. H. *et al.* Hit and lead generation: beyond high-throughput screening. **Nature Reviews Drug Discovery 2003 2:5**, v. 2, n. 5, p. 369–378, maio 2003.

BORBA, J. V. B. *et al.* Pred-Skin: A Web Portal for Accurate Prediction of Human Skin Sensitizers. **Chem. Res. Toxicol.**, v. 34, n. 2, p. 258–267, 15 fev. 2021.

BROWN, A. C.; FRASER, T. R. On the Connection between Chemical Constitution and Physiological Action; with special reference to the Physiological Action of the Salts of the Ammonium Bases derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. **Journal of Anatomy and Physiology**, v. 2, n. 2, p. 224, 1868.

ÇELIK, T. A. Introductory Chapter: Cytotoxicity. *Em*: **Cytotoxicity**. [s.l.] InTech, 2018. .

CHERKASOV, A. *et al.* QSAR Modeling: Where Have You Been? Where Are You Going To? 2013.

CITELINE. **Pharma R&D Annual Review 2024**, 2024.

COSTA, R. A. P. *et al.* The role of mitochondrial DNA damage in the citotoxicity of reactive oxygen species. **Journal of Bioenergetics and Biomembranes**, v. 43, n. 1, p. 25–29, 1 fev. 2011.

CRAMER, R. D.; PATTERSON, D. E.; BUNCE, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. **Journal of the American Chemical Society**, v. 110, n. 18, p. 5959–5967, 1988.

CROUCH, S. P. M. *et al.* The use of ATP bioluminescence as a measure of cell proliferation and cytotoxicity. **Journal of immunological methods**, v. 160, n. 1, p. 81–88, 1993.

DEVLIN, J. *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference**, v. 1, p. 4171–4186, 11 out. 2018.

EDMONDSON, R. *et al.* Three-Dimensional Cell Culture Systems and Their Applications in Drug Discovery and Cell-Based Biosensors. **Assay and Drug Development Technologies**, v. 12, n. 4, p. 207, 1 maio 2014.

FRESHNEY, R. I. Cytotoxicity. *Em*: **Culture of Animal Cells**. [s.l.] Wiley, 2010. p. 365–381.

GUPTA, P. K. Preclinical toxicological investigations of pharmaceutical products. **Fundamentals of Toxicology**, p. 165–171, 1 jan. 2016.

HAMMETT, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. **Journal of the American Chemical Society**, v. 59, n. 1, p. 96–103, 1 jan. 1937.

HANSCH, C.; FUJITA, T. ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. **Journal of the American Chemical Society**, v. 86, n. 8, p. 1616–1626, 1 abr. 1964.

HASSAN, S. *et al.* Liver-on-a-Chip Models of Fatty Liver Disease. **Hepatology (Baltimore, Md.)**, v. 71, n. 2, p. 733, 1 fev. 2020.

HINKSON, I. V.; MADEJ, B.; STAHLBERG, E. A. Accelerating Therapeutics for Opportunities in Medicine: A Paradigm Shift in Drug Discovery. **Frontiers in Pharmacology**, v. 11, p. 501961, 30 jun. 2020.

HOPFINGER, A. J. *et al.* Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. **Journal of the American Chemical Society**, v. 119, n. 43, p. 10509–10524, 1997.

KLEBE, G.; ABRAHAM, U.; MIETZNER, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. **Journal of medicinal chemistry**, v. 37, n. 24, p. 4130–4146, 1 nov. 1994.

KUMAR, P.; NAGARAJAN, A.; UCHIL, P. D. Analysis of Cell Viability by the Lactate Dehydrogenase Assay. **Cold Spring Harbor protocols**, v. 2018, n. 6, p. 465–468, 1 jun. 2018.

LI, M. S. M.; FILICE, F. P.; DING, Z. A time course study of cadmium effect on membrane permeability of single human bladder cancer cells using scanning electrochemical microscopy. **Journal of Inorganic Biochemistry**, v. 136, p. 177–183, 1 jul. 2014.

LI, R. *et al.* Organ-on-a-chip devices advance to market. **Lab on a Chip**, v. 17, n. 14, p. 2395–2420, 11 jul. 2017.

LIESCHKE, G. J.; CURRIE, P. D. Animal models of human disease: zebrafish swim into view. **Nature Reviews Genetics 2007 8:5**, v. 8, n. 5, p. 353–367, maio 2007.

LOMBARDINO, J. G.; LOWE, J. A. The role of the medicinal chemist in drug discovery — then and now. **Nature Reviews Drug Discovery**, v. 3, n. 10, p. 853–862, out. 2004.

LONGHIN, E. M. *et al.* The alamar blue assay in the context of safety testing of nanomaterials. **Frontiers in Toxicology**, v. 4, p. 981701, 2022.

MENZ, J. *et al.* Genotoxicity assessment: opportunities, challenges and perspectives for quantitative evaluations of dose–response data. **Archives of Toxicology 2023 97:9**, v. 97, n. 9, p. 2303–2328, 5 jul. 2023.

MEYER, H. Zur Theorie der Alkoholnarkose. **Archiv für Experimentelle Pathologie und Pharmakologie**, v. 42, n. 2–4, p. 109–118, 1 mar. 1899.

NEVES, B. J. *et al.* QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. **Frontiers in pharmacology**, v. 9, n. NOV, 13 nov. 2018.

NEVES, B. J. *et al.* Relações quantitativas entre estrutura química e atividade biológica. *Em*: **Fundamentos de química farmacêutica medicinal**. 1. ed. [s.l.] Manole, 2022. .

OECD. **Test No. 129: Guidance Document On Using Cytotoxicity Tests To Estimate Starting Doses For Acute Oral Systemic Toxicity Tests**. [s.l: s.n.].

OECD. **Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models**. [s.l.] OECD, 2014.

OECD. **Guidance Document on the Reporting of Defined Approaches to be Used Within Integrated Approaches to Testing and Assessment**. [s.l.] OECD, 2017.

OVERTON, C. E. **Studien über die Narkose, zugleich ein Beitrag zur allgemeinen Pharmakologie**. Jena: [s.n.].

RAVAL, K. Y.; KANSAGRA, J. J.; GANATRA, T. H. A brief review of high throughput screening in drug discovery process. **Current Trends in Pharmacy and Pharmaceutical Chemistry**, v. 4, n. 3, p. 120–122, 28 ago. 2022.

RAVI, M. *et al.* 3D cell culture systems: advantages and applications. **Journal of cellular physiology**, v. 230, n. 1, p. 16–26, 1 jan. 2015.

RISS, T. *et al.* Cytotoxicity Assays: In Vitro Methods to Measure Dead Cells. *Em*: **Assay Guidance Manual [Internet]**. [s.l.] Bethesda (MD): Eli Lilly & Company and the National Center for Advancing Translational Sciences, 2019. .

ROGERS, D.; HAHN, M. Extended-Connectivity Fingerprints. **Journal of Chemical Information and Modeling**, v. 50, n. 5, p. 742–754, 24 maio 2010.

SANCHES, I. H. *et al.* QSAR models for predicting cardiac toxicity of drugs. **QSAR in Safety Evaluation and Risk Assessment**, p. 351–362, 1 jan. 2023.

SANCHES. Enhancing hERG Risk Assessment with Interpretable Classificatory and Regression Models. **Chemical Research in Toxicology**, v. 37, n. 6, p. 910–922, 17 jun. 2024.

SEAL, S. *et al.* Comparison of Cellular Morphological Descriptors and Molecular Fingerprints for the Prediction of Cytotoxicity- And Proliferation-Related Assays. **Chemical Research in Toxicology**, v. 34, n. 2, p. 422–437, 15 fev. 2021.

SHOU, W. Z. Current status and future directions of high-throughput ADME screening in drug discovery. **Journal of Pharmaceutical Analysis**, v. 10, n. 3, p. 201, 1 jun. 2020.

SIES, H.; BERNDT, C.; JONES, D. P. Oxidative Stress. v. 22, p. 5, 2024.

SINGH, N. *et al.* Drug discovery and development: introduction to the general public and patient groups. **Frontiers in Drug Discovery**, v. 3, p. 1201419, 24 maio 2023.

SOARES, T. A. *et al.* The (Re)-Evolution of Quantitative Structure-Activity Relationship (QSAR) Studies Propelled by the Surge of Machine Learning Methods. **Journal of Chemical Information and Modeling**, v. 62, n. 22, p. 5317–5320, 28 nov. 2022.

STACIARINI, J. H. S. **A Consolidação do Setor Farmacêutico na Economia Global: crescimento, influência, desvios e marketing**. [s.l.] Universidade Federal de Goiás, 2023.

SUN, D. *et al.* Why 90% of clinical drug development fails and how to improve it? **Acta Pharmaceutica Sinica B**, v. 12, n. 7, p. 3049–3062, 1 jul. 2022.

SYLVESTER, P. W. Optimization of the Tetrazolium Dye (MTT) Colorimetric Assay for Cellular Growth and Viability. **Methods in Molecular Biology**, v. 716, p. 157–168, 2011.

TAFT, R. W. Polar and Steric Substituent Constants for Aliphatic and o-Benzoate Groups from Rates of Esterification and Hydrolysis of Esters. **Journal of the American Chemical Society**, v. 74, n. 12, p. 3120–3128, 1 jun. 1952.

TANNENBAUM, J.; BENNETT, B. T. Russell and Burch's 3Rs Then and Now: The Need for Clarity in Definition and Purpose. **Journal of the American Association for Laboratory Animal Science : JAALAS**, v. 54, n. 2, p. 120, 1 mar. 2015.

TODESCHINI, R.; CONSONNI, V. **Handbook of Molecular Descriptors**. [s.l.] Wiley, 2000.

TRAJANOSKA, K. *et al.* From target discovery to clinical drug development with human genetics. **Nature 2023 620:7975**, v. 620, n. 7975, p. 737–745, 23 ago. 2023.

VAN-NORMAN, G. A. Drugs, Devices, and the FDA: Part 1. **JACC: Basic to Translational Science**, v. 1, n. 3, p. 170–179, abr. 2016.

VASWANI, A. *et al.* Attention Is All You Need. **Advances in Neural Information Processing Systems**, v. 2017- December, p. 5999–6009, 12 jun. 2017.

VEDANI, A.; DOBLER, M. 5D-QSAR: the key for simulating induced fit? **Journal of medicinal chemistry**, v. 45, n. 11, p. 2139–2149, 23 maio 2002.

VEDANI, A.; DOBLER, M.; LILL, M. A. Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. **Journal of medicinal chemistry**, v. 48, n. 11, p. 3700–3703, 2 jun. 2005.

XUE, L.; BAJORATH, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. **Combinatorial chemistry & high throughput screening**, v. 3, n. 5, p. 363–372, 4 out. 2000.

# 4 APÊNDICES

## APÊNDICE A - Graphical Abstract

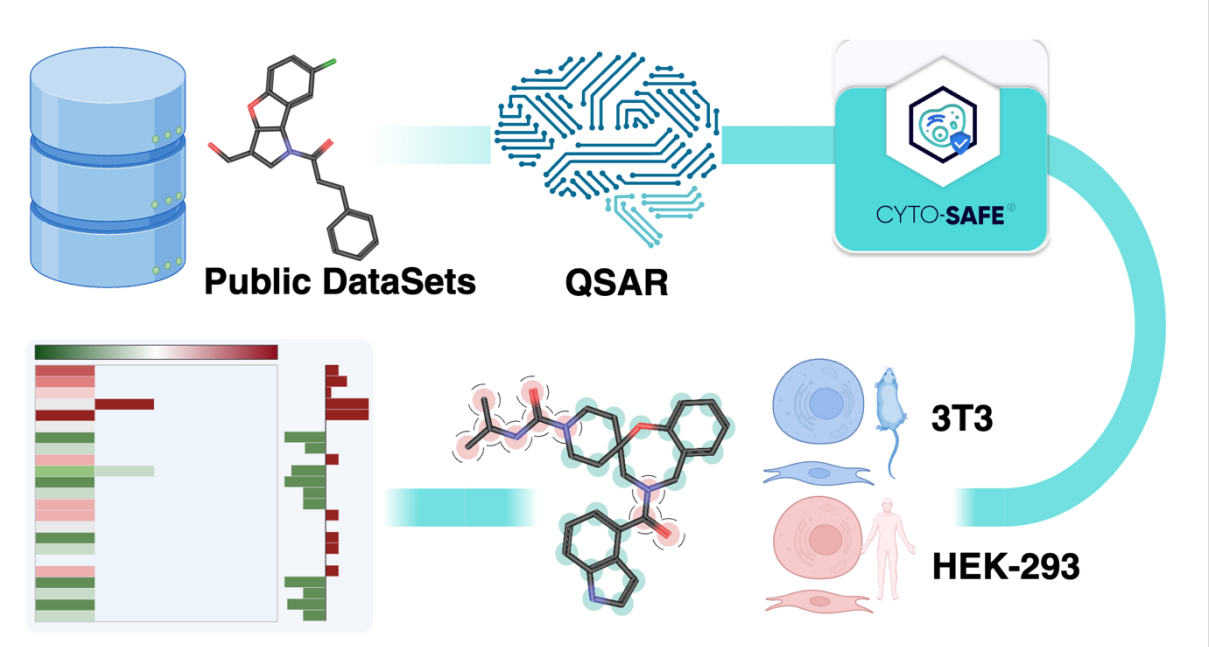


**Figure 3** - Graphical Abstract

## APÊNDICE B - Supporting Information

### Hyperparameters selected

**3T3 model:**

{'boosting_type': 'gbdt',
'class_weight': None,
'colsample_bytree': 0.838657087222196,
'importance_type': 'split',
'learning_rate': 0.058602817860809286,
'max_depth': 256,
'min_child_samples': 20,
'min_child_weight': 0.001,
'min_split_gain': 0.0,
'n_estimators': 3897,
'n_jobs': -1,
'num_leaves': 64,
'objective': None,
'random_state': 0,
'reg_alpha': 1e-09,
'reg_lambda': 0.005217445896529637,
'subsample': 0.8532738462167552,
'subsample_for_bin': 200000,
'subsample_freq': 10,
'reg_sqrt': True}

**HEK 293 model:**

{'boosting_type': 'gbdt',
'class_weight': None,

'colsample_bytree': 0.4365853826089555,
'importance_type': 'split',
'learning_rate': 0.10987760679510937,
'max_depth': 99,
'min_child_samples': 20,
'min_child_weight': 0.001,
'min_split_gain': 0.0,
'n_estimators': 4516,
'n_jobs': -1,
'num_leaves': 392,
'objective': None,
'random_state': 0,
'reg_alpha': 0.05622601988408812,
'reg_lambda': 0.2105202460086872,
'subsample': 0.7476139566402403,
'subsample_for_bin': 200000,
'subsample_freq': 6,
'reg_sqrt': False}

**Activity cliffs analysis**

We performed a clustering analysis to identify pairs of compounds that have a high degree of similarity but belong to different activity classes (e.g., toxic vs. Non-toxic). The methodology employed for clustering and calculating proportions involved several systematic steps. Initially, molecular fingerprints (ECFP4) were computed for all compounds using their SMILES representations to encode structural features. These fingerprints served as input for the clustering algorithm BitBIRCH[1], which grouped structurally similar compounds based on hierarchical subcluster relationships. The resulting clusters were assigned unique identifiers, and each compound was mapped to its respective cluster. Subsequently, the consistency of biological outcomes (cytotoxic or non-cytotoxic) within each cluster was assessed by analyzing the distribution of outcomes. Clusters were classified as having "homogeneous outcomes" when all compounds shared the same activity and as "heterogeneous outcomes" when compounds exhibited both activities. Finally, the proportions of classified clusters were calculated for each cell type, providing insights into the structural similarity and activity distribution of compounds in the dataset.

BitBIRCH uses a new similarity index known as instant similarity (iSIM)[2] to process binary fingerprints, enabling the application of Tanimoto similarity while reducing memory requirements. Instead of utilizing each compound's features directly, BitBIRCH calculates cluster features (CF) and employs them for clustering, enhancing both efficiency and effectiveness in managing large datasets.

As a result, 5,715 homogeneous and 765 heterogeneous clusters were found for 3T3 cells, and HEK293, 9,201 homogeneous and 1,402 heterogeneous clusters. The heterogeneous clusters (probably representing activity cliffs) represent 11.8% and 13.2% of the training set, respectively. These results indicate that while most clusters show uniform outcomes, a small fraction contains compounds with diverging outcomes, reflecting potential activity cliffs and representing challenging regions for machine learning models to generalize, thus enriching the dataset's complexity and reliability for predictive modeling.

**Analysis of Chemical Space**

We conducted a chemical space analysis of the datasets to visualize and interpret the distribution of the chemical compounds. We utilized t-distributed stochastic neighbor embedding (t-SNE)[3], available as a module *sklearn.manifold.TSNE* on scikit-learn[4], to transform the original 1024-bit molecular fingerprint matrix into a two-dimensional coordinate system. This transformation facilitated the visualization of complex, high-dimensional data by reducing it to a more interpretable form. We then plotted these coordinates into a scatter plot to examine the spatial distribution and clustering of compounds. To assess the impact of data representation on the analysis, we tested the different under sampling ratios conducted previously. This approach allowed us to evaluate how varying the sample size affected the clustering patterns and overall visualization of the chemical space. Figures S1 and S2 show the chemical space analysis of the 3T3 and HEK 292 datasets, respectively.

## Supplementary Figures

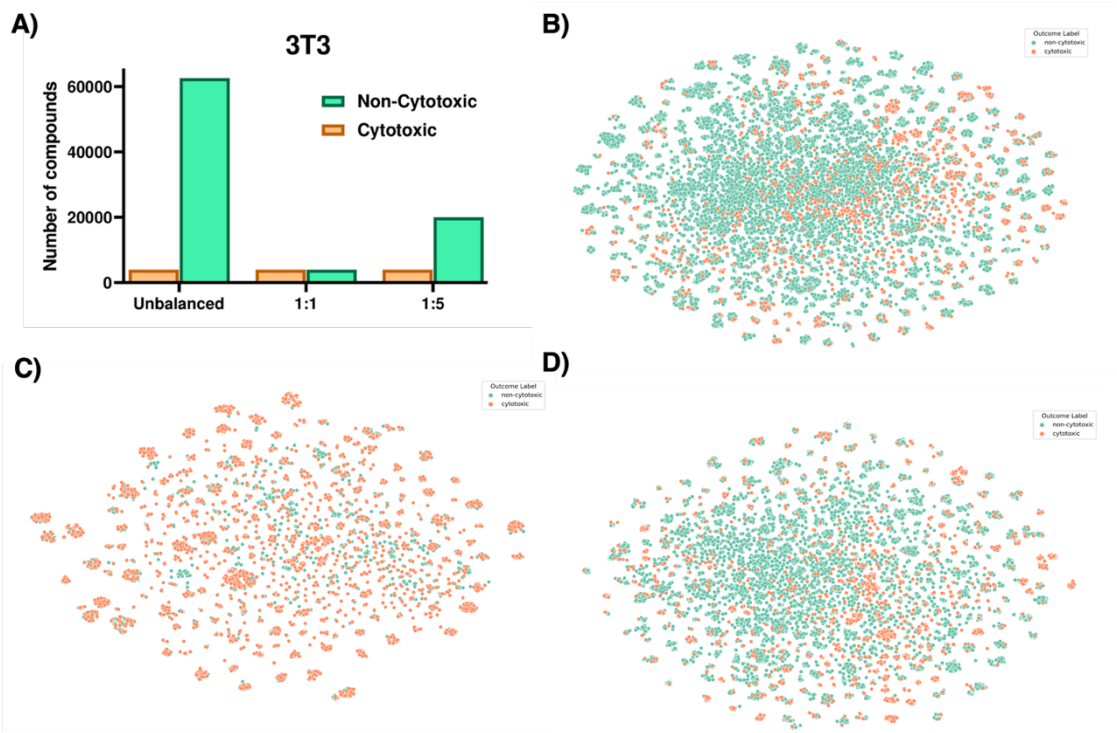


**Figure S 1 -** Chemical space analysis of 3T3 datasets. A) Distribution of compounds; B) Visualization of unbalanced data; C) Visualization of a 1:1 under-sampling proportion; D) Visualization of a 1:5 under-sampling proportion. Note: Green indicates non-cytotoxic compounds, while orange represents cytotoxic compounds.

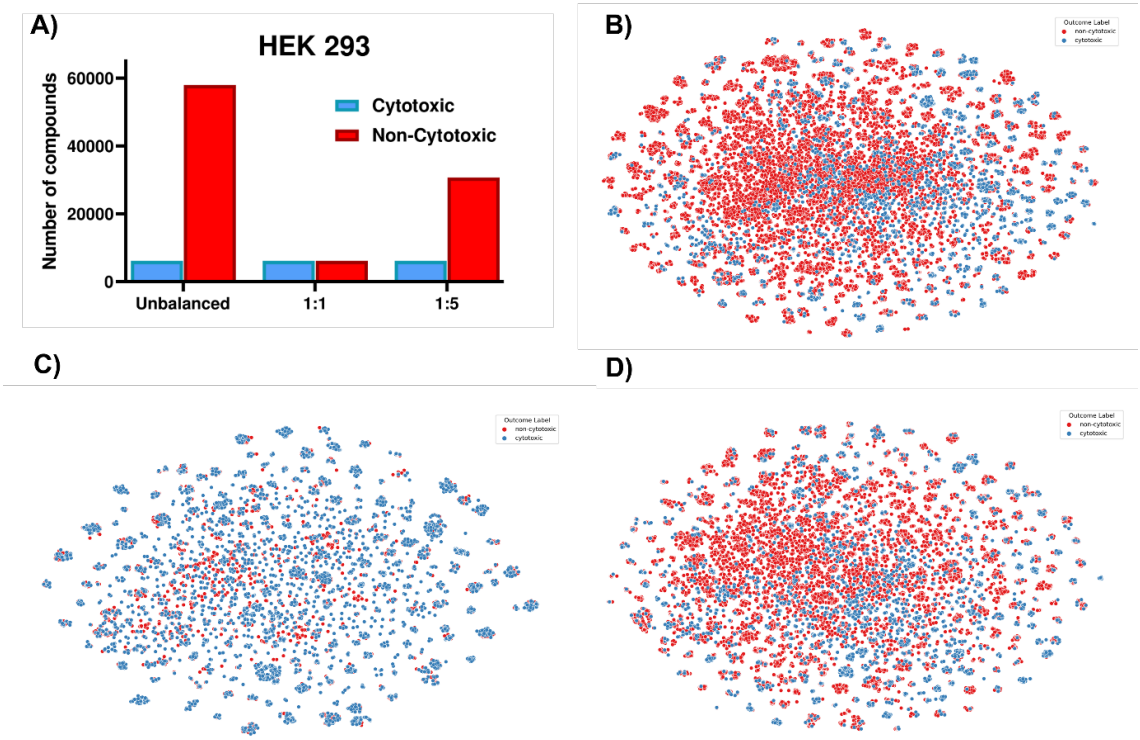


**Figure S 2 -** Chemical space analysis of HEK 293 datasets. A) Distribution of compounds; B) Visualization of unbalanced data; C) Visualization of a 1:1 under-sampling proportion; D) Visualization of a 1:5 under-sampling proportion. Note: Green indicates non-cytotoxic compounds, while orange represents cytotoxic compounds.

## Structural Diversity and Applicability Domain Threshold

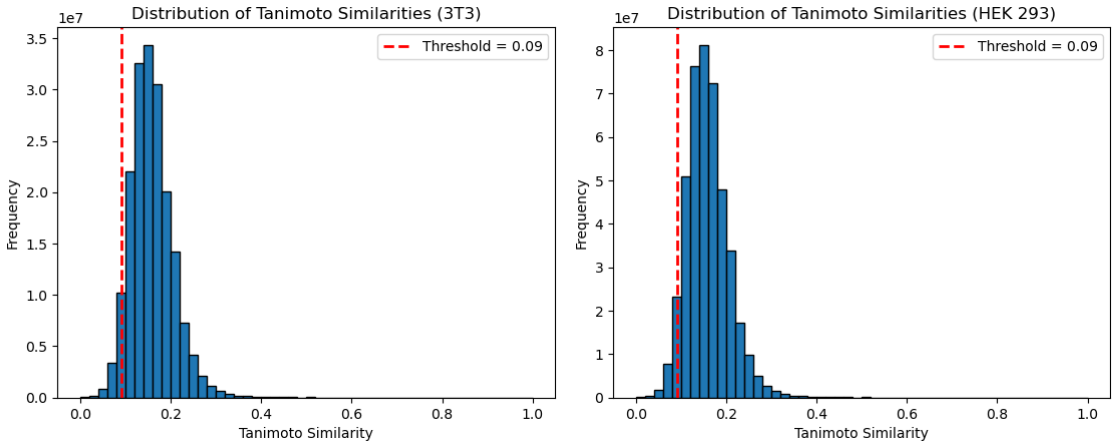


**Figure S 3 -** Distribution of Tanimoto Similarities within the training set for compounds tested on 3T3 (left) and HEK-293 (right) cell lines. The red dashed line represents the threshold (Tanimoto similarity = 0.09) corresponding to the 5$^{th}$ percentile of the similarity distribution.

## Explainable AI (XAI) heatmaps for the case study

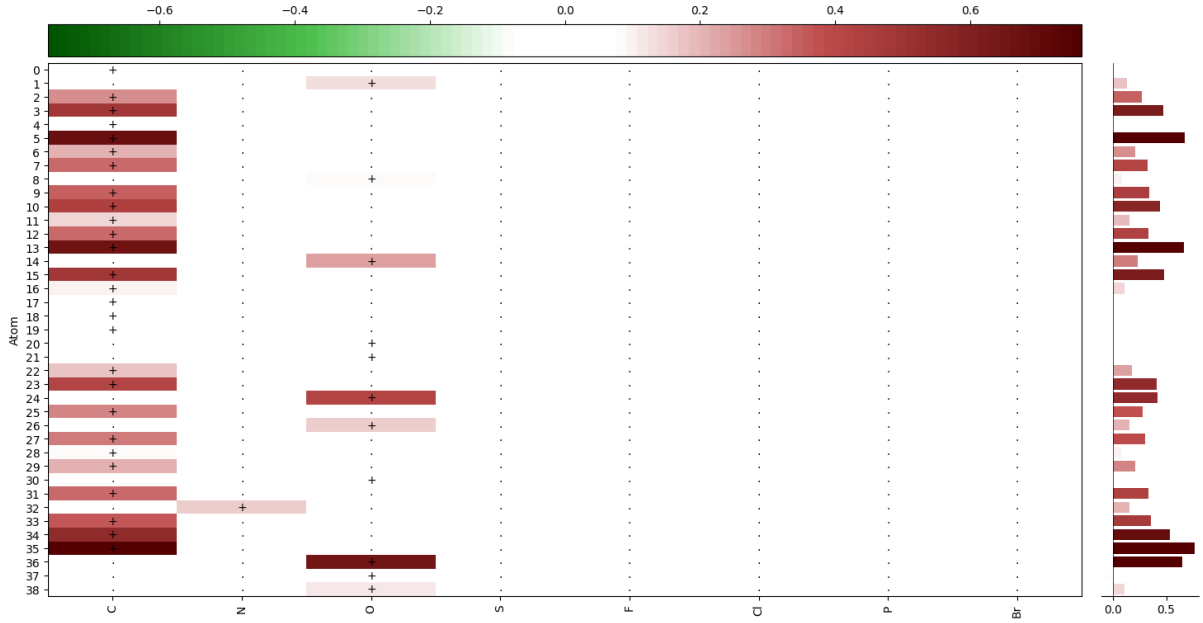


**Figure S 4 -** Heatmap contribution of 3T3 model's prediction for Doxorubicin. Atoms with a positive influence on cytotoxicity are highlighted in shades of red, while atoms with a positive influence on non-cytotoxicity are highlighted in shades of green. The intensity of the color corresponds to the strength of the influence. The *y*-axis corresponds to the atom ID, and the *x*-axis represents the atom type.

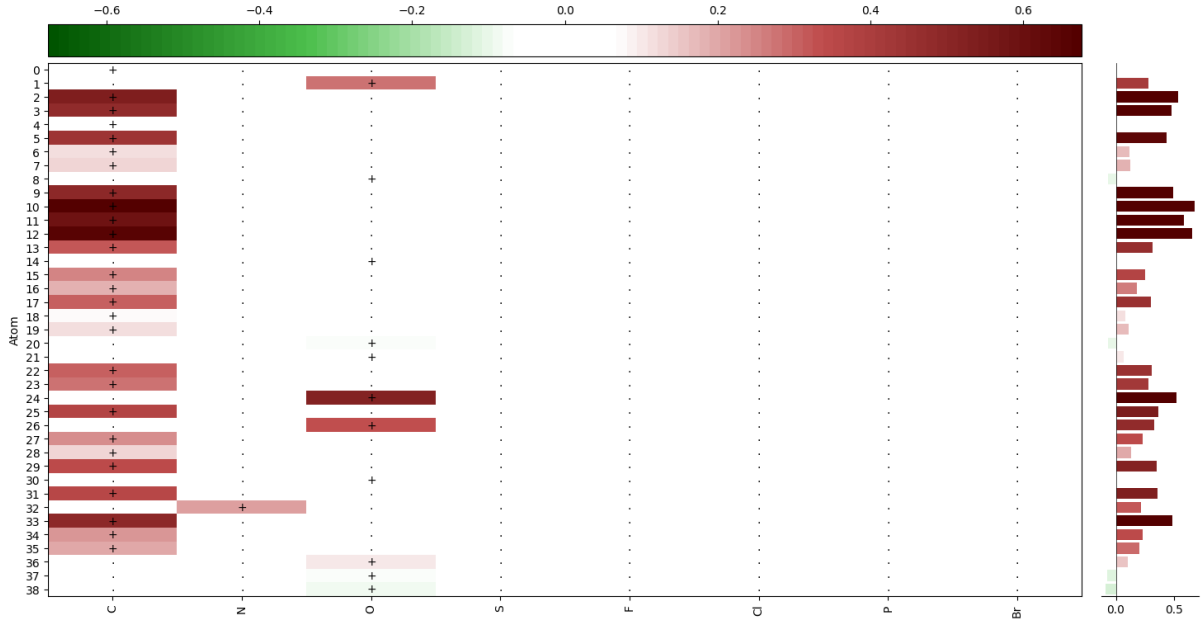


**Figure S 5 -** Heatmap contribution of HEK 293 model's prediction for Doxorubicin. Atoms with a positive influence on cytotoxicity are highlighted in shades of red, while atoms with a positive influence on non-cytotoxicity are highlighted in shades of green. The intensity of the color corresponds to the strength of the influence. The *y*-axis corresponds to the atom ID, and the *x*-axis represents the atom type.

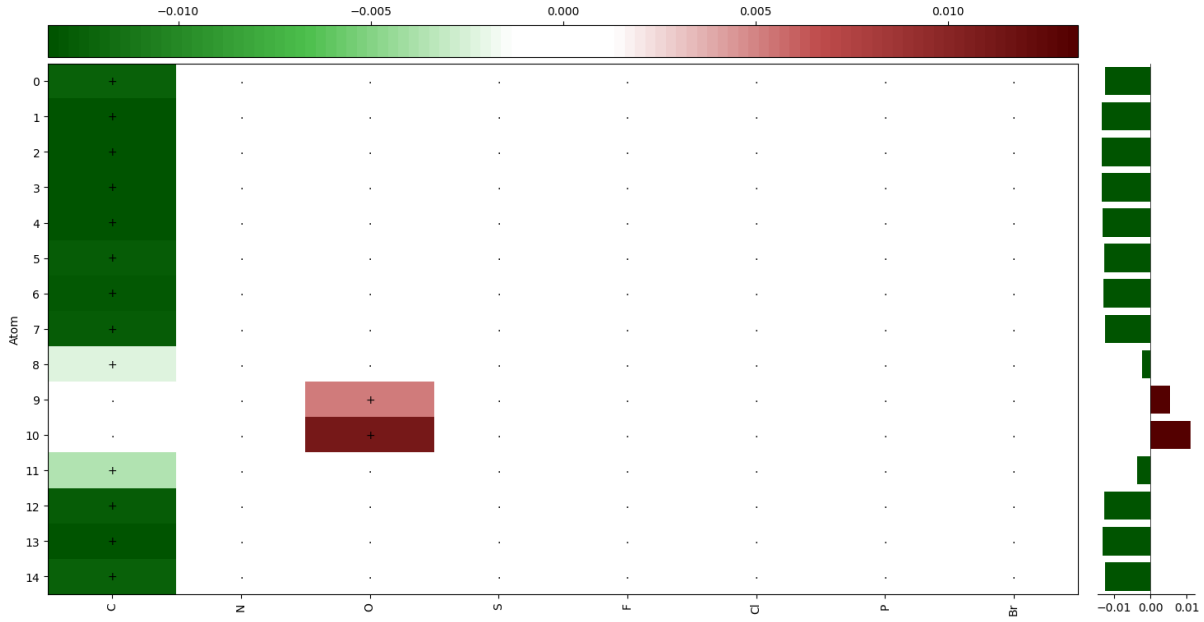


**Figure S 6 -** Heatmap contribution of 3T3 model's prediction for Ibuprofen. Atoms with a positive influence on cytotoxicity are highlighted in shades of red, while atoms with a positive influence on non-cytotoxicity are highlighted in shades of green. The intensity of the color corresponds to the strength of the influence. The *y*-axis corresponds to the atom ID, and the *x*-axis represents the atom type.

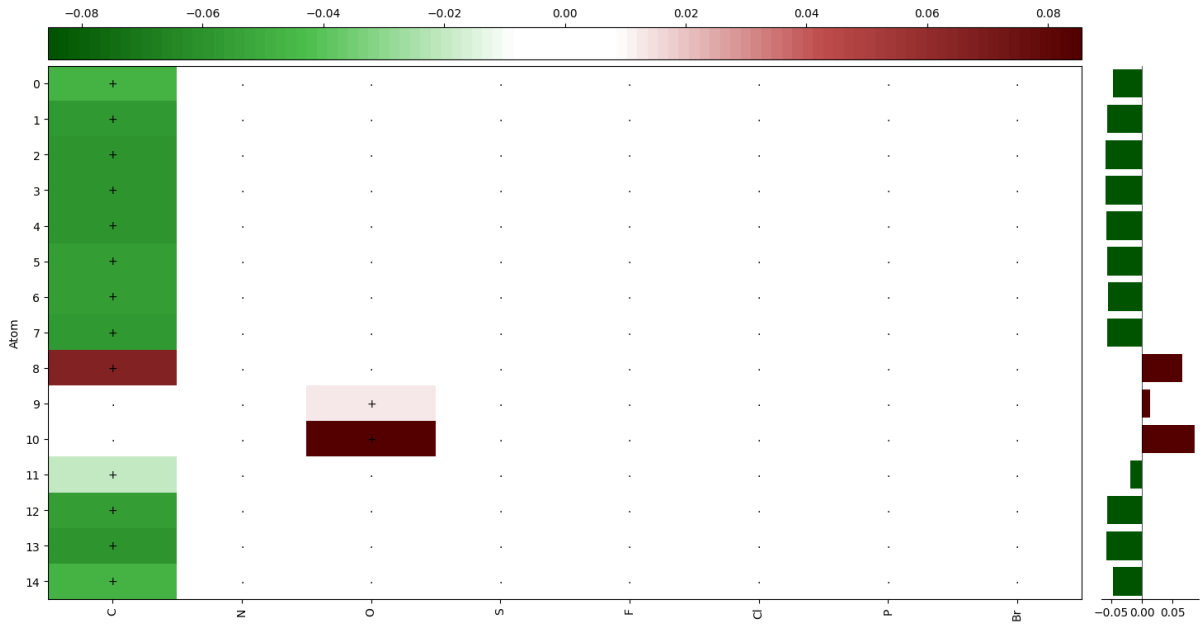


**Figure S 7 -** Heatmap contribution of HEK 293 model's prediction for Ibuprofen. Atoms with a positive influence on cytotoxicity are highlighted in shades of red, while atoms with a positive influence on non-cytotoxicity are highlighted in shades of green. The intensity of the color corresponds to the strength of the influence. The *y*-axis corresponds to the atom ID, and the *x*-axis represents the atom type.

## Supplementary References

(1) Pérez, K. L.; Jung, V.; Chen, L.; Huddleston, K.; Miranda-Quintana, R. A. Efficient Clustering of Large Molecular Libraries. August 10, 2024. https://doi.org/10.1101/2024.08.10.607459.

(2) Pérez, K. L.; Kim, T. D.; Miranda-Quintana, R. A. ISIM: Instant Similarity. *Digit Discov* **2024**, *3* (6), 1160–1171. https://doi.org/10.1039/D4DD00041B.

(3) van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *Journal of Machine Learning Research* **2008**, *9* (86), 2579–2605.

(4) Scikit-Learn TSNE Module. https://scikit-learn.org/dev/modules/generated/sklearn.manifold.TSNE.html.

# 5 ANEXO

## ANEXO A - Regras de submissão do *Journal of Chemical Information and Modeling*

**Manuscript Type**

Application Notes: Application notes are informative peer-reviewed reports on novel software packages, databases, and web servers. Submissions should be no longer than 5000 words and contain at least one figure. This word count includes the abstract, text, and graphics only. References do not count towards the word limit. You may approximate the length of figures, schemes, and tables by counting single-column images as 300 words and double-column as 600 words (this count assumes that they will occupy ¼ and ½ page, respectively, at final production size). If a table was created using the Tables function in Word, it will be included in the word count; subtract the word count for the table and estimate the space as you would for a scheme or figure. Authors will be expected to abide by this rule and to submit a statement from the corresponding author indicating the word count of the article and how it was obtained. Articles with word counts over 5000 words will be returned to the authors for editing before going through the review process. The name of the application being described should be clearly stated in the manuscript title. The scientific, technical, or other usability advancements of the software should be clearly described. As far as possible, the software packages and web servers should be operating system agnostic (Windows, Linux, and OSX). All submitted manuscripts will be reviewed by an editor prior to being sent out for peer-review. The software should be generally available for evaluation or purchase at the time of publication for academic and commercial use. The software must be made available for testing by reviewers to address specific data or claims in the manuscript, upon request by reviewers, while preserving reviewer anonymity. Articles viewed to be unsuitable for the journal or inconsistent with the above guidelines will be returned after editorial review.

**Manuscript Preparation**

**Submit with Fast Format**

All ACS journals and partner journals have simplified their formatting requirements in favor of a streamlined and standardized format for an initial manuscript submission. Read more about the requirements and the benefits these serves authors and reviewers here.

Manuscripts submitted for initial consideration must adhere to these standards:

- Submissions must be complete with clearly identified standard sections used to report original research, free of annotations or highlights, and include all numbered and labeled components.
- Figures, charts, tables, schemes, and equations should be embedded in the text at the point of relevance. Separate graphics can be supplied later at revision, if necessary.
- When required by a journal's structure or length limitations, manuscript templates should be used.
- References can be provided in any style, but they must be complete, including titles. For information about the required components of different reference types, please refer to the ACS Style Quick Guide.
- Supporting Information must be submitted as a separate file(s).

The complete and up to date Guidelines can be accessed on: https://researcher-resources.acs.org/publish/author_guidelines?coden=jcisd8