

UNIVERSIDADE FEDERAL DE GOIÁS  
FACULDADE DE FARMÁCIA  
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS FARMACÊUTICAS

VINÍCIUS DE MEDEIROS ALVES

**Desenvolvimento de modelos de QSAR e análise  
quimioinformática da sensibilização e permeabilidade da pele**

Goiânia  
2014

VINÍCIUS DE MEDEIROS ALVES

**Desenvolvimento de modelos de QSAR e análise  
quimioinformática da sensibilização e permeabilidade da pele**

Dissertação apresentada ao Programa de Pós-Graduação em Ciências Farmacêuticas da Faculdade de Farmácia da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre em Ciências Farmacêuticas.

Área de Concentração: Fármacos e Medicamentos

Orientadora: Profa. Dra. Carolina Horta Andrade

Goiânia

2014

## Folha de Aprovação

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciências Farmacêuticas da Universidade Federal de Goiás, em 17 de março de 2014, pelo mestrando Vinícius de Medeiros Alves.

Banca Examinadora:



---

Prof. Dra. Carolina Horta Andrade (FE/UFG)

*Presidente*



---

Prof. Dra. Elizabeth Igne Ferreira (USP)



---

Prof. Dr. Ademir J. Camargo (IQ/UEG)

## DEDICATÓRIA

*Aos meus pais eu dedico este trabalho pelos exemplos de conduta moral, simplicidade e persistência; por todo esforço e sacrifício que tiveram para me dar a oportunidade de seguir meus sonhos e realizar minhas conquistas.*

## AGRADECIMENTOS

*À Profa. Dra. Carolina Horta Andrade, pelo aprendizado, estímulo constante, confiança, suporte, orientação, amizade e infinitas oportunidades, meu eterno agradecimento.*

*Ao MSc. Rodolpho de Campos Braga, pelas orientações, discussões científicas, suporte, confiança e amizade.*

*Ao Prof. Dr. Alexander Tropsha, pela oportunidade de estágio e aprendizado em seu laboratório na University of North Carolina (UNC).*

*Aos Drs. Eugene Muratov e Denis Fourches, pelo aprendizado, discussões, broncas e amizade.*

*Às Dras. Judy Strickland e Nicole Kleinstreuer pelas contribuições ao trabalho.*

*À todos do LabMol, pelas discussões científicas, apoio e amizade.*

*Aos meus pais, Nivaldo e Neide, e minha irmã, Lívia, por sempre me apoiarem nas minhas decisões e por acreditarem em mim.*

*Aos meus grandes amigos Jivago, Túlio, Murilo, Fernanda, Juarez, entre vários outros, simplesmente por serem meus amigos.*

*À Rebeca Tipple, por todo carinho, parceria, paciência e compreensão nessa reta final.*

*À todos da Technicolor, por nunca deixarem meu espírito adolescente morrer.*

*Ao Urso, pela companhia marota durante (quase) todo tempo que trabalhei em casa.*

*Aos amigos que fiz durante a graduação e que fazem questão de manter a amizade.*

*À todos os professores e funcionários da Faculdade de Farmácia da UFG.*

*À CAPES, FAPEG e CNPq, pelo apoio financeiro.*

*À todos que, direta ou indiretamente, contribuíram para a realização deste trabalho.*

## EPÍGRAFE

*“Extraordinary claims require extraordinary evidence.”*

*Carl Sagan*

## SUMÁRIO

<b>LISTA DE FIGURAS.....</b>	<b>X</b>
<b>LISTA DE TABELAS.....</b>	<b>XI</b>
<b>LISTA DE ABREVIATURAS E SIGLAS .....</b>	<b>XII</b>
<b>RESUMO.....</b>	<b>XV</b>
<b>ABSTRACT .....</b>	<b>XVI</b>
<b>1 INTRODUÇÃO .....</b>	<b>17</b>
<b>1.1 Panorama atual da indústria farmacêutica .....</b>	<b>17</b>
<b>1.2 Planejamento e desenvolvimento de novos fármacos.....</b>	<b>17</b>
<b>1.3 Química medicinal e o planejamento racional de novos fármacos .....</b>	<b>20</b>
<b>1.4 Triagem virtual: além da farmacodinâmica .....</b>	<b>23</b>
<b>1.5 Relações quantitativas entre estrutura química e atividade/propriedade (QSAR/QSPR).....</b>	<b>25</b>
1.5.1 História e evolução .....	25
1.5.2 Aplicações .....	28
1.5.3 Princípios.....	29
1.5.4 Descritores moleculares.....	30
1.5.5 Métodos de aprendizado de máquina .....	32
1.5.6 Boas práticas de desenvolvimento e validação.....	34
1.5.6.1 Preparo do conjunto de dados .....	34
1.5.6.2 Detecção e remoção de outliers antes da geração dos modelos.....	35
1.5.6.3 Validação dos modelos.....	36
1.5.6.4 Definição do domínio de aplicabilidade .....	36
<b>1.6 Sensibilização da pele .....</b>	<b>37</b>
1.6.1 Avaliação experimental da sensibilização da pele .....	39
1.6.2 Revisão bibliográfica de modelos de QSAR a respeito da sensibilização da pele .....	40
<b>1.7 Permeabilidade da pele .....</b>	<b>46</b>
<b>2 JUSTIFICATIVA E OBJETIVOS .....</b>	<b>52</b>
<b>2.1 Justificativa e Objetivo Geral.....</b>	<b>52</b>
<b>2.2 Objetivos específicos.....</b>	<b>52</b>
<b>3 MATERIAIS E MÉTODOS .....</b>	<b>53</b>
<b>3.1 Conjuntos de dados .....</b>	<b>53</b>
3.1.1 Conjunto de dados de sensibilização da pele (conjunto de dados A).....	53

3.1.2	Conjunto de validação externa para sensibilização da pele (conjunto de dados B) .....	53
3.1.3	Conjunto de dados da Scorecard (conjunto de dados C) .....	53
3.1.4	Conjunto de dados da permeabilidade da pele humana (conjunto de dados D) .....	54
3.1.5	Conjunto de dados da permeabilidade da pele de roedores (conjunto de dados E).....	54
<b>3.2</b>	<b>Preparo do conjunto de dados .....</b>	<b>54</b>
<b>3.3</b>	<b>Cálculo dos descritores.....</b>	<b>55</b>
3.3.1	Descritores DRAGON .....	55
3.3.2	Descritores SiRMS .....	55
<b>3.4</b>	<b>Análise de agrupamentos hierárquicos .....</b>	<b>55</b>
<b>3.5</b>	<b>Análise do vizinho mais próximo .....</b>	<b>56</b>
<b>3.6</b>	<b>Geração dos modelos de QSAR.....</b>	<b>56</b>
<b>3.7</b>	<b>Random Forest.....</b>	<b>59</b>
<b>3.8</b>	<b>OECD QSAR Toolbox .....</b>	<b>60</b>
<b>4</b>	<b>RESULTADOS E DISCUSSÃO .....</b>	<b>61</b>
<b>4.1</b>	<b>Sensibilização da pele .....</b>	<b>61</b>
4.1.1	Caracterização dos conjuntos de dados .....	61
4.1.1.1	<i>Conjunto de dados de sensibilização da pele (conjunto de dados A) .....</i>	<i>61</i>
4.1.1.2	<i>Conjunto de validação externa para sensibilização da pele (conjunto de dados B).....</i>	<i>62</i>
4.1.1.3	<i>Conjunto de dados da Scorecard (conjunto de dados C).....</i>	<i>62</i>
4.1.2	Análise de agrupamentos hierárquicos do conjunto de dados A .....	62
4.1.3	Modelos de QSAR gerados para sensibilização da pele.....	66
4.1.4	Comparação dos modelos de QSAR desenvolvidos para sensibilização da pele com o algoritmo de sensibilização do QSAR Toolbox .....	69
4.1.5	Triagem virtual do conjunto de dados da Scorecard .....	82
<b>4.2</b>	<b>Permeabilidade da pele .....</b>	<b>88</b>
4.2.1	Caracterização dos conjuntos de dados .....	88
4.2.1.1	<i>Conjunto de dados da permeabilidade da pele humana (conjunto de dados D) .....</i>	<i>88</i>
4.2.1.2	<i>Conjunto de dados da permeabilidade da pele de roedores (conjunto de dados E).....</i>	<i>88</i>
4.2.2	Análise da sobreposição dos dados de permeabilidade humana e de roedores .....	89
4.2.3	Análise da sobreposição dos dados de sensibilização e permeabilidade.....	89
4.2.4	Análise de agrupamentos hierárquicos do conjunto de dados D .....	91
4.2.5	Modelos de QSAR gerados para permeabilidade da pele .....	93
4.2.6	Relação entre o potencial sensibilizador e a permeabilidade de compostos químicos usando dados atribuídos .....	94

<b>5</b>	<b>CONCLUSÃO.....</b>	<b>97</b>
	<b>REFERÊNCIAS BIBLIOGRÁFICAS .....</b>	<b>98</b>
	<b>ANEXOS .....</b>	<b>117</b>

## LISTA DE FIGURAS

<b>Figura 1.</b> Esquema simplificado do processo de planejamento e desenvolvimento de um novo fármaco (modificado de GUIDO, 2008 e LOMBARDINO; LOWE, 2004). .....	18
<b>Figura 2.</b> Identificação, seleção e otimização de moléculas bioativas (fase pré-clínica) (modificado de BELFIELD; DELANEY, 2006 e GUIDO, 2008). .....	19
<b>Figura 3.</b> Esquema representando a geração de um modelo de QSAR (modificado de TROPSHA, 2010). .....	30
<b>Figura 4.</b> Fluxo de trabalho para preparo de conjunto de dados químicos (modificado de Fourches, Muratov e Tropsha, 2010). .....	35
<b>Figura 5.</b> Via de efeito adverso da sensibilização da pele (modificado de KARLBERG et al., 2008 e OECD, 2012). .....	38
<b>Figura 6.</b> Fluxo de trabalho para desenvolvimento dos modelos de QSAR (modificado de Tropsha (2010). .....	57
<b>Figura 7.</b> Resumo do protocolo realizado para comparação dos modelos gerados nesse estudo com o módulo de sensibilização da pele do QSAR Toolbox. ....	60
<b>Figura 9.</b> Correlação entre os dados experimentais de permeabilidade de pele humana vs. de roedores ( $r^2 = 0,44$ ). .....	89
<b>Figura 10.</b> Análise agrupamentos hierárquicos para o conjunto de dados D: dendograma e mapa de calor da matriz de distâncias organizadas com base na similaridade estrutural (violeta/azul = similar; amarelo, vermelho = dissimilar). .....	91
<b>Figura 11.</b> Dados experimentais e preditos de permeabilidade vs. sensibilização, contendo 174 sensibilizadores e 109 não sensibilizadores. ....	95
<b>Figura 12.</b> Dados experimentais e preditos de permeabilidade vs. sensibilização para o grupo 1, contendo 4 sensibilizadores e 7 não sensibilizadores. ....	96

## LISTA DE TABELAS

<b>Tabela 1.</b> Representação genérica de uma matriz de dados para um estudo de QSAR/QSPR. .....	32
<b>Tabela 2.</b> Estudos prévios de QSAR para sensibilização da pele.....	41
<b>Tabela 3.</b> Estudos prévios de QSAR para permeabilidade da pele.....	47
<b>Tabela 4.</b> Par de estruturas duplicadas encontradas no estudo de Chauhan e Shakya (2010). 51	
<b>Tabela 5.</b> Número de compostos testados em LLNA por veículo.....	61
<b>Tabela 6.</b> Exemplos de compostos “suspeitos” no conjunto de dados A identificados por HCA. .....	64
<b>Tabela 7.</b> Valores estatísticos dos modelos de QSAR gerados para sensibilização da pele avaliados pelo método de validação externa <i>5-fold</i> .....	67
<b>Tabela 8.</b> Valores estatísticos dos modelos de QSAR para sensibilização da pele quando aplicados a um conjunto de validação externa (133 sensibilizadores do conjunto de dados A + conjunto de dados B). .....	69
<b>Tabela 9.</b> Comparação entre os modelos de QSAR para sensibilização da pele e o QSAR Toolbox.....	71
<b>Tabela 10.</b> Vizinho mais próximo dos sensibilizadores erroneamente preditos pelo modelo 5. .....	73
<b>Tabela 11.</b> Vizinhos mais próximos dos potenciais sensibilizadores preditos pelo modelo 7.83	
<b>Tabela 12.</b> Dados experimentais de sensibilização e permeabilidade da pele.....	90
<b>Tabela 13.</b> Exemplos de compostos “suspeitos” no conjunto de dados D identificados por HCA. .....	92
<b>Tabela 14.</b> Valores estatísticos dos modelos de QSAR gerados para permeabilidade da pele humana e de roedores avaliados pelo método de validação externa <i>5-fold</i> .....	94
<b>Tabela A1.</b> Lista de compostos da Scorecard e predições correspondentes feitas pelos modelos 5 e 7 em arquivo SDF. ....	117
<b>Tabela A2.</b> Conjunto de dados experimentais e preditos de sensibilização e permeabilidade para 283 compostos. ....	117
<b>Tabela A3.</b> Dados de sensibilização e permeabilidade experimentais e preditos para 11 compostos identificados por HCA.....	117

## LISTA DE ABREVIATURAS E SIGLAS

2D	Bidimensional ou segunda dimensão
3D	Tridimensional ou terceira dimensão
Å	Angstrom
AC	Acurácia
ACB	Acurácia balanceada
AC <sub>LOO</sub>	Acurácia baseada no método <i>leave-one-out</i>
ADME/Tox	Absorção, distribuição, metabolismo, excreção e toxicidade
ANN	<i>Artificial neural networks</i>
CoMFA	<i>Comparative Molecular Fields Analysis</i>
CoMSIA	<i>Comparative Molecular Similarity Index Analysis</i>
CT	Coefficiente de Tanimoto
D	Difusibilidade pela membrana
DA	Domínio de aplicabilidade
DPM	Desintegrações por minuto
EC <sub>3</sub>	Concentração necessária para produzir um índice de estímulo $\geq 3$
ES	Especificidade
FDA	<i>Food and Drug Administration</i>
FL	Fator de lipofilicidade
GPMT	<i>Guinea pig maximization test</i>
GPR	Gaussian process regression
H <sub>a</sub>	Aceptor de ligação de hidrogênio
H <sub>b</sub>	Habilidade de formar ligações de hidrogênio
HCA	<i>Hierarchical clustering analysis</i>
H <sub>d</sub>	Doador de ligação de hidrogênio
HiT QSAR	<i>Hierarchical QSAR technology</i>
HTS	<i>High-throughput screening</i>
ICCVAM	<i>Interagency Coordinating Committee on the Validation of Alternative Methods</i>
K <sub>ow</sub> ou P	Partição octanol/água
LBDD	<i>Ligand-based drug design</i>
LDA	<i>Linear discriminant analysis</i>
LLNA	<i>Local lymph node assay</i>

logK <sub>p</sub>	Coeficiente de permeabilidade (cm/h)
LR	<i>Logistic regression</i>
LSRA	<i>Least-square regression analysis</i>
MAE	<i>Mean absolute error</i>
MM	Massa molecular
MLP	<i>Multilayer perceptron</i>
MLR	<i>Multiple linear regression</i>
MRA	<i>Multivariate regression analysis</i>
MSE	<i>Mean squared error</i>
NCE	<i>New chemical entities</i>
OECD	<i>Organization for Economic Co-operation and Development</i>
OOB	<i>Out-of-bag</i>
PCA	<i>Principal component analysis</i>
PCR	<i>Principal component regression</i>
PD&I	Pesquisa, desenvolvimento e inovação
PF	Ponto de fusão
PLSR	<i>Partial least squares regression</i>
PSO	<i>Particle swarm optimization</i>
$q^2$	Coeficiente de correlação da validação cruzada
QSAR	<i>Quantitative-structure-activity relationship</i>
RF	<i>Random forest</i>
(r)LLNA	<i>(Reduced) local lymph node assay</i>
RM	Refratividade molar
RMSE	<i>Root mean squared error</i>
RPT	<i>Recursive partitioning tree</i>
S(M)RA	<i>Stepwise (multiple) regression analysis</i>
SALI	<i>Structure-activity landscape index</i>
SAM1	Modelo 1 <i>ab initio</i> semiempírico
SAR	<i>Structure-activity relationship</i>
SBDD	<i>Structure-based drug design</i>
SDA	<i>Stepwise discriminant analysis</i>
SE	Sensibilidade
SiRMS	<i>Simplex representation of molecular structure</i>

SVM	<i>Support vector machine</i>
TFN	Taxa de falsos negativos
TFP	Taxa de falsos positivos
VM	Volume molecular
VP	Verdadeiros positivos
VPN	Valor de preditividade negativa
VPP	Valor de preditividade positive
VN	Verdadeiros negativos
$\delta$	Difusão pelo comprimento do caminho

## RESUMO

A exposição repetida a agentes químicos pode induzir a sensibilização da pele em indivíduos inerentemente suscetíveis e desencadear uma resposta imunológica exacerbada. Apesar de muitos compostos químicos estarem implicados na sensibilização da pele, existem poucos estudos analisando as relações entre a estrutura molecular e o potencial sensibilizador desses compostos, incluindo a conexão com a permeabilidade pela pele, a qual é referida como sendo primordial para o processo de sensibilização. Neste estudo foram compilados, integrados e preparados os maiores conjuntos de dados disponíveis publicamente relacionados tanto com a sensibilização da pele quanto à permeabilidade. Inesperadamente, não se encontrou correlação entre essas duas propriedades. Modelos de QSAR robustos e preditivos foram gerados e validados para ambas as propriedades usando um fluxo de trabalho totalmente complacente com as recomendações da OECD. As taxas de acerto dos modelos discriminaram estruturas sensibilizadoras de não sensibilizadoras com uma taxa de 0,68-0,88 de sucesso, quando avaliadas em vários conjuntos de validação externa. Quando comparados com o módulo de sensibilização da pele implementado na ferramenta QSAR Toolbox da OECD, os modelos tiveram baixa cobertura do espaço químico, mas precisão preditiva mais elevada para os mesmos conjuntos de compostos externos avaliados pelo valor de preditividade positiva e valor de preditividade negativa assim como pela acurácia balanceada. O coeficiente de correlação cruzada entre os dados de permeabilidade da pele humana e de roedores apresentou-se baixo ( $r^2=0,44$ ); assim, apenas o conjunto de dados de pele humana foi considerado para geração de modelos de permeabilidade, que apresentaram precisão externa de  $q^2_{\text{ext}} = 0,87$  (para 62% dos compostos dentro do domínio de aplicabilidade). Modelos de sensibilização da pele foram empregados para identificação de toxicantes putativos no banco de dados de possíveis agentes toxicantes da Scorecard que podem ser considerados como candidatos para validação experimental.

Palavras-chave: QSAR, sensibilização da pele, permeabilidade da pele, triagem virtual.

## ABSTRACT

Repetitive exposure to a chemical agent can induce an immune reaction in inherently susceptible individuals that leads to skin sensitization. Although many chemicals have been reported as skin sensitizers, there have been very few reports analyzing the relationships between their molecular structure and the sensitization potential including the connection to skin permeability, which is widely considered to be mechanistically implicated in sensitization. In this study, we have compiled, curated, and integrated the largest publicly available datasets related to chemically-induced skin sensitization and skin permeability. Unexpectedly, no correlation between sensitization and permeability has been found. Predictive QSAR models have been developed and validated for both skin sensitization and skin permeability using a standardized workflow fully compliant with the OECD guidelines. The classification accuracies of QSAR models discriminating sensitizers from non-sensitizers were 0.68-0.88 when evaluated on several external validation sets. When compared to the predictions generated by the OECD QSAR Toolbox skin sensitization module, our models had significantly higher prediction accuracy for the same sets of external compounds as evaluated by Positive Predicted Rate and Negative Predicted Rate as well as Correct Classification Rate. We have also developed QSAR models of skin permeability measured quantitatively. Cross-species correlation between human and rodent permeability data was found to be low ( $r^2=0.44$ ); thus, skin permeability models were developed using human data only and their external accuracy was  $q^2_{\text{ext}} = 0.87$  (for 62% of external compounds found within the model applicability domain). Skin sensitization models have been employed to identify putative chemical hazards in the Scorecard database of possible skin or sense organ toxicants that should be regarded as primary candidates for the experimental validation.

Keywords: QSAR modeling, skin sensitization, skin permeability, virtual screening.

## 1 INTRODUÇÃO

### 1.1 Panorama atual da indústria farmacêutica

A indústria farmacêutica é responsável por desenvolver, produzir, comercializar fármacos e outros produtos farmacêuticos, controlando um mercado que movimenta cerca US\$880 bilhões por ano (THOMSON REUTERS, 2012). A manutenção desse mercado depende do desenvolvimento de fármacos mais eficazes e seguros, o que leva a um investimento anual de aproximadamente US\$48,5 bilhões em pesquisa, desenvolvimento e inovação (PD&I) (PHRMA, 2013).

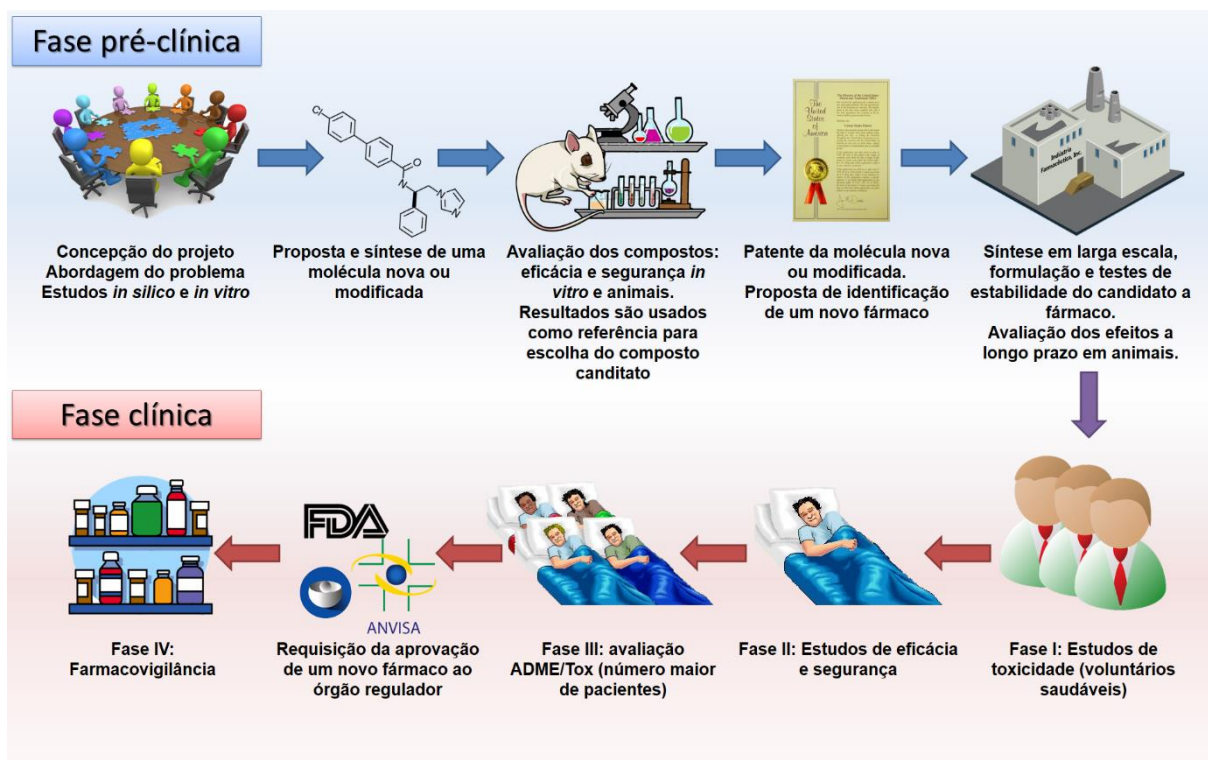
A introdução de um fármaco no mercado é um processo complexo, competitivo e que depende da integração de várias áreas estratégicas, como: inovação, conhecimento, tecnologia, gerenciamento e investimento (PHRMA, 2007). Desde a concepção do projeto até a introdução de um único fármaco no mercado, se gastam, em média, de 12 a 15 anos em PD&I, podendo ultrapassar o valor de US\$ 1,2 bilhão (PAUL et al., 2010).

Inovações terapêuticas de grande sucesso podem garantir o retorno do investimento em menos de um ano. Alguns fármacos, denominados de *blockbusters*, garantem rendimentos anuais superiores a US\$1 bilhão (ADEUSI, 2011; PHRMA, 2013). Como exemplo, cita-se o hipolipemiante atorvastatina (Liptor®, Pfizer), que só em 2008 gerou extraordinários US\$12,401 bilhões e rendeu desde seu lançamento no mercado em 1997 até 2011 aproximadamente US\$130 bilhões à Pfizer (NATURE, 2011).

Entretanto, o processo de PD&I de novos fármacos é desafiador. Em 2012, o FDA (do inglês, *Food and Drug Administration*) aprovou 44 novas entidades químicas (NCE, do inglês, *New Chemical Entities*) (PHRMA, 2013), o maior número em 15 anos. Contudo, apenas 27 NCE foram aprovadas em 2013 (FDA, 2013), mantendo a média do número de NCE aprovadas entre os anos de 2002 e 2011 (THOMSON REUTERS, 2012).

### 1.2 Planejamento e desenvolvimento de novos fármacos

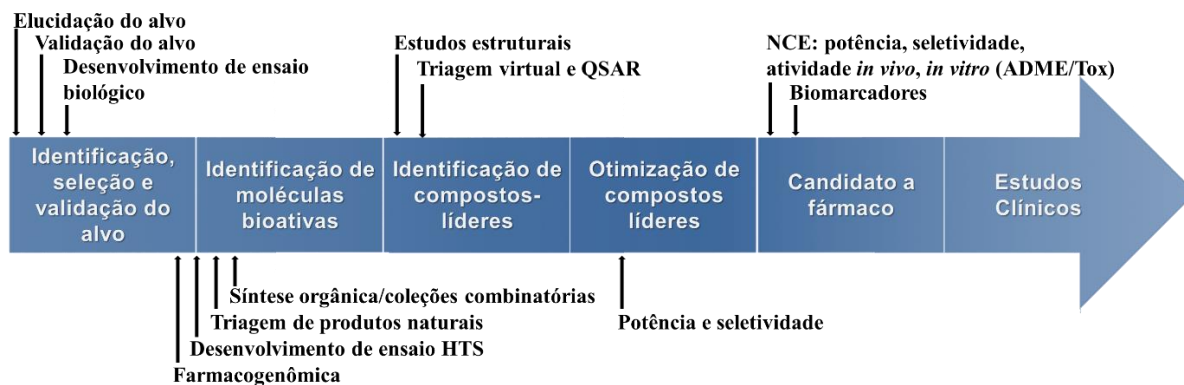
O processo de planejamento e desenvolvimento novos fármacos é complexo, envolve várias áreas do conhecimento e tipos de ensaios (Figura 1). Os ensaios desenvolvidos durante essas fases visam responder uma série de questionamentos científicos sobre um novo composto, como a avaliação de sua potência, eficácia, segurança e estabilidade. Cada uma dessas etapas soma custos e demanda tempo, modelando todo o panorama da indústria farmacêutica (PHRMA, 2007).



**Figura 1.** Esquema simplificado do processo de planejamento e desenvolvimento de um novo fármaco (modificado de GUIDO, 2008 e LOMBARDINO; LOWE, 2004).

Mesmo que o processo de PD&I de novos fármacos seja exemplificado linearmente para fins didáticos, o modelo representado na Figura 1 é simplificado. Na prática, uma rede com diferentes fases do processo ocorre concomitantemente, visto que, da concepção do projeto até a introdução do produto no mercado, diferentes compostos são avaliados e inúmeras modificações estruturais são propostas com o objetivo de encontrar uma estrutura ideal (BAXTER et al., 2013).

Nos estágios iniciais, as pesquisas se concentram na identificação e validação de alvos moleculares; e identificação e otimização de moléculas pequenas capazes de modular a atividade do alvo macromolecular selecionado (Figura 2). Nessa etapa, o estudo das interações entre os ligantes com os diferentes sistemas biológicos ou com a biofase permite a compreensão em nível molecular de processos bioquímicos, farmacológicos, toxicológicos e farmacocinéticos (WALTERS et al., 2011).



**Figura 2.** Identificação, seleção e otimização de moléculas bioativas (fase pré-clínica) (modificado de BELFIELD; DELANEY, 2006 e GUIDO, 2008).

O efeito terapêutico de um composto, parâmetros físico-químicos e suas propriedades farmacocinéticas (absorção, distribuição, metabolismo e excreção, juntamente com toxicidade - ADME/Tox) são avaliados *in silico*, *in vitro* e *in vivo*. Essa avaliação poupa tempo e recursos financeiros ao processo de PD&I, visto que elimina candidatos com propriedades inadequadas antes de seguirem para estudos clínicos (HUGHES et al., 2011).

As moléculas aprovadas nos ensaios pré-clínicos são submetidas aos órgãos reguladores, requerendo autorização para iniciar os estudos em seres humanos. Nos Estados Unidos, o órgão responsável é o FDA (PHRMA, 2013), no Brasil, a Agência Nacional de Vigilância Sanitária (ANVISA) (PIOVESAN; LABRA, 2007) e na Europa, a *European Medicine Agency* (EMA) (REGNSTROM et al., 2010).

A fase clínica visa garantir a segurança e eficácia de um candidato a fármaco em seres humanos. Comumente, é dividida em quatro fases (I-IV). Porém, o termo fase “zero” vem sendo incorporado à literatura e conduzido por algumas companhias nos últimos anos. Durante essa fase, doses subterapêuticas são administradas em um pequeno número de indivíduos (10 a 15) para coletar informações farmacocinéticas e farmacodinâmicas do composto no organismo humano (MENDOZA, 2011).

Na fase clínica I, o candidato a fármaco é administrado em um grupo ainda pequeno de voluntários (20 a 100) para avaliar a toxicidade. Na fase clínica II os estudos são iniciados em pacientes que apresentam a doença (100 a 300 indivíduos) para se avaliar a dosagem, eficácia e segurança da composição. Durante a fase III, os estudos são aplicados a um grupo maior de indivíduos (300 a  $\geq 3,000$ ), aleatorizados e distribuídos em regiões geográficas distintas, a fim de aumentar a credibilidade dos estudos. Essa fase determina o quão efetivo o candidato é, pois possibilita sua comparação com o atual tratamento padrão. A fase III é a mais onerosa, mais

demorada e mais difícil de projetar e desenvolver, especialmente em terapias para doenças crônicas (FRIEDMAN; FURBERG; DEMETS, 2010).

Uma vez que os ensaios de fase III se mostrem satisfatórios, os resultados são combinados a documentos contendo uma descrição compreensiva dos métodos e resultados dos estudos realizados em humanos e animais, processos de produção, forma farmacêutica e prazo de validade. Essa coleção de documentos é fornecida para revisão dos órgãos regulatórios, onde se espera que seja dada ao investidor a autorização de se comercializar o novo fármaco (MCGEE, 2006). No mercado, o novo fármaco passa pela fase IV, chamada de farmacovigilância, na qual é realizada a detecção, avaliação e prevenção de efeitos adversos não detectados anteriormente através da notificação espontânea e intensivo monitoramento das agências regulatórias e da própria indústria farmacêutica (HÄRMARK; VAN GROOTHEEST, 2008).

### **1.3 Química medicinal e o planejamento racional de novos fármacos**

O ser humano busca desde os primórdios algo que possa aliviar suas dores e tratar suas doenças. Esse objetivo foi alcançado durante milhares de anos de forma primitiva e não científica, usando-se apenas produtos naturais (NOGUEIRA; MONTANARI; DONNICI, 2009). Com o advento da escrita e a formação de sociedades mais complexas, a descoberta de medicamentos passou a ter como base testes empíricos usando-se, ainda, somente a natureza como fonte. A partir do meio do século XIX, técnicas mais modernas de extração e purificação de produtos naturais, levaram, mais tarde, ao desenvolvimento do medicamento industrializado como conhecemos hoje (CHAST, 2008).

Atualmente, a natureza ainda é um importante objeto de estudo para a descoberta de novos compostos bioativos. Entretanto, o número de fármacos derivados de produtos naturais tem diminuído significativamente entre os anos de 1981 e 2007 (LI; VEDERAS, 2009). De fato, moléculas promissoras podem surgir em vários estágios do planejamento, por diferentes estratégias e, inclusive, através de uma observação de um evento imprevisto (serendipidade). Contudo, a identificação de uma molécula inovadora pela triagem empírica do vasto espaço químico de estruturas fármaco-semelhante (ver item 1.4) possíveis é praticamente impossível (LIPINSKI; HOPKINS, 2004).

Por esse motivo, o espaço químico investigado precisa ser reduzido. O conjunto de estratégias e ferramentas que se baseiam no conhecimento de um alvo molecular e seus ligantes a fim de se identificar moléculas mais eficazes e seguras caracteriza o planejamento racional de

fármacos. A área da ciência que racionaliza o estudo de propriedades biológicas, farmacológicas e das relações estrutura-atividade de fármacos é denominada química medicinal (NOGUEIRA; MONTANARI; DONNICI, 2009).

Segundo a IUPAC, a química medicinal é definida como “uma disciplina baseada na química, envolvendo aspectos das ciências biológica, médica e farmacêutica, cuja missão é o planejamento, descoberta, invenção, identificação e preparação de compostos biologicamente ativos, o estudo do metabolismo, interpretação do mecanismo de ação a nível molecular e o desenvolvimento de relações entre a estrutura química e a atividade biológica (SAR)” (WERMUTH; GANELLIN; LINDBERG, 1998). A química medicinal possui papel imprescindível na identificação e otimização de estruturas, assim como na compreensão das bases moleculares da ação de fármacos, e se molda como uma ciência inter-, multi- e transdisciplinar, que permeia entre a química, especialmente a química orgânica, a farmacologia, bioquímica, biologia molecular, bioinformática, toxicologia, física, entre outras (IMMING, 2008).

Os grandes avanços da genômica e proteômica nas últimas décadas, associados à evolução das técnicas de cristalografia de raios-X e ressonância magnética nuclear (RMN), contribuíram bastante para elucidação das estruturas de proteínas envolvidas nos mais diversos processos biológicos (BERMAN et al., 2013). Além disso, a introdução de técnicas e recursos computacionais permitiu a visualização e o estudo de proteínas e ligantes, o que revolucionou todo o processo de PD&I de novos fármacos (YOUNG, 2009).

O conhecimento da estrutura do alvo molecular ou do complexo ligante-alvo permite o planejamento e desenvolvimento de inibidores enzimáticos ou agonistas/antagonistas de receptores, através do processo de complementaridade molecular (estéreo/eletrostática). Neste processo, planeja-se um candidato com propriedades estruturais adequadas para o reconhecimento molecular e aumento da afinidade pelo receptor biológico. Esta estratégia é conhecida como planejamento baseado na estrutura (SBDD, do inglês, *Structure-Based Drug Design*) (ANDRICOPULO; SALUM; ABRAHAM, 2009). Caso nenhum ligante seja conhecido, a estrutura do alvo é utilizada para a triagem de fragmentos os quais podem ser utilizados para construção de blocos e geração de novas moléculas. Este processo é denominado planejamento *de novo* (BAKER, 2013).

Por outro lado, a estrutura do alvo biológico pode não ser conhecida. Neste caso, o processo de planejamento inicia-se pelo estudo das características multifuncionais dos ligantes conhecidos ou do ligante endógeno (substrato natural) envolvido na fisiopatologia/farmacologia do processo sob investigação. Esta estratégia é conhecida como

planejamento baseado no ligante (LBDD, do inglês, *Ligand-Based Drug Design*) (RIPPHAUSEN; NISIUS; BAJORATH, 2011). Em muitos casos, o uso combinado das duas estratégias mencionadas, SBDD e LBDD, pode ser útil no processo de descoberta de novos fármacos, gerando informações adicionais fruto do sinergismo entre as técnicas (COI; BIANUCCI, 2013; SWANN et al., 2011).

O não conhecimento do(s) alvo(s) molecular(es) envolvido(s) em determinada patologia ou desordem e ligantes diminui o leque de estratégias que possam ser utilizadas. Comumente, a triagem biológica automatizada em alta escala (HTS, do inglês, *high-throughput screening*) associada à química combinatória compõem estratégias iniciais, que servem como ponto de partida para estratégias mais racionais, como o LBDD (BLEICHER et al., 2003). Mesmo assim, o HTS é uma técnica onerosa e com uma taxa de sucesso pequena na identificação de ligantes (CLARK, 2008; RIPPHAUSEN et al., 2010). Para contornar o alto custo do processo da descoberta de novos fármacos (KESSEL, 2011), a triagem virtual tem se mostrado uma alternativa ao HTS (CHENG et al., 2012).

A triagem virtual é uma técnica computacional usada na busca e descoberta de novos candidatos a fármacos a partir de um banco de dados de compostos. Seu objetivo é direcionar a busca por compostos com propriedades desejadas (potência, seletividade, propriedades farmacocinéticas e toxicológicas favoráveis) e eliminar compostos com propriedades indesejáveis (inativos, reativos, tóxicos, efeitos farmacocinéticos inapropriados) (KAR; ROY, 2013). Estratégias de LBDD usam dados de estrutura e atividade de um conjunto de estruturas com atividade biológica conhecida em via de se identificar estruturas potencialmente mais ativas/menos tóxicas, usando abordagens de busca de similaridade estrutural (WILLETT, 2006), aprendizado de máquina (item 1.5) (MELVILLE; BURKE; HIRST, 2009) e modelos farmacofóricos baseados no ligante (YANG, 2010). Por outro lado, estratégias de SBDD incluem o *docking* molecular (WASZKOWYCZ; CLARK; GANCIA, 2011) e modelos farmacofóricos baseados na estrutura do receptor (CAPORUSCIO; TAFI, 2011).

Os compostos selecionados da triagem virtual seguem para ensaios *in vitro* e *in vivo*. Caso os compostos sejam promissores, eles seguem no processo de desenvolvimento (ver item 1.2 e Figura 1). Após avaliação experimental *in vivo* e atividade promissora, os ligantes passam a se chamar protótipos (BLEICHER et al., 2003). Nesse momento, inicia-se a fase de otimização dos protótipos. Propriedades farmacodinâmicas favoráveis não implicam propriedades farmacocinéticas, farmacêuticas e toxicológicas adequadas. Por essa razão, protótipos precisam ser otimizados quanto a essas propriedades. Estudos de relação entre

estrutura atividade (QSAR, do inglês, *Quantitative Structure-Activity Relationships*) (item 1.5) contribuem significativamente para a síntese planejada de novos análogos.

#### 1.4 Triagem virtual: além da farmacodinâmica

Guiados pelo conceito da “bala mágica” cunhada por Paul Ehrlich, os químicos medicinais acreditaram por décadas que fármacos mais seguros eram sinônimos de fármacos mais seletivos (STREBHARDT; ULLRICH, 2008). Uma análise retrospectiva de falhas e sucessos de fármacos inovadores provoca uma mudança de paradigma. Como exemplo, podem-se citar os inibidores seletivos de cicloxigenase 2 (COX2) e o natalizumabe (retirado do mercado), os quais apresentam efeitos adversos potencialmente graves, não justificando seu uso em detrimento de fármacos mais antigos e seguros (DREWS, 2006). Em contrapartida, o imatinibe (Glivec<sup>®</sup>, Novartis), planejado para ser inibidor de quinase, se mostrou inicialmente um inibidor potente da BCR-ABL quinase, proteína que inibe a proliferação celular na leucemia mielóide crônica, doença não muito comum. Entretanto, posteriormente, descobriu-se que o imatinibe não era seletivo para a BCR-ABL, e sim atuava em outros alvos moleculares, como c-KIT e PDGF. Isso proveu base para seu uso em outros tipos de câncer, como tumor gastrointestinal estromal, aumentando assim o seu leque de aplicações e também os rendimentos da indústria farmacêutica que o desenvolveu (CAPDEVILLE et al., 2002).

O “aprendizado por fazer” tem levado a comunidade científica a repensar estratégias de planejamento de novos fármacos. Fármacos seletivos não atuam em outros processos-chave que estão inevitavelmente conectados por redes complexas e com alto nível hierárquico natural em sistemas biológicos (MAGGIORA, 2011). Por esse motivo, a identificação de protótipos que funcionem como uma “chave-mestra” e favoravelmente interajam com vários alvos-moleculares envolvidos em determinada patologia/desordem tem se mostrado encorajador (MEDINA-FRANCO et al., 2013).

A interação de um ligante com múltiplos alvos moleculares é denominada de polifarmacologia. Quando um ligante atua em vários alvos de determinada doença, a eficácia terapêutica pode ser melhorada, prevenindo-se, assim, o surgimento de resistência e/ou reduzindo efeitos adversos. Porém, a polifarmacologia comumente é inesperada e, na maioria das vezes, indesejada, visto que a interação com alvos moleculares não envolvidos no processo biológico em questão podem levar a efeitos adversos (PETERS, 2013).

A maioria dos fármacos que entram em fase clínica não chega ao mercado, sendo aproximadamente 50% devido à falta de eficácia, 25% por alta toxicidade e 15% por outra razão

(PLENGE; SCOLNICK; ALTSHULER, 2013). Muitas falhas por eficácia decorrem do fato de modelos animais não corresponderem totalmente à realidade biológica do organismo humano (MAGGIORA, 2011); já as falhas por toxicidade correspondem a interações com alvos moleculares indesejados (PETERS, 2013).

Historicamente, potência e seletividade de protótipos eram otimizadas primeiramente e propriedades farmacocinéticas e toxicológicas posteriormente. Atualmente, tanto a farmacodinâmica quanto a farmacocinética devem ser otimizadas concomitantemente (OPREA; MATTER, 2004). Essa mudança de paradigma sofreu bastante influência com a “regra dos cinco” de Lipinski (Ro5), que identificou parâmetros favoráveis à absorção e permeabilidade de compostos, que são eles:  $\leq 5$  doadores de ligação de hidrogênio;  $\leq 10$  aceptores de ligação de hidrogênio; massa molecular  $\leq 500$  Da; e  $\log P \leq 5$  (LIPINSKI et al., 1997). Posteriormente, Veber e colaboradores (2002) incluíram que o número de ligações rotacionáveis deveria ser  $\leq 10$  e que a área de superfície polar  $\leq 140 \text{ \AA}^2$ .

Essas regras levaram ao desenvolvimento do conceito qualitativo de *druglikeness*, o qual pode ser definido como um conjunto complexo de propriedades moleculares e características estruturais que determinam se uma estrutura em particular é similar a fármacos existentes (i.e., fármaco-semelhante) (URSU et al., 2011). Todavia, estima-se que 20% dos fármacos absorvidos por via oral falham em pelo menos uma das cinco regras propostas por Lipinski, o que torna reducionista a abordagem de aplicar filtros moleculares que retirem ligantes que não seguem essas regras nos estágios iniciais de pesquisa e desenvolvimento. Entretanto, tendo em vista o oneroso e complexo processo de desenvolvimento de novos fármacos, a triagem precoce de propriedades que vão além das propriedades farmacodinâmicas se torna muito mais interessante, visto que diminui os riscos de falhas em estágios avançados ou após a introdução do fármaco no mercado (KIRKPATRICK, 2012; OVERINGTON; ALLAZIKANI; HOPKINS, 2006).

Dada a complexidade dos organismos biológicos, triar todos os alvos pode parecer uma tarefa árdua. Todavia, existe uma série de alvos moleculares, denominados antialvos, os quais sua interação é comumente indesejada (RECANATINI; BOTTEGONI; CAVALLI, 2004). Entre os mais conhecidos antialvos, pode-se citar o canal de potássio hERG (do inglês, *human Ether-à-go-go Related Gene*), no qual sua inibição causa arritmias cardíacas potencialmente fatais (VANDENBERG et al., 2012), e receptores 5-HT<sub>2B</sub>, em que sua inibição pode levar à fibrose cardíaca (SHYU, 2009). Outras propriedades interessantes de avaliação precoce são: solubilidade em meio aquoso, biodisponibilidade, permeabilidade pela barreira hematoencefálica, interações com isoformas do citocromo P450 (GLEESON, 2008), avaliação

da mutagenicidade de Ames (XU et al., 2012), sensibilização (OECD, 2012) e permeabilidade da pele (XIA, 2011).

A avaliação de propriedades ADME/Tox *in vitro* e *in vivo* é realizada há décadas, mas ainda são experimentos caros e que demandam tempo (ADLER et al., 2011; HUTT, 1997). Além disso, existe uma crescente pressão pública e política no que diz respeito ao uso de animais em laboratório para a avaliação de propriedades farmacodinâmicas, farmacocinéticas e toxicológicas de compostos (EUROPEAN UNION, 2003; ROLLIN, 2006). Métodos *in silico*, como estudos de QSAR/QSPR podem ser úteis na avaliação de substâncias que carecem de dados experimentais (MODA et al., 2008), mesmo enfrentado a desafiadora complexidade e, frequentemente, variabilidade de dados (GLEESON et al., 2012).

A avaliação toxicológica *in silico* é importante não apenas para o desenvolvimento de novos fármacos, mas também para produtos químicos em geral, como cosméticos, conservantes, materiais de construção, produtos agrários, entre outros. Em 2006, a União Europeia aprovou uma lei chamada REACH (*Regulation, Evaluation and Authorization of Chemicals*) que exige a avaliação de propriedades físico-químicas e efeitos adversos de todo e qualquer composto que é produzido em escala de mais de 1 ton/ano na Europa. A REACH recomenda o uso de métodos que não utilizem animais, como abordagens de relações quantitativas entre estrutura e atividade/propriedade (QSAR/QSPR) (item 1.5), a fim de diminuir o número e o custo de testes em animais (EUROPEAN UNION, 2007, 2010).

## **1.5 Relações quantitativas entre estrutura química e atividade/propriedade (QSAR/QSPR)**

### **1.5.1 História e evolução**

Na década de 1930, Hammett publicou um estudo que representa um marco no entendimento de propriedades moleculares de compostos orgânicos. Ele estudou o efeito de substituintes nas reações de derivados do ácido benzoico a fim de se quantificar relações lineares de energia livre de Gibbs. Uma equação (Equação 1) foi proposta, a qual relaciona linearmente os logaritmos da constante de ionização do ácido benzoico meta- ou para-substituído ( $\sigma$ ) e o logaritmo da constante de ionização do ácido benzoico ( $\rho$ ) com a constante do equilíbrio de reação com o substituinte R ( $K$ ) e a constante de equilíbrio quando R é um átomo de hidrogênio ( $K_0$ ) (HAMMETT, 1937).

$$\rho\sigma = \log(K/K_0)$$

Equação 1

A primeira abordagem teórica que relacionava atividades biológicas com a estrutura molecular data do final dos anos 1940. O índice de Wiener e o número de Platt, propostos em 1947 para modelar o ponto de ebulição de hidrocarbonetos foram os primeiros descritores moleculares teóricos (ver item 1.5.4) baseados na teoria dos grafos (PLATT, 1947; WIENER, 1947). Outro estudo inovador que relacionava propriedades físico-químicas e as interações de energia soluto-solvente foi publicado no final dos anos 1950 (TAFT, 1952, 1953).

Na década de 1960, um estudo inovador publicado por Hansch e colaboradores (1962) consolidava as bases dos estudos das relações quantitativas entre estrutura química e atividade/propriedade (QSAR/QSPR). Nesse estudo, foram relatados os esforços para entender as relações estrutura-atividade (SAR, do inglês, *structure-activity relationships*) de reguladores do crescimento de plantas, mas nenhuma relação robusta foi produzida.

Logo depois, Fujita, Iwasa e Hansch (1964) começaram a analisar o coeficiente de partição octanol-água ( $\log P$ ) em 203 benzenos mono e dissustituídos e perceberam que essa propriedade era aditiva: a contribuição parcial do  $\log P$  do substituinte em uma molécula é comumente a mesma em outra. Essa contribuição hidrofóbica foi chamada de  $\pi$ , a qual é definida pela diferença entre o  $\log P$  da molécula derivada substituída e da não substituída ( $\pi = \log P_x - \log P_H$ ). Hansch e Fujita (1964) combinaram a constante hidrofóbica ( $\pi$ ) com a constante de Hammett ( $\sigma$ ) e tentaram relacionar o efeito de substituintes na atividade biológica de ácidos benzoicos usando regressão linear múltipla através da Equação 2.

$$\log(1/C) = a\sigma + b\pi + c$$

Equação 2

No qual  $\log(1/C)$  corresponde à concentração molar do composto necessária para produzir uma resposta biológica definida,  $\sigma$  ao efeito eletrônico,  $\pi$  à contribuição hidrofóbica e  $a$ ,  $b$  e  $c$  aos respectivos coeficientes determinados por análise de regressão. Entretanto, Hansch notou que a resposta biológica de compostos com caráter hidrofóbico alcançava um platô ou diminuía após determinado valor. Este comportamento se deve à retenção de compostos lipofílicos nos interstícios das membranas, impedindo que o composto chegue ao seu sítio de ação. Uma nova equação (Equação 3), mais robusta que expressa que o  $\log(1/C)$  depende do  $\log P$  parabolicamente foi proposta (HANSCH, 1969).

$$\log(1/C) = a(\log P)^2 + b \log P + c \quad \text{Equação 3}$$

Além da abordagem de Hansch, outros pesquisadores buscaram descrever a atividade biológica de compostos através de equações; a análise de Free e Wilson (1964) não considera qualquer parâmetro dos substituintes seja definido. A ideia básica dessa abordagem é que a atividade biológica de uma molécula pode ser descrita como a soma das contribuições de atividade das subestruturas (núcleo comum + grupo R). Na década de 1970, Kubinyi refinou o modelo parabólico de Hansch, desenvolvendo um modelo (Equação 4) que descreve a dependência não linear da atividade biológica sobre o caráter hidrofóbico. Nesse caso, o termo  $\beta$  é um vetor e corresponde ao movimento do composto em sistemas multicompartimentados (DEBNATH, 2001; KUBINYI, 1977, 1979).

$$\text{Log}(1/C) = a \log P - b \log(\beta P + 1) + c \quad \text{Equação 4}$$

Os trabalhos citados tiveram significativo impactor para a química medicinal, por introduzir a ideia multiparamétrica da correlação da atividade biológica, o que permitiu que os químicos medicinais refletissem sobre como substituintes mudariam propriedades químicas e conseqüentemente a atividade biológica (GANELLIN, 2004). Diferentes abordagens e técnicas foram desenvolvidas desde então.

A abordagem descrita até aqui é comumente denominada de QSAR clássico, em que a atividade biológica é correlacionada apenas com propriedades físico-químicas (experimentais ou calculadas). No início dos anos 1980, abordagens de geração de fragmentos moleculares em duas dimensões (2D) para conjuntos de treinamento grandes (centenas a milhares) foram desenvolvidas na tentativa de relacionar esses fragmentos com a atividade biológica, o que foi uma inovação para a época e deu início à desafiadora tarefa de se desenvolver métodos que pudessem ser preditivos para conjuntos de dados diversos (compostos de séries não análogas), gerando-se assim modelos chamados de “globais”. Modelos construídos para séries congêneres ou subconjuntos análogos de uma série diversa são chamados de “loais” (CHERKASOV et al., 2014; GAVAGHAN et al., 2007).

Ainda na década de 1980, várias abordagens que consideravam parâmetros em três dimensões (3D) foram desenvolvidas. Hopfinger desenvolveu descritores 3D baseados na forma molecular, criando o método denominado Análise da Forma Molecular (MSA, do inglês, *Molecular Shape Analysis*), o qual gera descritores baseados na forma da conformação mais provável de ser ativa para o conjunto de moléculas estudadas (HOPFINGER, 1980).

Um dos maiores avanços foi o QSAR-3D, com a técnica de Análise Comparativa de Campos Moleculares (CoMFA, do inglês, *Comparative Molecular Fields Analysis*) descrita por Cramer, Patterson e Bunce (1988). Essa técnica correlaciona mudanças na atividade biológica com efeitos estéricos e eletrostáticos empregando o método dos mínimos quadrados parciais (PLS, do inglês, *partial least squares*). Em 1989, CoMFA foi implementado na plataforma comercial do SYBYL (<http://www.tripos.com>) e se tornou um dos métodos mais utilizados e difundidos nas últimas duas décadas (ZHANG et al., 2011). Um inconveniente do CoMFA é que ele é dependente do alinhamento das estruturas do conjunto de dados, o que torna o processo de modelagem lento e tedioso durante a tentativa de que garantir a preditividade dos modelos (CHERKASOV et al., 2014). A Análise Comparativa de Índices de Similaridade Molecular (CoMSIA, do inglês, *Comparative Molecular Similarity Index Analysis*) estende as variedades estéricas e eletrostáticas do CoMFA para ligações de hidrogênio e efeitos hidrofóbicos (KLEBE; ABRAHAM; MIETZNER, 1994).

Na tentativa de desenvolver métodos cada vez mais preditivos, Hopfinger e colaboradores propuseram o QSAR-4D, que é uma abordagem tridimensional que utiliza uma abordagem conformacional obtida por meio de simulação de dinâmica molecular (HOPFINGER et al., 1997). O Quasar QSAR-5D proposto por Vedani e Dobler (2002) foi uma extensão do QSAR-4D, adicionando liberdade conformacional, permitindo uma representação múltipla da topologia dos ligantes no sítio ativo. O mesmo grupo propôs em seguida o Quasar QSAR-6D, que considera vários modelos de solvatação simultaneamente (VEDANI; DOBLER; LILL, 2005).

Na última década, o QSAR apresentou vários avanços e crescente interesse da academia e indústria farmacêutica (CRAMER, 2012). Inúmeros descritores moleculares, métodos de aprendizado de máquina e parâmetros de validação foram desenvolvidos e vem sendo aplicados. O desenvolvimento das “ômicas”, o aprimoramento de recursos computacionais e o crescente aumento da disponibilidade de conjuntos de dados de alta qualidade aliados ao desenvolvimento de modelos preditivos levam a crer que o campo de estudos de modelagem de QSAR e suas aplicações tendem a aumentar (CHERKASOV et al., 2014).

### 1.5.2 Aplicações

Nas últimas décadas, a capacidade de coletar, analisar, armazenar e disseminar todos os tipos de dados se desenvolveu em velocidade surpreendente. Técnicas modernas como a química combinatória e o HTS levaram ao acúmulo de bancos de dados moleculares imensos

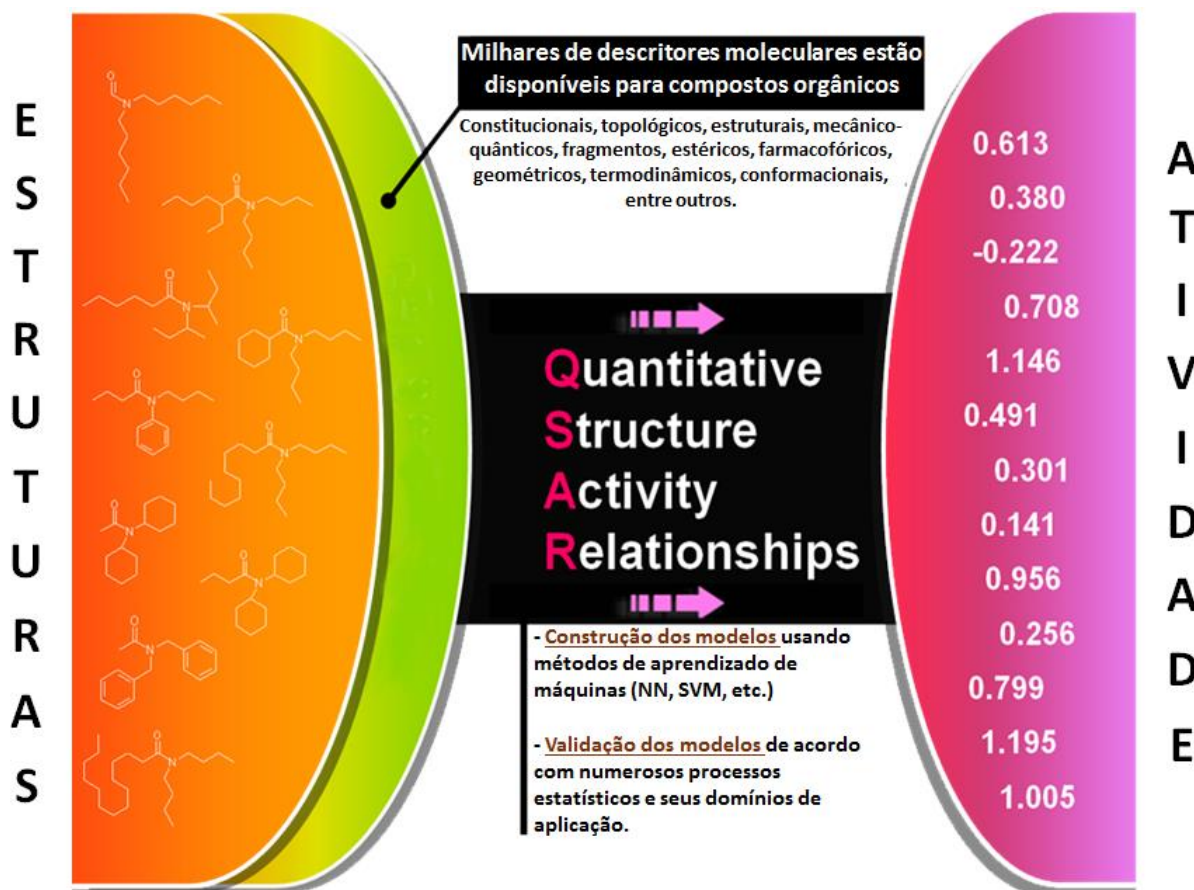
(OPREA; TROPSHA, 2006). Conjuntos de dados contendo milhares de compostos testados em múltiplos ensaios biológicos disponíveis em bases de dados como PubChem (WANG et al., 2012) e ChEMBL (GAULTON et al., 2012) estão disponíveis *online* para a comunidade científica. Para se identificar novos protótipos nesses conjuntos de dados, faz-se necessário o uso de procedimentos que sejam robustos e eficientes para triagem virtual.

Estudos computacionais permitem a predição da atividade biológica e de propriedades físico-químicas por meios racionais, assim como a compreensão e racionalização do mecanismo de ação entre séries congêneres. Subsequentemente, estratégias *in silico* contribuem para redução do custo do desenvolvimento do produto; redução do número de animais utilizados em ensaios experimentais; promoção da química verde por aumentar a eficiência do processo de PD&I, diminuindo resíduos que seriam descartados por compostos improváveis de terem sucesso (CRONIN, 2010).

Nesse contexto, estudos de QSAR/QSPR apresentam várias aplicações, como: (i) identificação racional de novos ligantes/protótipos com atividade/propriedade desejada; (ii) otimização da atividade/propriedade; e a (iii) identificação de compostos potencialmente perigosos em estágios preliminares do desenvolvimento (TROPSHA, 2010).

### 1.5.3 Princípios

A abordagem de QSAR/QSPR pode ser descrita como um método estatístico de análise de dados para desenvolver modelos que possam prever corretamente determinada atividade biológica ou propriedade de compostos baseados em sua estrutura química. Para tal fim, dois tipos de informações são necessários para o desenvolvimento desses modelos: (i) a atividade biológica/propriedade e (ii) descritores químicos (Figura 3). Essa relação é gerada pela aplicação de métodos estatísticos que, genericamente, pode ser definida como  $P_i = k'(D_1, D_2, \dots, D_n)$ , na qual  $P_i$  é a atividade biológica ou propriedade da molécula,  $D_1, D_2, \dots, D_n$  são propriedades estruturais (descritores) calculadas (ou, em alguns casos, medidas experimentalmente) e  $k'$  é um peso definido pelo algoritmo computacional e atribuído aos descritores para se calcular as propriedades das moléculas (CRONIN, 2010; TROPSHA, 2010).



**Figura 3.** Esquema representando a geração de um modelo de QSAR (modificado de TROPSHA, 2010).

Atualmente, milhares de descritores moleculares estão disponíveis (TODESCHINI; CONSONNI, 2000). O desenvolvimento manual de um modelo matemático considerando a quantidade de dados disponíveis demoraria, possivelmente, uma vida inteira de trabalho. Para lidar com essa imensa quantidade de informação, métodos matemáticos e estatísticos se tornaram imprescindíveis. Esses métodos, denominados de aprendizado de máquina, estabelecem peso aos descritores, ajustando a equação que relaciona a estrutura química com a atividade biológica ou propriedade (Figura 3) (TROPSHA, 2010).

#### 1.5.4 Descritores moleculares

Nas duas primeiras décadas da era QSAR, apenas alguns parâmetros físico-químicos e mecânico-quânticos estavam disponíveis para serem usados como descritores (SELASSIE; ONE, 2003). Pesquisadores perceberam que nem toda informação molecular era extraída por ensaios experimentais, o que impulsionou esforços para extrair informações codificadas na

estrutura da molécula em números – chamados de descritores moleculares – através de um caminho teórico (CONSONNI; TODESCHINI, 2010).

Um descritor molecular é o resultado final de um procedimento matemático e lógico que transforma informação química codificada em uma representação simbólica de uma molécula em um número útil ou o resultado de algum experimento padronizado. Descritores moleculares contribuem para a compreensão de propriedades moleculares e/ou podem ser utilizados na geração de um modelo matemático para a predição de determinada propriedade de outras moléculas (TODESCHINI; CONSONNI, 2000).

O campo de estudo de descritores moleculares é multi- e interdisciplinar, envolvendo imensa quantidade de diferentes teorias, que envolvem conhecimentos em álgebra, teoria dos grafos, teoria da informação, química computacional, química orgânica e físico-química. Para uso dos descritores, conhecimentos de estatística, quimiometria e princípios de QSAR/QSPR são necessários, assim como conhecimento específico no problema estudado (CONSONNI; TODESCHINI, 2010).

Diferentes tipos de descritores químicos refletem diferentes níveis de representação estrutural. Esses descritores podem ser classificados quanto à sua “dimensionalidade” em unidimensionais (1D), baseados em propriedades físico-químicas e da fórmula molecular (ex., massa molecular, refratividade molar, logP, entre outros); bidimensionais (2D), que descrevem propriedades que podem ser calculadas de uma representação 2D (ex., número de átomos, número de ligações, índices de conectividade, entre outros); e tridimensionais (3D), que dependem da conformação das moléculas (ex., volume de Van der Waals, área de superfície acessível ao solvente, entre outros) (XUE; BAJORATH, 2000).

Outra classificação diz respeito à natureza desses descritores, que podem ser: (i) constitucionais, que são derivados da composição atômica do composto (ex., peso molecular, números de átomos e ligações); (ii) topológicos (ex., índices de Randić); (iii) geométricos, que são derivados de coordenadas 3D (ex., volume molecular, área de superfície polar, entre outros); (iv) eletrostáticos, que são derivados da substituição de cargas parciais (ex., índices de polaridade, carga parciais, entre outros); e (v) quanto-mecânicos, que são derivados das funções de onda dos elétrons (ex., energia dos orbitais moleculares) (CONSONNI; TODESCHINI, 2010).

Para desenvolver modelos de QSAR/QSPR, descritores e os dados de atividade/propriedade são armazenados em uma tabela, como demonstrado na Tabela 1, na qual os dados de atividade/propriedade são armazenados na matriz **Y** e os descritores na matriz **X**. Vários tipos de relações podem ser obtidos dessas matrizes. Modelos de SAR podem ser

gerados de forma categórica (ex., ativos/inativos, tóxico/não tóxico) ou uma relação quantitativa, na qual a propriedade estudada (**Y**) é representada por uma função de um ou mais descritores (**X**) (CRONIN, 2010).

**Tabela 1.** Representação genérica de uma matriz de dados para um estudo de QSAR/QSPR.

Identificador químico	Atividade/ Propriedade	Descritor 1	Descritor 2	Descritor 3	...	Descritor n
Molécula 1	$Y_1$	$X_{11}$	$X_{12}$	$X_{13}$	...	$X_{1n}$
Molécula 2	$Y_2$	$X_{21}$	$X_{22}$	$X_{23}$	...	$X_{2n}$
Molécula 3	$Y_3$	$X_{31}$	$X_{32}$	$X_{33}$	...	$X_{3n}$
Molécula 4	...	...	...	...	...	...
Molécula 5	$Y_n$	$X_{n1}$	$X_{n2}$	$X_{n3}$	...	$X_{nn}$

Muitos *softwares* estão disponíveis para o cálculo de descritores moleculares, como DRAGON<sup>®</sup> (Talete SRL, Milan, Italy), CDK<sup>®</sup> (<http://www.rguha.net/code/java/cdkdesc.html>), CODESSA<sup>®</sup> (<http://www.semichem.com/codessa/>), entre outros.

### 1.5.5 Métodos de aprendizado de máquina

Químicos medicinais enfrentaram ao longo das últimas décadas o desafio de lidar com a imensa quantidade de dados químicos e biológicos de forma a se identificar novos compostos biologicamente ativos (OPREA; TROPSHA, 2006). O ramo da química que usa técnicas computacionais para transformar dados em informação e informação em conhecimento, com a intensão de aprimorar a identificação e otimização de compostos no processo de desenvolvimento de fármacos é denominado de quimioinformática (VARNEK; BASKIN, 2011).

Um modelo de QSAR é uma equação matemática que relaciona a estrutura química com a atividade biológica. A geração desses modelos não necessita, obrigatoriamente, do uso de recursos computacionais (*vide* análise de Hansch). Contudo, parece inviável esse trabalho considerando a gama de descritores moleculares e conjuntos de dados de atividade biológica disponíveis. Para contornar essa dificuldade, métodos de aprendizado têm sido amplamente empregados e são praticamente tão importantes quanto os dados de atividade biológica e descritores moleculares (MELVILLE; BURKE; HIRST, 2009).

O aprendizado de máquina é um ramo da inteligência artificial que diz respeito ao estudo e construção de sistemas que podem aprender com dados. Esse aprendizado é construído em três estágios: (i) representação dos dados; (ii) otimização da hipótese; (iii) generalização. Na geração de um modelo de QSAR, os dados são representados conforme demonstrado na Tabela 1. Uma hipótese (equação) é gerada na tentativa de estabelecer uma relação dos descritores com a atividade biológica e essa relação é otimizada. Por exemplo: supõe-se que determinada atividade biológica  $Y$  é descrita linearmente em função de único descritor molecular  $x$ . Portanto, uma boa hipótese para esta equação poderia ser estabelecida pela Equação 5 (WELLING, 2011).

$$Y(x) = a + bx$$

Equação 5

Com o objetivo de encontrar valores para os parâmetros  $a$  e  $b$ , algoritmos computacionais (ex. gradiente descendente) são utilizados para minimizar esses valores, de forma que se ajustem e sejam capazes de descrever os dados de entrada (conjunto treinamento). Em seguida, é testada a capacidade de generalização da hipótese gerada. Generalizar diz respeito à habilidade da hipótese em prever corretamente um exemplo/tarefa não usado na construção dos modelos (conjunto teste) (WELLING, 2011).

Métodos de aprendizado supervisionados (os algoritmos são treinados com variável  $Y$  definida, na tentativa de gerar uma função matemática, que generalize a mesma) são utilizados para gerar modelos de QSAR/QSPR que sejam capazes de discriminar ativos/inativos (modelos binários) e previsões numéricas da propriedade biológica (modelos contínuos). Para execução dessa tarefa, descritores moleculares são calculados para um conjunto de compostos com propriedade biológica definida experimentalmente. Esses descritores são usados para treinar o algoritmo, o qual busca fazer uma representação da atividade biológica ou propriedade com base nos descritores, seguido de uma otimização dessa representação e, por fim, uma generalização da propriedade predita com base na equação gerada (Figura 3). Os modelos gerados, uma vez validados, constituem um ponto de partida para seleção de novos compostos (DOMINGOS, 2012; TROPSHA, 2010).

Métodos não supervisionados (os algoritmos são treinados sem definição da variável  $Y$ , na tentativa de identificar padrões nos conjuntos de dados com base nos descritores) são frequentemente utilizados para análise de um conjunto de dados químico. Essa abordagem, quando aplicada aos conjuntos de dados químicos, identifica subgrupos homogêneos entre um conjunto de dados heterogêneo. Alguns dos vários usos dessa abordagem incluem (i) a

verificação da diversidade estrutural do conjunto de dados; *(ii)* a avaliação da consistência de dados experimentais; e *(iii)* a exploração de possíveis interferências que influenciam na atividade, contribuindo para a revelação de novas regras de SAR (DOWNS; BARNARD, 2003; MERCIER, 2003).

#### 1.5.6 Boas práticas de desenvolvimento e validação

Várias diretrizes e recomendações de boas práticas de desenvolvimento e validação de modelos de QSAR foram publicadas nos últimos anos (CHERKASOV et al., 2014; DEARDEN; CRONIN; KAISER, 2009; TROPSHA, 2010). Em 2004, a OECD (*Organization for Economic Co-operation and Development*) publicou princípios para validação de modelos de QSAR para uso prático de agências de regulamentação. Esses princípios são: *(i)* atividade biológica ou propriedade definida; *(ii)* algoritmo claro; *(iii)* domínio de aplicabilidade (DA) definido; *(iv)* avaliação apropriada da robustez e preditividade; *(v)* e interpretação mecanística, que significa encontrar relações entre os descritores e a atividade biológica ou propriedade, em via de se compreender melhor o mecanismo de ação de uma estrutura química ou aprofundar o conhecimento biológico sobre a propriedade em estudo (OECD, 2004).

Apesar desses princípios terem sido estabelecidos para se garantir que modelos de QSPR relativos à propriedade de toxicantes fossem usados por agências de regulamentação de produtos químicos, eles têm sido amplamente considerados para desenvolvimentos de quaisquer modelos de QSAR/QSPR. Em consideração a esses princípios, vários elementos-chave devem ser considerados durante o desenvolvimento e validação de modelos de QSAR/QSPR e serão discutidos a seguir (CHERKASOV et al., 2014; DEARDEN; CRONIN; KAISER, 2009; TROPSHA, 2010).

##### 1.5.6.1 Preparo do conjunto de dados

Comumente, pesquisadores que trabalham com quimioinformática e desenvolvimento de modelos de QSAR não são os mesmos que geram os dados experimentais. Por consequência, os conjuntos de dados necessitam ser cuidadosamente preparados para se evitar que erros sejam perpetuados e/ou interfiram na preditividade dos modelos (Figura 4).



**Figura 4.** Fluxo de trabalho para preparo de conjunto de dados químicos (modificado de Fourches, Muratov e Tropsha, 2010).

Nesse processo, compostos que não podem ser adequadamente tratados por técnicas de quimioinformática usadas no estudo precisam ser removidos. Por exemplo, alguns *softwares* (ex., DRAGON, [http://www.taletе.mi.it/products/dragon\\_molecular\\_descriptors.htm](http://www.taletе.mi.it/products/dragon_molecular_descriptors.htm)) não calculam descritores moleculares para compostos inorgânicos e organometálicos. Manter esses compostos no conjunto de dados interfere negativamente no desenvolvimento dos modelos, visto que a atividade biológica para tais compostos não poderá ser completamente correlacionada.

Quimiotipos específicos, como anéis aromáticos e grupos nitro ( $\text{RNO}_2$ ), e formas tautoméricas precisam ser padronizados e os contraíons removidos, para evitar que a mesma estrutura permaneça no conjunto de dados. Assim, duplicatas necessitam ser identificadas e removidas. Uma inspeção manual é requerida ao final do processo para se garantir que todas as estruturas estejam corretas (FOURCHES; MURATOV; TROPSHA, 2010).

#### 1.5.6.2 Detecção e remoção de outliers antes da geração dos modelos

*Outliers* são compostos que apresentam atividade/propriedade inesperada ou não se ajustam ao modelo. *Outliers* podem resultar de erros experimentais ou podem apontar falhas no

modelo gerado. Por esse motivo, não podem ser removidos sem explicação, visto que sua remoção para melhorar os parâmetros estatísticos do modelo é considerada manipulação e essa melhoria pode não representar a real melhoria na preditividade externa do modelo. Se um composto possui um dado de atividade/propriedade dentro do intervalo do conjunto de dados e esse composto está dentro do domínio de aplicabilidade (ver item 1.5.6.4) do conjunto de dados, ele não pode ser removido. Esse composto pode representar um *cliff* de atividade e sua má predição pelo modelo precisa ser interpretada e discutida (CHERKASOV et al., 2014). Um *cliff* de atividade é definido pela razão da diferença na atividade de dois compostos pela diferença da similaridade de dois compostos no espaço químico e pode ser determinado pelo SALI (do inglês, *Structure-Activity Landscape Index*) (Equação 6), ou seja, estruturas químicas semelhantes com atividade/propriedade muito diferentes (GUHA; VAN DRIE, 2008; MAGGIORA, 2006).

$$SALI = \frac{|Y_i - Y_j|}{1 - sim(i, j)} \quad \text{Equação 6}$$

no qual Y representa a atividade biológica e *sim* a similaridade entre as estruturas *i* e *j*.

### 1.5.6.3 Validação dos modelos

Um modelo de QSAR/QSPR é uma representação teórica, portanto, sua utilidade depende de sua capacidade de generalizar (predizer), com taxa de acerto elevada, determinada atividade biológica ou propriedade de compostos que não foram utilizados para gerar os modelos. Essa capacidade é denominada preditividade. Dois métodos principais são usados para determinar a preditividade dos modelos: validação interna e validação externa. Para validação externa um conjunto teste não usado na construção dos modelos deve ser utilizado e parâmetros de avaliação devem ser empregados (GOLBRAIKH; TROPSHA, 2002; GRAMATICA, 2007).

### 1.5.6.4 Definição do domínio de aplicabilidade

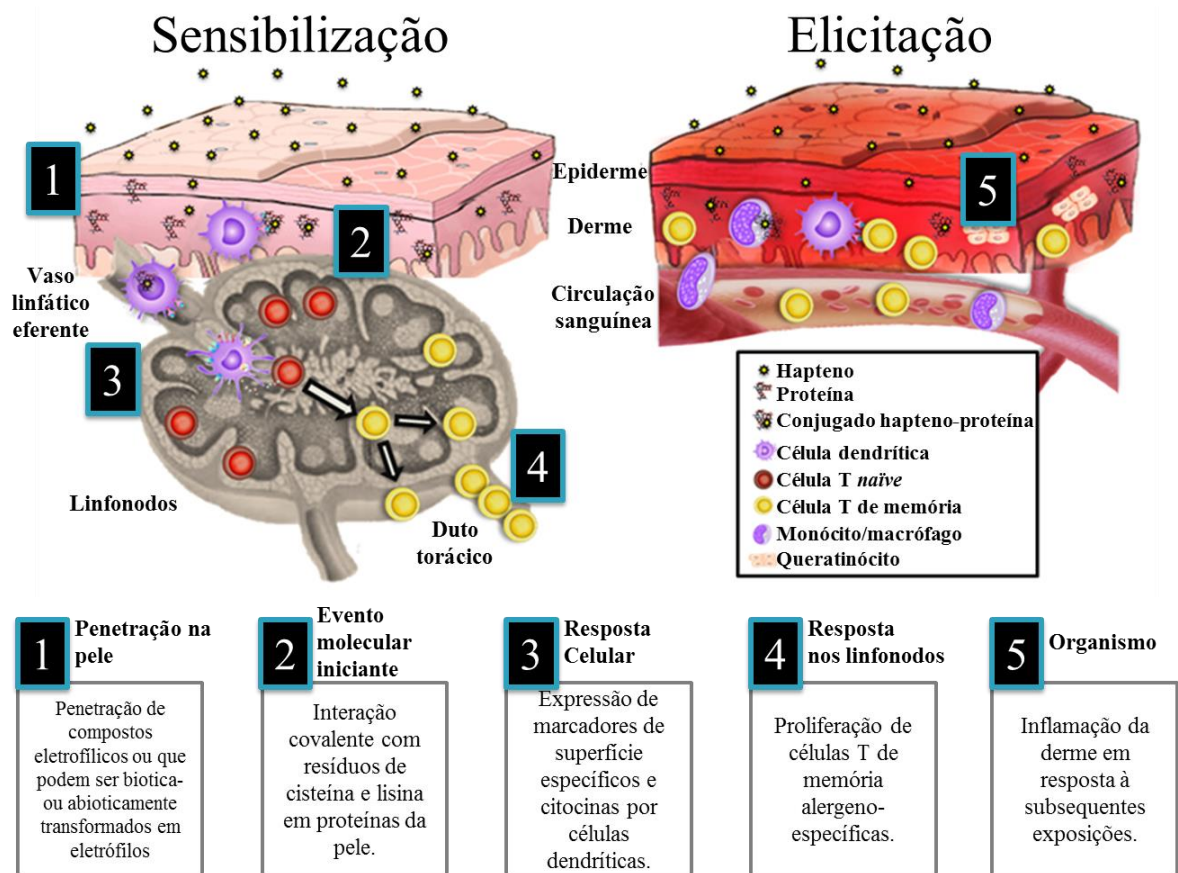
Um modelo de QSAR pode predizer a atividade de qualquer composto químico. Contudo, se sua estrutura química for muito diferente das estruturas usadas na construção do modelo, essa predição pode ser incorreta, visto que o modelo não cobre o espaço químico o

qual se insere o composto em questão. Portanto, a determinação do domínio de aplicabilidade é obrigatória para aceitabilidade de um modelo (TROP SHA, 2010).

## 1.6 Sensibilização da pele

Seres humanos estão expostos diariamente a uma variedade de substâncias naturais e sintéticas que nunca foram testadas em qualquer tipo de protocolo toxicológico. As informações sobre os riscos direcionados à saúde humana e ao meio ambiente para esses compostos químicos são limitadas e muitas vezes inadequadas, mesmo entre compostos com alto volume de produção (CHUPRINA et al., 2010; EGEHY et al., 2012; MUIR; HOWARD, 2006). Muitos compostos químicos perigosos causam efeitos adversos pelo contato direto com a pele; o fenômeno associado inclui a sensibilização e permeabilidade desses compostos na pele (DICKEL et al., 2002; GRANDJEAN et al., 1988; KIMBER et al., 2011). Cada um desses fenômenos tem sido amplamente estudado independentemente, mesmo que seja possível uma conexão entre eles (LEPOITTEVIN, 2011; MAGNUSSON; PUGH; ROBERTS, 2004; STRID; STROBEL, 2005).

A sequência de respostas biológicas que se inicia com os eventos moleculares e que leva a uma resposta adversa *in vivo* é chamada de via de efeito adverso (ANKLEY et al., 2010; KNUDSEN; KLEINSTREUER, 2011; OECD, 2012; WATANABE et al., 2011). O evento molecular iniciante para a sensibilização da pele resulta de um tipo de hipersensibilidade tardia, denominada dermatite de contato alérgica (AEBY et al., 2010; HENNINO et al., 2005). A dermatite de contato alérgica é uma doença ocupacional e um importante problema de saúde ambiental (KEEGEL et al., 2009; KIMBER et al., 2002). Sua via de efeito adverso consiste em duas fases (Figura 5): (i) sensibilização da pele e (ii) eliciação da resposta imune. A primeira fase, sensibilização da pele, se inicia com o contato e penetração da substância na pele (KARLBERG et al., 2008).



**Figura 5.** Via de efeito adverso da sensibilização da pele (modificado de KARLBERG et al., 2008 e OECD, 2012).

Durante a passagem pelas camadas da pele, os compostos químicos podem estar sujeitos a diferentes biotransformações, que podem aumentar ou diminuir seu potencial alergênico (OECD, 2012). Muitos haptenos (moléculas pequenas que podem elicitar uma resposta imune apenas quando agregada a um grande carregador, como uma proteína) são conhecidos por possuir grupos lipofílicos e terem baixa massa molecular (<500 Da), o que lhes permite atravessar a camada mais externa da pele, o estrato córneo (BOS; MEINARDI, 2000). Haptenos podem possuir grupos eletrofílicos que se ligam covalentemente a resíduos nucleofílicos de proteínas na pele para formar conjugados estáveis, caracterizando o evento molecular iniciante, o qual parece ser o maior determinante, dependente da estrutura para o potencial sensibilizador da molécula (ROBERTS; APTULA, 2008). Esses conjugados, também chamados de complexo hapteno-proteína, são processados por células dendríticas (células de Langerhans), as quais subsequentemente maturam e migram para os linfonodos (OECD, 2012; SAINT-MEZARD et al., 2004). Os conjugados são apresentados para linfócitos T *naïve* (precursores de células T hapteno-específicas) resultando em uma proliferação de células T hapteno-específicas que

emigram dos linfonodos e entram na corrente sanguínea pelo duto torácico (HENNINO et al., 2005). A segunda fase, elicitação, ocorre depois de um contato subsequente com o mesmo hapteno. Haptenos difundem para a pele e formam complexos hapteno-proteína, os quais são capturados por células dendríticas. As células T hapteno-específicas são ativadas pelas células dendríticas na derme e epiderme, iniciando o processo inflamatório responsável por lesões (HENNINO et al., 2005; OECD, 2012; SAINT-MEZARD et al., 2004).

#### 1.6.1 Avaliação experimental da sensibilização da pele

Modelos *in vivo* para avaliação da sensibilização da pele incluem o *occluded patch test* (BUEHLER, 1965), o *guinea pig maximization test* (GPMT) (MAGNUSSON; KLIGMAN, 1969) e o *local lymph node assay* (LLNA) (BASKETTER et al., 2002); sendo o último atualmente considerado como o teste padrão-ouro (OECD, 2010). O LLNA foi desenvolvido por Kimber, Mitchell e Griffin (1986) e validado por Basketter e colaboradores (1992). O propósito deste ensaio é identificar potenciais sensibilizadores da pele quantificando a proliferação de linfócitos, o qual está correlacionado com a extensão da sensibilização provocada após reexposição tópica a uma substância sensibilizadora.

Nesse teste, a substância-teste é administrada topicamente na orelha do camundongo. Dois dias depois da última aplicação,  $^3\text{H}$ -timidina ou  $^{125}\text{I}$ -iododesoxiuridina (em solução salina tamponada com fosfato; 250  $\mu\text{L}$ /camundongo) é administrada na veia caudal. Cinco horas depois, os linfonodos auriculares são excisados e uma suspensão celular dos linfonodos é preparada para se quantificar a incorporação dos nucleosídeos radioativos, os quais se correlacionam com a proliferação celular nos linfonodos. A incorporação de nucleotídeos radioativos para cada camundongo é expressa em desintegrações por minuto (DPM). O índice de estímulo é calculado como a relação da média DPM/camundongo para cada grupo de tratamento contra a média DPM/camundongo do grupo controle. O limiar para uma dose positiva é um índice de estímulo  $\geq 3$ . A concentração necessária para produzir um índice de estímulo  $\geq 3$  é chamada de EC3 (BASKETTER et al., 2002).

Uma modificação do LLNA, o *reduced local lymph node assay* (rLLNA), que reduz o número de animais envolvidos em 40%, foi recentemente validada pelo ICCVAM (do inglês, *Interagency Coordinating Committee on the Validation of Alternative Methods*). Nessa modificação, apenas uma dose da substância é testada, sendo aquela que possui maior solubilidade sem causar toxicidade sistêmica e/ou irritação/corrosão da pele (ICCVAM, 2009). Apesar de uma considerável redução no uso de animais, esses testes continuam caros e com

baixo rendimento, enquanto métodos *in silico* estão emergindo como solução prática para a avaliação de substâncias que carecem de dados experimentais (RAUNIO, 2011).

#### 1.6.2 Revisão bibliográfica de modelos de QSAR a respeito da sensibilização da pele

Apesar de muitos modelos de sensibilização da pele serem descritos na literatura (Tabela 2), e aparentemente estarem bem ajustados e serem robustos, uma análise crítica dos mesmos revela vários problemas importantes. Especificamente, a maioria dos modelos de QSAR publicados para a sensibilização da pele não obedeceu a procedimentos e critérios estatísticos, assim como recomendações de validação externa, que constituem boas práticas comuns (CHERKASOV et al., 2014; TROPSHA, 2010), não estando, portanto, de acordo com as orientações da OECD (2004) para o desenvolvimento e validação de modelos de QSAR. Como consequência, apesar do grande número de estudos de QSAR reportados na literatura, nenhum modelo pode realmente ser empregado para predição confiável do potencial sensibilizador de compostos químicos.

**Tabela 2.** Estudos prévios de QSAR para sensibilização da pele.

Tipo de ensaio	Número de compostos	Fonte do conjunto de dados	Descritores utilizados	Método estatístico	Validação	Resultados dos modelos	DA	Y-rand	Referência
GPMT	259	(CRONIN; BASKETTER, 1994)	MolConn-X Medchem	SDA	Interna	SE=0,78 ES=0,88	Não	Não	(CRONIN; BASKETTER, 1994)
Alertas clínicos	72	(ZIEGLER et al., 1989)	LogP, RM, pH, FL Fragmentos Eletrônico	LDA	Interna	SE=0,79 ES=0,88	Não	Não	(MAGEE et al., 1994)
Alertas clínicos	72	(MAGEE et al., 1994)	RM, FL Fragmentos Eletrônico	LDA	Interna	SE=0,90 ES=1,00	Não	Não	(HOSTÝNEK; MAGEE, 1997b)
	89	(HOSTÝNEK; MAGEE; MAIBACH, 1996)	RM, FL Fragmentos Eletrônico	Teste <i>t</i>	Interna	SE=0,95 ES=0,60			
GPMT	259	(CRONIN; BASKETTER, 1994)	(CRONIN; BASKETTER, 1994)	ANN	Externa	SE=0,83 ES=0,94	Não	Não	(DEVILLERS, 2000)
LLNA	93	Várias	TOPS-MODE (Topológico)	LDA (Forte / moderado)	Externa	ACB=0,80	Não	Não	(ESTRADA et al., 2003)
	42		FL, RM, cargas atômicas, van der Waals	LDA (fraco / não sensibilizadores)	Interna	ACB=0,80			

Tabela 2. (Continuação).

Tipo de ensaio	Número de compostos	Fonte do conjunto de dados	Descritores utilizados	Método estatístico	Validação	Resultados dos modelos	DA	Y-rand	Referência
LLNA	54	(ASHBY et al., 1995; HANEKE et al., 2001)	Dragon	LR	Interna	AC=0,79 SE=0,68 ES=0,90	Não	Não	(FEDOROWICZ et al., 2004)
LLNA	178	(HANEKE et al., 2001)			DEREK	SE=0,87 ES=0,32 ACB=0,73			
			Dragon MolconnZ	LR	Externa	SE=0,95 ES=0,46 ACB=0,82	Não	Não	(FEDOROWICZ et al., 2005)
					DEREK	ACB=0,83			
					TOPKAT	ACB=0,73			
GPMT	105	(HANEKE et al., 2001)	Dragon MolconnZ	LR	Externa	SE=0,98 ES=0,43 ACB=0,86			
LLNA	178	(ICCVAM; NICEATM, 1999)	Dragon Molconn-Z	Árvore de classificação	10-fold	SE=0,99 ES=0,34	Não	Não	(LI et al., 2005)
				RF	10-fold	SE=0,84 ES=0,83			
LLNA	67		Codessa	SAM1	LOO	$r^2_{cv}=0,88$	Não	Não	(MILLER et al., 2005)

Tabela 2. (Continuação).

Tipo de ensaio	Número de compostos	Fonte do conjunto de dados	Descritores utilizados	Método estatístico	Validação	Resultados dos modelos	DA	Y-rand	Referência
LLNA	40	(ROBERTS et al., 2007)		TIMES	Externa	SE=0,56 ES=0,88	Sim	Não	(ROBERTS et al., 2007)
LLNA	131	(GERBERICK et al., 2005)	4D-fingerprints	LR	LOO	SE=0,80 ES=0,80 AC <sub>LOO</sub> =0,89	Não	Não	(LI et al., 2007)
			4D-fingerprints	PLS-LR	LOO <sub>c</sub>	ES=0,70 AC <sub>LOO</sub> =0,89			
LLNA	162	(FEDOROWICZ et al., 2005)	Dragon	SVM/PSO	Externa	AC=0,89 SE=0,93 ES=0,73	Não	Não	(YUAN; HUANG; CAO, 2009)
GPMT	92	(FEDOROWICZ et al., 2005)	Dragon	SVM/PSO	Externa	AC=0,90 SE=0,96 ES=0,71	Não	Não	
LLNA	358	(GERBERICK et al., 2005; ICCVAM; NICEATM, 1999)	Codessa			AC=0,90			
GPMT	307	(CRONIN; BASKETTER, 1994; DEVILLERS, 2000)	Dragon	ANN	Interna	AC=0,95	Não	Não	(GOLLA et al., 2009)
BgVV	251	(SCHLEDE et al., 2003)				AC=0,90			

Tabela 2. (Continuação).

Tipo de ensaio	Número de compostos	Fonte do conjunto de dados	Descritores utilizados	Método estatístico	Validação	Resultados dos modelos	DA	Y-rand	Referência
LLNA	209	(GERBERICK et al., 2005)	Dragon	AFP	Externa	AC=0,90	Sim	Não	(CHAUDHRY et al., 2010)
						SE=1,00			
						ES=0,75			
				MLP	Externa	AC=0,71	Sim	Não	
						SE=0,70			
						ES=0,67			
LLNA	255	Várias	Bio-suite	GA-kNN	Externa	SE=0,75	Não	Não	(GUNTURI et al., 2010)
						ES=0,86			
						ACB=0,80			
LLNA	161	Várias	E-Dragon	SVM	Externa	1-TFN=0,903	Não	Não	(CHENG et al., 2010)
						1-TFP=0,778			
						AC=0,857			
				AdaBoost	Externa	1-TFN=0,968			
						1-TFP=0,778			
						AC=0,898			
LLNA	393	Várias	Fragmentos, logP, mecânico-quânticos, eletrônicos	RPT	Externa	SE=0,895	Não	Não	(LU et al., 2011)
						ES=0,722			
						AC=0,809			

Nota: AC: acurácia; AC<sub>LoO</sub>: acurácia baseada no método *leave-one-out*; ACB: acurácia balanceada; ANN: *artificial neural networks*; DA: domínio de aplicabilidade; ES: especificidade; FL: fator de lipofilicidade; GPMT: *Guinea Pig Maximization Test*; LLNA: *Local Lymph Node Assay*; SDA: *stepwise discriminant analysis*; LDA: *linear discriminant analysis*; LR: *logistic regression*; MLP: *multilayer perceptron*; PSO: *particle swarm optimization*; RM: refratividade molar RF: *random forest*; RPT: *recursice partitioning tree*; SAM1: *semi-empirical ab initio model 1*; SE: sensibilidade; SVM: *support vector machine*; TFN: taxa de falsos negativos; TFP: taxa de falsos positivos; Y-rand: *Y-randomization*.

As principais falhas dos modelos apresentados na Tabela 2 são: (i) a preditividade não foi avaliada adequadamente e/ou testada em compostos não usados na construção dos modelos (CRONIN; BASKETTER, 1994; FEDOROWICZ et al., 2004; HOSTÝNEK; MAGEE, 1997b; MAGEE et al., 1994); (ii) o DA não foi estimado; (iii) não há prova de passagem pelo teste de *Y-randomization* (aleatorização da variável Y) (CHENG et al., 2010; CRONIN; BASKETTER, 1994; DEVILLERS, 2000; ESTRADA et al., 2003; FEDOROWICZ et al., 2004, 2005; GOLLA et al., 2009; GUNTURI et al., 2010; HOSTÝNEK; MAGEE, 1997b; LI et al., 2007, 2005; LU et al., 2011; MAGEE et al., 1994; MILLER et al., 2005; YUAN; HUANG; CAO, 2009); e (iv) muitos usaram conjuntos de dados desbalanceados (CHAUDHRY et al., 2010; FEDOROWICZ et al., 2005; LI et al., 2005, 2007; LU et al., 2011), o que resultou em modelos tendenciosos em direção à classe com maior número de compostos.

Essas falhas comprometem o uso prático e confiável de métodos de avaliação prévia para sensibilização da pele. Por exemplo, o conjunto de dados estudado por Cronin e Basketter (1994) continha muitos *cliffs* de atividade (fenóis e acetatos, no caso); isso explica porque fenóis e acetatos foram tão mal preditos. Magee e colaboradores (1994) e Hostýnek e Magee (1997) investigaram a viabilidade de construir modelos usando abordagens de classificação binária e contínua, mas a decisão de usar apenas compostos usados como fragrâncias diminuiu bastante o espaço químico desses modelos. Devillers (2000) experimentou comparar o poder de predição de redes neurais artificiais e análise discriminante linear, mas selecionou um conjunto teste contendo apenas 7% do número total de compostos, o que não é suficiente para uma validação externa adequada. A abordagem TOPS-MODE (*Topological Substructural Molecular Descriptors*) (ESTRADA et al., 2003) demonstrou uma precisão relativamente boa, mas que pode ter sido sobrestimada devido ao pequeno número de compostos nos dois conjuntos de validação externa (15 e 6 compostos, respectivamente). Similarmente, em uma análise simplificada, o modelo desenvolvido por Miller e colaboradores (2005) pareceu ser altamente preciso; contudo, uma análise mais detalhada revelou que 20 compostos foram considerados *outliers* e assim removidos do conjunto modelagem simplesmente por terem sido mal preditos pelo modelo. A remoção desses compostos pode ter artificialmente superestimado a habilidade preditiva desse modelo. Roberts e colaboradores (2007) descreveram uma validação externa que foi realizada na plataforma *Tissue Metabolism Simulator for Skin Sensitization* (TIMES-SS). Os autores testaram experimentalmente 40 compostos químicos no ensaio LLNA e então compararam os resultados com predições computacionalmente derivadas feitas pelo TIMES-SS. Apesar da alta especificidade (ca. 87,5%), a sensibilidade do modelo foi baixa (ca. 56%). Outro exemplo de um estudo questionável foi apresentado por Golla e colaboradores (2009)

que desenvolveram um modelo usando um conjunto de dados do *Federal Institute for Health Protection of Consumers and Veterinary Medicine* (SCHLEDE et al., 2003). O conjunto de dados foi coletado de dados clínicos e experimentais em humanos e diferentes testes em animais. Os autores dividiram os compostos identificados em três grupos: (i) efeito alergênico significativo por contato; (ii) indicação sólida de efeito alergênico por contato; (iii) efeito alergênico insignificante por contato. Esse sistema de classificação não é claro e pode fazer os esforços descritos pelos autores menos prático para uso futuro e mais difícil de comparar com outros estudos.

## 1.7 Permeabilidade da pele

A habilidade de moléculas passarem por membranas é chamada de permeabilidade. A permeabilidade de um composto químico na pele é determinada por propriedades físico-químicas dos compostos e pelas propriedades físico-químicas e biológicas da membrana (XIA, 2011). Apesar da alta importância da permeabilidade da pele, a quantidade de dados experimentais disponíveis para domínio público é bastante limitada. No começo dos anos 1990, Flynn (1990) publicou uma coleção de dados para permeabilidade da pele humana compilada de várias fontes, que foi enriquecida ao longo das últimas décadas por outros cientistas. Como disponível na Tabela 3, a maioria dos estudos publicados não tem definição do DA e nenhuma prova de passar no teste de *Y-randomization* (ABRAHAM et al., 1999; BARRATT, 1995; CHEN; LIAN; HAN, 2007, 2010; CRONIN et al., 1999; HOSTÝNEK; MAGEE, 1997a; LIEN; GAO, 1995; MAGNUSSON et al., 2004; MOSS; CRONIN, 2002; MOSS et al., 2011; PATEL; TEN BERGE; CRONIN, 2002; POTTS; GUY, 1992, 1995).

**Tabela 3.** Estudos prévios de QSAR para permeabilidade da pele.

<b>Tipo de ensaio</b>	<b>Número de compostos</b>	<b>Fonte do conjunto de dados</b>	<b>Descritores utilizados</b>	<b>Método estatístico</b>	<b>Validação</b>	<b>Resultados dos modelos</b>	<b>DA</b>	<b>Y-rand</b>	<b>Referência</b>
Pele Humana LogKp (cm/h)	93	(FLYNN, 1990)	VM, MM, $K_{ow}$	MRA	Interna	$r^2=0,670$	Não	Não	(POTTS; GUY, 1992)
Pele Humana LogKp (cm/h)	31	(SCHEUPLEIN; BLANK, 1973; SCHEUPLEIN; BRONAUGH, 1983; SCHEUPLEIN et al., 1969)	LogP, MM, $H_b$	SRA	Externa	$q^2=0,960$	Não	Não	(LIEN; GAO, 1995)
Pele Humana LogKp (cm/h)	60	(FLYNN, 1990)	VM, logP, Mpt	MRA PCA	Interna	$r^2=0,885$	Não	Não	(BARRATT, 1995)
Pele Humana LogKp (cm/s)	37	(FLYNN, 1990)	VM, logP, $\log D_0/\delta$ , $H_d$ , $H_a$	MRA	Interna	$r^2=0,940$	Não	Não	(POTTS; GUY, 1995)
Pele Humana LogKp (cm/s)	28	(FLYNN, 1990)	logP, RM, $H_b$ , $H_a$ , veículo, oclusão	MRA	Interna	$r^2=0,893$	Não	Não	(HOSTÝNEK; MAGEE, 1997a)
Pele Humana LogKp (cm/h)	53	(FLYNN, 1990; JOHNSON; BLANKSCHTEIN; LANGER, 1995)	Abraham	MRA	Interna	$r^2=0,957$	Não	Não	(ABRAHAM et al., 1999)

Tabela 3. (Continuação).

<b>Tipo de ensaio</b>	<b>Número de compostos</b>	<b>Fonte do conjunto de dados</b>	<b>Descritores utilizados</b>	<b>Método estatístico</b>	<b>Validação</b>	<b>Resultados dos modelos</b>	<b>DA</b>	<b>Y-rand</b>	<b>Referência</b>
Pele Humana LogKp (cm/h)	107	(KIRCHNER et al., 1997)	Físico-químicos, topológicos, eletrônicos	SRA/LSRA	Interna	$r^2=0,927$	Não	Não	(CRONIN et al., 1999)
Pele Humana LogKp (cm/s)	116	(DEGIM; PUGH; HADGRAFT, 1998; FLYNN, 1990; JOHNSON; BLANKSCHTEIN; LANGER, 1995; KIRCHNER et al., 1997)	LogP, MM	Análise de regressão	Interna	$r^2=0,810$	Não	Não	(MOSS; CRONIN, 2002)
Pele Humana LogKp (cm/h)	143	(FLYNN, 1990; WILSCHUT et al., 1995)	ClogP, VAMP, TSAR, QSARis, AMSOL, Dragon, Índice de linearidade, $\log K_{ow}$	SRA	Interna	$r^2=0,900$	Não	Não	(PATEL; TENBERGE; CRONIN, 2002)
Pele Humana LogJ <sub>max</sub> (mol/cm <sup>2</sup> /h)	178	Várias	MM	SMRA	Externa	$r^2=0,847$ $q^2=0,784$	Não	Não	(MAGNUSSON et al., 2004)

Tabela 3. (Continuação).

Tipo de ensaio	Número de compostos	Fonte do conjunto de dados	Descritores utilizados	Método estatístico	Validação	Resultados dos modelos	DA	Y-rand	Referência
Pele Humana LogKp (cm/h)	215	(FLYNN, 1990; JOHNSON; BLANKSCHTEIN; LANGER, 1997; PATEL; TEN BERGE; CRONIN, 2002)	ABSOLV	ANN	Externa	$q^2=0,698$	Não	Não	(CHEN; LIAN; HAN, 2007)
				MLR		$q^2=0,792$			
Pele Humana LogKp (cm/h)	>200	Várias	Brick and Mortar		Externa	$q^2=0,74$	Não	Não	(CHEN; LIAN; HAN, 2010)
Pele Humana LogKp (cm/h)	208	(CHEN; LIAN; HAN, 2007, 2010; KIELHORN; MELCHING-KOLLMUS; MANGELSDORF, 2005)	e-Dragon, Pharma, ADME/Tox web	MLR	Externa	$q^2= 0,932$	Sim	Não	(CHAUHAN; SHAKYA, 2010)
				PCR		$q^2= 0,886$			
				PLSR		$q^2= 0,936$			

Tabela 3. (Continuação).

Tipo de ensaio	Número de compostos	Fonte do conjunto de dados	Descritores utilizados	Método estatístico	Validação	Resultados dos modelos	DA	Y-rand	Referência
Pele Humana logKp (cm/h)	140	Várias				MSE=1,13			
Pele de roedores logKp (cm/h)	103	Várias	logP, MM, H <sub>d</sub> , H <sub>a</sub> , solubilidade	GPR	Externa	MSE=0,53	Não	Não	(MOSS et al., 2011)
Pele suína logKp (cm/h)	15	Várias				MSE=0,83			
Membrana sintética logKp (cm/h)	19	Várias				MSE=3,57			

Nota: ANN: *artificial neural networks*; D: difusibilidade pela membrana; DA: domínio de aplicabilidade; GPR: *gaussian process regression*; H<sub>a</sub>: acceptor de ligação de hidrogênio; H<sub>b</sub>: habilidade de formar ligações de hidrogênio; H<sub>d</sub>: Doador de ligação de hidrogênio; K<sub>ow</sub> ou P: partição octanol/água; LSRA: *least-square regression analysis*; MLR: *multiple linear regression*; MRA: *multivariate regression analysis*; MSE: *mean-squared error*; PF: ponto de fusão; MM: massa molecular; PCA: *principal component analysis*; PCR: *principal component regression*; PLSR: *Partial least squares regression*; RM: refratividade molar; S(M)RA: *stepwise (multiple) regression analysis*; SVM: *support vector machine*; VM: volume molecular; Y-rand: *Y-randomization*; δ: difusão pelo comprimento do caminho.

Chauhan e Shakya (2010) desenvolveram um modelo que aparentemente seguiu as melhores práticas para construção de modelos de QSAR (TROPSHA, 2010); contudo, foram encontradas várias duplicatas no conjunto de dados utilizado (Tabela 4), o que indica uma potencial tendenciosidade do modelo e provável sobrestimação de sua capacidade preditiva.

**Tabela 4.** Par de estruturas duplicadas encontradas no estudo de Chauhan e Shakya (2010).

Nome	logKp	Duplicatas	log Kp
<i>p</i> -Fenilenodiamina	-3,62	4-Fenilenodiamina	-3,62
Cortexona	-3,35	Deoxicorticosterona	-3,34
2-Fenilenodiamina	-3,35	<i>o</i> -Fenilenodiamina	-3,35
Propanol	-2,85	<i>n</i> -Propanol	-2,85
Ácido butanoico	-2,58	Ácido butírico	-3,00
Lidocaina	-2,40	Lignocaina	-2,40
Butanona	-2,35	2-Butanona	-2,35
<i>p</i> -Nitrofenol	-2,25	4-Nitrofenol	-2,25
3-Nitrofenol	-2,25	<i>m</i> -Nitrofenol	-2,25
2-Metilfenol	-1,81	<i>o</i> -Cresol	-1,81
3-Metilfenol	-1,81	3-Cresol; <i>m</i> -cresol	-1,82
Éter dietílico	-1,79	Éter de etila	-1,80
4-Metilfenol	-1,75	4-Cresol; <i>p</i> -cresol	-1,76
2-Naftol	-1,55	$\beta$ -Naftol	-1,56
2-Clorofenol	-1,48	<i>o</i> -Clorofenol	-1,48
4-Etil fenol	-1,46	<i>p</i> -Etilfenol	-1,45
3,4-Xilenol	-1,44	3,4-Dimetilfenol	-1,44
Timol	-1,26	2-Isopropil-5-metilfenol	-1,27
4-Cloro-3,5-dimetilfenol	-1,23	4-Cloro-3,5-xilenol	-1,28

## **2 JUSTIFICATIVA E OBJETIVOS**

### **2.1 Justificativa e Objetivo Geral**

Os estudos de QSAR resumidos nas tabelas 2 e 3 não obtiveram sucesso em prover à comunidade modelos preditivos e robustos capazes de avaliar precisamente o potencial sensibilizador e a permeabilidade na pele humana de compostos químicos. Além disso, apesar da importância da permeabilidade da pele para o processo de sensibilização (KARLBERG et al., 2008; OECD, 2012), não foi possível identificar qualquer estudo no qual ambas propriedades foram analisadas conjuntamente no possível contexto de interdependência.

Face ao exposto, o objetivo geral do presente trabalho foi realizar um estudo sistemático para elucidar características químicas que influenciam os riscos provocados por contato com a pele associados à sensibilização e permeabilidade de compostos químicos, assim como gerar modelos de QSAR que sejam capazes de prever o potencial sensibilizador e penetrante de compostos sem dados experimentais.

### **2.2 Objetivos específicos**

Os objetivos específicos deste trabalho foram:

- Compilar, integrar e preparar o maior conjunto de dados publicamente disponível para sensibilização e permeabilidade da pele;
- Desenvolver e validar modelos de QSAR para sensibilização e permeabilidade da pele;
- Aplicar esses modelos ao conjunto de dados da Scorecard para identificação de potenciais agentes sensibilizadores;
- Estudar a relação global (se existente) entre sensibilização e permeabilidade usando tanto dados experimentais quanto preditos.

## 3 MATERIAIS E MÉTODOS

### 3.1 Conjuntos de dados

#### 3.1.1 Conjunto de dados de sensibilização da pele (conjunto de dados A)

Dados de sensibilização da pele baseados no LLNA contendo 471 registros de compostos foram obtidos do relatório do ICCVAM (2009). Como os dados desse relatório vinham de fontes diferentes, verificou-se na literatura a variabilidade interlaboratorial do LLNA. Em seguida, analisou-se se a concordância da atividade biológica entre estruturas duplicadas a fim de se confirmar a baixa variabilidade do teste entre diferentes veículos e laboratórios. Para evitar que modelos de QSAR apresentassem preditividade tendenciosa em direção ao conjunto maior (260 sensibilizadores), esse conjunto foi balanceado antes da modelagem para equalizar o número de sensibilizadores e não sensibilizadores. Para o balanceio, usou-se o grupo menor, contendo 127 não sensibilizadores, como sonda para buscar e ordenar os sensibilizadores mais similares. A seleção foi executada no módulo *Methods of Data Analysis* do *software HiT QSAR (Hierarchical QSAR technology)* (KUZ'MIN; ARTEMENKO; MURATOV, 2008) em três etapas: (i) gerou-se a matriz da confusão da distância Euclidiana entre todos os pares de compostos; (ii) selecionou-se metade dos sensibilizadores (64 compostos) com menor distância Euclidiana ao não sensibilizador mais próximo; (iii) selecionou-se a outra metade (63 compostos) aleatoriamente considerando-se toda distribuição das distâncias dos sensibilizadores restantes.

#### 3.1.2 Conjunto de validação externa para sensibilização da pele (conjunto de dados B)

Jaworska e colaboradores (2011) desenvolveram uma estratégia intitulada *Bayesian Network Integrated Testing* na tentativa de estimar o potencial sensibilizador de compostos químicos combinando-se dados *in silico* e *in vitro* relacionados à penetração na pele, reatividade peptídica e ativação de células dendríticas. Eles compilaram dados de LLNA da literatura (GERBERICK et al., 2005; KERN et al., 2010) e não publicados de vários laboratórios. foram recuperadas desse estudo 153 estruturas únicas (conjunto de dados B). Estruturas presentes também no conjunto de dados A foram identificadas no *software HiT QSAR* e removidas.

#### 3.1.3 Conjunto de dados da Scorecard (conjunto de dados C)

Scorecard é uma página da *web* financiada pela empresa GoodGuide<sup>®</sup> dedicada à coleta e publicação de informação relacionada com poluição ambiental e toxicantes. A equipe da Scorecard compilou dados de várias fontes – por exemplo, *Environmental Protection Agency of United States* (<http://www.epa.gov/>) e Haz-Map<sup>®</sup> (<http://hazmap.nlm.nih.gov>) – e rotulou 867 compostos suspeitos de serem toxicantes de pele ou órgãos internos. Os compostos foram recuperados das bases de dados do PubMed, ChemSpider e SciFinder e preparados. Compostos presentes no conjunto de dados C que estavam presentes no conjunto de dados A foram identificados no *software* HiT QSAR e removidos.

#### 3.1.4 Conjunto de dados da permeabilidade da pele humana (conjunto de dados D)

Dados *in vitro* de permeabilidade humana foram coletados do estudo de Chauhan e Shakya (2010) envolvendo 211 registros expressos em logK<sub>p</sub> (cm/h).

#### 3.1.5 Conjunto de dados da permeabilidade da pele de roedores (conjunto de dados E)

Dados *in vitro* da permeabilidade da pele de roedores consistindo de 103 compostos foram retirados do estudo de Moss e colaboradores (2011).

### 3.2 Preparo do conjunto de dados

Informações das estruturas de compostos químicos foram retiradas das bases de dados do PubChem (<http://pubchem.ncbi.nlm.nih.gov/>), ChemSpider (<http://www.chemspider.com/>) e SciFinder ([scifinder.cas.org/scifinder/](http://scifinder.cas.org/scifinder/)) usando-se o número de registro do *Chemical Abstracts Services* (CAS) ou nomes químicos. Compostos sem estrutura química definida/depositada em uma dessas bases de dados foram removidos. Cada conjunto de dados foi independentemente preparado (FOURCHES; MURATOV; TROPSHA, 2010). Contraíons foram removidos e quimiotipos específicos como grupos aromáticos e nitro foram normalizados usando o *software* Standardizer (v.5.3, ChemAxon, Budapest, Hungary, <http://www.chemaxon.com>). Sais inorgânicos, compostos organometálicos, polímeros e misturas também foram removidos. A presença de duplicatas, ou seja, compostos idênticos reportados mais de uma vez no mesmo conjunto de dados, é conhecida por levar à uma sobrestimação da preditividade dos modelos de QSAR gerados. Contudo, a análise de tais registros também permite estimar a qualidade do conjunto de dados: se os dados de atividade para o mesmo composto são semelhantes, a qualidade dos dados é alta; se existe um grande

desvio dos valores experimentais, a qualidade é baixa. As duplicatas foram identificadas usando o *software* HiT QSAR (KUZ'MIN; ARTEMENKO; MURATOV, 2008) e cuidadosamente analisadas. Quando os valores experimentais associados a um mesmo composto eram idênticos, apenas um dos registros era mantido no conjunto de dados. Caso contrário, todos os registros eram removidos.

### 3.3 Cálculo dos descritores

#### 3.3.1 Descritores DRAGON

Os seguintes tipos de descritores 2D foram gerados usando o DRAGON (v.5.5, Talete SRL, Milan, Italy): constitucionais (contagem de átomos e grupos químicos), grupos funcionais, fragmentos centrados nos átomos, topológicos, índices de conectividade, índices de informação, índices de carga topológica, autovalores de Burden, *fingerprints* e propriedades moleculares (TODESCHINI; CONSONNI, 2000).

#### 3.3.2 Descritores SiRMS

Descritores SiRMS (do inglês, *Simplex Representation of Molecular Structure*) contam fragmentos tetratômicos com composição e estrutura topológica fixa. Em nível bidimensional, a conectividade dos átomos, tipo de átomo e de ligação (simples, dupla, tripla ou aromática) foram consideradas. Esses descritores consideram não apenas o tipo de átomo, mas também características atômicas que podem ter importância na atividade biológica das moléculas, como carga parcial, lipofilicidade, refratividade molar e a habilidade do átomo em questão de ser doador/aceptor de ligação de hidrogênio. Características atômicas com valores contínuos (carga, lipofilicidade e refração) possuem as seguintes divisões: (i) carga parcial  $I \leq -0,05 < II \leq 0 < III \leq 0,05 < IV$ , (ii) lipofilicidade  $I \leq -0,5 < II \leq 0 < III \leq 0,5 < IV$ ; (iii) refratividade molar  $I \leq 1,5 < II \leq 3 < III \leq 8 < IV$ . Para ligação de hidrogênio, os átomos foram divididos em três categorias: aceptor de ligação de hidrogênio, doador de ligação de hidrogênio e átomo indiferente (KUZ'MIN; ARTEMENKO; MURATOV, 2008; MURATOV et al., 2010).

### 3.4 Análise de agrupamentos hierárquicos

A HCA (do inglês, *hierarchical clustering analysis*) é um método de aprendizado não supervisionado. Nesse estudo empregou-se o método *Sequential Agglomerative Hierarchical*

*Non-overlapping* implementado no *software* ISIDA Cluster (<http://infochim.u-strasbg.fr>) (VARNEK et al., 2007). Nesse método, cada composto representa um grupo (do inglês, *cluster*) no início. Então, os  $n$  compostos são agrupados iterativamente em grupos usando suas distâncias euclidianas armazenadas em uma matriz quadrada de distâncias simétricas ( $n * n$ ). Os dois objetos mais próximos (compostos ou grupos) são iterativamente identificados e agrupados para formar um novo grupo. A matriz de distâncias é atualizada com as novas distâncias separando o novo grupo dos outros, de acordo com o tipo de acoplamento selecionado pelo usuário (nesse estudo foi usado o acoplamento completo). O processo se repete até que reste apenas um grupo. A relação sequencial de formação dos grupos resulta em um gráfico hierárquico chamado de dendograma. Esse *software* gera um dendograma e um mapa de calor da matriz de distância colorida de acordo com a similaridade/dissimilaridade entre os compostos.

### 3.5 Análise do vizinho mais próximo

A análise do vizinho mais próximo (do inglês, *k-nearest neighbor*) foi realizada no *software* ISIDA Similarity (<http://infochim.u-strasbg.fr>), usando o Coeficiente de Tanimoto (CT, Equação 7) para determinar a similaridade de determinado composto com todos os compostos de determinado conjunto de dados. Os compostos do conjunto de dados são ordenados pela similaridade e o composto mais similar (maior CT) é o vizinho mais próximo (VARNEK et al., 2007).

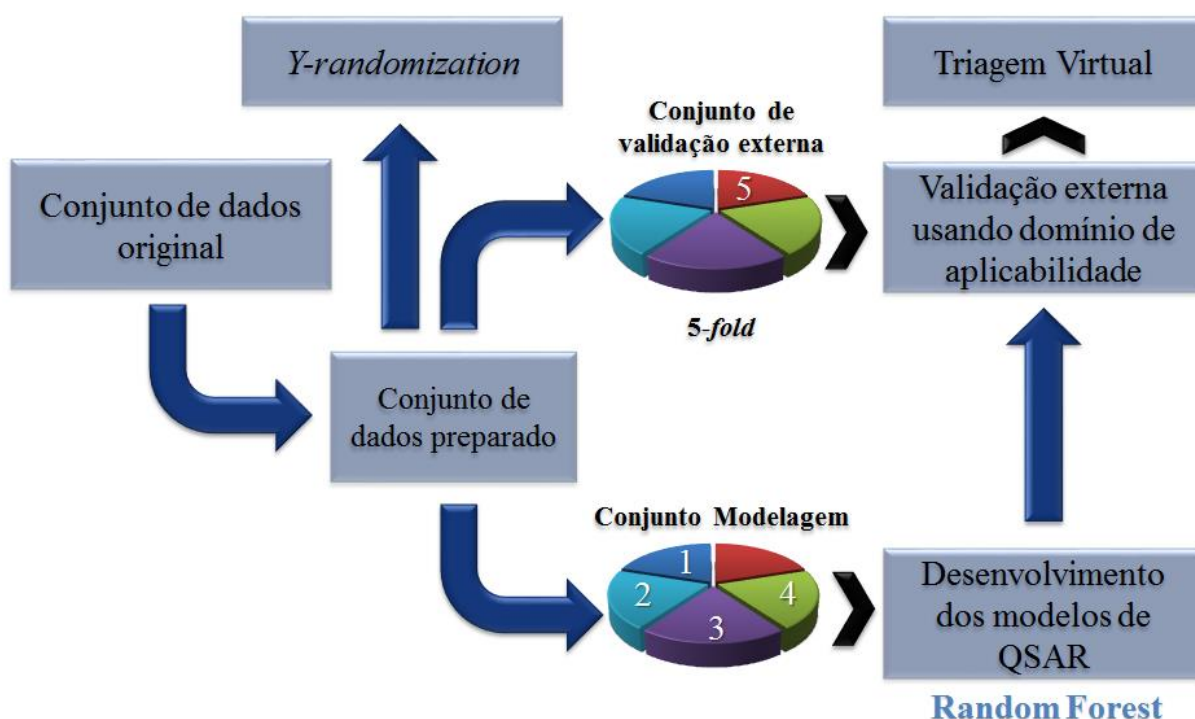
$$CT(I, J) = \frac{I \cdot J}{\|I\|^2 + \|J\|^2 - I \cdot J} \quad \text{Equação 7}$$

no qual I e J são objetos representados por vetores.

### 3.6 Geração dos modelos de QSAR

O fluxo de trabalho para geração dos modelos de QSAR incluiu três grandes passos (CHERKASOV et al., 2014; TROPSHA, 2010): (i) preparo/análise do conjunto de dados (seleção de descritores e compostos), (ii) construção dos modelos e (iii) validação/seleção dos modelos. Nesse trabalho foi empregado o método de validação cruzada externa de *5-fold* (Figura 6): o conjunto de dados total com atividade experimental definida é aleatoriamente dividido por Kennard-Stone em cinco subgrupos de tamanhos iguais; então um desses subgrupos (20% de todos os compostos) é definido como conjunto de validação externa e os

quatro conjuntos restantes formam o conjunto de modelagem (80% de todo o conjunto de dados). Esse procedimento é repetido cinco vezes permitindo que cada um dos cinco subconjuntos seja usado como conjunto de validação externa. Os modelos são gerados usando o conjunto modelagem apenas. É importante enfatizar que o conjunto de validação externa nunca é empregado para construção e/ou seleção dos modelos. Cada conjunto modelagem é dividido em vários conjuntos de treinamento e teste; então os modelos são gerados usando compostos de cada conjunto treinamento e aplicados aos conjuntos testes para avaliar a preditividade.



**Figura 6.** Fluxo de trabalho para desenvolvimento dos modelos de QSAR (modificado de Tropsha (2010)).

Os seguintes parâmetros foram utilizados para avaliar diferentes aspectos da preditividade dos modelos classificatórios de sensibilização (Equações 7-11) e modelos contínuos de permeabilidade (Equações 12-14):

Acurácia balanceada (ACB)

$$ACB = \frac{(Sensibilidade + especificidade)}{2} \quad \text{Equação 8}$$

$$Sensibilidade = \frac{N_{\text{verdadeiros positivos}}}{N_{\text{verdadeiros positivos}} + N_{\text{falsos negativos}}} \quad \text{Equação 9}$$

$$\text{Especificidade} = \frac{N_{\text{verdadeiros negativos}}}{N_{\text{verdadeiros negativos}} + N_{\text{falsos positivos}}} \quad \text{Equação 10}$$

Valor de preditividade positiva (VPP):

$$\text{VPP} = \frac{N_{\text{verdadeiros positivos}}}{N_{\text{verdadeiros positivos}} + N_{\text{falsos positivos}}} \quad \text{Equação 11}$$

Valor de preditividade negativa (VPN):

$$\text{VPN} = \frac{N_{\text{verdadeiros negativos}}}{N_{\text{verdadeiros negativos}} + N_{\text{falsos negativos}}} \quad \text{Equação 12}$$

Coefficiente de correlação da validação cruzada ( $q^2$ ):

$$q^2 = 1 - \frac{\sum_{i=1}^n (Y_{i,exp} - Y_{i,pred})^2}{\sum_{i=1}^n (Y_{i,exp} - \langle Y \rangle_{exp})^2} \quad \text{Equação 13}$$

Raiz quadrada do erro médio (RMSE, do inglês, *root mean squared error*):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (Y_{i,exp} - Y_{i,pred})^2}{n - 1}} \quad \text{Equação 14}$$

Erro absoluto médio (MAE, do inglês, *mean absolute error*):

$$\text{MAE} = \frac{\sum_{i=1}^n |Y_{i,exp} - Y_{i,pred}|}{n} \quad \text{Equação 15}$$

no qual  $n$  é o número total de compostos,  $Y_{i,exp}$  é a atividade experimental do  $i$ -ésimo composto,  $Y_{i,pred}$  é a atividade predita para o  $i$ -ésimo composto e  $\langle Y \rangle_{exp}$  é a média da atividade experimental para todos os  $n$  compostos.

Os melhores modelos foram identificados e selecionados de acordo com valores aceitáveis de ACB ( $ACB \geq 0,6$ ) e  $q^2$  ( $q^2 \geq 0,6$ ) (TROPSHA, 2010) para os múltiplos conjuntos testes (chamados de conjunto *out-of-bag* no *Random Forest* (RF), veja próxima seção). Os modelos selecionados foram aplicados aos conjuntos de validação externa para prever suas propriedades experimentais. Esse procedimento foi repetido cinco vezes para garantir que cada composto passasse uma vez pelo conjunto de validação externa. Como a taxa de acerto de cada modelo é estimada apenas nos compostos do conjunto de validação externa, os quais nunca são usados para derivar, tendenciar, ou selecionar os modelos, esse protocolo garante uma estimativa objetiva da real preditividade externa dos modelos. Além disso, 1000 execuções de *Y-randomization* foram realizadas para garantir que a preditividade dos modelos gerados foi

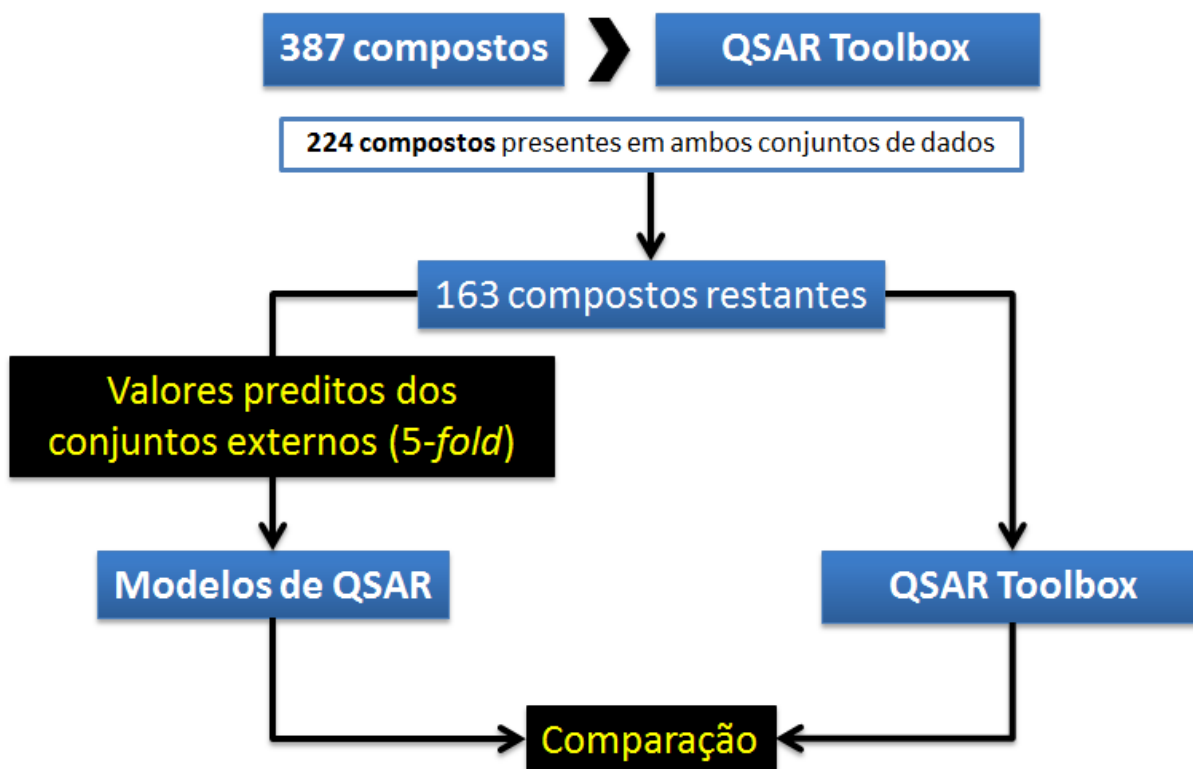
devida a correlações aleatórias. Nesse procedimento, a variável Y é aleatorizada e modelos são gerados. Caso a preditividade dos modelos com variável Y aleatória obtenham resultados melhores que os modelos de QSAR, então os modelos devem ser descartados, visto os descritores moleculares não descrevem bem Y.

### 3.7 Random Forest

Modelos foram gerados de acordo com o algoritmo original de RF (BREIMAN, 2001) usando o *software* CF versão 2.12 (POLISHCHUK et al., 2009). RF é um conjunto de árvores de decisão. O resultado de todas as árvores é agregado para obter uma predição final. Cada árvore cresce da seguinte maneira: (i) uma amostra extraída por *bootstrap* do conjunto modelagem contendo  $n$  compostos forma o conjunto treinamento para a árvore em questão. Os compostos que não estão no conjunto treinamento compõem o conjunto *out-of-bag* (OOB), com tamanho de aproximadamente  $n/3$ ; (ii) a melhor divisão dos descritores pelo algoritmo de classificação e regressão de árvores (CART) (BREIMAN et al., 1984) a partir dos  $m$  descritores selecionados de todo o conjunto de descritores é escolhida; (iii) cada árvore é gerada até a máxima extensão possível. Os valores preditos são definidos pela média das predições de todas as árvores. Assim, cada árvore prediz o valor para os compostos que não estão no conjunto treinamento daquela árvore (apenas OOB). O modelo final é escolhido pelo menor erro para predição do conjunto OOB. A abordagem de árvores locais para definição do DA foi usada para todos os modelos desenvolvidos nesse estudo. Nessa abordagem, uma matriz de distância entre os compostos do conjunto treinamento é gerada no espaço químico de cada modelo de QSAR. Em seguida, as menores distâncias entre os compostos são detectadas e o contorno do espaço químico é definido. A distância média ( $d_m$ ) e a raiz quadrada do desvio médio ( $\sigma$ ) dos compostos são calculadas e esses parâmetros definem a densidade média da distribuição molecular do espaço químico desse modelo. Por fim, os compostos do conjunto teste são projetados no espaço químico definido. Se a distância entre determinada estrutura e o ponto mais próximo do conjunto treinamento for maior que  $d_m + 3\sigma$ , então esse composto está fora do DA, visto que qualquer ponto três vezes além está  $\sigma$  muito distante da distribuição dos compostos. Caso contrário, a molécula está dentro do DA (ARTEMENKO et al., 2011).

### 3.8 OECD QSAR Toolbox

A eficiência dos modelos gerados nesse estudo foram comparados com o módulo de sensibilização da pele do OECD QSAR Toolbox versão 3.1. O *software* foi baixado da página da internet <http://www.qsartoolbox.org/download.html> e implementado de acordo com as instruções retiradas do manual do usuário. Para se fazer uma comparação justa entre os modelos gerados e ao módulo de sensibilização do *software*, excluíram-se os 224 compostos que estavam presentes tanto no conjunto de dados A quanto no banco de dados do QSAR Toolbox (com dados de LLNA). Posteriormente, aplicou-se o método de *read-across* implementado no *software* aos 163 compostos restantes do conjunto de dados A usando os caracterizadores para ligação proteica OASIS v1.1 e OECD. Os valores preditos do *read-across* foram comparados com predições feitas pelos modelos de QSAR desenvolvidos neste trabalho quando esses compostos estavam em conjuntos de validação externa (Figura 7).



**Figura 7.** Resumo do protocolo realizado para comparação dos modelos gerados nesse estudo com o módulo de sensibilização da pele do QSAR Toolbox.

## 4 RESULTADOS E DISCUSSÃO

### 4.1 Sensibilização da pele

#### 4.1.1 Caracterização dos conjuntos de dados

##### 4.1.1.1 Conjunto de dados de sensibilização da pele (conjunto de dados A)

O conjunto de dados retirado do relatório do ICCVAM (2009) continha dados recuperados de diferentes laboratórios e diferentes veículos para se atingir solubilidade adequada e conseqüente penetração na pele (Tabela 5). Contudo, a variabilidade interlaboratorial do LLNA é considerada baixa (SCHOLES et al., 1992) e apesar de existir uma variância nos valores causados por diferentes veículos, ela é insuficiente para mudar a categoria de sensibilização de um dado composto químico (sensibilizador/não sensibilizador) (ANDERSON; SIEGEL; MEADE, 2011). Por essa razão, decidiu-se integrar os dados obtidos com diferentes veículos no mesmo conjunto de dados. A análise do potencial binário para o registro de duplicatas mostrou boa concordância entre todos os dados, apresentando apenas três compostos com diferentes anotações. Buscando informação adicional, encontrou-se que o acrilato de etila (DEARMAN et al., 2007) e aldeído hexilcinâmico (DEARMAN et al., 2001) eram, de fato, sensibilizadores. Assim, apenas linalol foi excluído do conjunto de dados por falta de evidências adicionais. Ao final, 387 compostos (260 sensibilizadores e 127 não sensibilizadores) foram mantidos. O conjunto modelagem consistiu em 254 compostos (127 sensibilizadores e 127 não sensibilizadores). Os 133 sensibilizadores restantes foram mantidos como conjunto de validação externa.

**Tabela 5.** Número de compostos testados em LLNA por veículo.

<b>Tipo de veículo</b>	<b>No. de sensibilizadores</b>	<b>No. de não sensibilizadores</b>	<b>Total</b>
ACE	31	14	45
AOO	178	51	229
dH <sub>2</sub> O	2	2	4
DMF	27	40	67
DMSO	15	16	31
PG	8	6	14

**Tabela 5.** (Continuação).

<b>Tipo de veículo</b>	<b>No. de sensibilizadores</b>	<b>No. de não sensibilizadores</b>	<b>Total</b>
Pluronic L92 (1%)	5	2	7
Outros	7	4	11
Total	273	135	408

Nota: AOO, acetona e óleo de oliva (4:1 por volume); ACE, acetona; DMF, dimetilformamida; DMSO, dimetilsulfóxido; PG, propilenoglicol.

#### 4.1.1.2 *Conjunto de validação externa para sensibilização da pele (conjunto de dados B)*

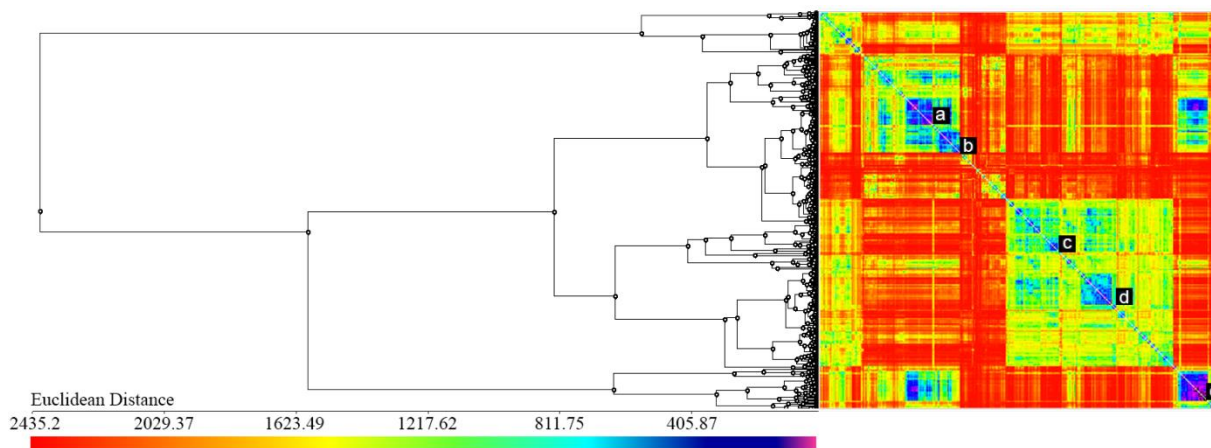
Dentre as 153 estruturas únicas recuperadas do trabalho de Jaworska e colaboradores (2011), 129 compostos estavam presentes também no conjunto de dados A, observando-se apenas duas divergências nas anotações (laurilsulfato e ácido tartárico), o que demonstra forte consistência dos dados. Os 24 compostos do conjunto de dados B que não estavam no conjunto modelagem (conjunto de dados A) foram selecionados para validação externa.

#### 4.1.1.3 *Conjunto de dados da Scorecard (conjunto de dados C)*

A análise inicial do conjunto de dados da Scorecard revelou 55 compostos já presentes no conjunto modelagem: 42 compostos sensibilizadores e 13 não sensibilizadores, o que mostra que nem todos os compostos contidos nesse conjunto são tóxicos para pele. Depois de remover duplicatas, sais inorgânicos e compostos organometálicos, 571 compostos foram mantidos para triagem virtual a fim de se identificar potenciais sensibilizadores.

#### 4.1.2 *Análise de agrupamentos hierárquicos do conjunto de dados A*

O mapa de calor representando a matriz de distâncias das estruturas e o dendograma do conjunto de dados A (utilizado na construção dos modelos de QSAR para sensibilização da pele) são apresentados na Figura 8. Foi possível identificar cinco grupos maiores (anotados como **a-e**) de compostos similares no mapa de calor. Todos os outros grupos identificados são bem menores (ca. 3-10 compostos). A análise do potencial sensibilizador para os compostos em grupos menores de alta similaridade estrutural revelou a consistência do conjunto de dados A e permitiu que fossem encontrados 24 compostos suspeitos de possuírem anotações de atividade diferentes do restante do conjunto de dados.



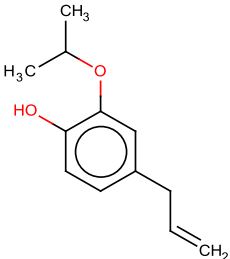
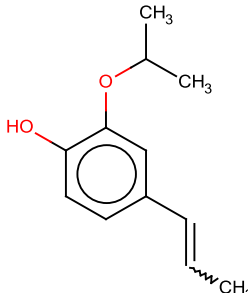
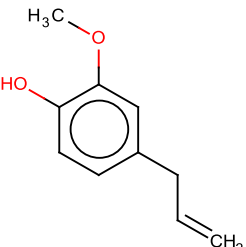
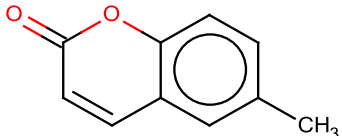
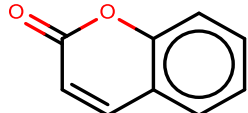
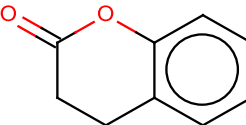
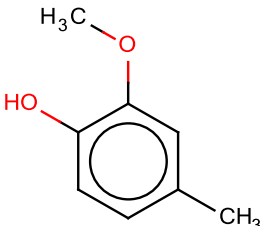
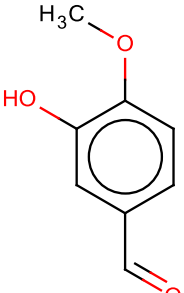
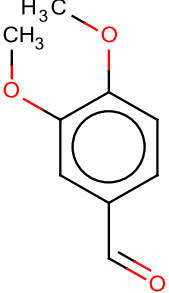
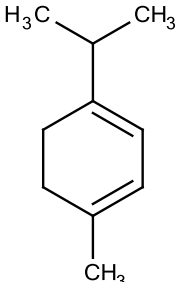
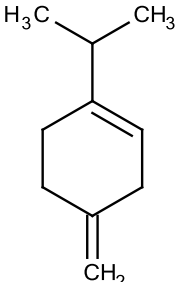
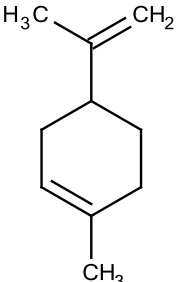
**Figura 8.** Análise agrupamentos hierárquicos para o conjunto de dados A: dendograma e mapa de calor da matriz de distâncias organizadas com base na similaridade estrutural (violeta/azul = similar; amarelo, vermelho = dissimilar).

Primeiramente, procurou-se por evidência adicional na literatura científica que poderia confirmar ou refutar a atividade dos compostos suspeitos. Os dados para o conjunto de dados A foram coletados do relatório do ICCVAM (2009) sobre o rLLNA, o qual consistiu de dados compilados de diversas fontes. A busca na literatura revelou que existe uma evidente carência de dados no potencial risco de compostos que causam sensibilização da pele. Muito comumente, as referências encontradas (ex., (GERBERICK et al., 2005)) foram citadas no relatório da ICCVAM. Foram encontrados dados adicionais para apenas cinco de 24 compostos [bromo-hexano (SIEGEL et al., 2009), bromononano (SIEGEL et al., 2009), 3,4-diidrocurarina (GERBERICK et al., 2004), 2-metoxi-4-metilfenol (GERBERICK et al., 2009) e ácido octanoico (JOHANSSON et al., 2011)]. As anotações para esses compostos foram todas confirmadas na literatura; assim eles representam casos interessantes de *cliffs* de atividade. Uma análise detalhada dos grupos mais interessantes está apresentada a seguir.

Foi possível observar que apenas um fenilpropanoide (entre 11) foi anotado como sensibilizador (Tabela 6 e grupo **a** na Figura 8). A única diferença entre isopropil eugenol (o único não sensibilizador) com seu vizinho estruturalmente mais próximo, isopropil isoeugenol, é a posição da dupla ligação no propileno ligado ao anel. O segundo vizinho mais próximo, eugenol, tem uma oximetila ao invés de uma oxi-isopropila na posição 1 do anel benzênico. Assim, o isopropil eugenol é um exemplo de um *cliff* de atividade e, como esperado, esse composto foi mal predito pelos modelos gerados. Pequenas alterações nos substituintes químicos e/ou disposições de ligações causam considerável mudança na atividade/toxicidade

de compostos. Por exemplo, a cafeína, que não é hepatotóxica, e a teofilina, hepatotóxica, diferem apenas por uma metila (LOW et al., 2011).

**Tabela 6.** Exemplos de compostos “suspeitos” no conjunto de dados A identificados por HCA.

		Compostos/Perfil toxicológico		
Fenil propanoides				
	Isopropil eugenol <b>Não sensibilizador</b>	Isopropil isoeugenol <b>Sensibilizador</b>	Eugenol <b>Sensibilizador</b>	
	Cumarinas			
6-Metilcumarina <b>Não sensibilizador</b>		Cumarina <b>Não sensibilizador</b>	3,4-Diidrocumarina <b>Sensibilizador</b>	
Anisóis				
	2-Metoxi-4-metilfenol <b>Sensibilizador</b>	3-Hidroxi-4-metoxibenzaldeído <b>Não sensibilizador</b>	Veratraldeído <b>Sensibilizador</b>	
	Terpenos			
$\alpha$ -Terpineno <b>Sensibilizador</b>		$\beta$ -Terpineno <b>Não sensibilizador</b>	Limoneno <b>Sensibilizador</b>	4-Isopropil-1-metileno-ciclohexano <b>Não sensibilizador</b>

Outro grupo foi gerado por ácidos dicarboxílicos (grupo **c**, Figura 8). Todos os compostos eram não sensibilizadores, exceto o ácido oxálico (também erroneamente predito pelos modelos gerados). Todos os bromoalcanos (grupos **d** e **e**, Figura 8) eram sensibilizadores, com exceção do bromononano. Suspeitou-se que ele poderia ter sido anotado erroneamente, então realizou-se uma busca por evidência publicada que poderia apoiar essa hipótese. Siegel e colaboradores (2009) realizaram um estudo para avaliar o impacto de propriedades físico-químicas e solubilidade de bromoalcanos na sensibilização da pele. Eles concluíram que o caso desse grupo de compostos é complicado, visto que, além da solubilidade lipídica, vários fatores, incluindo veículo, solvatação e retenção do composto na superfície da pele contribuem para o potencial sensibilizador aparente de 1-bromoalcanos no LLNA. Os autores também encontraram que logP não tem influência direta nos bromoalcanos e que os resultados dessa classe no LLNA pode ser afetado pela volatilidade dos compostos.

Em outras palavras, a falta de potência ou qualquer estímulo observável no LLNA por vários dos bromoalcanos semi-voláteis (C6-C11) é mais provável de ser devido à perda por volatilidade na superfície da pele competindo com absorção/distribuição dentro da epiderme e até mesmo haptenação. Li e colaboradores (2005) também enfrentaram problemas com o potencial sensibilizador de bromoalcanos. Bromononano foi mal predito em todos seus modelos devido às incertezas com sua real atividade. Assim, observa-se que bromoalcanos representam um caso interessante, e difícil, e que a diferença do bromononano com os demais representantes desse grupo pode se dar por particularidades experimentais dessa classe química com o LLNA. Entre os iodoalcanos, tanto o iodoexano (6C) quanto iodo-octadecano (18C) eram não sensibilizadores, enquanto iodononano (9C), iododecano (10C), iodotetradecano (14C) e iodoexadecano (16C) eram sensibilizadores. À primeira vista, parece que a cadeia longa do iodo-octadecano é responsável pela diminuição da permeabilidade pela pele (ou da afinidade de ligação por proteínas da pele). Como resultado, se essa teoria estivesse correta, o bromodocosano (22C) seria não sensibilizador, visto que o comprimento da cadeia é maior que o iodo-octadecano (18C) e o bromo tem características eletrônicas similares ao iodo. Contudo, o bromodocosano está anotado como sensibilizador no conjunto de dados A. Essa inconsistência dentro do grupo de haloalcanos levou à má predição de alguns compostos dentro desse grupo.

O composto 3,4-diidrocurmarina foi anotado como sensibilizador, enquanto dois compostos similares – curmarina e 6-metilcurmarina – são não sensibilizadores (Tabela 6 e grupo **a**, Figura 8). A única diferença entre a 3,4-diidrocurmarina e a curmarina é a falta de uma dupla ligação entre os átomos de carbono na posição 3 e 4. Gerberick e colaboradores (2004)

demonstraram que a 3,4-diidrocurarina é um pró-hapteno, ou seja, requer biotransformação para iniciar sua habilidade em reagir com aminoácidos nucleofílicos, como a cisteína, e proteínas da pele, causando conseqüentemente a sensibilização da pele. Assim, a 3,4-diidrocurarina representa um exemplo clássico de *cliff* de atividade e esse composto foi mal predito pelos modelos.

Os únicos dois sensibilizadores no grupo **a** (Figura 8) foram o 2-metoxi-4-metilfenol e veratraldeído. O primeiro apresenta evidências de ser um pró-hapteno e que requer biotransformação na pele para causar sensibilização e foi corretamente predito pelo modelo 5, enquanto o veratraldeído foi mal predito (GERBERICK et al., 2009). Dois compostos do grupo **b** na Figura 8, o 2-bromo-5-hidroxi-benzaldeído e o brometo de benzila, são sensibilizadores. O primeiro esteve fora do DA, enquanto o último foi mal predito. Os compostos no grupo de terpenos (Tabela 6) são sensibilizadores e não sensibilizadores dependendo das mútuas posições de substituintes do anel terpênico. O ácido octanoico (grupo **e**, Figura 8) foi o único não sensibilizador de seis ácidos e foi mal predito pelos modelos neste trabalho; a anotação foi posteriormente apoiada pela literatura (JOHANSSON et al., 2011).

O composto 2-benzil-*tert*-butilamino-3'-hidroximetil-4'-hidroxiaceto-fenona foi anotado como sensibilizador, enquanto seu vizinho mais próximo, 2-(benzil)*tert*-butilamino)-1-(alfa-4-diidroxim-*tolil*)etano, é um não sensibilizador e a diferença na atividade pode ser causado pela protonação do segundo. O 5-metil-6-(trifluorometil)indolina foi marcado como não sensibilizador, enquanto o composto 5-metoxi-6-trifluorometil-2,3-diidro-1*H*-indol é um sensibilizador. Enquanto o primeiro tem uma metila na posição 5, o último tem uma metoxila, o qual é provável responsável pela sensibilização na pele. Uma característica similar foi observada com 2,6-dimetoxi-4-metil-8-nitro-5-[3-(trifluorometil)fenoxi]quinolona (sensibilizador) e 2,6-dimetoxi-4-metil-5-[3-(trifluorometil)fenoxi]-8-quinolinamina (não sensibilizador). O primeiro tem um grupo nitro na posição 8, enquanto o último tem uma aminoquinolina.

#### 4.1.3 Modelos de QSAR gerados para sensibilização da pele

O desenvolvimento paralelo de múltiplos modelos de QSAR usando todas as combinações possíveis de diferentes tipos de descritores e vários algoritmos de aprendizado de máquina tem se mostrado superior ao uso de um único modelo (KUZ'MIN et al., 2009; ZHU et al., 2008). Contudo, apesar de evidente as vantagens de se empregar métodos de modelagem de QSAR em consenso (TROPSHA, 2010), estudos recentes (FOURCHES; MURATOV;

TROPSHA, 2010) mostraram que não existe necessidade de uma superabundância de modelos. Além disso, estudos mostraram que a influência da qualidade do conjunto de dados, dos descritores moleculares e dos algoritmos usados para gerar os modelos diminuem de acordo com a seguinte ordem: (i) precisão dos dados; (ii) descritores moleculares; (iii) algoritmo de aprendizado de máquina (FOURCHES; MURATOV; TROPSHA, 2010; YOUNG; YUAN; ZHU, 2012). Por essa razão, foi conduzido um cuidadoso preparo do conjunto de dados e cálculo de descritores SiRMS (baseados em fragmentos) e DRAGON (baseados em toda estrutura). Ademais, gerou-se modelos de RF separados para os descritores SiRMS e DRAGON e uniu-se os modelos em vários consensos que serão descritos a seguir.

Utilizou-se o método de validação externa cruzada *5-fold* para estimar a capacidade preditiva dos modelos gerados. Depois de dividir em conjunto treinamento e conjunto de validação externa, mas antes da modelagem, os descritores constantes (variância menor que 1%) e variáveis intercorrelacionadas foram removidos. O limiar de correlação foi determinado como  $|r| \geq 0,5$  para descritores SiRMS e  $|r| \geq 0,9$  para DRAGON. Esses valores foram determinados com base na experiência do grupo adquirida a partir de tentativa e erro na geração de melhores modelos usando esses descritores. Como cada RF é um consenso de um conjunto de 1000 árvores (modelos) de classificação, o valor final é dado pela média do resultado das árvores. Assim, o valor predito para cada composto foi um número contínuo entre 0 e 1. O resultado da predição foi arredondado para 0 (não sensibilizador) se a média das predições fosse menor que 0,5; caso contrário, foi considerado como 1 (sensibilizador).

Os resultados estatísticos dos modelos de QSAR gerados estão resumidos na Tabela 7. Como pode ser observado, a combinação de descritores SiRMS ou DRAGON com RF produziu modelos de QSAR robustos e preditivos (ACB = 0,71-0,76) para sensibilização da pele. O uso do DA resultou em uma alta diminuição da cobertura dos modelos (42-48%), mas aumentou o ACB dos modelos consideravelmente (9-12%).

**Tabela 7.** Valores estatísticos dos modelos de QSAR gerados para sensibilização da pele avaliados pelo método de validação externa *5-fold*.

Modelos de QSAR	Sensibilidade	Especificidade	ACB	Cobertura
Modelo 1 – SiRMS (DA)	0,83	0,83	0,83	0,58
Modelo 2 – SiRMS	0,69	0,73	0,71	1,00
Modelo 3 – Dragon (DA)	0,84	0,87	0,85	0,52
Modelo 4 – Dragon	0,70	0,81	0,76	1,00
Modelo 5 – Consenso (DA)	0,79	0,85	0,82	0,70

**Tabela 7.** (Continuação).

<b>Modelos de QSAR</b>	<b>Sensibilidade</b>	<b>Especificidade</b>	<b>ACB</b>	<b>Cobertura</b>
Modelo 6 – Consenso	0,74	0,83	0,79	0,82
Modelo 7 – Consenso Rigor	0,91	0,85	0,88	0,39

Como vários modelos foram gerados, a predição por consenso foi usada calculando-se a média dos valores preditos de cada modelo. Dessa maneira, o valor predito para cada composto poderia ser igual a 0, 0,5 ou 1. Todas as predições iguais a 0,5 foram consideradas inconclusivas e foram descartadas. No total, foram desenvolvidos três modelos por consenso: (i) modelo 5, que foi a média das predições feitas pelos modelos 1 e 3 (SiRMS e DRAGON, considerando o DA); (ii) modelo 6, que foi uma média das predições feitas pelos modelos 2 e 4 (SiRMS e DRAGON, sem estimação do DA); (iii) modelo 7, que foi a média das predições feita pelos modelos 1 e 3, mas apenas considerando os compostos que caíram dentro do DA para ambos os modelos. Em outras palavras, para o modelo 5 é suficiente que apenas um composto esteja dentro do DA de um único modelo (tanto 1 quanto 3), enquanto o modelo 7 é mais restritivo e considera apenas resultados concordantes quando cada composto cai dentro do DA de ambos os modelos 1 e 3. O modelo 6 foi um conjunto geral, que não considera o DA. O resumo dos valores estatísticos na Tabela 8 mostraram que o DA aumenta o ACB para os modelos SiRMS e DRAGON, assim como para o modelo de consenso 5, resultando em um aumento do ACB e diminuição da cobertura. O modelo mais preditivo, modelo 7, foi gerado usando a estimação do DA mais precisa, o que diminuiu a cobertura do modelo para 39%.

Os 133 compostos sensibilizadores excluídos do conjunto de dados A, baseado em uma busca de similaridade foram usados como validação externa juntamente com o conjunto de dados B. O conjunto de dados B foi recuperado do estudo de Jaworska e colaboradores (2011), no qual 24 compostos (19 sensibilizadores, divididos em forte, moderado e fraco e 5 não sensibilizadores) foram usados para validação externa adicional. Como pode ser observado na Tabela 8, o modelo 5 foi capaz de prever corretamente 65% dos compostos (cobertura de 50%), enquanto o modelo 7 conseguiu prever 73% dos compostos sensibilizadores corretamente. Infelizmente, a maioria dos compostos esteve fora do DA (cobertura de 24%) e o número pequeno de não sensibilizadores foi insuficiente para estimar o poder de predição do modelo para essa classe.

**Tabela 8.** Valores estatísticos dos modelos de QSAR para sensibilização da pele quando aplicados a um conjunto de validação externa (133 sensibilizadores do conjunto de dados A + conjunto de dados B).

<b>Modelo</b>	<b>Sensibilidade</b>	<b>VP</b>	<b>Especificidade</b>	<b>VN</b>	<b>ACB</b>	<b>Cobertura</b>
<b>Modelo 1</b> (SiRMS DA)	0,66	43	1,00	1	0,83	0,42
<b>Modelo 2</b> (SiRMS)	0,51	78	1,00	5	0,76	1,00
<b>Modelo 3</b> (Dragon DA)	0,69	36	0,50	1	0,60	0,34
<b>Modelo 4</b> (Dragon)	0,47	72	0,60	3	0,54	1,00
<b>Modelo 5</b> (Consenso DA)	0,65	51	1,00	1	0,83	0,5
<b>Modelo 6</b> (Consenso)	0,49	59	1,00	3	0,75	0,78
<b>Modelo 7</b> (Consenso Rigor)	0,73	27	1,00	1	0,86	0,24

Nota: Número total de positivos = 152; Número total de negativos = 5; VP = verdadeiros positivos; VN = verdadeiros negativos.

#### 4.1.4 Comparação dos modelos de QSAR desenvolvidos para sensibilização da pele com o algoritmo de sensibilização do QSAR Toolbox

A OECD desenvolveu um *software* chamado QSAR Toolbox para facilitar a aplicação prática de modelos de QSAR em contextos regulatórios por agências governamentais e indústrias a fim de se acelerar o processo de registros relacionados à segurança de compostos químicos. Esse *software* foi planejado para incorporar uma variedade de informações sobre compostos químicos de múltiplas fontes e para agrupar esses compostos baseados em suas estruturas moleculares, características e efeitos biológicos/toxicológicos relevantes. A sensibilização da pele é uma das propriedades incluídas no QSAR Toolbox. Apesar desse *software* ser principalmente usado por agências governamentais e indústrias para avaliação do potencial perigoso de compostos, é possível obter predições categóricas (sensibilizador ou não sensibilizador) para novos compostos.

Encontraram-se 224 compostos que estavam presentes tanto no conjunto de dados A quanto no conjunto de dados do QSAR Toolbox. Dentro do conjunto de dados do *software*, 30 compostos foram registrados como “ambíguo”, ou seja, eles possuíam potenciais sensibilizadores contraditórios reportados em diferentes estudos. Esses compostos foram excluídos da análise de duplicatas. 183 compostos dos 194 restantes possuíam anotações semelhantes entre os conjuntos de dados A e do QSAR Toolbox, resultando em uma concordância de 0,94. Aplicou-se o método de *read-across* implementado no *software* para o conjunto de 163 estruturas únicas do conjunto de dados A que não estavam presentes no banco de dados do QSAR Toolbox e compararam-se os resultados com as predições cegas feitas pelos modelos de QSAR gerados enquanto esses compostos estavam nos conjuntos de validação externa.

Os resultados da comparação (Tabela 9) mostram que o módulo de sensibilização do QSAR Toolbox superou o modelo 5 na sensibilidade (0,82 vs. 0,48, respectivamente), mas falhou em prever não sensibilizadores corretamente, como ilustrado pelo alto número de falsos positivos como mostra o baixo valor de preditividade positiva (VPP, Tabela 9). O modelo 5 atingiu uma especificidade de 0,98, contra 0,20 do QSAR Toolbox. A ACB do modelo 5 foi mais alta que a do QSAR Toolbox (0,73 vs. 0,51), mas com uma cobertura menor (0,60 vs. 0,92, respectivamente). Dessa maneira, observa-se que as estatísticas para esses 163 compostos estão em desacordo com aquelas da validação por *5-fold* (Tabela 7) e para o conjunto de dados externo (Tabela 8). Porém, são os resultados das mesmas predições. Essa discrepância principalmente na sensibilidade é devido à má amostragem do conjunto de dados A para comparação com o QSAR Toolbox. Em outras palavras, as estruturas químicas do conjunto de dados A que não estavam presentes na base de dados do QSAR Toolbox não representam a real preditividade dos modelos gerados.

Por essa razão, decidiu-se melhor avaliar esses dados estatísticos. Calcularam-se valores de preditividade positiva e negativa (VPP e VPN, respectivamente) para se estimar a probabilidade de uma predição correta de um novo composto como sensibilizador ou não sensibilizador. Como se observa na Tabela 9, se um composto é predito como sensibilizador pelos modelos 5 e 7 é quase garantido que ele realmente seja sensibilizador (VPP = 0,95-1,00). Ao mesmo tempo, existe 0,71 de chance de que um composto não sensibilizador seja predito corretamente. Enquanto isso, observa-se que o QSAR Toolbox gera uma alta taxa de falsos positivos e que a probabilidade de um composto sensibilizador ser corretamente predito é praticamente aleatória (VPP = 0,51). Surpreendentemente, apesar da baixa fração de falsos negativos, a probabilidade de um composto não sensibilizador ser predito corretamente também

é ruim (VPN = 0,52). Esse fato se dá devido à vasta maioria (0,80) de compostos não sensibilizadores terem sido preditos como sensibilizadores e ao número de verdadeiros negativos ser bem próximo ao número de falsos negativos.

**Tabela 9.** Comparação entre os modelos de QSAR para sensibilização da pele e o QSAR Toolbox.

Modelo	Sensibilidade	VPP	Especificidade	VPN	ACB	Cobertura
<b>Modelo 5</b> (Consenso DA)	0,48	0,95	0,98	0,71	0,73	0,60
<b>Modelo 7</b> (Consenso Rigor)	0,41	1,00	1,00	0,71	0,82	0,26
<b>QSAR Toolbox</b>	0,82	0,51	0,20	0,52	0,51	0,92

Investigando-se as razões pelas quais houve baixa sensibilidade dos modelos 5 e 7 para os compostos avaliados nesse exercício, encontrou-se que um terço dos compostos no conjunto de comparação poderiam ser considerados *cliffs* de atividade. Esses compostos são: (i) sete sensibilizadores fracos (4-bromo-1-ftalimidopentano, 2-benzil-*tert*-butilamino-3'-hidroximetil-4'-hidroxiaceto-fenona, 1,1-dimetiletil3-[[[(3s)-2,3,4,5-tetraidro-1-[2-[(1-metiletil)fenilamino]-2-oxoetil]-2,4-dioxo-5-fenil-1*H*-1,5-benzodiazepin-3-il]amino]carbonil]amino]benzoato, 5-metoxi-6-trifluorometil-2,3-diidro-1*H*-indol, ácido oxálico, bromo-hexano, e 3,4-dicloroanilina); (ii) dois sensibilizadores moderados (2,6-dimetoxi-4-metil-8-nitro-5-[3-(trifluorometil)-fenoxi]quinolona e veratraldeído); (iii) um sensibilizador forte (brometo de benzila); e (iv) e quatro compostos sem concentração especificamente determinada (1-(prop-2-enoiloxi)-2-[(prop-2-enoiloxi)metil]pentan-3-il-prop-2-enoato, azitromicina, 5-cloro-2,6-dimetoxi-4-metil-8-nitroquinolona e 1,5-dimetil-3-(1-oxo-2-propenil)-4-fenil-2-imidazolidinona). O coeficiente de Tanimoto (CT) para esses compostos e seus vizinhos mais próximos ao conjunto modelagem foi usualmente maior que 0,9 e nenhum foi menor que 0,75 (veja Tabela 10 para mais detalhes).

A maioria desses 14 compostos (ex., bromo-hexano e brometo de benzila) já foi discutida na seção 4.1.2 como sendo muito difíceis de predizer seu potencial sensibilizador. A diferença entre o sensibilizador veratraldeído e seus vizinhos mais próximos, anotados como não sensibilizadores, vanilina e 3-hidroxi-4-metoxibenzaldeído, é a substituição de um grupo químico (oximetila ao invés de hidroxila). O composto 2-benzil-*tert*-butilamino-3'-hidroximetil-4'-hidroxiaceto-fenona foi marcado como sensibilizador, enquanto seu vizinho 2-

(benzil)*tert*-butilamino)-1-( $\alpha$ ,4-diidroxi-*m*-tolil)etano (CT=0,92) é não sensibilizador. Essa mudança na atividade se deve pela protonação do último. Outro exemplo interessante de *cliff* de atividade é o 5-cloro-2,6-dimetoxi-4-metil-8-nitroquinolina. A única diferença entre esse composto e seu vizinho mais próximo, o não sensibilizador 5-cloro-2,6-dimetoxi-4-metilquinolina, é um grupo nitro, o qual pode ser o responsável direto pelo aparecimento do potencial sensibilizador do primeiro. Certamente, veratraldeído, brometo de benzila, 5-cloro-2,6-dimetoxi-4-metil-8-nitroquinolina assim como todos os outros casos de sensibilizadores que são *cliffs* de atividade serão preditos erroneamente por qualquer modelo de QSAR. Esse pode ser o principal motivo de vários sensibilizadores terem sido mal preditos pelos modelos gerados.

Além dos *cliffs* de atividade, encontraram-se cinco casos interessantes:  $\beta$ -felandreno (sensibilizador moderado), lauril lactilato sódico (sensibilizador fraco), 2,2-*bis*-[4-(2-hidroxi-3-metacriloxipropoxi)fenil]-propano (sensibilizador fraco), acrilato de 2-hidroxietila (sensibilizador moderado) e 3-etoxi-1-(2',3',4',5'-tetrametilfenil)propano-1,3-diona (sensibilizador fraco) (veja Tabela 10), os quais foram preditos como não sensibilizadores. Apesar desses compostos pertencerem a grupos que contêm tanto sensibilizadores quanto não sensibilizadores, eles possuem vizinhos (CT=0,75-0,96) que são sensibilizadores.  $\beta$ -felandreno pertencente ao grupo dos terpenos é um bom exemplo de um sensibilizador que foi predito erroneamente. O potencial sensibilizador desses compostos depende de diferentes disposições de ligações duplas e substituintes. Nesse caso, limoneno – o vizinho estruturalmente mais similar (CT=0,96) do  $\beta$ -felandreno – é um sensibilizador, mas o segundo vizinho mais próximo, 4-isopropil-1-metileno-ciclo-hexano (CT=0,94) não é sensibilizador. Essa alta similaridade entre um sensibilizador e um não sensibilizador reforça a dificuldade de se predizer corretamente os terpenos.

Sete outros compostos sensibilizadores erroneamente preditos: base de Bandrowski, 3,3',4',5-tetracloro-salicilanilida, Dinocap, ácido monometanossulfonato 7-[(4z)-3-(aminometil)-4-(metoxiimino)-1-pirrolidinil]-1-ciclopropil-6-fluoro-1,4-diidro-4-oxo-1,8-naftiridina-3-carboxílico, etilenodiamina, 8-cloro-3-pentil-3,7-diidro-1*H*-purina-2,6-diona e sulfóxido de dimetila (veja Tabela 10) foram bem dissimilares do conjunto modelagem (CT<0,5). Apesar de uma dissimilaridade alta poder ser a razão das predições erradas, esses compostos ainda estavam dentro do DA.

**Tabela 10.** Vizinho mais próximo dos sensibilizadores erroneamente preditos pelo modelo 5.

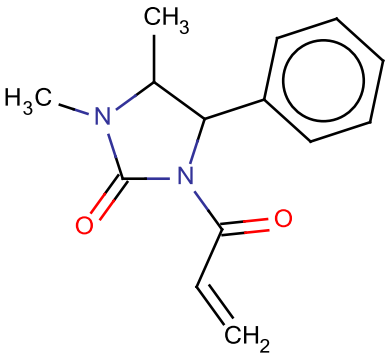
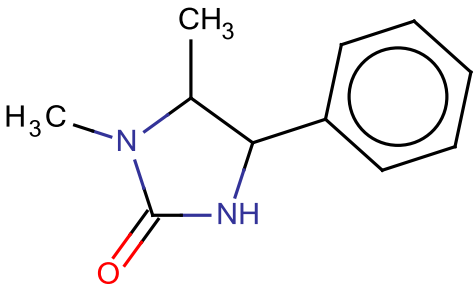
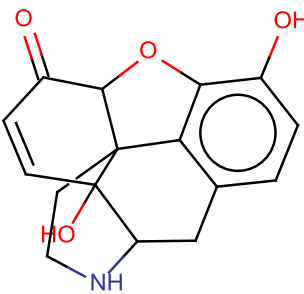
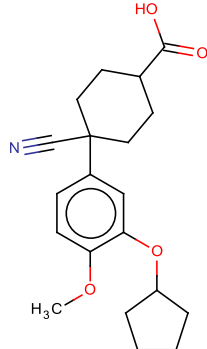
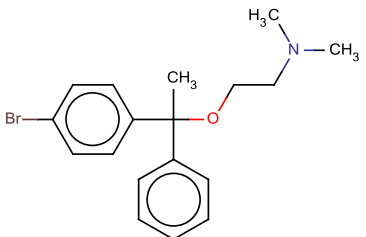
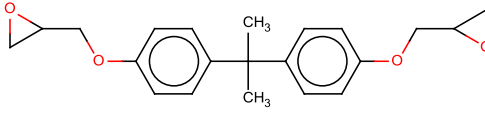
Sensibilizadores mal preditos	CT	Primeiro vizinho
<p><b>1</b></p>  <p>1,5-dimetil-3-(1-oxo-2-propenil)-4-fenil-2-imidazolidinona (potência não definida)</p>	0,765	 <p>1,5-dimetil-4-fenil-2-imidazolidinona (não sensibilizador)</p>
<p><b>2</b></p>  <p>14-hidroxinor-morfinona (fraco)</p>	0,758	 <p>Ácido 4-ciano-4-[3-(ciclopentiloxi)-4-metoxifenil]ciclo-hhexanocarboxílico (não sensibilizador)</p>
<p><b>3</b></p>  <p>2-[1-(4-bromofenil)-1-feniletoxi]-N,N-dimetil-etanamina (moderado)</p>	0,578	 <p>Diglicidil éter de bisfenol A (sensibilizador)</p>

Tabela 10. (Continuação).

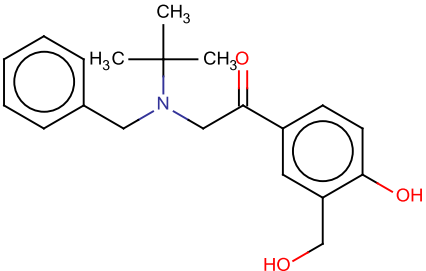
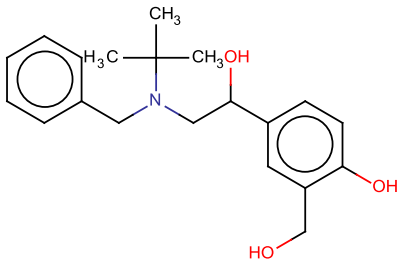
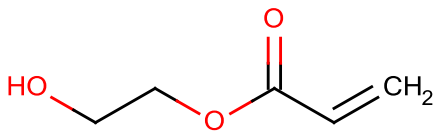
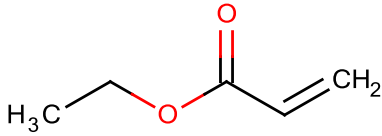
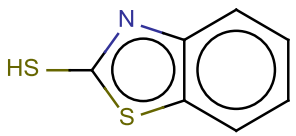
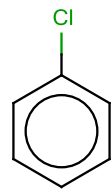
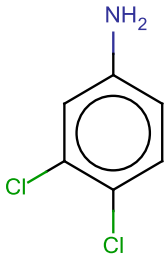
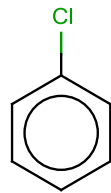
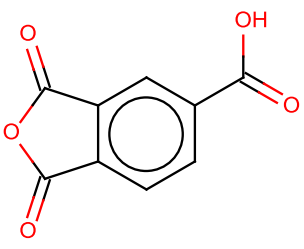
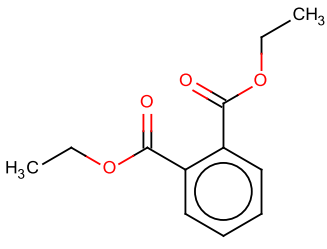
	Sensibilizadores mal preditos	CT	Primeiro vizinho
4	 <p>2-benzil-<i>tert</i>-butilamino-3'-hidroximetil-4'-hidroxiaceto-fenona (fraco)</p>	0,924	 <p>2-(benzil)<i>tert</i>-butilamino)-1-(alfa,4-diidroxi-<i>m</i>-tolil)etano (não sensibilizador)</p>
5	 <p>Acrilato de 2-hidroxi etila (moderado)</p>	0,825	 <p>Acrilato de etila (sensibilizador)</p>
6	 <p>2-mercapto-benzotiazola (moderado)</p>	0,609	 <p>Clorobenzeno (não sensibilizador)</p>
7	 <p>3,4-Dicloroanilina (fraco)</p>	0,815	 <p>Clorobenzeno (não sensibilizador)</p>
8	 <p>Anidrido benzeno-1,3,4-tricarboxílico (moderado)</p>	0,623	 <p>Ftalato de dietila (não sensibilizador)</p>

Tabela 10. (Continuação).

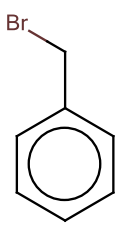
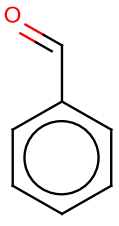
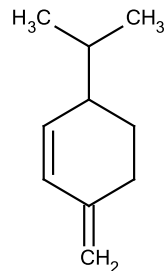
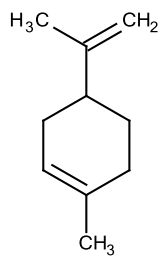
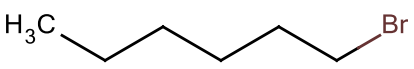
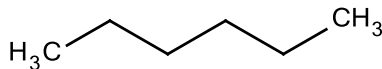
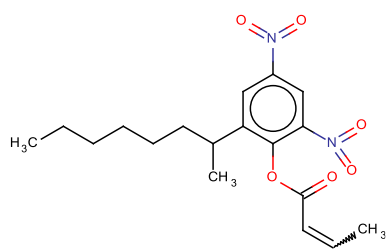
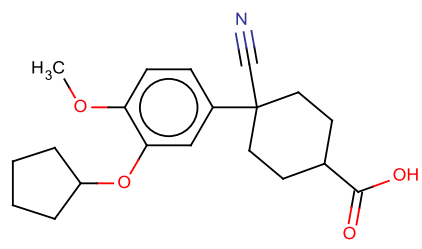
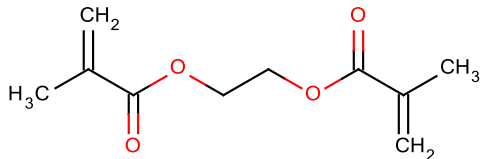
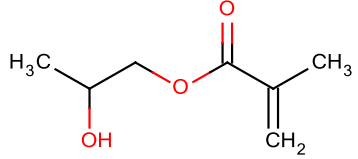
	Sensibilizadores mal preditos	CT	Primeiro vizinho
9	 <p>Brometo de benzila (forte)</p>	0,805	 <p>Benzaldeído (não sensibilizador)</p>
10	 <p><math>\beta</math>-Felandreno (moderado)</p>	0,961	 <p>Limoneno (sensibilizador)</p>
11	 <p>Bromo-hexano (fraco)</p>	0,902	 <p>Hexano (não sensibilizador)</p>
12	 <p>Dinocap (forte-moderado)</p>	0,432	 <p>Ácido 4-ciano-4-[3-(ciclopentiloxy)-4-metoxifenil]ciclohexanocarboxílico (não sensibilizador)</p>
13	 <p>Dimetacrilato de etileno (fraco)</p>	0,745	 <p>Metacrilato de 2-hidroxipropila (não sensibilizador)</p>

Tabela 10. (Continuação).

	Sensibilizadores mal preditos	CT	Primeiro vizinho
14	<p>Fluoresceina isotiocianato (extremo)</p>	0,628	<p>Diglicidil éter de bisfenol A (sensibilizador)</p>
15	<p>Lauril galato (extremo)</p>	0,69	<p>Miristato de isopropila (sensibilizador)</p>
16	<p><i>N</i>-(4-metoxifenil)-3-oxobutanamida (moderado)</p>	0,674	<p>6-Metoxi-4-metil-2(1<i>H</i>)-quinolinona (não sensibilizador)</p>
17	<p>Hipofluorito de biciclo[2.2.1]hept-5-em-2-il (fraco)</p>	0,64	<p><math>\beta</math>-terpineno (não sensibilizador)</p>
18	<p>(4<i>E</i>)-4-(etoximetileno)-2-fenil-1,3-oxazol-5(4<i>H</i>)-ona (potência não definida)</p>	0,556	<p>Benzocaína (não sensibilizador)</p>

Tabela 10. (Continuação).

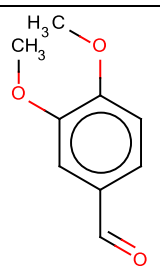
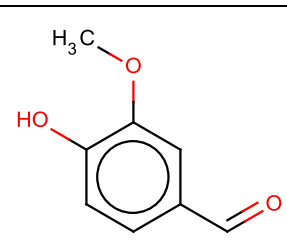
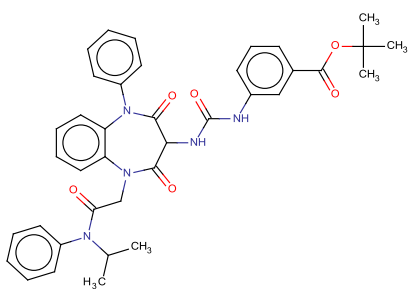
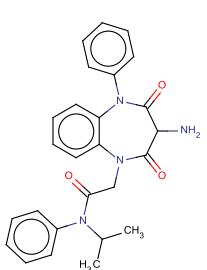
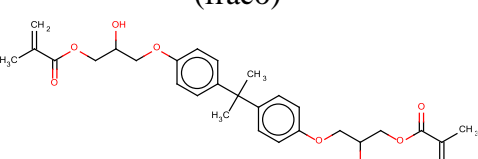
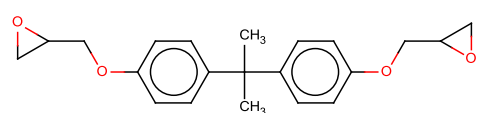
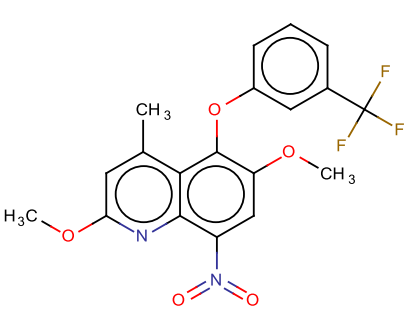
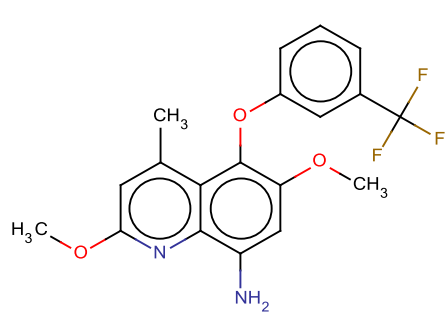
	Sensibilizadores mal preditos	CT	Primeiro vizinho
19	 <p>Veratraldeído (moderado)</p>	0,927	 <p>Vanilina (não sensibilizador)</p>
20	 <p>1,1-dimetiletil-3-[[[(3s)-2,3,4,5-tetrahydro-1-[2-[(1-metiletil)fenilamino]-2-oxoetil]-2,4-dioxo-5-fenil-1H-1,5-benzodiazepin-3-il]amino]carbonil]amino]benzoato (fraco)</p>	0,863	 <p>3-amino-2,3,4,5-tetrahydro-N-(1-metiletil)-2,4-dioxo-N,5-difenil-1H-1,5-benzodiazepina-1-acetamida (não sensibilizador)</p>
21	 <p>2,2-bis-[4-(2-hidroxi-3-metacriloxipropoxi)fenil]-propano (fraco)</p>	0,832	 <p>Diglicidil éter de Bisfenol A (sensibilizador)</p>
22	 <p>2,6-dimetoxi-4-metil-8-nitro-5-[3-(trifluorometil)-fenoxi]quinolina (moderado)</p>	0,943	 <p>2,6-dimetoxi-4-metil-5-[3-(trifluorometil)-fenoxi]-8-quinolinamina (não sensibilizador)</p>

Tabela 10. (Continuação).

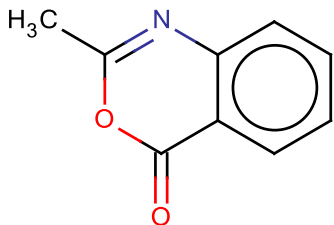
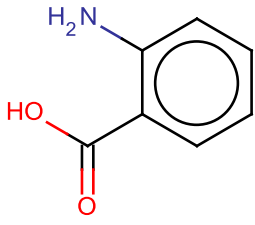
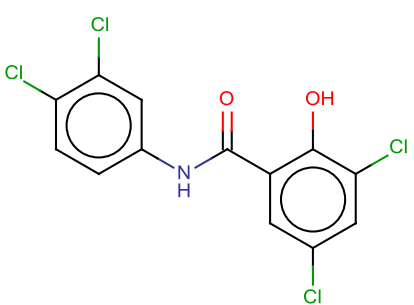
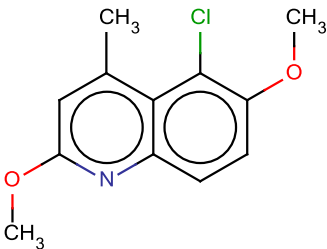
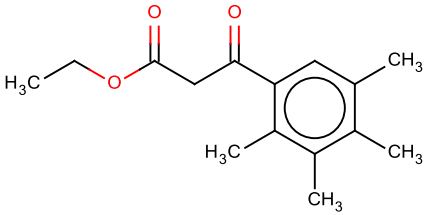
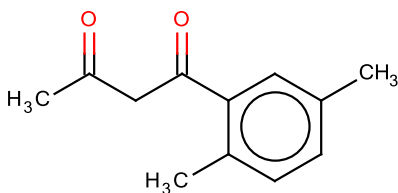
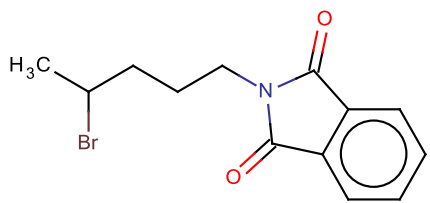
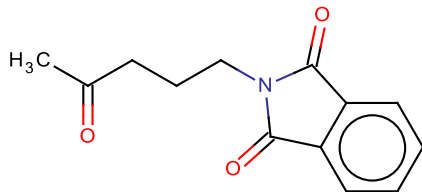
	Sensibilizadores mal preditos	CT	Primeiro vizinho
23	 <p>2-metil-4<i>H</i>-3,1-benzoxazin-4-ona (extremo)</p>	0,77	 <p>Ácido antranilínico (não sensibilizador)</p>
24	 <p>3, 3', 4', 5-tetracloro-salicilanilida (extremo)</p>	0,496	 <p>5-cloro-2,6-dimetoxi-4-metilquinolina (não sensibilizador)</p>
25	 <p>3-Etoxi-1-(2',3',4',5'-tetrametilfenil)propano-1,3-diona (fraco)</p>	0,749	 <p>1-(2,5-dimetilfenil)-1,3-butanodiona (sensibilizador)</p>
26	 <p>4-Bromo-1-ftalimidopentano (fraco)</p>	0,935	 <p>2-(4-oxopentil)-1<i>H</i>-isindol-1,3(2<i>H</i>)-diona (não sensibilizador)</p>

Tabela 10. (Continuação).

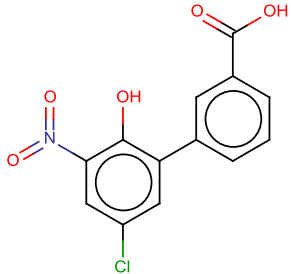
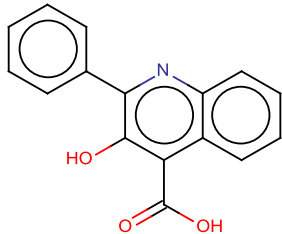
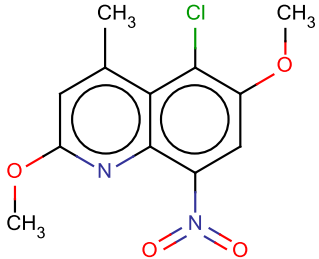
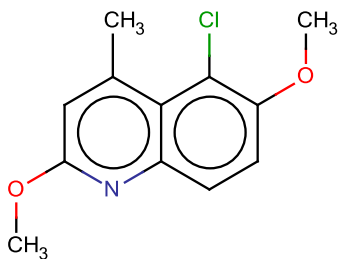
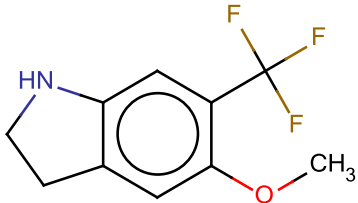
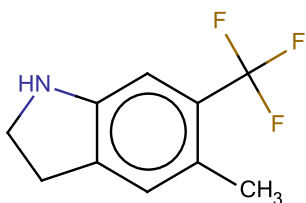
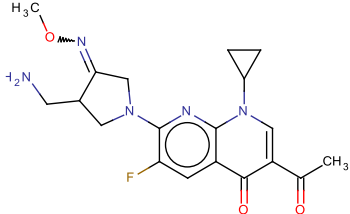
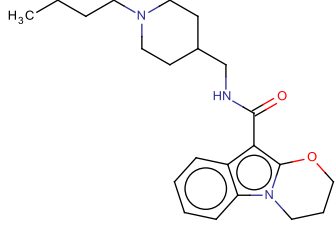
	Sensibilizadores mal preditos	CT	Primeiro vizinho
27	 <p>Ácido 5'-cloro-2'-hidroxi-3'-nitro-[1,1'-bifenil]-3-carboxílico (fraco)</p>	0,696	 <p>Ácido 3-hidroxi-2-fenil-4-quinolinecarboxílico (não sensibilizador)</p>
28	 <p>5-cloro-2,6-dimetoxi-4-metil-8-nitroquinolina (potência não definida)</p>	0,797	 <p>5-cloro-2,6-dimetoxi-4-metilquinolina (não sensibilizador)</p>
29	 <p>5-metoxi-6-trifluorometil-2,3-diidro-1H-indol (fraco)</p>	0,809	 <p>5-metil-6-(trifluorometil)indolina (não sensibilizador)</p>
30	 <p>Ácido monometano sulfonato 7-[(4z)-3-(aminometil)-4-(metoxiimino)-1-pirrolidinil]-1-ciclopropil-6-fluoro-1,4-diidro-4-oxo-1,8-naftiridine-3-carboxílico (moderado)</p>	0,357	 <p>N-[(1-butil-4-piperidinil)metil]-3,4-diidro-2H-[1,3]oxazino[3,2-a]indol-10-carboxamida (não sensibilizador)</p>

Tabela 10. (Continuação).

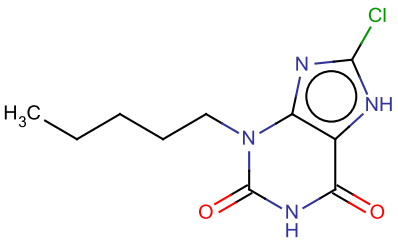
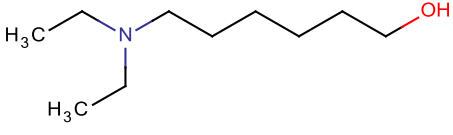
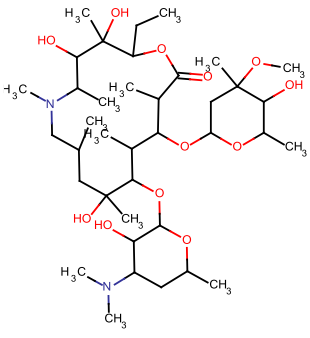
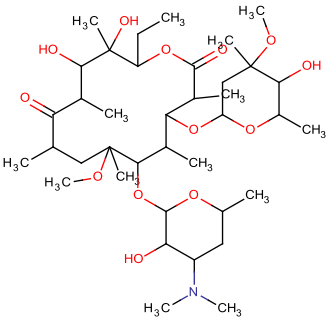
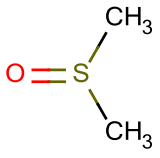
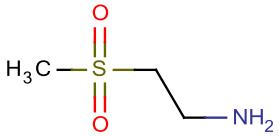
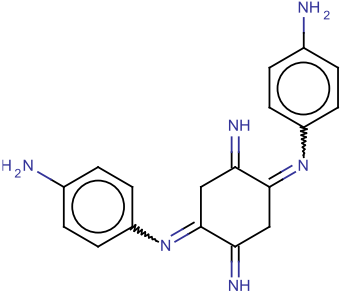
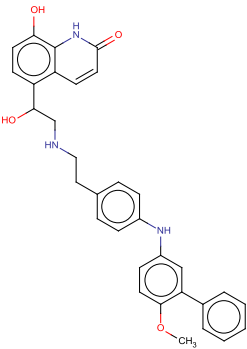
	Sensibilizadores mal preditos	CT	Primeiro vizinho
31	 <p>8-cloro-3-pentil-3,7-diidro-1H-purine-2,6-diona (fraco)</p>	0,376	 <p>6-(dietilamino)-1-hexanol (sensibilizador)</p>
32	 <p>Azitromicina (potência não definida)</p>	0,918	 <p>Claritromicina (não sensibilizador)</p>
33	 <p>Sulfóxido de dimetila (fraco)</p>	0,417	 <p>2-aminoetil metil sulfona (não sensibilizador)</p>
34	 <p>Base de Bandrowski (extremo)</p>	0,311	 <p>8-hidroxi-5-{1-hidroxi-2-[(2-{4-[(4-metoxi-3-fenilfenil)amino]etil)amino]etil}amino]etil}amino]quinolin-2-ona (não sensibilizador)</p>

Tabela 10. (Continuação).

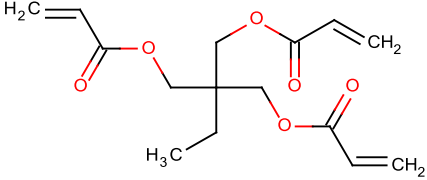
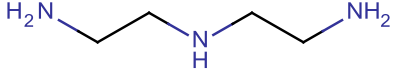
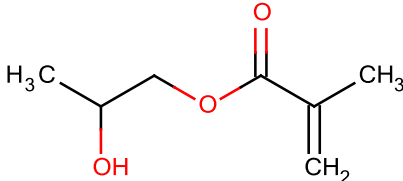
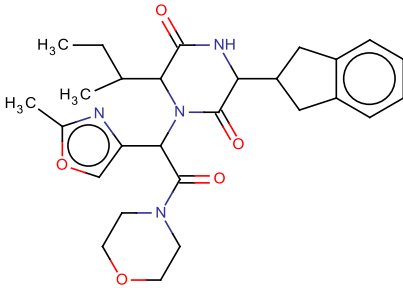
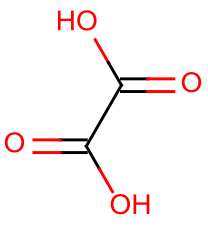
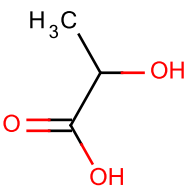
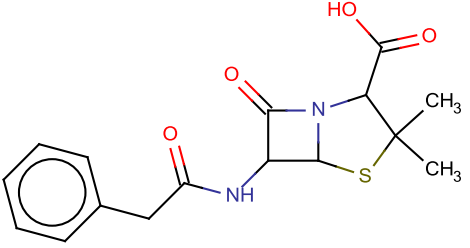
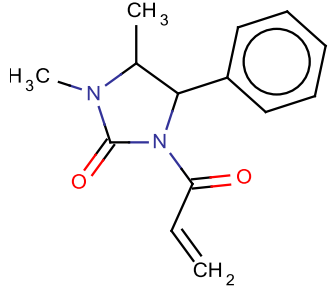
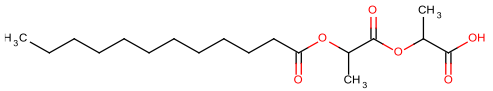
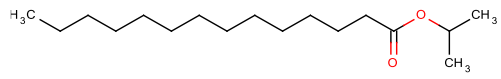
Sensibilizadores mal preditos	CT	Primeiro vizinho
35	0,948	
<p>1-(prop-2-enoiloxi)-2-[(prop-2-enoiloxi)metil]pentan-3-il prop-2-enoato (potência não definida)</p>	<p>Triacrilato de etil-2-(hidroximetil)-1,3-propanodiol (não sensibilizador)</p>	
36	0,444	
Etilenediamina (moderado)	Dietilenetriamina (sensibilizador)	
37	0,683	
Gliceril tioglicolato (moderado)	Metacrilato de 2-hidroxiopropila (não sensibilizador)	
38	0,561	
<p><i>N</i>-[2-(dietilamino)etil]-2-[[4-fluorofenil)-metil]tio]-4,5,6,7-tetrahidro-4-oxo-<i>N</i>-[[4'-(trifluorometil)-[1,1'-bifenil]-4-il]metil]-1<i>H</i>-ciclopentapirimidina-1-acetamida (fraco)</p>	<p>3-(2,3-diidro-1<i>H</i>-inden-2-il)-1-[1-(2-metil-1,3-oxazol-4-il)-2-(4-morpholinil)-2-oxoetil]-6-[1-metilpropil]-2,5-piperazinediona (não sensibilizador)</p>	

Tabela 10. (Continuação).

	Sensibilizadores mal preditos	CT	Primeiro vizinho
39	 <p>Ácido oxálico (fraco)</p>	0,677	 <p>Ácido láctico (não sensibilizador)</p>
40	 <p>Penicilina G (fraco)</p>	0,508	 <p>1,5-dimetil-3-(1-oxo-2-propenil)-4- fenil-2-imidazolidinona (sensibilizador)</p>
41	 <p>Lauril lactilato sódico (fraco)</p>	0,833	 <p>Miristato de isopropila (sensibilizador)</p>

#### 4.1.5 Triagem virtual do conjunto de dados da Scorecard

O propósito principal da análise de QSAR é a aplicação dos modelos desenvolvidos para a identificação de novos compostos com propriedades desejadas ou eliminar estruturas com propriedades desfavoráveis (CHERKASOV et al., 2014). O conjunto de dados da Scorecard foi selecionado para aplicação dos modelos de sensibilização com o objetivo de se identificar potenciais toxicantes. Os 571 compostos foram triados pelos modelos de QSAR gerados. De acordo com o modelo 5, 82 compostos foram preditos como sensibilizadores e 166 como não sensibilizadores (o restante dos compostos ficou fora do DA). O modelo 7 resultou em 27 sensibilizadores e 29 não sensibilizadores. Uma estimativa mais rigorosa do DA diminuiu a cobertura para 11%, mas a preditividade é possivelmente mais alta. Todos os compostos e seus

resultados correspondentes das predições feitas pelos modelos 5 e 7 estão apresentadas na Tabela A1 (Anexos).

A análise do vizinho mais próximo para os 27 compostos preditos como sensibilizadores pelo modelo 7 (Tabela 11) mostra alta similaridade entre esses compostos com os compostos presentes no nosso conjunto modelagem. Como podem ser observados, todos os compostos do conjunto modelagem (conjunto de dados A) são sensibilizadores. Esses compostos são candidatos para validação experimental dos modelos.

**Tabela 11.** Vizinhos mais próximos dos potenciais sensibilizadores preditos pelo modelo 7.

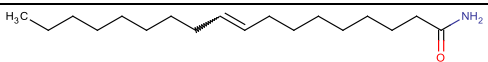
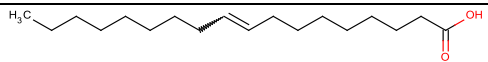

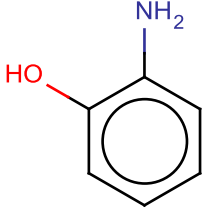
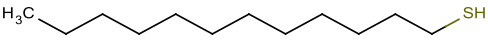
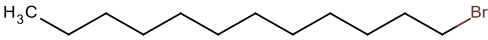
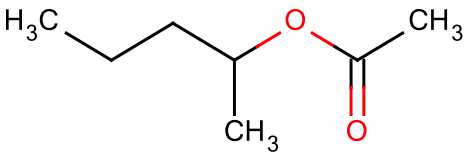
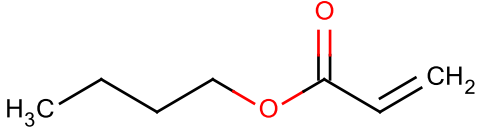
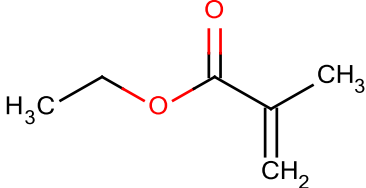
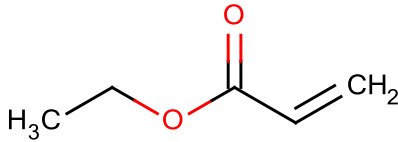
	Scorecard	CT	Conjunto modelagem (A)
1	 (Z)-9-octadecenamida	0,984	 Ácido oleico (sensibilizador)
2	 1-amino-4-hidroxibenzeno	0,981	 2-aminofenol (sensibilizador)
3	 1-dodecanotiol	0,974	 Bromododecano (sensibilizador)
4	 1-metilbutil acetato	0,803	 Acrilato de butila (sensibilizador)
5	 2-metil-2-propenoico ácido, etil éster	0,821	 Acrilato de etila (sensibilizador)

Tabela 11. (Continuação).

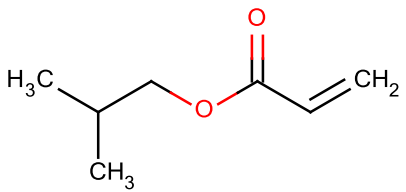
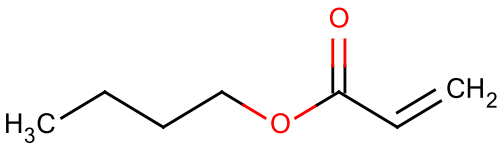
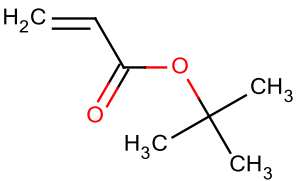
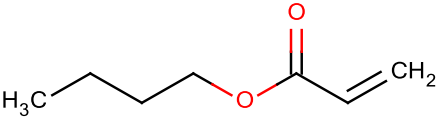
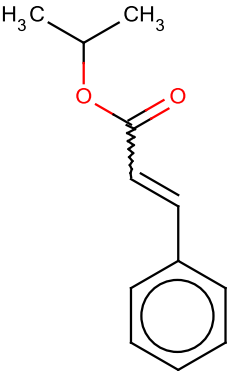
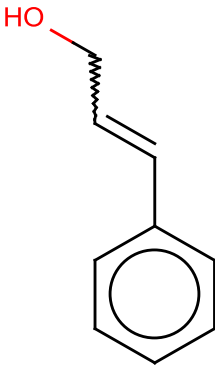
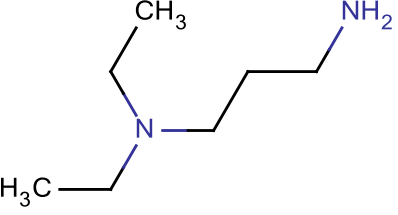
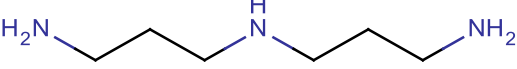
	Scorecard	CT	Conjunto modelagem (A)
6	 <p>2-metilpropil 2-propenoato</p>	0,838	 <p>Acrilato de butila (sensibilizador)</p>
7	 <p>2-propenoico ácido, 1,1-dimetil etil éster</p>	0,663	 <p>Acrilato de butila (sensibilizador)</p>
8	 <p>Ácido 2-propenoico, 3-fenil-, 1-metil etil éster</p>	0,710	 <p>Álcool cinâmico (sensibilizador)</p>
9	 <p>3-aminopropil dietilamina</p>	0,752	 <p>Dipropileno triamina (sensibilizador)</p>

Tabela 11. (Continuação).

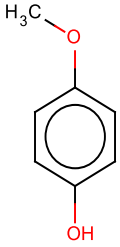
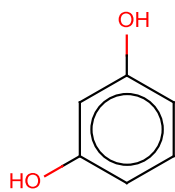
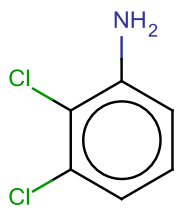
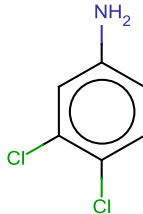
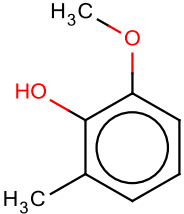
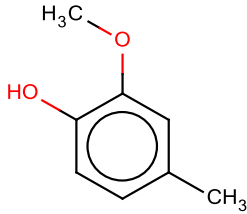
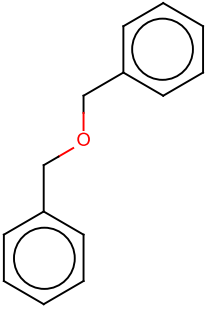
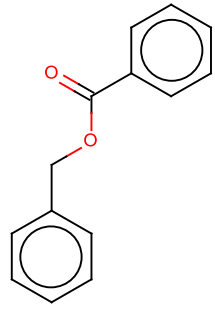
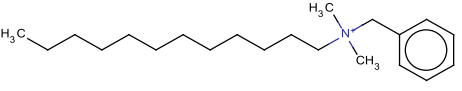
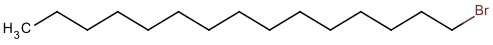
	Scorecard	CT	Conjunto modelagem (A)
10	 <p>4-metoxifenol</p>	0,898	 <p>Resorcinol (sensibilizador)</p>
11	 <p>Dicloroanilina</p>	0,985	 <p>3,4-Dicloroaniline (sensibilizador)</p>
12	 <p>2-metoxi-2-metilfenol</p>	0,987	 <p>2-metoxi-4-metilfenol (sensibilizador)</p>
13	 <p>Éter de hidroquinona monobenzílica</p>	0,958	 <p>Benzoado de benzila (sensibilizador)</p>
14	 <p>Lauril dimetil benzil amônio</p>	0,624	 <p>Bromopentadecano (sensibilizador)</p>

Tabela 11. (Continuação).

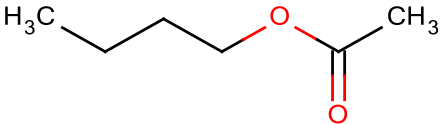
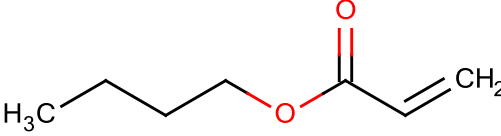
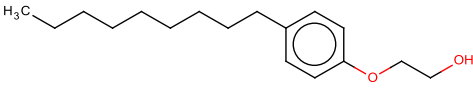
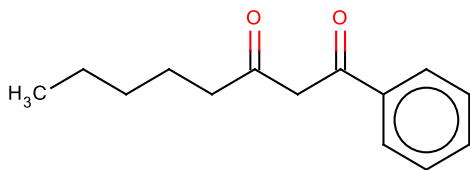
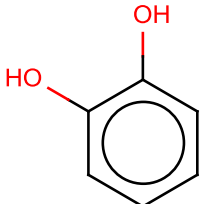
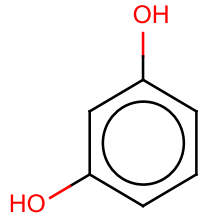
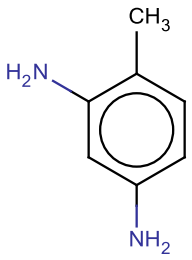
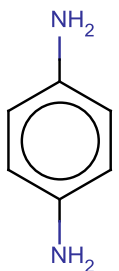
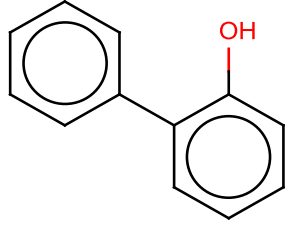
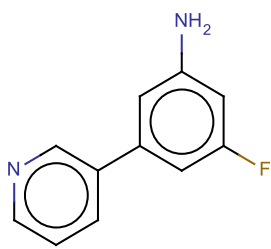
	Scorecard	CT	Conjunto modelagem (A)
15	 <p><i>N</i>-butil acetato</p>	0,873	 <p>Acrilato de butila (sensibilizador)</p>
16	 <p>Nonoxinol</p>	0,738	 <p>1-feniloctano-1,3-diona (sensibilizador)</p>
17	 <p>Catecol</p>	0,984	 <p>Resorcinol (sensibilizador)</p>
18	 <p>2,4-diaminotolueno</p>	0,896	 <p>4-fenileno-diamina (sensibilizador)</p>
19	 <p>2-fenilfenol</p>	0,798	 <p>3-fluoro-5-(3-piridinil)benzenamina (sensibilizador)</p>

Tabela 11. (Continuação).

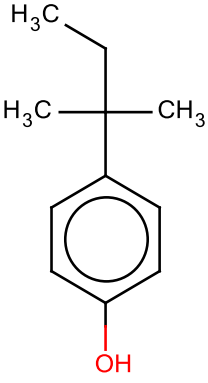
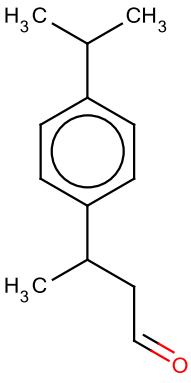
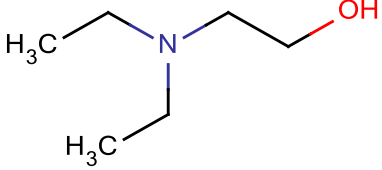
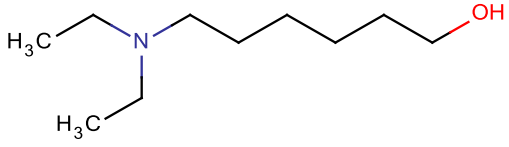
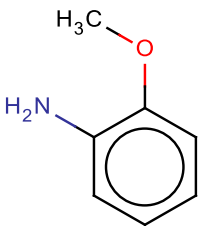
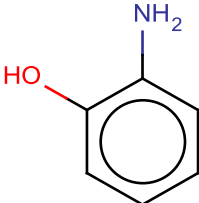
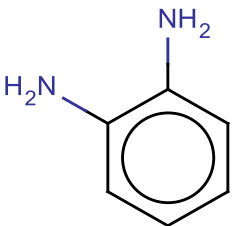
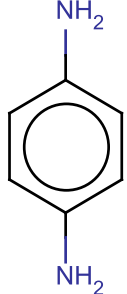
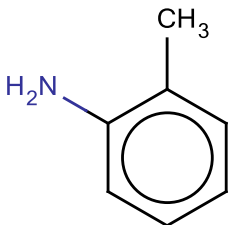
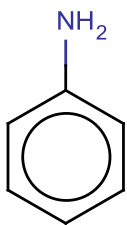
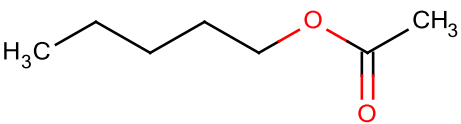
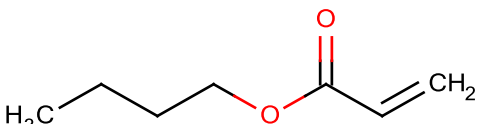
	Scorecard	CT	Conjunto modelagem (A)
20	 <p><i>p</i>-tert-amifenol</p>	0,803	 <p>Aldeído cyclamen (sensibilizador)</p>
21	 <p>2-(dietilamino)etanol</p>	0,534	 <p>6-(dietilamino)-1-hexanol (sensibilizador)</p>
24	 <p><i>o</i>-anisidine</p>	0,896	 <p>2-aminofenol (sensibilizador)</p>
25	 <p><i>o</i>-fenilenediamina</p>	0,984	 <p>4-fenilene-diamina (sensibilizador)</p>

Tabela 11. (Continuação).

	Scorecard	CT	Conjunto modelagem (A)
26	 <p><i>o</i>-toluidina</p>	0,892	 <p>Anilina (sensibilizador)</p>
27	 <p>Acetato de amila</p>	0,809	 <p>Acrilato de butila (sensibilizador)</p>

## 4.2 Permeabilidade da pele

### 4.2.1 Caracterização dos conjuntos de dados

#### 4.2.1.1 Conjunto de dados da permeabilidade da pele humana (conjunto de dados D)

Durante o processo de preparo desse conjunto de dados, três compostos foram removidos; tanto o estireno ( $\log K_p = -0,19$ ) quanto o etil benzeno ( $\log K_p = 0,08$ ) foram identificados como *outliers*, ou seja, seus valores de  $\log K_p$  são muito distantes da extensão de atividade dos outros compostos do conjunto de dados, que variam de -5,52 a -0,69. O RF é um algoritmo que é bastante sensível à lacunas de atividade (lacuna  $\geq 0,5$  unidades logarítmicas). Por esse motivo compostos com atividades extremas ( $\geq 0,5$  escalas logarítmicas) precisam ser removidos (BREIMAN, 2001). A digitoxina (produto natural) se apresentou como um *outlier* estrutural. Ao final, 186 compostos foram mantidos para modelagem.

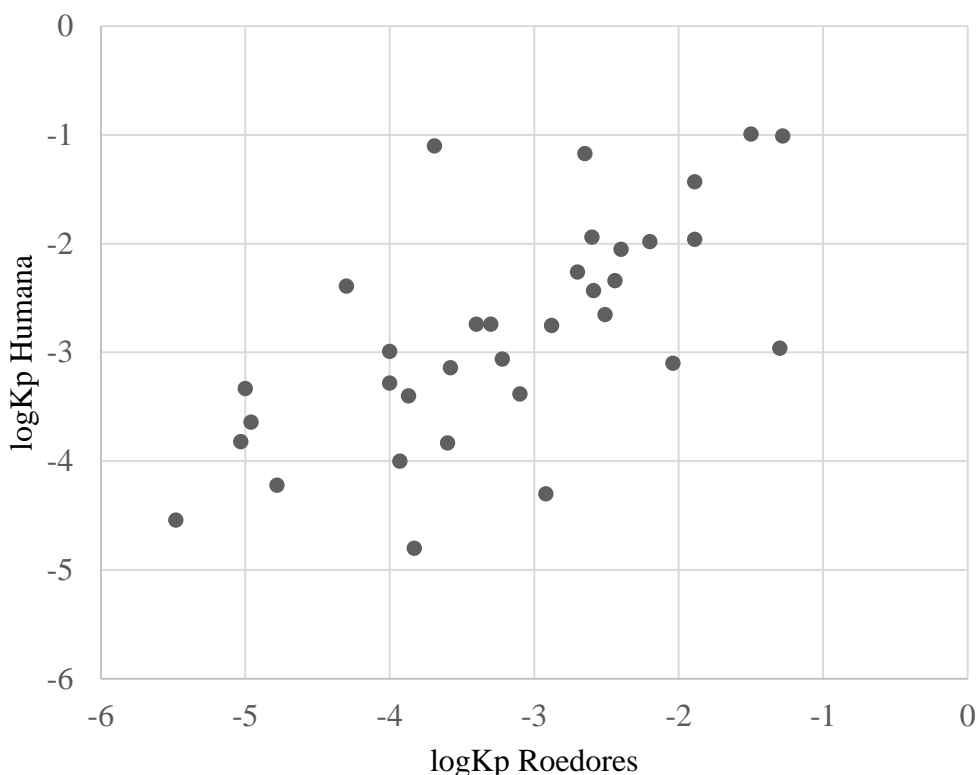
#### 4.2.1.2 Conjunto de dados da permeabilidade da pele de roedores (conjunto de dados E)

Após o preparo desse conjunto de dados, 96 compostos foram mantidos para modelagem. Os seguintes compostos foram removidos do conjunto de dados: diglicidil éter de bisfenol A (-5,26), óxido de decabromodifenila (-5,15), 4-*N*-butilamina (-0,64), bufexamaca (-

0,57) e triclosana (0,13). A extensão geral do  $\log K_p$  para esse conjunto de dados variou de -4,85 a -0,94.

#### 4.2.2 Análise da sobreposição dos dados de permeabilidade humana e de roedores

Encontrou-se um pequeno número de compostos (34) que possuíam dados de permeabilidade da pele humana e de roedores. A correlação geral da permeabilidade da pele medida nas duas espécies não foi alta ( $r^2 = 0,44$ , Figura 9 **Error! Reference source not found.**).



**Figura 9.** Correlação entre os dados experimentais de permeabilidade de pele humana vs. de roedores ( $r^2 = 0,44$ ).

#### 4.2.3 Análise da sobreposição dos dados de sensibilização e permeabilidade

A sensibilização da pele é um processo complexo que é influenciado por vários parâmetros biológicos. A permeabilidade da pele é frequentemente relatada como essencial para o processo de sensibilização (JAWORSKA et al., 2011; JOHANSEN; FROSCHE; MENNÉ, 2011). Assim, decidiu-se explorar a relação entre sensibilização e permeabilidade da pele. Identificou-se um subconjunto de 20 compostos para os quais tanto dados de sensibilização quanto de permeabilidade da pele humana estavam disponíveis. Dados de permeabilidade da

pele de roedores não foram usados devido à baixa correlação com a permeabilidade da pele humana. Esses dados experimentais foram compilados de vários estudos, que incluem o relatório do ICCVAM (2009) e os artigos de Jaworska e colaboradores (2011) e Chauhan e Shakya (2010). Como pode ser visto na Tabela 12, não existe correlação direta entre ambas as propriedades, ou seja, forte sensibilizadores poderiam ter uma permeabilidade ruim e vice-versa. Apesar de ter um entendimento comum que um composto deveria penetrar na pele para causar sensibilização, visto que a penetração é frequentemente referida como o primeiro passo para via de efeito adverso de sensibilização (KARLBERG et al., 2008; OECD, 2012), um penetrante ruim pode ainda ser um agente sensibilizador (ex., *p*-Fenilenodiamina ). Essa observação reforça a complexidade do processo de sensibilização da pele e destaca a necessidade de se considerar componentes adicionais na atual via de efeito adverso. Essa consideração foi confirmada por um recente estudo (JAWORSKA et al., 2013), o qual mostrou que, na verdade, o acúmulo de um composto na epiderme está muito mais relacionado à sensibilização do que a permeabilidade pela pele.

**Tabela 12.** Dados experimentais de sensibilização e permeabilidade da pele.

<b>Composto químico</b>	<b>Resultado do LLNA</b>	<b>logKp</b>
Éter de 1,6-Hexanodiol diglicidila [1-3]	Sensibilizador (não especificado)	-3,87
Resorcinol [1-3]	Sensibilizador (moderado)	-3,62
<i>p</i> -Fenilenodiamina [1-3]	Sensibilizador (forte)	-3,62
2-Nitro-4-fenilenodiamina [1-3]	Sensibilizador (forte)	-3,30
Álcool isopropílico [1-3]	Não sensibilizador	-3,05
Piridina [1-3]	Sensibilizador (fraco)	-2,74
Clotrimazol [1-3]	Sensibilizador (não especificado)	-2,70
Acrilato de metila [1-3]	Sensibilizador (fraco)	-2,68
Formaldeído [1-3]	Sensibilizador (forte)	-2,65
Anilina [1-3]	Sensibilizador (fraco)	-2,65
n-Butanol [1-3]	Não sensibilizador	-2,60
Ácido metil acrílico [1-3]	Sensibilizador (não especificado)	-2,58
Acrilato de etila [1-3]	Sensibilizador (fraco)	-2,39
Ácido salicílico [1-3]	Não sensibilizador	-2,20
Cumarina [1-3]	Não sensibilizador	-2,04
Metil 4-hidroxibenzoato [1-3]	Não sensibilizador	-2,04

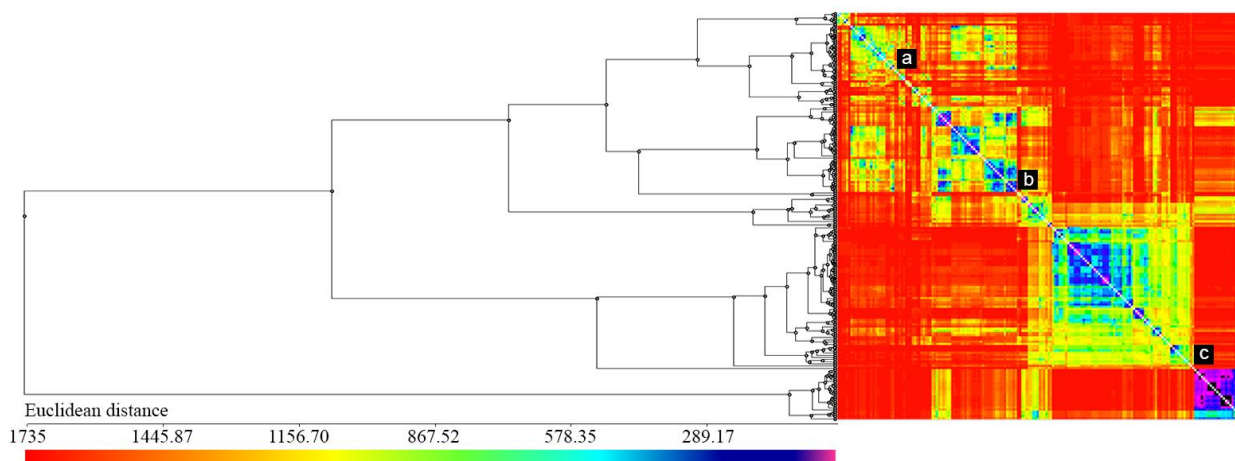
**Tabela 12.** (Continuação).

Composto químico	Resultado do LLNA	logKp
Acrilato de butila [1-3]	Sensibilizador (fraco)	-2,00
Ácido octanoico [1-3]	Não sensibilizador	-1,60
1-Naftol [1-3]	Sensibilizador (moderado)	-1,55
n-Octanol [1, 2]	Sensibilizador (não especificado)	-1,28

Nota: [1] (CHAUHAN; SHAKYA, 2010); [2] (ICCVAM, 2009); [3] (JAWORSKA et al., 2011); Kp: coeficiente de permeabilidade.

#### 4.2.4 Análise de agrupamentos hierárquicos do conjunto de dados D

Devido à baixa concordância entre os dados de permeabilidade da pele humana e de roedores, além do maior interesse em se prever a permeabilidade da pele humana, essa análise só foi aplicada ao conjunto de dados D. De forma geral, o agrupamento de todo o conjunto revelou boa consistência dos dados experimentais. Como pode se observar na Figura 10, o dendograma para o conjunto de dados D não é tão extenso quanto para o conjunto de dados A (Figura 8), ou seja, os compostos do conjunto de dados D são estruturalmente mais similares do que aqueles do conjunto de dados A. Os grupos mais interessantes são comentados a seguir.

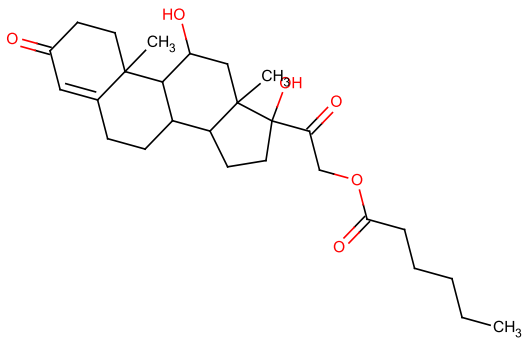
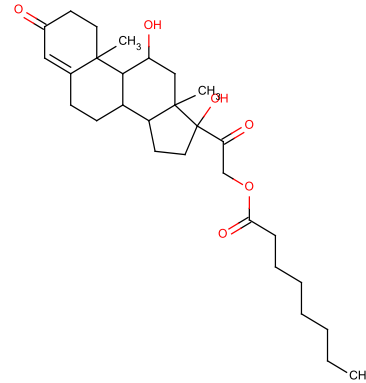
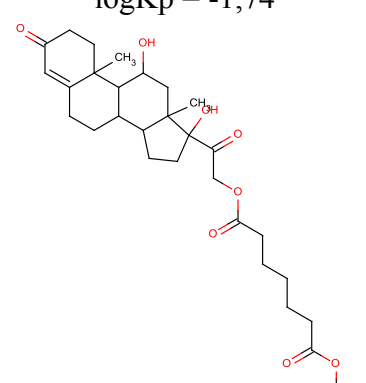
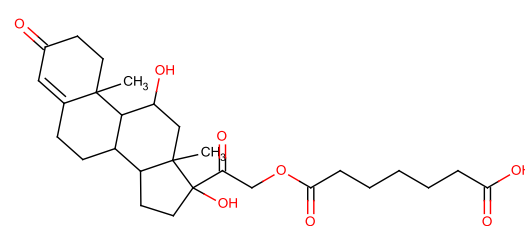
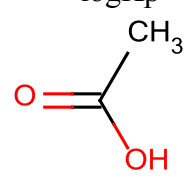
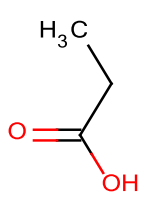
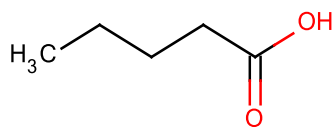
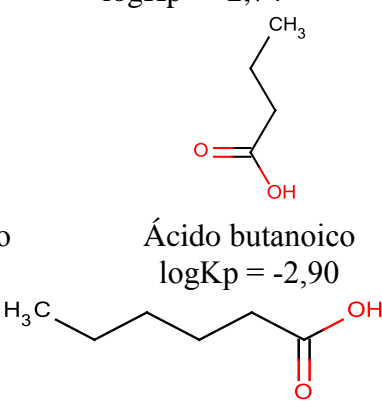


**Figura 10.** Análise agrupamentos hierárquicos para o conjunto de dados D: dendograma e mapa de calor da matriz de distâncias organizadas com base na similaridade estrutural (violeta/azul = similar; amarelo, vermelho = dissimilar).

A permeabilidade de ácidos carboxílicos (grupo **a**, Figura 10) aumenta de forma proporcional ao tamanho da cadeia hidrocarbônica, ou seja,  $\log K_p$  (ácido acético) <  $\log K_p$  (ácido propiônico) <  $\log K_p$  (ácido butanoico) <  $\log K_p$  (ácido pentanoico) <  $\log K_p$  (ácido

hexanoico) (Tabela 13). Álcoois de cadeia curta possuem permeabilidade similar ( $\log K_p \approx -3,00$ ), exceto o etilenoglicol ( $\log K_p = -4,07$ ). Para éteres glicólicos (grupo **b**, Figura 10), a diminuição da permeabilidade é provavelmente devido ao tamanho da cadeia carbônica (2-(2-metoxietoxi)etanol > 2-(2-etoxietoxi)etanol > 2-(2-butoxietoxi)etanol).

**Tabela 13.** Exemplos de compostos “suspeitos” no conjunto de dados D identificados por HCA.

Perfil dos compostos		
Derivados da hidrocortisona	 <p>Hexanoato de hidrocortisona <math>\log K_p = -1,74</math></p>	 <p>Octanoato de hidrocortisona <math>\log K_p = -1,21</math></p>
	 <p>Metilpimelato de hidrocortisona <math>\log K_p = -2,27</math></p>	 <p>Hemipimelato de hidrocortisona <math>\log K_p = -2,74</math></p>
	 <p>Ácido acético <math>\log K_p = -3,21</math></p>	 <p>Ácido propiônico <math>\log K_p = -2,74</math></p>
	 <p>Ácido pentanoico <math>\log K_p = -2,70</math></p>	 <p>Ácido hexanoico <math>\log K_p = -1,85</math></p>

Nota:  $K_p$ : coeficiente de permeabilidade.

Baseado nessa análise encontraram-se três compostos “suspeitos”, um barbitúrico (amilobarbital) e dois esteroides derivados da hidrocortisona (metilsuccinato de hidrocortisona e succinamato de hidrocortisona), mas não foi possível encontrar dados adicionais sobre esses compostos que pudessem confirmar ou refutar os dados dessas observações. Dentro dos esteroides (grupo **c**, Figura 10), a progesterona é o melhor penetrante ( $\log K_p = -1,89$ ), mas a adição de um grupo hidroxila na posição 17 do esqueleto esteroide diminui consideravelmente a permeabilidade (hidroxiprogesterona,  $\log K_p = -3,22$ ). O grupo hidroxila na posição 17 parece crítico para a permeabilidade de esteroides, visto que a testosterona ( $\log K_p = -3,40$ ) contém a hidroxila, mas carece da etoxila. A posição de outros substituintes e duplas ligações parecem ser menos importante. Os efeitos de substituintes na posição 21 na permeabilidade dos compostos dependem de seus tamanhos: inicialmente a permeabilidade diminui quando se substitui uma metila (metilsuccinato;  $\log K_p = -3,68$ ) por um fragmento maior (succinamato;  $\log K_p = -4,59$ ), mas aumenta drasticamente para quatro derivados da hidrocortisona que tem uma cadeia maior nessa posição (hexanoato de hidrocortisona, octanoato de hidrocortisona, metilpimelato de hidrocortisona e hemipimelato de hidrocortisona) (Tabela 13).

#### 4.2.5 Modelos de QSAR gerados para permeabilidade da pele

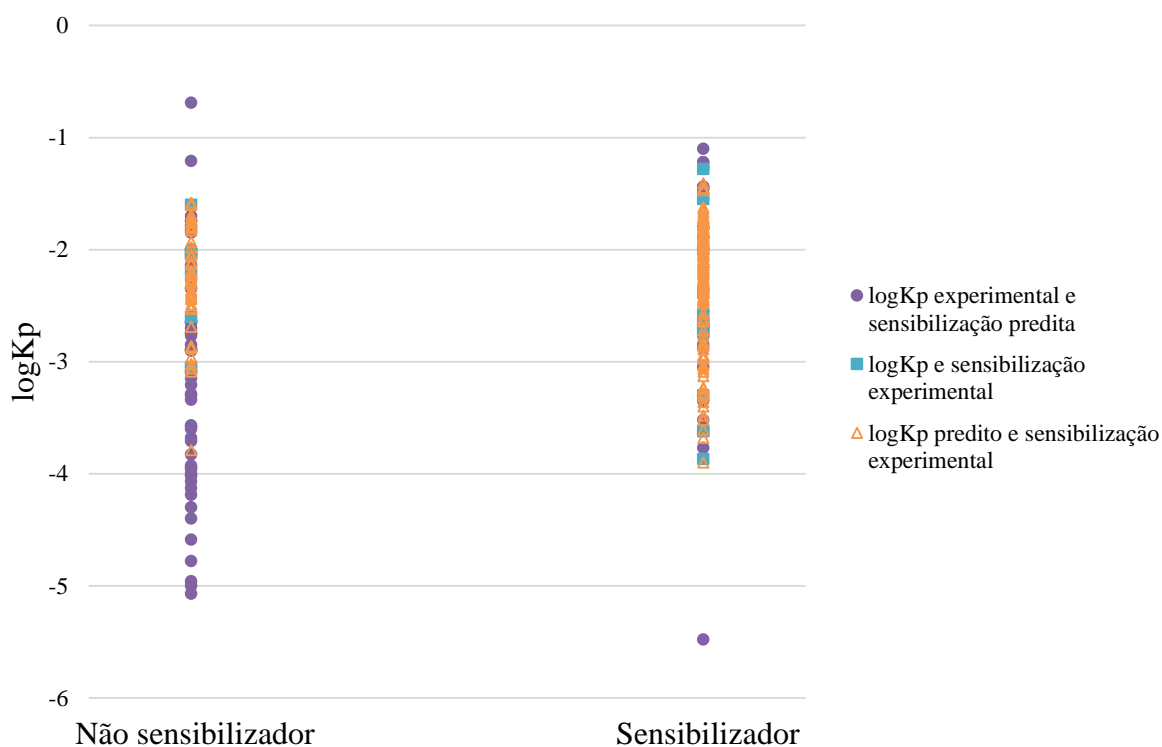
Empregou-se o conjunto de dados compilado por Chauhan e Shakya (2010) compreendendo 211 compostos (conjunto de dados D). Várias estruturas duplicadas foram identificadas durante o preparo do conjunto de dados. A lista completa contendo 17 duplicatas e duas triplicadas está disponível na Tabela 4. Como podem ser observados, os valores de permeabilidade são os mesmos, ou quase os mesmos, para as duplicatas. Esse achado sugere que os modelos desenvolvidos por Chauhan e Shakya (2010) podem ser excessivamente otimistas, visto que a presença de duplicatas com anotações idênticas geralmente levam à sobrestimação da preditividade dos modelos. Modelos baseados em dados de pele de roedores usando-se os 96 compostos do conjunto de dados E foram gerados separadamente. Posteriormente, usaram-se esses modelos para uma predição cruzada entre as espécies. Os resultados estatísticos dos modelos gerados para permeabilidade são apresentados na Tabela 15, na qual se pode observar que modelos preditivos para ambos os conjuntos de dados foram gerados e que a qualidade dos modelos 12 e 14 baseados em dados da pele humana são superiores aos baseados na pele de roedores (19 e 21).

**Tabela 14.** Valores estatísticos dos modelos de QSAR gerados para permeabilidade da pele humana e de roedores avaliados pelo método de validação externa *5-fold*.

	<b>Modelos de QSAR</b>	<b>q<sup>2</sup><sub>ext</sub></b>	<b>RMSEP</b>	<b>MAE</b>	<b>Cobertura</b>
PELE HUMANA	Modelo 8 – SiRMS (DA)	0,79	0,53	0,41	0,73
	Modelo 9 – SiRMS	0,49	0,74	0,54	1,00
	Modelo 10 – Dragon (DA)	0,85	0,47	0,34	0,66
	Modelo 11 – Dragon	0,55	0,69	0,49	1,00
	Modelo 12 – Consenso (DA)	0,79	0,50	0,39	0,77
	Modelo 13 – Consenso	0,55	0,70	0,50	1,00
	Modelo 14 - Consenso Rigor	0,87	0,47	0,35	0,62
PELE DE ROEDORES	Modelo 15 – SiRMS (DA)	0,57	0,67	0,50	0,71
	Modelo 16 – SiRMS	0,35	0,82	0,61	1,00
	Modelo 17 – Dragon (DA)	0,33	0,75	0,56	0,69
	Modelo 18 – Dragon	0,32	0,84	0,63	1,00
	Modelo 19 – Consens (DA)	0,41	0,77	0,56	0,86
	Modelo 20 – Consenso	0,38	0,80	0,61	1,00
	Modelo 21 – Consenso Rigor	0,61	0,58	0,43	0,53

#### 4.2.6 Relação entre o potencial sensibilizador e a permeabilidade de compostos químicos usando dados atribuídos

Dado o pequeno número de compostos com atividade definida para sensibilização e permeabilidade da pele, decidiu-se aplicar os modelos gerados para predição cruzada dessas propriedades. O uso de dados atribuídos por predição dos modelos poderia de fato permitir examinar uma possível relação entre essas duas propriedades levando em consideração um número muito maior de compostos. O modelo 5 foi usado para prever a sensibilização da pele para o conjunto de dados de permeabilidade humana (conjunto de dados D), enquanto o modelo 12 foi utilizado para prever o valores de permeabilidade para os compostos do conjunto de dados de sensibilização da pele (conjunto de dados A). Em seguida, combinou-se todas as predições para os compostos com dados experimentais não disponíveis, obtendo-se 283 compostos para análise (Figura 11 e Tabela A2).



**Figura 11.** Dados experimentais e preditos de permeabilidade vs. sensibilização, contendo 174 sensibilizadores e 109 não sensibilizadores.

Como pode se observar na Figura 11, a maioria dos penetrantes fracos ( $\log K_p < -5$ ) eram não sensibilizadores. A única exceção foi o éter de 1-dodecil glicidila, o qual pode ter sido mal predito pelo modelo 5. Em um estudo recente, os autores levantaram a hipótese de que o acúmulo do composto na pele contribui mais para a sensibilização do que sua permeabilidade pela pele (JAWORSKA et al., 2013). Mesmo que essa hipótese seja verdadeira, o resultado mostrado na Figura 11 não suporta a hipótese de que bons penetrantes podem não se acumular na pele e, assim, não induzem resposta imune significativa. Essa evidente ausência de correlação entre permeabilidade e sensibilização foi, de fato, surpreendente, visto que a permeabilidade da pele é comumente vista como etapa primordial para via de efeito adverso da sensibilização (BASKETTER et al., 2007; JAWORSKA et al., 2013).

Para investigar se a concordância entre sensibilização e permeabilidade da pele poderia ser confirmada para algumas classes de compostos, foi realizada uma análise de grupos para o conjunto de dados de 283 estruturas com dados experimentais e atribuídos pelos modelos. A análise detalhada das relações entre as duas propriedades em grupos de compostos estruturalmente similares mostrou que a conexão direta entre sensibilização e permeabilidade poderia ser feita apenas para um grupo entre 34. Esse grupo específico é formado por 11

compostos representados por quatro barbitais (barbital, amilobarbital, butobarbital e fenobarbital) e sete outras estruturas (escopolamina, ibuprofeno, atropina, griseofulvina, aldeído cyclamen, *p-tert*-butil- $\alpha$ -etil-hidrocinalmal e 3,3,5-trimetil-benzenosulfonato). A comparação entre permeabilidade e sensibilização para compostos nesse grupo está apresentada na Figura 12 e Tabela A3 (Anexos).



**Figura 12.** Dados experimentais e preditos de permeabilidade vs. sensibilização para o grupo 1, contendo 4 sensibilizadores e 7 não sensibilizadores.

Os dados mostram que todos os barbitais estão associados com  $\log K_p < -2,5$  e são não sensibilizadores, enquanto sete outros compostos com  $\log K_p > -2,5$  são sensibilizadores. Contudo, nenhuma tendência pode ser observada para nenhum dos outros 34 grupos. Assim, esses resultados mostram que não existem evidências que suportem a hipótese de que o potencial sensibilizador de um composto depende de sua permeabilidade na pele para ocorrer.

## 5 CONCLUSÃO

Foram compilados, integrados e preparados os maiores conjuntos de dados publicamente disponíveis para sensibilização e permeabilidade da pele para compostos químicos diversos. A análise dos dados experimentais para compostos contendo dados tanto para sensibilização quanto para permeabilidade indicaram que, com algumas exceções, não existe concordância global entre essas duas propriedades, ou seja, penetrantes fracos podem ser fortes sensibilizadores e *vice versa*. A análise dos agrupamentos hierárquicos contribuiu para revelar a alta consistência dos dados experimentais, assim como identificar 24 compostos suspeitos de possuírem dados incorretos.

Modelos de QSAR robustos e preditivos para sensibilização (modelo de classificação) e permeabilidade (modelo contínuo) foram obtidos. Os modelos de sensibilização mostraram especificidade e ACB significativamente mais altos que o módulo de sensibilização do QSAR Toolbox, mas menor cobertura e sensibilidade. Contudo, a análise dos valores de preditividade positiva e negativa mostraram que os modelos gerados neste trabalho garantiram que praticamente todos os compostos sensibilizadores (VPP=95%) classificados como sensibilizadores são, de fato, sensibilizadores. Enquanto isso, existe uma chance de 71% de que um composto não sensibilizador seja predito como não sensibilizador. Ao mesmo tempo, a probabilidade de um composto sensibilizador ou não sensibilizador combinar com a predição feita pelo QSAR Toolbox é praticamente aleatória (VPP=0,51, VPP=0,52, predição aleatória = 0,5).

O uso de valores de permeabilidade e sensibilidade imputados por modelos de QSAR permitiu examinar as relações entre essas duas propriedades usando um conjunto significativamente enriquecido de 283 compostos. Os resultados indicaram que não existe concordância global entre essas duas propriedades. Ainda, ao se analisar algumas estruturas similares identificadas pela análise de grupos, nenhuma correlação foi encontrada para 33 de 34 grupos.

Os modelos de sensibilização da pele estão disponíveis para serem usados pela comunidade científica, indústrias farmacêuticas e de produtos químicos e agências de regulamentação para identificar compostos potencialmente sensibilizadores. A triagem virtual do banco de dados da Scorecard revelou 27 potenciais sensibilizadores que podem ser utilizados para validação experimental. Todos os modelos desenvolvidos estão publicamente disponíveis para uso no *Chembench Web Portal* (<http://chembench.mml.unc.edu>). Esse trabalho resultou em dois manuscritos que foram submetidos para o *Environmental Health Perspectives* (Anexos).

## REFERÊNCIAS BIBLIOGRÁFICAS

- ABRAHAM, M. H.; CHADHA, H. S.; MARTINS, F.; MITCHELL, R. C.; BRADBURY, M. W.; GRATTON, J. A. Hydrogen bonding part 46: a review of the correlation and prediction of transport properties by an lfer method: physicochemical properties, brain penetration and skin permeability. **Pesticide Science**, v. 55, n. 1, p. 78–88, 26 jan. 1999.
- ADEUSI, S. **Pharmaceutical R&D: an organizational design approach to enhancing productivity**. [s.l.] Massachusetts Institute of Technology, 2011.
- ADLER, S.; BASKETTER, D.; CRETON, S.; PELKONEN, O.; VAN BENTHEM, J.; ZUANG, V.; ANDERSEN, K. E.; ANGERS-LOUSTAU, A.; APTULA, A.; BAL-PRICE, A.; BENFENATI, E.; BERNAUER, U.; BESSEMS, J.; BOIS, F. Y.; BOOBIS, A.; BRANDON, E.; BREMER, S.; BROSCARD, T.; CASATI, S.; COECKE, S.; CORVI, R.; CRONIN, M.; DASTON, G.; DEKANT, W.; FELTER, S.; GRIGNARD, E.; GUNDERT-REMY, U.; HEINONEN, T.; KIMBER, I.; KLEINJANS, J.; KOMULAINEN, H.; KREILING, R.; KREYSA, J.; LEITE, S. B.; LOIZOU, G.; MAXWELL, G.; MAZZATORTA, P.; MUNN, S.; PFUHLER, S.; PHRAKONKHAM, P.; PIERSMA, A.; POTH, A.; PRIETO, P.; REPETTO, G.; ROGIERS, V.; SCHOETERS, G.; SCHWARZ, M.; SERAFIMOVA, R.; TÄHTI, H.; TESTAI, E.; VAN DELFT, J.; VAN LOVEREN, H.; VINKEN, M.; WORTH, A.; ZALDIVAR, J.-M. Alternative (non-animal) methods for cosmetics testing: current status and future prospects-2010. **Archives of toxicology**, v. 85, n. 5, p. 367–485, maio 2011.
- AEBY, P.; ASHIKAGA, T.; BESSOU-TOUYA, S.; SCHEPKY, A.; GERBERICK, F.; KERN, P.; MARREC-FAIRLEY, M.; MAXWELL, G.; OVIGNE, J.-M.; SAKAGUCHI, H.; REISINGER, K.; TAILHARDAT, M.; MARTINOZZI-TEISSIER, S.; WINKLER, P. Identifying and characterizing chemical skin sensitizers without animal testing: Colipa's research and method development program. **Toxicology in vitro : an international journal published in association with BIBRA**, v. 24, n. 6, p. 1465–73, set. 2010.
- ANDERSON, S. E.; SIEGEL, P. D.; MEADE, B. J. The LLNA: A brief review of recent advances and limitations. **Journal of allergy**, v. 2011, p. 424203, jan. 2011.
- ANDRICOPULO, A. D.; SALUM, L. B.; ABRAHAM, D. J. Structure-based drug design strategies in medicinal chemistry. **Current topics in medicinal chemistry**, v. 9, n. 9, p. 771–90, jan. 2009.
- ANKLEY, G. T.; BENNETT, R. S.; ERICKSON, R. J.; HOFF, D. J.; HORNUNG, M. W.; JOHNSON, R. D.; MOUNT, D. R.; NICHOLS, J. W.; RUSSOM, C. L.; SCHMIEDER, P. K.; SERRRANO, J. A.; TIETGE, J. E.; VILLENEUVE, D. L. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. **Environmental toxicology and chemistry / SETAC**, v. 29, n. 3, p. 730–41, mar. 2010.

- ARTEMENKO, A. G.; MURATOV, E. N.; KUZ'MIN, V. E.; MURATOV, N. N.; VARLAMOVA, E. V.; KUZ'MINA, A. V.; GORB, L. G.; GOLIUS, A.; HILL, F. C.; LESZCZYNSKI, J.; TROPSHA, A. QSAR analysis of the toxicity of nitroaromatics in *Tetrahymena pyriformis*: structural factors and possible modes of action. **SAR and QSAR in environmental research**, v. 22, n. 5-6, p. 575–601, 2011.
- ASHBY, J.; BASKETTER, D. A.; PATON, D.; KIMBER, I. Structure activity relationships in skin sensitization using the murine local lymph node assay. **Toxicology**, v. 103, n. 3, p. 177–94, 10 dez. 1995.
- BAKER, M. Fragment-based lead discovery grows up. **Nature reviews. Drug discovery**, v. 12, n. January, p. 5–7, 2013.
- BARRATT, M. D. Quantitative structure-activity relationships for skin permeability. **Toxicology in vitro : an international journal published in association with BIBRA**, v. 9, n. 1, p. 27–37, fev. 1995.
- BASKETTER, D. A.; EVANS, P.; FIELDER, R. J.; GERBERICK, G. F.; DEARMAN, R. J.; KIMBER, I. Local lymph node assay - validation, conduct and use in practice. **Food and chemical toxicology : an international journal published for the British Industrial Biological Research Association**, v. 40, n. 5, p. 593–8, maio 2002.
- BASKETTER, D. A.; ROBERTS, D. W.; CRONIN, M.; SCHOLES, E. W. The value of the local lymph node assay in quantitative structure-activity investigations. **Contact dermatitis**, v. 27, n. 3, p. 137–42, set. 1992.
- BASKETTER, D.; PEASE, C.; KASTING, G.; KIMBER, I.; CASATI, S.; CRONIN, M.; DIEMBECK, W.; GERBERICK, F.; HADGRAFT, J.; HARTUNG, T.; MARTY, J.-P.; NIKOLAIDIS, E.; PATLEWICZ, G.; ROBERTS, D.; ROGGEN, E.; ROVIDA, C.; VAN DE SANDT, J. Skin sensitisation and epidermal disposition: the relevance of epidermal disposition for sensitisation hazard identification and risk assessment. The report and recommendations of ECVAM workshop 59. **Alternatives to laboratory animals : ATLA**, v. 35, n. 1, p. 137–54, mar. 2007.
- BAXTER, K.; HORN, E.; GAL-EDD, N.; ZONNO, K.; O'LEARY, J.; TERRY, P. F.; TERRY, S. F. An end to the myth: there is no drug development pipeline. **Science translational medicine**, v. 5, n. 171, p. 171cm1, 6 fev. 2013.
- BELFIELD, G. P.; DELANEY, S. J. The impact of molecular biology on drug discovery. **Biochemical Society transactions**, v. 34, n. Pt 2, p. 313–6, abr. 2006.
- BERMAN, H. M.; KLEYWEGT, G. J.; NAKAMURA, H.; MARKLEY, J. L. How community has shaped the Protein Data Bank. **Structure (London, England : 1993)**, v. 21, n. 9, p. 1485–91, 3 set. 2013.

- BLEICHER, K. H.; BÖHM, H.-J.; MÜLLER, K.; ALANINE, A. I. Hit and lead generation: beyond high-throughput screening. **Nature reviews. Drug discovery**, v. 2, n. 5, p. 369–78, maio 2003.
- BOS, J. D.; MEINARDI, M. M. The 500 Dalton rule for the skin penetration of chemical compounds and drugs. **Experimental dermatology**, v. 9, n. 3, p. 165–9, jun. 2000.
- BREIMAN, L. E. O. Random Forests. **Machine Learning**, v. 45, p. 5–32, 2001.
- BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and Regression Trees**. Belmont: Wadsworth Publishing, 1984. p. 358
- BUEHLER, E. V. Delayed contact hypersensitivity in the guinea pig. **Archives of dermatology**, v. 91, p. 171–7, fev. 1965.
- CAPDEVILLE, R.; BUCHDUNGER, E.; ZIMMERMANN, J.; MATTER, A. Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. **Nature reviews. Drug discovery**, v. 1, n. 7, p. 493–502, jul. 2002.
- CAPORUSCIO, F.; TAFI, A. Pharmacophore modelling: a forty year old approach and its modern synergies. **Current medicinal chemistry**, v. 18, n. 17, p. 2543–53, jan. 2011.
- CHAST, F. A history of drug discovery. In: WERMUTH, C. (Ed.). **The practice of Medicinal Chemistry**. 3. ed. [s.l.] Elsevier, 2008. p. 3–62.
- CHAUDHRY, Q.; PICLIN, N.; COTTERILL, J.; PINTORE, M.; PRICE, N. R.; CHRÉTIEN, J. R.; RONCAGLIONI, A. Global QSAR models of skin sensitizers for regulatory purposes. **Chemistry central journal**, v. 4 Suppl 1, n. Suppl 1, p. S5, jan. 2010.
- CHAUHAN, P.; SHAKYA, M. Role of physicochemical properties in the estimation of skin permeability: in vitro data assessment by Partial Least-Squares Regression. **SAR and QSAR in environmental research**, v. 21, n. 5-6, p. 481–94, jul. 2010.
- CHEN, L. L.; LIAN, G. G.; HAN, L. L. Prediction of human skin permeability using artificial neural network (ANN) modeling. **Acta pharmacologica Sinica**, v. 28, n. 4, p. 591–600, abr. 2007.
- CHEN, L.; LIAN, G.; HAN, L. Modeling transdermal permeation. Part I. Predicting skin permeability of both hydrophobic and hydrophilic solutes. **AIChE journal**, v. 56, n. 5, p. 1136–1146, 2010.
- CHENG, T.; LI, Q.; ZHOU, Z.; WANG, Y.; BRYANT, S. H. Structure-based virtual screening for drug discovery: a problem-centric review. **The AAPS journal**, v. 14, n. 1, p. 133–41, mar. 2012.

- CHENG, Z.; ZHANG, Y.; ZHOU, C.; ZHANG, W.; GAO, S. Classification of skin sensitizers on the basis of their effective concentration 3 values by using adaptive boosting method. **International journal of digital content technology and its applications**, v. 4, n. 2, p. 109–121, 30 abr. 2010.
- CHERKASOV, A.; MURATOV, E. N.; FOURCHES, D.; VARNEK, A.; BASKIN, I. I.; CRONIN, M.; DEARDEN, J.; GRAMATICA, P.; MARTIN, Y. C.; TODESCHINI, R.; CONSONNI, V.; KUZ'MIN, V. E.; CRAMER, R.; BENIGNI, R.; YANG, C.; RATHMAN, J.; TERFLOTH, L.; GASTEIGER, J.; RICHARD, A.; TROPSHA, A. QSAR Modeling: Where Have You Been? Where Are You Going To? **Journal of medicinal chemistry**, v. Epub ahead, 6 jan. 2014.
- CHUPRINA, A.; LUKIN, O.; DEMOISEAUX, R.; BUZKO, A.; SHIVANYUK, A. Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers. **Journal of chemical information and modeling**, v. 50, n. 4, p. 470–479, 26 abr. 2010.
- CLARK, D. E. What has virtual screening ever done for drug discovery? **Expert opinion on drug discovery**, v. 3, n. 8, p. 841–51, ago. 2008.
- COI, A.; BIANUCCI, A. M. Combining structure- and ligand-based approaches for studies of interactions between different conformations of the hERG K(+) channel pore and known ligands. **Journal of molecular graphics & modelling**, v. 46, p. 93–104, nov. 2013.
- CONSONNI, V.; TODESCHINI, R. Molecular descriptors. In: PUZYN, T.; LESZCZYNSKI, J.; CRONIN, M. T. (Eds.). **Recent Advances in QSAR Studies**. Dordrecht: Springer, 2010. p. 29–.
- CRAMER, R. D. The inevitable QSAR renaissance. **Journal of computer-aided molecular design**, v. 26, n. 1, p. 35–8, jan. 2012.
- 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. 5959–67, 1 ago. 1988.
- CRONIN, M. T. Quantitative structure-activity relationships (QSARs) -- applications and methodology. In: PUZYN, T.; LESZCZYNSKI, J.; CRONIN, M. T. (Eds.). **Recent Advances in QSAR Studies**. Dordrecht: Springer, 2010. p. 3–11.
- CRONIN, M. T.; BASKETTER, D. A. Multivariate QSAR analysis of a skin sensitization database. **SAR and QSAR in environmental research**, v. 2, n. 3, p. 159–79, jan. 1994.
- CRONIN, M. T.; DEARDEN, J. C.; MOSS, G. P.; MURRAY-DICKSON, G. Investigation of the mechanism of flux across human skin in vitro by quantitative structure-permeability relationships. **European journal of pharmaceutical sciences : official**

**journal of the European Federation for Pharmaceutical Sciences**, v. 7, n. 4, p. 325–30, mar. 1999.

- DEARDEN, J. C.; CRONIN, M. T. D.; KAISER, K. L. E. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). **SAR and QSAR in environmental research**, v. 20, n. 3-4, p. 241–66, jan. 2009.
- DEARMAN, R. J.; BETTS, C. J.; FARR, C.; MCLAUGHLIN, J.; BERDASCO, N.; WIENCH, K.; KIMBER, I. Comparative analysis of skin sensitization potency of acrylates (methyl acrylate, ethyl acrylate, butyl acrylate, and ethylhexyl acrylate) using the local lymph node assay. **Contact dermatitis**, v. 57, n. 4, p. 242–7, out. 2007.
- DEARMAN, R. J.; WRIGHT, Z. M.; BASKETTER, D. A.; RYAN, C. A.; GERBERICK, G. F.; KIMBER, I. The suitability of hexyl cinnamic aldehyde as a calibrant for the murine local lymph node assay. **Contact dermatitis**, v. 44, n. 6, p. 357–61, jun. 2001.
- DEBNATH, A. K. Quantitative structure-activity relationship (QSAR) paradigm--Hansch era to new millennium. **Mini reviews in medicinal chemistry**, v. 1, n. 2, p. 187–95, jul. 2001.
- DEGIM, I. T.; PUGH, W. J.; HADGRAFT, J. Skin permeability data: anomalous results. **International journal of pharmaceutics**, v. 170, n. 1, p. 129–133, ago. 1998.
- DEVILLERS, J. A neural network SAR model for allergic contact dermatitis. **Toxicology methods**, v. 10, p. 181–193, 2000.
- DICKEL, H.; KUSS, O.; SCHMIDT, A.; KRETZ, J.; DIEPGEN, T. L. Importance of irritant contact dermatitis in occupational skin disease. **American journal of clinical dermatology**, v. 3, n. 4, p. 283–9, jan. 2002.
- DOMINGOS, P. A few useful things to know about machine learning. **Communications of the ACM**, v. 55, n. 10, p. 78, 1 out. 2012.
- DOWNS, G. M.; BARNARD, J. M. Clustering Methods and Their Uses in Computational Chemistry. In: LIPKOWITZ, K. B.; BOYD, D. B. (Eds.). **Reviews in Computational Chemistry**. Hoboken: John Wiley & Sons, Inc., 2003. v. 18p. 1–40.
- DREWS, J. Case histories, magic bullets and the state of drug discovery. **Nature Reviews Drug Discovery**, v. 5, n. August, p. 635–640, 2006.
- EGEGHY, P. P.; JUDSON, R.; GANGWAL, S.; MOSHER, S.; SMITH, D.; VAIL, J.; HUBAL, E. A. C. The exposure data landscape for manufactured chemicals. **The Science of the total environment**, v. 414, p. 159–66, 1 jan. 2012.

- ESTRADA, E.; PATLEWICZ, G.; CHAMBERLAIN, M.; BASKETTER, D.; LARBHEY, S. Computer-aided knowledge generation for understanding skin sensitization mechanisms: the TOPS-MODE approach. **Chemical research in toxicology**, v. 16, n. 10, p. 1226–35, out. 2003.
- EUROPEAN UNION. **Directive 2003/15/EC of the European Parliament and of the Council of 27 February 2003 amending Council Directive 76/768/EEC on the approximation of the laws of the Member States relating to cosmetic products** Official journal of the european union, 2003.
- EUROPEAN UNION. Regulation (EC) No 1907/2006. **Official journal of the European Union**, n. L 136, p. 3–280, 2007.
- EUROPEAN UNION. Directive 2010/63/EU. **Official Journal of the European Union**, n. L 276, p. 33–79, 2010.
- FDA. **New Drugs at FDA: CDER's New Molecular Entities and New Therapeutic Biological Products of 2013**. Disponível em: <<http://www.fda.gov/drugs/developmentapprovalprocess/druginnovation/default.htm>>. Acesso em: 2 jan. 2013.
- FEDOROWICZ, A.; SINGH, H.; SODERHOLM, S.; DEMCHUK, E. Structure-activity models for contact sensitization. **Chemical research in toxicology**, v. 18, n. 6, p. 954–69, jun. 2005.
- FEDOROWICZ, A.; ZHENG, L.; SINGH, H.; DEMCHUK, E. QSAR study of skin sensitization using local lymph node assay data. **International journal of molecular sciences**, v. 5, n. 2, p. 56–66, 30 jan. 2004.
- FLYNN, G. L. Physicochemical determinants of skin absorption. In: GERRITY, T. R.; HENRY, C. J. (Eds.). **Principles of Route-to-Route Extrapolation for Risk assessment**. New York, NY: Elsevier, 1990. p. 93–117.
- FOURCHES, D.; MURATOV, E.; TROPSHA, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. **Journal of chemical information and modeling**, v. 50, n. 7, p. 1189–204, 26 jul. 2010.
- FREE, S. M.; WILSON, J. W. A Mathematical Contribution to Structure-Activity Studies. **Journal of Medicinal Chemistry**, v. 7, n. 4, p. 395–399, jul. 1964.
- FRIEDMAN, L. M.; FURBERG, C. D.; DEMETS, D. L. **Fundamentals of Clinical Trials**. 4. ed. New York, NY: Springer, 2010. p. 400
- FUJITA, T.; IWASA, J.; HANSCH, C. A New Substituent Constant,  $\pi$ , Derived from Partition Coefficients. **Journal of the American Chemical Society**, v. 86, n. 23, p. 5175–5180, dez. 1964.

- GANELLIN, C. R. Robin Ganellin gives his views on medicinal chemistry and drug discovery. Interview by Stephen L. Carney. **Drug discovery today**, v. 9, n. 4, p. 158–60, 15 fev. 2004.
- GAVAGHAN, C. L.; ARNBY, C. H.; BLOMBERG, N.; STRANDLUND, G.; BOYER, S. Development, interpretation and temporal evaluation of a global QSAR of hERG electrophysiology screening data. **Journal of computer-aided molecular design**, v. 21, n. 4, p. 189–206, abr. 2007.
- GERBERICK, G. F.; RYAN, C. A.; KERN, P. S.; SCHLATTER, H. Compilation of historical local lymph node data for evaluation of skin sensitization alternative methods. **Dermatitis**, v. 16, n. 4, p. 157–202, 2005.
- GERBERICK, G. F.; TROUTMAN, J. A.; FOERTSCH, L. M.; VASSALLO, J. D.; QUIJANO, M.; DOBSON, R. L. M.; GOEBEL, C.; LEPOITTEVIN, J.-P. Investigation of peptide reactivity of pro-hapten skin sensitizers using a peroxidase-peroxide oxidation system. **Toxicological sciences : an official journal of the Society of Toxicology**, v. 112, n. 1, p. 164–74, nov. 2009.
- GERBERICK, G. F.; VASSALLO, J. D.; BAILEY, R. E.; CHANEY, J. G.; MORRALL, S. W.; LEPOITTEVIN, J.-P. Development of a peptide reactivity assay for screening contact allergens. **Toxicological sciences : an official journal of the Society of Toxicology**, v. 81, n. 2, p. 332–43, out. 2004.
- GLEESON, M. P. Generation of a set of simple, interpretable ADMET rules of thumb. **Journal of medicinal chemistry**, v. 51, n. 4, p. 817–34, 28 fev. 2008.
- GLEESON, M. P.; MODI, S.; BENDER, A.; ROBINSON, R. L. M.; KIRCHMAIR, J.; PROMKATKAEW, M.; HANNONGBUA, S.; GLEN, R. C. The challenges involved in modeling toxicity data in silico: a review. **Current pharmaceutical design**, v. 18, n. 9, p. 1266–91, jan. 2012.
- GOLBRAIKH, A.; TROPSHA, A. Beware of q<sup>2</sup>! **Journal of molecular graphics & modelling**, v. 20, n. 4, p. 269–76, jan. 2002.
- GOLLA, S.; MADIHALLY, S.; ROBINSON, R. L.; GASEM, K. A. M. Quantitative structure-property relationship modeling of skin sensitization: a quantitative prediction. **Toxicology in vitro : an international journal published in association with BIBRA**, v. 23, n. 3, p. 454–65, abr. 2009.
- GRAMATICA, P. Principles of QSAR models validation: internal and external. **QSAR & Combinatorial Science**, v. 26, n. 5, p. 694–701, maio 2007.
- GRANDJEAN, P.; BERLIN, A.; GILBERT, M.; PENNING, W. Preventing percutaneous absorption of industrial chemicals: the “skin” denotation. **American journal of industrial medicine**, v. 14, n. 1, p. 97–107, jan. 1988.

- GUHA, R.; VAN DRIE, J. H. Structure-activity landscape index: identifying and quantifying activity cliffs. **Journal of chemical information and modeling**, v. 48, n. 3, p. 646–58, mar. 2008.
- GUIDO, R. V. C. **Planejamento de inibidores da enzima gliceraldeído-3-fosfato desidrogenase de Trypanosoma cruzi: biologia estrutural e química medicinal**. [s.l.] Universidade de São Paulo, 2008.
- GUNTURI, S. B.; THEERTHALA, S. S.; PATEL, N. K.; BAHL, J.; NARAYANAN, R. Prediction of skin sensitization potential using D-optimal design and GA-kNN classification methods. **SAR and QSAR in environmental research**, v. 21, n. 3-4, p. 305–35, abr. 2010.
- 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, jan. 1937.
- HANEKE, K. E.; TICE, R. R.; CARSON, B. L.; MARGOLIN, B. H.; STOKES, W. S. ICCVAM evaluation of the murine local lymph node assay. Data analyses completed by the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods. **Regulatory toxicology and pharmacology : RTP**, v. 34, n. 3, p. 274–86, dez. 2001.
- HANSCH, C. Quantitative approach to biochemical structure-activity relationships. **Accounts of Chemical Research**, v. 2, n. 8, p. 232–239, ago. 1969.
- HANSCH, C.; FUJITA, T.  $\rho$ - $\sigma$ - $\pi$  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, abr. 1964.
- HANSCH, C.; MALONEY, P. P.; FUJITA, T.; MUIR, R. M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. **Nature**, v. 194, n. 4824, p. 178–180, 14 abr. 1962.
- HÄRMARK, L.; VAN GROOTHEEST, A. C. Pharmacovigilance: methods, recent developments and future perspectives. **European journal of clinical pharmacology**, v. 64, n. 8, p. 743–52, ago. 2008.
- HENNINO, A.; VOCANSON, M.; CHAVAGNAC, C.; SAINT-MEZARD, P.; DUBOIS, B.; KAISERLIAN, D.; NICOLAS, J. Update on the pathophysiology with special emphasis on CD8 effector T cells and CD4 regulatory T cells. **Anais brasileiros de dermatologia**, v. 80, n. 4, p. 335–347, ago. 2005.
- HOPFINGER, A. J.; WANG, S.; TOKARSKI, J. S.; JIN, B.; ALBUQUERQUE, M.; MADHAV, P. J.; DURAISWAMI, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. **Journal of the American Chemical Society**, v. 119, n. 43, p. 10509–10524, out. 1997.

- HOPFINGER, A. J. A QSAR investigation of dihydrofolate reductase inhibition by Baker triazines based upon molecular shape analysis. **Journal of the American Chemical Society**, v. 102, n. 24, p. 7196–7206, set. 1980.
- HOSTÝNEK, J. J.; MAGEE, P. S. Modelling in vivo human skin absorption. **Quantitative structure-activity relationships**, v. 16, n. 6, p. 473–479, 1997a.
- HOSTÝNEK, J. J.; MAGEE, P. S. Fragrance allergens: Classification and ranking by QSAR. **Toxicology in vitro : an international journal published in association with BIBRA**, v. 11, n. 4, p. 377–84, ago. 1997b.
- HOSTÝNEK, J. J.; MAGEE, P. S.; MAIBACH, H. I. QSAR predictive of contact allergy: scope and limitations. **Current problems in dermatology**, v. 25, p. 18–27, jan. 1996.
- HUGHES, J. P.; REES, S.; KALINDJIAN, S. B.; PHILPOTT, K. L. Principles of early drug discovery. **British journal of pharmacology**, v. 162, n. 6, p. 1239–49, mar. 2011.
- HUTT, P. B. **The Historical Development of Animal Toxicity Testing**. Disponível em: <<http://nrs.harvard.edu/urn-3:HUL.InstRepos:8889439>>.
- ICCVAM. **The reduced murine local lymph node assay: an alternative test method using fewer animals to assess the allergic contact dermatitis potential of chemicals and products**. Disponível em: <[http://ntp.niehs.nih.gov/iccvam/docs/immunotox\\_docs/LLNA-LD/TMER.pdf](http://ntp.niehs.nih.gov/iccvam/docs/immunotox_docs/LLNA-LD/TMER.pdf)>. Acesso em: 20 fev. 2012.
- ICCVAM; NICEATM. **The murine local lymph node assay: a test method for assessing the allergic contact dermatitis potential of chemicals/compounds**. Disponível em: <[http://iccvam.niehs.nih.gov/docs/immunotox\\_docs/llna/llnarep.pdf](http://iccvam.niehs.nih.gov/docs/immunotox_docs/llna/llnarep.pdf)>. Acesso em: 8 jun. 2012.
- IMMING, P. Medicinal Chemistry: definitions and objectives, drug activity phases, drug classification systems. In: WERMUTH, C. (Ed.). **The practice of Medicinal Chemistry**. 3. ed. [s.l.] Elsevier, 2008. p. 63–72.
- JAWORSKA, J.; DANCİK, Y.; KERN, P.; GERBERICK, F.; NATSCH, A. Bayesian integrated testing strategy to assess skin sensitization potency: from theory to practice. **Journal of applied toxicology : JAT**, v. 33, n. 11, p. 1353–64, 14 maio 2013.
- JAWORSKA, J.; HAROL, A.; KERN, P. S.; GERBERICK, G. F. Integrating non-animal test information into an adaptive testing strategy - skin sensitization proof of concept case. **Altex**, v. 28, n. 3, p. 211–25, jan. 2011.
- JOHANSEN, J. D.; FROSCH, P. J.; MENNÉ, T. Allergic contact dermatitis in humans: experimental and quantitative aspects. In: JOHANSEN, J. D.; FROSCH, P. J.;

LEPOITTEVIN, J.-P. (Eds.). **Contact Dermatitis**. Berlin: Springer, 2011. p. 241–251.

JOHANSSON, H.; LINDSTEDT, M.; ALBREKT, A.-S.; BORREBAECK, C. A. K. A genomic biomarker signature can predict skin sensitizers using a cell-based in vitro alternative to animal tests. **BMC genomics**, v. 12, n. 1, p. 399, jan. 2011.

JOHNSON, M. E.; BLANKSCHTEIN, D.; LANGER, R. Permeation of steroids through human skin. **Journal of pharmaceutical sciences**, v. 84, n. 9, p. 1144–6, set. 1995.

JOHNSON, M. E.; BLANKSCHTEIN, D.; LANGER, R. Evaluation of solute permeation through the stratum corneum: lateral bilayer diffusion as the primary transport mechanism. **Journal of pharmaceutical sciences**, v. 86, n. 10, p. 1162–72, out. 1997.

KAR, S.; ROY, K. How far can virtual screening take us in drug discovery? **Expert opinion on drug discovery**, v. 8, n. 3, p. 245–61, mar. 2013.

KARLBERG, A.-T.; BERGSTRÖM, M. A.; BÖRJE, A.; LUTHMAN, K.; NILSSON, J. L. G. Allergic contact dermatitis - formation, structural requirements, and reactivity of skin sensitizers. **Chemical research in toxicology**, v. 21, n. 1, p. 53–69, jan. 2008.

KEEGEL, T.; MOYLE, M.; DHARMAGE, S.; FROWEN, K.; NIXON, R. The epidemiology of occupational contact dermatitis (1990-2007): a systematic review. **International journal of dermatology**, v. 48, n. 6, p. 571–8, jun. 2009.

KERN, P. S.; GERBERICK, G. F.; RYAN, C. A.; KIMBER, I.; APTULA, A.; BASKETTER, D. A. Local lymph node data for the evaluation of skin sensitization alternatives: a second compilation. **Dermatitis : contact, atopic, occupational, drug : official journal of the American Contact Dermatitis Society, North American Contact Dermatitis Group**, v. 21, n. 1, p. 8–32, fev. 2010.

KESSEL, M. The problems with today's pharmaceutical business--an outsider's view. **Nature biotechnology**, v. 29, n. 1, p. 27–33, jan. 2011.

KIELHORN, J.; MELCHING-KOLLMUS, S.; MANGELSDORF, I. **The Dermal Absorption**. Hanover: Fraunhofer Institute Toxicology and Experimental Medicine, 2005. p. 90–91

KIMBER, I.; BASKETTER, D. A.; GERBERICK, G. F.; DEARMAN, R. J. Allergic contact dermatitis. **International immunopharmacology**, v. 2, n. 2-3, p. 201–11, fev. 2002.

KIMBER, I.; BASKETTER, D. A.; GERBERICK, G. F.; RYAN, C. A.; DEARMAN, R. J. Chemical allergy: translating biology into hazard characterization. **Toxicological**

**sciences : an official journal of the Society of Toxicology**, v. 120(S1), p. S238–S268, mar. 2011.

KIMBER, I.; MITCHELL, J. A.; GRIFFIN, A. C. Development of a murine local lymph node assay for the determination of sensitizing potential. **Food and Chemical Toxicology**, v. 24, n. 6-7, p. 585–586, jun. 1986.

KIRCHNER, L. A.; MOODY, R. P.; DOYLE, E.; BOSE, B.; JEFFRY, J.; CHU, I. The prediction of skin permeability by using physicochemical data. **ATLA**, v. 25, p. 359–370, 1997.

KIRKPATRICK, P. An audience with... Chris Lipinski. **Nature Reviews Drug Discovery**, v. 11, p. 900–901, 2012.

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. 4130–46, nov. 1994.

KNUDSEN, T. B.; KLEINSTREUER, N. C. Disruption of embryonic vascular development in predictive toxicology. **Birth defects research. Part C, Embryo today : reviews**, v. 93, n. 4, p. 312–23, dez. 2011.

KUBINYI, H. Quantitative structure-activity relationships. V. A simple simple algorithm for Fujita-Ban and Free-Wilson analyses. **Arzneimittel-Forschung**, v. 27, n. 4, p. 750–8, jan. 1977.

KUBINYI, H. Nonlinear dependence of biological activity on hydrophobic character: the bilinear model. **Il Farmaco; edizione scientifica**, v. 34, n. 3, p. 248–76, mar. 1979.

KUZ'MIN, V. E.; ARTEMENKO, A. G.; MURATOV, E. N. Hierarchical QSAR technology based on the Simplex representation of molecular structure. **Journal of computer-aided molecular design**, v. 22, n. 6-7, p. 403–21, 2008.

KUZ'MIN, V. E.; MURATOV, E. N.; ARTEMENKO, A. G.; VARLAMOVA, E. V.; GORB, L.; WANG, J.; LESZCZYNSKI, J. Consensus QSAR modeling of phosphor-containing chiral AChE inhibitors. **QSAR & combinatorial science**, v. 28, n. 6-7, p. 664–677, jul. 2009.

LEPOITTEVIN, J. **Contact Dermatitis**. 5. ed. Berlin: Springer Berlin Heidelberg, 2011. p. 91–110

LI, J. W.-H.; VEDERAS, J. C. Drug discovery and natural products: end of an era or an endless frontier? **Science (New York, N.Y.)**, v. 325, n. 5937, p. 161–5, 10 jul. 2009.

- LI, S.; FEDOROWICZ, A.; SINGH, H.; SODERHOLM, S. C. Application of the random forest method in studies of local lymph node assay based skin sensitization data. **Journal of chemical information and modeling**, v. 45, n. 4, p. 952–64, 2005.
- LI, Y.; TSENG, Y. J.; PAN, D.; LIU, J.; KERN, P. S.; GERBERICK, G. F.; HOPFINGER, A. J. 4D-fingerprint categorical QSAR models for skin sensitization based on the classification of local lymph node assay measures. **Chemical research in toxicology**, v. 20, n. 1, p. 114–28, jan. 2007.
- LIEN, E. J.; GAO, H. QSAR analysis of skin permeability of various drugs in man as compared to in vivo and in vitro studies in rodents. **Pharmaceutical research**, v. 12, n. 4, p. 583–7, maio 1995.
- LIPINSKI, C. A.; LOMBARDO, F.; DOMINY, B. W.; FEENEY, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings IPII of original article: S0169-409X(96)00423-1. The article was originally published in *Advanced Drug Delivery Reviews* 23 (1997) 3. **Advanced Drug Delivery Reviews**, v. 23, n. 1-3, p. 3–26, 1 mar. 1997.
- LIPINSKI, C.; HOPKINS, A. Navigating chemical space for biology and medicine. **Nature**, v. 432, n. 7019, p. 855–61, 16 dez. 2004.
- 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. 853–62, out. 2004.
- LOW, Y.; UEHARA, T.; MINOWA, Y.; YAMADA, H.; OHNO, Y.; URUSHIDANI, T.; SEDYKH, A.; MURATOV, E.; KUZ'MIN, V.; FOURCHES, D.; ZHU, H.; RUSYN, I.; TROPSHA, A. Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches. **Chemical research in toxicology**, v. 24, n. 8, p. 1251–62, 15 ago. 2011.
- LU, J.; ZHENG, M.; WANG, Y.; SHEN, Q.; LUO, X.; JIANG, H.; CHEN, K. Fragment-based prediction of skin sensitization using recursive partitioning. **Journal of computer-aided molecular design**, v. 25, n. 9, p. 885–93, set. 2011.
- MAGEE, P. S.; HOSTÝNEK, J. J.; MAIBACH, H. I.; FRANCISCO, S. A classification model for allergic contact dermatitis. **QSAR & combinatorial science**, v. 13, n. 1, p. 22–33, jun. 1994.
- MAGGIORA, G. M. On outliers and activity cliffs--why QSAR often disappoints. **Journal of chemical information and modeling**, v. 46, n. 4, p. 1535, 2006.
- MAGGIORA, G. M. The reductionist paradox: are the laws of chemistry and physics sufficient for the discovery of new drugs? **Journal of computer-aided molecular design**, v. 25, n. 8, p. 699–708, ago. 2011.

- MAGNUSSON, B.; KLIGMAN, A. M. The identification of contact allergens by animal assay. The guinea pig maximization test. **The Journal of investigative dermatology**, v. 52, n. 3, p. 268–76, mar. 1969.
- MAGNUSSON, B. M.; ANISSIMOV, Y. G.; CROSS, S. E.; ROBERTS, M. S. Molecular size as the main determinant of solute maximum flux across the skin. **The Journal of investigative dermatology**, v. 122, n. 4, p. 993–9, abr. 2004.
- MAGNUSSON, B. M.; PUGH, W. J.; ROBERTS, M. S. Simple rules defining the potential of compounds for transdermal delivery or toxicity. **Pharmaceutical research**, v. 21, n. 6, p. 1047–54, jun. 2004.
- MCGEE, P. Clinical trials on the move. **Drug discovery & development**, v. 9, n. 6, p. 16–22, 2006.
- MEDINA-FRANCO, J. L.; GIULIANOTTI, M. A.; WELMAKER, G. S.; HOUGHTEN, R. A. Shifting from the single to the multitarget paradigm in drug discovery. **Drug discovery today**, v. 18, n. 9-10, p. 495–501, maio 2013.
- MELVILLE, J. L.; BURKE, E. K.; HIRST, J. D. Machine learning in virtual screening. **Combinatorial chemistry & high throughput screening**, v. 12, n. 4, p. 332–43, maio 2009.
- MENDOZA, L. Phase 0 clinical trials will overcome stagnation of anticancer drug development? **Klinická onkologie : casopis České a Slovenské onkologické společnosti**, v. 24, n. 2, p. 143–5, jan. 2011.
- MERCIER, D. Clustering large datasets. **Electronic review - Linacre College**, 2003.
- MILLER, M. D.; YOURTEE, D. M.; GLAROS, A. G.; CHAPPELOW, C. C.; EICK, J. D.; HOLDER, A. J. Quantum mechanical structure-activity relationship analyses for skin sensitization. **Journal of chemical information and modeling**, v. 45, n. 4, p. 924–9, 2005.
- MODA, T. L.; TORRES, L. G.; CARRARA, A. E.; ANDRICOPULO, A. D. PK/DB: database for pharmacokinetic properties and predictive in silico ADME models. **Bioinformatics (Oxford, England)**, v. 24, n. 19, p. 2270–1, 1 out. 2008.
- MOSS, G. P.; CRONIN, M. T. D. Quantitative structure-permeability relationships for percutaneous absorption: re-analysis of steroid data. **International journal of pharmaceutics**, v. 238, n. 1-2, p. 105–9, 15 maio 2002.
- MOSS, G. P.; SUN, Y.; WILKINSON, S. C.; DAVEY, N.; ADAMS, R.; MARTIN, G. P.; PRAPOPOPOLOU, M.; BROWN, M. B. The application and limitations of mathematical modelling in the prediction of permeability across mammalian skin and polydimethylsiloxane membranes. **The Journal of pharmacy and pharmacology**, v. 63, n. 11, p. 1411–27, nov. 2011.

- MUIR, D. C. G.; HOWARD, P. H. Are there other persistent organic pollutants? A challenge for environmental chemists. **Environmental science & technology**, v. 40, n. 23, p. 7157–66, 1 dez. 2006.
- MURATOV, E. N.; ARTEMENKO, A. G.; VARLAMOVA, E. V.; POLISCHUK, P. G.; LOZITSKY, V. P.; FEDCHUK, A. S.; LOZITSKA, R. L.; GRIDINA, T. L.; KOROLEVA, L. S.; SIL'NIKOV, V. N.; GALABOV, A. S.; MAKAROV, V. A.; RIABOVA, O. B.; WUTZLER, P.; SCHMIDTKE, M.; KUZ'MIN, V. E. Per aspera ad astra: application of Simplex QSAR approach in antiviral research. **Future medicinal chemistry**, v. 2, n. 7, p. 1205–26, jul. 2010.
- NATURE. End of the Lipitor era. **Nature Reviews Drug Discovery**, v. 10, n. 12, p. 889–889, 1 dez. 2011.
- NOGUEIRA, L. J.; MONTANARI, C. A.; DONNICI, C. L. Histórico da evolução da química medicinal e a importância da lipofilia: de Hipócrates e Galeno a Paracelsus e as contribuições de Overton e de Hansch Histórico da evolução da química medicinal e a importância da lipofilia : de Hipócrates e Galeno a Para. **Revista virtual de química**, v. 1, n. 3, p. 227–240, 2009.
- OECD. **OECD principles for the validation, for regulatory purposes, of (Quantitative) Structure-Activity Relationship models**. Disponível em: <<http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf>>. Acesso em: 11 abr. 2013.
- OECD. **Test No. 429: Skin Sensitisation**. Disponível em: <<http://iccvam.niehs.nih.gov/SuppDocs/FedDocs/OECD/OECD-TG429-2010.pdf>>. Acesso em: 23 jan. 2013.
- OECD. **The adverse outcome pathway for skin sensitisation initiated by covalent binding to proteins part 1: scientific evidence**. Disponível em: <[http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono\(2012\)10/part1&doclanguage=en](http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2012)10/part1&doclanguage=en)>. Acesso em: 11 abr. 2012.
- OPREA, T. I.; MATTER, H. Integrating virtual screening in lead discovery. **Current opinion in chemical biology**, v. 8, n. 4, p. 349–58, ago. 2004.
- OVERINGTON, J. P.; AL-LAZIKANI, B.; HOPKINS, A. L. How many drug targets are there? **Nature reviews. Drug discovery**, v. 5, n. 12, p. 993–6, dez. 2006.
- PATEL, H.; TEN BERGE, W.; CRONIN, M. T. D. Quantitative structure-activity relationships (QSARs) for the prediction of skin permeation of exogenous chemicals. **Chemosphere**, v. 48, n. 6, p. 603–13, ago. 2002.
- PAUL, S. M.; MYTELKA, D. S.; DUNWIDDIE, C. T.; PERSINGER, C. C.; MUNOS, B. H.; LINDBORG, S. R.; SCHACHT, A. L. How to improve R&D productivity: the

pharmaceutical industry's grand challenge. **Nature reviews. Drug discovery**, v. 9, n. 3, p. 203–14, mar. 2010.

PETERS, J.-U. Polypharmacology - Foe or Friend? **Journal of medicinal chemistry**, 22 ago. 2013.

PHRMA. **Drug Discovery and Development: Overview** (M. S. Chorghade, Ed.). Washington, DC: John Wiley & Sons, Inc., 14 abr. 2007. Disponível em: <[http://www.innovation.org/drug\\_discovery/objects/pdf/RD\\_Brochure.pdf](http://www.innovation.org/drug_discovery/objects/pdf/RD_Brochure.pdf)>.

PHRMA. **Pharmaceutical research and manufacturers of America, 2013 biopharmaceutical research industry profile**. Washington, DC: [s.n.]. Disponível em: <<http://www.phrma.org/profiles-reports>>.

PIOVESAN, M. F.; LABRA, M. E. Institutional change and political decision-making in the creation of the Brazilian National Health Surveillance Agency. **Cadernos de Saúde Pública**, v. 23, n. 6, p. 1373–1382, jun. 2007.

PLATT, J. R. Influence of Neighbor Bonds on Additive Bond Properties in Paraffins. **The Journal of Chemical Physics**, v. 15, n. 6, p. 419, 1947.

PLENGE, R. M.; SCOLNICK, E. M.; ALTSHULER, D. Validating therapeutic targets through human genetics. **Nature reviews. Drug discovery**, v. 12, n. 8, p. 581–94, ago. 2013.

POLISHCHUK, P. G.; MURATOV, E. N.; ARTEMENKO, A. G.; KOLUMBIN, O. G.; MURATOV, N. N.; KUZ'MIN, V. E. Application of random forest approach to QSAR prediction of aquatic toxicity. **Journal of chemical information and modeling**, v. 49, n. 11, p. 2481–8, nov. 2009.

POTTS, R. O.; GUY, R. H. Predicting skin permeability. **Pharmaceutical research**, v. 9, n. 5, p. 663–9, maio 1992.

POTTS, R. O.; GUY, R. H. A predictive algorithm for skin permeability: the effects of molecular size and hydrogen bond activity. **Pharmaceutical research**, v. 12, n. 11, p. 1628–33, dez. 1995.

RAUNIO, H. In silico toxicology - non-testing methods. **Frontiers in pharmacology**, v. 2, n. June, p. 33, jan. 2011.

RECANATINI, M.; BOTTEGONI, G.; CAVALLI, A. In silico antitarget screening. **Drug Discovery Today: Technologies**, v. 1, n. 3, p. 209–215, dez. 2004.

REGNSTROM, J.; KOENIG, F.; ARONSSON, B.; REIMER, T.; SVENDSEN, K.; TSIGKOS, S.; FLAMION, B.; EICHLER, H.-G.; VAMVAKAS, S. Factors associated with success of market authorisation applications for pharmaceutical drugs submitted

to the European Medicines Agency. **European journal of clinical pharmacology**, v. 66, n. 1, p. 39–48, jan. 2010.

RIPPHAUSEN, P.; NISIUS, B.; BAJORATH, J. State-of-the-art in ligand-based virtual screening. **Drug discovery today**, v. 16, n. 9-10, p. 372–6, maio 2011.

RIPPHAUSEN, P.; NISIUS, B.; PELTASON, L.; BAJORATH, J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. **Journal of medicinal chemistry**, v. 53, n. 24, p. 8461–7, 23 dez. 2010.

ROBERTS, D. W.; APTULA, A. O. Determinants of skin sensitisation potential. **Journal of applied toxicology : JAT**, v. 28, n. 3, p. 377–87, abr. 2008.

ROBERTS, D. W.; PATLEWICZ, G.; DIMITROV, S. D.; LOW, L. K.; APTULA, A. O.; KERN, P. S.; DIMITROVA, G. D.; COMBER, M. I. H.; PHILLIPS, R. D.; NIEMELÄ, J.; MADSEN, C.; WEDEBYE, E. B.; BAILEY, P. T.; MEKENYAN, O. G. TIMES-SS—a mechanistic evaluation of an external validation study using reaction chemistry principles. **Chemical research in toxicology**, v. 20, n. 9, p. 1321–30, set. 2007.

ROLLIN, B. E. The regulation of animal research and the emergence of animal ethics: a conceptual history. **Theoretical medicine and bioethics**, v. 27, n. 4, p. 285–304, jan. 2006.

SAINT-MEZARD, P.; ROSIERES, A.; BERARD, F.; DUBOIS, B.; KAISERLIAN, D.; NICOLAS, J.; KRASTEVA, M. Allergic contact dermatitis. **European journal of dermatology : EJD**, v. 14, n. 5, p. 284–95, 2004.

SCHEUPLEIN, R. J.; BLANK, I. H. Mechanism of percutaneous absorption. IV. Penetration of nonelectrolytes (alcohols) from aqueous solutions and from pure liquids. **The Journal of investigative dermatology**, v. 60, n. 5, p. 286–296, maio 1973.

SCHEUPLEIN, R. J.; BLANK, I. H.; BRAUNER, G. J.; MACFARLANE, D. J. Percutaneous absorption of steroids. **The Journal of investigative dermatology**, v. 52, n. 1, p. 63–70, jan. 1969.

SCHEUPLEIN, R. J.; BRONAUGH, R. L. Percutaneous Absorption. In: GOLDSMITH, L. A. (Ed.). **Biochemistry and physiology of the skin**. New York, NY: Oxford University Press, 1983. p. 1255–1295.

SCHLEDE, E.; ABERER, W.; FUCHS, T.; GERNER, I.; LESSMANN, H.; MAURER, T.; ROSSBACHER, R.; STROPP, G.; WAGNER, E.; KAYSER, D. Chemical substances and contact allergy—244 substances ranked according to allergenic potency. **Toxicology**, v. 193, n. 3, p. 219–259, dez. 2003.

SCHOLES, E. W.; BASKETTER, D. A.; SARLL, A. E.; KIMBER, I.; EVANS, C. D.; MILLER, K.; ROBBINS, M. C.; HARRISON, P. T.; WAITE, S. J. The local lymph

node assay: results of a final inter-laboratory validation under field conditions. **Journal of applied toxicology : JAT**, v. 12, n. 3, p. 217–22, jun. 1992.

SHYU, K.-G. Serotonin 5-HT<sub>2B</sub> receptor in cardiac fibroblast contributes to cardiac hypertrophy: a new therapeutic target for heart failure? **Circulation research**, v. 104, n. 1, p. 1–3, 2 jan. 2009.

SIEGEL, P. D.; FEDOROWICZ, A.; BUTTERWORTH, L.; LAW, B.; ANDERSON, S. E.; SNYDER, J.; BEEZHOLD, D. Physical-chemical and solvent considerations in evaluating the influence of carbon chain length on the skin sensitization activity of 1-bromoalkanes. **Toxicological sciences : an official journal of the Society of Toxicology**, v. 107, n. 1, p. 78–84, jan. 2009.

STREBHARDT, K.; ULLRICH, A. Paul Ehrlich's magic bullet concept: 100 years of progress. **Nature reviews. Cancer**, v. 8, n. 6, p. 473–80, jun. 2008.

STRID, J.; STROBEL, S. Skin barrier dysfunction and systemic sensitization to allergens through the skin. **Current drug targets. Inflammation and allergy**, v. 4, n. 5, p. 531–41, out. 2005.

SWANN, S. L.; BROWN, S. P.; MUCHMORE, S. W.; PATEL, H.; MERTA, P.; LOCKLEAR, J.; HAJDUK, P. J. A unified, probabilistic framework for structure- and ligand-based virtual screening. **Journal of medicinal chemistry**, v. 54, n. 5, p. 1223–32, 10 mar. 2011.

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

TAFT, R. W. Linear Steric Energy Relationships. **Journal of the American Chemical Society**, v. 75, n. 18, p. 4538–4539, set. 1953.

THOMSON REUTERS. **2012 CMR International pharmaceutical R&D FactBook**. [s.l.: s.n.].

TODESCHINI, R.; CONSONNI, V. **Handbook of Molecular Descriptors**. Weinheim, Germany: Wiley-VCH, 2000.

TROPSHA, A. Best practices for QSAR model development, validation, and exploitation. **Molecular informatics**, v. 29, n. 6-7, p. 476–488, 6 jul. 2010.

URSU, O.; RAYAN, A.; GOLDBLUM, A.; OPREA, T. I. Understanding drug-likeness. **Wiley Interdisciplinary Reviews: Computational Molecular Science**, v. 1, n. 5, p. 760–781, 28 set. 2011.

- VANDENBERG, J. I.; PERRY, M. D.; PERRIN, M. J.; MANN, S. A.; KE, Y.; HILL, A. P. hERG K<sup>+</sup> Channels: Structure, Function, and Clinical Significance. **Physiological Reviews**, v. 92, n. 3, p. 1393–1478, 18 jul. 2012.
- VARNEK, A.; FOURCHES, D.; SIEFFERT, N.; SOLOV'EV, V. P.; HILL, C.; LECOMTE, M. QSPR Modeling of the Am III /Eu III Separation Factor: How Far Can we Predict? **Solvent extraction and ion exchange**, v. 25, n. 1, p. 1–26, mar. 2007.
- VEBER, D. F.; JOHNSON, S. R.; CHENG, H.-Y.; SMITH, B. R.; WARD, K. W.; KOPPLE, K. D. Molecular properties that influence the oral bioavailability of drug candidates. **Journal of medicinal chemistry**, v. 45, n. 12, p. 2615–23, 6 jun. 2002.
- VEDANI, A.; DOBLER, M. 5D-QSAR: the key for simulating induced fit? **Journal of medicinal chemistry**, v. 45, n. 11, p. 2139–49, 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. 3700–3, 2 jun. 2005.
- WALTERS, W. P.; GREEN, J.; WEISS, J. R.; MURCKO, M. A. What do medicinal chemists actually make? A 50-year retrospective. **Journal of medicinal chemistry**, v. 54, n. 19, p. 6405–16, 13 out. 2011.
- WASZKOWYCZ, B.; CLARK, D. E.; GANCIA, E. Outstanding challenges in protein-ligand docking and structure-based virtual screening. **Wiley Interdisciplinary Reviews: Computational Molecular Science**, v. 1, n. 2, p. 229–259, 16 mar. 2011.
- WATANABE, K. H.; ANDERSEN, M. E.; BASU, N.; CARVAN, M. J.; CROFTON, K. M.; KING, K. A.; SUÑOL, C.; TIFFANY-CASTIGLIONI, E.; SCHULTZ, I. R. Defining and modeling known adverse outcome pathways: domoic acid and neuronal signaling as a case study. **Environmental toxicology and chemistry / SETAC**, v. 30, n. 1, p. 9–21, jan. 2011.
- WELLING, M. **A first encounter with Machine Learning**. Irvine, CA.: University of California, 2011. p. 1–93
- WERMUTH, C.; GANELLIN, C.; LINDBERG, P. GLOSSARY OF TERMS USED IN MEDICINAL CHEMISTRY (IUPAC Recommendations 1998). **Pure and applied chemistry**, v. 70, n. 5, p. 1129–1143, 1998.
- WIENER, H. Influence of Interatomic Forces on Paraffin Properties. **The Journal of Chemical Physics**, v. 15, n. 10, p. 766, 1947.
- WILLETT, P. Similarity-based virtual screening using 2D fingerprints. **Drug discovery today**, v. 11, n. 23-24, p. 1046–53, dez. 2006.

- WILSCHUT, A.; TEN BERGE, W. F.; ROBINSON, P. J.; MCKONE, T. E. Estimating skin permeation. The validation of five mathematical skin permeation models. **Chemosphere**, v. 30, n. 7, p. 1275–96, abr. 1995.
- XIA, X. R. Quantitative Structure-Permeability Relationships. In: RIVIERE, J. E. (Ed.). **Comparative Pharmacokinetics: Principles, Techniques, and Applications**. 2. ed. [s.l.] John Wiley & Sons, Inc., 2011. p. 27–38.
- XU, C.; CHENG, F.; CHEN, L.; DU, Z.; LI, W.; LIU, G.; LEE, P. W.; TANG, Y. In silico Prediction of Chemical Ames Mutagenicity. **Journal of chemical information and modeling**, v. 52, n. 11, p. 2840–7, 17 out. 2012.
- XUE, L.; BAJORATH, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. **Combinatorial chemistry & high throughput screening**, v. 3, n. 5, p. 363–72, out. 2000.
- YANG, S.-Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. **Drug discovery today**, v. 15, n. 11-12, p. 444–50, jun. 2010.
- YOUNG, D. C. **Computational drug design: A guide for computational and medicinal chemists**. Hoboken, NJ: John Wiley & Sons, Inc., 2009. v. 14p. 1–307
- YOUNG, S. S.; YUAN, F.; ZHU, M. Chemical Descriptors Are More Important Than Learning Algorithms for Modelling. **Molecular informatics**, v. 31, n. 10, p. 707–710, 11 out. 2012.
- YUAN, H.; HUANG, J.; CAO, C. Prediction of skin sensitization with a particle swarm optimized support vector machine. **International journal of molecular sciences**, v. 10, n. 7, p. 3237–54, jun. 2009.
- ZHANG, L.; TSAI, K.-C.; DU, L.; FANG, H.; LI, M.; XU, W. How to generate reliable and predictive CoMFA models. **Current medicinal chemistry**, v. 18, n. 6, p. 923–30, jan. 2011.
- ZHU, H.; TROPSHA, A.; FOURCHES, D.; VARNEK, A.; PAPA, E.; GRAMATICA, P.; OBERG, T.; DAO, P.; CHERKASOV, A.; TETKO, I. V. Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. **Journal of chemical information and modeling**, v. 48, n. 4, p. 766–84, abr. 2008.
- ZIEGLER, V.; RICHTER, C.; ZIEGLER, B.; HAUSTEIN, U. F. INPRET--database on predictive tests (allergy). **Seminars in dermatology**, v. 8, n. 2, p. 80–2, jun. 1989.

## ANEXOS

### Conjuntos de dados

Todos os conjuntos de dados usados nesse estudo estão disponíveis em formato SDF na mídia digital que acompanha essa dissertação.

### Tabelas

As tabelas A1-3 estão disponíveis em PDF na mídia digital que acompanha essa dissertação.

**Tabela A1.** Lista de compostos da Scorecard e predições correspondentes feitas pelos modelos 5 e 7 em arquivo SDF.

**Tabela A2.** Conjunto de dados experimentais e preditos de sensibilização e permeabilidade para 283 compostos.

**Tabela A3.** Dados de sensibilização e permeabilidade experimentais e preditos para 11 compostos identificados por HCA.

### Artigos

A capa dos manuscritos submetidos para o Toxicology and Applied Pharmacology segue nas próximas páginas.

**Novel computational tools to predict chemically-induced skin reactions. Part I: QSAR Models of Skin Sensitization and their application to identify potentially hazardous compounds**

*Vinícius Medeiros Alves,<sup>a,b</sup> Eugene Muratov,<sup>b,c</sup> Denis Fourches,<sup>b</sup> Judy Strickland,<sup>d</sup> Nicole Kleinstreuer,<sup>d</sup> Carolina Horta Andrade,<sup>a</sup> and Alexander Tropsha<sup>b,\*</sup>.*

<sup>a</sup>Laboratory of Molecular Modeling and Design, Faculty of Pharmacy, Federal University of Goiás, Goiânia, GO, 74605-220, Brazil.

<sup>b</sup>Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, 27599, USA.

<sup>c</sup>Laboratory of Theoretical Chemistry, A.V. Bogatsky Physical-Chemical Institute NAS of Ukraine, Odessa, 65080, Ukraine.

<sup>d</sup>ILS, Inc./Contractor supporting the NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM), P.O. Box 13501, Research Triangle Park, NC, 27709, USA.

**Corresponding Author**

\* Address for correspondence: 100K Beard Hall, Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, 27599, USA; Telephone: (919) 966-2955; FAX: (919) 966-0204; E-mail: [alex\\_tropsha@unc.edu](mailto:alex_tropsha@unc.edu)

**Novel computational tools to predict chemically-induced skin reactions. Part II: QSAR Models of Skin Permeability and the Relationship between Skin Permeability and Skin Sensitization**

*Vinícius Medeiros Alves,<sup>a,b</sup> Eugene Muratov,<sup>b,c</sup> Denis Fourches,<sup>b</sup> Judy Strickland,<sup>d</sup> Nicole Kleinstreuer,<sup>d</sup> Carolina Horta Andrade,<sup>a</sup> and Alexander Tropsha<sup>b,\*</sup>.*

<sup>a</sup>Laboratory of Molecular Modeling and Design, Faculty of Pharmacy, Federal University of Goiás, Goiânia, GO, 74605-220, Brazil.

<sup>b</sup>Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, 27599, USA.

<sup>c</sup>Laboratory of Theoretical Chemistry, A.V. Bogatsky Physical-Chemical Institute NAS of Ukraine, Odessa, 65080, Ukraine.

<sup>d</sup>ILS, Inc./Contractor supporting the NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM), P.O. Box 13501, Research Triangle Park, NC, 27709, USA.

**Corresponding Author**

\* Address for correspondence: 100K Beard Hall, Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, 27599, USA; Telephone: (919) 966-2955; FAX: (919) 966-0204; E-mail: [alex\\_tropsha@unc.edu](mailto:alex_tropsha@unc.edu)