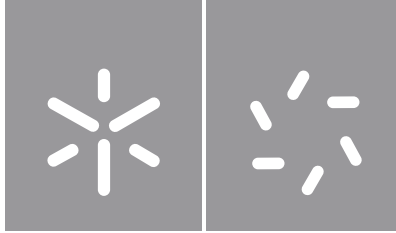


Universidade do Minho
Escola de Ciências

João Miguel Inácio

Desenvolvimento *in silico* de novos agentes antimicrobianos derivados da polimixina B



Universidade do Minho
Escola de Ciências

João Miguel Inácio

Desenvolvimento *in silico* de novos agentes antimicrobianos derivados da polimixina B

Dissertação de Mestrado
Mestrado em Química Medicinal

Trabalho efetuado sob a orientação do
Doutor Filipe Carlos Teixeira Gil
e da
Doutora Paula Alexandra da Silva Jorge

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho



Atribuição-NãoComercial
CC BY-NC

<https://creativecommons.org/licenses/by-nc/4.0/>

AGRADECIMENTOS

Chegado ao fim do desenvolvimento do trabalho, gostaria de agradecer a todos os que contribuíram de forma direta e indireta para o meu percurso académico.

Aos meus orientadores, Doutor Filipe Teixeira e Doutora Paula Jorge, que serviram de guias no desenvolvimento deste trabalho, quero agradecer-lhes pelo conhecimento que me proporcionaram, espírito de incentivo e disponibilidade ao longo do desenvolvimento do trabalho.

À minha família, particularmente aos meus pais e irmãos, que mesmo à distância me transmitiram espírito de conforto ao longo da minha formação.

Ao laboratório LIBRO do Centro de Engenharia biológica da Universidade do Minho, por me ter facultado as condições e materiais necessários para a realização dos ensaios biológicos.

Ao meu colega de turma e amigo Júdse Zeca pela partilha de conhecimento, apoio e incentivos.

Agradecimentos especiais também vão para Camões, Instituto de Cooperação e da Língua, e Instituto de Bolsas de Estudos de Moçambique, entidades financiadoras da minha bolsa de estudos. De modo extensivo, gostaria de agradecer à Fundação para a Ciência e Tecnologia (FCT), a entidade financiadora deste trabalho, no âmbito do projeto POLYmix-POLYmic (2022.06595.PTDC), do financiamento programático ao CQUM (UID/QUI/00686/2020), do financiamento estratégico ao CEB (UIDB/04469/2020), e do contrato CEECIND/00194/2020.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

RESUMO

Titulo: Desenvolvimento *in silico* de novos agentes antimicrobianos derivados da polimixina B.

A resistência antimicrobiana (RAM) é um dos principais problemas de saúde pública da atualidade, provocando uma morbimortalidade significativa. A evolução deste problema, agravada pelo lento desenvolvimento de novos antimicrobianos, levou à reconsideração do uso das polimixinas, fármacos que já estavam em desuso devido a sua alta toxicidade. Na tentativa de diminuir a sua toxicidade e/ou melhorar a sua atividade antimicrobiana, vários análogos de polimixinas são gerados através de diferentes estratégias, principalmente experimentais. Como tal, estão em falta abordagens mais rápidas e fiáveis para tornar o *design* de análogos mais eficaz, a fim de combater a RAM o mais rápido possível. A solução para acelerar a descoberta de novos fármacos provavelmente está no uso de abordagens *in silico*, com métodos de *machine learning* (ML), devido ao seu ritmo mais rápido e baixo custo.

Neste trabalho, a atividade de análogos da polimixina B foi modelada usando modelos semi-quantitativos de relação estrutura-atividade baseados em ML. Neste contexto, foram aplicados três algoritmos diferentes de ML (árvore de decisão, floresta aleatória e AdaBoost) em dez famílias diferentes de descritores moleculares ao conjunto de dados de 413 pares molécula/microrganismo proveniente da PubChem e dos ensaios laboratoriais.

O modelo DT/Estate_VSA destacou-se como promissor, com exatidões e previsões verdadeiras altas, bem como previsões falsas negativas e falsas positivas muito baixas. Este modelo foi aplicado para prever a atividade antimicrobiana de seis análogos das polimixinas B e E, sendo que todos são previstos como promissores para *Pseudomonas* e não promissores para *Acinetobacter*. Para *Escherichia*, os três análogos mais hidrofílicos foram previstos como promissores e os outros três como não promissores. Estes análogos estão a ser sintetizados e posteriormente serão testados quanto a sua atividade *in vitro*.

Palavras-chave: actividade antimicrobiana; análogos de polimixinas; estudos *in silico*; *machine learning*; modelos QSAR.

ABSTRACT

Title: In *silico* development of novel antimicrobial agents derived from polymyxin B.

Antimicrobial resistance (AMR) is one of the main public health problems today, causing significant morbidity and mortality. The evolution of this problem, aggravated by the slow development of new antimicrobials, led to the reconsideration of the use of polymyxins, drugs that were already in disuse due to their high toxicity. In an attempt to decrease its toxicity and/or improve its antimicrobial activity, several polymyxin analogues are generated through different strategies, mainly experimental. As such, faster and more reliable approaches to make analogue *design* more effective in order to tackle AMR as quickly as possible are lacking. The solution to accelerate the discovery of new drugs probably lies in the use of in *silico* approaches, with *machine learning* (ML) methods, due to their faster pace and low cost.

In this work, the activity of polymyxin B analogues was modelled using semi-quantitative structure-activity relationship models based on ML. In this context, three different ML algorithms (decision tree, random forest, and AdaBoost) were applied in ten different families of molecular descriptors to the dataset of 413 molecule/microorganism pairs from PubChem and laboratory assays.

The DT/Estate_VSA model stood out as promising, with high true accuracies and predictions, as well as very low false negative and false positive predictions. This model was applied to predict the antimicrobial activity of six polymyxin B and E analogues, all of which are predicted to be promising for *Pseudomonas* and not promising for *Acinetobacter*. For *Escherichia*, the three most hydrophilic analogues were predicted to be promising and the other three to be unpromising. These analogues are being synthesized and will later be tested for their in *vitro* activity.

Keywords: antimicrobial activity; *in silico* studies; *machine learning*; QSAR models; polymyxin analogues.

Índice

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS	ii
AGRADECIMENTOS	iii
DECLARAÇÃO DE INTEGRIDADE.....	iv
RESUMO	v
ABSTRACT	vi
LISTA DE ABREVIATURAS E SIGLAS.....	x
LISTA DE FIGURAS	xii
LISTA DE TABELAS.....	xv
1. INTRODUÇÃO.....	1
1.1. Antimicrobianos.....	2
1.1.1. Conceito.....	2
1.1.2. História dos antimicrobianos.....	2
1.1.3. Resistência aos antimicrobianos	6
1.1.4. Polimixinas.....	7
1.2. Métodos de pesquisa de novos fármacos	10
1.3. Modelos QSAR.....	12
1.3.1. QSAR baseado em modelos lineares.....	12
1.3.2. QSAR baseado em métodos ML.....	15
1.3.3. Modelos de classificação em ML.....	16
1.3.4. Validação cruzada de modelos.....	20
1.3.5. Métricas de avaliação de modelos.....	20
1.4. Descritores Moleculares.....	22
1.4.1. Descritores 0D	23
1.4.2. Descritores 1D	23
1.4.3. Descritores 2D	24
1.4.4. Descritores 3D	35

1.5. Representação molecular	35
1.6. Testes de suscetibilidade aos antimicrobianos	36
1.7. Objetivos	36
1.7.1. Objetivo geral	36
1.7.2. Objetivos específicos	37
2. METODOLOGIA	38
2.1. Colheita inicial dos dados	39
2.1.1. Caracterização dos dados.....	39
2.1.2. Cura de dados.....	39
2.2. Desenvolvimento do primeiro modelo.....	40
2.2.1. Geração de descritores moleculares.....	40
2.2.2. Exploração dos algoritmos	41
2.2.3. Análise do modelo.....	42
2.3. Ensaio de suscetibilidade <i>in vitro</i>	42
2.3.1. Microrganismos e reagentes	42
2.3.2. Meios de cultura e soluções.....	43
2.3.3. Curva de calibração para ajuste da concentração celular.....	44
2.3.4. Determinação da MIC.....	44
2.4. Treino do novo modelo	45
2.5. Aplicação do modelo	45
3. RESULTADOS E DISCUSSÃO	46
3.1. Caracterização dos dados colhidos	47
3.1.1. Distribuição dos valores da MIC.....	51
3.1.2. Categorização dos quartis da MIC.....	51
3.2. Caracterização dos modelos da primeira série.	52
3.3. Caracterização do melhor modelo da 1ª série	56

3.3.1. Desempenho do modelo AdaBoost/CKP	56
3.3.2. Importância cada variável no modelo por permuta	57
3.3.3. Influência dos descritores moleculares.....	57
3.3.4. Influência do alvo biológico	59
3.4. Suscetibilidade da <i>Shigella sonnei</i> , <i>Proteus mirabilis</i> e <i>Listeria monocytogenes</i> à colistina e à polimixina B	61
3.5. Caracterização dos dados da 2ª série	63
3.5.1. Distribuição dos valores da MIC	64
3.5.2. Categorização dos quartis da MIC.....	65
3.6. Caracterização dos modelos da 2ª série.....	65
3.7. Caracterização do melhor modelo da 2ª série	69
3.7.1. Desempenho dos melhores modelos da segunda série	69
3.7.2. Importância de cada variável no modelo por permuta	70
3.7.3. Influência dos descritores moleculares no modelo Adaboost/PEOE_VSA	71
3.7.4. Influência dos descritores moleculares no modelo DT/Estate_VSA	74
3.7.5. Influência do alvo biológico nos modelos AdaBoost/PEOE_VSA e DT/Estate_VSA.....	76
3.8. Mutações sistemáticas da polimixina B.....	79
3.9. Aplicação do modelo	81
4. CONCLUSÕES	86
5. PERSPETIVAS DO TRABALHO.....	89
6. REFERÊNCIAS BIBLIOGRÁFICAS.....	90
7. ANEXOS	101

LISTA DE ABREVIATURAS E SIGLAS

6-APA	Ácido 6-aminopenicilâmico (do inglês, <i>6-aminopenicillanic acid</i>)
AdaBoost	Reforço Adaptável (do inglês, <i>Adaptive Boosting</i>)
AMP	Péptidos Antimicrobianos (do inglês, <i>Antimicrobial Peptides</i>)
ATCC	Coleção Americana Culturas Tipo (<i>do inglês, American Type Culture Collection</i>)
BHI	Caldo de Cérebro e Coração (do inglês, <i>Brain Heart Infusion</i>)
CAS	Serviço de Resumos Químicos (do inglês, <i>Chemical Abstracts Service</i>)
CDC	Centro de Controlo e Prevenção de Doenças (do inglês, <i>Center for Disease Control and Prevention</i>)
CEB	Centro de Engenharia Biológica
CFU	Unidade Formadora de Colónias (do inglês, <i>Colony Forming Unit</i>)
CID	Número de Identificação do Composto (do inglês, <i>Compound Identifier</i>)
DO	Densidade Ótica
DT	Árvore de Decisão (do inglês, <i>Decision Tree</i>)
EMA	Agência Europeia de Medicamentos (do inglês, <i>European Medicines Agency</i>)
E-state	Estado Eletrotopológico (do inglês, <i>Eletrotopological State</i>)
EUCAST	Comitê Europeu de Testes de Suscetibilidade Antimicrobiana (do inglês, <i>European Commitee on Antimicrobial Susceptibility Testing</i>)
FCT	Fundação para a Ciência e a Tecnologia
ICA	Análise de Componentes Independentes (do inglês, <i>Independent Component Analysis</i>)
LB	Caldo de Lisogenia (do inglês, <i>Lysogeny Broth</i>)
LDA	Alocação Latente de Dirichlet (do inglês, <i>Latent Dirichlet Allocation</i>)
L-Dab	Ácido L-diaminobutírico
Log p	Coefficiente de partição

LPS	Lipopolissacarídeos
MCR	Resistência Mobilizada à Colistina (do inglês, <i>Mobilized Colistin Resistance</i>)
MHB	Caldo Muller-Hinton (do inglês, <i>Muller-Hinton Broth</i>)
MIC	Concentração Mínima Inibitória (do inglês, <i>Minimal Inhibitory concentration</i>)
ML	Aprendizagem Automática (do inglês, <i>Machine Learning</i>)
MR	Multirresistente
NaCl	Cloreto de sódio
OMS	Organização Mundial da Saúde
PCA	Análise de Componentes Principais (do inglês, <i>Principal Component Analysis</i>)
PEOE	Equalização Parcial de Eletronegatividades Orbitais (do inglês, <i>Partial Equalization of Orbital Electronegativities</i>)
PI	Importância de Permutação (do inglês, <i>Permutation Importance</i>)
PIB	Produto Interno Bruto
QSAR	Relação Quantitativa Estrutura-Atividade (do inglês, <i>Quantitative Structure-Activity Relationship</i>)
RAM	Resistência Antimicrobiana
RF	Floresta Aleatória (do inglês, <i>Random Forest</i>)
rpm	Rotações Por Minuto
SAR	Relação Estrutura-Actividade (do inglês, <i>Structure-Activity Relationship</i>)
SMILE	Representação Simplificada do Sistema Molecular de Entrada de Linha (do inglês, <i>Simplified Molecular-Input Line-Entry system</i>)
SMR	Refratividade Molecular
TPSA	Área de Superfície Polar Topológica (do inglês, <i>Topological Polar Surface Area</i>)
TSB	Caldo Triptico de Soja (do inglês, <i>Tryptic Soy Broth</i>)
VSA	Área de Superfície de Van der Waals (do inglês, <i>Van der Waals Surface Area</i>)

LISTA DE FIGURAS

Figura 1.1. Linha temporal da descoberta das principais classes de agentes antimicrobianos	3
Figura 1.2. Estrutura química do Salvarsan.....	3
Figura 1.3. Estrutura geral das penicilinas, onde é destacado a vermelho a estrutura do anel β -lactâmico.	4
Figura 1.4. Síntese de derivados semissintéticos da penicilina.....	4
Figura 1.5. Estruturas químicas da meticilina (a) e da ampicilina (b).....	4
Figura 1.6. Metabolismo do prontossil (a) em sulfanilamida (b) e no metabolito inativo triaminobenzeno (c).	5
Figura 1.7. Estrutura química da gramicidina (formil-L-Val-Gly-L-Ala-D-Leu-L-Ala-D-Val-L-Val-	5
Figura 1.8. Estrutura química da tirocidina (ciclo [Asn-Gln-tyr-Val-Orn-Leu-D-phe-Pro-Phe-D-Phe).	6
Figura 1.9. Estruturas químicas da polimixina B (a), onde se destaca a vermelho a cadeia lateral benzil do aminoácido fenilalanina e a azul o ácido gordo 6-metiloctanolílo ligado ao terminal N, e da colistina (b), onde se destaca a vermelho a cadeia lateral isopropil do aminácido leucina da 6 ^a posição e a azul o ácido gordo 5-metileptanoílo ligado ao terminal N.....	8
Figura 1.10. Esquema de Toplis aplicado à síntese de agentes antimicrobianos derivados de benzofuroxano.....	14
Figura 1.11. Diagrama de Craig.....	15
Figura 1.12. Estrutura de uma DT.....	17
Figura 1.13. Esquema do algoritmo da RF.	18
Figura 1.14. Esquema de um modelo AdaBoost.....	19
Figura 1.15. Matriz de confusão para um problema de classificação multiclasse. com três classes (A, B e C).....	22
Figura 1.16. Representação do isopropano como grafo (a) e matriz de adjacência (b).....	24
Figura 3.1. Caracterização dos dados colhidos por tipo de microrganismo (a) e por género taxonómico (b).....	48
Figura 3.2. Caracterização dos dados colhidos por compostos anotados.....	49
Figura 3.4. Histograma da distribuição dos valores da MIC por M_{TYP}	51
Figura 3.5. Valores de diferentes métricas (exatidão, $f(Q1 Q1)$ e $f(Q1 Q4)$) para cada família de descritores e algoritmos: DT (a); RF (b); e AdaBoost (c).	55
Figura 3.6. Matriz de confusão dos conjuntos treino e teste do modelo AdaBoost/CKP.....	56

Figura 3.7. Importância de Permutação (PI) média dos descritores no modelo AdaBoost/CPK calculada usando 10 réplicas do conjunto de dados para cada recurso. As barras de erro representam o desvio padrão da PI nas 10 réplicas e o eixo dos yy à direita indica o valor da PI média, normalizado para percentagem.	57
Figura 3.8. Gráficos de dependência parcial do modelo AdaBoost/CKP: em relação a $^1\chi$ (a), em relação a $^0\chi$ (b) e em relação a $^3\kappa$ (c).....	58
Figura 3.9. Gráfico de dependência parcial do modelo AdaBoost/CPK por M_{Typ}	60
Figura 3.10. Gráficos de dependência parcial do modelo AdaBoost/ CPK por T_xG	61
Figura 3.11. Caracterização de novos dados por M_{Typ} (a) e por T_xG (b).	63
Figura 3.12. Caracterização de novos dados por compostos anotados.....	64
Figura 3.13. Histograma da distribuição dos valores da MIC por M_{Typ}	65
Figura 3.14. Valores de diferentes métricas (exatidão, $f(Q1 Q1)$ e $f(Q1 Q4)$) para cada família de descritores e algoritmos: DT (a), RF (b), e AdaBoost (c). As barras de erro representam o desvio-padrão dos scores.....	67
Figura 3.15. Matriz de confusão dos melhores modelos da segunda série: modelo AdaBoost/PEOE_VSA (a), modelo DT/Estate_VSA (b).	69
Figura 3.16. Importância de variáveis por permuta dos modelos AdaBoost /PEOE_VSA (a) e DT/ Estate_VSA (b). As barras de erro representam o desvio padrão da PI nas 10 réplicas e o eixo dos yy à direita indica o valor da PI média, normalizado para percentagem.	71
Figura 3.17. Gráfico de dependência parcial do modelo AdaBoost usando o descritor molecular PEOE_VSA ₇ (a), PEOE_VSA ₈ (b) e PEOE_VSA ₆ (c).....	73
Figura 3.18. Gráficos de dependência parcial do modelo DT usando os descritores moleculares Estate_VSA ₉ (a), Estate_VSA ₄ (b) e Estate_VSA ₁ (c).	75
Figura 3.19. Gráficos de dependência parcial dos modelos AdaBoost/PEOE_VSA (a) e DT/Estate_VSA (b) em relação a M_{Typ}	77
Figura 3.20. Gráfico de dependência parcial dos modelos AdaBoost /PEOE_VSA (a) e DT/ Estate_VSA (b) em relação a T_xG	78
Figura 3.21. Classificação mais provável quanto à atividade antimicrobiana contra Escherichia de variantes mutadas da polimixina B ao alterar sistematicamente cada resíduo de aminoácido para: Gly (a), Leu (b), Lys (c) e Glu (d).	80
Figura 3.22. Estruturas químicas dos candidatos análogos da polimixina B: Bmim1 (a), Bmim2 (b) e Bmim3 (c).	82

Figura 3.23. Estruturas químicas dos candidatos análogos da polimixina E: Emim1 (a), Emim2 (b) e Emim3 (c).	83
Figura 3.24. Classificação mais provável dos análogos das polimixinas B e E quanto à atividade antimicrobiana contra bactérias dos géneros: <i>Acinetobacter</i> (a), <i>Pseudomonas</i> (b) e <i>Escherichia</i> (c). 84	
Figura A1. Curvas de calibração da concentração celular vs DO para: <i>S. sonnei</i> (a), <i>P. mirabilis</i> (b) e <i>L. monocytogenes</i> (c).	101
Figura A2. Classificação mais provável quanto à atividade antimicrobiana contra <i>Acinetobacter</i> de variantes mutadas da polimixina B ao alterar sistematicamente cada resíduo de aminoácido para: Gly (a), Leu (b), Lys (c) e Glu (d).	109
Figura A3. Classificação mais provável quanto à atividade antimicrobiana contra <i>Pseudomonas</i> de variantes mutadas da polimixina B ao alterar sistematicamente cada resíduo de aminoácido para: Gly (a), Leu (b), Lys (c) e Glu (d).	110

LISTA DE TABELAS

Tabela 1.1. Exemplo da matriz de confusão da classificação binária.....	21
Tabela 1.2. Ilustração de ordem de conectividade do isopropano.....	28
Tabela 1.3. Descritores de propriedade físicas	31
Tabela 2.1. Classificação dos descritores, com indicação do número de descritores (n,) gerado em cada família.....	40
Tabela 2.2. Valores testados para cada parâmetro nos modelos RF e AdaBoost.....	41
Tabela 2.3. Espécies bacterianas usadas na colheita de novos dados de MIC.....	42
Tabela 3.1. Separação dos quartis da MIC.	52
Tabela 3.2. Valores da MIC da colistina e da polimixina B para E. coli, S. sonnei, P. mirabilis e L. monocytogenes.	62
Tabela 3.3 Separação dos quartis da MIC para o conjunto de dados da segunda série.....	65
Tabela A1. Resultados detalhados dos modelos da primeira série.....	102
Tabela A2. Otimização dos hiper-parâmetros nos modelos RF e AdaBoost da primeira série	103
Tabela A3. Otimização dos hiper-parâmetros nos modelos RF e AdaBoost da segunda série	104
Tabela A4. Resultados detalhados dos modelos da segunda série.....	105
Tabela A5. Resultados detalhados da aplicação do modelo DT/Estate_VSA	107

1. INTRODUÇÃO

Este capítulo é dividido em duas partes. Na primeira parte é feita a introdução ao problema, discutindo-se a história dos antimicrobianos, a problemática da resistência aos antimicrobianos, as características estruturais das polimixinas, com enfoque na polimixina B e na colistina (polimixina E), e os seus mecanismos de ação e de toxicidade. Finalmente, apresentam-se novas perspectivas na pesquisa e desenvolvimento de novos fármacos antimicrobianos. Na segunda parte, é feita uma introdução aos métodos usados, em particular os estudos quantitativos da relação estrutura-atividade, o uso de descritores moleculares e métodos de classificação em Aprendizagem Automática (doravante referida pela sigla inglesa ML, *Machine Learning*), assim como os métodos usados em ensaios de atividade biológica.

1.1. Antimicrobianos

1.1.1. Conceito

Um antimicrobiano é definido como uma substância natural, semissintética ou sintética que mata ou inibe o crescimento de microrganismos. Estes dividem-se em antibacterianos, antivirais, antifúngicos e antiparasitários. Por outro lado, um antibiótico é definido como uma substância de origem natural, semissintética ou sintética, que apresenta atividade seletiva contra bactérias, sendo portanto, de potencial uso no tratamento de infeções.¹⁻³ Assim sendo, os termos antibacteriano e antibiótico são sinónimos, sendo que o termo antibiótico é mais usado para os fármacos já usados na prática clínica.

1.1.2. História dos antimicrobianos

Antes da descoberta de antimicrobianos, as doenças infecciosas eram a principal causa de morbimortalidade em populações humanas, com taxa de letalidade de 40 % a 97 %.⁴ Por exemplo, durante a Primeira Guerra Mundial, 70 % das feridas infetadas resultavam em amputação, pois nenhum antimicrobiano estava disponível.⁴⁻⁷ A introdução dos antimicrobianos na prática clínica revolucionou o tratamento das doenças infecciosas, salvando milhões de vidas, e permitiu procedimentos médicos importantes, incluindo cirurgia e quimioterapia contra o cancro, sendo, por isso, considerada uma das descobertas mais importantes na história da medicina.^{8,9} Isto permitiu reduzir a taxa de mortalidade em 25 % a 30 % para pneumonia adquirida na comunidade e associada ao meio hospitalar, 75 % para a endocardite, 60 % para as infeções meningéas e em 11 % para a celulite infecciosa.^{6,10-13}

A evolução da descoberta dos principais antimicrobianos é ilustrada na Figura 1.1, no entanto, será relatada a história até a descoberta dos péptidos antimicrobianos (AMPs, do inglês *Antimicrobial Peptides*), que é o foco deste trabalho.

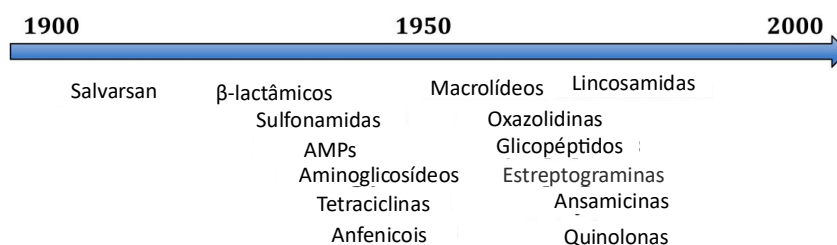


Figura 1.1. Linha temporal da descoberta das principais classes de agentes antimicrobianos (adaptada de Aminov R.).⁶

A história dos antimicrobianos remonta a milênios, com o uso de microrganismos produtores de antimicrobianos para prevenir e tratar doenças infecciosas, havendo relatos do uso de cataplasmas de pão mofado para tratar feridas abertas na China, Grécia e Egito há mais de 2000 anos.¹⁴ No entanto, só em 1910 foi desenvolvido o primeiro antibiótico moderno, o Salvarsan (Arsfenamina) (Figura 1.2), por Paul Ehrlich à base do arsênio para tratar a sífilis.¹⁵

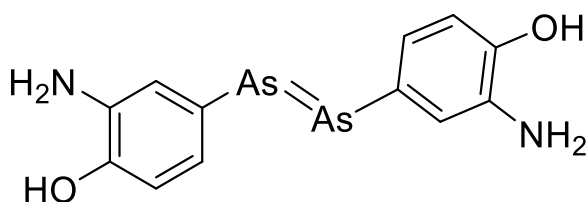


Figura 1.2. Estrutura química do Salvarsan.

O Salvarsan foi substituído pela penicilina, a qual foi descoberta em 1928 a partir do fungo do género *Penicillium* por Alexander Fleming.¹⁴ Em 1945, foi determinada a estrutura β-lactâmica da penicilina (Figura 1.3) por Dorothy Hodgkin,¹⁴ e em 1957 foi isolado o ácido 6-aminopenicilâmico (6-APA, do inglês *6-aminopenicillanic acid*) por Beechams.¹⁶ Estas descobertas permitiram o desenvolvimento de derivados semissintéticos da penicilina (Figura 1.4) para contornar a resistência à penicilina (introdução de grupos volumosos na cadeia lateral, como na meticilina) e sintetizar penicilinas de amplo espectro de

ação (introdução de grupos hidrofílicos no carbono α da cadeia lateral, como na ampicilina) (Figura 1.5).^{14,17}

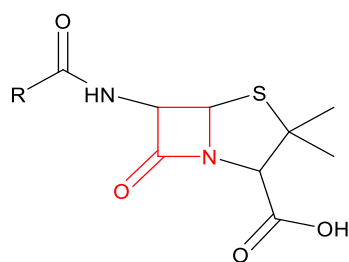


Figura 1.3. Estrutura geral das penicilinas, onde é destacado a vermelho a estrutura do anel β -lactâmico.

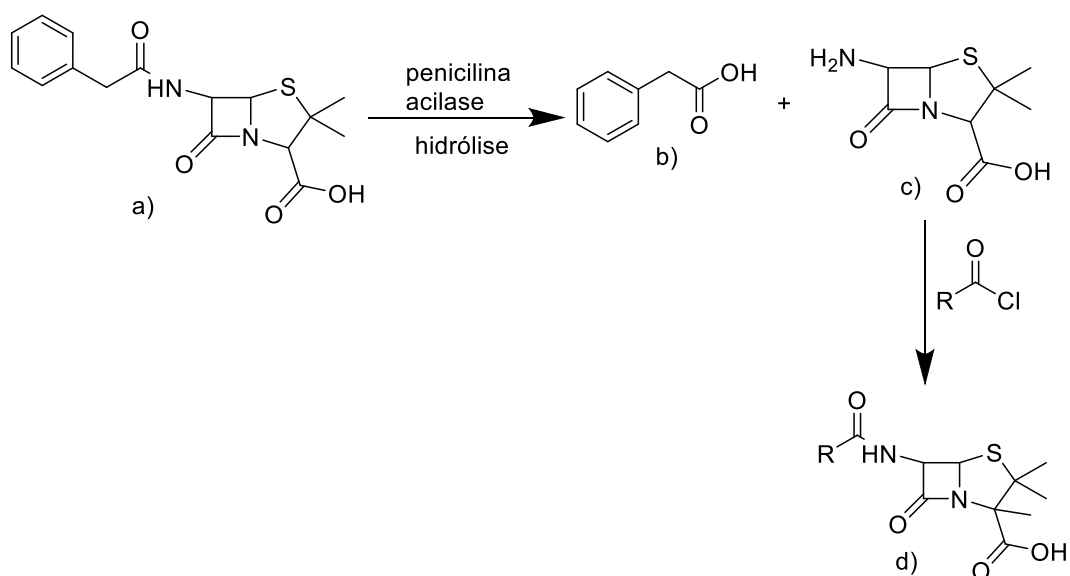


Figura 1.4. Síntese de derivados semissintéticos da penicilina. A síntese começa com a hidrólise da benzilpenicilina (a) para formar o ácido 2- fenilacético (b) e o 6-APA (c). Posteriormente, faz-se reagir o 6-APA com a cadeia lateral para formar um derivado da penicilina (d).

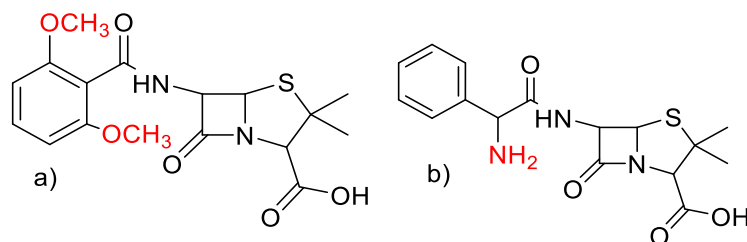


Figura 1.5. Estruturas químicas da metilicina (a), onde se destacam os grupos metoxilo que criam impedimento estérico para resistir à penicilinase, e da ampicilina (b), onde se destaca o grupo amino na cadeia lateral, que aumenta a hidrofílicidade da molécula de modo a favorecer um aumento na atividade contra bactérias Gram- negativas.

Em 1935, Gerhard Domagk descobriu o Prontosil (sulfamidocrisoidina) que era eficaz *in vivo* contra infecções provocadas por *Streptococcus*. Este foi reconhecido como pró-fármaco (fármaco administrado em forma inativa, que é ativado durante a sua metabolização) da sulfanilamida (*p*-aminofenilsulfonamida) (Figura 1.6).^{15,18,19}

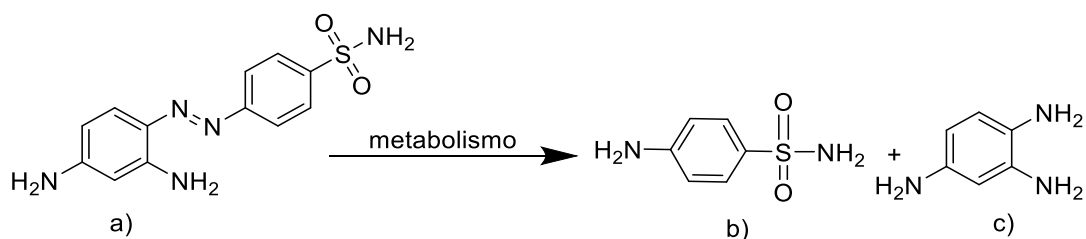


Figura 1.6. Metabolismo do prontossil (a) em sulfanilamida (b) e no metabolito inativo triaminobenzeno (c).

As sulfonamidas foram os primeiros antibióticos eficazes a serem produzidos num laboratório farmacêutico, e a sua introdução na prática clínica teve um impacto positivo, reduzindo em 24 % a 36 % da mortalidade materna, em 17 % a 32 % da mortalidade por pneumonia e em 52 % a 52 % da mortalidade por escarlatina.²⁰

No final da década de 1930 e início da década de 1940, foi descoberto que várias estirpes da bactéria do solo *Brevibacillus brevis* produziam substâncias lineares e cíclicas, usando proteínas sintetases não-ribossômicas, que inibiam bactérias e fungos patogênicos.⁶ Em 1939, o microbiólogo americano René Dubos conseguiu isolar pela primeira vez desta bactéria a tirotricina, que é uma mistura de AMPs cíclicos e lineares, sendo a gramicidina (Figura 1.7) e a tirocidina (Figura 1.8) os seus principais componentes.^{6,21,22} Vários AMPs foram descobertos desde então tanto em procarionotas como em eucariotas.^{21,23}

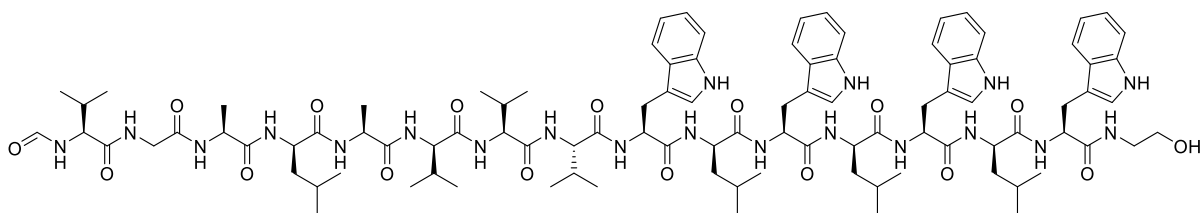


Figura 1.7. Estrutura química da gramicidina (formil-L-Val-Gly-L-Ala-D-Leu-L-Ala-D-Val-L-Val-d-Val-L-Trp-D-Leu-L-trp-D-Leu-L-Trp-D-Leu-L-Trp-etanolamina).

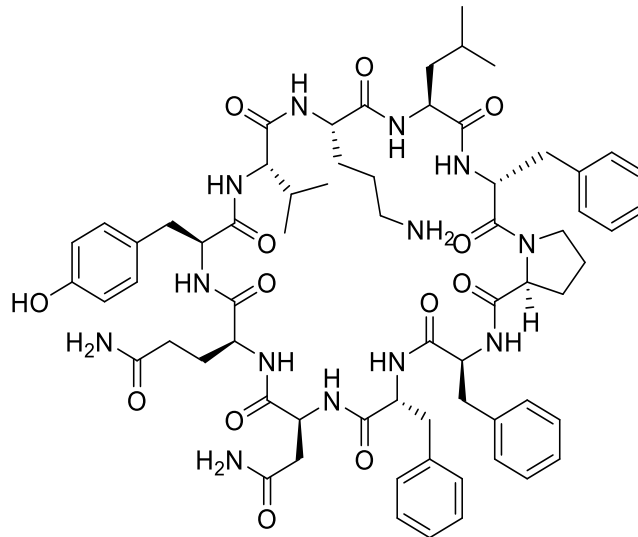


Figura 1.8. Estrutura química da tirocidina (ciclo [Asn-Gln-tyr-Val-Orn-Leu-D-phe-Pro-Phe-D-Phe]).

Os AMPs são definidos como oligopéptidos cíclicos ou lineares com um número variável de aminoácidos (geralmente menos de 100), com atividades imuno-moduladoras e antimicrobianas de amplo espectro contra bactérias, vírus, fungos e parasitas.^{23,24} Geralmente, os AMPs possuem resíduos hidrofóbicos e hidrofílicos, fornecendo-lhes uma estrutura anfipática, e apresentam uma carga total positiva devido à presença de resíduos aminoácídicos básicos na suas estruturas.^{24,25}

1.1.3. Resistência aos antimicrobianos

Globalmente, as infecções causadas por bactérias multirresistentes (MR) estão a aumentar, reduzindo, deste modo, a eficácia dos antimicrobianos disponíveis.^{26,27} Define-se a resistência antimicrobiana (RAM) como a capacidade desenvolvida pelos microrganismos de evadir a ação dos antimicrobianos desenhados para matá-los, e ela ocorre à medida que os microrganismos se adaptam à presença contínua dos antimicrobianos.²⁸ A RAM é um fenómeno natural, sendo inevitável que se desenvolva para todos os antimicrobianos em algum momento, mas o uso indevido e excessivo destes acelera o seu desenvolvimento.^{28,29}

A RAM é uma grande ameaça global para humanos, animais, plantas e o meio ambiente.^{26,27,30} O relatório do sistema global de vigilância de resistência e uso dos antimicrobianos da Organização Mundial da Saúde (OMS) confirma que a RAM está a aumentar, especificamente em países de baixo e médio rendimento, causando morbidade e mortalidade significativas.²⁹ De acordo com O'Neill (2016), a RAM causa anualmente 700000 mortes e prevê-se um aumento contínuo até 2050, se não for controlada, altura em que se preveem 10 milhões de mortes anuais associadas à RAM e uma redução de 2 % a

3,5 % no Produto Interno Bruto (PIB) mundial, o que custaria ao mundo até 100 trilhões de dólares.³¹ Segundo estimativas feitas pelo Centro de Controle e Prevenção de Doenças (CDC, do inglês *Center for Disease Control and Prevention*), houve pelo menos 2 milhões de infecções com microrganismos resistentes a antimicrobianos nos Estados Unidos de América (EUA) em 2013, causando pelo menos 23000 mortes. Em 2019, estes números aumentaram para mais de 2,6 milhões de infecções resistentes a antimicrobianos que resultaram em cerca de 44000 mortes.²⁸ A Agência Europeia de Medicamentos (EMA, do inglês *European Medicines Agency*) estima que as infecções causadas por bactérias MR causem 33000 mortes na UE (União europeia) por ano.³² Estimativas feitas por Murray *et al*, indicam que, em 2019, cerca de 1,27 e 4,95 milhões de mortes foram diretamente atribuíveis ou associadas a infecções MR, respetivamente, sendo que a maioria foi reportada na África Subsaariana e na Ásia.³³

Pensa-se que a evolução da RAM esteja a ser agravada com a escassez de novos antimicrobianos. Isto possibilita que infecções intratáveis e MR se tornem cada vez mais comuns, visto que, após o seu aparecimento, a RAM é irreversível.³⁴ Por esta razão, projetar e produzir novos agentes antimicrobianos eficazes que limitem a disseminação de patógenos resistentes a antimicrobianos é uma das medidas para controlar a RAM.²⁹ Neste contexto, a OMS apresentou uma lista de patógenos que necessitam urgentemente de novos antimicrobianos, que inclui: (i) prioridade crítica: *Acinetobacter baumannii* resistente a carbapenemas, *Pseudomonas aeruginosa* resistente a carbapenemas, e todas as bactérias pertencentes à família *Enterobacteriaceae* resistentes a carbapenemas e cefalosporinas da terceira; (ii) prioridade alta: *Staphylococcus aureus* resistente a vancomicina e meticilina, *Campylobacter spp* resistente a fluoroquinolonas, *Neisseria gonorrhoeae* resistente a cefalosporinas da terceira geração e fluoroquinolonas, *Enterococcus faecium* resistente a vancomicina, *Helicobacter pylori* resistente a claritromicina, e *Salmonella spp* resistente a fluoroquinolonas; e (iii) prioridade média: *Streptococcus pneumoniae* resistente a penicilinas, *Haemophilus influenzae* resistente a ampicilina, e *Shigella spp* resistente a fluoroquinolonas.³⁵

1.1.4. Polimixinas

As polimixinas são AMPs cíclicos, de origem natural produzidas pela bactéria Gram-positiva *Paenibacillus polymyxa*, descobertas em 1947 e introduzidas na prática clínica para tratar infecções causadas pelas bactérias Gram-negativas na década de 1950.^{36,37} As polimixinas incluem cinco compostos quimicamente distintos (polimixinas A, B, C, D e E), sendo as polimixinas B e E (esta última também conhecida por colistina) as duas que têm sido utilizadas na prática clínica.^{36,37} A principal entre estes dois fármacos consiste na presença de L-fenilalanina (polimixina B) ou D-leucina (polimixina E) na 6ª posição

da cadeia de aminoácidos, assim como no ácido gordo ligado ao terminal N da cadeia peptídica (6-metiloctanolilo e 5-metileptanoílo), respectivamente (Figura 1.9).

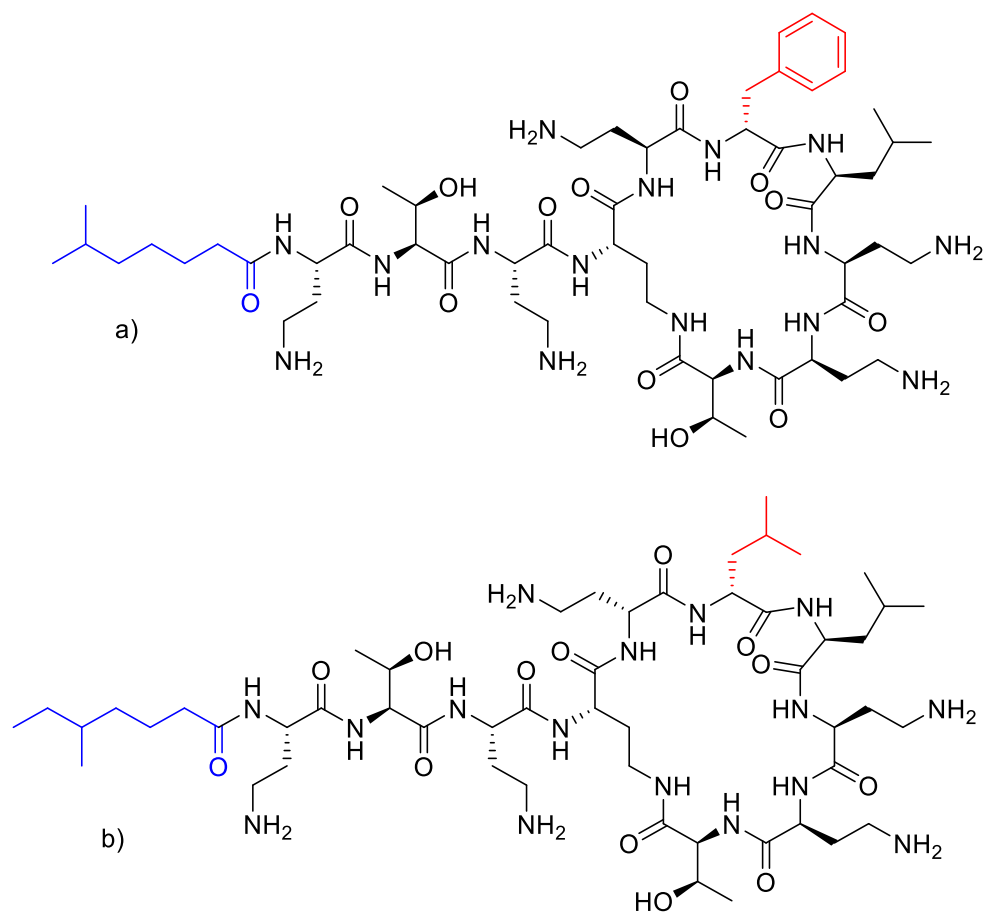


Figura 1.9. Estruturas químicas da polimixina B (a), onde se destaca a vermelho a cadeia lateral benzil do aminoácido fenilalanina e a azul o ácido gordo 6-metiloctanolilo ligado ao terminal N, e da colistina (b), onde se destaca a vermelho a cadeia lateral isopropil do aminoácido leucina da 6ª posição e a azul o ácido gordo 5-metileptanoílo ligado ao terminal N.

Uma vez que os antimicrobianos disponíveis atualmente, incluindo as penicilinas de amplo espectro, carbapenemas, cefalosporinas, fluoroquinolonas, monobactamas e aminoglicosídeos, muitas vezes não conseguem tratar infecções causadas por bactérias MR, e dada a ausência de novos antimicrobianos no mercado, as polimixinas (entretanto retiradas da clínica devido à sua toxicidade) foram reintroduzidas como tratamento de último recurso.³⁷ A maioria das bactérias Gram-negativas são sensíveis às polimixinas, incluindo *Acinetobacter baumannii* MR, *Pseudomonas aeruginosa*, *Klebsiella spp.*, *Enterobacter spp.*, *Escherichia coli*, *Salmonella spp.*, *Shigella spp.*, *Citrobacter spp.*, *Yersinia pseudotuberculosis*, e *Haemophilus spp.* Também foi relatado que a colistina é potencialmente ativa

contra várias espécies de micobactérias, incluindo *M. xenopi*, *M. intracellulare*, *M. tuberculosis*, *M. fortuitum*, *M. phlei* e *M. smegmatis*.³⁸⁻⁴¹

1.1.4.1. Mecanismo de ação das polimixinas

A polimixina B e a colistina partilham o mesmo mecanismo de ação, que se baseia em interações electrostáticas. Estas interações ocorrem devido às moléculas de ácido L-diaminobutírico (L-Dab) das polimixinas carregadas positivamente e aos lipopolissacarídeos (LPS) presentes na membrana externa das bactérias Gram-negativas, os quais têm cargas negativas. Esta interação desloca os catiões Mg^{2+} e Ca^{2+} produzindo um efeito disruptivo que leva a alterações na permeabilidade da membrana celular e, por fim, à morte da bactéria.^{38-40,42}

Para além do mecanismo acima exposto, a enzima respiratória NADH-quinona oxidoreductase do tipo II (NDH-2) da membrana bacteriana interna é um alvo secundário das polimixinas. Esta enzima faz parte da cadeia transportadora de eletrões e estudos mostram que as polimixinas inibem esta enzima em bactérias Gram positivas, bactérias Gram-negativas e fungos.^{38,43}

1.1.4.2. Desvantagens das polimixinas

As polimixinas são antimicrobianos valiosos para o tratamento de infeções causadas por bactérias Gram-negativas MR, mas o seu valor clínico é limitado devido à sua alta nefrotoxicidade e neurotoxicidade, bem como à baixa permeabilidade e absorção no trato gastrointestinal.⁴⁴⁻⁴⁶ A polimixina B tem sido tradicionalmente considerada mais tóxica do que a colistina.⁴⁷⁻⁴⁹ Vardakas e Falagas, numa metanálise, constataram que quatro estudos relataram uma taxa de mortalidade por nefrotoxicidade entre 8 % e 56 % em pacientes tratados com colistina e de 31 % a 61 % em pacientes tratados com polimixina B.⁵⁰ Um outro estudo realizado por Ouderkerk *et al.* em 2003 verificou uma mortalidade de 20 % (n= 60) em indivíduos tratados com a polimixina B.⁵¹ Zeng *et al.* verificaram uma taxa de nefrotoxicidade de 31,6 % (n= 107) em indivíduos tratados com a polimixina B, dos quais 20,1 % mostraram toxicidade precoce (3 dias após a administração do fármaco).⁵² Outro estudo, realizado por Montero *et al.*, revelou que apenas 8,3 % (n= 121) dos pacientes tratados com colistina que sofreram de nefrotoxicidade insuficiência renal crónica prévia, diabetes *mellitus* e/ou uso de aminoglicosídeos.⁵³

As polimixinas causam nefrotoxicidade por um mecanismo de apoptose celular induzida por fármacos. A reabsorção das polimixinas ocorre no túbulo proximal dos rins e, nesse processo, há acumulação intracelular substancial do fármaco que é mediado pelo recetor endocítico megalina (para

além de outros transportadores), resultando na alta concentração intracelular das polimixinas que leva à apoptose celular, diminuindo, deste modo, a função renal.^{44,54} Um modelo alternativo para o mecanismo da nefrotoxicidade das polimixinas baseia-se na possibilidade das polimixinas induzirem danos no DNA, levando à segregação cromossômica e instabilidade do genoma.⁵⁵

Um outro problema é o aumento de espécies bacterianas resistentes às polimixinas. Isto é considerado um problema sério, gerando preocupação devido ao baixo número de antimicrobianos eficazes atualmente disponíveis.^{56,57} Algumas bactérias, como *Klebsiella pneumoniae*, *Pseudomonas aeruginosa* e *Acinetobacter baumannii*, desenvolvem resistência às polimixinas num processo conhecido como resistência adquirida, enquanto outras bactérias, como a *Neisseria spp.*, a *Stenotrophomonas maltophilia*, a *Brucella spp.*, a *Proteus spp.*, a *Providencia spp.*, a *Morganella spp.*, a *Serratia spp.* e a *Burkholderia spp.*, são naturalmente resistentes a estes fármacos.^{56,57}

A maioria dos casos da resistência às polimixinas mostraram ser devidos ao gene *mcr-1* (resistência mobilizada à colistina, do inglês *mobilized colistin resistance*) mediada por plasmídeo.⁵⁸ A MCR-1 é uma enzima da família da fosfoetanolamina transferase que faz a adição de fosfoetanolamina ao lípido A, levando à modificação da estrutura membranar do microrganismo, o que resulta na redução da afinidade com as polimixinas.^{41,58} Encontram-se descritos casos de disseminação de várias espécies de *Enterobacteriaceae* positivas para MCR-1 na África, Ásia, Europa, América do Sul e América do Norte.^{41,59} Lu *et al.* identificaram o gene *mcr-1* em 21% de *E. coli* isolada em animais e 1% de *E. coli* isolada de amostras de pacientes.⁵⁸ Num outro estudo realizado por Ohene *et al.*, foi verificada a resistência fenotípica à colistina em 8% de *E. coli* e *Enterobacter spp.*, isoladas do porco no Gana.⁶⁰

Portanto, melhorar as propriedades das polimixinas de forma a que apresentem maior eficácia e segurança será de grande interesse para o desenvolvimento de novos fármacos antimicrobianos.⁴⁵

1.2. Métodos de pesquisa de novos fármacos

O desenvolvimento de novos agentes antimicrobianos é uma das medidas traçadas pela OMS para combater a RAM.²⁹ Para esse fim, é importante selecionar candidatos a fármacos, que exibam uma série de propriedades desejadas, particularmente em relação à biodisponibilidade, bioatividade e toxicidade.⁶¹ Tradicionalmente, tanto a identificação como a otimização de novos fármacos são realizadas através de triagem experimental. Apesar de vários esforços para melhorar a sua eficiência, este continua a ser um processo caro e demorado.⁶² Nas últimas décadas, a química computacional tem vindo a

consolidar a sua posição como uma ferramenta útil para o processo de descoberta e *design* de novos fármacos.⁶² Atualmente, a simulação computacional e estudos experimentais complementam-se para encontrar moléculas que possam ser usadas como fármacos de forma rápida e eficaz.^{63,64}

As estratégias usadas no *design* de fármacos em química computacional podem ser divididas em dois grandes grupos: (i) *design* de fármacos baseado na estrutura do alvo terapêutico, e (ii) *design* de fármacos baseado na estrutura dos ligandos. A primeira abordagem aplica-se quando se dispõe de informações sobre a estrutura do alvo bioquímico da ação dos compostos bioativos, em que se usam os métodos baseados em modelos físicos da matéria, como as simulações de dinâmica molecular, e os métodos de *docking* molecular. Por outro lado, a segunda abordagem pode ser considerada em casos para os quais se dispõe de informações sobre a estrutura dos compostos bioativos e a sua respetiva atividade biológica. Esta estratégia baseia-se em métodos de análise de dados chamados QSAR (Relação Quantitativa Estrutura-Atividade, do inglês *Quantitative Structure-Activity Relationship*), que relaciona parâmetros calculados de uma série de compostos conhecidos com as atividades biológicas determinadas experimentalmente.^{65,66} Atualmente, enormes bases de dados podem ser pesquisadas num tempo relativamente curto e a um custo muito baixo, resultando numa economia de tempo, dinheiro e recursos humanos, pelo que os estudos QSAR tornaram-se numa ferramenta importante para a descoberta de novos fármacos.^{63,67-70}

Mais recentemente, os algoritmos da inteligência artificial (IA) têm vindo a ganhar popularidade para identificar novos compostos com atividades promissoras contra um determinado alvo. O termo IA foi cunhado por John McCarthy em 1956 para descrever a ciência e a engenharia de fazer máquinas inteligentes.⁷¹ A IA envolve várias ferramentas e algoritmos que visam imitar a inteligência humana^{72,73} e já é hoje aplicada com sucesso numa ampla gama de domínios particularmente desafiantes, tais como a robótica, o processamento de linguagem natural, a análise de imagens, otimização de processos de logística, assim como no *design* molecular.⁷⁴ Dentro da IA, a ML é uma subdisciplina que usa algoritmos capazes de reconhecer padrões dentro de um conjunto de dados.^{72,73} A IA e a ML floresceram nas últimas décadas, impulsionadas por avanços revolucionários na tecnologia computacional. Isso levou a melhorias consideráveis na capacidade da colheita e processamento de grandes volumes de dados.^{72,74} Devido à sua natureza automatizada e capacidades preditivas, as técnicas de ML são atraentes para a indústria farmacêutica, para avaliar fármacos quanto às suas potenciais atividades biológicas e toxicidades, com vantagens a nível da eficiência do processo de desenvolvimento de novos fármacos.^{72,75}

O objetivo final da aplicação dos métodos da IA na descoberta de fármacos é trazer os melhores fármacos para a clínica, de modo a satisfazer necessidades médicas não atendidas. No contexto da descoberta de fármacos, a IA envolve tarefas na identificação do alvo do fármaco, identificação de principais compostos, otimização do protótipo em relação a vários perfis de propriedades de interesse e identificação dos mecanismos de síntese.⁷⁴

1.3. Modelos QSAR

A hipótese geral dos modelos QSAR é que as mudanças na estrutura molecular são proporcionais às mudanças da sua atividade, e que é possível obter um modelo que relaciona descritores moleculares (também conhecidos como preditores) e a sua atividade biológica, usando métodos estatísticos.^{76,77}

O progresso dos métodos QSAR obteve um sucesso notável em vários campos, tais como a química medicinal, a ciência de materiais e a toxicologia.⁷⁸ Speck-Planche *et al*/desenvolveram um modelo QSTR (Relação Quantitativa Estrutura-Toxicidade, do inglês, *Quantitative Structure-Toxicity Relationship*) para prever os perfis ecotoxicológicos de fungicidas agroquímicos contra 20 espécies aquáticas usando os descritores de grupos funcionais.⁷⁹ Milicevic e Sinko desenvolveram um modelo QSAR para prever a atividade das oximas, N-hidroxiiminoacetamidas, 4-aminoquinolinas e flavonoides na inibição da enzima acetilcolinesterase usando descritores topológicos, especificamente os índices de conectividade molecular de valência de ordem zero ($^0\chi$).⁸⁰ Roy e Pandey desenvolveram o modelo QSPR (Relação Quantitativa Estrutura-Propriedade, do inglês *Quantitative Structure-Property Relationship*) para o coeficiente de partição octanol/água (para reconhecer as características que podem aumentar ou diminuir o coeficiente de partição de classes químicas) e para o coeficiente de sorção (teor do carbono orgânico do solo), usando os índices topológicos simples.⁸¹

1.3.1. QSAR baseado em modelos lineares

O uso de modelos QSAR foi introduzido por Corwin Hansch em 1964. A intenção inicial não era prever análogos bioativos mais potentes, mas aumentar a compreensão dos mecanismos químicos e/ou bioquímicos de interação entre fármacos e recetores.^{82,83} Para este fim, usaram-se modelos de regressão linear aplicados a um conjunto de pequenas moléculas que compartilham atividade biológica semelhante.⁷⁶ Para este fim, usaram-se três tipos de descritores: propriedades eletrônicas de Hammett (σ), propriedades hidrofóbicas ($\log P$ e hidrofobicidade do substituinte (π)) e propriedades estéricas de Taft (E_s).^{82,84}

Tradicionalmente, a relação quantitativa entre estrutura e atividade biológica pode ser determinada usando as equações de Hansch, que geralmente relacionam a atividade biológica com as propriedades físico-químicas mais comumente usadas ($\log P$, π , σ e E_s).⁸⁴ Usa-se a equação 1.1 se $\log P$ for limitado a uma pequena faixa de valores e a equação 1.2 se os valores de $\log P$ estiverem mais dispersos.

$$\log \left(\frac{1}{C} \right) = K_1\pi + K_2\sigma + K_3E_s + K_4 \quad (1.1)$$

$$\log \left(\frac{1}{C} \right) = K_1\pi^2 + K_2\pi + K_3\sigma + K_4E_s + K_5 \quad (1.2)$$

onde C é a concentração à qual um composto que produz uma dada resposta biológica, π é a constante de hidrofobicidade do substituinte, σ é a constante de Hammett (efeitos eletrônicos do substituinte) e E_s é a constante de Taft (efeitos estéricos do substituinte). Os coeficientes k_1 a k_5 são determinados usando o método dos mínimos quadrados.

Outro método usado no QSAR tradicional é o de Free e Wilson, que é baseado na suposição de que cada substituinte faz uma contribuição aditiva e constante para a atividade biológica, independentemente da variação do substituinte no resto da molécula (equação 1.3), e o peso das contribuições individuais do grupo é calculado por análise de regressão.^{84,85} Nesta abordagem, a atividade biológica do composto original é medida e comparada com a atividade de uma variedade de análogos substituídos. A análise de Free-Wilson usa métodos de regressão estatística para gerar modelos que são baseados em fragmentos ou geração de contribuição dos substituintes. Estes modelos baseados nas contribuições de fragmentos permitem uma interpretação do modelo no mesmo esquema que as SAR (Relação Estrutura-Atividade, do inglês *Structure-Activity Relationship*) para identificar os farmacóforos.^{84,86}

$$\log \left(\frac{1}{C} \right) = \sum_i a_i + \mu \quad (1.3)$$

onde a_i é a contribuição do grupo substituinte X_i , baseado em $a_H = 0$, e μ é a atividade biológica do composto substituído.

Um método de modelação QSAR tradicional alternativo e mais interpretável foi introduzido por John Topliss com base na análise de Hansch, sendo geralmente empregado para guiar a escolha de substituintes mais potentes.^{87,88} Os esquemas de Topliss são baseados numa suposição fundamental do

método de Hansch de que um substituinte particular pode modificar a atividade em relação ao composto original em virtude das mudanças resultantes nos efeitos hidrofóbicos, eletrônicos e estéricos.^{84,86,88} O método é estruturado como uma árvore de decisão de Topliss e os análogos são sintetizados ao longo desta árvore, considerando a mudança na atividade biológica para sugerir o próximo grupo de substituinte a ser explorado, como ilustrado na Figura 1.10.^{87,89}

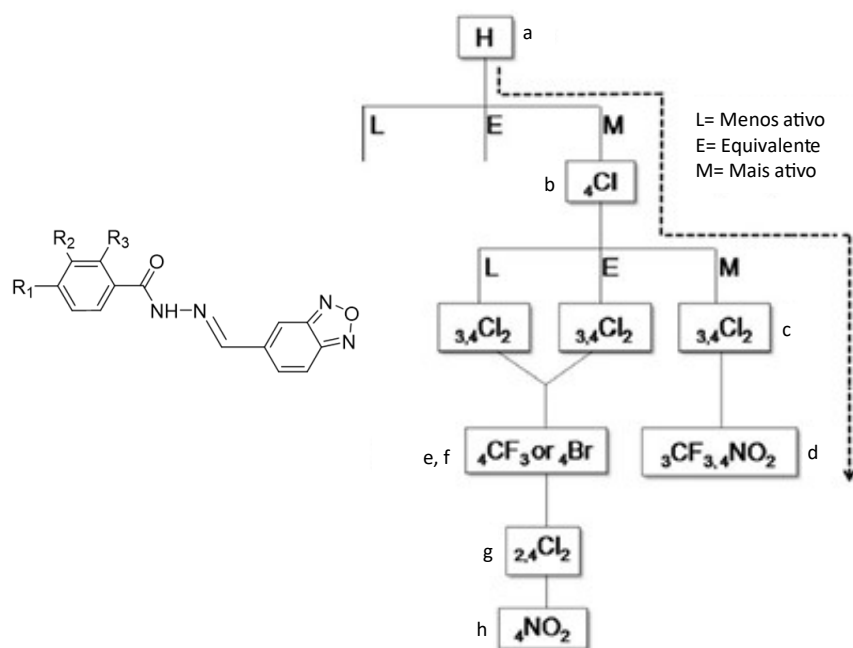


Figura 1.10. Esquema de Topliss aplicado à síntese de agentes antimicrobianos derivados de benzofuroxano (adaptado de Jorge, SD et al.).⁸⁷

Ao iniciar um estudo QSAR, é importante decidir que parâmetros físico-químicos vão ser estudados e projetar os análogos de modo que os parâmetros em estudo sejam adequadamente variados. Também é importante criar estruturas suficientes para tornar os resultados estatisticamente significativos. Como regra geral, pelo menos cinco estruturas devem ser feitas para cada parâmetro estudado. Normalmente, um estudo QSAR inicial envolve dois parâmetros, π e σ , e, possivelmente, E_s . O diagrama de Craig (Figura 1.11) pode ser usado para escolher substituintes adequados.⁸⁴

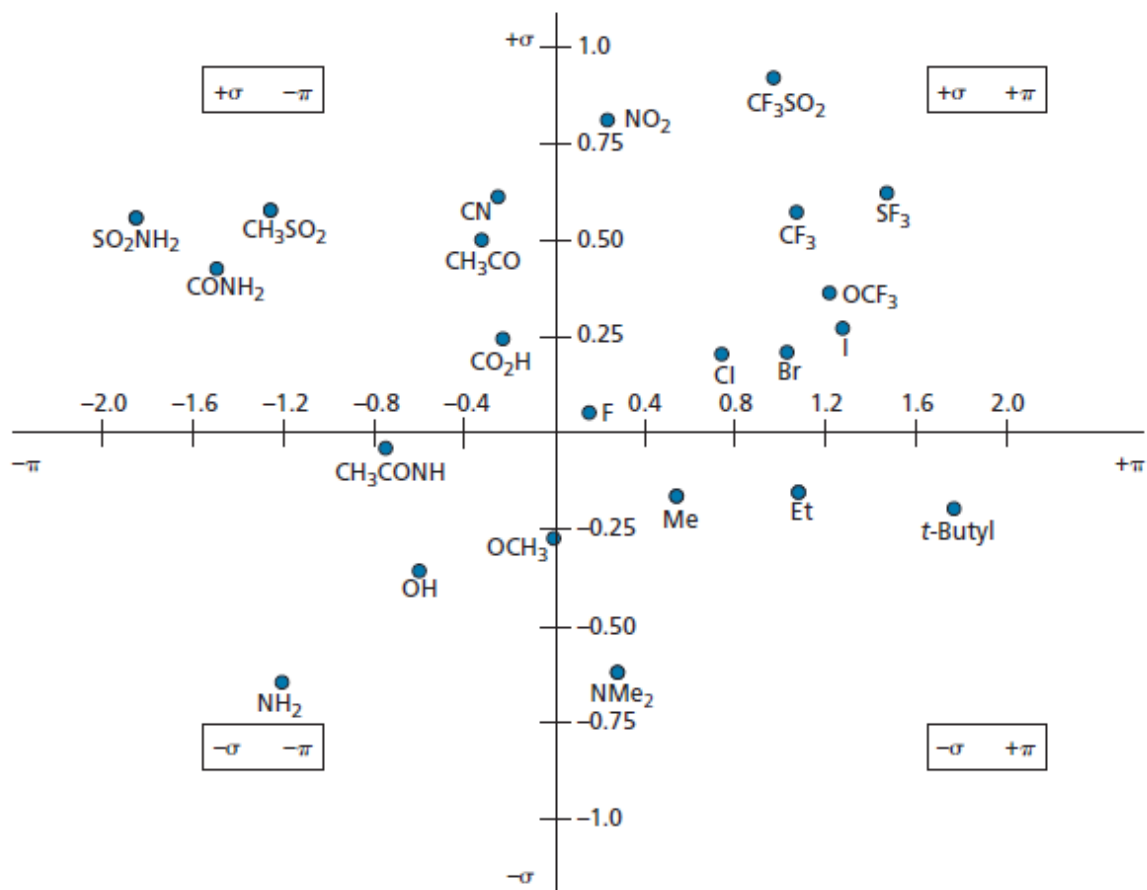


Figura 1.11. Diagrama de Craig (retirado de Craig, PN).⁹⁰

1.3.2. QSAR baseado em métodos ML

Atualmente, os métodos QSAR evoluíram da análise de regressão simples para técnicas de ML.⁹¹ A construção de modelos QSAR baseados em ML consiste em quatro etapas principais envolvendo técnicas de quimioinformática (uso de recursos das ciências de computação para resolver problemas no campo da química): (i) codificação molecular, onde as características e propriedades químicas são derivadas das estruturas químicas ou pesquisa de resultados experimentais; (ii) seleção de variáveis, onde as técnicas de aprendizagem não supervisionada são usadas para identificar as propriedades mais relevantes e reduzir a dimensionalidade do vetor de recursos; (iii) treino do algoritmo, onde um método de ML supervisionado é aplicado para descobrir uma função empírica que pode alcançar um mapeamento ideal entre os vetores de recursos de entrada e as respostas biológicas; e (iv) validação, onde se analisa a robustez e a capacidade preditiva do modelo, sendo avaliadas métricas como precisão, exatidão, sensibilidade e especificidade.^{92,93}

1.3.3. Modelos de classificação em ML

Tradicionalmente, os métodos de ML dividem-se entre métodos supervisionados e não-supervisionados.^{73,94} Os métodos supervisionados são aqueles que tentam descobrir a relação entre atributos de entrada (variáveis independentes ou preditores) e um ou mais atributos de destino (variáveis dependentes ou alvos). Na aprendizagem supervisionada, o algoritmo observa pares de entrada-saída e aprende uma função que mapeia da entrada para a saída. Esses métodos podem ser subdivididos relativamente ao tipo de saída entre modelos de classificação (classificadores) e modelos de regressão.^{73,95} Os modelos de classificação têm como objetivo identificar a categoria à qual um objeto pertence. Por outro lado, os modelos de regressão têm como objetivo prever um valor interpolável associado a um dado conjunto de dados de entrada.⁹⁶

Os métodos de aprendizagem não supervisionada agrupam instâncias dos dados fornecidos sem um atributo dependente pré-especificado e o agente aprende padrões na entrada sem nenhum *feedback* explícito. Esses métodos incluem técnicas de redução de dimensionalidade, como a análise de componentes principais (PCA, do inglês *Principal Components Analysis*), a análise de componentes independentes (ICA, do inglês *Independent Components Analysis*), a alocação latente de Dirichlet (LDA do inglês *Latent Dirichlet Allocation*), entre outros.^{73,92,95}

1.3.3.1. Árvore de decisão

A árvore de decisão (DT, do inglês *Decision Tree*), desenvolvido por Breiman *et al*,⁹⁷ é um método supervisionado que pode ser usado tanto em problemas de classificação como de regressão. Este é um algoritmo hierárquico de decisões e suas consequências (Figura 1.12), em que o objetivo é criar um modelo que preveja o valor de uma variável de destino aprendendo regras de decisão simples inferidas a partir dos dados fornecidos.⁹⁸

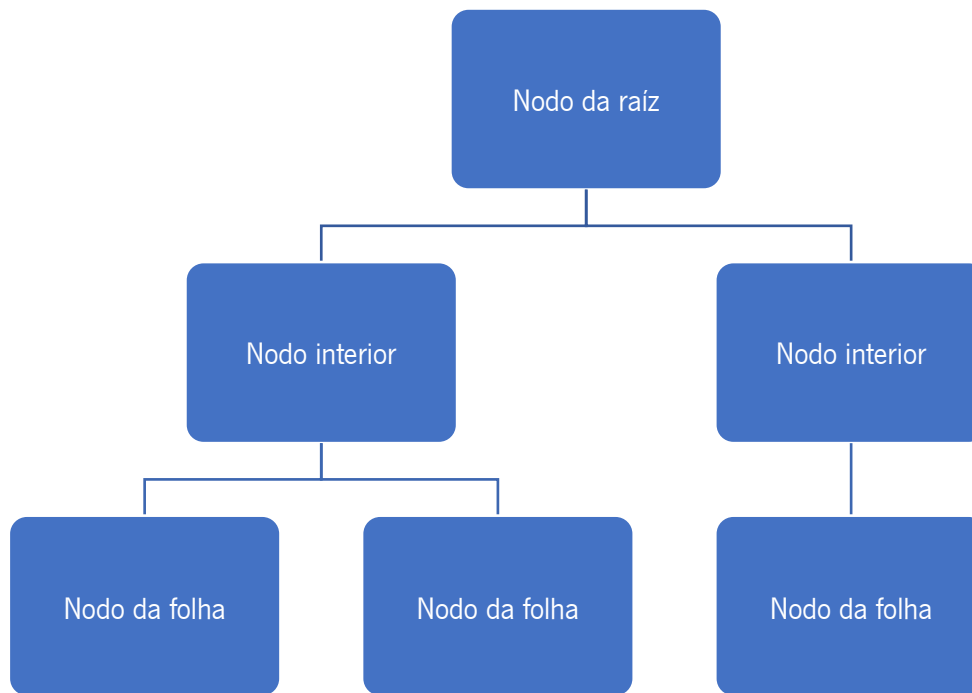


Figura 1.12. Estrutura de uma DT.

Uma DT chega à sua decisão realizando a seguinte sequência: (i) construir um nó raiz, que contem todos dados de treino; (ii) selecionar uma variável ideal e dividir os dados treinos em subconjuntos de acordo com a variável ideal selecionada, para que cada subconjunto possa alcançar a melhor classificação; (iii) construir os nós da folha para os subconjuntos que foram classificados corretamente; (iv) se houver subconjuntos classificados incorretamente, novas variáveis opcionais são selecionadas para continuarem a ser divididos, e os nós correspondentes continuam a ser construídos até que todos os subconjuntos treino sejam classificados corretamente ou até que não haja variáveis adequadas; (v) finalmente, cada subconjunto é atribuído a um nó principal. Cada nó interno da árvore corresponde a um teste do valor de um das variáveis de entrada e os nós das folhas especificam qual valor deve ser retornado pela função.^{73,95,99}

O índice de Gini (entropia cruzada) é frequentemente usado como critério para divisão de nós durante o treino de uma DT. A variável que resultaria na maior queda do índice de Gini é selecionada para divisão de todas as variáveis candidatas consideradas para a divisão, um de cada vez. A divisão binária divide sucessivamente cada um dos dois nós filhos produzidos pela divisão anterior, até que algum critério de paragem seja atendido.^{97,100,101}

1.3.3.2. Floresta aleatória

Segundo Breiman, “a floresta aleatória (RF, do inglês *Random Forest*) é uma combinação de preditores de árvores, de modo que cada árvore depende dos valores de um vetor aleatório amostrado independentemente e com a mesma distribuição para todas as árvores da floresta”.¹⁰² A RF é um método de conjunto, cujo objetivo é combinar as previsões de várias DTs a fim de melhorar a generalização ou robustez que seria obtida usando apenas um estimador.⁹⁶ Numa RF, cada árvore do conjunto é construída a partir de uma amostra extraída com reposição do conjunto de treino. Ao dividir cada nodo durante a construção de uma árvore, a melhor divisão é encontrada de entre todas as variáveis independentes (ou de entre um subconjunto aleatório das mesmas).^{96,102,103} A RF é robusta contra *overfitting* (quando um modelo se ajusta muito bem ao conjunto treino, mas se mostra ineficaz para prever novos resultados).¹⁰³

O algoritmo da RF (Figura 1.13) é o seguinte: (i) colheita de um conjunto com $n_{\text{árvore}}$ amostragens dos dados originais; (ii) para cada uma das amostras do conjunto treino de uma árvore de classificação ou regressão, escolha da melhor divisão entre todos os preditores; e (iii) revisão da resposta a novos dados, agregando as previsões das árvores $n_{\text{árvore}}$ (maioria de votos para classificação, média para regressão).¹⁰³

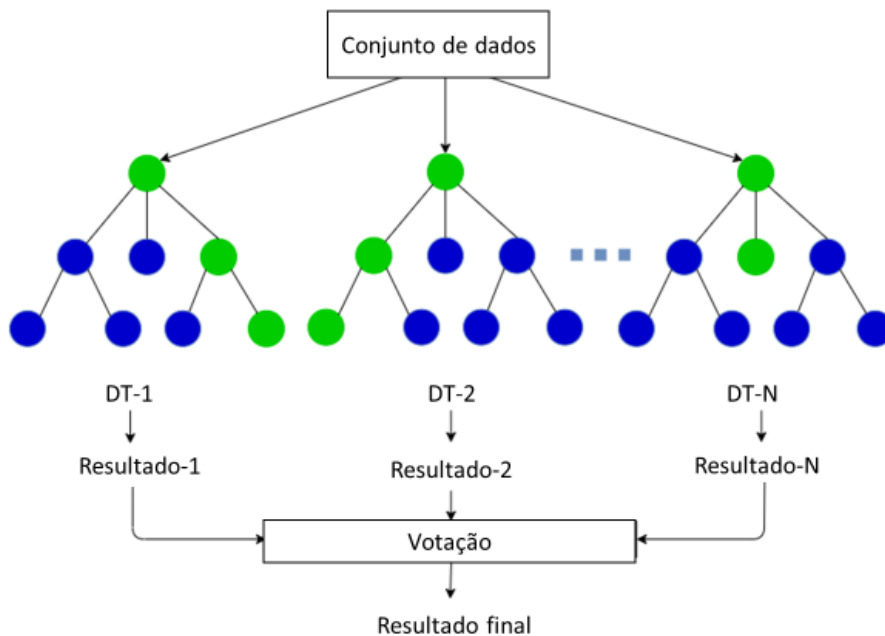


Figura 1.13. Esquema do algoritmo da RF (adaptado de Awasthi, S).¹⁰⁴

1.3.3.3. AdaBoost

O AdaBoost (reforço adaptável, do inglês *Adaptive Boosting*) é um meta-algoritmo de classificação estatística formulado por Yoav Freund e Robert Schapire em 1995.¹⁰⁵ O algoritmo AdaBoost é um procedimento iterativo que tenta aproximar o classificador de Bayes (que minimiza a probabilidade de erro de classificação) combinando vários classificadores fracos, conforme ilustrado na Figura 1.14.¹⁰⁶ O princípio central do AdaBoost é ajustar uma sequência de algoritmos fracos (ou seja, métodos que são apenas ligeiramente melhores do que as adivinhações aleatórias, como pequenas DTs) em versões repetidamente modificadas dos dados.⁹⁶ As previsões dos algoritmos fracos são combinadas por meio de uma votação majoritária ponderada (ou soma ponderada, nos problemas de regressão) para produzir a previsão final. As modificações de iteração do reforço consistem na aplicação de pesos (W_1, W_2, \dots, W_n) para cada uma das amostras de treino. Inicialmente, esses pesos são todos definidos para $W_i=1/N$, de modo que, na primeira etapa, simplesmente treine um algoritmo fraco nos dados originais. Para cada iteração sucessiva, os pesos da amostra são modificados individualmente e o algoritmo do ML é reaplicado aos dados reponderados. Os dados de treino preditos incorretamente pelo modelo na etapa anterior têm os seus pesos aumentados, ao passo que os pesos são diminuídos para aqueles que foram preditos corretamente.⁹⁶

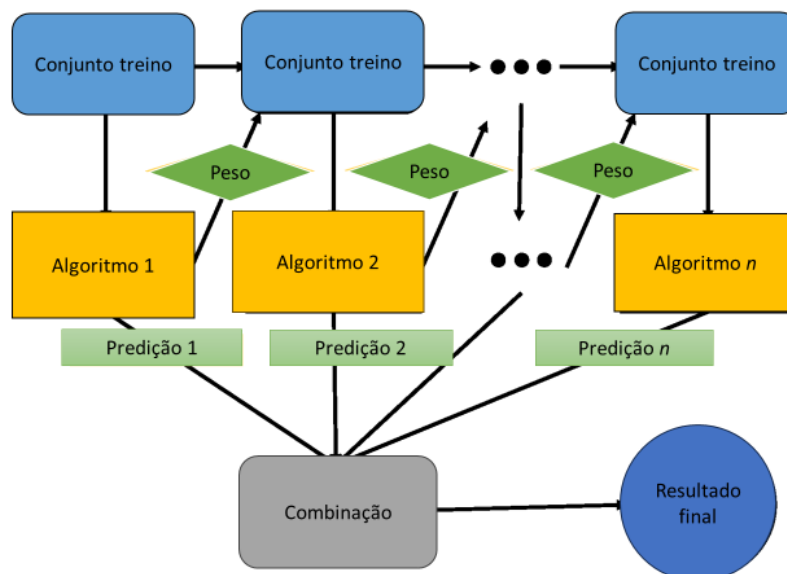


Figura 1.14. Esquema de um modelo AdaBoost (adaptado de Wang, C et al.).⁹⁹

1.3.4. Validação cruzada de modelos

O método de validação cruzada foi proposto por Stone em 1974 e é definido como uma técnica computacional intensiva que usa todos os dados disponíveis como dados de treino e teste.^{107,108} Numa validação cruzada k-fold, o conjunto de dados original é particionado aleatoriamente em k subconjuntos de tamanhos aproximadamente iguais. A cada uma das k iterações da validação cruzada, $k-1$ partes são usadas para treinar o modelo e a parte restante é usada como conjunto de teste. Findas as k iterações, a medida do desempenho é calculada com base na distribuição da métrica de desempenho ao longo dos k conjuntos de teste.¹⁰⁹

1.3.5. Métricas de avaliação de modelos

O desempenho do modelo de classificação é avaliado por valores escalares usando diferentes métricas, como precisão, exatidão, sensibilidade e especificidade.¹¹⁰

1.3.5.1. Exatidão

A exatidão, ou seja, a fração de previsões corretas, é definida como uma razão entre as amostras classificadas corretamente e o número total de amostras (equação 1.4),

$$Exatidão = \frac{VP + VN}{VP + VN + FP + FN} \quad (1.4)$$

onde VP são os verdadeiros positivos, VN são os verdadeiros negativos, FN são os falsos negativos e FP são os falsos positivos.¹¹¹

1.3.5.2. Sensibilidade e Especificidade

A sensibilidade é a medida da proporção dos positivos reais que foram corretamente previstos (equação 1.5), ao passo que a especificidade é a medida dos negativos reais que foram corretamente previstos (equação 1.6).^{86,92,112}

$$Sensibilidade = \frac{VP}{(VP + FN)} \quad (1.5)$$

$$Especificidade = \frac{VN}{VN + FP} \quad (1.6)$$

Numa classificação binária, um modelo perfeito teria 100 % de sensibilidade e 100 % de especificidade, o que significa que todos os ativos e inativos são previstos corretamente.^{86,92}

1.3.5.3. Precisão

A precisão representa a proporção de amostras positivas que foram classificadas corretamente em relação ao número total de amostras positivas previstas (equação 1.7).¹¹⁰ Intuitivamente, a precisão é a capacidade do classificador de não rotular como positiva uma amostra negativa.¹¹¹

$$Precisão = \frac{VP}{VP - FP} \quad (1.7)$$

1.3.5.4. Matriz de confusão

A matriz de confusão, ou matriz de erro, é um dos métodos mais simples de resumir a classificação dos objetos nas suas classes observadas usando um modelo preditivo. A matriz de confusão permite uma análise confiável e detalhada do poder preditivo de um modelo classificador.⁸⁶ Os exemplos de matrizes de confusão são apresentados na Tabela 1.1, para a classificação binária, e na Figura 1.15, para a classificação multiclasse.

Tabela 1.1. Exemplo da matriz de confusão da classificação binária.

		Real	
		Verdade	Falso
Previsão	Verdade	VP	FP
	Falso	FN	VN

	A	B	C	
A	VP _A	E _{BA}	E _{CA}	Previsão
B	E _{AB}	VP _B	E _{CB}	
C	E _{AC}	E _{BC}	VP _C	
	Real			

Figura 1.15. Matriz de confusão para um problema de classificação multiclasse, com três classes (A, B e C), onde VP_A é o número de amostras corretamente classificadas na classe A e E_{AB} são as amostras da classe A que foram classificadas incorretamente como da classe B. Assim, o falso negativo na classe A (FN_A) é a soma de E_{AB} e E_{AC} (FN_A = E_{AB} + E_{AC}) que indica a soma de todas as amostras da classe A que foram incorretamente classificadas como sendo da classe B ou C. O falso positivo na classe A (FP_A) é a soma de E_{BA} e E_{CA} (FP_A = E_{BA} + E_{CA}), que corresponde a amostras das classes B e C classificadas como sendo da classe A.^{110,113}

1.4. Descritores Moleculares

Como referido anteriormente, os métodos QSAR buscam fazer uma associação entre a estrutura de compostos de uma dada família e a sua atividade biológica mediante o uso de descritores moleculares. Assim sendo, a escolha de um bom conjunto de descritores é fundamental para o sucesso do modelo. Define-se como descritor molecular a representação numérica da molécula que descreve quantitativamente as suas informações físico-químicas.¹¹⁴

Nas últimas décadas, tem havido um foco crescente em como representar numericamente a informação codificada na estrutura molecular de forma a estabelecer relações quantitativas entre estruturas e propriedades biológicas. Dessa forma, os descritores moleculares tornaram-se uma ferramenta muito útil para realizar a busca de semelhanças moleculares, pois podem encontrar moléculas com propriedades físico-químicas idênticas, de acordo com sua semelhança, com os valores dos descritores calculados.¹¹⁴

Um descritor molecular deve apresentar as seguintes características fundamentais:^{115,116}

1. Ser invariante para a rotulagem e numeração dos átomos;
2. Ser invariante para a roto-translação da molécula;
3. Ser definido por um algoritmo bem definido;
4. Ter uma aplicabilidade bem definida em estruturas moleculares.

Além disso, um descritor molecular deve apresentar outras características que são consideradas desejáveis, mas não fundamentais, tais como: ^{115,116}

1. Ter uma interpretação estrutural;
2. Ter uma boa correlação com pelo menos uma propriedade experimental;
3. Ser numericamente contínuo;
4. Apresentar degeneração mínima;
5. Ser simples;
6. Ser aplicável a uma ampla classe de moléculas;
7. Ser capaz de discriminar entre isômeros;
8. Ter valores calculados numa faixa numérica adequada para o conjunto de moléculas onde é aplicável;
9. Não ter uma relação trivial com outros descritores moleculares; e (x) não ser baseado em propriedades experimentais.

1.4.1. Descritores 0D

Os descritores 0D são baseados na fórmula molecular, codificam os tipos de átomos e as suas ocorrências dentro de uma molécula, e não fornecem nenhuma informação sobre as conexões dos átomos. São exemplos de descritores 0D a fórmula química, o peso molecular, a contagem de átomos e a contagem de ligações.^{114,115}

1.4.2. Descritores 1D

Os descritores 1D são baseados na lista de subestruturas (que consiste numa lista de fragmentos estruturais que podem fazer parte de uma molécula). Os descritores 1D baseiam-se na contagem de grupos funcionais, como o número total de átomos de carbono primários, número de cianatos, número de grupos nitrilo, etc. Estes descritores não representam toda a topologia da estrutura química, mas

podem ser facilmente calculados e interpretados. São exemplos de descritores 1D as contagens de fragmentos e as contagens de grupos funcionais.^{114,115}

1.4.3. Descritores 2D

A representação bidimensional de uma molécula contém as informações sobre a conectividade dos seus átomos em termos de um grafo molecular.^{115,116} Um grafo é um conjunto de pontos, chamados vértices, que são conectados por linhas, chamadas arestas. Num grafo molecular, os vértices representam os átomos e as arestas representam as ligações: cada aresta une dois vértices. Normalmente, a teoria de grafos usa representações matriciais para traduzir informações contidas no grafo, dependendo das conexões e da definição de invariantes do mesmo.¹¹⁷ Os descritores moleculares derivados da representação bidimensional da molécula são referidos como descritores moleculares 2D. Os descritores moleculares 2D são propriedades numéricas que podem ser calculadas a partir da representação da tabela de conexão de uma molécula. São exemplos de descritores 2D o índice de Wiener, o índice de Balaban e o índice de Randić.^{114,115}

1.4.3.1. Matriz de adjacência

A matriz de adjacência (matriz A), representada na Figura 1.16, é uma matriz quadrada simétrica em que cada elemento a_{ij} é igual a um se os átomos i e j estão conectados e zero caso contrário, conforme descrito na equação 1.8. A matriz A pode ser construída enumerando em qualquer ordem.^{115,117,118}

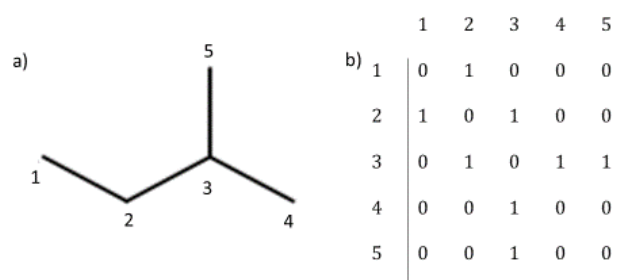


Figura 1.16. Representação do isopropano como grafo (a) e matriz de adjacência (b).

$$[A]_{ij} \equiv a_{ij} = \begin{cases} 1 & \text{se } (i, j) \in E \\ 0 & \text{se } (i, j) \notin E \end{cases} \quad (1.8)$$

onde E é a aresta ou ligação.

A maioria das matrizes de grafos são calculadas a partir de um grafo que não inclui os átomos de hidrogênio, exceto quando necessário para melhor representar a estrutura molecular.¹¹⁵

1.4.3.2. Matriz de distância

A matriz de distância (matriz D) é uma matriz quadrada simétrica cujos elementos D_{ij} correspondem à distância topológica entre os vértices i e j .¹¹⁵ A equação 1.9 descreve uma matriz D,

$$[D]_{ij} \equiv dij = \begin{cases} |^{min}p_{ij}| se <> j \\ 0 se i = j \end{cases} \quad (1.9)$$

onde $|^{min}p_{ij}|$ é o número de arestas ao longo do caminho mais curto entre i e j . O valor zero significa que átomos i e j fazem parte dos componentes do sistema não conectados.¹¹⁹

1.4.3.3. Índice de Wiener

O índice de Wiener (índice W) é o índice topológico mais conhecido e é definido como a metade da soma de todos os elementos d_{ij} da matriz D (equação 1.10).¹¹⁵

$$W = \frac{1}{2} * \sum_{i=1}^{|v|} \sum_{j=1}^{|v|} d_{ij} \quad (1.10)$$

1.4.3.4. Matriz Laplaciana

A matriz Laplaciana (matriz L), também conhecida como matriz de Kirchhoff, é uma matriz simétrica quadrada que é definida como a diferença entre a matriz de grau do vértice e a matriz A.¹¹⁵ A equação 1.11 descreve as condições de determinação de cada elemento da matriz L,

$$[L]_{ij} = \begin{cases} \delta_i se i = j \\ -1 se (i,j) \in E \\ 0 se (i,j) \notin E \end{cases} \quad (1.11)$$

onde δ_i é o grau, conectividade, valência, ou n° de coordenação do átomo i .

Os graus dos vértices determinam quantos átomos pesados ($Z > 1$) estão ligados ao átomo i e podem ser calculados como as somas das linhas da matriz A (equação 1.12),

$$\delta_i = \sum_{j=1}^n a_{ij} \quad (1.12)$$

onde n é o número de vértices e a ordem da matriz A .¹¹⁵

1.4.3.5. Matriz de desvio

A matriz de desvio (matriz Δ), também conhecida como matriz de caminho máximo, é uma matriz de vértices simétrica, como descrita na equação 1.13,

$$[\Delta]_{ij} = \begin{cases} |^{max} P_{ij} | se i <> j \\ 0 se i = j \end{cases} \quad (1.13)$$

onde $|^{max} P_{ij} |$ é o número de arestas ao longo do caminho mais longo entre os vértices i e j , ou seja, o desvio da distância.¹¹⁵

1.4.3.6. Matrizes de termos ponderados

As matrizes de termos ponderados são obtidas associando cada vértice ou aresta a uma propriedade definida (o peso) que serve para diferenciar os diversos tipos de átomos ou ligações. Os pesos são utilizados para diferenciar os diversos tipos de átomos ou ligações. Normalmente, os pesos dos vértices são propriedades atômicas, como números atômicos, massa ou qualquer propriedade físico-química (cargas atômicas, volume ou área acessível ao solvente) que possa ser associada a cada um dos átomos.¹¹⁵

A matriz de conectividade de átomos C é o exemplo mais conhecido de matrizes ponderadas. Esta é uma matriz A ponderada cujos elementos C_{ij} são definidos pela equação 1.14,

$$[C]_{ij\delta} \equiv c_{ij} = \begin{cases} \pi_{ij}^* se (i,j) \in E \\ 0 se (i,j) \notin E \end{cases} \quad (1.14)$$

onde π_{ij}^* é a ordem de ligação convencional.¹¹⁵

Outro exemplo de matrizes ponderadas é a matriz de Randic (matriz X), que é definida pela equação 1.15,

$$[X]_{ij} = \begin{cases} \frac{1}{\sqrt{\delta_i * \delta_j}} se (i,j) \in E \\ 0 se (i,j) \notin E \end{cases} \quad (1.15)$$

onde δ é o grau do vértice dos dois átomos conectados.¹¹⁵

1.4.3.7. Índice de conectividade de Randic

O índice de conectividade de Randic (X_R) pode ser calculado a partir da matriz X (equação 1.15) como um índice do tipo Wiener usando a equação 1.16,

$$X_R \equiv {}^1X = \frac{1}{2} * \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} [X]_{ij} = \frac{1}{2} * \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} a_{ij} (\delta_i * \delta_j)^{-1/2} \quad (1.16)$$

onde a_{ij} são os elementos da matriz A .¹¹⁵

1.4.3.8. Matriz de adjacência de arestas

A matriz de adjacência de arestas é uma matriz simétrica que colhe informações sobre as conectividades das ligações moleculares. A matriz de adjacência de arestas é definida pela equação 1.17.¹¹⁵

$$[E]_{ij} \equiv e_{ij} = \begin{cases} 1 & \text{se } |{}^{min}P_{i,j}| = 1 \\ 0 & \text{se } |{}^{min}P_{i,j}| \neq 1 \end{cases} \quad (1.17)$$

1.4.3.9. Índice de conectividade de Kier e Hall

Os índices de conectividade de Kier e Hall foram propostos por Kier e Hall através de um esquema geral (equação 1.18) para calcular a conectividade com base no índice de Randic (equação 1.16), mas considerando fragmentos maiores do que as ligações, como por exemplo pares de átomos adjacentes,¹¹⁵

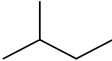
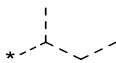
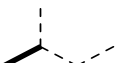
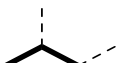
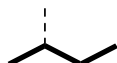
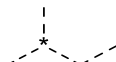
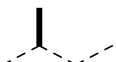
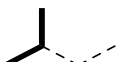
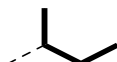
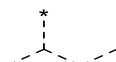
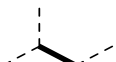
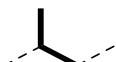

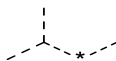
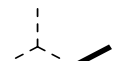
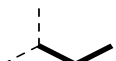

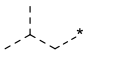



$${}^m\chi = \sum_{k=1}^K \left(\prod_{i=1}^{n_k} \delta_i \right)_k^{-1/2} \quad (1.18)$$

onde δ_i é o grau do vértice do átomo i .

Os índices de conectividade de Kier e Hall são calculados para todos os subgrafos de ordem m pré-definidos. Os subgrafos de ordem zero ($m = 0$) são todos os átomos na molécula, e K é igual ao número de átomos e $n_k = 1$. Analogamente, os subgrafos de primeira ordem ($m = 1$) são todos os pares de átomos adjacentes, ou seja, todos os caminhos de comprimento 1 (ligações) e, para $m = 1$, o índice de conectividade de Randic é obtido com K igual ao número de arestas e $n_k = 2$. O índices de conectividade de segunda ordem ($m = 2$) são calculados em todos os caminhos de ordem 2, em que K é igual ao

número do caminho de comprimento 2 no grafo e $n_i = 3$, pois qualquer caminho de comprimento 2 envolve três átomos.¹¹⁵ A Tabela 1.2 apresenta os exemplos da ordem de conexão do isopropano apresentado em forma de grafo molecular.

Tabela 1.2. Ilustração de ordem de conectividade do isopropano.

				
Ordem m				
Ordem 0	Ordem 1	Ordem 2	Ordem 3	
				
				
				
				
				

1.4.3.9.1. Índice χ de Ordem zero (${}^0\chi$)

A subdivisão mais simples de um grafo molecular é o conjunto de vértices. O número de subgrafos de ordem zero é simplesmente o número de átomos ou vértices da estrutura química. Cada átomo da estrutura química é caracterizado pelos seus valores δ e de valência. O peso atribuído a cada vértice é a raiz quadrada recíproca do valor δ . O termo de conectividade do subgrafo C é definido para a ordem zero para ambas as formas simples (equação 1.19) e de valência (equação 1.20) para o átomo s .

$${}^0C_s = (\delta_s)^{-1/2} \quad (1.19)$$

$${}^0C_s^v = (\delta_s^v)^{-1/2} \quad (1.20)$$

O índice χ de ordem zero (${}^0\chi$ e ${}^0\chi^v$) é a soma desses termos para todos os átomos no grafo (equações 1.21 e 1.22). O índice de conectividade de ordem zero aumenta com o aumento das ramificações.¹²⁰

$${}^0\chi = \sum {}^0C_s \quad (1.21)$$

$${}^0\chi^v = \sum {}^0C_s^v \quad (1.22)$$

1.4.3.9.2. Índice χ de Ordem um (${}^1\chi$)

Os termos do subgrafo 1C_s podem ser definidos em analogia com aqueles das equações 1.19 e 1.20 para a aresta s , ou seja, a aresta entre os vértices i e j (equações 1.23 e 1.24).¹²⁰

$${}^1C_s = (\delta_i\delta_j)_s^{-1/2} \quad (1.23)$$

$${}^1C_s^v = (\delta_i^v\delta_j^v)_s^{-1/2} \quad (1.24)$$

Os índices χ de primeira ordem (${}^1\chi$ e ${}^1\chi^v$) são definidos como a soma de todas as arestas do grafo (equações 1.25 e 1.26), diminuindo com o aumento das ramificações.^{118,120}

$${}^1\chi = \sum {}^1C_s \quad (1.25)$$

$${}^1\chi^v = \sum {}^1C_s^v \quad (1.26)$$

O número de arestas do grafo varia com o tipo da Estrutura. Existe uma relação geral entre o número de arestas (1P), o número de átomos do esqueleto (A) e o número de anéis (R) (equação 1.27).

$$R = {}^1P + 1 - A \quad (1.27)$$

Os índices ${}^1\chi$ contêm mais informações sobre a estrutura do que os índices ${}^0\chi$.¹²⁰

1.4.3.10. Índices de forma κ

Os índices de forma κ são projetados para caracterizar aspectos da forma molecular, comparando uma molécula com as formas extremas que são possíveis para aquele número de átomos.¹²¹ Na abordagem que leva aos índices de forma κ , é necessário transformar caminhos próximos de comprimento m (mP) num índice que contém informações de número de átomos existentes na molécula. Os termos selecionados para qualquer atributo de ordem m são a contagem mínima (${}^mP_{min}$) e a contagem máxima (${}^mP_{max}$) de caminhos nos grafos moleculares com uma contagem de átomos comum.¹²⁰

1.4.3.10.1. Atributo de forma κ de ordem um (${}^1\kappa$)

O atributo da ${}^1\kappa$ é descrita pela equação 1.28,

$${}^1\kappa = 2 \cdot {}^1P_{max} * {}^1P_{min} / ({}^1P_i)^2 \quad (1.28)$$

onde ${}^1\kappa$ está relacionado com a ciclicidade de uma molécula e 1P_i é número de fragmentos com uma ligação adjacente.¹²⁰

1.4.3.10.2. Atributo de forma κ de ordem dois (${}^2\kappa$)

O atributo da ${}^2\kappa$ é descrita pela equação 1.29.

$${}^2\kappa = 2 \cdot {}^2P_{max} * {}^2P_{min} / ({}^2P_i)^2 \quad (1.29)$$

onde 2P_i denota o número de fragmentos com duas ligações adjacentes.

No caso de ${}^2\kappa$, a informação é codificada em relação ao grau de semelhança a uma estrutura em forma de estrela (valores mais elevados/reduzidos) ou a uma estrutura linear (valores mais reduzidos/elevados).¹²⁰

1.4.3.10.3. Atributo de forma κ de ordem três (${}^3\kappa$)

O atributo de ${}^3\kappa$ é descrito pela equação 1.30,

$${}^3K = 4 \cdot {}^3P_{max} * {}^3P_{min} / ({}^3P_i)^2 \quad (1.30)$$

onde o fator de escala 4 é usado no numerador para trazer o ${}^3\kappa$ aproximadamente para a mesma escala numérica que os outros valores κ e 3P_i indica o número de fragmentos com três ligações adjacentes. Os valores de ${}^3\kappa$ podem ser expressos pela equação 1.31,

$${}^3\kappa = \begin{cases} \frac{(n-1)(n-3)^2}{P_3^2} & \text{se } n \text{ for ímpar,} \\ \frac{(n-3)(n-2)^2}{P_3^2} & \text{se } n \text{ for par.} \end{cases} \quad (1.31)$$

onde n é o número de átomos ($Z > 1$) e p_3 é o número de caminhos de comprimento três (grupos de átomos conectados por três ligações).

O $^3\kappa$ codifica a informação sobre a centralidade da ramificação. Os valores de $^3\kappa$ são maiores quando a ramificação é inexistente ou quando está localizada nas extremidades da estrutura, ou seja, os valores $^3\kappa$ diminuem quanto mais globular a forma da molécula.¹²⁰

1.4.3.11. Descritores de propriedades físicas

Algumas propriedades físicas podem ser calculadas a partir da tabela de conexão (sem dependência da conformação) de uma molécula.¹¹⁸ A Tabela 1.3 descreve os principais descritores moleculares baseados em propriedades físicas usados neste trabalho.

O coeficiente de partição ($\log P$) e a refratividade molar (SMR) de pequenas moléculas podem ser calculados como a soma das contribuições de cada um dos átomos nas moléculas (equação 1.33),

$$P_{cal} = \sum_i n_i a_i \quad (1.32)$$

onde P_{calc} é a propriedade a ser calculada ($\log P$ ou SMR), n_i é o número de átomos do tipo i presentes na molécula e a_i é a contribuição dos átomos do tipo i .¹²²

Tabela 1.3. Descritores de propriedade físicas

Acrónimo	Descrição
SMR	Refratividade molecular (incluindo átomos de hidrogénio) é calculada usando um modelo de contribuição atómica que assume o estado de protonação correto. ¹¹⁸
$\log P(o/w)$	Coefficiente de partição octanol/água (incluindo átomos de hidrogénio). Esta propriedade é calculada a partir de um modelo linear de contribuições atómicas com $r^2 = 0,931$, RMSE=0,393 em 1827 moléculas. ¹¹⁸
SlogP	Coefficiente de partição octanol/água (incluindo átomos de hidrogénio). Esta propriedade é um modelo de contribuição atómica que calcula $\log P$ a partir do estado de protonação correto. Os resultados podem variar do descritor $\log P(o/w)$. ¹¹⁸

Acrónimo	Descrição
TPSA	Área de superfície pertencente a átomos polares é um descritor que correlaciona o transporte passivo da molécula através das membranas e, portanto, permite a previsão das propriedades de transporte de fármacos ¹²³ . Área de superfície polar (Å^2) é calculada usando contribuições de grupo para aproximar a área de superfície polar apenas a partir das informações da tabela de conexão. ^{118,123}
A_{vdW}	Área da superfície de van der Waals (Å^2) calculada usando uma aproximação da tabela de conexão. ¹¹⁸
V_{vdW}	Volume de van der Waals (Å^3) calculado usando uma aproximação da tabela de conexão. ¹¹⁸

1.4.3.12. Áreas de superfície subdivididas

A área de superfície associada a um átomo é aquela não contida em nenhum outro átomo da molécula. Considerando cada átomo como uma esfera de raio igual ao raio de van der Waals, obtém-se a área de superfície de Van der Waals (VSA, do inglês *Van der Waals Surface Area*).¹²⁴

As áreas de superfície subdivididas são descritores baseados num cálculo aproximado de VSA para cada átomo (v_i), junto com alguma outra propriedade atômica (p_i) de acordo com a equação 1.33,

$$P_VSA_K = \sum_i v_i \delta(p_i \in [a_{K-1}, a_K]) \quad k = 1, 2, \dots, n \quad (1.33)$$

onde v_i é a contribuição atômica do átomo i para o VSA da molécula e $a_0 < a_k < a_n$ são limites de intervalo que limitam todos os valores de p_i em qualquer molécula. Os v_i são calculados usando uma aproximação da tabela de conexão. Cada descritor numa série é definido como sendo a soma de v_i sobre todos os átomos i , tal que, p_i está num intervalo específico.^{118,124} São alguns exemplos dos descritores baseados em áreas de superfície subdivididas:

- SlogP_VSA₁- Soma de v_i para $L_i \in] -0,4; -0,2]$.
- SlogP_VSA₂- soma de v_i para $L_i \in] -0,2; 0]$.

- SlogP_VSA₅- Soma de v_i para $L_i > 0,40$.
- SMR_VSA₆- Soma de v_i para $R_i \in [0; 0,11]$.
- SMR_VSA₇- Soma de v_i para $R_i > 0,56$.

em que L_i denota a contribuição para $\log P$ para o átomo i e R_i denota a contribuição para a SMR para o átomo i .

1.4.3.13. Descritores de carga parcial

Os descritores que dependem da carga parcial de cada átomo de uma estrutura química normalmente usam o método da equalização parcial de eletronegatividades orbitais (PEOE, do inglês *Partial Equalization of Orbital Electronegativities*) para o cálculo da mesma. A equação 1.3 descreve a fórmula de cálculo de cargas parciais atômicas utilizando o método PEOE. As cargas PEOE dependem apenas da conectividade das estruturas, tais como: elementos, cargas formais e ordens de ligação.^{118,125}

$$dq_{ij} = \frac{(1/2^k)(X_i - X_j)}{X_j^+} \quad (1.34)$$

onde dq_{ij} é a quantidade de carga transferida entre os átomos i e j ($X_i > X_j$), X_j^+ é a eletronegatividade do ião positivo do átomo j , X_i é a eletronegatividade do átomo i (quadraticamente dependente da carga parcial) e k é o número de iteração do algoritmo.

1.4.3.14. Descritores eletrotológicos

O estado eletrotológico (E-State, do inglês *Electrotopological State*) representa a acessibilidade de um elétron de valência num dado átomo. Esses índices codificam acessibilidade de elétrons, presença ou ausência de grupos e contagem de grupos.^{126,127}

Para calcular um índice E-state, um valor intrínseco do átomo é atribuído a cada átomo de acordo com a equação 1.35,

$$I = \frac{\delta^V + 1}{\delta} \quad (1.35)$$

onde δ^V e δ são as contagens de valência e elétrons σ de átomo associado ao grafo molecular.¹²⁶⁻¹²⁸

O valor do E-state para o átomo i , S_i , é definido pela equação 1.36 e a influência do átomo j no átomo i é dada pela equação 1.37,

$$S_i = I_i + \Delta I_i \quad (1.36)$$

$$\Delta I_i = \sum_{j=1}^N (I_i - I_j) / r_{ij}^2 \quad (1.37)$$

onde r_{ij} é a separação entre o átomo i e o átomo j no grafo molecular, contado como o número de átomos.¹²⁶⁻¹²⁸

Os índices E-state são normalmente combinados com as VSA, formando uma série de descritores E-state_VSA, os quais são fracionados por intervalos do valor de S_i .

1.4.3.15. Descritores de auto-correlação

Os descritores de auto-correlação denotam uma classe heterogénea de descritores moleculares cuja fórmula geral é apresentada na equação 1.38,

$$\mathbf{D5}_{(\mathcal{L}; \alpha, \lambda, \kappa)} = \alpha * \sum_{i=1}^A \sum_{j=1}^A (\mathcal{L}_i * \mathcal{L}_j)_{ij}^\lambda * \delta(d_{ij}; \kappa) \quad (1.38)$$

onde \mathcal{L} é um invariante do vértice, α é um fator de escala, λ é um parâmetro de potência, $\delta(d_{ij}; \kappa)$ é uma função δ de Kronecker igual a um para pares de subestruturas centrais na distância topológica $d_{ij} = \kappa$ e zero caso contrário, e A é o número de subestruturas centrais que normalmente são os átomos da molécula.¹¹⁵

Os descritores de auto-correlação estão relacionados com a distribuição topológica de uma propriedade molecular genérica e medem a força de uma relação entre átomos a uma distância predefinida igual a κ .¹¹⁵

1.4.3.16. Descritores BCUT

Os descritores BCUT derivam da sugestão de Burden (B) e na validação dessa sugestão pelo Serviço de Resumos Químicos (CAS, do inglês, *Chemical Abstracts Service*) (C) e de Pearlman da Universidade do Texas (UT) que adicionou algumas extensões significativas aos descritores.¹²⁹

Descritores BCUT foram projetados para codificar propriedades atômicas relevantes para interações intermoleculares.¹²¹ Os valores BCUT são baseados num descritor desenvolvido por Burden, que é calculado a partir duma representação matricial da tabela de conexão de uma molécula, em que

os números atômicos dos átomos pesados são colocados na diagonal da matriz, os elementos fora da diagonal recebem o valor 0,1 vezes a ordem de ligação se os átomos estiverem ligados e 0,001 se os átomos não estiverem ligados. O método foi estendido por Pearlman para gerar uma família de descritores chamados de descritores BCUT que podem ser usados para definir um espaço químico de baixa dimensão.^{121,129} Em vez de usar o número atômico, os elementos da diagonal nas matrizes BCUT codificam carga atômica, polaridade atômica e capacidade do átomo fazer a ligação de hidrogênio.¹²¹

Para além dos números atômicos, a diagonal principal da matriz de conexão da molécula pode conter o valor das cargas parciais PEOE (neste caso, cada entrada ij da matriz de adjacência assume o valor $1/b_{ij}^{1/2}$, onde b_{ij} é a ordem de ligação formal entre os átomos i e j ligados) ou as contribuições atômicas para $\log P$ ou para a refratividade molar, dando origem a variantes destes descritores.¹¹⁸

1.4.4. Descritores 3D

Descritores moleculares 3D derivam da informação espacial sobre as posições dos átomos, para além das relações de conectividade entre os átomos. Esta representação é referida como representação geométrica de uma molécula. São alguns exemplos de descritores 3D: *3D fragment screens* (que codifica relações espaciais, como distâncias e ângulos entre as diferentes características de uma molécula, como átomos e centro geométrico de anéis); *pharmacophore keys* (baseados em características farmacofóricas, ou seja, átomos ou subestruturas relevantes para a ligação ao alvo biológico); e índices topográficos 3D (calculados a partir da matriz de distância da molécula).^{115,130}

1.5. Representação molecular

Uma vez que os métodos QSAR são baseados na correlação entre os descritores moleculares e a atividade biológica, no cálculo desses descritores moleculares, especificamente nos 2D, é tradicionalmente usada a forma de representação molecular por SMILES (representação simplificada do sistema molecular de entrada de linha, do inglês *Simplified Molecular-Input Line-Entry System*). A primeira versão do SMILES foi proposta em 1988 por David Weininger. O SMILES é um protocolo para a escrita de estruturas químicas usando apenas caracteres de texto, números e alguns símbolos simples, tais como: "-" para representar as ligações simples, mas o símbolo pode ser omitido, "=" para representar ligações duplas, "#" para representar ligações triplas, "\$" para representar ligações quadruplas, "@" para representar os centros quirais R e "@@" para representar os centros quirais S . Cada átomo é representado pelo seu símbolo na tabela periódica, átomos de hidrogênio normalmente

não são explicitamente representados, parênteses são usados para indicar pontos de ramificação e os rótulos numéricos designam pontos de conexão de anéis.^{121,131,132}

1.6. Testes de suscetibilidade aos antimicrobianos

Os testes de suscetibilidade antimicrobiana são procedimentos laboratoriais *in vitro* realizados para selecionar um antimicrobiano especificamente eficaz para um determinado microrganismo. Eles são usados em epidemiologia, na prática clínica (para selecionar os antimicrobianos mais eficazes) e na descoberta de fármacos. Os objetivos destes testes são detetar possível resistência a fármacos em microrganismos específicos e assegurar a suscetibilidade aos fármacos de escolha para infeções específicas.¹³³⁻¹³⁵

De entre os vários métodos usados para avaliar a suscetibilidade de um microrganismo a um determinado agente antimicrobiano,^{3,136,137} o Comité Europeu de Testes de Suscetibilidade Antimicrobiana (EUCAST, do inglês *European Committee on Antimicrobial Susceptibility Testing*) recomenda o método de microdiluição em caldo como teste de referência para as polimixinas, sendo que os outros métodos, incluindo diluição em ágar, difusão em disco e difusão em gradiente, não são recomendados por não haver evidências sobre a eficácia destes métodos na determinação da suscetibilidade dos microrganismos às polimixinas.¹³⁸

No método de microdiluição em caldo, a suscetibilidade do microrganismo é determinada por meio da concentração mínima inibitória (MIC, do inglês *Minimal Inhibitory Concentration*), que é a concentração mais baixa de um agente antimicrobiano que, sob condições *in vitro* estritamente controladas, impede o crescimento visível dos microrganismos. A MIC define os níveis *in vitro* de suscetibilidade ou resistência de espécies bacterianas ao agente antimicrobiano aplicado.¹³⁹

1.7. Objetivos

Este trabalho tem como objetivo encontrar análogos da polimixina B que sejam mais ativos e menos tóxicos utilizando métodos da química computacional.

1.7.1. Objetivo geral

- Desenhar novas moléculas baseadas na polimixina B com propriedades antimicrobianas e baixa citotoxicidade, usando uma abordagem *in silico* envolvendo vários passos de *design* de fármacos.

1.7.2. Objetivos específicos

- Tratar dados pré-existentes de atividade antimicrobiana e citotoxicidade;
- Desenvolver modelos preditivos de atividade antimicrobiana e citotoxicidade;
- Descobrir novas estruturas com propriedades antimicrobianas promissoras e com baixa toxicidade.

2. METODOLOGIA

Este capítulo descreve a metodologia usada para o desenvolvimento do trabalho. O capítulo é dividido em 5 seções: (2.1) colheita inicial de dados; (2.2) desenvolvimento do primeiro modelo preditivo; (2.3) colheita de novos dados experimentais para alimentar o modelo; (2.4) desenvolvimento do modelo com novos dados experimentais; (2.5) aplicação do melhor modelo desenvolvido na previsão da atividade antimicrobiana dos análogos da polimixina B.

2.1. Colheita inicial dos dados

Os dados foram recolhidos da base de dados PubChem.¹⁴⁰ Inicialmente, foi pesquisado o termo de interesse “*polymyxin B*” e, através do número de identificação do composto (CID, do inglês *Compound Identifier*) 123978 do composto (2S, 3R) -2-amino- N -[(2S) -4-amino-1-oxo-1-[[[(3S, 6S, 9S, 12S, 15R, 18S, 21S) -6,9,18-tris(2-aminoetil)-15-benzil-3-[(1R) -1-hidroxiethyl]-1,2-(2-metilpropil)-2,5,8,11,14,17,20-heptaóxido-1,4,7,10,13,16,19-heptazacilobutano-21-il]amino]butano-2-il]-3-hidroxi]butanamida) que corresponde a parte peptídica da polimixina B, procurou-se estruturas similares à polimixina B existentes na base de dados selecionando-se no separador “*Similarity*” (a busca da similaridade baseou-se no coeficiente de Tanimoto e no *finger print* 2D da estrutura), para encontrar até 1000 estruturas semelhantes.^{141,142} Os compostos similares encontrados foram primeiro filtrados pela disponibilidade de dados de ensaios biológicos e de toxicidade, e depois os dados com ensaios biológicos foram filtrados pela disponibilidade de um valor definido para MIC, MIC₅₀ ou MIC₉₅. Os dados encontrados foram compilados numa tabela e a informação relativa ao composto foi suplementada pela representação SMILES do mesmo.

2.1.1. Caracterização dos dados

Os dados foram processados e analisados usando a linguagem *Python* versão 3.10. através da plataforma Google Colab, tendo sido usados os pacotes *Pandas* versão 2.1.0,¹⁴³ *NumPy* versão 1.26.0,¹⁴⁴ *Matplotlib* versão 3.7.2,¹⁴⁵ *seaborn* versão 0.13,¹⁴⁶ *Scikit-learn* versão 1.0.2,⁹⁶ e *RDKit* versão 2023.03.1.¹⁴⁷

2.1.2. Cura de dados

Os dados foram curados para a remoção de duplicados, entradas sem descrição do microrganismo alvo ou pertencentes a estudos de combinação de fármacos. A análise prévia dos dados revelou a existência de uma distribuição irregular das espécies do alvo biológico, com poucos dados para algumas delas. Dado que isto poderia inviabilizar o desenvolvimento do modelo, e para colmatar esta situação, criou-se uma variável que contém o género taxonómico do microrganismo alvo (T.G) (ao invés

da espécie). Para além disso, uma vez que as polimixinas têm maior atividade em bactérias Gram-negativas e pouca atividade em fungos e bactérias Gram-positivas, foi criada uma variável referente a uma classificação mais ampla do microrganismo (M_{Typ}), que descreve o tipo de microrganismo (bactérias Gram-negativas, bactérias Gram-positivas e fungos), de modo a verificar a contribuição deste parâmetro no modelo.

2.2. Desenvolvimento do primeiro modelo

2.2.1. Geração de descritores moleculares

Foram calculados vários descritores moleculares usando o pacote RDKit versão 2023.03.1.¹⁴⁷ Estes descritores foram calculados a partir da representação SMILES de cada molécula, providenciada pela base de dados da PubChem, e divididos por famílias de descritores, conforme a classificação detalhada Tabela 2.1.

Tabela 2.1. Classificação dos descritores, com indicação do número de descritores (n_i) gerado em cada família.

Família de descritores moleculares	Descritores moleculares	n_i
Gen	BalabanJ, Ipc, RingCount, NumRotatableBonds, MolLogP, FractionCSP3, HallKierAlpha, BertzCT, TPSA, LabuteASA, HeavyAtomCount, MolWt, NumValenceElectrons, NumHeteroatoms	14
Hb	Descritores relacionados à formação de ligação de hidrogénio: NumHAcceptors, NHOHCoun; NumHDonors, NOCount	4
CKP	${}^1\chi, {}^0\chi, {}^3\chi, {}^1\chi, {}^0\chi^n, {}^2\chi^v, {}^4\chi^n, {}^2\chi, {}^1\chi^n, {}^3\chi^n, {}^1\chi^v, {}^3\chi^v, {}^4\chi^v, {}^0\chi^v, {}^2\chi^n$	15
PEOE_VSA	PEOE_VSA ₁ -PEOE_VSA ₁₄	14
SMR_VSA	SMR_VSA ₁ -SMR_VSA ₁₀	10
SlogP_VSA	SlogP_VSA ₁ -SlogP_VSA ₁₂	12
Estate_VSA	Estate_VSA ₁ -Estate_VSA ₁₁	11
AC2D	Função bidimensional de autocorrelação ¹¹⁹	192
BCUT2D	Métricas BCUT de Perlman ¹⁴⁸	8
FG	Contagem de fragmentos de grupos funcionais	85

2.2.2. Exploração dos algoritmos

Foram considerados três algoritmos: um algoritmo baseado em árvores de decisão¹⁴⁹ e dois algoritmos de conjunto, RF e *AdaBoost*,¹⁵⁰ com as diferentes famílias de descritores moleculares (Tabela 2.1) dando origem a total de 30 modelos classificadores. As variáveis T_xG e M_{Typ} foram adicionadas a cada conjunto de descritores moleculares de forma a providenciar informação relativa ao alvo biológico. Cada modelo (par algoritmo/descritor) foi treinado visando uma predição multiclasse do quartil da MIC, usando uma divisão de 65:35 entre os dados de treino e teste.

Todos os modelos foram criados como um *pipeline* de dados, onde todos os campos numéricos foram primeiro normalizados para média zero desvio-padrão unitário, todas as variáveis não numéricas foram codificadas, usando codificação *one-hot*, antes de serem fornecidas ao algoritmo principal. Os modelos de DT foram treinados nos dados transformados usando os hiper-parâmetros padrão definidos na sua implementação. Por outro lado, alguns dos hiper-parâmetros dos algoritmos RF e AdaBoost foram alvo de otimização. Para os modelos de RF, o número de estimadores (árvores) (n_{est}), bem como a fração de amostras (n_s) e *features* (n_f) consideradas por cada estimador, foram otimizados usando o método de validação cruzada *5-fold*, considerando 100 combinações aleatórias de n_{est} , n_s e n_f . Um esquema semelhante de validação cruzada também foi usado no caso dos modelos AdaBoost, otimizando o número de estimadores (árvores) (n_{est}), a profundidade dos estimadores da base (d_{est}) e a taxa de aprendizagem (r_L). Todos os hiper-parâmetros foram testados numa gama de valores apresentados na Tabela 2.2

Tabela 2.2. Valores testados para cada parâmetro nos modelos RF e AdaBoost

Parâmetros	Valores testados	
	RF	AdaBoost
n_{est}	5; 10; 20; 50; 100	5; 10; 20; 50; 100
n_s	0,1; 1; 5	NA
n_f	0,05; 1; 10	NA
d_{es}	NA	1; 2; 5; 10; 10
r_L	NA	0,01; 2; 20

2.2.3. Análise do modelo

Após a execução de cada modelo, analisou-se o seu desempenho com base na matriz de confusão, tendo-se calculado a exatidão, a taxa de verdadeiros positivos para Q1, $f(Q1|Q1)$, e as $f(Q1|Q4)$ e $f(Q4|Q1)$, considerando-se um bom modelo aquele que possui uma exatidão elevada (sem grande quebra entre os valores do treino e do teste), uma maior taxa de verdadeiros Q1 e $f(Q1|Q)$ e $f(Q4|Q1)$ menores. Também uma análise mais aprofundada de cada modelo foi realizada usando *scripts Python* desenvolvidos internamente para analisar a importância de cada variável, a dependência parcial da resposta de cada modelo às variáveis mais importantes, bem como a resposta de cada modelo a mutações sistemáticas da estrutura da polimixina B.

2.3. Ensaio de suscetibilidade *in vitro*.

No desenvolvimento do trabalho, verificou-se que algumas espécies bacterianas possuíam poucas entradas nos dados recolhidos na PubChem, o que poderia influenciar negativamente o treino do modelo. Para resolver este problema e tentar melhorar a precisão e exatidão do modelo, foi decidido analisar a suscetibilidade de algumas das espécies bacterianas com poucas entradas às polimixinas (B e E), usando o método de microdiluição em caldo, de acordo com as normas da EUCAST.¹⁵¹

2.3.1. Microrganismos e reagentes

Neste trabalho, foram determinadas as MICs da colistina e da polimixina B em duas espécies de bactérias Gram-negativas e uma espécie bacteriana Gram-positiva, conforme descrito na Tabela 2.3. A *E. coli* foi usada como controlo de qualidade para os ensaios com *Enterobacterales* (*Shigella sonnei* e *Proteus mirabilis*). Todas as espécies usadas pertencem à coleção de microrganismos do Centro de Engenharia Biológica (CEB) da Universidade do Minho.

Tabela 2.3. Espécies bacterianas usadas na colheita de novos dados de MIC.

Espécie bacteriana	Estirpe	Classificação de Gram
<i>Escherichia coli</i> (controlo)	ATCC 25922	Negativa
<i>Shigella sonnei</i>	ATCC 25931	Negativa
<i>Proteus mirabilis</i>	CECT 4168	Negativa
<i>Listeria monocytogenes</i>	ATCC 15313T	Positiva

O caldo BHI (Infusão de Cérebro e Coração, do inglês *Brain Heart Infusion*), triptona, extrato de levedura, TSB (Caldo Triptico de Soja, do inglês *Tryptic Soy Broth*) e agar foram obtidos na Liofilchem (Itália). MHB II (Caldo Muller-Hinton II, do inglês *Muller-Hinton Broth II*), corante vermelho congo e sulfato de colistina foram obtidos na Sigma-Aldrich (EUA). NaCl (cloreto de sódio) foi obtido na Fisher Chemical (Reino Unido). Sulfato de polimixina B foi obtido na PanReac AppliChem (Espanha).

2.3.2. Meios de cultura e soluções

Para o crescimento da *Listeria monocytogenes*, o meio líquido usado foi o caldo BHI, preparado a uma concentração de 37 g/dm³, e o meio sólido foi o caldo BHI suplementado com agar (15 g/dm³). Foi ainda usado o meio líquido MH-F (Muller-Hinton suplementado com 5 % de sangue de cavalo lisado e 20 mg/dm³ de β -NAD),¹⁵² previamente preparado pelo grupo de investigação, no ensaio da MIC.

Para o crescimento da *Shigella sonnei*, foi usado o MHB II preparado a uma concentração de 22 g/dm³ como meio líquido e o TSB a uma concentração de 30 g/dm³, suplementado com corante vermelho congo (0,1 g/dm³) e agar (15 g/dm³) como meio sólido.

Para o crescimento em meio líquido da *Proteus mirabilis*, foi usado o meio *low salt* LB (caldo de lisogenia com baixo teor de sal, do inglês *low salt Lysogeny broth*), preparado a partir de triptona (10 g/dm³), extrato de levedura (5 g/dm³) e NaCl (0,5 g/dm³) na determinação da curva de calibração, e o meio líquido MBH II (22 g/dm³) foi usado no ensaio da MIC. Para o crescimento em meio sólido foi usado o *low salt* LB suplementado com agar (15 g/dm³).

Nas diluições para a determinação da curva de calibração, foi usada a solução de NaCl a 0,9 % (m/v). Todos os meios de cultura e a solução salina foram preparados com água destilada como solvente e esterilizados em autoclave durante 15 minutos a 121 °C.

Foram, ainda, usados dois antibióticos, o sulfato de colistina e o sulfato de polimixina B. Foram preparadas as soluções *stock* de ambos os agentes antimicrobianos a uma concentração de 1 g/dm³ de acordo com as recomendações do fabricante, usando água esterilizada como solvente e conservando-as a -20 °C. Para cada ensaio da MIC, foram preparadas diferentes diluições a partir das soluções *stock* usando o meio de crescimento adequado à espécie a testar.

2.3.3. Curva de calibração para ajuste da concentração celular

Ao longo do trabalho, houve a necessidade de ajustar a concentração celular das espécies bacterianas para os ensaios da MIC. Para tal, foi necessário elaborar previamente uma curva de calibração que relacione a concentração celular em unidades formadoras de colónias (CFU, do inglês *Colony Forming Units*) com a DO (Densidade Óptica). Inicialmente, prepararam-se os pré-inóculos em 15 cm³ do meio líquido inoculado com algumas colónias obtidas a partir do espalhamento das bactérias em placas de Petri. Estes foram incubados a 37 °C, sob agitação, a 120 rpm (rotações por minuto), durante a noite (aproximadamente 17 horas). Após este período, transferiu-se 1 cm³ da suspensão bacteriana para um microtubo, sendo posteriormente centrifugada durante 5 minutos à temperatura ambiente e a 9000 g. De seguida, descartou-se o sobrenadante, adicionou-se 1 cm³ do meio líquido, e ressuspendeu-se o *pellet* no vortex. De seguida, prepararam-se cinco diluições em meio líquido e mediram-se as suas DO num espectrofotómetro de microplacas (EZ Read 800 Plus) a 620 nm. O objetivo foi encontrar 5 valores de DO entre 0,1 e 1 para posteriormente traçar a curva de calibração. A partir das 5 soluções de bactéria preparadas anteriormente, foram feitas diluições decimais de 10⁻¹ a 10⁻⁸ em solução salina. De seguida, inoculou-se 10 mm³ de cada diluição numa placa de Petri contendo meio sólido e incubou-se durante 24 horas a 37 °C. Após este período, fez-se a contagem de colónias correspondente a cada diluição e determinou-se o número de CFU correspondente a cada valor da DO, usando a equação 2.1. Por fim, procedeu-se à construção da curva de calibração, usando a regressão linear, que relaciona a concentração celular (CFU/cm³) com os valores da DO (Figura A1 em Anexos). Todos os ensaios foram realizados pelo menos em duplicado.

$$\text{CFU/cm}^3 = \frac{\text{CFU} * \text{fator inverso de diluição}}{10 \text{ mm}^3} * \frac{1000 \text{ mm}^3}{1 \text{ cm}^3} \quad (2.1)$$

onde 1 CFU equivale a uma colónia.

2.3.4. Determinação da MIC

A atividade antimicrobiana das polimixinas B e E foi avaliada mediante a determinação da MIC, usando o método da microdiluição em caldo, de acordo com os procedimentos descritos pelo EUCAST.¹⁵¹ Inicialmente, foram preparados os pré-inóculos como descrito anteriormente. Posteriormente, foram utilizadas placas de microdiluição de 96 poços de fundo redondo de poliestireno estéril (lifesciences, EUA), cujos poços foram preenchidos inicialmente com 100 mm³ do meio líquido com concentrações

crescentes dos antibióticos, aos quais foram adicionados 100 mm³ de suspensão bacteriana a uma concentração de 10⁶ CFU/cm³. De seguida, as placas foram incubadas a 37 °C, durante 24 horas. Após este período, fez-se a leitura da MIC mediante a observação a olho nu da inibição do crescimento bacteriano nos poços com o agente antimicrobiano, comparando com o controlo negativo (poço com meio e células e sem agente antimicrobiano) e o branco (poço só com meio de cultura). Para fazer a leitura da MIC dos agentes antimicrobianos em *Listeria monocytogenes*, dado que as características do meio utilizado não permitem a determinação a olho nu, foram inoculados 10 mm³ de cada poço da microplaca em placas de Petri com meio sólido. Posteriormente, as placas foram incubadas durante 24 horas a uma temperatura de 37 °C. Após este período, fez-se a observação a olho nu da inibição do crescimento das bactérias, tendo em consideração a concentração do antimicrobiano do poço em que o inóculo foi retirado. Todos os ensaios da MIC foram realizados pelo menos em duplicado. Os valores da MIC foram convertidos de g/dm³ para µM, unidade usada nos dados colhidos na PubChem.

2.4. Treino do novo modelo

Após terem sido obtidos os valores da MIC para algumas das espécies menos representadas nos dados iniciais, foram desenvolvidos 30 novos modelos (referidos como modelos da segunda série), explorando os algoritmos DT, RF e AdaBoost, conforme descrito na subsecção 2.2, de modo a analisar o impacto de novos dados.

2.5. Aplicação do modelo

Após o desenvolvimento dos modelos, foi selecionado o melhor modelo como aquele que apresenta o melhor desempenho e foi usado para prever a potencial atividade antimicrobiana para os géneros de microrganismos *Acinetobacter*, *Pseudomonas* e *Escherichia* de novos derivados sintéticos das polimixinas B e E, propostos pelo grupo de Química Biomolecular Aplicada do Centro de Química da Universidade do Minho. Além disto, foram feitas mutações sistemáticas nas posições 1 a 3 e 5 a 10 da polimixina B, usando glicina (Gly), leucina (Leu), lisina (Lys) e ácido glutâmico (Glu). Essas mutações têm como objetivo analisar o comportamento do modelo ao variar o impedimento estérico numa posição específica (Gly vs Leu) ou mediante a introdução de um aminoácido básico ou ácido (Lys vs Glu). Em todas as execuções preditivas do modelo, foram usados os géneros de microrganismos *Acinetobacter*, *Pseudomonas* e *Escherichia* como alvos. Finalmente, o mesmo modelo foi usado para

3. RESULTADOS E DISCUSSÃO

O presente capítulo apresenta os resultados e a discussão dos mesmos. Este capítulo é dividido em 8 seções: (3.1) caracterização dos dados iniciais extraídos da PubChem; (3.2) caracterização dos modelos da primeira série; (3.3) caracterização do melhor modelo da primeira série; (3.4) apresentação e discussão dos resultados laboratoriais. Nas seções 3.5, 3.6 e 3.7, faz-se a mesma análise que nas seções 3.1, 3.2 e 3.3, mas usando os dados da segunda série (dados dos ensaios laboratoriais adicionados aos dados extraído da Pubchem). Finalmente, na seção 3.8, discute-se a aplicação do melhor modelo.

3.1. Caracterização dos dados colhidos

Foi obtido, da PubChem, um conjunto de dados com 399 pares de moléculas/microrganismos constituído por 56 compostos similares (ou iguais) à polimixina B, cuja actividade foi testada contra 40 espécies de microrganismos (bactérias e fungos) distribuídas em 24 géneros taxonómicos. Além disto, também foram colhidos compostos similares à polimixina B com dados de toxicidade, os quais, após a cura, resultaram um conjunto de dados com 24 entradas e com um total de 9 compostos. Estes dados são insuficientes para desenvolver um modelo preditivo da toxicidade de acordo com as metodologias propostas neste trabalho.

Antes de proceder à análise dos modelos desenvolvidos, foi necessário caracterizar os dados obtidos. Na Figura 3.1, foi feita a caracterização dos dados quanto ao tipo de microrganismo e quanto ao género taxonómico. Na Figura 3.1a, é possível verificar que a maioria dos dados estão relacionados com actividade antibacteriana para bactérias Gram-negativas, o que seria de esperar, uma vez que as polimixinas são usadas para tratar infeções causadas por bactérias Gram-negativas. Além disso, as bactérias Gram-negativas dominam a lista das bactérias que necessitam urgentemente de novos antibióticos,³⁵ daí haver mais interesse em investigar a atividade dos compostos similares a polimixina B contra bactérias Gram-negativas.

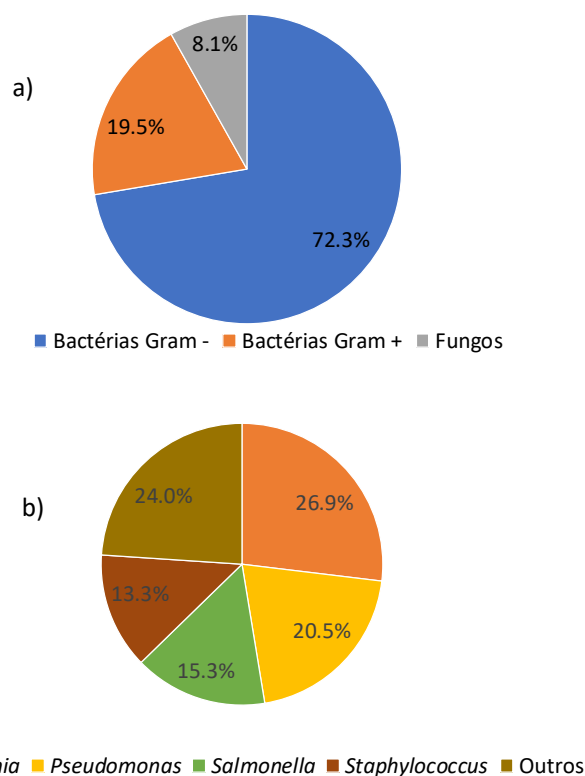


Figura 3.1. Caracterização dos dados colhidos por tipo de microrganismo (a) e por género taxonómico (b).

Relativamente à caracterização dos dados por género taxonómico, os resultados revelaram que quatro géneros de microrganismos (*Escherichia*, *Pseudomonas*, *Salmonella* e *Staphylococcus*) são os mais representados no conjunto de dados, representando mais de 75% das entradas (Figura 3.1a). O maior número de dados com estes géneros poderá ser devido a estes serem patógenos comuns em humanos. Por exemplo, bactérias do género *Staphylococcus* são das principais fontes de infeção em lesões e nas vias respiratórias.¹⁵³ De modo particular, a espécie *Staphylococcus aureus* é ainda uma das principais causas de infeções da corrente sanguínea com risco de vida, como a sepsis, e de endocardites.¹⁵⁴ A *E. coli* é uma das causas mais frequentes de várias infeções bacterianas comuns em humanos e animais, tais como: infeção urinária, doenças diarreicas, sepsis e meningite.^{153,155} As bactérias do género *Salmonella* estão associadas a febre tifoide, bacteremia e intoxicação alimentar.¹⁵³ A *Pseudomonas aeruginosa* é a principal causa das infeções hospitalares, infetando as vias respiratórias e urinaria, queimaduras e uma causa comum da otite externa.^{153,156} Por outro lado, os géneros *Priestia*, *Yersinia*, *Enterobacter*, *Cryptococcus*, *Shigella*, *Vibrio* e *Proteus* foram os menos representados, com apenas uma entrada (0,25%) por género.

Além da caracterização dos dados quanto variáveis relacionadas com o alvo biológico, também foi feita a caracterização de acordo com os compostos usados (Figura 3.2).

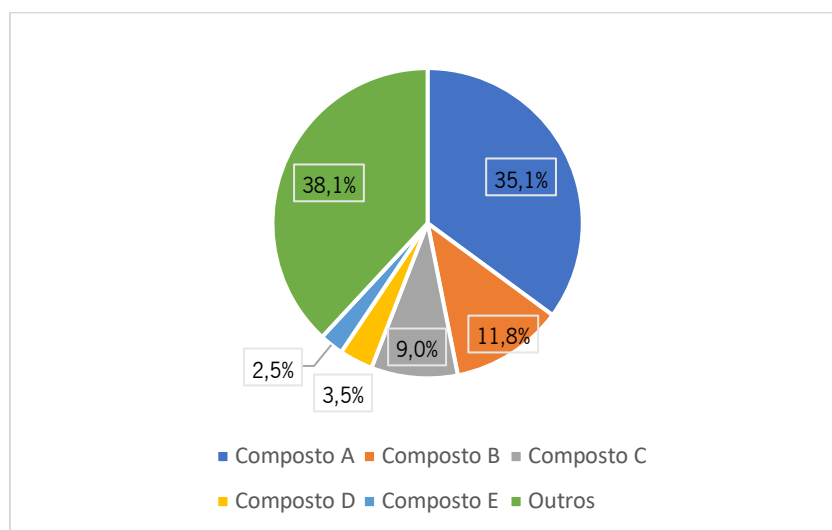


Figura 3.2. Caracterização dos dados colhidos por compostos anotados. Composto A: polimixina B1; composto B: Mistura de polimixina B1 e polimixina B2 (R -Dab-Thr-Dab-Dab(1)-D-Dab-D-Phe-Leu-Dab-Dab-Thr-(1). R -Dab-Thr-Dab-Dab(2)-D-Dab-D-Phe-Leu-Dab-Dab-Thr-(2)), onde R é o ácido gordo ligado ao terminal N da cadeia peptídica; composto C: H-Ala-Ala-Arg-Ile-Ile-Leu-Arg-Thr-Arg-Phe-Arg-NH₂; composto D: H-Phe-Leu-Gln-Leu-Ile-Gly-Arg-Val-Leu-Ser-Gly-Ile-Leu-NH₂; composto E: ciclo[Ala-Ser-Pro-D-Thr-Pro-Phe-Ile].

De acordo com a Figura 3.2, verifica-se que cinco compostos (representados na Figura 3.3) foram mais proeminentes, de entre eles a polimixina B1 foi a estudada, seguida de um análogo da polimixina B que corresponde a uma mistura de duas polimixinas B (polimixina B1 e polimixina B2) (Figura 3.3b).

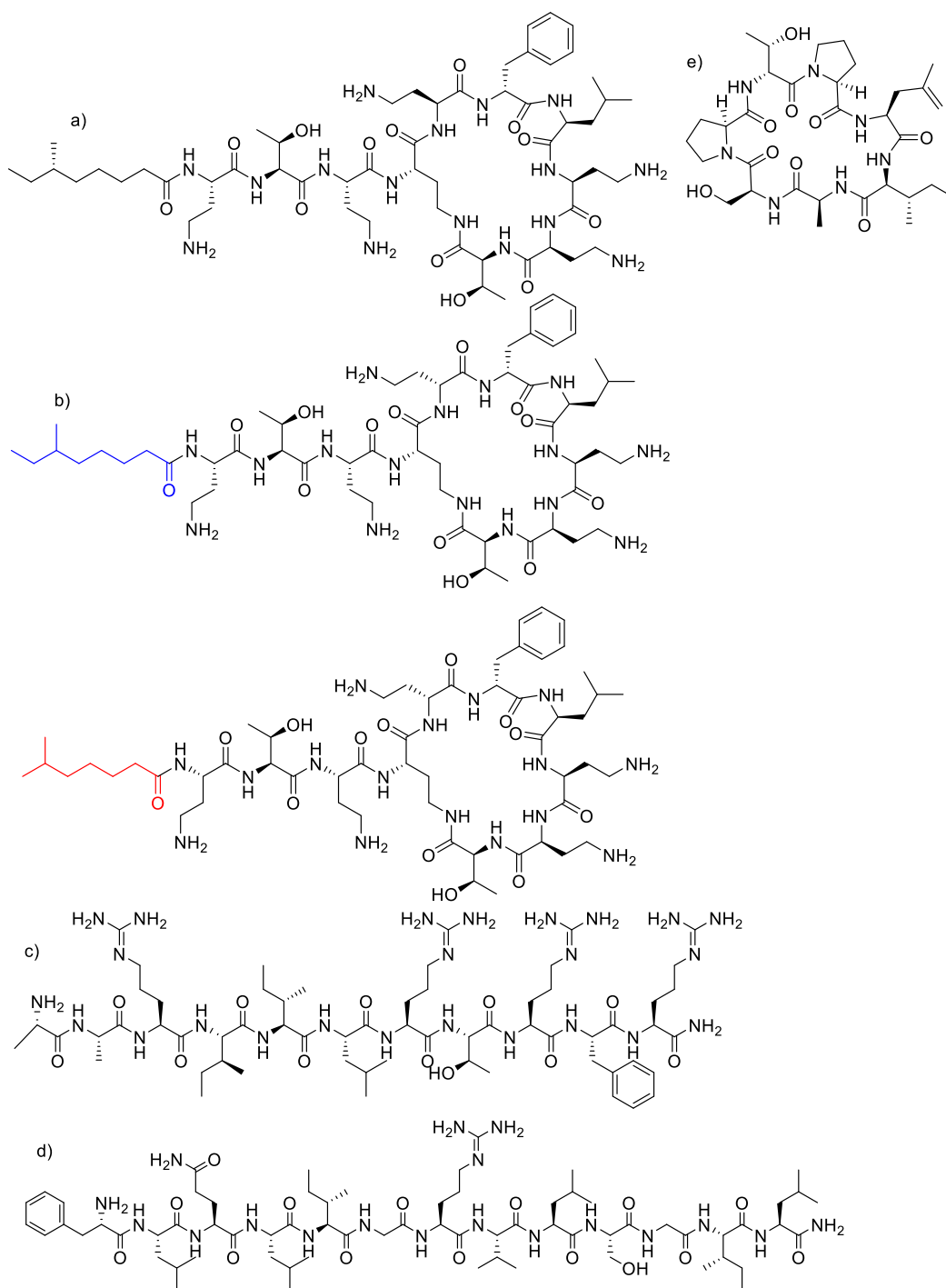


Figura 3.3. Estruturas químicas dos compostos mais representativos no conjunto de dados: polimixina B1 (a), mistura de polimixina B1 e polimixina B2 (R-Dab-Thr-Dab-Dab(1)-D-Dab-D-Phe-Leu-Dab-Dab-Thr-(1).R-Dab-Thr-Dab-Dab(2)-D-Dab-D-Phe-Leu-Dab-Dab-Thr-(2)) (b), onde se destaca a azul o ácido gordo 6-metiloctanolilo na polimixina B1 e a vermelho o ácido gordo 6-metileptatanolilo na polimixina B2, H-Ala-Ala-Arg-Ile-Ile-Leu-Arg-Thr-Arg-Phe-Arg-NH₂ (c), H-Phe-Leu-Gln-Leu-Ile-Gly-Arg-Val-Leu-Ser-Gly-Ile-Leu-NH₂ (d) e ciclo[L-alaniil-L-seriil-L-proliil-D-treoniil-L-proliil-L-fenilalaniil-L-isoleuciil] (e).

3.1.1. Distribuição dos valores da MIC

Os valores da MIC colhidos variam entre 0,006 μM e 256 μM , com média de $26,2 \pm 43,5 \mu\text{M}$. Estes dados são bastante assimétricos (como pode ser visto na Figura 3.4). Os valores da MIC para as bactérias Gram-negativas variam entre 0,006 μM e 256 μM e têm o seu pico modal num valor baixo (2 μM), o que indica maior suscetibilidade deste tipo de microrganismo a polimixinas. O limite máximo dos valores da MIC para esta classe de microrganismo provavelmente é derivado dos ensaios realizados com bactérias resistentes a polimixinas. Por outro lado, os valores da MIC para bactérias Gram-positivas variaram entre 0,096 μM e 128 μM , e têm o seu pico modal por volta de 16 μM . Os valores acima de 120 μM foram comuns a vários ensaios realizados com fungos (com moda de 128 μM) e refletem provavelmente o limite máximo de concentração usada nos mesmos.

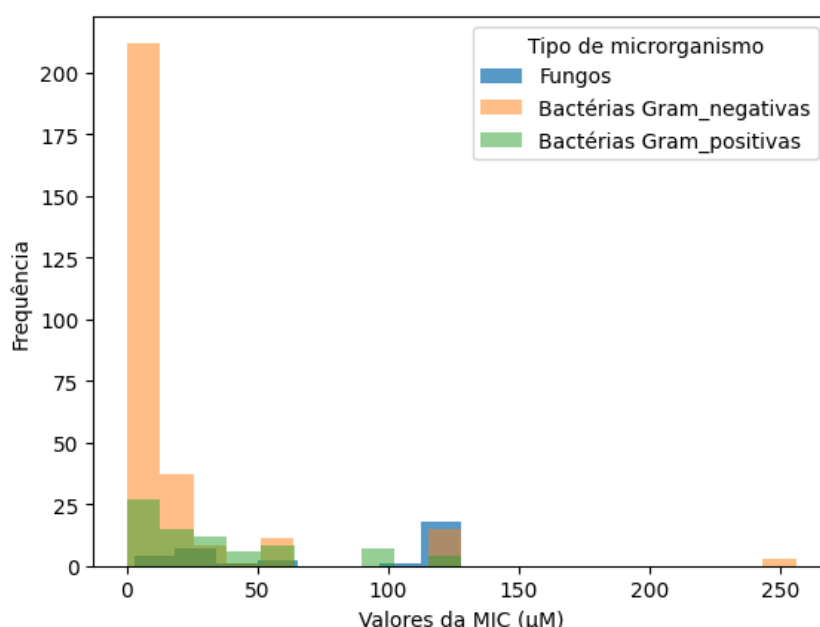


Figura 3.4. Histograma da distribuição dos valores da MIC por MTyp.

3.1.2. Categorização dos quartis da MIC

Este trabalho teve como objetivo original o desenvolvimento de um modelo de regressão capaz de relacionar quantitativamente a estrutura química e a atividade biológica, mas os cálculos preliminares com vista num modelo de regressão dos valores da MIC falharam, levando a considerar uma abordagem semiquantitativa baseada nos quartis dos valores da MIC relatados para um determinado par composto/alvo. Esta abordagem não só permitiu fazer uma avaliação semiquantitativa da atividade

inibitória de novos compostos, mas também garantiu que as diferentes categorias da variável-alvo fossem igualmente representadas nos dados.

A Tabela 3.1 apresenta a separação dos valores da MIC dos dados obtidos em quartis. Os valores da MIC encontram-se distribuídos de forma bastante assimétrica e com uma distância considerável entre a mediana e o Q3. Isto pode ser parcialmente atribuído aos valores da MIC determinados em ensaios com fungos.

Tabela 3.1. Separação dos quartis da MIC.

Quartil	Intervalo da MIC
Q1	$MIC < 1,25 \mu M$
Q2	$1,25 \mu M \leq MIC < 4 \mu M$
Q3	$4 \mu M \leq MIC < 32 \mu M$
Q4	$MIC \geq 32 \mu M$

3.2. Caracterização dos modelos da primeira série.

No total, foram treinados 30 modelos combinando cada um de três algoritmos (DT, RF e AdaBoost) com as 10 famílias de descritores moleculares (Tabela 2.1). Todos os modelos foram treinados visando a posição do quartil de cada entrada no conjunto de dados, com o objetivo final de entender quais modificações na estrutura da polimixina B produziriam maior atividade antimicrobiana. Uma vez que quanto menor for a MIC de um composto, maior é a sua atividade antimicrobiana, e levando em consideração o *breakpoint* da colistina definido para *Pseudomonas aeruginosa*, que é de 4 mg/dm³ (aproximadamente 4 μM), uma classificação de Q1 ou Q2 para uma nova estrutura fará dela um candidato promissor para síntese e testes *in vitro*, sendo os mais promissores aqueles classificados como Q1.

Por causa disso, a avaliação de cada modelo levou em consideração não apenas os valores da exatidão (equação 1.4) para os conjuntos treino e teste, mas também a taxa de verdadeiros positivos para Q1 (ou seja, a fração de casos em que um composto foi corretamente previsto como Q1, ou $f(Q1 | Q1)$). Além disso, avaliou-se também as métricas indesejáveis, como $f(Q1 | Q4)$ (compostos muito ativos classificados como pouco ativos) e $f(Q4 | Q1)$ (compostos pouco ativos classificados como muito ativos), dado que essas métricas se traduzem em desperdiçar uma boa estrutura proposta e promover

uma molécula particularmente inativa (pelo menos para o alvo microbiano selecionado) , respectivamente.¹⁵⁷ No entanto, os valores da $f(Q4|Q1)$ foram sempre muito baixos para todos os modelos (Tabela A1 em Anexos) e, portanto, foi excluída de considerações posteriores.

Após a otimização dos hiper-parâmetros, a maioria dos modelos RF requereram florestas relativamente pequenas (n_{est} entre 10 e 25), o que é adequado considerando o número de entradas. A maioria dos modelos favoreceu o uso de pelo menos 50 % das características disponíveis para cada árvore (n_f entre 0,58 e 1) e o modelo usando a família de descritores CKP apresentou a melhor exatidão de validação cruzada para $n_f = 0,16$ no conjunto de dados. Por outro lado, a maioria dos modelos RF optaram por em cada árvore considerar toda fração dos dados apresentados, n_{s_i} entre 0,78 e 1 (Tabela A2 em Anexo).

Os modelos AdaBoost exibiram uma preferência semelhante aos modelos RF para valores menores de n_{est} (entre 5 e 10), com exceção do modelo usando o conjunto de descritores SLogP_VSA, que exigiu $n_{est} = 50$. Cada uma das árvores do modelo AdaBoost foi geralmente limitada a uma profundidade máxima (d_{est}) de 10, com exceção dos modelos AdaBoost usando a família de descritores FG ($d_{est} = 2$), bem como dos modelos com as famílias de descritores SLogP_VSA e AC2D, que atingiram exatidão máxima para $d_{est} = 100$ (Tabela A2 em Anexo).

Os resultados representados na Figura 3.5 mostram o comportamento dos modelos considerados neste trabalho através das métricas detalhadas acima. No geral, os modelos RF apresentaram o pior desempenho, mostrando um comportamento de *overfitting* muito acentuado (Figura 3.5b). Os modelos DT são os que apresentaram os valores das métricas desejáveis mais baixos, em comparação com os outros modelos, um comportamento de *overfitting* considerável, bem como as $f(Q1|Q1)$ relativamente baixas nos dados de teste (Figura 3.5a). Os modelos AdaBoost apresentaram os melhores resultados, mostrando-se com um *overfitting* relativamente baixo, comparado aos outros modelos, e com valores da exatidão e da $f(Q1|Q1)$ elevados e os valores de $f(Q1|Q4)$ baixos (Figura 3.5c).

Com relação ao desempenho de cada família de descritores moleculares, a fração de grupos funcionais (FG) destacou-se negativamente ao apresentar um *overfitting* significativo nos métodos DT e RF, afetando a métrica $f(Q1|Q1)$. Além disso, FG foi a família de descritores com pior desempenho para os modelos AdaBoost. Da mesma forma, a família Hb também produziu alguns dos modelos mais fracos,

geralmente aumentando os valores de $f(Q1|Q4)$ nos conjuntos treino e teste, independentemente do algoritmo usado (Figura 3.5).

O desempenho dos conjuntos de descritores derivados das contribuições da área de superfície (descritores baseados em VSA) variou significativamente entre os algoritmos, mas, em geral, essas famílias de descritores tiveram um bom desempenho, embora com uma tendência geral de aumentar significativamente os valores de $f(Q1|Q4)$ no conjunto teste. Destes, o modelo AdaBoost/SLogP_VSA destacou-se como o melhor com uma exatidão alta, uma $f(Q1|Q1)$ alta e uma $f(Q1|Q4)$ baixa (Figura 3.5c).

Por sua vez, as famílias de descritores com raízes topológicas (CPK, AD2D e BCUT2D) tiveram um bom desempenho ao se observar uma exatidão e uma $f(Q1|Q1)$ elevadas e uma $f(Q1|Q4)$ baixa, em paridade com os encontrados ao usar os descritores baseados em VSA. Caso digno de nota é o modelo AdaBoost/CKP que mostrou uma exatidão geral aceitável, valores da $f(Q1|Q1)$ altos e valores de $f(Q1|Q4)$ muito baixos. Isso levou a selecionar-se este modelo para estudos adicionais.

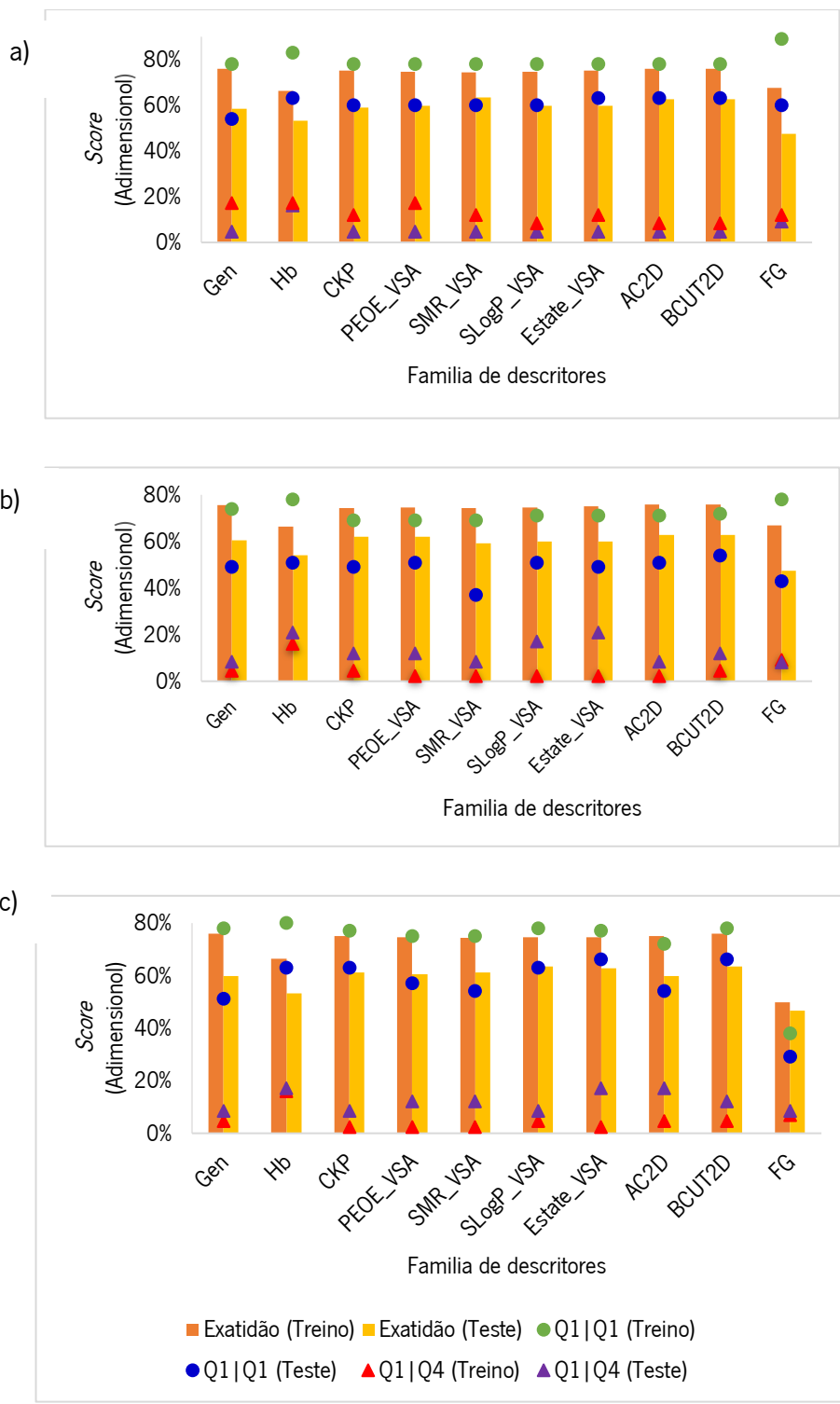


Figura 3.5. Valores de diferentes métricas (exatidão, $f(Q1|Q1)$ e $f(Q1|Q4)$) para cada família de descritores e algoritmos: DT (a); RF (b); e AdaBoost (c).

3.3. Caracterização do melhor modelo da 1ª série

O modelo AdaBoost/CKP apresentou o melhor desempenho: uma exatidão aceitável (75,1 % e 61,3 % nos conjuntos de treino e teste, respetivamente), valores da $f(Q1|Q1)$ elevados (77 % e 63 %, nos conjuntos de treino e teste, respetivamente) e valores da $f(Q1|Q4)$ baixos (2,3 % e 8,3 % nos conjuntos de treino e teste, respetivamente). Por isso, foi considerado como o melhor modelo da primeira série.

3.3.1. Desempenho do modelo AdaBoost/CKP

Como referido anteriormente, o melhor modelo é caracterizado pela sua qualidade preditiva (alta taxa de verdadeiros positivos e baixa taxa de falsos positivos e negativos). De acordo com as matrizes de confusão do modelo AdaBoost/CKP, ilustradas na Figura 3.6, verifica-se que este apresenta uma fração de verdadeiros positivos acima de 50 %, com exceção de Q4|Q4 no conjunto teste, o que é irrelevante, uma vez que não se levou em consideração essa métrica por não haver interesse nos compostos classificados como Q4. O modelo apresenta valores ligeiramente elevados de falsos positivos e negativos de quartis próximos, como por exemplo uma $f(Q1|Q2)$ (Q1 previsto como Q2) de 20 % e 23 % nos conjuntos teste e treino, respetivamente, e uma $f(Q2|Q1)$ (Q2 previsto como Q1) de 20 % e 17 % nos conjuntos treino e teste, respetivamente. Isto não constitui uma grande preocupação, uma vez que os compostos classificados como Q2 também são considerados promissores, pelo que não haverá maior probabilidade de sintetizar compostos não promissores e desperdiçar compostos promissores.

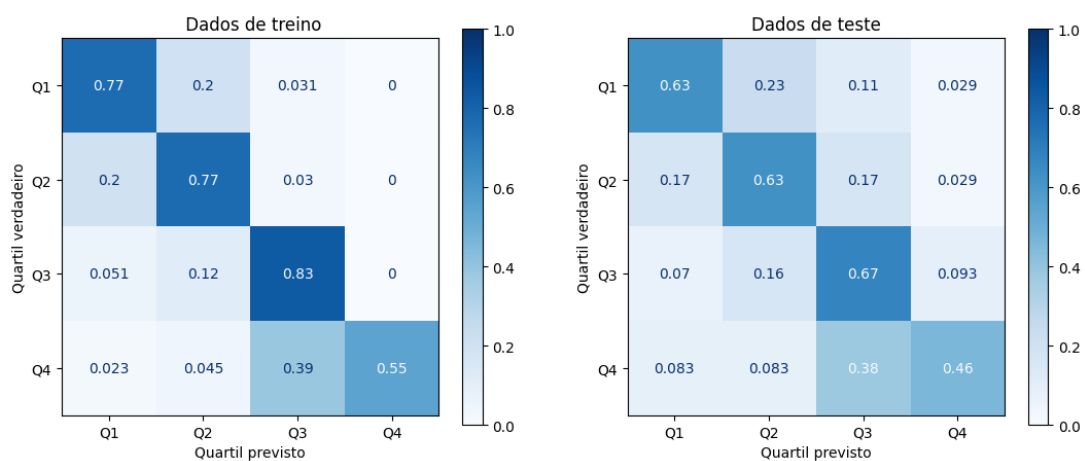


Figura 3.6. Matriz de confusão dos conjuntos treino e teste do modelo AdaBoost/CKP.

3.3.2. Importância cada variável no modelo por permuta

A importância relativa de cada variável considerada no modelo AdaBoost/CKP foi avaliada através da importância por permuta (PI, do inglês *Permutation Importance*), no qual o peso de cada variável é aferido com base na mudança do resultado do modelo quando esta é substituída por dados aleatórios com a distribuição observada nos dados de treino.¹⁵⁸

A importância por permuta de cada variável no modelo AdaBoost/CKP é representada na Figura 3.7 e sugere que o modelo é particularmente sensível a cinco variáveis: duas variáveis que descrevem o microrganismo ($T_{\chi G}$ e M_{typ}), bem como a três descritores moleculares (${}^1\chi$, ${}^0\chi$ e ${}^3\kappa$). Essas cinco variáveis representam 71 % da PI total.

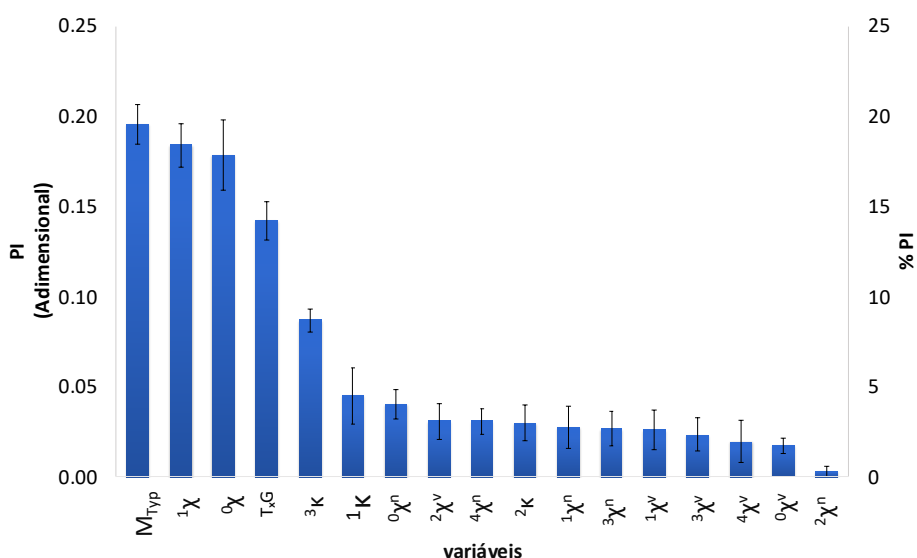


Figura 3.7. Importância de Permutação (PI) média dos descritores no modelo AdaBoost/CPK calculada usando 10 réplicas do conjunto de dados para cada recurso. As barras de erro representam o desvio padrão da PI nas 10 réplicas e o eixo dos yy à direita indica o valor da PI média, normalizado para percentagem.

3.3.3. Influência dos descritores moleculares

Do ponto de vista dos descritores moleculares, a resposta do modelo é dominada por ${}^1\chi$, ${}^0\chi$ e, em menor grau, por ${}^3\kappa$. Estes três descritores moleculares reúnem cerca de 40 % da PI total. Para mostrar o efeito que estas variáveis têm no resultado previsto do modelo, foram gerados gráficos de dependência parcial do modelo com esses três descritores (Figuras 3.8), onde é possível observar a

probabilidade de um determinado resultado de classificação (Q1, Q2, Q3 ou Q4) com valores variados do descritor em questão, quando todas as outras variáveis são substituídas por valores aleatórios.

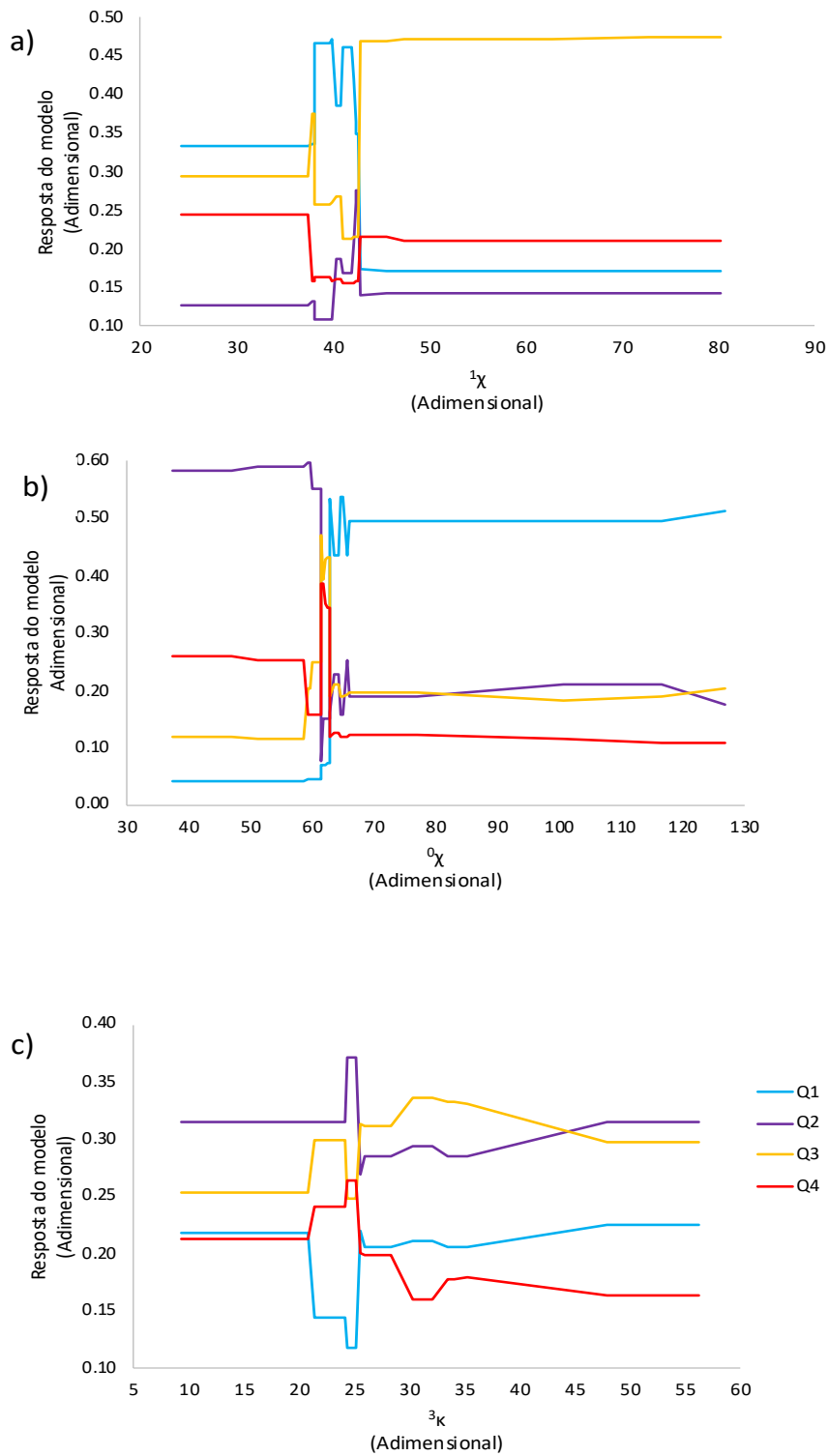


Figura 3.8. Gráficos de dependência parcial do modelo AdaBoost/CKP: em relação a $^1\chi$ (a), em relação a $^0\chi$ (b) e em relação a $^3\kappa$ (c).

O efeito da ramificação pode ser observado quando se considera a dependência parcial em relação a ${}^1\chi$, mostrada na Figura 3.8a. Os valores baixos de ${}^1\chi$ são usados para favorecer a classificação Q1 e valores mais elevados (a partir de 43-45) são usados para favorecer a classificação Q3. Uma vez que o ${}^1\chi$ diminui com as ramificações, a dependência parcial do modelo sugere que quanto mais localmente ramificada for a estrutura, maior será a atividade biológica. O mesmo aspeto da influência da ramificação é verificado na dependência parcial em relação a ${}^0\chi$ (Figura 3.8b), uma vez que os valores do descritor molecular ${}^0\chi$ aumentam com o aumento das ramificações. A Figura 3.8b revela que valores elevados de ${}^0\chi$ (superior a 65) favorecem a classificação Q1. Tal como em ${}^1\chi$, há uma zona de transição bastante caótica entre os dois regimes. Por outro lado, o descritor molecular ${}^3\kappa$ faz a caracterização da forma da molécula (codifica a informação sobre a centralidade da ramificação). Os valores do descritor ${}^3\kappa$ diminuem quanto mais localmente globular for a molécula e são maiores quando a ramificação é inexistente ou quando está localizada nas extremidades da estrutura. O gráfico da dependência molecular deste descritor molecular (Figura 3.8c) parece não esclarecer a relação entre a forma da molécula e a atividade antimicrobiana, mas sugere que os valores baixos de ${}^3\kappa$ favorecem a classificação Q2 e a probabilidade do modelo classificar uma dada estrutura como Q1 aumenta quando as ramificações estão localizadas nas extremidades (Figura 3.8c).

3.3.4. Influência do alvo biológico

Tal como descrito na Figura 3.7, das 5 variáveis que mais influenciam a resposta do modelo, duas ($M_{T_{yp}}$ e T_xG) descrevem o microrganismo, tendo um peso de 30,5 % na PI total. Destas, a característica mais importante foi $M_{T_{yp}}$, com uma PI de 17,7 %. Este resultado reflete a estrutura dos dados discutida na seção 3.1, em que os valores da MIC dos fungos e das bactérias Gram-positivas foram significativamente maiores do que os das bactérias Gram-negativas. Foram gerados gráficos de dependência parcial que descrevem a resposta do modelo em relação a $M_{T_{yp}}$ (Figura 3.9 e em relação a T_xG (Figura 3.10). Neste último, foram considerados os géneros mais representativos no conjunto de dados.

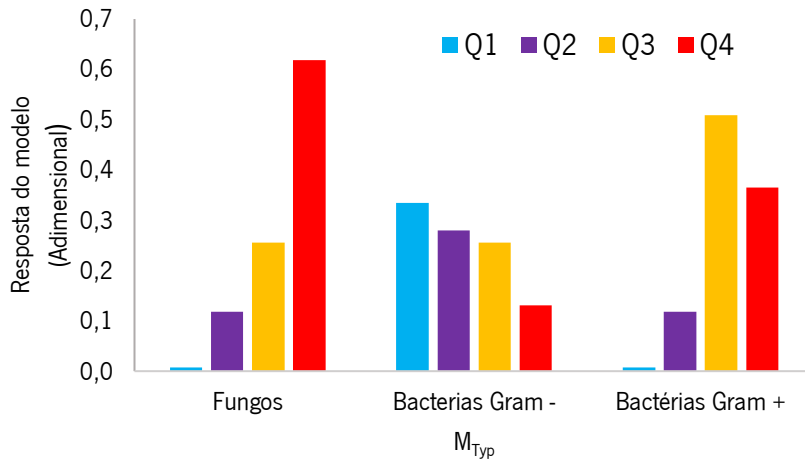


Figura 3.9. Gráfico de dependência parcial do modelo AdaBoost/CPK por M_{Typ} .

Em relação à resposta do modelo usando a variável M_{Typ} , a Figura 3.9 revela que os ensaios usando as bactérias Gram-negativas favorecem a classificação Q1, ao passo que as bactérias Gram-positivas e fungos favorecem a classificação Q3 e Q4. A resposta do modelo segue o conhecimento empírico de que as polimixinas são particularmente eficazes em bactérias Gram-negativas, daí o modelo dar prevalência a Q1 e Q2 para esta categoria. A parede celular e membrana plasmática dos fungos são bastante diferentes das bactérias. Os fungos são seres eucariotas que têm uma parede celular rígida composta por quitina, glicana e uma membrana plasmática em que o ergosterol (esterol) é o principal componente, ao passo que uma bactéria é um ser procarionte que tem uma parede celular constituída por proteínas, lípidos e peptidoglicanos e uma membrana constituída por fosfolípidos e proteínas,¹⁵³ pelo que o modelo penaliza o emparelhamento de derivados das polimixinas com fungos, atribuindo-lhe uma maior propensão para Q3 e especialmente para Q4.¹⁵³

A variável T_xG é a quarta mais influente no modelo, com uma PI de 12,8%. O gráfico da dependência parcial do modelo usando esta variável é mostrado na Figura 3.10, e este revela que o modelo parece estar a usar M_{Typ} como um primeiro filtro para classificar os dados recebidos e T_xG como um filtro secundário.

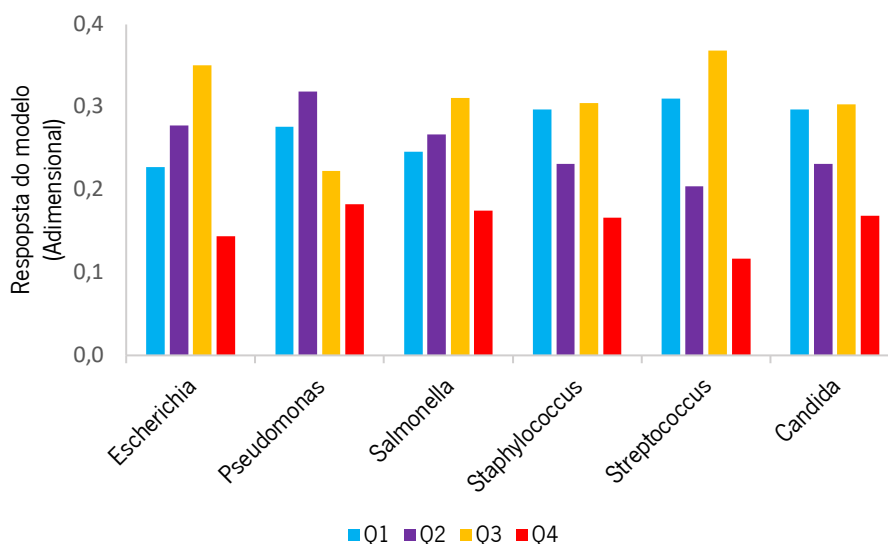


Figura 3.10. Gráficos de dependência parcial do modelo AdaBoost/CPK por T,G.

De facto, os dados da dependência parcial ilustrados na Figura 3.10 sugerem que, apesar da tendência do modelo de classificar os ensaios realizados com bactérias Gram-negativas como Q1, aqueles que têm como alvo *Escherichia* e *Salmonella* sofreram uma correção no sentido de uma classificação Q3.

3.4. Suscetibilidade da *Shigella sonnei*, *Proteus mirabilis* e *Listeria monocytogenes* à colistina e à polimixina B

Ao analisar os dados extraídos da PubChem, foram identificados 7 géneros de microrganismos (*Priestia*, *Yersinia*, *Enterobacter*, *Cryptococcus*, *Shigella*, *Vibrio* e *Proteus*) com apenas 1 entrada e 5 géneros (*Listeria*, *Micrococcus*, *Erwinia*, *Trichophyton* e *Saccharomyces*) com 2 entradas, que poderiam enviar o modelo. Das que foram identificadas, foi possível selecionar as espécies bacterianas *S. sonnei*, *P. mirabilis* e *L. monocytogenes*, pois eram as que estavam disponíveis na coleção do CEB, e a *E. coli* foi usada como controlo para validar os ensaios da MIC realizados em *Enterobacterales*, conforme sugerido pela EUCAST.¹⁵⁹ Apesar de não ter sido possível usar a *Streptococcus pneumoniae*, espécie usada no controlo de qualidade para *L. monocytogenes*,¹⁵⁹ neste ensaio, foi possível deduzir a validade do ensaio no geral, uma vez que o controlo para as outras duas espécies (*E. coli*) foi válido, ao se verificar que o valor da MIC da colistina obtido no ensaio de suscetibilidade para esta espécie encontra-se dentro do intervalo estabelecido pelo Instituto de Padrões Clínicos e Laboratoriais, e validado pela EUCAST (0,25 mg/dm³ a 2 mg/dm³).¹⁶⁰ Na Tabela 3.2, encontram-se os valores das concentrações da colistina e

polimixina B para as quais foi possível verificar a inibição do crescimento (MIC) após 24 horas de incubação.

Tabela 3.2. Valores da MIC da colistina e da polimixina B para *E. coli*, *S. sonnei*, *P. mirabilis* e *L. monocytogenes*.

Estirpes bacterianas	Agentes antimicrobianos			
	Colistina		Polimixina B	
	MIC/ mg/dm ³	MIC/ μ M	MIC/ mg/dm ³	MIC/ μ M
<i>E. coli</i> (ATCC 25922)	0,25	0,2	0,5	0,38
<i>S. sonnei</i> (ATCC 2593)	0,25	0,2	0,5	0,38
<i>P. mirabilis</i> (CECT 4168)	>256	>204,2	>256	>196,7
<i>L. monocytogenes</i> (ATCC 15313T)	16	12,8	6	6,8

Os resultados apresentados na tabela 3.2 mostram que a colistina é mais potente contra *S. sonnei* e *E. coli* comparado à polimixina B. Por outro lado, a polimixina B é mais potente contra *L. monocytogenes* do que a colistina. Estes resultados estão de acordo com os resultados obtidos por Sader *et al*, em que a colistina exibiu valores da MIC duas vezes menores que a polimixina B contra espécies bacterianas mais suscetíveis a polimixinas ($MIC \leq 2 \text{ mg/dm}^3$), e a polimixina B exibiu os valores da MIC mais baixos que a colistina contra espécies bacterianas menos suscetíveis a qualquer polimixina ($MIC \geq 4 \text{ mg/dm}^3$).¹⁶¹ Ambos os agentes antimicrobianos não apresentaram actividade contra *P. mirabilis*, o que era de esperar uma vez que a *P. mirabilis* é naturalmente resistente a polimixinas,^{57,162} e este resultado assemelha-se aos obtidos por Chew *et al*, em que *P. mirabilis* foi resistente a ambas polimixinas.¹⁶³ Estudos relatam que a resistência da *P. mirabilis* às polimixinas, se deve à alteração da estrutura do LPS, devida à presença de alguns genes, como por exemplo *eptC*¹⁶⁴ e *rppA*.¹⁶⁵ Estes genes direcionam a incorporação da fosfoetanolamina no lípido A, reduzindo deste modo as cargas negativas do LPS. De acordo com a EUCAST, as estirpes de *S. sonnei* e *E. coli* utilizadas neste trabalho são classificadas como suscetíveis à colistina, uma vez que os valores registados da MIC foram inferiores ao *breakpoint* definido (2 mg/dm³ para *Enterobacterales*), enquanto que *P. mirabilis* é classificada como resistente, uma vez que o valor de MIC é superior ao *breakpoint* definido (2 mg/dm³ para *Enterobacterales*). Por outro lado,

a EUCAST não apresenta um *breakpoint* da polimixina B para qualquer espécie bacteriana usada neste estudo e de ambas as polimixinas para *L. monocytogenes*.

3.5. Caracterização dos dados da 2ª série

Os dados extraídos da PubChem foram suplementados com novos dados de MIC, constituindo os dados da segunda série. No total, foram introduzidas oito novas entradas provenientes dos dados dos ensaios laboratoriais de suscetibilidade de 4 espécies bacterianas a 2 compostos (sulfato de polimixina B e sulfato de colistina), de acordo com a tabela 3.2. Para além da introdução de novos dados provenientes dos ensaios laboratoriais, foram também introduzidas 6 novas entradas (com novas espécies de *Salmonella* e *Pseudomonas*) provenientes de uma revisão dos dados excluídos durante a cura inicial dos dados. Após a adição destes novos dados, obteve-se um conjunto de 413 entradas, constituído por 58 compostos e 42 espécies de microrganismos distribuídas em 24 géneros taxonómicos. A Figura 3.11 caracteriza o novo conjunto de dados por M_{typ} (tipo de microrganismos) e T_xG (Género taxonómico do microrganismo).

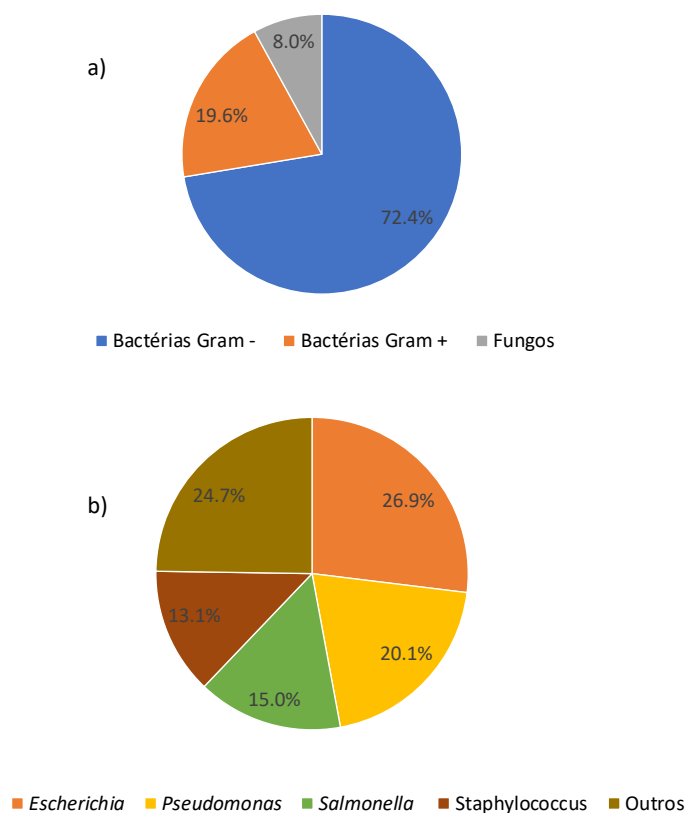


Figura 3.11. Caracterização de novos dados por M_{typ} (a) e por T_xG (b).

Como referido anteriormente, foram usadas 3 bactérias Gram-negativas e uma bactéria Gram-positiva nos ensaios laboratoriais, e foram introduzidas duas espécies através do processo de cura dos dados para suplementar os dados extraídos da PubChem. Isto aumentou o peso das bactérias Gram-negativas no conjunto de dados e reduziu ligeiramente o peso de bactérias Gram-positivas e fungos (Figura 3.11a, Cf. Figura 3.1a). Com a introdução de novos dados, houve uma ligeira redução do peso dos 4 géneros mais predominantes (Figura 3.11b, Cf Figura 3.1b) devido ao uso de géneros com menos entradas no conjunto de dados iniciais nos ensaios laboratoriais.

Assim como na primeira série, foi feita a caracterização dos novos dados por compostos usados, onde verifica-se que a introdução de novos dados não alterou significativamente o peso dos cinco compostos mais proeminentes (Figura 3.12 Cf Figura 3.2).

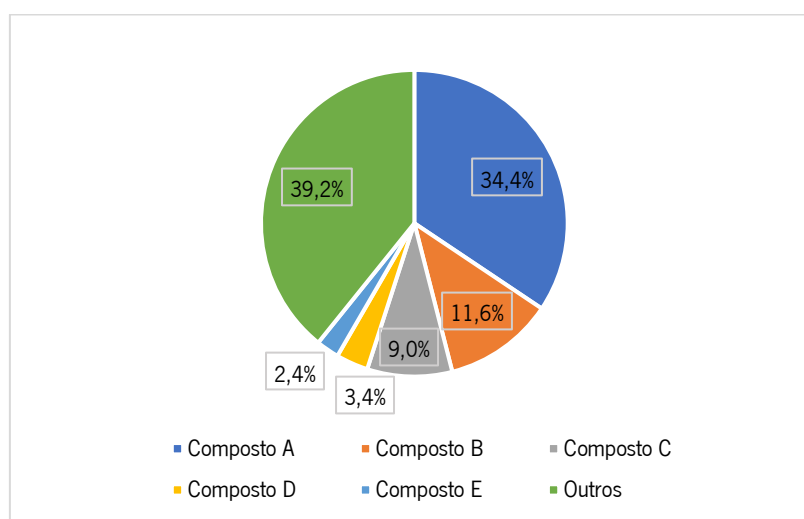


Figura 3.12. Caracterização de novos dados por compostos anotados. Composto A: polimixina B1; composto B: Mistura de polimixina B1 e polimixina B2 ($R\text{-Dab-Thr-Dab-Dab}(1)\text{-D-Dab-D-Phe-Leu-Dab-Dab-Thr}(1)$. $R\text{-Dab-Thr-Dab-Dab}(2)\text{-D-Dab-D-Phe-Leu-Dab-Dab-Thr}(2)$), onde R é o ácido gordo ligado ao terminal N da cadeia peptídica; composto C: $H\text{-Ala-Ala-Arg-Ile-Ile-Leu-Arg-Thr-Arg-Phe-Arg-NH}_2$; composto D: $H\text{-Phe-Leu-Gln-Leu-Ile-Gly-Arg-Val-Leu-Ser-Gly-Ile-Leu-NH}_2$; composto E: $\text{ciclo}[\text{Ala-Ser-Pro-D-Thr-Pro-Phe-Ile}]$.

3.5.1. Distribuição dos valores da MIC

Os dados introduzidos não alteraram as distribuições dos valores da MIC, mas é notório o ligeiro aumento dos valores elevados da MIC para bactérias Gram-negativas (Figura 3.13, Cf Figura 3.4). Isto deveu-se ao uso de uma bactéria naturalmente resistente nos ensaios laboratoriais.

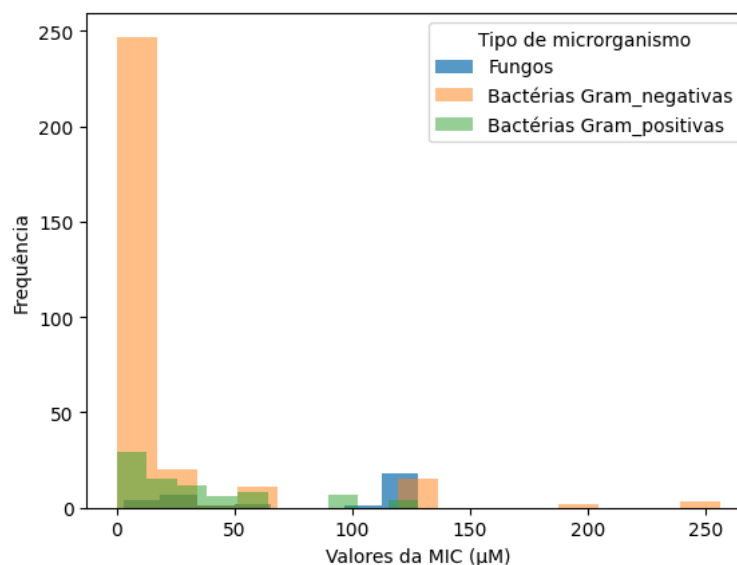


Figura 3.13. Histograma da distribuição dos valores da MIC por MTyp.

3.5.2. Categorização dos quartis da MIC

A Tabela 3.3 revela que a introdução de novos dados apenas alterou a localização do primeiro quartil (Cf. Tabela 3.1).

Tabela 3.3 Separação dos quartis da MIC para o conjunto de dados da segunda série

Quartil	Intervalo da MIC
Q1	MIC < 1 µM
Q2	1 µM ≤ MIC < 4 µM
Q3	4 µM ≤ MIC < 32 µM
Q4	MIC ≥ 32 µM

3.6. Caracterização dos modelos da 2ª série

Tal como na primeira série, foram treinados 30 modelos combinando cada um dos três algoritmos (DT, RF e AdaBoost) com 10 famílias de descritores moleculares, e foram avaliadas as métricas exatidão, $f(Q1|Q1)$ e $f(Q1|Q4)$ dos conjuntos treino e teste de modo a avaliar o impacto da introdução dos novos dados da MIC. Novamente, verificou-se que a métrica $f(Q4|Q1)$ permaneceu muito baixa para todos os modelos.

Após a otimização dos hiper-parâmetros dos dez modelos de RF desenvolvidos, cinco modelos aumentaram o número de árvores na floresta (n_{est}) e dois sofreram uma redução do n_{est} . Destes dez modelos, quatro (Hb, CKP, SLogP_VSA e FG) requereram uma floresta relativamente grande ($n_{est} = 100$) e seis requereram uma floresta relativamente pequena (n_{est} entre 10 e 20). Nesta série, a maioria dos modelos apresentou um aumento do número de características disponíveis para cada árvore (n_f) e o modelo usando a família de descritores FG apresentou a melhor pontuação de exatidão da validação cruzada para $n_f = 0,16$, seguido da família de descritores BCUT2D com $n_f = 0,47$ (Tabela A3 em Anexos).

Contrariamente aos modelos da primeira série, nos modelos AdaBoost da segunda série, quatro famílias de descritores (Gen, Hb, CKP e AC2D) exibiram uma preferência por valores maiores de n_{est} (entre 60 e 100) ao passo que seis famílias de descritores (PEOE_VSA, SMR_VSA, SlogP_VSA, Estate_VSA e FG) exigiram valores menores de n_{est} (entre 5 e 20). Em relação à profundidade dos estimadores da base, cada árvore foi limitada a uma profundidade máxima de cinco, com exceção dos modelos usando as famílias de descritores Hb e Estate_VSA ($d_{est} = 2$), assim como nos modelos usando as famílias de descritores PEOE_VSA, SMR_VSA e SlogP_VSA ($d_{est} = 10$) (Tabela A3 em Anexos).

No geral, a introdução de novos dados da MIC melhorou significativamente o desempenho dos modelos desenvolvidos, com um aumento da exatidão e de $f(Q1 | Q1)$ e uma diminuição de $f(Q1 | Q4)$, sendo também notória a redução do problema de *overfitting* (Figura 3.14, Cf. Figura 3.4).

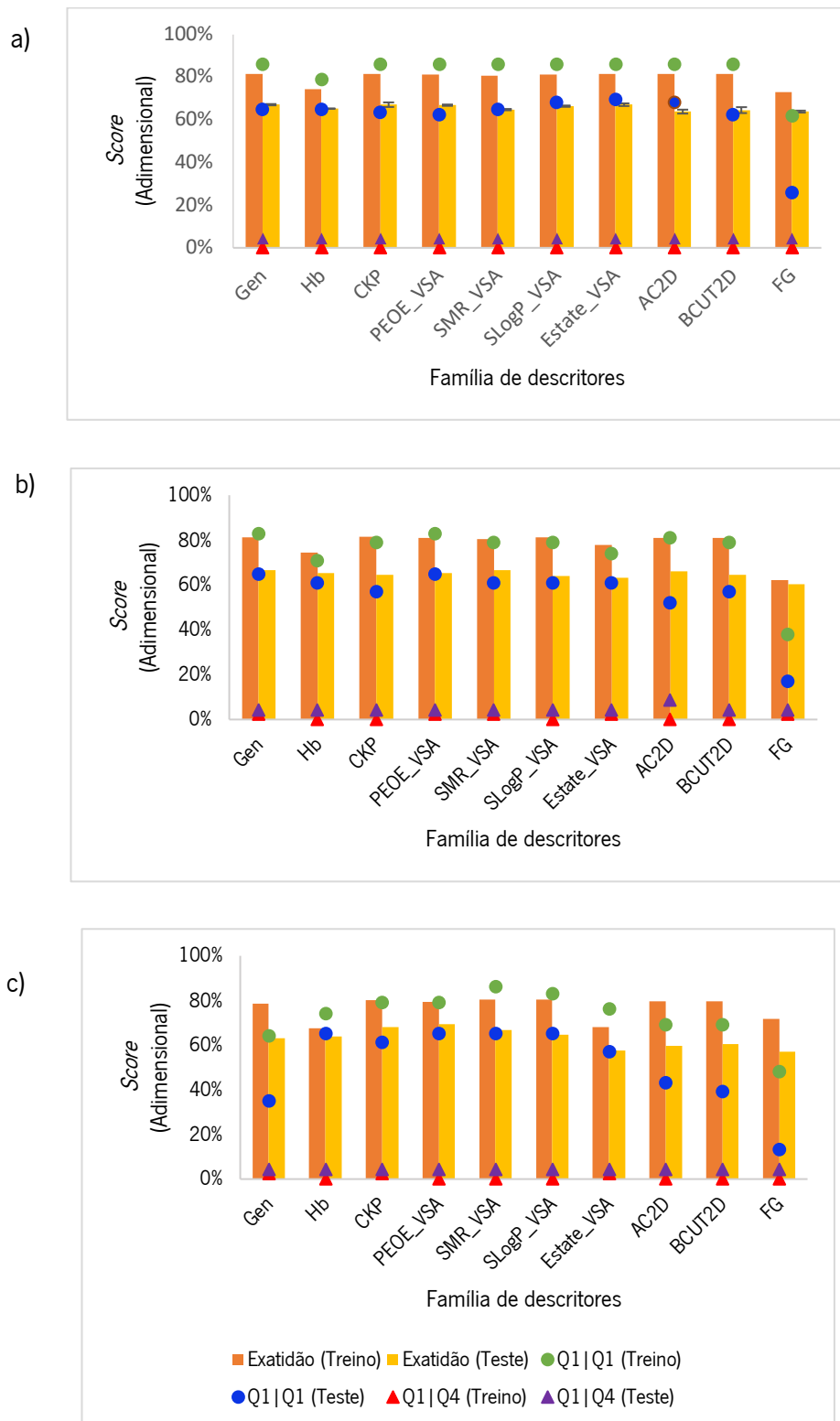


Figura 3.14. Valores de diferentes métricas (exatidão, $f(Q1|Q1)$ e $f(Q1|Q4)$) para cada família de descritores e algoritmos: DT (a), RF (b), e AdaBoost (c). As barras de erro representam o desvio-padrão dos scores.

Assim como na primeira série, os modelos RF sofreram bastante do problema de *overfitting* (Figura 3.14b). No geral, os modelos de conjunto (RF e AdaBoost) sofreram bastante com o problema de *overfitting*, principalmente ao considerar os valores da $f(Q1|Q1)$ (Figuras 3.14b e 3.14c). Isto provavelmente advém do número reduzido de dados, porque estes métodos requerem número elevado de dados. Nesta série, os modelos AdaBoost destacaram-se por apresentarem as exatidões mais elevadas (Figura 3.14c) e os modelos de DT destacaram-se por apresentarem valores da $f(Q1|Q1)$ mais elevados (Figura 3.14a). Vale a pena ressaltar que, ao treinar os modelos da DT, verificou-se que, ao repetir o treino, os valores de todas as métricas alteraram no conjunto teste. Por esta razão, os modelos foram treinados três vezes e foram anotadas as médias e os desvios-padrão para cada uma das métricas (Figura 3.14a). Apesar de haver alteração dos valores da $f(Q1|Q1)$ nos modelos DT, estas são pouco significativas como pode ser observado nos valores do desvio-padrão apresentados como barras de erro na Figura 3.14a.

Em relação ao desempenho de cada família de descritores moleculares, assim como na primeira série, a família dos descritores FG destacou-se negativamente ao apresentar um *overfitting* significativo na exatidão do modelo AdaBoost e de $f(Q1|Q1)$ em todos os modelos. Além disso, FG foi a família de descritores com pior desempenho para todos os modelos, apresentando valores muito baixos de $f(Q1|Q1)$ (Figura 3.14).

As famílias de descritores com raízes topológicas (CPK, AD2D e BCUT2D) apresentaram um desempenho bastante diferenciado, apesar de apresentarem valores mais elevados da exatidão. Particularmente, a família dos descritores CKP apresentou um bom desempenho, sobretudo nos modelos DT e AdaBoost, ao se observarem valores elevados da exatidão e de $f(Q1|Q1)$ e valores baixos de $f(Q1|Q4)$ (Figuras 3.14a e 3.14c). Por outro lado, as famílias AC2D e BCUT2D destacaram-se negativamente nos modelos AdaBoost ao se observar um valor baixo de $f(Q1|Q1)$ no conjunto teste (Figura 3.14c).

Por sua vez, o desempenho das famílias de descritores derivados das contribuições da área de superfície (descritores baseados em VSA) destacaram-se como os melhores em todos os modelos, ao apresentarem valores da exatidão e de $f(Q1|Q1)$ mais elevados. Destes, destacaram-se os modelos Adaboost/PEOE_VSA, com a exatidão mais elevada no conjunto teste, e DT/Estate_VSA, com o valor de $f(Q1|Q1)$ mais elevado no conjunto teste. Por estes motivos, estes foram considerados os melhores modelos da segunda série.

Analisando o desempenho das famílias de descritores moleculares em cada algoritmo, constatou-se que as famílias de descritores Estate_VSA, Gen e PEOE_VSA foram os melhores nos modelos DT, RF e AdaBoost, respetivamente, ao se observarem melhores métricas, nomeadamente valores da exatidão e de $f(Q1|Q1)$ mais elevados, e valores de $f(Q1|Q4)$ mais baixos.

Nesta série, o modelo AdaBoost/PEOE_VSA, o qual apresentou a maior exatidão (69,4 %) no conjunto teste, e o modelo DT/Estate_VSA, o qual foi particularmente bom em identificar os Q1, com um valor da $f(Q1|Q1)$ de aproximadamente 70 % no conjunto teste, destacaram-se como os mais promissores para prever a actividade dos novos análogos da polimixina B.

3.7. Caracterização do melhor modelo da 2ª série

3.7.1. Desempenho dos melhores modelos da segunda série

A qualidade preditiva dos melhores modelos desenvolvidos na segunda série pode ser comprovada pelas matrizes de confusão ilustradas na Figura 3.15, onde é possível verificar os valores elevados dos verdadeiros positivos e valores baixos dos falsos positivos e negativos.

O modelo DT/Estate_VSA foi o melhor a classificar os Q1 (Figura 3.15b), ao passo que o modelo AdaBoost/PEOE_VSA classificou melhor os Q2 (Figura 3.15a). Ambos os modelos apresentaram valores baixos das métricas indesejáveis ($f(Q1|Q4)$ e $f(Q4|Q1)$).

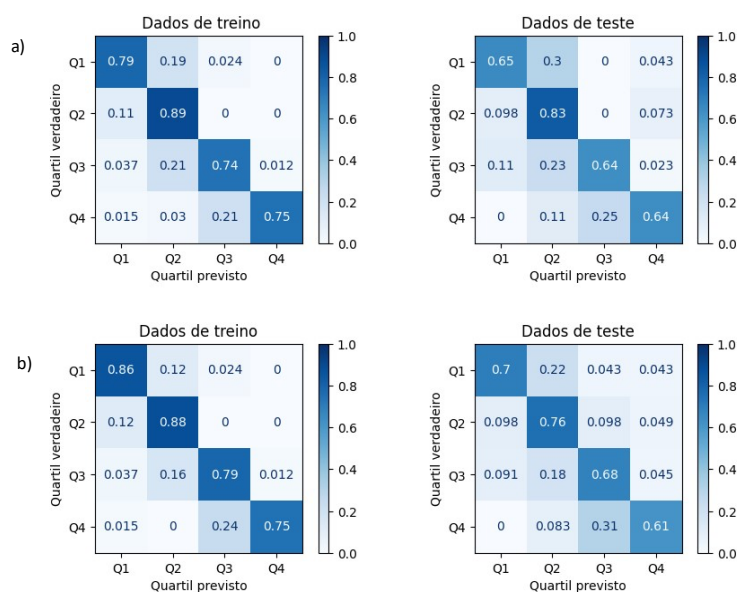


Figura 3.15. Matriz de confusão dos melhores modelos da segunda série: modelo AdaBoost/PEOE_VSA (a), modelo DT/Estate_VSA (b).

3.7.2. Importância de cada variável no modelo por permuta

As duas famílias dos descritores moleculares que se destacaram nos melhores modelos da 2ª série englobam os descritores da área de superfície subdividida que são baseados no cálculo da VSA de um átomo junto com uma outra propriedade, que neste caso foram as cargas parciais calculadas pelo método PEOE e o E-State, que codifica a acessibilidade de elétrons.^{126,127} Como referido anteriormente, as polimixinas são carregadas positivamente e interagem com os LPS das membranas bacterianas (carregadas negativamente) por meio de uma interação eletrostática. Por este motivo, o bom desempenho destes conjuntos de descritores espelham o mecanismo de ação das polimixinas, uma vez que estes conjuntos buscam a acessibilidade de cargas (PEOE_VSA) e a acessibilidade de elétrons (Estate_VSA) nos átomos.¹²⁴

Como referido na subseção 3.3.2, a importância relativa de cada variável considerada num modelo é avaliada por meio da PI. A importância de cada variável nos dois melhores modelos avaliada pela PI é ilustrada na Figura 3.16.

Relativamente ao modelo Adaboost/PEOE_VSA, este foi sensível a cinco variáveis, duas que descrevem o microrganismo ($M_{T_{yp}}$ e T_xG) e três descritores moleculares (PEOE_VSA₇, PEOE_VSA₈ e PEOE_VSA₆). Estas cinco variáveis representaram 82,2 % da PI total. Por outro lado, o modelo DT/Estate_VSA também foi sensível a cinco variáveis, duas que descrevem o microrganismo ($M_{T_{yp}}$ e T_xG) e três descritores moleculares (Estate_VSA₉, Estate_VSA₄ e Estate_VSA₁), representando 73,7 % da PI total.

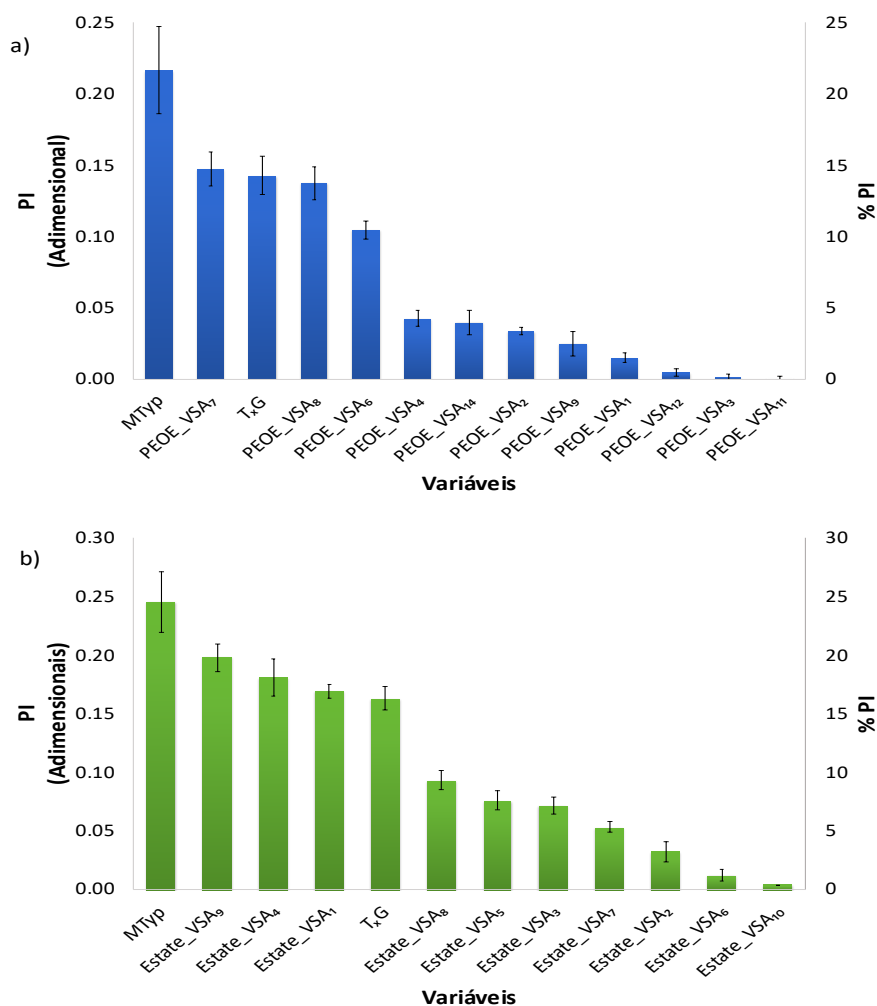


Figura 3.16. Importância de variáveis por permuta dos modelos AdaBoost /PEOE_VSA (a) e DT/ Estate_VSA (b). As barras de erro representam o desvio padrão da PI nas 10 réplicas e o eixo dos yy à direita indica o valor da PI média, normalizado para percentagem.

3.7.3. Influência dos descritores moleculares no modelo Adaboost/PEOE_VSA

Relativamente à contribuição dos descritores moleculares, a resposta do modelo AdaBoost/PEOE_VSA é dominada por PEOE_VSA₇, PEOE_VSA₈ e PEOE_VSA₆. Estes três descritores moleculares reúnem cerca de 42,8% da PI. Para mostrar os efeitos que estas variáveis têm nos resultados previstos do modelo, foram gerados gráficos de dependência parcial do modelo com esses três descritores (Figuras 3.17), onde é possível observar a probabilidade de um determinado resultado de classificação (Q1, Q2, Q3 ou Q4) com valores variados do descritor em questão, quando todas as outras variáveis são substituídas por valores aleatórios.

De acordo com a Figura 3.17, os valores mais baixos do descritor PEOE_VSA₇ são usados para favorecer a classificações Q2, Q3 ou Q4. A probabilidade do modelo identificar um composto com valor baixo do PEOE_VSA₇ (até aproximadamente 120) como sendo Q1 é baixa, e essa probabilidade aumenta a partir do valor 120 (Figura 3.17a). Por outro lado, os valores baixos do descritor PEOE_VSA₈ favorecem a classificação Q3 e servem para distinguir os Q3 dos demais quartis. A partir de 20, o modelo passa a favorecer as classificações Q4 e Q2 (Figura 3.17b). No descritor PEOE_VSA₆, os valores baixos (até aproximadamente 50) são usados para favorecer as classificações Q3 e Q4, valores de 50 a 80 favorecem a classificação Q2, e a partir de 80 o descritor favorece a classificação Q3 (Figura 3.17b). No conjunto destes três descritores moleculares que mais contribuíram no modelo AdaBoost/PEOE_VSA, nenhum é usado para favorecer a classificação Q1, sendo que provavelmente os Q1 são favorecidos por um conjunto de descritores e não apenas um.

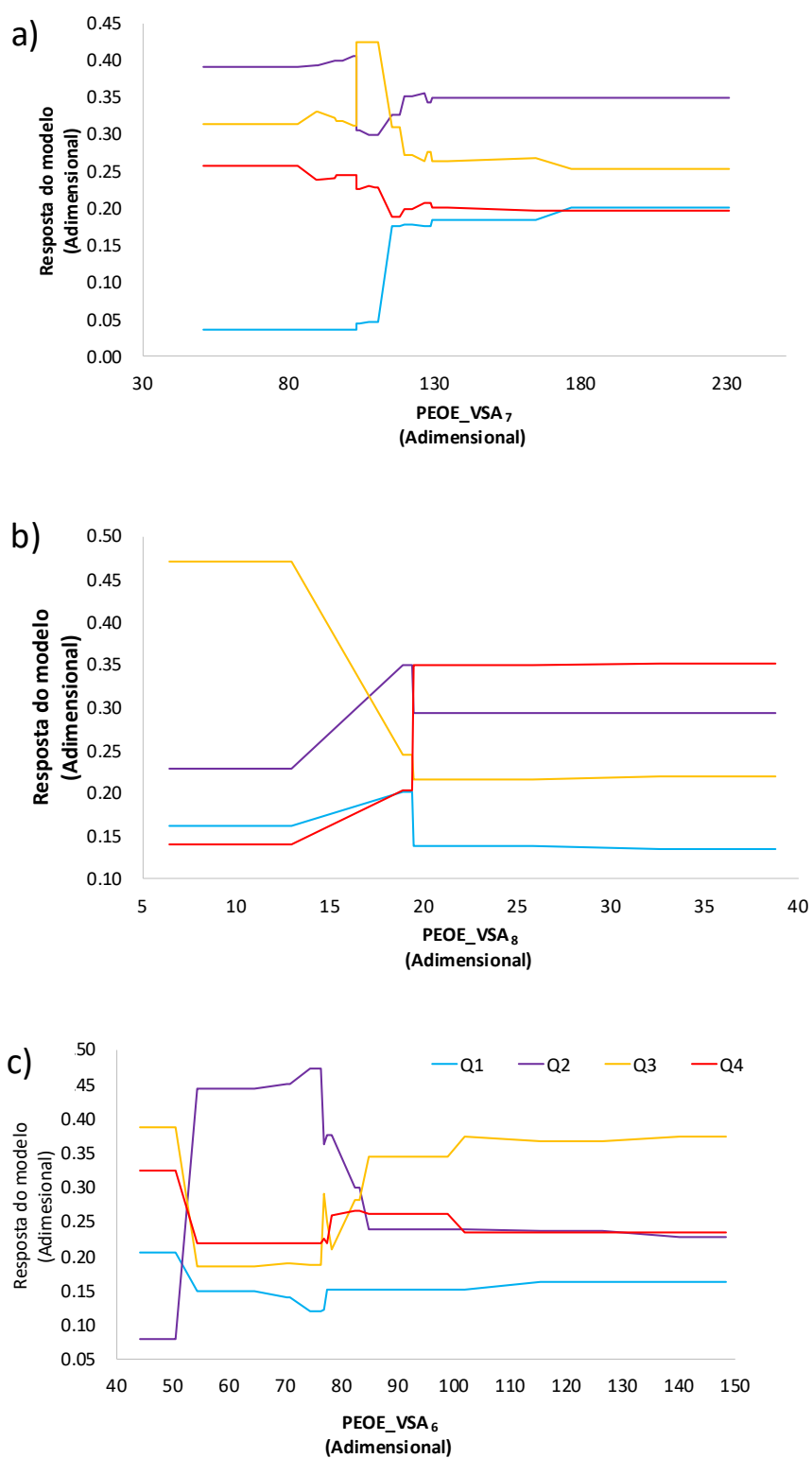


Figura 3.17. Gráfico de dependência parcial do modelo AdaBoost usando o descritor molecular PEOE_VSA₇ (a), PEOE_VSA₈ (b) e PEOE_VSA₆ (c).

3.7.4. Influência dos descritores moleculares no modelo DT/Estate_VSA

A resposta do modelo DT/Estate_VSA é dominada pelos descritores Estate_VSA₉, Estate_VSA₄ e Estate_VSA₁. Estes três descritores moleculares reúnem cerca de 42,2 % da PI total. A Figura 3.18 ilustra a dependência parcial do modelo DT/Estate_VSA com esses três descritores moleculares, onde é possível observar a probabilidade de um determinado resultado de classificação (Q1, Q2, Q3 ou Q4) com valores variados do descritor em questão, quando todas as outras variáveis são substituídas por valores aleatórios.

De acordo com a Figura 3.18a, os valores baixos do descritor Estate_VSA₉ favorecem a classificação Q3, enquanto que valores elevados (a partir de 40) favorecem a classificação Q1. Por outro lado, para moléculas com valores de Estate_VSA₉ entre 34 e 40, o modelo DT/Estate_VSA terá dificuldades em distinguir entre Q1, Q2, Q3 ou Q4. Relativamente à dependência parcial do modelo com o descritor molecular Estate_VSA₄, a Figura 3.18b ilustra que os valores mais baixos favorecem as classificações Q3 e Q4, e que os valores acima de 10 favorecem a classificação Q2. Este modelo aparenta ter dificuldade em distinguir as moléculas com valores entre 7 e 11 do descritor molecular Estate_VSA₄ entre Q2, Q3 ou Q4. Apesar de nenhum ponto do descritor Estate_VSA₄ favorecer a classificação Q1, é notório o aumento da probabilidade de classificação Q1 com o aumento dos valores do descritor, ou seja, moléculas com valores mais elevados do descritor Estate_VSA₄ têm maior probabilidade de serem classificados como Q1 em relação àqueles com valores baixos (Figura 3.18b). Por outro lado, valores baixos do descritor Estate_VSA₁ são usados para favorecer as classificações Q4 e Q3 e valores a partir de aproximadamente 150 favorecem a classificação Q3. Assim como em outros descritores (Estate_VSA₉ e Estate_VSA₄), existe uma zona confusa dos valores do descritor Estate_VSA₁ (valores entre 138 e 156) para classificar as moléculas (Figura 3.18c).

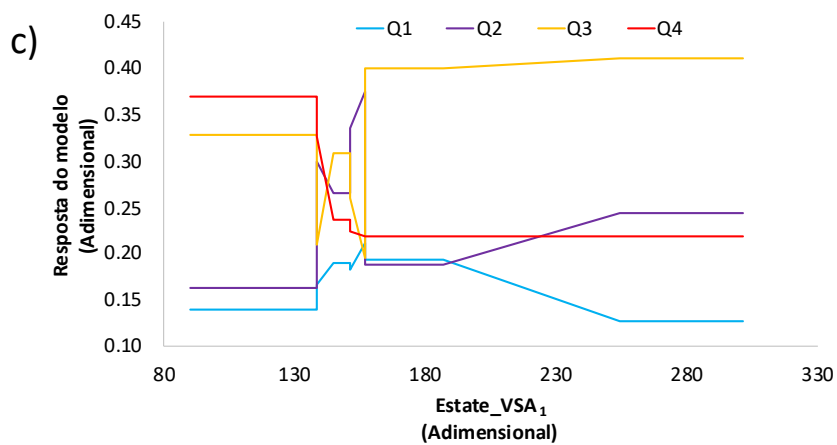
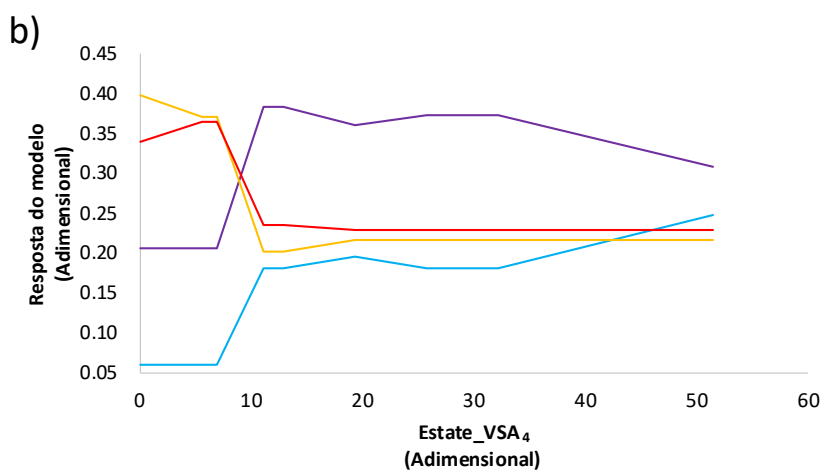
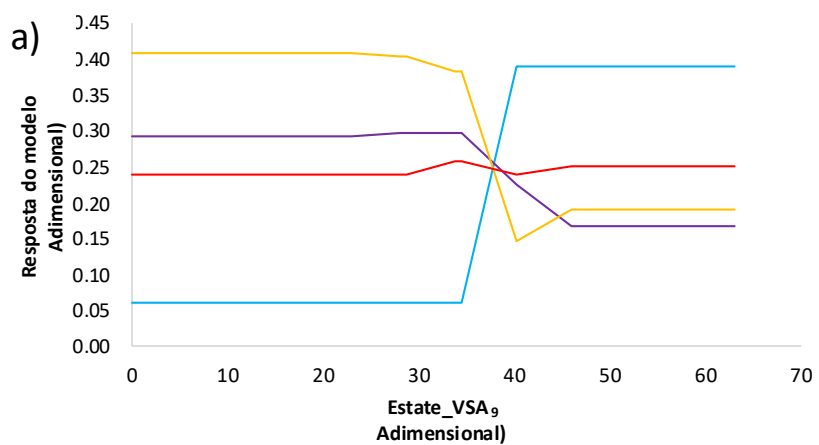


Figura 3.18. Gráficos de dependência parcial do modelo DT usando os descritores moleculares $Estate_VSA_9$, (a), $Estate_VSA_4$, (b) e $Estate_VSA_1$, (c).

3.7.5. Influência do alvo biológico nos modelos AdaBoost/PEOE_VSA e DT/Estate_VSA

De acordo com a Figura 3.16, a resposta dos modelos AdaBoost/PEOE_VSA e DT/Estate_VSA também é dominada pelas duas variáveis que descrevem o alvo biológico (M_{typ} e T_xG). Essas duas variáveis tem um peso de 39,5 % da PI total no modelo AdaBoost/PEOE_VSA e 31,4 % da PI total no modelo DT/Estate_VSA. Foram gerados gráficos de dependência parcial que descrevem a resposta do modelo em relação a M_{typ} (Figura 3.19) e a T_xG (Figura 3.20).

A Figura 3.19 evidencia que, nos dois modelos, os ensaios usando fungos e bactérias Gram-positivas são mais propensos as serem classificados como Q4, ao passo que os ensaios usando bactérias Gram-negativas são mais propensos a serem classificados como Q2 e Q3. Analisando o tipo de microrganismo que mais favorece a classificação Q1, observa-se que, no modelo DT/Estate_VSA, esta é mais favorecida pelas bactérias Gram-negativas (Figura 3.19b), ao passo que o modelo AdaBoost/PEOE_VSA sugere que os três tipos de microrganismos favorecem a classificação Q1 na mesma proporção (Figura 3.19a), o que é pouco provável, uma vez que as polimixinas são pouco ativas em bactérias Gram-positivas e em fungos. Isto provavelmente resulta de um erro do modelo ao tentar compensar os erros de um determinado descritor pela variável M_{typ} .

A Figura 3.20 ilustra os gráficos da dependência parcial dos modelos AdaBoost/PEOE_VSA e DT/Estate_VSA em relação ao gênero taxonômico do microrganismo. De acordo com a Figura 3.20b, no geral, as bactérias Gram-negativas (*Escherichia*, *Pseudomonas* e *Salmonelas*) favorecem a classificação Q2 e as bactérias Gram-positivas e fungos favorecem a classificação Q3. Os dois modelos divergem no gênero *Pseudomonas*, para o qual o modelo AdaBoost/PEOE_VSA sugere que este gênero é mais propenso a ser classificado como Q2 do que como Q4, ao passo que o modelo DT/Estate_VSA sugere o contrário.

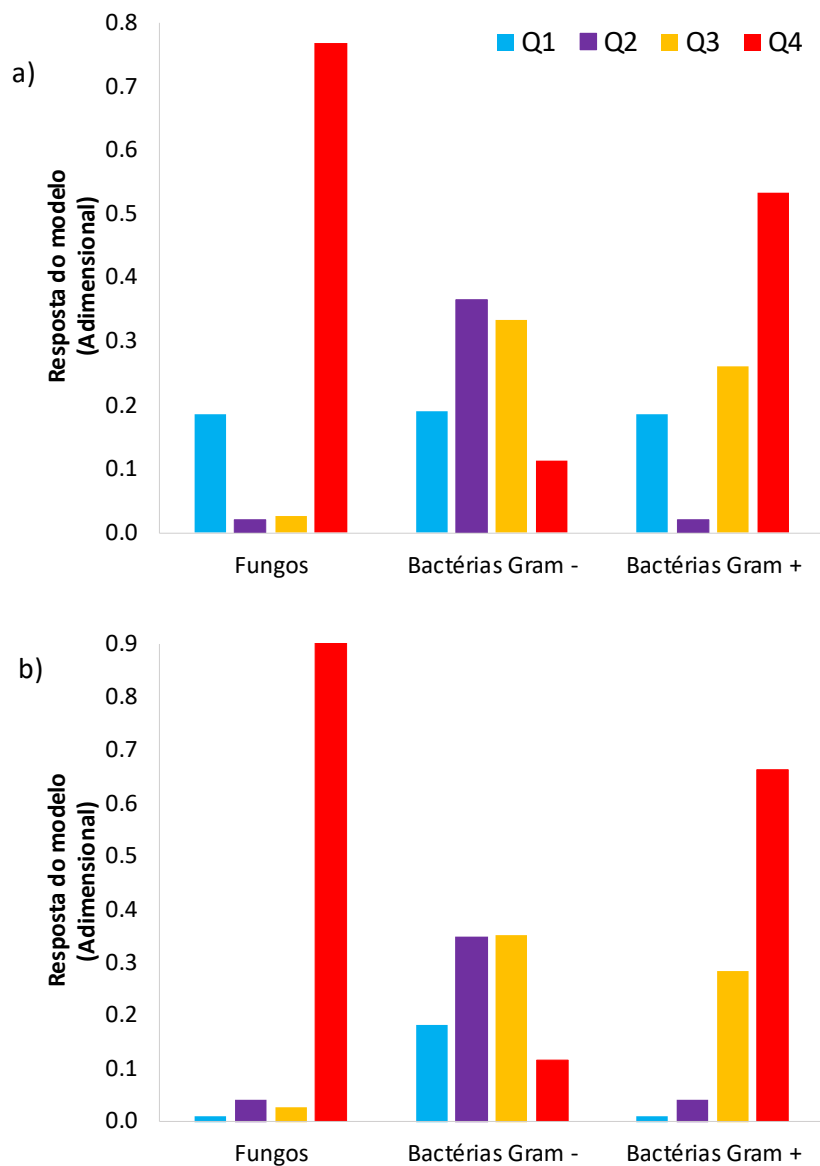


Figura 3.19. Gráficos de dependência parcial dos modelos AdaBoost/PEOE_VSA (a) e DT/Estate_VSA (b) em relação a M_{tipo} .

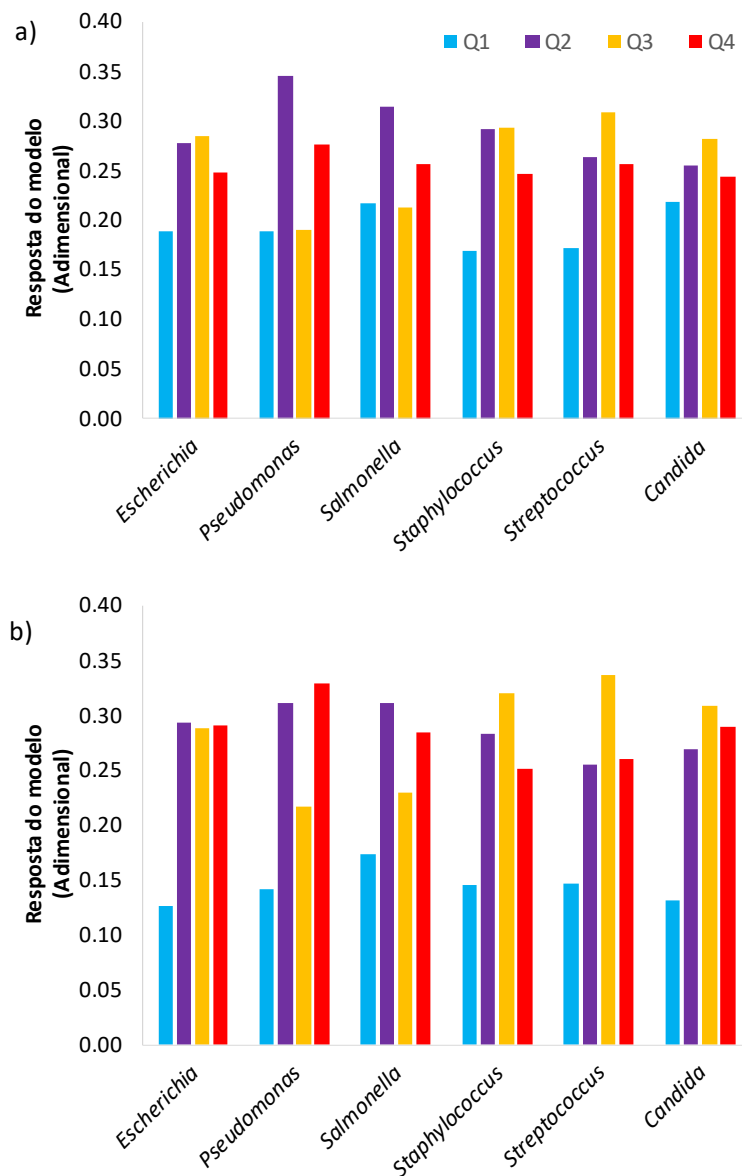


Figura 3.20. Gráfico de dependência parcial dos modelos AdaBoost/PEOE_VSA (a) e DT/Estate_VSA (b) em relação a T.G.

Analisados os desempenhos dos dois melhores modelos da segunda série (AdaBoost/PEOE_VSA e DT/Estate_VSA), percebe-se que o modelo DT/Estate_VSA é mais promissor para prever a atividade de novas moléculas análogas da polimixina B em relação ao modelo AdaBoost/PEOE_VSA. Isto pode ser verificado ao analisar a resposta dos modelos em relação aos descritores moleculares, que no modelo AdaBoost/PEOE_VSA é de que nenhum descritor molecular com mais peso favorece a classificação Q1 (Figura 3.17). O mesmo modelo também comete erro na resposta em relação à variável M_{Typ} ao apresentar a mesma probabilidade de classificar as moléculas como Q1 usando qualquer tipo de microrganismo (bactérias Gram-negativas e Gram-positivas e fungos) (Figura 3.19). Por estas razões, o

modelo DT/Estate_VSA foi selecionado como o melhor modelo desenvolvido neste trabalho e o mesmo foi aplicado para prever atividade de novas moléculas.

3.8. Mutações sistemáticas da polimixina B

Para ter uma noção mais imediata da resposta do modelo DT/Estate_VSA, a estrutura da polimixina B foi sistematicamente mutada nas posições 1 a 3 e 5 a 10 usando Gly, Leu, Lys e Glu. Em todas as execuções preditivas do modelo, foram usados os alvos microbianos *Acinetobacter*, *Pseudomonas* e *Escherichia*.

As mutações da polimixina B apresentaram resultados diferentes em cada alvo usado. Ao usar o género *Acinetobacter* como alvo, todas as estruturas mutadas tiveram maior probabilidade de serem classificadas como Q3 (MIC entre 4 e 32 μM) (Figura A2 em anexos), ou seja, nenhuma das alterações propostas na estrutura da polimixina B daria num composto promissor para tratar infeções causadas pelos microrganismos pertencente a este género taxonómico. Este resultado reflete ao conjunto dados usados neste trabalho, verificando-se que a maioria dos compostos estudados usando a *Acinetobacter* apresentaram MICs elevadas (com uma media de 16,1 μM e um pico modal de 25 μM), mas o valor da MIC da molécula nativa (polimixina B1) variou entre 0,012 μM e 1,9 μM . Por outro lado, ao usar o género *Pseudomonas*, todas as estruturas tiveram maior probabilidade de serem classificadas como Q2 (MIC entre 1 μM e 4 μM) (Figura A3 em Anexos), Por sua vez, as estruturas propostas tiveram a probabilidade de serem classificadas como Q2 ou Q3 usando o género *Escherichia* como alvo, dependendo da alteração feita (Figura 3.21).

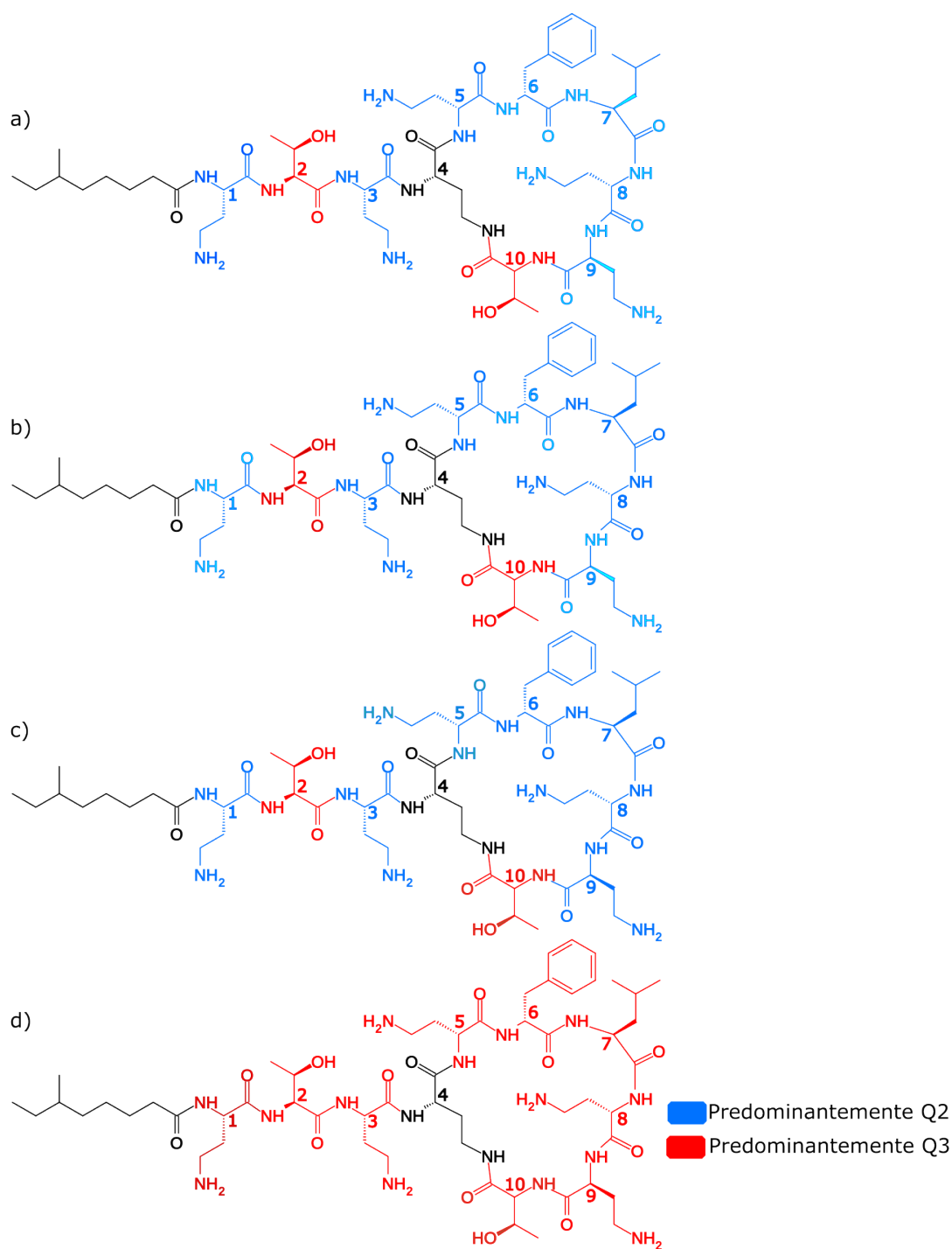


Figura 3.21. Classificação mais provável quanto à atividade antimicrobiana contra *Escherichia* de variantes mutadas da polimixina B ao alterar sistematicamente cada resíduo de aminoácido para: Gly (a), Leu (b), Lys (c) e Glu (d).

No que diz respeito à troca sistemática de cada um dos aminoácidos constituintes por Gly usando como alvo o género *Pseudomonas*, a tendência geral é favorecer a classificação Q2.

No que diz respeito às mutações sistemáticas de cada aminoácido existente na estrutura da polimixina B pelos aminoácidos Gly, Leu, Glu e Lys usando o género *Escherichia* como alvo, os resultados

foram bastante interessantes. As substituições por Gly (Figura 3.21a), Leu (Figura 3.21b) e Lys (Figura 3.21c) favoreceram a classificação Q2, com a exceção da substituição da treonina (Thr) nas posições 2 e 10, a qual teve um impacto negativo sobre a atividade antimicrobiana ao favorecer a classificação Q3. Este resultado revela que a Thr não deve ser substituída, provavelmente por ser importante na hidrofobicidade da molécula. Por outro lado, as alterações sistemáticas de cada aminoácido da polimixina B por Glu prejudicaram a atividade antimicrobiana, favorecendo a classificação como Q3 (Figura 3.21d). Ao comparar os efeitos dos aminoácidos Gly e Leu, os resultados revelaram que a Leu é melhor substituinte que o Gly por apresentar uma probabilidade maior das mutações serem classificadas como Q1. Os mesmos resultados das mutações sistemáticas por Leu são observados nas mutações por Lys. Os melhores resultados destas mutações refletem o mecanismo de ação das polimixinas ao se constatar que as cargas positivas favorecem a atividade antimicrobiana. No que diz respeito a substituição por Leu favorecer positivamente a atividade antimicrobiana, isto também pode ser observado ao comparar a polimixina B (que contém Phe na posição 6 e Leu na posição 7) com a colistina (que contém Leu nas posições 6 e 7), pois a colistina é mais ativa que a polimixina B contra bactérias Gram-negativas.

3.9. Aplicação do modelo

Além da aplicação do modelo DT/Estate_VSA na previsão da atividade das estruturas mutadas sistematicamente de cada aminoácido existente na polimixina B por Gly, Leu, Lys e Glu, o modelo também foi aplicado na previsão da atividade antimicrobiana de seis candidatos propostos pelo grupo de Química Biomolecular Aplicada do Centro de Química da Universidade do Minho, que serão sintetizados e testada a sua atividade *in vitro*. As estratégias usadas na elaboração destes candidatos consistiram na alteração das estruturas da polimixina B e da colistina, como ilustrado nas Figuras 3.22 e 3.23. Na previsão da atividade antimicrobiana destes candidatos, foram usados como alvos os géneros de bactérias que constituem a primeira prioridade no desenvolvimento de novos antibióticos, de acordo com a lista de prioridades publicada pela OMS,³⁵ e com dados mais representativos no conjunto de dados usado para treinar o modelo. Assim, foram selecionados os géneros *Acinetobacter*, *Pseudomonas* e *Escherichia* como alvos na previsão da atividade antimicrobiana dos candidatos propostos. Os resultados da previsão da atividade antimicrobiana dos candidatos propostos são apresentados na Figura 3.24.

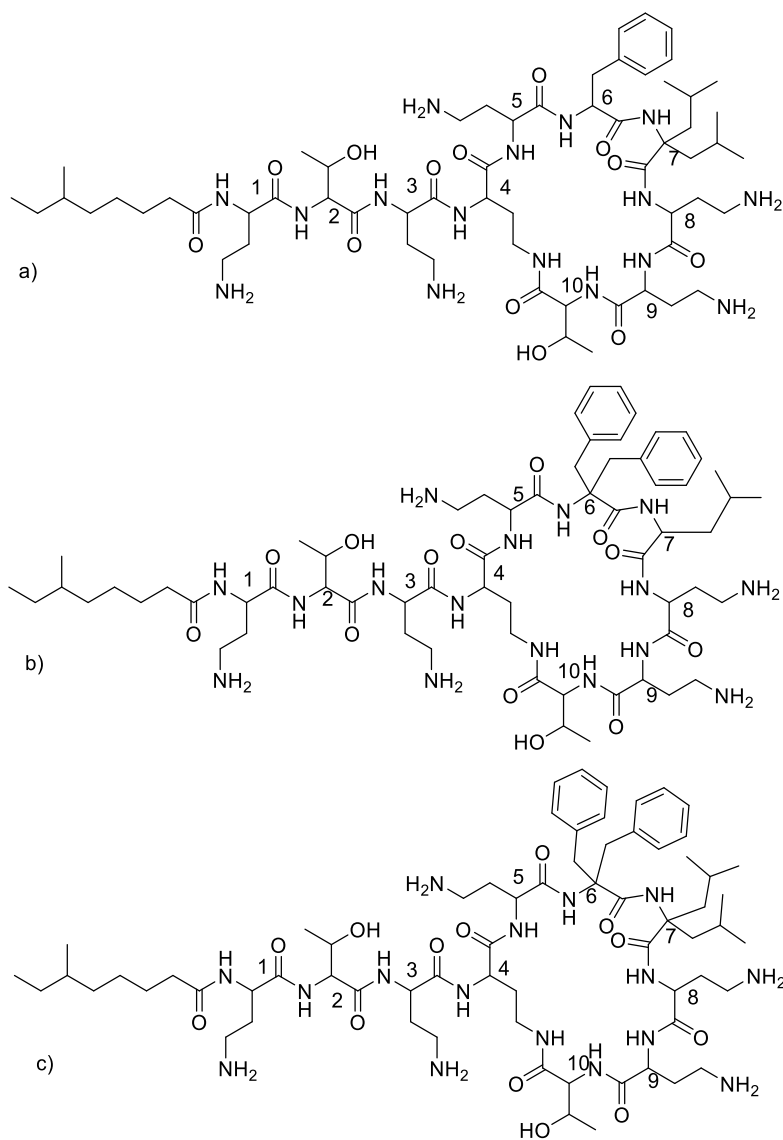


Figura 3.22. Estruturas químicas dos candidatos análogos da polimixina B: Bmim1 (a), Bmim2 (b) e Bmim3 (c).

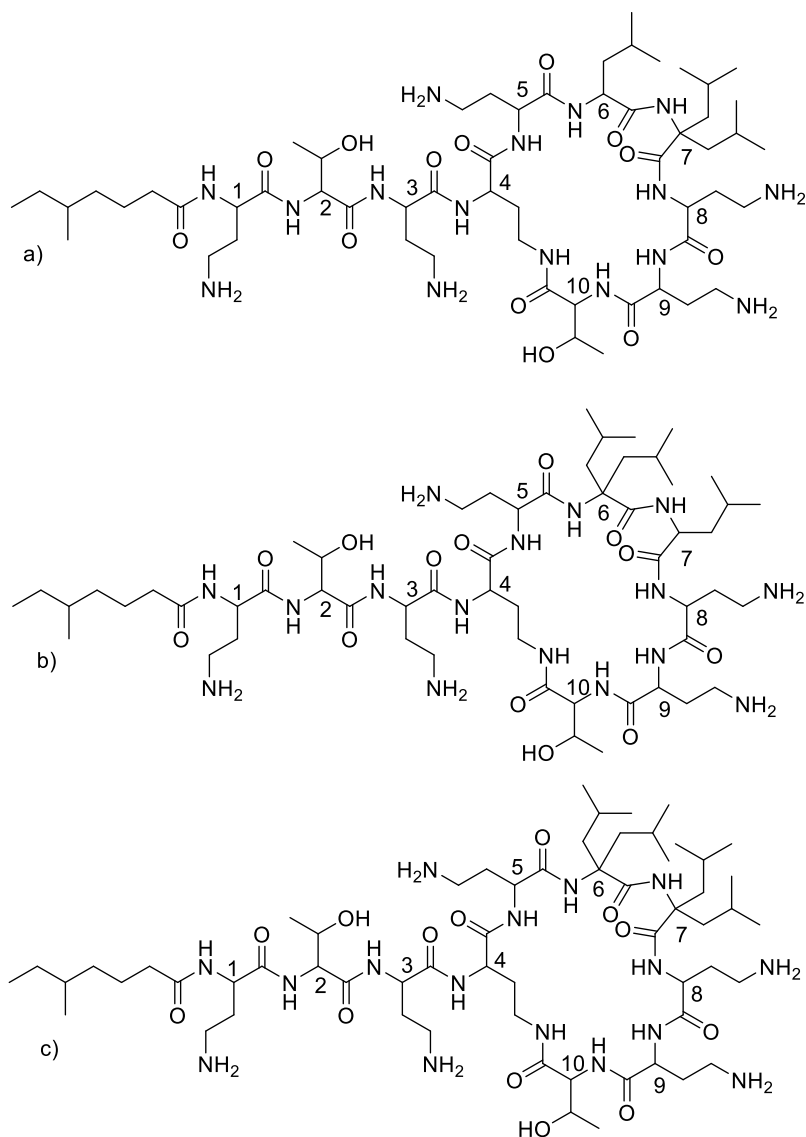


Figura 3.23. Estruturas químicas dos candidatos análogos da polimixina E: Emim1 (a), Emim2 (b) e Emim3 (c).

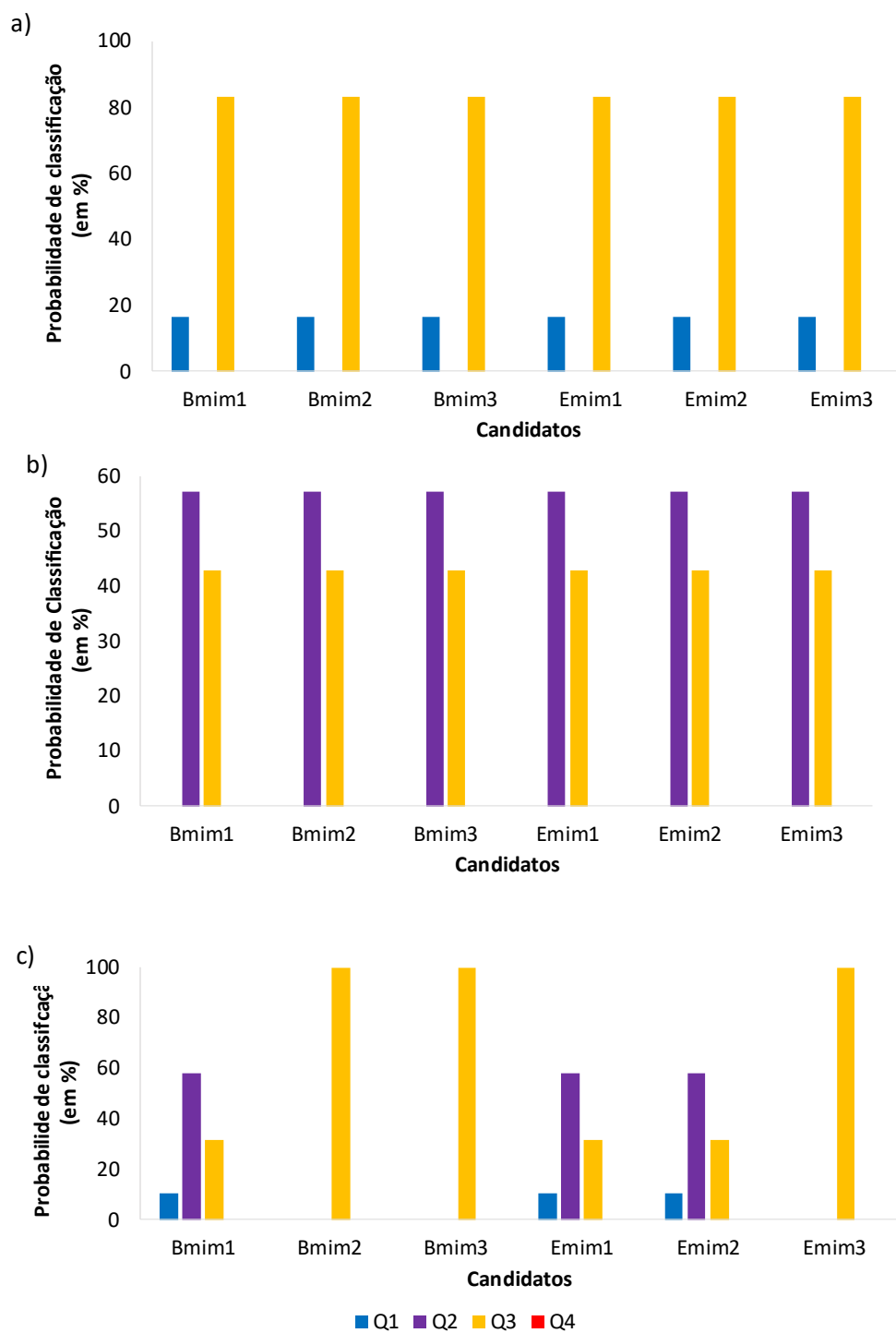


Figura 3.24. Classificação mais provável dos análogos das polimixinas B e E quanto à atividade antimicrobiana contra bactérias dos géneros: *Acinetobacter* (a), *Pseudomonas* (b) e *Escherichia* (c).

O modelo DT/Estate_VSA sugere que todos os candidatos têm maior probabilidade de serem menos ativos (classificados como Q3) e menor probabilidade de serem muito ativos (classificados como Q1) contra espécies bacterianas pertencentes ao gênero *Acinetobacter* (Figura 3.24a). Por outro lado, o modelo ilustra que todos os candidatos têm maior probabilidade de serem classificados como Q2 e menor probabilidade de serem classificados como Q3 contra bactérias do gênero *Pseudomonas* (Figura 3.24b), ou seja, todos os candidatos são previstos como promissores para combaterem infecções causadas por *Pseudomonas*. Os mesmos candidatos são previstos com resultados diferentes contra *Escherichia* (Figura 3.24c). Os análogos 1 e 2 da polimixina E (Figura 3.23a e 3.23b, respectivamente) são mais promissores contra *Escherichia*, ao passo que, nos análogos da polimixina B, apenas o Bmim1 (Figura 3.22a) revelou ser promissor contra bactérias deste gênero (Figura 3.24c). Os resultados em *Escherichia* revelam que o grupo isopropil é melhor para a atividade antimicrobiana ao se observar que a adição de mais um grupo deste na posição 7 dos análogos da polimixina B e E (Figuras 3.22a e 3.23a) favorece a atividade. Por outro lado, o grupo benzil prejudica a atividade antimicrobiana, pois pode ser observado, ao comparar a atividade dos análogos 2 da polimixina B e E (Figuras 3.22b e 3.23b, respectivamente), que a duplicação do grupo benzil na posição 6 não é boa para a atividade, ao passo que a duplicação do isopropil na mesma posição favorece a atividade antimicrobiana contra *Escherichia*. Os resultados também revelaram que a duplicação do isopropil nas duas posições (posições 6 e 7) não é boa para a atividade antimicrobiana contra *Escherichia* (Figura 3.24c).

Os resultados da aplicação do modelo usando o gênero *Escherichia* como alvo provavelmente refletem o efeito da hidrofobicidade da molécula na atividade antimicrobiana, uma vez que as estruturas mais hidrofóbicas tiveram a probabilidade de serem menos ativas. Isto era esperado, uma vez que os compostos mais hidrofílicos têm mais facilidade em atravessar a membrana externa das bactérias Gram-negativas em comparação aos compostos hidrofóbicos.

4. CONCLUSÕES

O trabalho teve como objetivo principal desenhar novas moléculas baseadas na polimixina B, com propriedades antimicrobianas e baixa citotoxicidade, usando uma abordagem *in silico*. Uma vez que não se obteve dados de citotoxicidade suficientes para desenvolver um modelo capaz de prever a citotoxicidade de novas moléculas, podemos considerar que os objetivos do trabalho foram parcialmente alcançados.

O plano de trabalho consistiu, inicialmente, em tratar dados pré-existentes de atividade antimicrobiana e citotoxicidade. No final, foi criada uma base de dados com 413 entradas de pares microrganismo/MIC, que compreende 58 compostos análogos da polimixina B e 42 espécies diferentes de microrganismos (bactérias e fungos). Esta base de dados foi primariamente elaborada a partir da recolha de dados disponíveis na literatura científica e suplementada com novos ensaios microbiológicos.

Na colheita inicial dos dados, foi constatado que alguns microrganismos alvo tinham poucas entradas na base de dados, o que poderia enviesar o modelo. Para resolver este problema e tentar melhorar a precisão e exatidão do modelo, foi decidido analisar a suscetibilidade *in vitro* de algumas das espécies bacterianas com poucas entradas às polimixinas (B e E), usando o método de microdiluição em caldo. Os resultados destes ensaios revelaram que a colistina parece ser mais potente contra bactérias mais susceptíveis a polimixinas, ao passo que a polimixina B parece ser mais potente contra bactérias menos susceptíveis a polimixinas, e ambos os agentes antimicrobianos não são ativos contra a *P. mirabilis*.

No treino dos modelos, foi feita uma pesquisa exaustiva do melhor conjunto de descritores moleculares para a criação de um modelo de classificação semi-quantitativo para a previsão da MIC de derivados da polimixina B usando os algoritmos DT, RF e AdaBoost, visando a posição do quartil de cada entrada no conjunto de dados. Nesta pesquisa, o algoritmo RF destacou-se pela negativa, ao se observar um problema de *overfitting* muito acentuado, ao passo que os algoritmos AdaBoost e DT se destacaram pela positiva, ao se observarem valores elevados de métricas desejáveis (exatidão e verdadeiros positivos) e valores baixos de métricas indesejáveis (falsos positivos e negativos).

Em relação às famílias de descritores moleculares, a contagem de grupos funcionais (FG) foi particularmente desadequada para estes modelos. Por outro lado, as famílias dos descritores topológicos (especificamente AC2D, BCUT e CKP) e famílias de descritores baseadas em área de superfície subdividida (VSA) foram mais propensas a apresentar modelos com bom desempenho.

O modelo DT/Estate_VSA foi o melhor modelo treinado neste trabalho, por apresentar melhor desempenho. Este modelo é sensível a três descritores moleculares (Estate_VSA₉, Estate_VSA₄ e Estate_VSA₁) e duas variáveis que descrevem o alvo microbiológico (tipo de microrganismo e género taxonómico do microrganismo). Neste modelo, os compostos mais ativos foram favorecidos pelos valores mais elevados dos descritores moleculares Estate_VSA₉ e Estate_VSA₄.

A exploração preliminar da resposta do melhor modelo às alterações sistemáticas da estrutura da polimixina B revelou que aminoácidos carregados negativamente e a Leu favorecem a atividade antimicrobiana, usando como alvo o género bacteriano *Escherichia*. Já ao usar os géneros *Acinetobacter* e *Pseudomonas*, a actividade não altera com a mutuação sistemática efetuada da estrutura da polimixina B.

O modelo promissor foi aplicado para prever a atividade antimicrobiana de três análogos da polimixina B e três análogos da polimixina E propostos pelo grupo do projeto onde este trabalho se insere, usando como alvos os géneros *Acinetobacter*, *Pseudomonas* e *Escherichia*. Todos os análogos propostos foram classificados como menos ativos contra *Acinetobacter* e mais ativos contra *Pseudomonas*. Por outro lado, os análogos mais hidrofóbicos foram previstos como menos ativos contra *Escherichia*.

5. PERSPETIVAS DO TRABALHO

Com vista a complementar o trabalho desenvolvido neste estudo, são aqui mencionadas algumas perspetivas para pesquisas futuras.

Como foi referido, não foram obtidos dados suficientes para desenvolver modelos preditivos da toxicidade dos análogos da polimixina B de acordo com as metodologias propostas neste trabalho. Propõe-se a realização um estudo futuro para prever a toxicidade destes análogos usando outras técnicas da química computacional, como *docking* molecular e dinâmica molecular.

Como constatado no conjunto de dados extraído da PubChem, alguns microrganismos tinham poucas entradas. Destes, apenas três foram usadas nos ensaios laboratoriais para determinar a MIC das polimixinas B e E. Para trabalho futuro, sugere-se usar outros microrganismos nos ensaios laboratoriais para suplementar o conjunto de dados.

6. REFERÊNCIAS BIBLIOGRÁFICAS

- (1) Burnett-Boothroyd, S. C.; McCarthy, B. J. Antimicrobial treatments of textiles for hygiene and infection control applications: an industrial perspective. *Text. Hyg. Infect. Control* **2011**, 196–209. <https://doi.org/10.1533/9780857093707.3.196>.
- (2) Food and Drug Administration. *The Judicious Use of Medically Important Antimicrobial Drugs in Food-Producing Animals*, 2012. <http://www.fda.gov/downloads/AnimalVeterinary/GuidanceComplianceEnforcement/GuidanceforIndustry/UCM216936.pdf> (acedido 2023-01-25).
- (3) EUCAST. Methods for the determination of susceptibility of bacteria to antimicrobial agents. Terminology. *Clin. Microbiol. Infect.* **1998**, 4 (5), 291–296. <https://doi.org/10.1111/j.1469-0691.1998.tb00061.x>.
- (4) Samanta, I.; Bandyopadhyay, S. History of antimicrobial resistance. *Antimicrob. Resist. Agric.* **2020**, N. 1874, 1–5. <https://doi.org/10.1016/b978-0-12-815770-1.00001-8>.
- (5) Jayachandran, S. Pre-antibiotics era to post-antibiotic era. *J. Indian Acad. Oral Med. Radiol.* **2018**, 30 (2), 100. https://doi.org/10.4103/JIAOMR.JIAOMR_29_18.
- (6) Aminov, R. History of antimicrobial drug discovery: Major classes and health impact. *Biochem. Pharmacol.* **2017**, 133, 4–19. <https://doi.org/10.1016/j.bcp.2016.10.001>.
- (7) Browne, K.; Chakraborty, S.; Chen, R.; Willcox, M. D. P.; Black, D. S.; Walsh, W. R.; Kumar, N. A new era of antibiotics: The clinical potential of antimicrobial peptides. *Int. J. Mol. Sci* **2020**, 21 (19), 1–23. <https://doi.org/10.3390/ijms21197047>.
- (8) Infectious Diseases Society of America. Position Paper: Recommended Design Features of Future Clinical Trials of Antibacterial Agents for Community-Acquired Pneumonia. *Clin. Infect. Dis.* **2008**, 47 (3), 249–265. <https://doi.org/https://doi.org/10.1086/591411>.
- (9) Dhingra, S.; Rahman, N. A. A.; Peile, E.; Rahman, M.; Sartelli, M.; Hassali, M. A.; Islam, T.; Islam, S.; Haque, M. Microbial Resistance Movements: An Overview of Global Public Health Threats Posed by Antimicrobial Resistance, and How Best to Counter. *Front. Public Heal.* **2020**, 8, 535668. <https://doi.org/10.3389/fpubh.2020.535668>.
- (10) Aljeldah, M. M. Antimicrobial Resistance and Its Spread Is a Global Threat. *Antibiotics* **2022**, 11, 1082. <https://doi.org/10.3390/antibiotics11081082>.
- (11) World Health Organization. *Antibacterial agents in clinical development: an analysis of the antibacterial clinical development pipeline, including tuberculosis*, 2017. <https://apps.who.int/iris/handle/10665/258965>.
- (12) Gajdács, M.; Albericio, F. Antibiotic Resistance: From the Bench to Patients. *Antibiotics* **2019**, 8 (3), 129. <https://doi.org/10.3390/antibiotics8030129>.
- (13) Stokes, J. M.; Lopatkin, A. J.; Lobritz, M. A.; Collins, J. J. Bacterial Metabolism and Antibiotic Efficacy. *Cell Metab.* **2019**, 30 (2), 251–259. <https://doi.org/10.1016/j.cmet.2019.06.009>.
- (14) Hutchings, M.; Truman, A.; Wilkinson, B. Antibiotics: past, present and future. *Curr. Opin. Microbiol.* **2019**, 51, 72–80. <https://doi.org/10.1016/j.mib.2019.10.008>.
- (15) Kaufmann, S. H. E. Paul Ehrlich: Founder of chemotherapy. *Nat. Rev. Drug Discov.* **2008**, 7 (5), 373. <https://doi.org/10.1038/nrd2582>.
- (16) Rolinson, G. N.; Geddes, A. M. The 50th anniversary of the discovery of 6-aminopenicillanic acid

- (6-APA). *Int. J. Antimicrob. Agents* **2007**, *29* (1), 3–8. <https://doi.org/10.1016/j.ijantimicag.2006.09.003>.
- (17) Plough, H. H.; Young, H. N.; Grimm, M. R. Penicillin-screened auxotrophic mutations in *Salmonella typhimurium* and their relation to x-ray dosage. *J. Bacteriol* **1950**, *60* (2), 145. <https://doi.org/10.1128/jb.60.2.145-157.1950>.
- (18) Otten, H. Domagk and the development of the sulphonamides. *J. Antimicrob. Chemother.* **1986**, *17*, 689–696. <https://doi.org/10.1093/jac/17.6.689>.
- (19) Gelpi, A.; Gilbertson, A.; Tucker, J. D. Magic bullet: Paul Ehrlich, Salvarsan and the birth of venereology. *Sex. Transm. Infect.* **2015**, *91* (1), 68–69. <https://doi.org/10.1136/sextrans-2014-051779>.
- (20) Jayachandran, S.; Lleras-Muney, A.; Smith, K. V. Modern medicine and the twentieth century decline in mortality: Evidence on the impact of sulfa drugs. *Am. Econ. J. Appl. Econ.* **2010**, *2* (2), 118–146. <https://doi.org/10.1257/app.2.2.118>.
- (21) Boparai, J. K.; Sharma, P. K. Mini Review on Antimicrobial Peptides, Sources, Mechanism and Recent Applications. *Protein Pept. Lett.* **2020**, *27* (1), 4–16. <https://doi.org/10.2174/0929866526666190822165812>.
- (22) Claro, B.; Bastos, M.; Garcia-Fandino, R. Design and applications of cyclic peptides. *Em Peptide Applications in Biomedicine, Biotechnology and Bioengineering*, Koutsopoulos, S., Ed.; Woodhead Publishing, 2018; pp 87–129. <https://doi.org/10.1016/B978-0-08-100736-5.00004-1>.
- (23) Bahar, A. A.; Ren, D. Antimicrobial peptides. *Pharmaceuticals* **2013**, *6* (12), 1543–1575. <https://doi.org/10.3390/ph6121543>.
- (24) Li, X.; Zuo, S.; Wang, B.; Zhang, K.; Wang, Y. Antimicrobial Mechanisms and Clinical Application Prospects of Antimicrobial Peptides. *Molecules* **2022**, *27* (9), 2675. <https://doi.org/10.3390/molecules27092675>.
- (25) Almeida, L. H. de O.; Oliveira, C. F. R. de; Rodrigues, M. de S.; Neto, S. M.; Boleti, A. P. de A.; Taveira, G. B.; Mello, É. de O.; Gomes, V. M.; Santos, E. L. dos; Crusca, E.; Franco, O. L.; Cardoso, M. H. e. S.; Macedo, M. L. R. Adepamycin: design, synthesis and biological properties of a new peptide with antimicrobial properties. *Arch. Biochem. Biophys.* **2020**, *691*, 108487. <https://doi.org/10.1016/J.ABB.2020.108487>.
- (26) Uddin, T. M.; Chakraborty, A. J.; Khusro, A.; Zidan, B. R. M.; Mitra, S.; Emran, T. Bin; Dhama, K.; Ripon, M. K. H.; Gajdács, M.; Sahibzada, M. U. K.; Hossain, M. J.; Koirala, N. Antibiotic resistance in microbes: History, mechanisms, therapeutic strategies and future prospects. *J. Infect. Public Health* **2021**, *14* (12), 1750–1766. <https://doi.org/10.1016/J.JIPH.2021.10.020>.
- (27) Magana, M.; Pushpanathan, M.; Santos, A. L.; Leanse, L.; Fernandez, M.; Ioannidis, A.; Giulianotti, M. A.; Apidianakis, Y.; Bradfute, S.; Ferguson, A. L.; Cherkasov, A.; Seleem, M. N.; Pinilla, C.; de la Fuente-Nunez, C.; Lazaridis, T.; Dai, T.; Houghten, R. A.; Hancock, R. E. W.; Tegos, G. P. The value of antimicrobial peptides in the age of resistance. *Lancet Infect. Dis.* **2020**, *20* (9), 216–230. [https://doi.org/10.1016/S1473-3099\(20\)30327-3](https://doi.org/10.1016/S1473-3099(20)30327-3).
- (28) CDC. *Antibiotic Resistance Threats in the United States, 2019*, 2019. <https://doi.org/10.15620/cdc:82532>. U.S.
- (29) World Health Organization. *2021 Antibacterial Agents in Clinical and Preclinical Development: an overview and analysis*. World Health Organization. <https://www.who.int/publications/i/item/9789240021303> (accedido 2015-08-22).

- (30) World Health Organization, Food and Agriculture Organization of the United Nations, U. N. E. P. and W. O. for A. H. *Antimicrobial Resistance Multi-Partner Trust Fund annual report 2021*; 2022.
- (31) O'Neill, J. *Antimicrobial Resistance : Tackling a crisis for the health and wealth of nations*; 2014; Vol. 20. https://amr-review.org/sites/default/files/AMR_Review_Paper_-_Tackling_a_crisis_for_the_health_and_wealth_of_nations_1.pdf (acedido 2022-08-15).
- (32) European Medicines Agency. *Antimicrobial resistance*. <https://www.ema.europa.eu/en/human-regulatory/overview/public-health-threats/antimicrobial-resistance> (acedido 2022-08-28).
- (33) Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **2022**, *399* (10325), 629–655. [https://doi.org/10.1016/S0140-6736\(21\)02724-0](https://doi.org/10.1016/S0140-6736(21)02724-0).
- (34) Organização Mundial da Saúde. *A crescente ameaça da resistência antimicrobiana: Opções de ação*, 2012. https://www.afro.who.int/sites/default/files/2017-06/OMS_IER_PSP_2012.2_por.pdf (acedido 2023-01-23).
- (35) *WHO publishes list of bacteria for which new antibiotics are urgently needed*. <https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed> (acedido 2023-09-10).
- (36) Falagas, M. E.; Kyriakidou, M.; Voulgaris, G. L.; Vokos, F.; Politi, S.; Kechagias, K. S. Clinical use of intravenous polymyxin B for the treatment of patients with multidrug-resistant Gram-negative bacterial infections: An evaluation of the current evidence. *J. Glob. Antimicrob. Resist.* **2021**, *24*, 342–359. <https://doi.org/10.1016/J.JGAR.2020.12.026>.
- (37) El-Sayed Ahmed, M. A. E. G.; Zhong, L. L.; Shen, C.; Yang, Y.; Doi, Y.; Tian, G. B. Colistin and its role in the Era of antibiotic resistance: an extended review (2000–2019). *Emerg. Microbes Infect.* **2020**, *9* (1), 868–885. <https://doi.org/10.1080/22221751.2020.1754133>.
- (38) Falagas, M.; Vardakas, K. *Polymyxins - Infectious Disease and Antimicrobial Agents*. <http://www.antimicrobe.org/d05.asp> (acedido 2022-12-21).
- (39) Jiang, X.; Zhang, S.; Azad, M. A. K.; Roberts, K. D.; Wan, L.; Gong, B.; Yang, K.; Yuan, B.; Uddin, H.; Li, J.; Thompson, P. E.; Velkov, T.; Fu, J.; Wang, L.; Li, J. Structure-Interaction Relationship of Polymyxins with the Membrane of Human Kidney Proximal Tubular Cells. *ACS Infect. Dis.* **2020**, *6* (8), 2110–2119. <https://doi.org/10.1021/acsinfectdis.0c00190>.
- (40) Newton, B. A. The properties and mode of action of the polymyxins. *Bacteriol. Rev.* **1956**, *20* (1), 14–27. <https://doi.org/10.1128/membr.20.1.14-27.1956>.
- (41) Liu, Y.; Chandler, C. E.; Leung, L. M.; Mcelheny, C. L.; Mettus, R. T.; Shanks, R. M. Q.; Liu, J.; Goodlett, D. R.; Ernst, R. K. Structural Modification of Lipopolysaccharide Conferred by mcr-1 in Gram-Negative ESKAPE Pathogens. *Antimicrob Agents Chemother* **2017**, *61* (6), 00580–17. <https://doi.org/10.1128/AAC.00580-17>.
- (42) Deris, Z. Z.; Swarbrick, J. D.; Roberts, K. D.; Azad, M. A. K.; Akter, J.; Horne, A. S.; Nation, R. L.; Rogers, K. L.; Thompson, P. E.; Velkov, T.; Li, J. Probing the penetration of antimicrobial polymyxin lipopeptides into gram-negative bacteria. *Bioconjug. Chem.* **2014**, *25* (4), 750–760. <https://doi.org/10.1021/bc500094d>.
- (43) Deris, Z. Z.; Akter, J.; Sivanesan, S.; Roberts, K. D.; Thompson, P. E.; Nation, R. L.; Li, J.; Velkov, T. A secondary mode of action of polymyxins against Gram-negative bacteria involves the inhibition of NADH-quinone oxidoreductase activity. *J. Antibiot. (Tokyo)*. **2014**, *67*, 147–151. <https://doi.org/10.1038/ja.2013.111>.

- (44) Nation, R. L.; Maria, M. H.; Falci, D. R.; Zavascki, A. P. Polymyxin acute kidney injury: Dosing and other strategies to reduce toxicity. *Antibiotics* **2019**, *8* (1), 24. <https://doi.org/10.3390/antibiotics8010024>.
- (45) Dubashynskaya, N. V.; Skorik, Y. A. Polymyxin Delivery Systems: Recent Advances and Challenges. *Pharmaceuticals* **2020**, *13* (5), 83. <https://doi.org/10.3390/ph13050083>.
- (46) Zavascki, A. P.; Nation, R. L. Nephrotoxicity of polymyxins: Is there any difference between colistimethate and polymyxin B? *Antimicrob. Agents Chemother.* **2017**, *61* (3), e02319-16. <https://doi.org/10.1128/AAC.02319-16>.
- (47) Molina, J.; Cordero, E.; Pachón, J. New information about the polymyxin/colistin class of antibiotics. *Expert Opin. Pharmacother.* **2009**, *10* (17), 2811–2828. <https://doi.org/10.1517/14656560903334185>.
- (48) Tuon, F. F.; Rigatto, M. H.; Lopes, C. K.; Kamei, L. K.; Rocha, J. L.; Zavascki, A. P. Risk factors for acute kidney injury in patients treated with polymyxin B or colistin methanesulfonate sodium. *Int. J. Antimicrob. Agents* **2014**, *43* (4), 349–352. <https://doi.org/10.1016/j.ijantimicag.2013.12.002>.
- (49) Michalopoulos, A.; Falagas, M. E. Colistin and Polymyxin B in Critical Care. *Crit. Care Clin.* **2008**, *24* (2), 377–391. <https://doi.org/10.1016/j.ccc.2007.12.003>.
- (50) Vardakas, K. Z.; Falagas, M. E. Colistin versus polymyxin B for the treatment of patients with multidrug-resistant Gram-negative infections: a systematic review and meta-analysis. *Int. J. Antimicrob. Agents* **2017**, *49* (2), 233–238. <https://doi.org/10.1016/j.ijantimicag.2016.07.023>.
- (51) Ouderkirk, J. P.; Nord, J. A.; Turett, G. S.; Kislak, J. W. Polymyxin B nephrotoxicity and efficacy against nosocomial infections caused by multiresistant gram-negative bacteria. *Antimicrob. Agents Chemother.* **2003**, *47* (8), 2659–2662. <https://doi.org/10.1128/AAC.47.8.2659-2662.2003>.
- (52) Zeng, H.; Zeng, Z.; Kong, X.; Zhang, H.; Chen, P.; Luo, H.; Chen, Y. Effectiveness and Nephrotoxicity of Intravenous Polymyxin B in Chinese Patients With MDR and XDR Nosocomial Pneumonia. *Front. Pharmacol.* **2021**, *11*, 1–10. <https://doi.org/10.3389/fphar.2020.579069>.
- (53) Montero, M.; Horcajada, J. P.; Sorlí, L.; Alvarez-Lerma, F.; Grau, S.; Riu, M.; Sala, M.; Knobel, H. Effectiveness and safety of colistin for the treatment of multidrug-resistant *Pseudomonas aeruginosa* infections. *Infection* **2009**, *37* (5), 461–465. <https://doi.org/10.1007/s15010-009-8342-x>.
- (54) Avedissian, S. N.; Scheetz, M. H. Does renal function matter for polymyxin B? *Br. J. Clin. Pharmacol.* **2021**, *87* (7), 2629–2632. <https://doi.org/10.1111/bcp.14675>.
- (55) Yun, B.; Zhang, T.; Azad, M. A. K.; Wang, J.; Nowell, C. J.; Kalitsis, P.; Velkov, T.; Hudson, D. F.; Li, J. Polymyxin B causes DNA damage in HK-2 cells and mice. *Arch. Toxicol.* **2018**, *92* (7), 2259–2271. <https://doi.org/10.1007/s00204-018-2192-1>.
- (56) Gonzalez-Avila, L. U.; Loyola-Cruz, M. A.; Hernández-Cortez, C.; Bello-López, J. M.; Castro-Escarpulli, G. Colistin resistance in aeromonas spp. *Int. J. Mol. Sci.* **2021**, *22* (11), 5974. <https://doi.org/10.3390/ijms22115974>.
- (57) Olaitan, A. O.; Morand, S.; Rolain, J. M. Mechanisms of polymyxin resistance: Acquired and intrinsic resistance in bacteria. *Front. Microbiol.* **2014**, *5*, 1–18. <https://doi.org/10.3389/fmicb.2014.00643>.

- (58) Liu, Y. Y.; Wang, Y.; Walsh, T. R.; Yi, L. X.; Zhang, R.; Spencer, J.; et al. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: A microbiological and molecular biological study. *Lancet Infect. Dis.* **2016**, *16* (2), 161–168. [https://doi.org/10.1016/S1473-3099\(15\)00424-7](https://doi.org/10.1016/S1473-3099(15)00424-7).
- (59) Tietgen, M.; Semmler, T.; Riedel-Christ, S.; Kempf, V. A. J.; Molinaro, A.; Ewers, C.; Göttig, S. Impact of the colistin resistance gene *mcr-1* on bacterial fitness. *Int. J. Antimicrob. Agents* **2018**, *51* (4), 554–561. <https://doi.org/10.1016/j.ijantimicag.2017.11.011>.
- (60) Ohene Larbi, R.; Adeapena, W.; Ayim-Akonor, M.; Ansa, E. D. O.; Tweya, H.; Terry, R. F.; Labi, A. K.; Harries, A. D. Antimicrobial, Multi-Drug and Colistin Resistance in Enterobacteriaceae in Healthy Pigs in the Greater Accra Region of Ghana, 2022: A Cross-Sectional Study. *Int. J. Environ. Res. Public Health* **2022**, *19* (16), 10449. <https://doi.org/10.3390/ijerph191610449>.
- (61) Chan, H. C. S.; Shan, H.; Dahoun, T.; Vogel, H.; Yuan, S. Advancing Drug Discovery via Artificial Intelligence. *Trends Pharmacol. Sci.* **2019**, *40* (8), 592–604. <https://doi.org/10.1016/j.tips.2019.06.004>.
- (62) Cavasotto, C. N.; Aucar, M. G. High-Throughput Docking Using Quantum Mechanical Scoring. *Front. Chem.* **2020**, *8*, 246. <https://doi.org/10.3389/fchem.2020.00246>.
- (63) Biswas, S. S.; Browne, R. B.; Borah, V. V.; Roy, J. D. In Silico Approach for Phytocompound-Based Drug Designing to Fight Efflux Pump-Mediated Multidrug-Resistant Mycobacterium tuberculosis. *Appl. Biochem. Biotechnol.* **2021**, *193* (6), 1757–1779. <https://doi.org/10.1007/s12010-021-03557-1>.
- (64) Singh, N.; Chaput, L.; Villoutreix, B. O. Virtual screening web servers: Designing chemical probes and drug candidates in the cyberspace. *Brief. Bioinform.* **2021**, *22* (2), 1790–1818. <https://doi.org/10.1093/bib/bbaa034>.
- (65) Barreiro, E. J.; Fraga, C. A. M. *Química medicinal: as bases moleculares da ação dos fármacos*, 3.ª ed.; Artmed, Ed.; 2015.
- (66) Lin, X.; Li, X.; Lin, X. A review on applications of computational methods in drug screening and design. *Molecules* **2020**, *25* (6), 1375. <https://doi.org/10.3390/molecules25061375>.
- (67) Oktay, L.; Erdemoğlu, E.; Tolu, İ.; Yumak, Y.; Özcan, A.; Acar, E.; Büyükkiliç, Ş.; Olkan, A.; Durdağ, S. Binary-QSAR guided virtual screening of FDA approved drugs and compounds in clinical investigation against SARS-CoV-2 main protease. *Turkish J. Biol.* **2021**, *45*, 459–468. <https://doi.org/10.3906/biy-2106-61>.
- (68) Skariyachan, S.; Gopal, D.; Kadam, S. P.; Muddebihalkar, A. G.; Uttarkar, A.; Niranjan, V. Carbon fullerene acts as potential lead molecule against prospective molecular targets of biofilm-producing multidrug-resistant *Acinetobacter baumannii* and *Pseudomonas aeruginosa*: computational modeling and MD simulation studies. *J. Biomol. Struct. Dyn.* **2021**, *39* (3), 1121–1137. <https://doi.org/10.1080/07391102.2020.1726821>.
- (69) Shakil, S.; Danish Rizvi, S. M.; Greig, N. H. High throughput virtual screening and molecular dynamics simulation for identifying a putative inhibitor of bacterial CTX-M-15. *Antibiotics* **2021**, *10* (5), 474. <https://doi.org/10.3390/antibiotics10050474>.
- (70) Uesawa, Y. AI-based QSAR Modeling for Prediction of Active Compounds in MIE/AOP. *Yakugaku Zasshi* **2020**, *140* (4), 499–505. <https://doi.org/10.1248/yakushi.19-00190-4>.
- (71) Yang, X.; Wang, Y.; Byrne, R.; Schneider, G.; Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* **2019**, *119* (18), 10520–10594.

- <https://doi.org/10.1021/acs.chemrev.8b00728>.
- (72) Kolluri, S.; Lin, J.; Liu, R.; Zhang, Y.; Zhang, W. Machine Learning and Artificial Intelligence in Pharmaceutical Research and Development: a Review. *AAPS J.* **2022**, *24* (1), 1–10. <https://doi.org/10.1208/s12248-021-00644-3>.
- (73) Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 4.^a ed.; Global, Ed.; 2022.
- (74) Sellwood, M. A.; Ahmed, M.; Segler, M. H. S.; Brown, N. Artificial intelligence in drug discovery. *Future Med. Chem.* **2018**, *10* (17), 2025–2028. <https://doi.org/10.4155/fmc-2018-0212>.
- (75) Zhu, H. Big data and artificial intelligence modeling for drug discovery. *Annu. Rev. Pharmacol. Toxicol.* **2020**, *60*, 573–589. <https://doi.org/10.1146/annurev-pharmtox-010919-023324>.
- (76) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; Winkler, D. A.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A. QSAR without borders. *Chem. Soc. Rev.* **2020**, *49* (11), 3525–3564. <https://doi.org/10.1039/d0cs00098a>.
- (77) Ramírez-Palma, L. G.; García-Jacas, C. R.; García-Ramos, J. C.; Almada-Monter, R.; Galindo-Murillo, R.; Cortés-Guzmán, F. Pharmacophoric sites of anticancer metal complexes located using quantum topological atomic descriptors. *J. Mol. Struct.* **2020**, *1204*, 127480. <https://doi.org/10.1016/j.molstruc.2019.127480>.
- (78) Wang, T.; Yuan, X.; Song, W.; Wu, M.; Bin, L.; Lin, J. P.; Yang, L. R. The advancement of multidimensional QSAR for novel drug discovery - where are we headed? *Expert Opin. Drug Discov.* **2017**, *12* (8), 769–784. <https://doi.org/10.1080/17460441.2017.1336157>.
- (79) Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. D. S. Predicting multiple ecotoxicological profiles in agrochemical fungicides: A multi-species chemoinformatic approach. *Ecotoxicol. Environ. Saf.* **2012**, *80*, 308–313. <https://doi.org/10.1016/j.ecoenv.2012.03.018>.
- (80) Milicevic, A.; Sinko, G. Development of a simple QSAR model for reliable evaluation of acetylcholinesterase inhibitor potency. *Eur. J. Pharm. Sci.* **2021**, *160*, 105757. <https://doi.org/10.1016/j.ejps.2021.105757>.
- (81) Pandey, S. K.; Roy, K. QSPR modeling of octanol-water partition coefficient and organic carbon normalized sorption coefficient of diverse organic chemicals using Extended Topochemical Atom (ETA) indices. *Ecotoxicol. Environ. Saf.* **2021**, *208*, 111411. <https://doi.org/10.1016/j.ecoenv.2020.111411>.
- (82) Fujita, T.; Winkler, D. A. Understanding the Roles of the «two QSARs». *J. Chem. Inf. Model.* **2016**, *56* (2), 269–274. <https://doi.org/10.1021/acs.jcim.5b00229>.
- (83) 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* **1962**, *194*, 178–180. <https://doi.org/10.1038/194178b0>.
- (84) Patrick, G. L. *An Introduction to Medicinal Chemistry*, 5.^a ed.; Oxford University, Ed.; 2013.
- (85) Kubinyi, H.; Kehrhaan, O. Quantitative Structure-Activity Relationships. 1. The Modified Free-Wilson Approach. *J. Med. Chem.* **1976**, *19* (5), 578–586. <https://doi.org/10.1021/jm00227a003>.
- (86) Brown, N. In *Silico Medicinal Chemistry: Computational Methods to Support Drug Design*. Em *Royal Society of Chemistry*, 2015; pp 3–9. <https://doi.org/10.1039/9781782622604-00003>.

- (87) Jorge, S. D.; Palace-Berl, F.; Masunari, A.; Cechinel, C. A.; Ishii, M.; Pasqualoto, K. F. M.; Tavares, L. C. Novel benzofuroxan derivatives against multidrug-resistant *Staphylococcus aureus* strains: Design using Topliss' decision tree, synthesis and biological assay. *Bioorganic Med. Chem.* **2011**, *19* (16), 5031–5038. <https://doi.org/10.1016/j.bmc.2011.06.034>.
- (88) Topliss, J. G. Utilization of Operational Schemes for Analog Synthesis in Drug Design. *J. Med. Química.* **1972**, *15*, 1006–1011. <https://doi.org/10.1021/jm00280a002>.
- (89) Richter, L. Topliss Batchwise Schemes Reviewed in the Era of Open Data Reveal Significant Differences between Enzymes and Membrane Receptors. *J. Chem. Inf. Model.* **2017**, *57* (10), 2575–2583. <https://doi.org/10.1021/acs.jcim.7b00195>.
- (90) Craig, P. N. Interdependence between Physical Parametess and Selection of Substituent Groups for Correlation Studies. *J. Med. Chem.* **1971**, *14* (8), 680–684. <https://doi.org/10.1021/jm00290a004>.
- (91) Mao, J.; Akhtar, J.; Zhang, X.; Sun, L.; Guan, S.; Li, X.; Chen, G.; Liu, J.; Jeon, H. N.; Kim, M. S.; No, K. T.; Wang, G. Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. *iScience* **2021**, *24* (9), 103052. <https://doi.org/10.1016/j.isci.2021.103052>.
- (92) Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **2018**, *23* (8), 1538–1546. <https://doi.org/10.1016/j.drudis.2018.05.010>.
- (93) Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S. P. A Critical Review of Machine Learning of Energy Materials. *Adv. Energy Mater.* **2020**, *10* (8), 1903242 (1). <https://doi.org/10.1002/aenm.201903242>.
- (94) Gong, Z.; Zhong, P.; Hu, W. Diversity in Machine Learning. *IEEE Access* **2019**, *7*, 64323–64350. <https://doi.org/10.1109/ACCESS.2019.2917620>.
- (95) Rokach, L.; Maimon, O. *Data mining with decision trees: theory and applications*, 2.^a ed.; World Scientific, Ed.; 2015.
- (96) *scikit-learn: machine learning in Python – scikit-learn 1.2.0 documentation*. <https://scikit-learn.org/stable/> (accedido 2022-12-27).
- (97) Hu, L.; Li, L. Using Tree-Based Machine Learning for Health Studies: Literature Review and Case Series. *Int. J. Environ. Res. Public Health* **2022**, *19* (23), 16080. <https://doi.org/10.3390/ijerph192316080>.
- (98) Rivest, R. L. Learning Decision Lists. *Mach. Learn.* **1987**, *2* (3), 229–246. <https://doi.org/10.1023/A:1022607331053>.
- (99) Wang, C.; Xu, S.; Yang, J. Adaboost algorithm in artificial intelligence for optimizing the IRI prediction accuracy of asphalt concrete pavement. *Sensors* **2021**, *21* (17), 5682. <https://doi.org/10.3390/s21175682>.
- (100) Ben-Gal, I.; Dana, A.; Shkolnik, N.; Singer, G. Efficient construction of decision trees by the dual information distance method. *Qual. Technol. Quant. Manag.* **2014**, *11* (1), 133–147. <https://doi.org/10.1080/16843703.2014.11673330>.
- (101) Javed Mehedi Shamrat, F. M.; Ranjan, R.; Hasib, K. M.; Yadav, A.; Siddique, A. H. Performance Evaluation Among ID3, C4.5, and CART Decision Tree Algorithm. *Lect. Notes Networks Syst.* **2022**, *317*, 127–142. https://doi.org/10.1007/978-981-16-5640-8_11.

- (102) Breiman, L. Random forests. Em *Machine Learning*, Schapire, R. E., Ed.; Kluwer Academic Publishers, 2001; Vol. 45, pp 5–32. <https://doi.org/10.1023/A:1010933404324>.
- (103) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2* (3), 18–22.
- (104) Awasthi, S. *Random Forests in Machine Learning: A Detailed Explanation - datamahadev.com*. <https://datamahadev.com/random-forests-in-machine-learning-a-detailed-explanation/> (acedido 2023-10-14).
- (105) Kégl, B. The return of AdaBoost.MH: multi-class Hamming trees. *arXiv Prepr. arXiv* **2013**, *1312*, 6086. <https://doi.org/10.48550/arXiv.1312.6086>.
- (106) Zhu, J.; Zou, H.; Rosset, S.; Hastie, T. Multi-class AdaBoost *. *Stat. Interface* **2009**, *2*, 349–360.
- (107) Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. *J. R. Stat. Soc.* **1974**, *36* (2), 111–147.
- (108) Bengio, Y.; Grandvalet, Y. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *J. Mach. Learn. Res.* **2004**, *5* (4), 1089–1105.
- (109) Couronné, R.; Probst, P.; Boulesteix, A. L. Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics* **2018**, *19*, 270. <https://doi.org/10.1186/s12859-018-2264-5>.
- (110) Tharwat, A. Classification assessment methods. *Appl. Comput. Informatics* **2021**, *17* (1), 168–192. <https://doi.org/10.1016/j.aci.2018.08.003>.
- (111) 3.3. Metrics and scoring: quantifying the quality of predictions — scikit-learn 1.1.2 documentation. https://scikit-learn.org/stable/modules/model_evaluation.html (acedido 2023-06-13).
- (112) Taylor, J. M. G.; Ankerst, D. P.; Andridge, R. R. Validation of biomarker-based risk prediction models. *Clin. Cancer Res.* **2008**, *14* (19), 5977–5983. <https://doi.org/10.1158/1078-0432.CCR-07-4534>.
- (113) Stehman, S. V. Selecting and Interpreting Measures of Thematic Classification Accuracy. *Remote Sens. Environ.* **1997**, *62* (1), 77–89. [https://doi.org/10.1016/S0034-4257\(97\)00083-7](https://doi.org/10.1016/S0034-4257(97)00083-7).
- (114) Carracedo-Reboredo, P.; Liñares-Blanco, J.; Rodríguez-Fernández, N.; Cedrón, F.; Novoa, F. J.; Carballal, A.; Maojo, V.; Pazos, A.; Fernandez-Lozano, C. A review on machine learning approaches and trends in drug discovery. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4538–4558. <https://doi.org/10.1016/j.csbj.2021.08.011>.
- (115) Mauri, A.; Consonni, V.; Todeschini, R. Molecular Descriptors. Em *Handbook of Computational Chemistry*, Leszczynski, J., Kaczmarek-Kedziera, A., Puzyn, T., Papadopoulos, M. G., Reis, H., Shukla, M. K., Eds.; Springer International Publishing Switzerland, 2017; p 2065. <https://doi.org/10.1007/978-3-319-27282-5>.
- (116) Petitjean, M. Applications of the Radius-Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (4), 331–337. <https://doi.org/10.1021/ci00008a012>.
- (117) Kier, L. B.; Hall, L. H. Molecular Connectivity in Chemistry and Drug Research. Em *Medicinal Chemistry: A Series of Monographs*; 1976. <https://doi.org/10.1016/b978-1-4832-3026-9.50001-7>.
- (118) *QuaSAR-Descriptor*. <https://cadaster.eu/sites/cadaster.eu/files/challenge/descr.htm> (acedido

- 2022-10-04).
- (119) Mauri, A.; Consonni, V.; Todeschini, R. Molecular Descriptors. Em *Springer International Publishing*, Leszczynski, J., Kaczmarek-Kedziera, A., Puzyn, T., Papadopoulos, M. G., Reis, H., Shukla, M. K., Eds.; 2017; pp 2065–2381. <https://doi.org/10.1007/978-3-319-27282-5>.
- (120) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. Em *Reviews in Computational Chemistry*, Lipkowitz, K. B., Boyd, D. B., Eds.; 2007; Vol. 2, pp 367–422. <https://doi.org/10.1002/9780470125793.ch9>.
- (121) Leach, A. R.; Gillet, V. J. *An introduction to chemoinformatics*, 2007. <https://doi.org/10.1007/978-1-4020-6291-9>.
- (122) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 868–873. <https://doi.org/10.1021/ci990307l>.
- (123) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43* (20), 3714–3717. <https://doi.org/10.1021/jm000942e>.
- (124) Labute, P. Derivation and applications of molecular descriptors based on approximate surface area. Em *Methods in Molecular Biology*, Bajorath, J., Ed.; 2004; Vol. 275, pp 261–278. <https://doi.org/10.1385/1-59259-802-1:261>.
- (125) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **1980**, *36* (22), 3219–3228. [https://doi.org/10.1016/0040-4020\(80\)80168-2](https://doi.org/10.1016/0040-4020(80)80168-2).
- (126) Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*, Taylor & Francis Group, Ed.; 1999.
- (127) Kier, L. B.; Hall, L. H. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7* (8), 801–807. <https://doi.org/10.1023/A:1015952613760>.
- (128) Hall, L. H.; Kier, L. B. The E-State as the Basis for Molecular Structure Space Definition and Structure Similarity. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (3), 784–791. <https://doi.org/10.1021/ci990140w>.
- (129) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. Em *Three-Dimensional Quantitative Structure Activity Relationships*, Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; 2002; Vol. 2, pp 339–353. https://doi.org/10.1007/0-306-46857-3_18.
- (130) Fourches, D.; Ash, J. 4D- quantitative structure–activity relationship modeling: making a comeback. *Expert Opin. Drug Discov.* **2019**, *14* (12), 1227–1235. <https://doi.org/10.1080/17460441.2019.1664467>.
- (131) Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. <https://doi.org/10.1021/ci00057a005>.
- (132) Alves, V.; Braga, R.; Muratov, E.; Andrade, C. Quimioinformática: Uma Introdução. *Quim. Nova* **2018**, *41* (2), 202–212. <https://doi.org/10.21577/0100-4042.20170145>.
- (133) Jorgensen, J. H.; Ferraro, M. J. Antimicrobial susceptibility testing: A review of general principles and contemporary practices. *Clin. Infect. Dis.* **2009**, *49* (11), 1749–1755. <https://doi.org/10.1086/647952>.

- (134) Balouiri, M.; Sadiki, M.; Ibsouda, S. K. Methods for in vitro evaluating antimicrobial activity: A review. *J. Pharm. Anal.* **2016**, *6* (2), 71–79. <https://doi.org/10.1016/j.jpha.2015.11.005>.
- (135) Bayot, M. L.; Bragg, B. N. Antimicrobial susceptibility testing. Em *Clinical Microbiology Procedures Handbook*; Garcia, L. S., Isenberg, H. D., Eds.; StatPearls Publishing, 2010. <https://doi.org/10.1128/9781555817435.ch5>.
- (136) Gajic, I.; Kabic, J.; Kekic, D.; Jovicevic, M.; Milenkovic, M.; Mitic Culafic, D.; Trudic, A.; Ranin, L.; Opavski, N. Antimicrobial Susceptibility Testing: A Comprehensive Review of Currently Used Methods. *Antibiotics* **2022**, *11* (4), 427. <https://doi.org/10.3390/antibiotics11040427>.
- (137) Vasoo, S. Susceptibility testing for the polymyxins: Two steps back, three steps forward? *J. Clin. Microbiol.* **2017**, *55* (9), 2573–2582. <https://doi.org/10.1128/JCM.00888-17>.
- (138) EUCAST. *Recommendations for MIC determination of colistin (polymyxin E) As recommended by the joint CLSI-EUCAST Polymyxin Breakpoints Working Group*. EUCAST. https://www.eucast.org/eucast_news/news_singleview?tx_ttnews%5Btt_news%5D=171&cHash=387b192a5e9d82d2e74ae72fdb456602 (acedido 2023-01-25).
- (139) Kowalska-Krochmal, B.; Dudek-Wicher, R. The minimum inhibitory concentration of antibiotics: Methods, interpretation, clinical relevance. *Pathogens* **2021**, *10* (2), 165. <https://doi.org/10.3390/pathogens10020165>.
- (140) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* **2019**, *8*, 1102–1109. <https://doi.org/10.1093/nar/gky1033>.
- (141) Kim, S. Exploring Chemical Information in PubChem. *Curr. Protoc.* **2021**, *1*, 217. <https://doi.org/10.1002/cpz1.217>.
- (142) Chen, X.; Reynolds, C. H. Performance of similarity measures in 2D fragment-based similarity searching: Comparison of structural descriptors and similarity coefficients. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1407–1414. <https://doi.org/10.1021/ci025531g>.
- (143) *Pandas*. <https://pandas.pydata.org/> (acedido 2023-02-17).
- (144) *NumPy*. <https://numpy.org/> (acedido 2023-02-17).
- (145) *Matplotlib – Visualization with Python*. <https://matplotlib.org/> (acedido 2023-02-17).
- (146) *Seaborn: statistical data visualization*. <http://seaborn.pydata.org/index.html> (acedido 2023-02-17).
- (147) *Getting Started with the RDKit in Python*. <https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors> (acedido 2022-12-21).
- (148) Stanton, D. T. Evaluation and use of BCUT descriptors in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (1), 11–20. <https://doi.org/10.1021/ci980102x>.
- (149) *1.10. Decision Trees – scikit-learn 1.2.2 documentation*. <https://scikit-learn.org/stable/modules/tree.html> (acedido 2023-04-05).
- (150) *1.11. Ensemble methods – scikit-learn 1.2.2 documentation*. <https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees> (acedido 2023-04-05).
- (151) *EUCAST*. <https://www.eucast.org/> (acedido 2023-04-03).
- (152) European Committee on Antimicrobial Susceptibility Testing. *Media preparation for EUCAST disk*

- diffusion testing and for determination of MIC values by the broth microdilution method*, 2022.
- (153) Brooks, G.; Carrol, K.; Butel, J.; Morse, S.; Timothy, M. *Microbiologia médica de Jawetz, Melnick e Adelberg*, 26.^a ed.; AMGH Editora Ltda, Ed.; 2014.
- (154) Kwiecinski, J. M.; Horswill, A. R. Staphylococcus aureus bloodstream infections: pathogenesis and regulatory mechanisms. *Curr. Opin. Microbiol.* **2020**, *53*, 51–60. <https://doi.org/10.1016/j.mib.2020.02.005>.
- (155) Allocati, N.; Masulli, M.; Alexeyev, M. F.; Di Ilio, C. Escherichia coli in Europe: An overview. *Int. J. Environ. Res. Public Heal.* **2013**, *10* (12), 6235–6254. <https://doi.org/10.3390/ijerph10126235>.
- (156) Maurice, N. M.; Bedi, B.; Sadikot, R. T. Pseudomonas aeruginosa biofilms: Host response and clinical implications in lung infections. *Am. J. Respir. Cell Mol. Biol.* **2018**, *58* (4), 428–439. <https://doi.org/10.1165/rcmb.2017-0321TR>.
- (157) Chicco, D.; Tötsch, N.; Jurman, G. The matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min.* **2021**, *14*, 1–22. <https://doi.org/10.1186/s13040-021-00244-z>.
- (158) Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2.^a ed.; 2022.
- (159) European Committee on Antimicrobial Susceptibility Testing. *Breakpoint tables for interpretation of MICs and zone diameters Version 12.0*. https://www.eucast.org/ast_of_bacteria/previous_versions_of_documents.
- (160) The European Committee on Antimicrobial Susceptibility Testing. *Routine and extended internal quality control for MIC determination and disk diffusion as recommended by EUCAST. Version 12.0*, 2022. <http://www.eucast.org>.
- (161) Sader, H. S.; Rhomberg, P. R.; Farrell, D. J.; Jones, R. N. Differences in potency and categorical agreement between colistin and polymyxin B when testing 15,377 clinical strains collected worldwide. *Diagn. Microbiol. Infect. Dis.* **2015**, *83* (4), 379–381. <https://doi.org/10.1016/j.diagmicrobio.2015.08.013>.
- (162) Medina, J.; Daniela, P.; Noceti, O.; Rieppi, G. R. Actualización acerca de colistina (polimixina E): aspectos clínicos, PK/PD y equivalencias. *Rev. Méd. Urug.* **2017**, *33* (3), 195. <https://doi.org/10.29193/rmu.33.3.5>.
- (163) Chew, K. L.; La, M.-V.; Lin, R. T. P.; Teo, J. W. P. Colistin and Polymyxin B Susceptibility Testing for Carbapenem-Resistant and mcr-Positive Enterobacteriaceae: Comparison of Sensititre, MicroScan, Vitek 2, and Etest with Broth Microdilution. *J. Clin. Microbiol.* **2017**, *55* (9), 2609–2616. <https://doi.org/10.1128/JCM.00268-17>.
- (164) Aquilini, E.; Merino, S.; Knirel, Y. A.; Regué, M.; Tomás, J. M. Functional identification of Proteus mirabilis eptC gene encoding a core lipopolysaccharide phosphoethanolamine transferase. *Int. J. Mol. Sci.* **2014**, *15* (4), 6689–6702. <https://doi.org/10.3390/ijms15046689>.
- (165) Wang, W. B.; Chen, I. C.; Jiang, S. S.; Chen, H. R.; Hsu, C. Y.; Hsueh, P. R.; Hsu, W. Bin; Liaw, S. J. Role of RppA in the regulation of polymyxin B susceptibility, swarming, and virulence factor expression in Proteus mirabilis. *Infect. Immun* **2008**, *76* (5), 2051–2062. <https://doi.org/10.1128/IAI.01557-07>.

7. ANEXOS

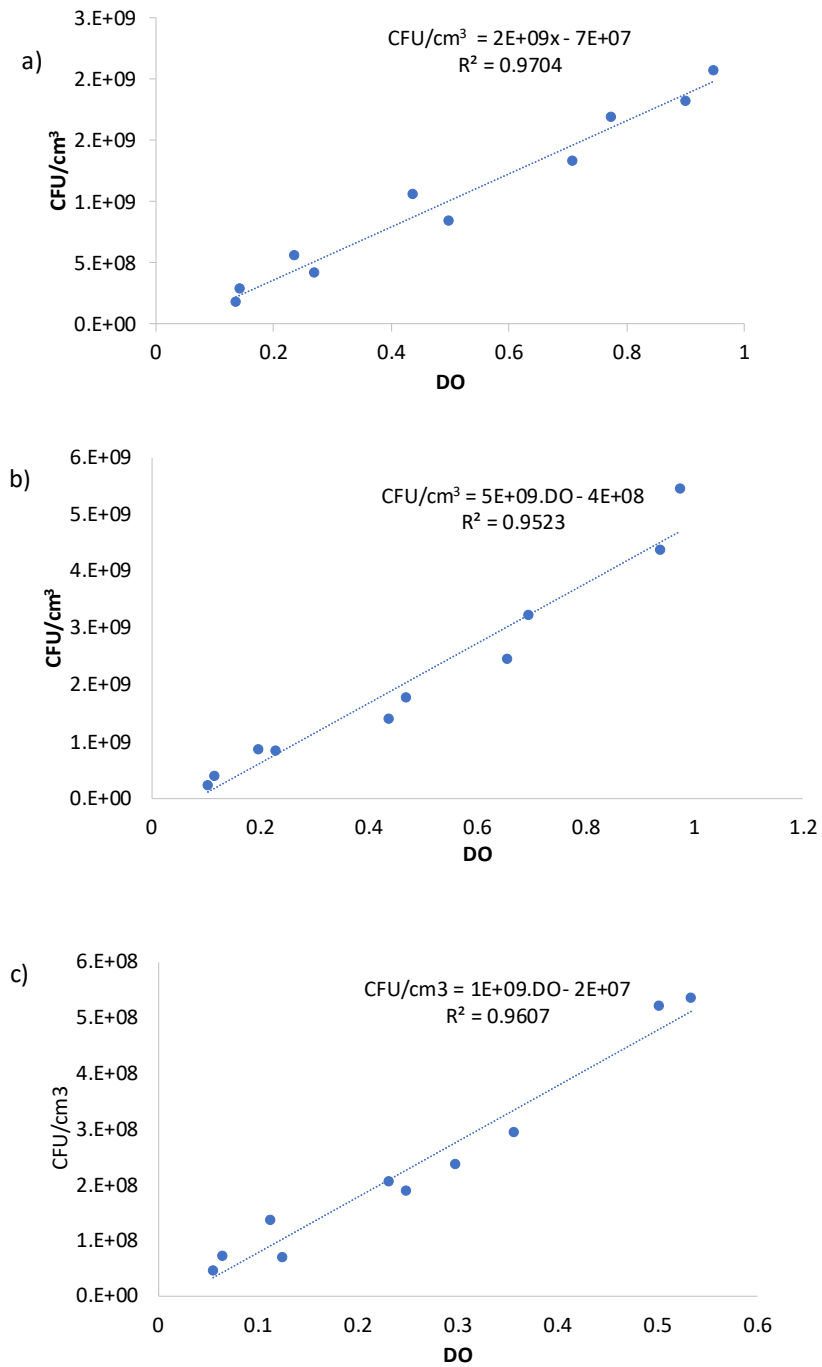


Figura A1. Curvas de calibração da concentração celular vs DO para: *S. sonnei* (a), *P. mirabilis* (b) e *L. monocytogenes* (c).

Tabela A1. Resultados detalhados dos modelos da primeira série

Algoritmos	Família de descritores moleculares	Conjunto treino			Conjunto teste		
		Exatidão	Q1 Q1	Q1 Q4	Exatidão	Q1 Q1	Q1 Q4
AdaBoost	AC2D	0,75	0,72	0,05	0,60	0,54	0,17
DT	AC2D	0,76	0,78	0,05	0,63	0,63	0,08
RF	AC2D	0,76	0,71	0,02	0,63	0,51	0,08
AdaBoost	BCUT2D	0,76	0,78	0,05	0,64	0,66	0,12
DT	BCUT2D	0,76	0,78	0,05	0,63	0,63	0,08
RF	BCUT2D	0,76	0,72	0,05	0,63	0,54	0,12
AdaBoost	CKP	0,75	0,77	0,02	0,61	0,63	0,08
DT	CKP	0,75	0,78	0,05	0,59	0,60	0,12
RF	CKP	0,74	0,69	0,05	0,62	0,49	0,12
AdaBoost	Estate_VSA	0,75	0,77	0,02	0,63	0,66	0,17
DT	Estate_VSA	0,75	0,78	0,05	0,60	0,63	0,12
RF	Estate_VSA	0,75	0,71	0,02	0,60	0,49	0,21
AdaBoost	FG	0,50	0,38	0,07	0,47	0,29	0,08
DT	FG	0,68	0,89	0,09	0,47	0,60	0,12
RF	FG	0,67	0,78	0,09	0,47	0,43	0,08
AdaBoost	Gen	0,76	0,78	0,05	0,60	0,51	0,08
DT	Gen	0,76	0,78	0,05	0,58	0,54	0,17
RF	Gen	0,75	0,74	0,05	0,61	0,49	0,08
AdaBoost	Hb	0,66	0,80	0,16	0,53	0,63	0,17
DT	Hb	0,66	0,83	0,16	0,53	0,63	0,17
RF	Hb	0,66	0,78	0,16	0,54	0,51	0,21
AdaBoost	PEOE_VSA	0,75	0,75	0,02	0,61	0,57	0,12
DT	PEOE_VSA	0,75	0,78	0,05	0,60	0,60	0,17
RF	PEOE_VSA	0,75	0,69	0,02	0,62	0,51	0,12

Algoritmos	Família de descritores moleculares	Conjunto treino			Conjunto teste		
		Exatidão	Q1 Q1	Q1 Q4	Exatidão	Q1 Q1	Q1 Q4
AdaBoost	SLogP_VSA	0,75	0,78	0,05	0,64	0,63	0,08
DT	SLogP_VSA	0,75	0,78	0,05	0,60	0,60	0,08
RF	SLogP_VSA	0,75	0,71	0,02	0,60	0,51	0,17
AdaBoost	SMR_VSA	0,74	0,75	0,03	0,61	0,54	0,12
DT	SMR_VSA	0,74	0,78	0,05	0,64	0,60	0,12
RF	SMR_VSA	0,74	0,69	0,02	0,59	0,37	0,08

Tabela A2. Otimização dos hiper-parâmetros nos modelos RF e AdaBoost da primeira série

Algoritmos	Família de descritores moleculares	Hiper-parâmetros				
		n_{est}	n_s	n_r	r_L	d_{est}
RF	AC2D	20	1	1	NA	NA
AdaBoost	AC2D	10	NA	NA	1	100
RF	BCUT2D	20	0,78	0,26	NA	NA
AdaBoost	BCUT2D	10	NA	NA	0,1	10
RF	CKP	25	0,78	0,16	NA	NA
AdaBoost	CKP	10	NA	NA	0,01	10
RF	Estate_VSA	25	0,78	0,58	NA	NA
AdaBoost	Estate_VSA	10	NA	NA	0,01	10
RF	FG	10	0,78	0,89	NA	NA
AdaBoost	FG	5	NA	NA	2	2
RF	Gn	25	0,78	0,47	NA	NA
AdaBoost	Gn	9	NA	NA	0,01	10
RF	Hb	25	0,6	0,58	NA	NA

Hiper-parâmetros						
Algoritmos	Família de descritores moleculares	n_{est}	n_s	n_r	r_L	d_{est}
AdaBoost	Hb	8	NA	NA	0,01	10
RF	PEOE_VSA	20	0,78	0,47	NA	NA
AdaBoost	PEOE_VSA	10	NA	NA	0,01	10
RF	SLogP_VSA	20	0,78	0,89	NA	NA
AdaBoost	SLogP_VSA	50	NA	NA	0,95	100
RF	SMR_VSA	15	1	0,89	NA	NA
AdaBoost	SMR_VSA	10	NA	NA	0,01	10

Tabela A3. Otimização dos hiper-parâmetros nos modelos RF e AdaBoost da segunda série

Hiper-parâmetros						
Algoritmos	Família de descritores moleculares	n_{est}	n_s	n_r	r_L	d_{est}
AdaBoost	AC2D	10	NA	NA	0,32	5
RF	AC2D	20	0,78	0,68	NA	NA
AdaBoost	BCUT2D	75	NA	NA	0,64	5
RF	BCUT2D	20	0,78	0,47	NA	NA
AdaBoost	CKP	60	NA	NA	0,01	5
RF	CKP	100	0,78	0,89	NA	NA
AdaBoost	ESTATE_VSA	100	NA	NA	0,53	2
RF	ESTATE_VSA	10	0,78	0,89	NA	NA
AdaBoost	FG	20	NA	NA	0,74	5
RF	FG	100	0,10	0,16	NA	NA
AdaBoost	Gen	75	NA	NA	1,267	5
RF	Gen	20	1	0,79	NA	NA
AdaBoost	Hb	75	NA	NA	0,43	2
RF	Hb	100	0,78	0,79	NA	NA

Hiper-parâmetros						
Algoritmos	Família de descritores moleculares	n_{est}	n_s	n_r	r_L	d_{est}
AdaBoost	PEOE_VSA	5	NA	NA	0,01	10
RF	PEOE_VSA	20	1	0,68	NA	NA
AdaBoost	SLOGP_VSA	10	NA	NA	0,01	10
RF	SLOGP_VSA	100	0,78	0,79	NA	NA
AdaBoost	SMR_VSA	10	NA	NA	0,01	10
RF	SMR_VSA	20	0,79	0,89	NA	NA

Tabela A4. Resultados detalhados dos modelos da segunda série

Algoritmos	Família de descritores moleculares	Conjunto treino			Conjunto teste		
		Exatidão	Q1 Q1	Q1 Q4	Exatidão	Q1 Q1	Q1 Q4
AdaBoost	AC2D	0,80	0,69	0	0,60	0,43	0,043
DT	AC2D	0,82	0,86	0	0,64	0,68	0,043
RF	AC2D	0,81	0,81	0	0,66	0,52	0,087
AdaBoost	BCUT2D	0,80	0,69	0	0,60	0,39	0,043
DT	BCUT2D	0,82	0,86	0	0,65	0,68	0,043
RF	BCUT2D	0,81	0,79	0	0,65	0,57	0,043
AdaBoost	CKP	0,80	0,79	0,02	0,68	0,61	0,043
DT	CKP	0,82	0,86	0	0,67	0,64	0,043
RF	CKP	0,82	0,79	0	0,65	0,57	0,043
AdaBoost	Estate_VSA	0,68	0,76	0,02	0,58	0,57	0,043
DT	Estate_VSA	0,82	0,86	0	0,68	0,69	0,043
RF	Estate_VSA	0,78	0,74	0,02	0,63	0,61	0,043
AdaBoost	FG	0,72	0,48	0	0,57	0,13	0,043

Algoritmos	Família de descritores moleculares	Conjunto treino			Conjunto teste		
		Exatidão	Q1 Q1	Q1 Q4	Exatidão	Q1 Q1	Q1 Q4
DT	FG	0,73	0,62	0	0,64	0,26	0,043
RF	FG	0,62	0,38	0,02	0,60	0,17	0,043
AdaBoost	Gen	0,79	0,64	0,02	0,63	0,35	0,043
DT	Gen	0,82	0,86	0	0,67	0,65	0,043
RF	Gen	0,81	0,83	0,02	0,67	0,65	0,043
AdaBoost	Hb	0,68	0,74	0	0,64	0,65	0,043
DT	Hb	0,74	0,79	0	0,65	0,65	0,043
RF	Hb	0,74	0,71	0	0,65	0,61	0,043
AdaBoost	PEOE_VSA	0,79	0,79	0	0,69	0,65	0,043
DT	PEOE_VSA	0,81	0,86	0	0,67	0,62	0,043
RF	PEOE_VSA	0,81	0,83	0,02	0,65	0,65	0,043
AdaBoost	SLogP_VSA	0,80	0,83	0	0,65	0,65	0,043
DT	SLogP_VSA	0,81	0,86	0	0,66	0,68	0,043
RF	SLogP_VSA	0,81	0,79	0	0,64	0,61	0,043
AdaBoost	SMR_VSA	0,80	0,86	0	0,67	0,65	0,043
DT	SMR_VSA	0,81	0,86	0	0,65	0,65	0,043
RF	SMR_VSA	0,80	0,79	0,02	0,67	0,61	0,043

Tabela A5. Resultados detalhados da aplicação do modelo DT/Estate_VSA

Moléculas	Alvos											
	<i>Pseudomonas</i>				<i>Acinetobacter</i>				<i>Escherichia</i>			
	%Q1	%Q2	%Q3	%Q4	%Q1	%Q2	%Q3	%Q4	%Q1	%Q2	%Q3	%Q4
Bmim1	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Bmim2	0	57,1	42,9	0	16,7	0	83,3	0	0	0	100	0
Bmim3	0	57,1	42,9	0	16,7	0	83,3	0	0	0	100	0
Emim1	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Emim2	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Emim3	0	57,1	42,9	0	16,7	0	83,3	0	0	0	100	0
Gly01	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Gly02	0	100	0	0	16,7	0	83,3	0	0	0	100	0
Gly03	0	100	0	0	16,7	0	83,3	0	0	100	0	0
Gly05	0	100	0	0	16,7	0	83,3	0	0	100	0	0
Gly06	0	100	0	0	16,7	0	83,3	0	0	100	0	0
Gly07	0	100	0	0	16,7	0	83,3	0	0	100	0	0
Gly08	0	100	0	0	16,7	0	83,3	0	0	100	0	0
Gly09	0	100	0	0	16,7	0	83,3	0	0	100	0	0
Gly10	0	100	0	0	16,7	0	83,3	0	0	0	100	0
Glu01	0	100	0	0	16,7	0	83,3	0	0	0	100	0
Glu02	0	100	0	0	16,7	0	83,3	0	0	0	100	0
Glu03	0	100	0	0	16,7	0	83,3	0	0	0	100	0
Glu05	0	100	0	0	16,7	0	83,3	0	0	0	100	0
Glu06	0	100	0	0	16,7	0	83,3	0	0	0	100	0
Glu07	0	100	0	0	16,7	0	83,3	0	0	0	100	0
Glu08	0	100	0	0	16,7	0	83,3	0	0	0	100	0
Glu09	0	100	0	0	16,7	0	83,3	0	0	0	100	0
Glu10	0	100	0	0	16,7	0	83,3	0	0	0	100	0

Moléculas	Alvos											
	<i>Pseudomonas</i>				<i>Acinetobacter</i>				<i>Escherichia</i>			
	%Q1	%Q2	%Q3	%Q4	%Q1	%Q2	%Q3	%Q4	%Q1	%Q2	%Q3	%Q4
Leu01	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Leu02	0	100	0	0	16,7	0	83,3	0	0	0	100	0
Leu03	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Leu05	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Leu06	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Leu07	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Leu08	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Leu09	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Leu10	0	100	0	0	16,7	0	83,3	0	0	0	100	0
Lys01	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Lys02	0	100	0	0	16,7	0	83,3	0	0	0	100	0
Lys03	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Lys05	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Lys06	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Lys07	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Lys08	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Lys09	0	57,1	42,9	0	16,7	0	83,3	0	10,5	57,9	31,6	0
Lys10	0	100	0	0	16,7	0	83,3	0	0	0	100	0

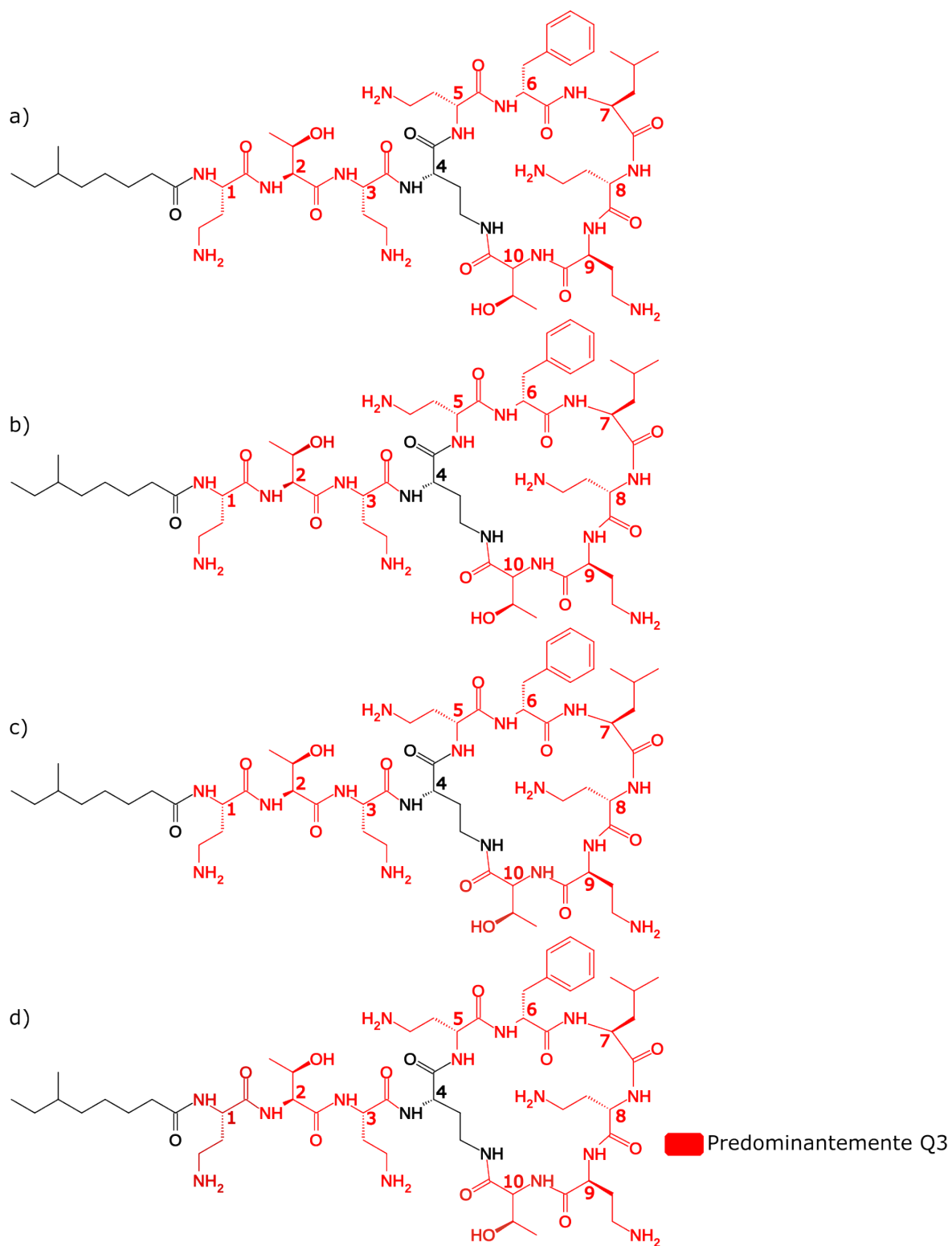


Figura A2. Classificação mais provável quanto à atividade antimicrobiana contra *Acinetobacter* de variantes mutadas da polimixina B ao alterar sistematicamente cada resíduo de aminoácido para: Gly (a), Leu (b), Lys (c) e Glu (d).

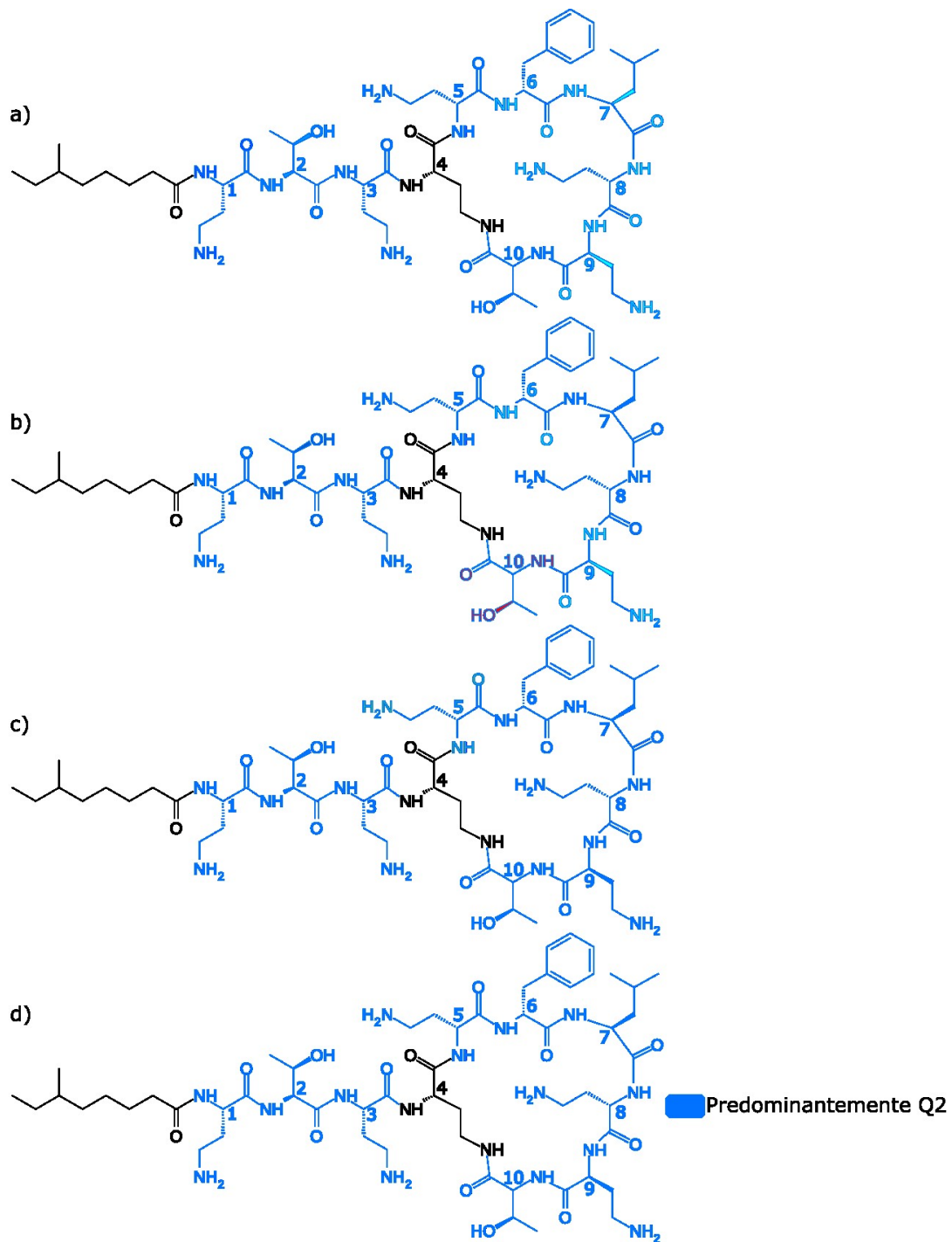


Figura A3. Classificação mais provável quanto à atividade antimicrobiana contra *Pseudomonas* de variantes mutadas da polimixina B ao alterar sistematicamente cada resíduo de aminoácido para: Gly (a), Leu (b), Lys (c) e Glu (d).