[Rui Pedro Santos Patrício]

[Licenciado em Bioquímica]

Desenvolvimento de abordagem de *design* de fármacos através de métodos computacionais para descoberta de ativadores da proteína p53 para terapia anti cancro

Dissertação para obtenção do Grau de Mestre em Bioquímica

Orientador: Doutora Florbela Pereira
Investigadora Pós-Doc, LAQV-REQUIMTE-FCT/NOVA

Co-orientadores: Doutora Paula Videira
Investigadora Pós-Doc, UCIBIO-REQUIMTE-FCT NOVA

Composição do Júri:

Presidente: Professor Doutor Pedro António de Brito Tavares

Vogais: Doutora Florbela Bento Martinho de Sá Pereira

Professor Doutor João Montargil Aires de Sousa

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Maio, 2020

[Rui Pedro Santos Patrício]

[Licenciado em Bioquímica]

Desenvolvimento de abordagem de *design* de fármacos através de métodos computacionais para descoberta de ativadores da proteína p53 para terapia anti cancro

Dissertação para obtenção do Grau de Mestre em
Bioquímica

Orientador: Doutora Florbela Pereira
Investigadora Pós-Doc, LAQV-REQUIMTE-FCT/NOVA

Co-orientadores: Doutora Paula Videira
Investigadora Pós-Doc, UCIBIO-REQUIMTE-FCT NOVA

Composição do Júri:

Presidente: Professor Doutor Pedro António de Brito Tavares

Vogais: Doutora Florbela Bento Martinho de Sá Pereira

Professor Doutor João Montargil Aires de Sousa

Maio, 2020

Direitos de autor

Dedico a todos aqueles que lutam todos os dias por um mundo melhor.

“Por que foi que cegámos, Não sei, Talvez um dia se chegue a conhecer a razão, Queres que te diga o que penso, Diz, Penso que não cegámos, penso que estamos cegos, Cegos que veem, Cegos que, vendo, não veem.”

José Saramago, em “Ensaio sobre a Cegueira”

Agradecimentos

Em primeiro lugar, gostaria de agradecer às minhas orientadoras: à professora Florbela Pereira por me ter recebido e ajudado em todo o processo, por ter compreendido todas as circunstâncias envolventes e por me ter apoiado numa área que não era, de todo, a minha "praia"; à professora Paula Videira por me ter recebido de igual forma, sempre disposta a ajudar e compreender os conceitos, por me ter incluído no seu grupo e por dar a possibilidade de aprender mais.

Gostaria de agradecer à Diana Sousa por me ter acompanhado durante a parte experimental da minha dissertação, por toda a ajuda, pela paciência, pelos conselhos e pela capacidade espetacular que tem em orientar pessoas.

Aos restantes colegas do grupo de Glicoimunologia, Zélia Silva, Danielle Almeida, Carlota Pascoal, Rita Francisco, Rita Lourenço, Michela Pucci e Gonçalo Mineiro, por todo o apoio e por todos os momentos de diversão que me proporcionaram.

À Mariana Cruzeiro, por me ter dado a possibilidade de a orientar no seu projeto de licenciatura, por me ter aturado e por me ter ajudado nas minhas (e também suas) dúvidas.

Aos meus amigos, sempre presentes, como em todas as fases da minha "carreira" universitária, Luís Pinheiro, Marta Furtado, Mariana Carmona, Margarida Custódio, Madalena Ayala, Catarina Pereira e Inês Anjos, para que saibam que os levarei sempre comigo e que estão sempre no meu pensamento. Agradeço-lhes todo o apoio e por me terem ajudado sempre que precisei.

À Cátia e ao Ricardo por me terem "abrigado" quando precisei, por me terem apoiado e por estarem sempre comigo nos momentos importantes. Agradeço pelas dores de cabeça e peço desculpa pelas dores de cabeça que lhes causei. Ao Senhor Zé por me ter "abrigado" por tempo indeterminado.

À Eva, a nova estrela da minha vida, pela oportunidade única de poder ser um tio/padrinho numa fase tão crucial e importante da minha vida.

À minha mãe, Isabel, por me ter sempre permitido seguir os meus sonhos, por me ter ajudado e compreendido e por ter estado sempre do meu lado em todo e qualquer momento ao longo de todos estes anos.

À minha fiel companheira, Carlota, por me aturar (quase) todos os dias, por me dar na cabeça, por me ajudar, por ralhar, por estar sempre lá quando mais preciso. Também sem ela, tudo isto teria sido muito mais difícil.

A toda a minha família, no geral (senão precisaria de mais duas páginas, no mínimo), por estarem sempre lá quando preciso e por me apoiarem em todos os momentos da minha vida. À Xana e aos restantes membros desta família maravilhosa que sempre me acolheu e sempre me ajudou em tudo.

Por último, gostaria de agradecer à REQUIMTE-FCT/NOVA, especialmente ao LAQV e à UCIBIO pela oportunidade de poder fazer este trabalho. Também gostaria de agradecer à Faculdade de Ciências da Universidade Nova de Lisboa por me ter deixado fazer parte desta experiência e por promover sempre os interesses em relação à pesquisa científica e por

estimular os estudantes a serem inovadores e a perseguirem os seus sonhos. Um último agradecimento à FCT/NOVA por ter sido a minha casa nos últimos 5 anos e meio.

Resumo

O cancro colorretal (CCR) é o terceiro tipo de cancro mais comum e a quarta maior causa de morte no mundo. A proteína p53 é um fator de transcrição induzível pelo stress, que regula um largo número de genes com a função de regular múltiplos processos de sinalização. Foi criada uma base de dados (com 10.505 moléculas) cujos dados nos davam informação acerca de atividade sobre a p53 (ativação ou inativação), dados estes que foram usados para construção de modelos usando várias técnicas de aprendizagem automática, tais como *Random Forest*, *Support Vector Machine*, Redes neuronais artificiais, *k-Nearest Neighbors* e Redes neuronais de *Kohonen*, sendo que com o algoritmo da *Random Forest* se obtiveram os melhores resultados (*Random Forest* com seleção dos 150 melhores descritores moleculares da categoria dos CDK/*Fingerprinter*).

A performance do modelo permitiu a distinção entre compostos ativos e inativos (classes utilizadas na construção do modelo de classificação) e foi avaliada internamente (com um conjunto de treino, usado para a construção do modelo, em validação cruzada) e externamente (esta através de um conjunto de teste), com uma previsibilidade geral (Q) de 0,808 e 0,814 para o treino e para o teste, respetivamente. Usando este modelo foi possível efetuar um *screening* virtual a partir da base de dados do ZINC de fármacos aprovados pela FDA (1442 compostos).

Foi selecionada uma lista de fármacos aprovados que apresentam maior probabilidade de serem ativadores (direta ou indiretamente) da p53, tendo como base vários limites estabelecidos nesta abordagem, designadamente: (1) probabilidade de ser ativo contra p53; (2) domínio de aplicabilidade do modelo; e (3) previsão da afinidade, em kcal/mol, para o alvo p53 através da técnica de *docking* molecular. O composto mais promissor - dihidroergocristina -, no que diz respeito a estes três pontos, encontra-se atualmente sob validação experimental.

Palavras-chave:

p53; Técnicas de aprendizagem automática; QSAR, *Screening* Virtual; *Docking*; HT29

Abstract

Colorectal cancer (CRC) is the third most common type of cancer and the fourth major cause of death in the world. The p53 protein is a transcriptional factor stress inducible that regulates a large number of genes and a lot of multiple signalling processes. It was created a data base (with 10.505 molecules) which data gave us information about the activity over p53 (activation or inactivation), that was used to build models using machine learning techniques as Random Forest, Support Vector Machine, Artificial Neural Networks, k-Nearest Neighbors and Kohonen neural networks.

The best algorithm was built with the Random Forest (with selection of the 150 best descriptors, selected by Random Forest, within the Fingerprint set and using CDK/Fingerprinter)

The model performance permitted the distinction between activated and inactivated compounds (classes that were used in the building of the classification model) and was evaluated internally (with the training set used to build the model, in cross validation) and externally (with a test set), obtaining an overall predictive activity of 0,808 and 0,814 to the training set and test set, respectively. Using the best model, it was possible to do a virtual screening test using approved drugs by FDA present in ZINC data base (1442 compounds).

It was used a list of approved drugs that presents a major probability of being activators (direct or indirectly) of p53, assented on some limits stablished in this approach, such as: (1) probability of being active against p53; (2) applicability domain of the model; (3) Prediction of the affinity (in kcal/mol) for the target, p53, through molecular docking.

The most promising molecule – dihydroergocristine –, with regard of the limits stablished above, is under experimental validation at this moment.

Keywords:

p53; Machine Learning Techniques; QSAR, Virtual Screening; Docking; HT29

Índice de matérias

Índice de Figuras

Índice de Tabelas

Índice de equações

Lista de abreviaturas e siglas

ANN- Artificial Neural Network
ASD- Average SOM Distance
CADD- Computer Aided Drug Design
Csc- Cost Sensitive Classifier
DA- Domínio de aplicabilidade
DHEC- Dihidroergocristina
DHET- Dihidroergotamina
DMSO- Dimetilsulfóxido
FDA- Food and Drug Administration
FN- Falsos Negativos
FP- Falsos Positivos
HTS- High-Throughput Screening
InChI- International Chemical Identifier
indW- Pesos individuais
ineW- Pesos de inércia
LB- Ligand-based
MAE- Erro máximo absoluto
MCC- Coeficiente de Correlação de Matthews
MDL- Molecular Designed Limited
MDM_{2-} Murine Double Minute 2
ML- Machine Learning
MLP- Multilayer Perceptron
MOL- Molecule file
m_{try}- Número de descritores utilizados em cada nódulo da Random Forest
NBO- Natural bond orbital
nF- nFeatures
nIt- nIterations
NN- Neural Network
n_{tree}- Número de árvores de decisão utilizadas na RandomForest
OOB- OutOfBag
PDB- Protein Data Bank
Q- Previsibilidade
QSAR- Relações quantitativas estrutura-atividade
QSPR- Relações quantitativas estrutura-propriedade
R^2- Quadrado do coeficiente de correlação de Pearson
R&D- Descoberta e desenvolvimento
RBF- Radial Basis Function
RF- Random Forest
RMN- Ressonância Magnética Nuclear
$RPMI_c$- Meio RPMI completo
SB- Structure-based
SDF- Structure data file
SE- Sensibilidade

socW- Pesos sociais
SOM- Self-Organized Maps
SP- Especificidade
SV- Screening Virtual
SVM- Support Vector Machine
TE- Tripsina-EDTA
TN- Verdadeiros negativos
TP- Verdadeiros Positivos
W- Peso

Capítulo 1 - INTRODUÇÃO

1.1 Contextualização

O cancro colorretal é o terceiro tipo de cancro mais comum no género masculino, sendo o segundo mais frequente relativamente ao género feminino. Em 2018, foram registados cerca de 1,8 milhões de novos casos, segundo dados do *World Cancer Research Fund*. No que se refere à incidência deste tipo de cancro, Portugal encontra-se no sexto lugar no que diz respeito a estatísticas de ambos os sexos, em quinto lugar no que diz respeito ao sexo masculino e no décimo segundo lugar em relação ao sexo feminino. Dados fornecidos pelo *National Cancer Institute* sugerem que são esperados 145 600 novos casos e 51 020 mortes em 2019. A percentagem de sobrevivência fixa-se nos 64,4%, segundo dados recolhidos de 2009 a 2015 pelo mesmo instituto.

Figura 1.1 – Estatísticas fornecidas pelo *National Cancer Institute* (visto a 19/07/2019)

O cancro colorretal é considerado a quarta causa de morte por cancro a nível mundial, sendo que, nas últimas décadas, houve um aumento considerável da sua incidência: de 783 000 casos em 1990 para 1 361 000 em 2012. Este tipo de cancro tem crescido em países industrializados com um Índice de Desenvolvimento Humano de moderado a alto. No que ao género diz respeito, este tem maior prevalência no sexo masculino do que no sexo feminino, sendo que o risco de desenvolver esta doença aumenta com a idade. Na verdade, mais de 90% dos pacientes diagnosticados têm mais de 50 anos, considerando-se os 64 anos como a idade média em que é diagnosticado.[1] Perante estes dados, podemos assumir que o estudo deste tipo de cancro é de extrema relevância. A fosfoproteína p53 revela-se muito importante neste tipo de doença, como demonstram os estudos apresentados de seguida. ***Yamaguchi et al., 1992*** reportaram que, em imunoensaios, se verificou a presença de imunoreatividade da p53 em 61% de espécimes de 100 pacientes com cancro coloretal, considerando esta proteína como um marcador biológico importante.[2] Os estudos de ***Scott et al., 1991*** permitiram detetar a proteína em 42% dos 52 adenocarcinomas colorretais, sugerindo que esta tem um papel importante na carcinogénese colorretal.[3]

1.2 p53: função, mecanismos e importância

A fosfoproteína p53 é uma proteína nuclear (também conhecida como supressor de tumor), com cerca de 53kD, que está relacionada com a entrada e a normal progressão do ciclo celular. Esta encontra-se altamente conservada em vertebrados e é induzida durante a transição da fase G_0 para a fase G_1. Este supressor é um fator de transcrição muito potente e que controla a maioria do ciclo celular, protegendo as células das transformações malignas, sendo a proteína com a maior frequência de inatividade em humanos que sofram de cancro. É sabido que a perda de função, algumas alterações oncogénicas e mutações que ocorrem nesta proteína levam a um aumento do desenvolvimento do tumor. O cancro colorretal é caracterizado por uma deleção no cromossoma 17p, perto do *locus* da p53, e foram encontrados níveis elevados desta proteína em 44% dos tumores, através de estudos feitos com radioimunoensaio. Estes estudos sugerem que estes níveis elevados podem estar associados a uma proteína mutada ou à estabilização desta com antigénios virais, enquanto em alguns tumores heterozigóticos a deleção de um alelo da p53 é acompanhada por mutação e sobre-expressão do outro alelo.[3-5]

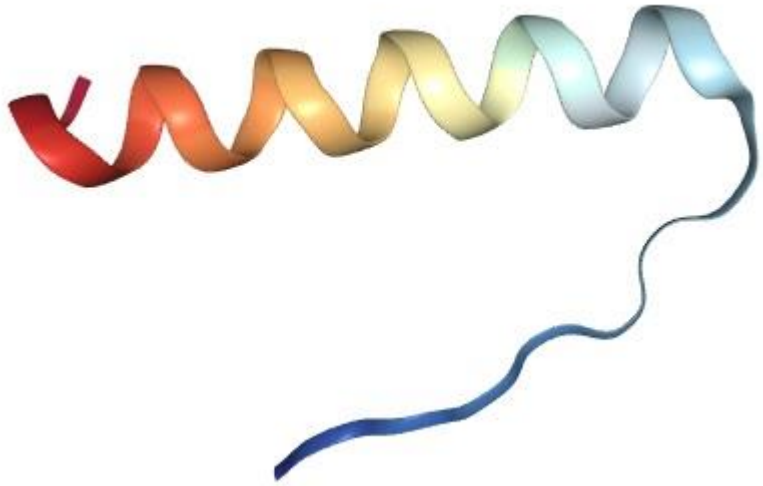

Figura 1.2 – Estrutura 3D da proteína p53 (PDB ID: 1AIE). Fonte: https://www.rcsb.org/3d-view/1AIE

Cerca de metade dos cancros em humanos contêm mutações na proteína, sendo que alguns estudos sugerem que muitos dos tumores que a retêm na forma nativa têm deficiência na capacidade de indução ou na capacidade de resposta à mesma. A p53 está no centro dos mecanismos de resposta ao stress que previnem o crescimento e a sobrevivência de células potencialmente malignas. Os sinais de stress podem ser encontrados em várias etapas durante a tumorigénese, desde a iniciação do tumor à invasão e metástase. Assim, é percetível o papel desta proteína na prevenção do crescimento das células tumorais em vários pontos deste processo, percebendo o porque da perda da sua função ter um efeito profundo no desenvolvimento do tumor. A ativação da p53 induz respostas celulares, entre as quais estão a diferenciação, senescência, reparação do DNA, inibição da angiogénese, sendo as mais bem estudadas a habilidade para induzir a paragem do ciclo celular e da morte celular apoptótica. Estas respostas permitem à proteína inibir o crescimento das células que sofreram determinado tipo de

stress, seja por paragem do ciclo, que pode ser irreversível ou transiente para permitir a reparação e recuperação antes de várias rondas de replicação, seja pela remoção permanente dessas células do organismo por apoptose. Ambas vão prevenir a replicação de células que sofreram essas mudanças oncogénicas e, portanto, vão inibir o desenvolvimento tumoral.[5]

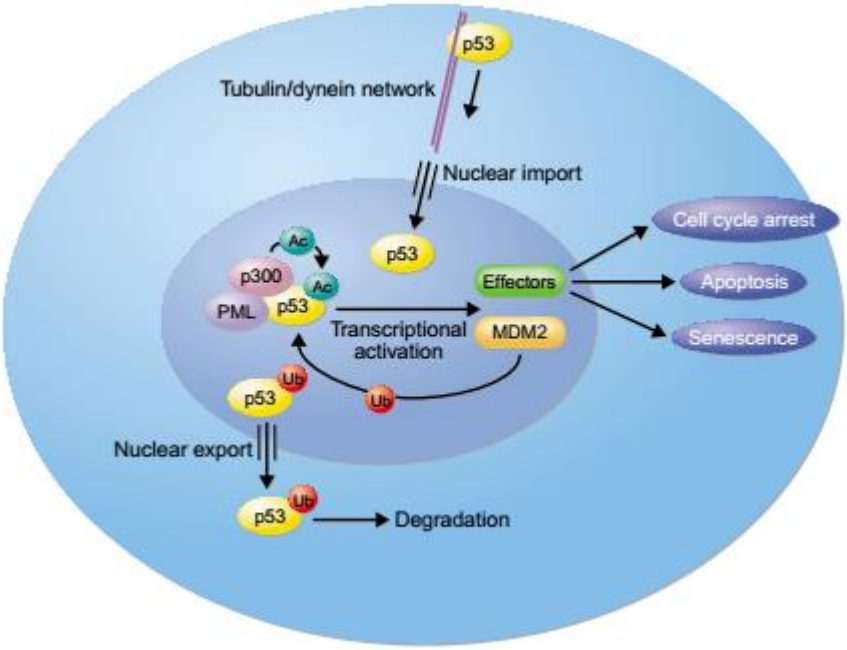

Figura 1.3 – Modelo de alguns mecanismos que podem regular a localização subcelular da p53, a estabilização e a atividade de transcrição.[5]

Em situações de resposta ao stress, o nível celular da p53 é elevado devido a um mecanismo de pós-tradução, que leva à paragem do ciclo celular ou à apoptose. Em condições normais, esta proteína encontra-se extremamente regulada pela proteína Murine Double Minute 2 (MDM_2), através de um ciclo autorregulatório. Este supressor de tumor pode ativar a expressão de MDM_2 que, por sua vez, leva à inibição da p53 por três possíveis mecanismos: primeiro, o MDM_2 liga-se à proteína no seu domínio de transativação e bloqueia a sua capacidade de ativar a transcrição; segundo, está envolvido no mecanismo de exportação nuclear, como mostra a figura 1.3; terceiro, o MDM_2 serve como uma ligase de ubiquitina que promove a degradação da p53. Sendo este inibidor o alvo transcricional da fosfoproteína, uma elevada atividade da mesma leva a uma elevada expressão do seu inibidor, devido ao *feedback* autorregulatório existente. Este *loop* é de extrema importância, pois a perda de expressão de MDM_2 pode levar a morte embrionária precoce causada por apoptose descontrolada via p53, como já foi demonstrado em ratos por ***Ryan, Phillips, & Vousden, 2001***. Na ausência da p53, ratos sem este inibidor têm uma evolução normal, o que ilustra o papel crítico que este tem na regulação negativa da proteína durante o desenvolvimento. Na maioria dos casos de resposta ao stress, a indução do supressor de tumor envolve a inibição do MDM_2. Alguns mecanismos podem regular, através do inibidor, a degradação da fosfoproteína, tais como repressão direta da expressão do mesmo, modificações pós-tradução de ambos, expressão de proteínas que inibem a função deste e regulação da localização subcelular das duas macromoléculas. A ubiquitinação da p53 pelo seu regulador ocorre

no terminal carboxilo e uma mutação nos três resíduos de lisina aí presentes inibe a exportação nuclear, sugerindo que este processo, que ocorre no núcleo, ativa a sequência de exportação, afetando o estado de oligomerização da proteína, o que resulta no seu transporte para o citoplasma. A sua capacidade de induzir a expressão do gene apoptótico, p53AIP1, e levar à morte celular, é dependente da fosforilação da Ser46. A paragem do ciclo celular não é dependente desta fosforilação, indicando que esta modificação no aminácido contribui para a resultado final da respota da p53.[3,5] É, portanto, de extrema importância o estudo desta proteína e, eventualmente, dos mecanismos de inibição da mesma, principalmente a interação com o seu inibidor, MDM_2. Existem estudos feitos com pequenas moléculas, mas nunca um estudo muito alargado e que usasse um grande número de moléculos (como é o caso deste). Grande parte dos mesmos atuavam sobre o MDM_x ou sobre a ligação do MDM_x à p53, tentado libertar a p53 do seu inibidor com o objetivo de lhe "devolver" a atividade apoptótica.[6]

1.3 Desenvolvimento e pesquisa de novos fármacos: metodologia CADD

O desenvolvimento e a pesquisa (R&D) de novos fármacos é um processo que consome imenso tempo, muito complexo, caro e que acarreta imensos riscos. Em 2016, um estudo reportou uma percentagem de sucesso clínico de cerca de 12%, desde a entrada do fármaco em testes clínicos até à sua aprovação. Na generalidade, o seu desenvolvimento integral até à chegada ao mercado pode durar entre 13 e 15 anos, com gastos entre os 2 e 3 biliões de dólares e embora os gastos tenham aumentados, o número de drogas aprovadas todos os anos por biliões de dólares investidos em R&D têm-se mantido ou até diminuido na passada década. Assim, as metodologias computacionais têm-se revelado importantes em várias etapas da descoberta destes compostos e continuam a ser indispensáveis na busca incessante por fármacos que possam salvar vidas. Métodos de ***Computer-Aided Drug Design*** (CADD) têm emergido nos últimos anos como uma ferramenta extremamente importante no desenvolvimento de moléculas com valor terapêutico, permitindo mais e melhores resultados que experiências de High-Throughput Screening (HTS) por si só. ***Mueller et al.*** construiram um modelo computacional em que usavam resultados de um HTS já feito (actividade metabotrópica do recetor 5 de glutamato), tendo sindo capazes de identificar moduladores numa experiência de *screening virtual* (HTS) com uma percentagem de 3,6% (16 vezes superior à do estudo original – 0,22%).[7] ***Wang, Weisi, et al*** em "***Identification of novel inhibitors of p53–MDM2 interaction facilitated by pharmacophore-based virtual screening combining molecular docking strategy.***" Testaram um protocolo de *screening* virtual baseado em farmacóforos para identificar inibidores da ligação entre a p53 e o MDM_2. Três derivados de aminotiofeno exibiram uma boa capacidade de inibição da ligação em estudo, tendo a maioria dos compostos mostrado uma atividade citotóxica potente contra as linhas celulares tumorais. Sugerem que o resultado indica que o screening virtual é uma estratégia racional e fiável para a descoberta deste tipo de inibidores. As metodologias CADD podem ser classificadas em duas categorias: baseadas na estrutura (SB) e baseadas no ligando (LB). A figura 1.4 ilustra o papel destes métodos num processo típico de descoberta de fármacos.[7]

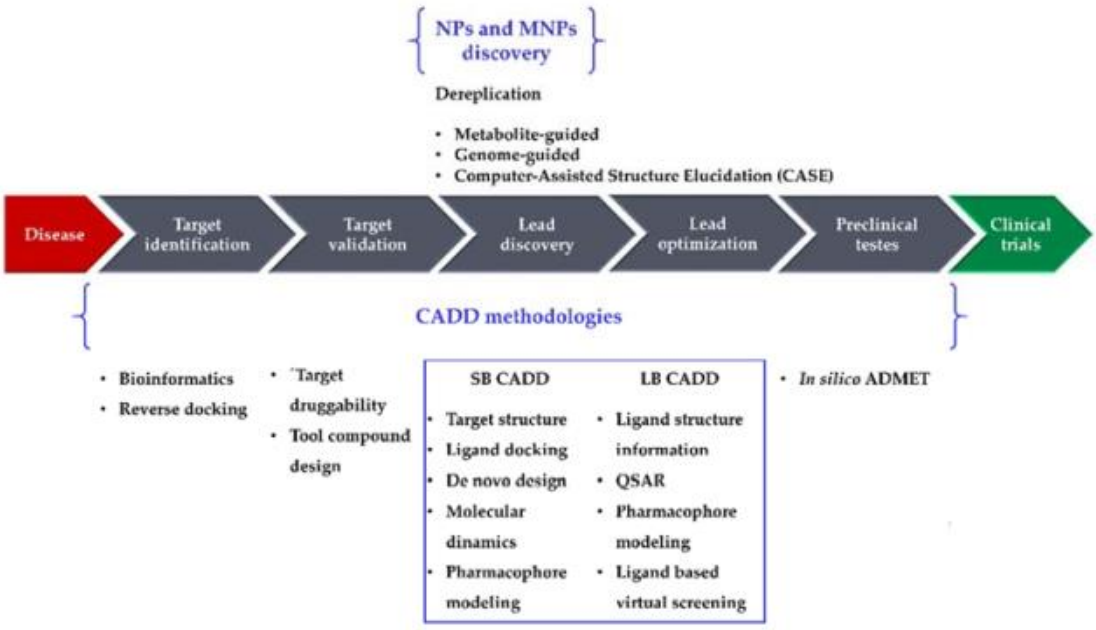

Figura 1.4 – Processo de descoberta de fármacos (CADD). SB – baseado na estrutura; LB, baseado no ligando; ADMET - absorção, distribuição, metabolismo, excreção e toxicidade; QSAR - *Quantitative Structure–Activity Relationship*.[7]

As decisões a serem tomadas requerem a previsão computacional da atividade biológica desses compostos, servindo também para ajudar no design de possíveis derivados que se revelem bioativos para determinado alvo e no *screening* virtual de bases de dados que contenham essas moléculas. Alguns avanços nas metodologias computacionais foram reportados e mostram a análise de redes onde há a conexão entre compostos ativos e os seus alvos, por forma a simular possíveis interações entre ligandos e o sítio de ligação para estabelecer relações de estrutura-atividade.

As metodologias LB são usadas para a descoberta de compostos quando grupos de moléculas são conhecidos para alvos específicos. Esta informação pode ser retirada das suas estruturas e da análise 2D e 3D das mesmas. As estratégias desenvolvidas podem incluir pesquisa em bases de dados de moléculas, alinhamento de estrutura para identificação de farmacóforos e *screening* virtual, algoritmos de *machine learning* (ML) para determinação de relações quantitativas estrutura-atividade (QSARs), previsão de propriedades dos possíveis compostos candidatos e desing de novas moléculas. No geral, o objetivo será que estes sejam representados por forma a que as suas propriedades fisico-químicas, relacionadas com as possíveis interações que estabelecem, se mantenham, enquanto a informação que não relevante é descartada. Esta metodologia é considerada uma aproximação indireta à descoberta de fármacos, pois não é necessário ter conhecimento da estrutura do alvo de interesse. As duas abordagens fundamentais são: **i.** seleção de compostos com base na semelhança química usando uma medida de similaridade; **ii.** construção de um modelo QSAR que prevê a atividade biológica a partir da estrutura química. A diferença entre ambas é que a **ii.** pesa as características da estrutura química de acordo com a influência na atividade biológica de interesse, sendo que a primeira abordagem não o faz. Os métodos LB, ao contrário dos métodos SB, podem ser aplicados quando não se sabe a estrutura do alvo biológico. Segundo ***Stumpfe et al., 2012***, compostos ativos identificados

por métodos de HTS virtual baseado no ligando são, geralmente, mais potentes que os que são identificados nos de HTS virtual baseado na estrutura. ***Cruz, Sara, et al*** em "***In silico HCT116 human colon cancer cell-based models en route to the discovery of lead-like anticancer drugs***" sugerem que a abordagem QSAR que se apoia no método baseado no ligando (seja através de estruturas moleculares ou espetros de Ressonância Magnética Nuclear), e quando corroborada por uma validação experimental, pode ser usada para prever novos inibidores contra o cancro do colon humano em linhas HTC116.

As metodologias SB recaem na capacidade de determinação e análise das estruturas 3D das moléculas biológicas. A principal hipótese desta abordagem é que a habilidade de uma molécula para interagir com uma determinada proteína (alvo), e produzir o efeito biológico desejado, depende da capacidade de se ligar ao sítio ativo da mesma. Moléculas que interajam da mesma forma terão efeitos biológicos similares. Assim, a descoberta de compostos vai depender da análise cuidada deste sítio ativo, sendo que é estritamente necessária informação acerca da estrutura do alvo. Os investigadores têm usado a estrutura do alvo, no que diz respeito à descoberta de novos fármacos, desde o início dos anos 80. Desde aí que esta técnica tem sido utilizada e, graças aos avanços em genómica e proteómica, foi permitido descobrir um grande número de candidatos. O uso de técnicas biofísicas como cristalografia de raio-X e espetroscopia de ressonância magnética nuclear (RMN) têm levado à elucidação de estruturas 3D de proteínas humanas e patogénicas, sucedendo que, no Protein Data Bank (PDB – local onde são depositadas todas as estruturas já descobertas e publicadas), existem mais de 81 000 estruturas de proteínas. A habilidade para rapidamente determinar o potencial de ligação ao alvo de interesse é um pré-requisito do processo de descoberta de novas drogas, sendo que, neste processo, os métodos computacionais permitem um *screening* de uma grande biblioteca de compostos e a determinação da capacidades de ligação através de técnicas de modelação/simulação e visualização. Docking molecular tem sido a metodologia mais utilizada para prever a afinidade de determinado fármaco ao alvo macromolecular, para interpretação das ligações efetuadas e para ajudar no desing de novas drogas.[7,8] ***Siddiquee, Khandaker, et al*** em "***Selective chemical probe inhibitor of Stat3, identified through structure-based virtual screening, induces antitumor activity.***" Demonstram a fiabilidade do uso de dados de cristalografia de raio-X e de modelação computacional através de *screening* virtual baseado na estrutura para identificação de inibidores de Stat3 a partir de bibliotecas químicas. Nesta dissertação foram usadas as duas abordagens aqui referidas: *screening* e *docking.*

1.4 Representação de estruturas moleculares

Neste tipo de metodologias foi necessário encontrar uma forma de representar as estruturas das moléculas por forma a aplicá-las em métodos computacionais. Esta representação é importante, por exemplo, para o cálculo de descritores moleculares (será explicada adiante a sua importância). Assim, houve o desenvolvimento de formatos que permitem mostrar as estruturas, como por exemplo, SMILES, SMARTS, InChI, InChIKey e ficheiros MDL (com ênfase em MOL e SDF).

SMILES (*Simplified Molecular Input Line System)* permite a representação de moléculas através de única linha de símbolos (que representa os átomos e as suas ligações), sendo um formato fácil de guardar e com o qual é fácil de trabalhar, utilizando-se, por exemplo, em bases de dados (como foi o caso desta dissertação).

SMARTS (*SMILES Arbitrary Target Specification*) é visto com uma extensão do SMILES, permitindo especificar subestruturas, como grupos funcionais, sendo importante para a obtenção de dados sobre a variabilidade das estruturas representadas.

InChI (*International Chemical Identifier*) e InChIKey têm o propósito de providenciar um código único para todas as estruturas químicas que possam ser indexadas, sem qualquer alteração, por motores de busca como a Google. Assim, os investigadores podem ter acesso a este tipo de compostos de uma maneira direta e rotineira. InChIKey é uma versão do código InChI que permite ter um comprimento fixo de 27 caracteres com uma resistência alta à colisão, isto é, a probabilidade de dois compostos terem o mesmo código é muito baixa. Esta particularidade torna este código muito usado quando se procuram duplicados numa base de dados, por exemplo.

MDL (*Molecular Design Limited*) serve para guardar estruturas através de tabelas de conectividade, que consistem numa linha e três blocos que representam átomos, ligações e propriedades. MOL (*molecule file*) e SDF (*structure data file*) são dois dos ficheiros que fazem parte desta designação, sendo que o primeiro permite guardar uma estrutura molecular e o segundo permite guardar várias.[8,9]

1.5 Descritores moleculares

Os descritores moleculares são, como o próprio nome indica, características que descrevem as moléculas. Estes estão associados a compostos químicos e são calculados através de processos matemáticos, que transformam as propriedades fisico-químicas em determinados valores associados a cada tipo de descritor, sendo que são posteriormente usados no desenvolvimento de modelos QSAR para previsão de atividade biológicas de novos compostos.[9] Nesta dissertação foi usado o software ***PaDEL-Descriptor***, que contém descritores 1D/2D, 3D e *Fingerprints*. Nesta dissertação foram calculados, através do mesmo, apenas descritores 1D/2D e *Fingerprints*, sendo que os descritores 3D foram calculados através de funções de distribuição radial (RDFs, em inglês), sendo que as estruturas 3D (utilizadas no *docking* molecular) foram otimizadas por métodos empíricos através do Corina e do Cxcalc. Um dos pontos importante da modelação QSAR é a seleção de descritores, que será explicada adiante.

1.5.1 Descritores 1D/2D

Este tipo de descritores são uni ou bidimensionais e representam a informação molecular sem ter em conta a sua conformação e a sua estrutura tridimensional, tendo em conta as suas distâncias topológicas. A contagem do número de átomos, do número de átomos de determinado elemento químico e do número de átomos da cadeia mais longa são exemplos deste tipo de descritores.

1.5.2 *Fingerprints*

A representação das propriedades e da estrutura de uma molécula por parte deste tipo de descritores é particularmente complexa, pois, geralmente, estes são codificados como sequências binárias cujas características produzem, de diferentes formas, um padrão de *bits* que é correspondente de determinada molécula. Neste caso, a presença de uma dada propriedade é representada por "1" e a sua ausência por "0". Os Fingerprints estão desenhados para se ter em conta diversos descritores moleculares, fragmentos estruturais, meios de conectividade pela molécula ou diferentes tipos de farmacóforos (parte de uma estrutura molecular que é responsável por uma determinada interação farmacológica ou biológica).[11] A tabela seguinte mostra alguns exemplos dos mesmos:

Tabela 1.1 – Exemplos do tipo de descritores do software PaDEL-Descriptor (http://www.yapcwsoft.com/dd/padeldescriptor/)

Descritor	Exemplos	Observação
1D/2D	*Atom Count*	Nº total de átomos e nº de átomos de um determinado tipo de elemento químico
	SV	Soma dos volumes atómicos de van der Waals
	MLFER_E	Refração molar excessiva
***Fingerprint* (circulares)**	*Fingerprinter* (CDK)	*Fingerprint* com 1024 características
	ExtendedFingerprinter (ExtCDK)	Extensão do *Fingerprinter* com bits adicionais que descrevem as características dos aneis aromáticos
***Fingerprint* (fragmentos)**	*SubstructureFingerprinter* (Sub)	Presença de padrões SMARTS para classificação de grupos funcionais
	SubstructureFingerprinterCount (SubC)	Contagem da presença de padrões SMARTS para classificação de grupos funcionais

1.5.3 Seleção de descritores

A seleção de descritores revela-se um ponto fulcral na otimização da modelação QSAR. A razão pela qual é efetuado este passo assenta no facto de nem todos os descritores moleculares contribuirem para o aumento da previsibilidade do modelo e, muito pelo contrário, o uso de um grande número de descritores pode introduzir ruído, levando a uma diminuição desta previsibilidade. Assim, esta seleção,

de uma forma geral, ajuda a aumentar a eficiência do modelo construído ao excluir os descritores menos relevantes, ajudando na interpretação do mesmo (mais simples e intuitivo) e diminuindo o ruído pelo facto de existirem menos variáveis a ser analisadas. Existem vários métodos de seleção de descritores: i) Métodos clássicos, como *Forward selection* e *Stepwise procedure*; ii) Métodos baseados em inteligência artificial, como o uso de redes neuronais artificiais ou do método do algoritmo genético; iii) Métodos diversos, como o método da substituição ou o método de aprendizagem automática *k-Nearest Neighbors*.[12]

1.6 Modelação QSAR

Os modelos QSAR são o resultado final de um processo que começa com a descrição de estruturas moleculares e culmina na inferência, hipótese e previsão de um efeito em determinados sistemas fisico-químicos, biológicos e/ou ambientais que estejam em análise. Esta metodologia baseia-se no facto de uma estrutura molecular poder conter características que são responsáveis pelas suas capacidades químicas, biológicas e físicas e na capacidade das mesmas poderem estarem presentes em 1 ou mais descritores moleculares. Através destes modelos, a atividade biológica de novos fármacos ou compostos não testados pode ser prevista a partir de moléculas com estrutura similar e cuja atividade já seja conhecida e testada.[13]

Este processo pode ser dividido em 5 etapas: **i)** seleção e construção da base de dados e posterior tratamento da informação; **ii)** cálculo de descritores; **iii)** seleção de descritores; **iv)** construção e otimização do modelo; **v)** validação do modelo.[8,12] A figura 1.5 representa o esquema que resume este tipo de metodologia.

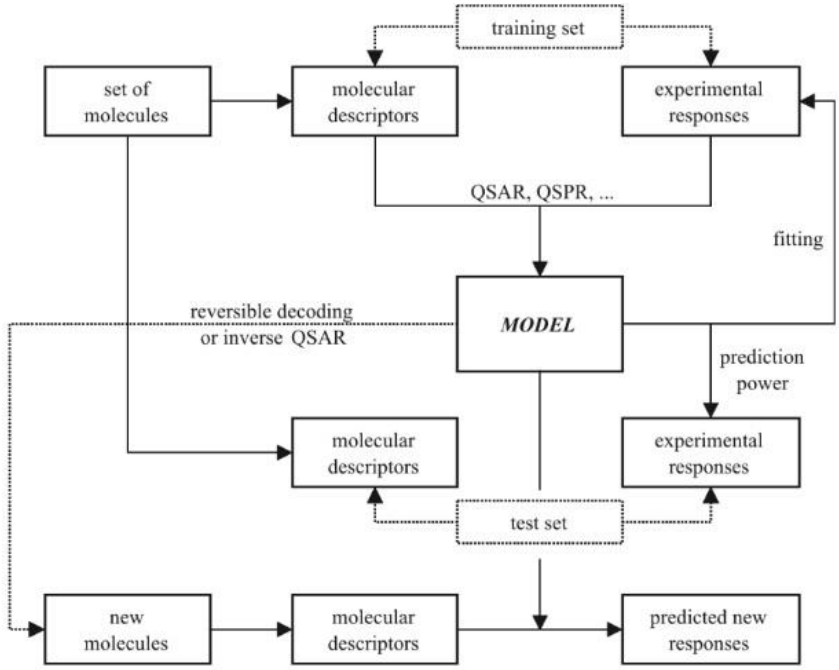

Figura 1.5 – Esquema geral da metodologia QSAR.[13]

Para se poder validar o nosso estudo, podem ser usados dois tipos de abordagens: através de modelos de regressão ou modelos de classificação. Nesta dissertação foi usado um modelo de classificação.

1.6.1 Modelos de Classificação

Neste tipo de abordagem, estes modelos são usados para prever se determinados compostos são ativos ou inativos para um dado estudo. Este tipo de método foi escolhido, pois, na base de dados, os valores de atividade não se apresentavam, na sua maioria, com IC50 e sim como Ativos ou Inativos. Caso se optasse por um modelo de regressão, teríamos de ter valores de IC50, maioritariamente, para o podermos utilizar. Optando, então, pelo modelo de classificação, é possível obter 4 tipos de resultados diferentes: **TP** – verdadeiros positivos, moléculas que são previstas, corretamente, como ativas -, **FP** – compostos que são previstos como ativos, mas que na realidade não o são -, funcionando de forma idêntica para os **TN** (verdadeiros negativos) e **FN** (falsos negativos). Claro que, para avaliar a capacidade de previsão de um determinado modelo, teremos de ter parâmetros estatísticos que ajudem nesta avaliação: especificidade (**SP** – capacidade de previsão para inativos), sensibilidade (**SE** – capacidade de previsão para ativos), previsibilidade (**Q**) e Coeficiente de Correlação de *Matthews* (**MCC**).[14] As equações seguintes mostram como calcular estes 4 parâmetros:

$$SP = \frac{TN}{TN+FP} \tag{1.1}$$

$$SE = \frac{TP}{TP+FN} \tag{1.2}$$

$$Q = \frac{TP+TN}{TP+TN+FP+FN} \tag{1.3}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \tag{1.4}$$

1.6.2 Modelos de Regressão

Este tipo de QSAR tem o mesmo objetivo que o método de classificação, com a particularidade de serem necessários valores (ao invés de classificações, como ativo ou inativo, por exemplo) para a construção deste tipo de modelos, como IC_{50} ou pIC_{50}. Estes podem ser avaliados tendo em conta vários parâmetros, como a raiz quadrada do erro quadrático médio (RMSE), quadrado do coeficiente de correlação de Pearson (R^2) e erro máximo absoluto (MAE). O primeiro indica o erro entre a raiz quadrada da média da atividade prevista e a raiz quadrada da média da atividade experimental para cada molécula. O R^2 mostra a correlação entre o valor da atividade prevista e o valor da atividade

experimental. O erro máximo absoluto mostra o desvio máximo do valor previsto em relação ao valor experimental. As equações 1.5, 1.6 e 1.7 mostram como se calculam estes parâmetros:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (1.5)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{N}} \quad (1.6)$$

$$MAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}|}{N} \quad (1.7)$$

onde yi é o valor da atividade experimental, ŷ é o valor de atividade previsto, ȳ é a média dos valores de atividade experimentais e N é o número de moléculas.[15]

1.7 Técnicas de aprendizagem automática

1.7.1 *Random Forest*

A árvore de decisão é uma metodologia supervisionada cujas características estão entre as mais desejadas: consegue tratar um grande conjunto de dados, ignora os descritores que são irrelevantes para o modelo, conseguindo ter múltiplos mecanismos de ação e sendo os resultados de fácil interpretação. Contrariamente, tem uma baixa capacidade de previsão, o que limita a sua aplicação no que se refere à previsão de atividade de novos compostos, por exemplo. Neste tipo de estudos, como o número de compostos submetidos ao modelo pode ser da ordem dos 10^5 e 10^6, uma pequena variação na capacidade de previsão pode fazer uma grande diferença nas classificações atribuídas.[16]

Como existe uma grande procura no que se refere a árvores de decisão, têm-se feito vários esforços para melhorar a sua previsibilidade, tendo estes resultado em vários algoritmos baseados nas mesmas. Foi, então, descoberto que uma das melhores abordagens seria usar combinações de árvores de decisão, como é o caso da *Random Forest* (RF). Ao contrário de outras combinações, A RF inclui características únicas e uma grande capacidade de previsão, fatores que a tornam aplicável para realizar modelos QSAR e QSPR.[16]

Esta é uma combinação de B árvores $\{T_1(X), \ldots, T_B(X)\}$, onde X={x1, …, xi} é o vetor de p dimensões correspondente às propriedades e aos descritores moleculares associados a uma molécula. O conjunto reproduz B outputs {K1=T1(X), …, KB=TB(X)} onde Ki, após agregação de todos os resultados das várias árvores, representa a previsão final para uma determinada molécula. Para modelos de classificação isto traduz-se na classe maioritariamente prevista pelas árvores de decisão. A figura 1.6 mostra um esquema deste método.[16]

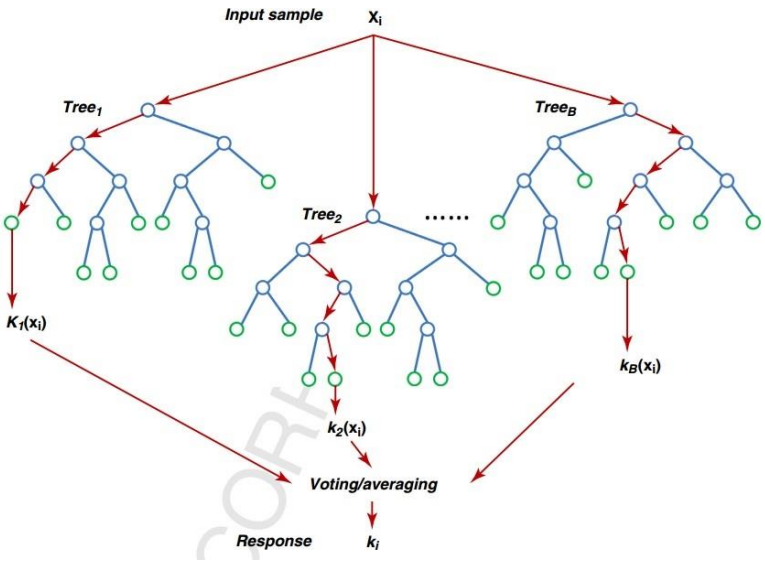

Figura 1.6 – Esquema de uma *Random Forest*.[17]

O processo de treino desta técnica de aprendizagem automática divide-se em três etapas:

i) Partição dos conjuntos iniciais

Os dados são colocados em grupos tendo em conta o número de árvores a usar, n_{tree}. Em cada grupo são usadas cerca de 2/3 das moléculas, sendo as restantes utilizadas para fazer uma validação *Out Of Bag* (OOB), que utiliza as que não são usadas no treino para determinar a capacidade de previsão do método.

ii) Treino do conjunto de árvores

Neste caso, em cada nódulo é escolhido o melhor resultado tendo em conta um determinado número de descritores, m_{try}, ao contrário do método de *bagging*, que escolhe o melhor entre todos os descritores.

iii) Previsão

As etapas 2 e 3 são repetidas até se achar o melhor resultado, sendo que o único parâmetro variável (na maioria das abordagens) será o m_{try}. Após essa variação, é registado o melhor modelo.[16,18]

Quando seria de esperar que a complexidade computacional deste modelo fosse maior comparada com o uso de uma única árvore de decisão (e não um conjunto delas), isso não se verifica, pois o algoritmo da *Random Forest* é extremamente eficiente, especialmente quando o número de descritores é grande. Esta eficiência assenta em duas diferenças nos algoritmos dos dois métodos: a primeira diferença reside no facto da RF usar apenas um número definido de descritores (m_{try}, que normalmente é dado como a raiz quadrada do número total de descritores usados), ao contrário de uma árvore normal que usará todos os descritores. Como este m_{try} será um número sempre mais reduzido, o processo será mais rápido. A segunda diferença diz respeito à previsibilidade da árvore de decisão: para que esta seja otimizada e se registe a melhor, serão necessários métodos que ajustem

e eliminem ramos de menor confiança, normalmente efetuados através de validação cruzada. A RF, como apenas utiliza um número específico de descritores (os melhores), não necessita de métodos de ajuste, sendo por isso um processo mais rápido em comparação ao treino de uma árvore de decisão.[16]

1.7.2 *Support Vector Machine*

Support Vector Machines (SVMs) são algoritmos de *machine learning* supervisionados que facilitam a classificação, o ranking e a previsão de um valor para determinada propriedade de um composto. A SVM pode ser usada para, por exemplo, distinguir moléculas que têm ou não uma atividade específica. Num primeiro passo, os compostos são projetados num espaço multidimensional, representados por vetores, onde, idealmente, se separam de forma linear, como mostra a figura 1.7.[17]

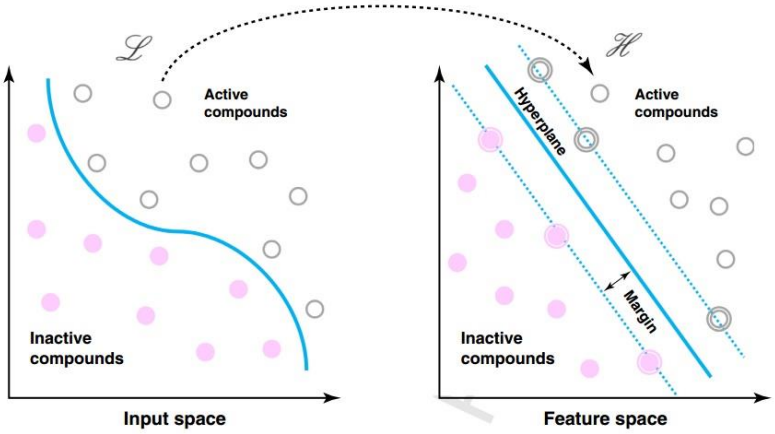

Figura 1.7 – Esquema de uma SVM para um modelo de classificação.[17]

A projeção é conseguida através do uso de uma função de *kernel*, uma das seguintes: linear, polinomial, sigmoidal e *radial basis* (RBF). AS primeiras três são funções globais, sendo que a RBF é local, tendo alguns resultados mostrado que a SVM baseada na função *radial basis* apresenta melhores resultados que as baseadas nas outras três, sendo por isso mais utilizada. A equação linear geral é dada por:

$$f_{(x,w)} = \sum_{j=1}^{m} w_j g_j(x) + b \qquad (1.8)$$

onde $g_j(x)$ representa as transformações não lineares com j=1, …, m e b o termo *bias*.[18] Após uma separação linear, as duas classes podem ser separadas por um hiperplano, que na realidade é escolhido, entre um infinito número deles, pela SVM que maximiza a margem entre as classes, assumindo que, quanto maior a margem, menor o erro de classificação quando se tratam dados desconhecidos (Fig. 1.7). Caso não seja possível haver uma separação linear, é aplicado um hiperplano

que maximiza a margem e mantém o número de compostos que não foram classificados o menor possível.[17]

A otimização dos parâmetros da SVM é essencial para a otimização do processo, principalmente no que se refere ao parâmetro *cost*, sendo este responsável pelo ajuste da margem, tentando maximiza-la, e do erro, tentando minimizá-lo.[19]

1.7.3 Redes neuronais artificiais

As redes neuronais artificiais (Artificial Neural Networks, ANNs) representam um sistema baseado na rede neuronal biológica, tentando mesmo recriar, por exemplo, o funcionamento do cérebro. Olhando um pouco para a biologia, o cérebro tem cerca de 86 mil milhões de neurónios[19] que comunicam através de sinais eletroquímicos. De uma forma geral, cada neurónio recebe sinais de outros neurónios (através das sinapses) até chegar a uma determinada célula do corpo. Uma resposta será enviada se a soma desses sinais ultrapassar um dado limite. O objetivo da ANN é criar um espelho daquilo que é a nossa rede neuronal biológica, não podendo haver uma comparação muito exigente, pois a complexidade e número de neurónios que estão presentes na rede biológica ultrapassam a rede neuronal artificial.[21]

Esta rede artificial é compostas por neurónio artificiais, também conhecidos como nódulos (que estão dispostos em camadas). Estes estão conectados uns aos outros e a sua conexão é dada por **peso** (W), consoante se trata de inibição ou excitação. Se o valor for alto, indica-nos uma conexão forte. Na passagem de nódulo para nódulo (a partir de camadas diferentes), é tido em conta o peso, também se estamos a falar de inibição ou excitação, e ainda se tem em conta as funções de transferência (que dependem das estruturas dos nódulos). Existem três tipos de neurónios (nódulos), que formam, por si, três tipos de camadas, numa ANN (como mostra a figura 1.8): camada de *input*, camada escondida e camada de *output*.[21,22]

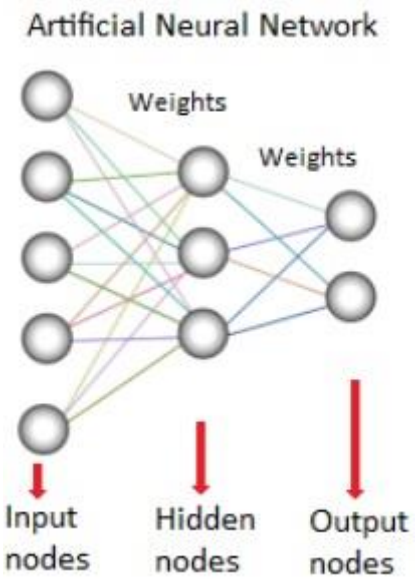

Figura 1.8 – Esquema de uma rede neuronal artificial. [21]

O nódulo de *input* recebe a informação de uma forma que esta possa ser numericamente expressa, sendo apresentada como valores de ativação (será de esperar que, quanto mais este valor, maior a ativação). A informação atravessa a rede e, baseando-se nos pesos (conexão entre neurónios), na inibição ou excitação e nas funções de transferência, este valor passa de nódulo para nódulo, sendo que cada um soma todos os valores de ativação que recebe e altera os mesmos com base nas funções de transferência respetivas. Esta ativação atravessa a rede, da camada de *input* para a camada escondida, até chegar à camada de *output*. Neste ponto, os nódulos aí presentes, transformam a informação dada como *input* em informação comum percetível para todas as pessoas, ao invés de nos mostrar apenas valores numéricos (por exemplo, a ativação ou inativação de uma determinada função celular por parte de um fármaco).

Claro que, como qualquer método experimental de previsão, existe um erro associado que, neste caso, é dado pela diferença entre o valor previsto e o valor atual. Esta diferença vai propagar-se de volta, distribuindo a mesma pelos pesos de cada nódulo de acordo com o erro associado a cada um (método do gradiente descente). Este método é uma otimização que minimiza funções: para uma função J definida por um conjunto de parâmetros (θ_0, θ_1, …), o gradiente descente encontra um mínimo local ou global, atribuindo um grupo de valores aos parâmetros e alterando os mesmos de forma iterativa proporcionalmente ao negativo do gradiente da função, assumindo que, se minimizar-mos a função J, o erro geral será o menor.[21,23] Nesta dissertação foram usados dois métodos de aprendizagem automática que replicam redes neuronais artificais supervisionadas: *Neural Network* (NN) e *MultilayerPerceptron* (MLP). Ao contrário de, por exemplo, redes neuronais de *Kohonen* (explicado adiante), que são redes não supervisionadas. De forma resumida, a aprendizagem automática supervisionado permite a realização de uma tarefa quando é providenciado um treino, padrões de *input* e *output* aos sistemas em estudo. A abordagem não supervisionada é uma técnica na qual o sistema tem de descobrir as características de um dado *input* populacional por si só, não sendo usado nenhum *set* de categorias *à priori* (como é o caso da rede neuronal de *Kohonen* usada nesta dissertação, que é usada para agrupar compostos em classes, como será visto posteriormente, sem nenhuma informação antecipada).

1.7.4 *k-Nearest Neighbors*

O algoritmo do k-Nearest Neighbors é um método que aglomera todos os casos possíveis e classifica novos objetos baseando-se numa medida de similaridade (funções de distância, por exemplo). O objeto é classificado pelos votos da maioria dos seus vizinhos, sendo atribuído a classe mais comum, dentro dos vizinhos, calculada pela função de distância. Por exemplo, se K=1, então ao objeto será atribuída a classe do seu vizinho. Neste método, uma das distâncias mais usadas é a Euclidiana: a distância entre um vetor não classificado xi e cada um dos vetores yi no conjunto de treino é calculado usando a seguinte fórmula:

$$d_e = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{1.9}$$

O total de k vetores próximos do vetor x_i é usado para determinar a classe do vetor não classificado, assim, a classe maioritária entre os vizinhos é decidida como a classe prevista do vetor não classificado x_i.[14,24]

1.7.5 Redes neuronais de *Kohonen*

As redes neuronais de *Kohonen* (ou mapas organizativos – SOM, *Self-Organized Maps,* método não supervisionado) são consideradas redes neuronais especiais e são uma excelente ferramenta para organizar e armazenar dados onde são definidas relações de similaridade ou distância – estes foram os primeiros estudos onde foram usados estes mapas. Ao invés de trabalhar de uma forma listada, este mapa guarda a informação de forma associativa e analisa os grupos de moléculas de forma não linear, permitindo reconhecer e classificar amostras desconhecidas. Este funciona com uma aprendizagem competitiva, o seu treino não é supervisionado e os dados são organizados em mapas 2D.[25]

Nesta dissertação foi usada uma matriz bidimensional de neurónios (os mapas podem ter formatos diferentes), onde cada um destes neurónios é um vetor $N_i = (n_{i_1,}\ n_{i_2}, \ldots, n_{i_n}$, bem como os compostos que foram utilizados para construir o mapa $X_i = (x_{i_1,}\ x_{i_2}, \ldots, x_{i_n}$, sendo que n representa o número de descritores que foram calculados e utilizados neste método.

A distância euclidiana (equação 1.9) é usada para determinar o neurónio mais próximo após comparação de cada neurónio X_i com todos os neurónios do mapa. Este passo é feito de treino para treino por forma a melhorar o peso de cada neurónio ao longo das várias iterações (os pesos são atribuídos de forma variável no início), sendo que o neurónio vencedor (com menor distância euclidiana) irá ver os seus pesos alterados por forma a assemelhar-se ao objeto em estudo.

A correção dos pesos é dada pela expressão:

$$w_{i_{novo}} = w_{i_{velho}} + (obj_i - w_{i_{velho}})(0.4995 \times \frac{(epoch_{max} - epoch)}{(epoch_{max} - 1)} + 0.0005)(\frac{1-d}{dL+1}) \qquad (1.10)$$

onde w_i é o peso do neurónio, sendo novo e velho referentes ao valor antes e depois da correção do mesmo, obj_i é a i variável do objeto, d é a distância entre o neurónio vencedor e o neurónio que está a ser corrigido e dL é o raio do neurónio vizinho.[25,26]

1.8 *Screening* Virtual

Hoje em dia é possível testar e selecionar milhares de ligandos de acordo com a sua capacidade de interação com um determinado alvo biológico associado a uma doença específica, existindo, para isso, métodos de *screening* automatizados. No que à descoberta de novos fármacos diz respeito, pode assumir-se que os métodos computacionais não substituem os métodos experimentais, no entanto, a junção de ambos pode ajudar os investigadores no *screening* e na síntese de compostos de uma maneira mais racional. A utilização das abordagens computacionais tem sofrido um desenvolvimento

notável no processo de descoberta de fármacos ao longo do tempo. O *screening* virtual (SV) será, através destas ferramentas automatizadas, a procura de um composto-líder que possa ter a melhor afinidade para um determinado alvo biológico em estudo. O *screening* biológico de milhares de compostos, através de métodos experimentais, é extremamente dispendioso, sendo que as abordagens CADD são consideradas alternativas de alto interesse para os substituir. Assim, o SV atingiu um estado em que se tornou altamente dinâmico e lucrativo na descoberta de novos fármacos ou os chamados *hits* (nome utilizado na indústria farmacêutica).

Consoante o objetivo final de um determinado estudo, podem ser usados dois tipos de *screening* virtual: baseado na estrutura (*Structured-based*, SB) ou baseado no ligando (*Ligand-based*, LB). Alguns exemplos deste tipo de *screening* encontram-se descritos no ponto 1.3.[27]

1.8.1 Baseado no ligando

Neste processo, o composto-líder é encontrado usando uma procura com base na similaridade topológica ou estrutural, tendo em conta diversos critérios, tais como a estrutura, os fragmentos individuais ou as propriedades electroestáticas do mesmo. As moléculas são classificadas por um *score* de similaridade, podendo este ser obtido a partir de métodos ou algoritmos diferentes. Um exemplo será o alinhamento de pequenas moléculas: neste caso, são sobrepostas cada uma das moléculas da base de dados com a molécula de referência, sendo o score atribuído tendo em conta a extensão da sua similaridade.[27]

1.8.2 Baseado na estrutura

Nesta abordagem, o primeiro passo é identificar o sítio de ligação da molécula alvo, geralmente um bolso ou protuberância com características hidrofóbicas, com uma variedade de hidrogénios dadores e aceitadores de ligação de hidrogénio e com superfícies de aderência molecular. Este sítio de ligação pode ser idêntico ao de uma enzima ou ser um local de comunicação que é necessário para o mecanismo da molécula. Mas para tudo isto ser possível, tem de ser possível determinar a estrutura do alvo em questão, sendo que esta determinação é feita através de Ressonância Magnética Nuclear ou Cristalografia de Raio-X. Um dos exemplos deste tipo de processo, usado nesta dissertação, é o *docking* molecular (que será explicado no ponto seguinte).[27]

1.9 *Docking* molecular

Um dos métodos que tem recebido mais atenção, no que se refere ao *screening* virtual, é o *Docking*. Este processo caracteriza-se pela capacidade de modulação da estrutura 3D de um complexo entre o alvo e o ligando e pela avaliação da estabilidade do mesmo num reconhecimento biológico específico. A premissa desta abordagem pode ser dividida em dois passos: **i)** Exploração do espaço conformacional dos ligandos que se ligam às moléculas alvo; **ii)** Pontuar em concordância com a afinidade de ligação estimada. Gera-se uma conformação do ligando que, com a ajuda das funções de pontuações, é comparada com conformações anteriores. Esta é aceite ou rejeitada com base na

pontuação para a respetiva conformação. Após isso, é gerada uma nova e o processo volta a começar, sendo que a procura e a pontuação podem estar acopladas no *docking*. Como seria de esperar, é importante obter melhores funções de pontuação para que a conformação com o melhor *score* possa ter a afinidade de ligação experimental mais alta com o alvo em estudo.

O programa usado nesta dissertação foi o AutoDock Vina: um método automático e de procura aleatória que altera um determinado ligando e as rotações das suas ligações para prever uma interação entre uma pequena molécula e a estrutura 3D do alvo. A robustez deste programa pode estar relacionada com o *annealing* simulado Monte Carlo e métodos de algoritmos genéticos e evolucionários, tendo a previsão de ligações entre compostos-líder e alvos mostrado um grande sucesso quando o AutoDock é utilizado.[27] Um dos exemplos apresentado no ponto 1.3 mostra que ***Wang, Weisi, et al*** em *"**Identification of novel inhibitors of p53–MDM2 interaction facilitated by pharmacophore-based virtual screening combining molecular docking strategy.***" testaram um protocolo de *screening* virtual baseado em farmacóforos para identificar inibidores da ligação entre a p53 e o MDM_2. Três derivados de aminotiofeno exibiram uma boa capacidade de inibição da ligação em estudo, tendo a maioria dos compostos mostrado uma atividade citotóxica potente contra as linhas celulares tumorais. Sugerem que o resultado indica que o screening virtual é uma estratégia racional e fiável para a descoberta deste tipo de inibidores.

Capítulo 2 - METODOLOGIA

2.1 Construção da base de dados e partição dos conjuntos de treino e de teste

Na construção da base de dados foram retirados os compostos depositados nas bases de dados ChEMBL[28], Reaxys[29] e ZINC[30], perante uma condição específica. Neste caso, a condição é considerar compostos com atividade contra a proteína p53 (tanto ativos como inativos). Como nestas bases de dados só existiam compostos até novembro de 2016, foi necessário fazer uma pesquisa bibliográfica que permitisse encontrar compostos promissores que não estariam aí presentes. Neste caso, foi vistos artigos até Setembro de 2019. As estruturas moleculares dos 105 compostos aí recolhidos foram desenhadas através do programa MarvinSketch[31] (Marvin, versão 17.28.0). Foram extraídos 13051 compostos na CHEMBL, 6 na ZINC e 126 na Reaxys. Após esta extração, foi necessário fazer uma revisão destes compostos, especialmente perceber se haveria duplicados, moléculas com falta de informação, com informações erradas ou com peso molecular superior a 1000 Da, e os que se encaixavam nestes requisitos foram removidos. A base de dados continha 10505 moléculas após estas correções. Posteriormente, foi necessário dividir esta base de dados num conjunto de treino e de teste tendo em conta as informações de atividade que esta continha: no conjunto de treino foram colocadas as moléculas cuja classificação era ativa ou inativa, sendo que as restantes foram, automaticamente, para um segundo conjunto de teste. Para o conjunto de treino, a partição tinha como objetivo tentar obter 50% de compostos ativos e 50% de inativos mas, como a base de dados continha mais compostos inativos, optou-se por dividir em cerca de 40% de ativos e 60% de inativos, sendo que os compostos que ficaram de fora da partição constituíram um primeiro conjunto de teste. Após esta divisão, o conjunto de treino continha 6028 compostos, o primeiro conjunto de teste 4211 moléculas e o segundo 266. O ficheiro SDF das moléculas (dos conjuntos) foram uniformizadas no Standardizer[32] (JChem, versão 18.23.0) seguindo os seguintes parâmetros: remover sais e solventes, adicionar hidrogénios, transformar os grupos seguintes nos mesómeros mais estáveis: nitro, sulfóxido e *N*-óxido terciário e aromatizar.

2.2 Cálculo de descritores moleculares

Os descritores moleculares foram calculados através do software PaDEL-Descriptor[33] (versão 2.21), tendo sido calculados todos os descritores 1D/2D, e alguns *fingerprints*: CDK, *ExtendedCDK, MACCS*, *Pubchem*, *Substructure e SubstructureCount*. Os descritores 3D foram calculados através de *Radial distribution functions* (RDFs)[34] e tendo em conta as cargas *jas*, através da equação (2.1) com 128 e 256 valores igualmente distribuídos de r entre 0 e 12,8/25,6 Å:

$$RDF(r) = \sum_{i=1}^{N-1} \quad \sum_{j=i+1}^{N} p_i p_j e^{-B(r-r_{ij})^2} \tag{2.1}$$

onde N é o número de átomos na molécula, pi é a propriedade atómica do átomo i (neste caso foi utilizada a carga atómica parcial), B é o parâmetro de indecisão e rij a distância 3D entre os átomos i e j. Um set de 128 e um set de 256 descritores RDF foram calculados separadamente, derivado de pares de átomos com i) uma carga negativa e positiva, ii) duas cargas positivas e iii) duas cargas negativas. As cargas atómicas parciais – cargas atómicas parciais *natural bond orbital* (NBO) – foram estimadas

usando um método de aprendizagem automática desenvolvida por Aires-de-Sousa e colegas (http://joao.airesdesousa.com/charges)[35]. A otimização das estruturas 3D (utilizadas no *docking*) feita por métodos empíricos pelos programas Corina e Cxcalc (ChemAxon).

2.3 Escolha dos sets de descritores, seleção dos descritores moleculares e do método de aprendizagem automática

Primeiramente, os métodos de aprendizagem automática usados foram: *Random Forest*, *Support Vector Machine* (SVM), *Multilayer Perceptron Neural Network* e *Neural Network*. Estes foram utilizados através do programa WEKA[36] (versão 3.8.3). Observando os resultados preliminares dos testes de aprendizagem automática, foi possível perceber quais os métodos com melhores resultados, bem como os melhores sets de descritores moleculares (1D/2D e *Fingerprints*). De todos os referidos no ponto anterior, foram escolhidos aqueles que apresentavam os melhores resultados para o conjunto treino em validação cruzada: 1D/2D, CDK *Fingerprint* e *PubchemFingerprinter*.

A seleção dos melhores descritores de cada set foi feita através do método de aprendizagem automática *k-Nearest Neighbor*, com os parâmetros K=10 e distância=1/d, sem d a distância Euclidiana. Foi usado o classificador *weka.classifier.meta.AttributeSelectedClassifier*, e vários métodos de procura: *BestFirst*; *GreedyStepwise*, com pesquisa *backwards*; *PSOsearch*, variando os pesos: individuais, inertia e social (por exemplo, 0,33 para os pesos inertia e social e 0,34 para o peso individual). Foi usado também outro método de seleção de descritores através do método de aprendizagem automática *Random Forest*, com a seleção dos 50, 100, 150 e 200 melhores descritores. A RF tem como um dos seus parâmetros a função “computeAttributeImportance”, que permite saber quais são os descritores que têm uma maior importância na construção do modelo. Foi a partir desta função que foram selecionados os 50, 100, 150 e 200 melhores.

A escolha do melhor método de aprendizagem automática teve por base os resultados da construção do modelo, tendo sido avaliados, no final, três métodos: *Random Forest*, *Neural Network* e SVM. O melhor resultado veio da *Random Forest*, com validação cruzada 10 vezes, com os parâmetros *nFeatures* e *nIterations* igual a 9 e 500, respetivamente, e tendo sido usada uma função que equilibra as classes, já que as mesmas não se encontravam igualmente balanceadas: função *meta.CostSensitiveClassifier*, sendo a primeira coluna da *cost matrix* (2x2) referente a um valor de 45% (classe dos inativos) e a segunda coluna a um valor de 55% (classe dos ativos). Esta função permitiu dar mais peso a uma classe (ativos) em relação à outra, por forma a equilibrar as duas, visto que o conjunto de treino continha um número ligeiramente maior de compostos inativos em relação ao número de compostos ativos.

2.4 Otimização do modelo

A otimização do modelo começa com variação dos parâmetros (nFeatures e nIterations, como referido no ponto anterior) e da função meta.CostSensitiveClassifier. Esta variação é feita até se atingir

o modelo com melhores resultados: *Random Forest*, com validação cruzada 10 vezes, com os parâmetros *nFeatures* e *nIterations* igual a 9 e 500, respetivamente, e tendo sido usada a função *meta.CostSensitiveClassifier*, sendo a primeira coluna da *cost matrix* (2x2) referente a um valor de 45% (classe dos inativos) e a segunda coluna a um valor de 55% (classe dos ativos). Posteriormente, é necessário fazer uma validação interna (com um primeiro conjunto de teste cujas atividades são conhecidas) e uma validação externa (com um segundo conjunto de teste cujas atividades são desconhecidas), por forma a avaliar a previsibilidade e eficácia do modelo. Foram também adicionados descritores 3D, calculados a partir de funções de distribuição radial (RDFs, em inglês) e foram comparados os resultados.

2.5 *Screening* virtual

O *screening* virtual consistiu em extrair 1442 (drogas/fármacos aprovados pela *Food and Drug Administration*, FDA) da base de dados do ZINC, os quais foram estandardizados da mesma forma que toda a base de dados referida no primeiro ponto da metodologia (através do programa Standardizer), e submetê-los ao modelo já criado e otimizado, por forma a obter uma previsão de atividade contra p53. Neste ponto também foi necessário agrupar os compostos em classes, definir o domínio de aplicabilidade e calcular a *average SOM distance* (ASD) para perceber que moléculas poderiam pertencer ao domínio de aplicabilidade, através das redes neuronais de *Kohonen*[25,26]. Foi preciso avaliar os compostos por classes e perceber quais continham mais erros, avaliando falsos positivos e falsos negativos por classes. As moléculas pertenceriam ao domínio de aplicabilidade se tivessem uma ASD maior que 0,34. Após toda esta análise, foram escolhidos os 100 compostos mais promissores que passariam à fase de *docking* molecular.

2.6 *Docking* molecular

O *docking* foi efetuado através do programa AutoDockVina[37] (versão 1.1.2), com o auxílio do programa Cygwin[38] (versão 3.8). O objetivo seria avaliar a interações dos compostos promissores com a estrutura da nossa molécula alvo, sempre com o objetivo de ver uma maior interação. Foram utilizadas 3 estruturas do *Protein Data Bank*[39] (PDB) neste estudo: 4HG7[40] (estrutura do MDM_2), 3DAB[41] (estrutura do MDM_2) e novamente 3DAB[41] (neste caso com a estrutura da p53), e foram feitas 4 experiências diferentes: uma com a estrutura 4HG7, uma com a estrutura 3DAB e duas com a estrutura 3DAB mas referente à proteína p53, onde se variou apenas a região onde seria estudada a interação (informação sobre a região de ligação em Anexos). Estas 4 experiências foram sempre feitas comparando a otimização das estruturas 3D dos ligandos através de dois programas Corina (Molecular Networks GmbH Computerchemie) e do Cxcalc (ChemAxon). O controlo positivo utilizado neste estudo foi a estrutura da *Nutlin-3*, cuja atividade contra p53 é conhecida. Foi selecionada uma lista de 10 moléculas que apresentavam os melhores resultados (previsão da menor energia de ligação) dentro de cada experiência, com o objetivo de se verificar se 1 ou 2 compostos se destacavam. Após a esta seleção, os dois compostos (nilotinib e dihidroergotamina) que se destacavam foram analisados com mais pormenor. Foi analisado ainda um derivado da dihidroergotamina (DHET), por ser

economicamente mais vantajoso: dihidroergocristina (DHEC). No final foram analisados a DHET e a DHEC com mais detalhe, para se perceber qual o composto que seria usado na validação experimental.

2.7 Validação experimental

2.7.1 Cultura Celular

Nesta dissertação foi usada uma linha celular de cancro do colon, mais precisamente a linha HT29 *wild type* (referência ATCC® HTB-38™). Esta linha é aderente e foi cultivada em frascos T25 e/ou T75 (Sarstedt) numa incubadora (Panasonic) a 37ºC, com atmosfera húmida e 5% de CO_2, tendo sido utilizado o meio RPMI completo ($RPMI_C$, Gibco) – composição em Anexos – e um profilático (Plasmocin) para prevenir a contaminação com micoplasma (mais informação em Anexos). As células foram soltas do frasco com recurso a Tripsina-EDTA (TE, Gibco) quando se observava uma confluência entre 80 a 90% e centrifugadas (Eppendorf) a 200 g durante 5 minutos. As células foram ressuspensas e colocadas em cultura tendo em conta a diluição desejada para usos subsequentes. Caso não fossem necessárias, as células eram congeladas a -80ºC ressuspendendo o pellet (após centrifugação) com o meio e 10% (v/v) de Dimetilsulfóxido (DMSO).

2.7.2 Teste de viabilidade pelo método da resazurina

Para estudar o efeito celular do composto em estudo – Dihidroergocristina (DHEC) -, foi usado um ensaio de viabilidade: Ensaio da Resazurina (procedimento em Anexos). Foram usadas placas de 24 poços, numa primeira abordagem, e, numa segunda fase, placas de 96 poços. As células foram plaqueadas em meio $RPMI_C$, tendo em conta um número de células já otimizado para cada poço: 250 mil células/poço nas placas de 24 poços (volume final de 2 mL) e 25 mil células/poço nas placas de 96 poços (volume final de 200 µL). Após 24h, o meio foi retirado e adicionado meio e o composto (DHEC), diluído em DMSO, em diversas concentrações: nesta dissertação são apresentadas imagens das células com concentrações de 100 µM (máximo de concentração usado), 50 µM e 0 µM, tendo as mesmas decrescido 10 µM até à concentração nula (mais informação em Anexos). Após 24h, o meio com composto foi retirado e os poços lavados com PBS. Após a lavagem, foi adicionado meio com resazurina, tendo a percentagem da mesma variado nas duas abordagem já mencionadas acima: na primeira fase foi usada uma percentagem de 10%, sendo que na segunda foi usada uma percentagem de 50%. Após a adição de resazurina, seguia-se uma incubação de 2h. As absorvâncias foram lidas através de um leitor de microplacas (SpectraMax 190 Microplate Reader) e os dados foram adquiridos usando o programa SoftMax Pro (versão 6.4). Os resultados foram analisados usando o Microsoft Office Excel e o GraphPad Prism (versão 6).

Capítulo 3 – RESULTADOS E DISCUSSÃO

3.1. Definição dos conjuntos de treino, teste e teste2

Tabela 3.1 – Informações extraídas da base de dados

Conjunto	Ativos	% Ativos	Inativos	% Inativos	Média Massa Molecular	Total
Treino	2025	33,6	4003	66.4	476,1	6028
Teste	797	18,9	3414	81.1	394,1	4211
Teste2	16	6	250	94	492,2	266
Global	2838	27	7667	73	454,1	10505

Analisando a base de dados, aquilo que salta à vista é a maior quantidade de compostos inativos. Este ponto revelou-se um desafio ao longo deste estudo e da construção do melhor modelo para o definir, pois a tentativa será sempre balancear as classes, como será visto adiante. No conjunto de treino, crucial para a construção do modelo, tentou balancear-se as classes de atividade, sendo por isso um conjunto mais equilibrado (ainda que não totalmente) que o conjunto de teste, por exemplo. Os dados para o Teste2 já são resultados provenientes do modelo final otimizado. Será de esperar que o nosso modelo tenha uma melhor capacidade para prever compostos inativos ao invés de ativos, pois estes correspondem a maioria do conjunto de treino. Analisando a globalidade dos dados, temos 2838 compostos ativos e 7667 inativos, correspondendo a 27% e 73%, respetivamente, do total de 10505 compostos da base de dados. A massa molecular média das moléculas é de 454,1 Da, sendo este um ponto positivo, pois este tipo de estudo não funciona de forma totalmente correta com moléculas com mais de 1000 Da porque existem erros no cálculo de descritores e nos métodos de aprendizagem automática (sendo por isso que as mesmas foram alvo de seleção, como explicado na metodologia).

3.2. Escolha do tipo de descritor, seleção dos descritores e escolha de método de aprendizagem automática

Num primeiro momento foi necessário avaliar os descritores moleculares e perceber quais obtinham melhores resultados. Existem descritores que representam, de forma geral, o mesmo tipo. Assim, podemos dividir os *sets* em 3 grupos: **i)** 1D/2D; **ii)** *Fingerprinter* (CDK) e *ExtendedFingerprinter* (ExtCDK) – dizem respeito a *fingerprints* circulares; **iii)** *Pubchem* (Pubchem), *Substructure* (Sub), *SubstructureCount* (SubC) e *MACCS* – dizem respeito a *fingerprints* que representam fragmentos das moléculas. Estes dados foram analisados utilizando o método de aprendizagem automática Random Forest (RF). A tabela 3.2 mostra os melhores resultados obtidos para este método tendo em conta os vários *sets* de descritores.

Tabela 3.2 – Melhores resultados utilizando a *RandomForest*.

Descritor ID	Nº de descritores	TP	TN	FP	FN	SE	SP	Q	MCC
1D 2D nF = 37 nIt = 200 Csc 40/60	1443	1346	3413	590	679	0,665	0,853	0,789	0,523
CDK nF = 34 nIt = 500 Csc 45/55	1024	1296	3566	437	729	0,640	0,891	0,807	0,554
ExtCDK nF = 33 nIt = 500 Csc 45/55	1024	1259	3596	407	766	0,621	0,898	0,805	0,549
Pubchem nF = 31 nIt = 200	881	1219	3637	366	806	0,602	0,909	0,806	0,548
Sub nF = 18 nIt = 200	307	1062	3516	487	963	0,524	0,878	0,759	0,435
SubC nF = 19 nIt = 500	307	1101	3699	304	924	0,544	0,924	0,796	0,523
MACCS nF = 13 nIt = 100 Csc 45/55	166	1303	3494	509	722	0,643	0,873	0,796	0,532

Na tabela, nF e nIt representam, respetivamente, *nFeatures* e *nIterations*, os parâmetros variáveis para otimização do modelo. Os únicos parâmetros comparáveis são a sensibilidade (SE), especificidade (SP), previsibilidade (Q) e o coeficiente de correlação de *Matthews* (MCC), sendo que é através destes que se podem comparar as diferentes abordagens utilizadas. O peso (também variável) representa a função *CostSensitiveClassifier* (Csc), já explicada na metodologia, e que permite o balanceamento da classe (ativo ou inativo). Analisando numericamente os resultados TP e TN (verdadeiros positivos e verdadeiros negativos) da tabela, observamos que temos mais compostos TN que TP, o que evidencia uma falta de balanço, que é também comprovado pela SE (capacidade que o modelo tem de prever compostos como ativos) e pela SP (capacidade que o modelo tem de prever compostos como inativos): podemos reparar que a SP está mais próxima de 1 e é maior que a SE, o que nos indica que no nosso modelo tem uma maior e melhor capacidade para prever compostos como inativos do que para prever compostos como ativos. E mesmo no caso específico em que se usa uma função que ajuda neste balanceamento, como é o caso do Csc nos descritores 1D/2D, CDK, ExtCDK e MACCS, continuamos a ter uma maior presença de compostos inativos (TN), bem como uma maior capacidade de previsão dos mesmos (SP). Olhando em pormenor para os dados e para os grupos definidos acima, percebemos que, como se encontram num grupo isolado, os descritores 1D/2D são escolhidos. No grupo ii), que diz respeito aos *fingerprints* circulares, temos os CDK e os ExtCDK. Ao analisar os resultados da RF, percebemos que os CDK apresentam melhores resultados, pois tem uma Q e um MCC superior. Por isso, estes descritores também são escolhidos para a fase seguinte.

Restando apenas o grupo iii), dizem respeito aos *fingerprints* de fragmentos, encontram-se os Pubchem, Sub, SubC e MACCS. Olhando para os valores obtidos, percebemos que os Pubchem apresentam os melhores resultados por uma larga margem, sendo também escolhidas para a próxima etapa.

Na etapa de seleção de descritores foi utilizado o método de aprendizagem automática *k-Nearest Neighbors*, tendo em conta diferentes abordagens, como referido na metodologia: ***BestFirst***; ***GreedyStepwise***, com pesquisa *backwards*; ***PSOsearch***, variando os pesos: individuais (indW), *inertia* (ineW) e social (socW) (por exemplo, 0,33 para os pesos *inertia* e social e 0,34 para o peso individual). Foi usado também outro método de seleção de descritores através da ***Random Forest***, com a seleção dos 50, 100, 150 e 200 melhores descritores. Esta seleção é importante, pois as redes não funcionam bem com um elevado número de descritores. As tabelas 3.3, 3.4 e 3.5 representam os resultados desta seleção, tendo em conta os *sets* de descritores selecionados para esta etapa.

Tabela 3.3 – Seleção de descritores para o *set* 1D/2D.

Método	Obs	nº descritores	SE	SP	Q	MCC
Best First	-	86	0,582	0,866	0,771	0,469
GreedyStepwise	-	92	0,573	0,858	0,762	0,450
PSOsearch	indW 0,34	471	0,588	0,851	0,762	0,454
	ineW 0,34	384	0,625	0,848	0,773	0,482
	socW 0,34	460	0,599	0,848	0,764	0,459
RF	-	50	0,589	0,869	0,775	0,479
RF	-	100	0,597	0,866	0,775	0,482
RF	-	150	0,617	0,866	0,782	0,499
RF	-	200	0,630	0,858	0,782	0,501

Tabela 3.4 – Seleção de descritores para o *set* CDK.

Método	Obs	nº descritores	SE	SP	Q	MCC
Best First	-	62	0,658	0,817	0,764	0,473
GreedyStepwise	-	68	0,686	0,807	0,767	0,485
PSOsearch	indW 0,34	277	0,703	0,808	0,773	0,502
	ineW 0,34	250	0,688	0,816	0,773	0,497
	socW 0,34	245	0,687	0,818	0,774	0,500
RF	-	50	0,665	0,826	0,772	0,491
RF	-	100	0,698	0,822	0,780	0,514
RF	-	150	0,705	0,823	0,783	0,521
RF	-	200	0,709	0,820	0,782	0,521

Tabela 3.5 – Seleção de descritores para o *set* Pubchem.

Método	Obs	nº descritores	SE	SP	Q	MCC
Best First	-	10	0,395	0,865	0,707	0,295
GreedyStepwise	-	10	0,395	0,865	0,707	0,295
PSOsearch	indW 0,34	227	0,611	0,837	0,761	0,457
	ineW 0,34	229	0,615	0,841	0,765	0,464
	socW 0,34	217	0,595	0,843	0,759	0,449
RF	-	50	0,528	0,847	0,740	0,395
RF	-	100	0,551	0,837	0,741	0,403
RF	-	150	0,550	0,831	0,736	0,393
RF	-	200	0,547	0,835	0,738	0,396

Analisando os resultados apresentados, foi escolhido o melhor método de procura, para cada *set* de descritores, que fará parte da etapa seguinte (otimização do modelo). Os métodos escolhidos, tendo em conta os valores apresentados nas tabelas 3.3, 3.4 e 3.5 foram: *PSOsearch* com parâmetros ineW igual a 0,34 e indW e socW iguais a 0,33 para o *set* Pubchem; RF com seleção dos 150 para o *set* CDK e RF com seleção dos 200 melhores descritores para o *set* 1D/2D. Na fase de otimização cada *set* será representado pela melhor seleção aqui apresentada.

Os métodos de aprendizagem automática utilizados no estudo: *Random Forest* (RF), *Support Vector Machine* (SVM), *Multilayer Perceptron* (MLP) e *Neural Network* (NN). Como a NN e o MLP são métodos que utilizam redes neuronais artificiais, foi necessário olhar para os seus resultados e optar por escolher uma para se juntar às restantes (RF e SVM), por forma a ter uma melhor diversidade de resultados para ajudar na análise e na complementaridade do estudo. As tabelas 3.6 e 3.7, MLP e NN, respetivamente, representam os melhores resultados obtidos tendo em conta os *sets* de descritores utilizados.

Tabela 3.6 – Melhores resultados utilizando o *Multilayer Perceptron*.

Descritor ID	TP	TN	FP	FN	SE	SP	Q	MCC
1D 2D H = 7 L = 0.01 M = 0.01	1222	3375	628	803	0,603	0,843	0,763	0,457
CDK H = 7 L = 0.2 M = 0.2 Csc 45/55	1234	3414	589	791	0,609	0,853	0,771	0,475
Pubchem H = 5 L = 0.2 M = 0.2	1114	3403	600	911	0,550	0,850	0,749	0,419

Tabela 3.7 – Melhores resultados utilizando a *Neural Network*.

Descritor ID	TP	TN	FP	FN	SE	SP	Q	MCC
1D2D hl = 150 hd = 0.5 id = 0.01	1362	3491	512	663	0,673	0,872	0,805	0,555
CDK hl = 100 hd = 0.2 id = 0.1	1376	3471	532	649	0,680	0,867	0,804	0,555
Pubchem hl = 150 hd = 0.4 id = 0.01	1323	3504	499	702	0,653	0,875	0,801	0,544

Na tabela 3.6, H, L e M representam, respetivamente, *hiddenLayers*, *LearningRate* e *momentum*. Na tabela 3.7, hl, hd e id representam, respetivamente, *hiddenLayers*, *hiddenLayersDropoutRate* e *inputLayerDropoutRate*. Estes representam os parâmetros que variam quando se procede à otimização de um determinado modelo e são independentes de método para método. O peso representa a função *CostSensitiveClassifier* (Csc) e que permite o balanceamento da classe (ativo ou inativo). Ao analisar os valores de TP e TN (verdadeiros positivos e verdadeiros negativos) das duas tabelas, observamos que temos mais compostos TN que TP, evidenciando, novamente, uma falta de balanço, que é também comprovado pela SE e pela SP: podemos reparar que a SP é maior que a SE, o que nos indica que no nosso conjunto de treino temos mais compostos inativos do que ativos. E mesmo no caso específico em que se usa uma função que ajuda neste balanceamento (tabela 3.6, descritor CDK, Csc 45/55), continuamos a ter uma maior presença de compostos inativos, bem como uma maior capacidade de previsão dos mesmos. Quanto à escolha do melhor método de aprendizagem automática, neste caso particular, e analisando os resultados das duas tabelas, podemos facilmente perceber que a *Neural Network* apresenta melhores resultados em relação ao *Multilayer Perceptron*, como podemos ver pelos valores de Q e MCC: são melhores em todos os sets de descritores usados para a NN, sendo por isso escolhida, juntamente com a RF e a SVM, para a etapa de otimização.

3.3. Otimização do modelo

Nesta fase do estudo, e após a seleção dos descritores com melhores resultados, é necessário olhar mais atentamente para os métodos de aprendizagem automática. A otimização do modelo final tem por base a análise dos resultados das três abordagens e a otimização dos seus parâmetros, tendo em conta os descritores selecionados no ponto anterior. As tabelas 3.8, 3.9 e 3.10 representam esses resultados.

Tabela 3.8 – Otimização do modelo construído pela *Random Forest* tendo em conta a seleção de descritores.

Descritor ID	TP	TN	FP	FN	SE	SP	Q	MCC	Obs	Seleção
1D 2D nF = 15 nIt = 200	1391	3401	602	634	0,687	0,850	0,795	0,539	Csc - 40 60 - RF com descritores otimizados	RF200
CDK nF = 9 nIt = 500	1301	3568	435	724	0,642	0,891	0,808	0,557	Csc - 45 55 - RF com descritores otimizados	RF150
Pubchem nF = 13 nIt = 200	1186	3556	447	839	0,586	0,888	0,787	0,504	RF com descritores otimizados	PSOsearch ineW0,34

Tabela 3.9 – Otimização do modelo construído pela *SupportVectorMachine* tendo em conta a seleção de descritores.

Descritor ID	TP	TN	FP	FN	SE	SP	Q	MCC	Obs	Seleção
1D 2D eps = 0.001 loss = 0.1 cost = 542	1383	3340	663	642	0,683	0,834	0,784	0,516	Csc - 60 40 - SVM com descritores otimizados	RF200
CDK eps = 0.001 loss = 0.1 cost = 11	1204	3619	384	821	0,595	0,904	0,800	0,535	SVM com descritores otimizados	RF150
Pubchem eps = 0.001 loss = 0.1 cost = 123	1088	3523	480	937	0,537	0,880	0,765	0,449	SVM com descritores otimizados	PSOsearch ineW0,34

Tabela 3.10 – Otimização do modelo construído pela *Neural Network* tendo em conta a seleção de descritores.

Descritor ID	TP	TN	FP	FN	SE	SP	Q	MCC	Obs	Seleção
1D 2D hd = 0.5 id = 0.01 hl = 200	1272	3563	440	753	0,628	0,890	0,802	0,543	NN com descritores otimizados	RF200
CDK hd = 0.2 id = 0.1 hl = 150	1448	3261	742	577	0,715	0,815	0,781	0,520	NN com descritores otimizados	RF150
Pubchem hd = 0.4 id = 0.01 hl = 200	1263	3409	594	762	0,624	0,852	0,775	0,486	NN com descritores otimizados	PSOsearch ineW0,34

Neste ponto é necessário escolher qual o melhor método de aprendizagem automática (tendo em conta os melhores parâmetros apresentados), o melhor *set* de descritores e a melhor seleção efetuado no ponto anterior. Tanto nos resultados com a *Random Forest* como com a SVM observamos que o melhor resultado é obtido com o *set* CDK e com a seleção dos 150 melhores descritores por parte da RF. Olhando para a NN, a melhor construção do modelo é obtida com o set 1D/2D e com a

seleção dos 200 melhores descritores por parte da RF. Podemos observar que o problema do balanceamento das classes não foi corrigido, mesmo após a otimização de todos os modelos promissores. O modelo com o melhor equilíbrio de classes é efetuado com a *Neural Network* como podemos ver pelos valores da SE e SP (0,715 e 0,815, respetivamente), tendo sido utilizado o set CDK com a seleção dos 150 melhores descritores pela RF. Ainda assim, este ponto não chega para ser considerado o melhor modelo, e analisando os resultados num todo, podemos ver que o melhor modelo, e já otimizado, é construído com a *Random Forest* (com os parâmetros nF igual a 9 e nIt igual a 500), com o set CDK e utilizando o método RF com a seleção dos 150 melhores descritores. Este método utiliza ainda a função *CostSensitiveClassifier* (Csc) com o balanceamento das classes, como já referido anteriormente, mas que acaba por não ser suficiente para um equilíbrio eficiente. Ainda assim, como é necessário analisar os restantes parâmetros, vemos que este é o modelo mais promissor e com melhor resultado.

Foram calculadas dois tipos de funções radiais de distribuição (RDF, sigla em inglês): 128 RDF e 256 RDF. Este passo serviu para perceber se este tipo de descritor 3D iria aumentar a performance do modelo, melhorando os seus parâmetros. A tabela 3.11 resume os resultados obtidos.

Tabela 3.11 – Resultados dos melhores modelos (construídos pela RF) tendo em conta três tipos de descritores: CDK, CDK + 128 RDF e CDK + 256 RDF.

Descritor ID	TP	TN	FP	FN	SE	SP	Q	MCC	Obs	
CDK nF = 9 nIt = 500	1301	3568	435	724	0,642	0,891	0,808	0,557	**Csc - 45 55 - RF com descritores otimizados**	**Sel.Attr - RF150**
CDK + 128 RDF nF = 26 nIt = 500	1220	3642	361	805	0,602	0,910	0,807	0,550	"	"
CDK + 256 RDF nF = 30 nIt = 500	1202	3647	356	823	0,594	0,911	0,804	0,544	"	"

Primeiramente, podemos ver que o problema do equilíbrio de classe não melhora, muito pelo contrário, sofre um decréscimo com o uso dos dois tipos de descritores 3D. No que ao modelo diz respeito, não houve uma melhoria nos parâmetros, pelo que se pode afirmar que os descritores 3D, calculados a partir das RDFs, não foram uteis na otimização do modelo. Logo, o modelo criado pela RF, com o *set* CDK, nF e nIt iguais a 9 e 500, respetivamente, com a função Csc e seleção dos 150 melhores descritores por parte da RF é considerado o modelo otimizado e com o melhor resultado. Este método apresenta uma capacidade de previsão de compostos como ativos de 0,642, uma capacidade de previsão de compostos como inativos de 0,891, pelo que podemos afirmar que não se ultrapassou o desafio de equilibrar as duas classes presentes no estudo. Apresenta, ainda, uma previsibilidade de 0.808 e um coeficiente de correlação de *Matthews* igual a 0,557. Assim, é este modelo, com os parâmetros já referidos, que será utilizado no *screening* virtual.

3.4. *Screening* Virtual

A etapa do *screening* virtual pode dividir-se em três partes: cálculo da *average SOM distance* (ASD) para definição do domínio de aplicabilidade, previsão da classe estrutural tendo por base as redes neuronais de *Kohonen* e análise e escolha dos compostos promissores para teste de *docking*. Inicialmente, os compostos retirados da base de dados ZINC contabilizavam 1442 moléculas (ver ficheiro Excel que serve de suporte digital a esta dissertação, com o título "**Excel de suporte**"). As redes neuronais de *Kohonen* serviram para o cálculo da ASD e para a previsão das classes dos 1442 compostos.

A ASD foi calculada para todos os compostos utilizando o programa JatoonSOM (versão b2.18g), sendo que os valores foram depois normalizados para valores entre 0 e 1. As classes estruturais (de A a J) foram previstas usando o programa JatoonSOM (versão 2.2a). A figura 3.1 representa o mapa neuronal com as classes estruturais do conjunto de treino. Este mapa foi usado para prever as classes estruturais dos compostos do teste e do *screening* virtual. A figura 3.2 representa o mapa neuronal representado na figura 3.1, com as classes estruturais previstas para o conjunto de teste distribuídas no mapa de *Kohonen*.

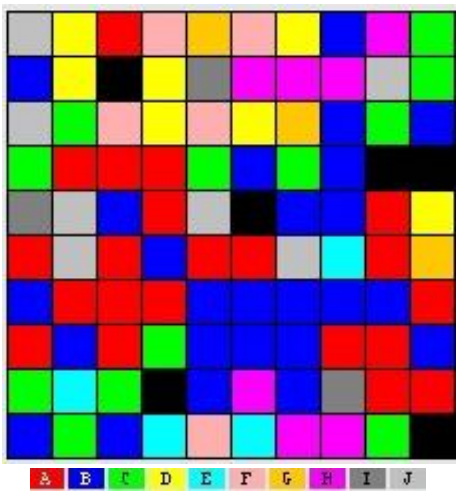

Figura 3.1 – Mapa neuronal das classes estruturais do conjunto de treino. O mapa tem um tamanho 10x10, foi usado um *Initial Learning Span* de 3 e um número de *Epochs* igual a 300. O número de classes definidas foi de 10 (de A a J), para evitar excesso de informação.

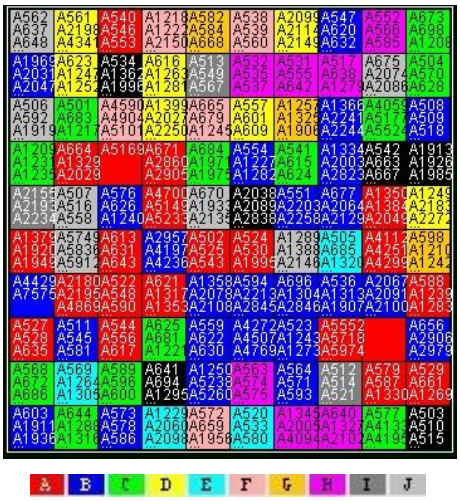

Figura 3.2 – Mapa neuronal do conjunto de treino com distribuição dos compostos do conjunto de teste, tendo em conta as classes estruturais definidas.

Cada uma das 10 classes é definida por um centróide: molécula que representa a classe estrutural em questão. A figura 3.3 representa as estruturas moleculares de cada um deles.

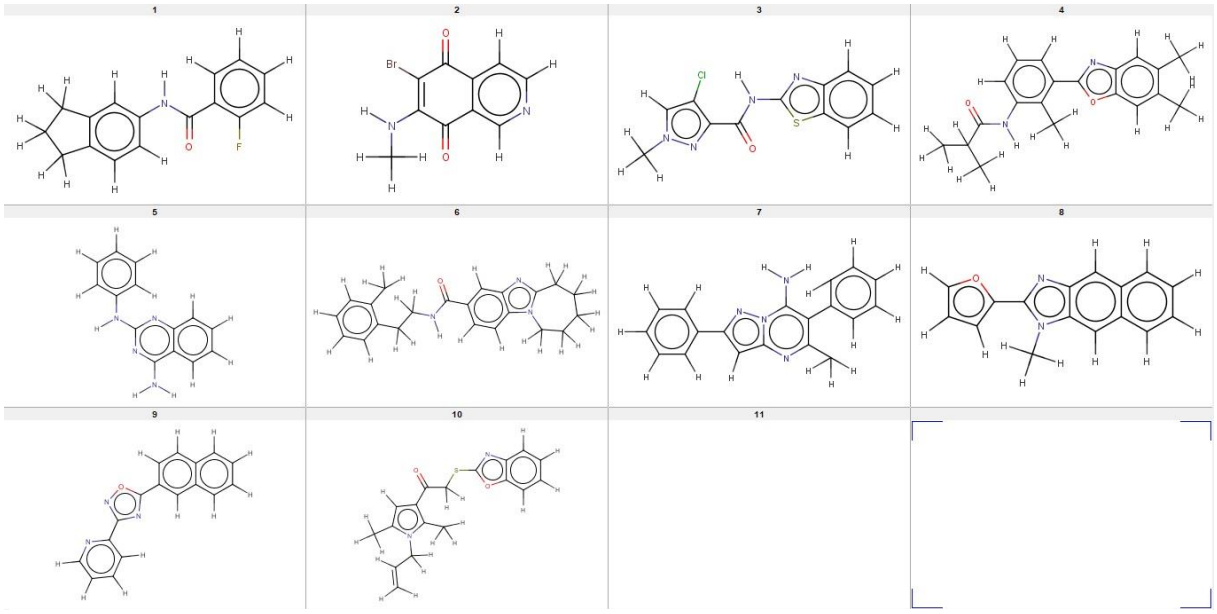

Figura 3.3 – Estruturas moleculares dos centróides que definem as 10 classes de A a J.

Estas moléculas representam cada uma das classes definidas e consegue perceber-se as diferenças existentes entre as mesmas.

Após estes passos, foi necessário definir um domínio de aplicabilidade (DA), tendo por base a *average SOM distance* (ASD). Neste caso, na lista de compostos, e após calculadas as probabilidades de um determinado composto ser ativo (obtidas no modelo RF com os CDK *Fingerprints*), foi encontrada uma molécula com uma ASD baixa (0,335) e uma alta probabilidade de ser ativa (0,928), mas que era considerada um resultado falso positivo. Tendo-se destacado pela ASD baixa e pela probabilidade alta, o domínio de aplicabilidade foi definido como: um composto pertence ao DA se apresentar uma ASD

superior a 0,34. Este foi o primeiro método de seleção: os compostos que não pertenciam ao domínio de aplicabilidade foram retirados do método de *screening* virtual. De seguida, foi necessário olhar para as classes. Como mostra a figura 3.1, os quadrados a preto representam neurónios que não definem nenhuma classe (que se denominou classe X – quadrados pretos nas figuras 3.1 e 3.2). Como seria de esperar, alguns compostos apresentavam esta classe e tiveram de ser retirados.

O passo seguinte consistiu em avaliar as classes e os seus erros, isto é, avaliar os falsos positivos e os falsos negativos. Esta avaliação consistiu em olhar para esses erros e, tendo em conta as classes com maior número de erros (E e J) e as moléculas com a menor probabilidade, escolher os compostos mais promissores. Desta seleção sobraram apenas 182 compostos, dos 1442 que faziam parte do conjunto ao início (ver "Excel de suporte"). Destes 182, foram escolhidas as 100 moléculas com maior probabilidade, evitando sempre as classes com mais erros (E e J, que ainda estavam presentes no conjunto final de 182 compostos). No final ficaram as 100 moléculas consideradas mais promissoras (sem classes E e J) que seguiram para a fase do *docking* molecular (ver "Excel de suporte").

A tabela 3.12 mostra os dez melhores resultados do *screening* virtual após todas as correções acima descritas terem sido efetuadas.

Tabela 3.12 – Os dez melhores resultados do *screening*

ID	Código ZINC	Probabilidade Ativo	ASD	Classe Prevista
SV710	ZINC000000008492	0,988	0,4081	H
SV1414	ZINC000000527386	0,956	0,4749	J
SV396	ZINC000003785268	0,954	0,4605	J
SV664	ZINC000003799072	0,954	0,4605	J
SV1337	ZINC000003775644	0,92	0,3879	H
SV1102	ZINC000001530688	0,884	0,4487	B
SV1339	ZINC000003798064	0,88	0,3547	H
SV410	ZINC000003798247	0,868	0,3637	H
SV255	ZINC000000599985	0,86	0,3879	H
SV392	ZINC000001853550	0,858	0,4762	E

Na tabela 3.12, o ID representado foi criado para efeitos do estudo em questão, não tendo qualquer significado. O código ZINC diz respeito ao código que representa os compostos na base de dados ZINC.

Estes resultados têm por base a maior probabilidade de ser ativo. Como também foi dito anteriormente, o objetivo seria evitar as classes com mais erros (E e J), mas ainda assim as mesmas estão presentes na lista final. Podemos ver que todos os compostos apresentados estão dentro do domínio de aplicabilidade (um dos requisitos) e apresentam, segundo a previsão do modelo construído, uma alta probabilidade de atividade.

3.5. *Docking* molecular

Os 100 compostos escolhidos pelo método de *screening* virtual foram utilizados em quatro ensaios de *docking*, como referenciado na metodologia. Estas quatro experiências têm em comum o facto de serem usados dois tipos de otimização da estrutura 3D dos ligandos: Corina e Cxcalc (Molecular Networks GmbH Computerchemie e ChemAxon, respetivamente).

Numa primeira experiência foi usada a estrutura do PDB 4HG7 (referente ao inibidor MDM_2). As coordenadas (x, y e z) da região de ligação que foi estudada neste ensaio de *docking* são, respetivamente, -24,12, 7,302 e -14,097. A figura 3.4 representa a estrutura PDB 4HG7.

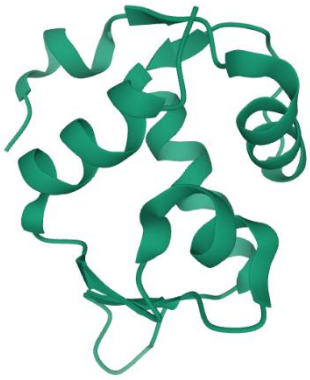

Figura 3.4 – Estrutura molecular do inibidor MDM_2 (PDB 4HG7).

Na segunda experiência foi utilizada a estrutura PDB 3DAB (referente ao inibidor MDM_2). As coordenadas x, y e z da região de ligação do ensaio são, respetivamente, 0,061, -24,677 e 8,477. A figura 3.5 representa a estrutura PDB 3DAB.

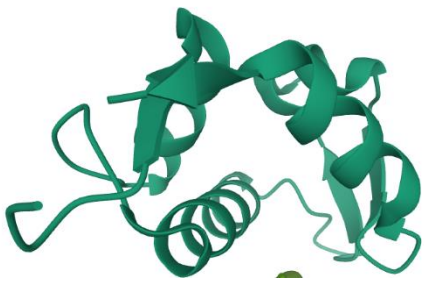

Figura 3.5 – Estrutura molecular do inibidor MDM_2 (PDB 3DAB).

Na terceira experiência foi usada a estrutura PDB 3DAB (neste caso já referente à p53). As coordenadas utilizadas foram as mesmas que na experiência 2, alterando-se apenas a estrutura alvo

do ensaio (ao invés do MDM_2, foi a estrutura da p53). A figura 3.6 representa a estrutura PDB 3DAB referente à p53.

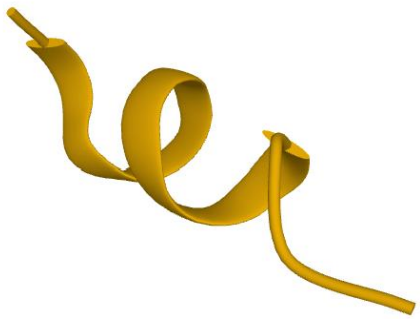

Figura 3.6 – Estrutura molecular da proteína p53 (PDB 3DAB).

Na última experiência foi utilizada a mesma estrutura do ensaio anterior (figura 3.6), tendo sido alteradas apenas as coordenadas da região de ligação, correspondendo a x, y e z os valores -2,223, -26,79 e 10,232, respetivamente.

Inicialmente foram escolhidos os 10 melhores resultados de cada uma das quatro experiências para que fosse possível analisar e perceber se havia 1 ou mais compostos que se destacassem (ver "Excel de suporte"). Segundo os dados obtidos pelos vários ensaios, duas moléculas revelaram-se promissoras na ativação (direta ou indireta) da p53. A tabela 3.13 apresenta os resultados obtidos pelas duas moléculas.

Tabela 3.13 – Resultados das experiências de *docking* molecular (valores em kcal/mol).

Composto	Corina Exp1	Cxcalc Exp1	Corina Exp2	Cxcalc Exp2	Corina Exp3	Cxcalc Exp3	Corina Exp4	Cxcalc Exp4	Código ZINC
Nilotinib	-9,5	-10	-8,4	-7,9	-5,8	-6,3	-5,8	-6,1	ZINC000006716957
Dihidroergotamina	-9,8	-9,3	-7,8	-7,4	-6,5	-6,4	-6,4	-6,3	ZINC000003978005
Nutlin-3 **(controlo)**	-8,3	-8,3	-6,4	-6,4	-4,6	-4,8	-5,1	-5,2	-

Os valores apresentados representam a energia da ligação entre os compostos e a estrutura 3D utilizada em cada experiência, como já referido acima, sendo que "Exp" corresponde a cada experiência enunciada anteriormente.

Em todas as experiências aqui representadas, os compostos apresentam melhores resultados que a molécula usada para controlo positivo (*Nutlin-3*, com atividade referenciada e estudada[42]). Como o composto dihidroergotamina (DHET) apresenta melhores resultados num maior número de experiências, foi o escolhido para passar para a fase seguinte: validação experimental. O composto

nilotinib não foi escolhido também por já apresentar resultados na literatura relacionados com o cancro, nomeadamente relacionados com o inibidor de p53, MDM_2.[43,44]

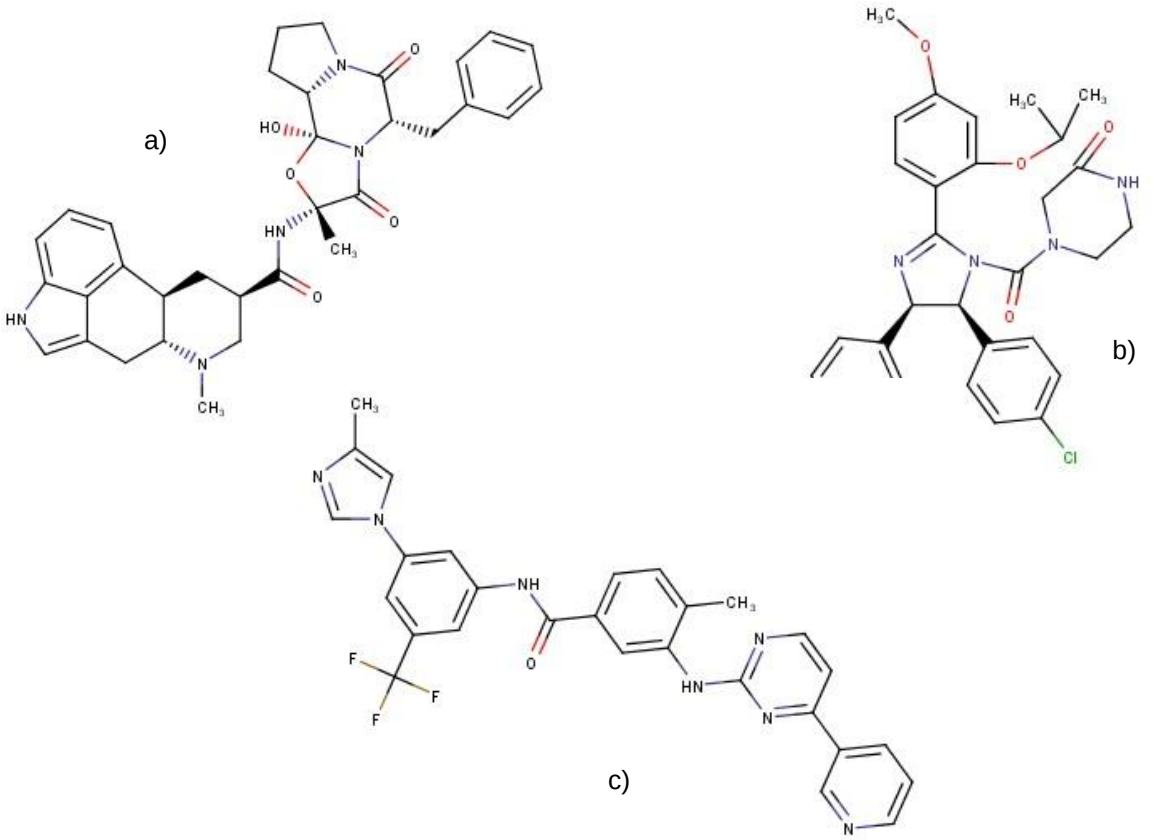

Figura 3.7 – Estruturas moleculares de a) Dihidroergotamina, b) Nutlin-3, c) Nilotinib.

Ainda assim, e sendo um composto caro, decidiu-se testar um derivado (mais barato) para tentar perceber se havia grandes alterações em relação aos valores apresentados. As tabelas 3.14 e 3.15 apresentam os resultados gerais dos dois compostos: DHET e dihidroergocristina (derivado, DHEC). A alteração que existe no derivado é um grupo isopropilo no lugar de um grupo metilo (este último presente na DHET).

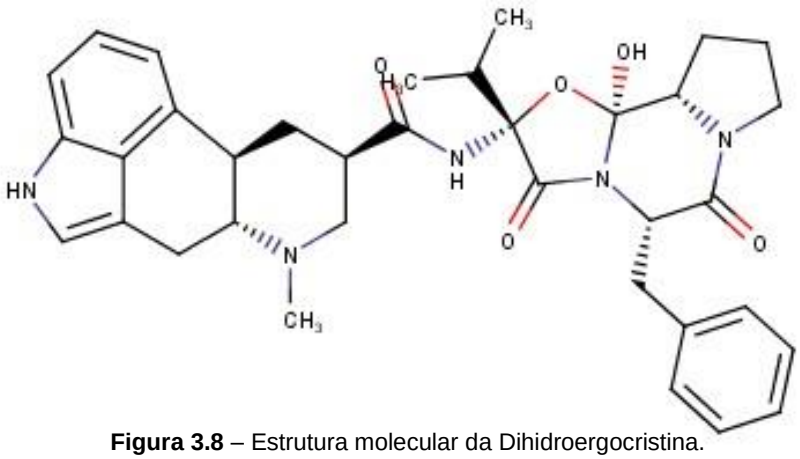

Figura 3.8 – Estrutura molecular da Dihidroergocristina.

Tabela 3.14 – Comparação dos resultados entre a DHET e a DHEC.

ID	Probabilidade Ativo	Classe prevista	ASD
DHEC	0,724	G	0,415
DHET	0,724	G	0,415

Tabela 3.15 – Comparação dos resultados entre a DHET, a DHEC e o controlo (*Nutlin-3*).

ID	Corina Exp1	Cxcalc Exp1	Corina Exp2	Cxcalc Exp2	Corina Exp3	Cxcalc Exp3	Corina Exp4	Cxcalc Exp4
DHET	-9,8	-9,3	-7,8	-7,4	-6,5	-6,4	-6,4	-6,3
DHEC	-10	-9,6	-7,5	-7,4	-6,2	-6,4	-6,2	-6,3
Controlo	**-8,3**	**-8,3**	**-6,4**	**-6,4**	**-4,6**	**-4,8**	**-5,1**	**-5,2**

Os valores apresentados dizem respeito à energia da ligação entre o ligando e o recetor de cada experiência, em kcal/mol. Como podemos perceber pelos resultados apresentados, o composto derivado, DHEC, obteve resultados também promissores tendo em conta o controlo utilizado. Como, em comparação com a DHET, os resultados são semelhantes, e tendo em conta que se revela uma solução economicamente mais favorável, optou-se por utilizar a dihidroergocristina (DHEC) nos ensaios de validação experimental.

3.6. Validação experimental

Na validação experimental foi usado o método da resazurina, como já referido nos métodos, para avaliar a viabilidade celular da linha HT29 WT quando o nosso composto é administrado. A descrição dos ensaios encontra-se nos Anexos. Neste ponto podemos dividir os dados obtidos em dois grupos: dados sem controlo e com controlo. Numa primeira fase não foi usado controlo de meio mas após uma análise cuidada dos primeiros resultados colocou-se a hipótese de usar esse controlo na análise de dados. Os dados sem controlo são feitos com 10% de resazurina (0,04 mg/mL) (resazurina com meio), sendo que os dados com controlo são feitos com 50% de resazurina (0,04 mg/mL), após se perceber que os valores de absorvância eram muito baixos. Todos estes dados foram analisados com recurso ao *GraphPad Prism* (versão 6).

3.6.1 Dados sem controlo (10% resazurina)

Esta análise teve em conta um conjunto de dados onde não havia controlo de meio, o qual poderia ajudar no tratamento da informação por forma a ser calculado um IC_{50}. Neste caso, foram feitas duas normalizações (uma no Excel e, posteriormente, no GraphPad).

A figura 3.9 representa o logaritmo da concentração (em µM) de composto (DHEC) em função da absorvância normalizada. As barras de erro representam o *standard error* (SE). Os resultados aí apresentados são referentes aos ensaios 1, 2 e 3.

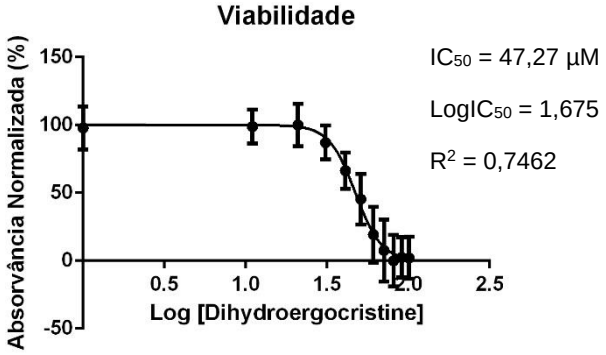

Figura 3.9 – Gráfico do logaritmo da concentração de Dihidroergocristina (em µM) em função da absorvância normalizada. As barras representam o *standard error*. Os resultados apresentados resultam de um *fit* não linear do logaritmo da concentração vs a resposta normalizada em percentagem. Ensaios 1, 2 e 3.

Segundo o *fit* não linear usado, obteve-se um IC_{50} de 47,27 µM, um $LogIC_{50}$ de 1,675 e um R^2 de 0,7462. Podemos perceber que existe morte celular ou inibição de crescimento (por ainda se ver algumas células nos poços de maior concentração, como ilustra a figura 3.10 a) pelo decréscimo da absorvância: como explicado nos anexos, a resazurina (composto que dá uma coloração azul/violeta) é metabolizada na presença de células, sendo convertida a resorufina (composto que dá uma coloração rosada). A absorvância que é retirada é fruto desta metabolização, pois o que vemos é a absorvância referente à resorufina: quando a concentração de composto é baixa, existem mais células e a resazurina é metabolizada, sendo que a absorvância lida (100%) diz respeito ao produto dessa metabolização (resorufina). À medida que a concentração de composto aumenta, o número de células diminui e a resazurina é menos metabolizada, logo será de esperar que a absorvância diminua devido ao facto de termos menos produto (resorufina). A figura 3.10 mostra o efeito do composto nas células (imagens referentes ao ensaio 1 – os restantes apresentam o mesmo tipo de efeito).

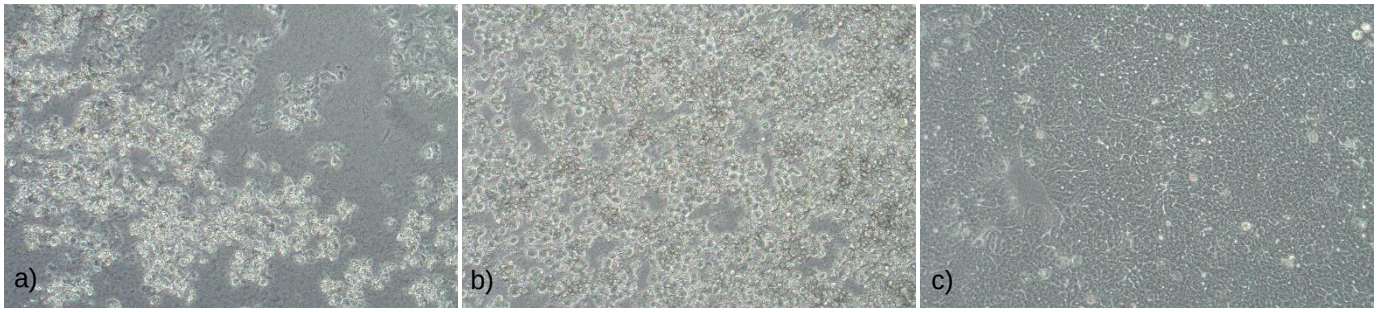

Figura 3.10 – Morfologia de células expostas a a) 100 µM de composto, b) 50 µM de composto e c) 0µM de composto. Imagens obtidas com a objetiva de 10X (Ensaio 1).

Podemos perceber que existe alteração da morfologia celular, o que nos indica que o composto tem efeito sobre as células. Percebe-se, também, que existe uma diminuição no número de células e de aglomerados, o que nos pode indicar que existe morte celular ou inibição do crescimento celular.

3.6.2 Dados com controlo (50% resazurina)

Esta análise teve em conta um conjunto de dados onde já havia controlo de meio. Neste caso, foi feita apenas uma normalização dos dados (já no GraphPad) o que, desde logo, ajuda no tratamento dos dados.

A figura 3.11 mostra o gráfico do logaritmo da concentração (em µM) em função da absorvância normalizada dos ensaios 4, 5 e 6.

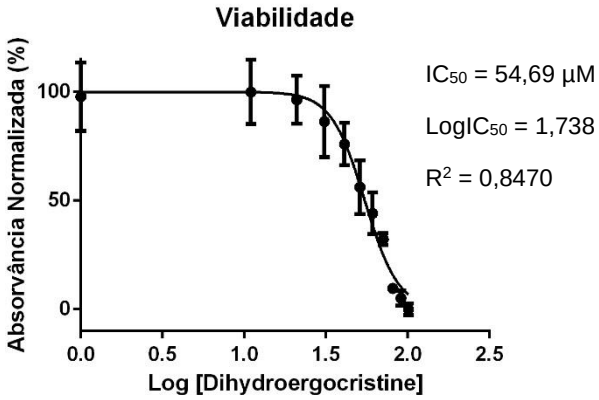

Figura 3.11 – Gráfico do logaritmo da concentração de *Dihydroergocristine* em função da absorvância normalizada. As barras representam o *standard error*. Os resultados apresentados resultam de um *fit* não linear do logaritmo da concentração vs a resposta normalizada em percentagem. Ensaios 4, 5 e 6.

Neste grupo obteve-se um IC_{50} de 54,69 µM, um $LogIC_{50}$ de 1,738 e um R^2 de 0,8470. O grupo de ensaios em questão obteve melhores resultados em relação aos dados sem controlo. Podemos depreender que esta melhoria se deveu tanto o aumento da percentagem de resazurina, como a utilização do controlo de meio, podendo assumir que a utilização deste controlo deve ser uma prioridade.

Analisando os resultados apresentados, percebe-se que existe morte celular ou inibição do crescimento celular, pois existe um decréscimo de absorvância e, se tivermos em atenção as imagens apresentadas na figura 3.9 (será idêntica no que se refere aos dados com controlo), conseguimos perceber que, em concentrações mais altas, existe uma diminuição do número de células. Além disso, consegue perceber-se que foi obtido um R^2 mais próximo de 1 e que é o mais alto no que se refere aos conjuntos de ensaios. Olhando ainda para o erro, neste grupo as barras de erro estão, no geral, diminuídas em relação aos dados da figura 3.9 e, por todas as razões já enumeradas, podemos assumir que este é o conjunto de ensaios com melhores resultados e mais rigorosos.

Podemos dizer que este primeiro valor de IC50, **54,60 µM**, já se encontra perto daquilo que seria um valor final, mas ainda há a necessidade de torna-lo mais rigoroso. Logo, pode assumir-se que seriam necessários mais ensaios para aumentar essa variável e obter resultados mais afinados. Neste caso não se pode dizer que o composto ativa, direta ou indiretamente, a proteína p53, ainda que se prove que o mesmo provoca morte celular ou inibição do crescimento celular, não descartando a hipótese de uma possível interação com a proteína (que só seria possível estudar com outros testes, por exemplo, Western Blot).

Capítulo 4 – CONCLUSÕES

4. Conclusões

Este estudo tinha como principal objetivo desenvolver uma metodologia computacional que permitisse prever e estudar se um ou mais determinados compostos teriam atividade (direta ou indireta) contra a proteína p53. De uma forma geral, pretendia criar-se um modelo que fosse capaz de prever esta atividade quando se lhe eram submetidas moléculas sobre as quais não havia essa informação.

Numa primeira fase, o modelo foi criado com sucesso, tendo conseguido parâmetros como um Q de 0,808 e um MCC de 0,557. Nesta construção ainda foi usada a função *CostSensitiveClassifier*, uma função que tinha o objetivo de balancear as classes em estudo (ativos e inativos). Foram selecionados os 150 melhores descritores através da técnica de aprendizagem automática *Random Forest*, que apresenta como um dos parâmetros a capacidade de atribuir importância aos descritores (*computeAttributeImportance*). Um dos desafios que não foi superado foi o facto de não se ter conseguido equilibrar as classes de compostos ativos e inativos, sendo que no modelo final, e segundo os parâmetros SE (capacidade de previsão de compostos como ativos) e SP (capacidade de previsão dos compostos como inativos), se obtiveram valores de 0,642 e 0,891, respetivamente, onde podemos ver que este equilíbrio ficou aquém do pretendido. Foram, ainda, utilizados descritores 3D (RDF) para testar se o aumento da complexidade e das características teria influência nos parâmetros do modelo. Podemos verificar que a introdução destes descritores não foi vantajosa, observando-se piores resultados quando comparando com os valores obtidos pelo *set* CDK apenas.

O processo de *screening* virtual, no geral, foi efetuado com sucesso. Foi necessário definir um domínio de aplicabilidade (compostos com ASD maior que 0,34) e foram definidas classes estruturais a partir do conjunto de treino (de A a J), tendo esse mapa de classes sido usado para prever as classes estruturais dos compostos do *screening* virtual. Apesar de haver compostos do conjunto de treino que não teriam classe (classe X) e apesar de não se enquadrarem, pode dizer-se que o mapa apresentava pouco erro. Nas classes previstas pelo mapa definido, havia classes com mais erros que outras (classes E e J), que sofreram uma tentativa de eliminação que acabou por não corresponder ao desejado, pois na lista de 100 compostos para *docking* molecular, ainda havia compostos que pertenciam a estas duas classes (alguns com probabilidade de serem ativos bastante alta).

No *docking* molecular foram testadas 100 moléculas, das quais foi escolhida apenas 1 para validação experimental. Ainda assim, este processo não correu exatamente como estava previsto. Inicialmente, houve dois compostos que se destacavam, claramente, dos restantes: nilotinib e dihidroergotamina (DHET). Após uma análise cuidada, percebeu-se que a DHET apresentava resultados ligeiramente superiores, pelo que foi escolhida. Entretanto, entrou nesta "equação" um derivado desta molécula: dihidroergocristina, DHEC (com a substituição de um grupo metilo por um grupo isopropilo). Esta apresentava resultados iguais no que se refere à previsão da atividade, classe estrutural e domínio de aplicabilidade (0,724, classe G, 0,415, respetivamente), sendo que no que ao docking se referia, os resultados eram bastante semelhantes (com pequenas diferenças, ainda que melhor do que o controlo utilizado). Pelo facto de ser uma molécula economicamente mais favorável, acabou por ser a escolhida para os ensaios celulares.

Na validação experimental, apesar de todos os percalços laboratoriais e alguns erros na execução do protocolo, chegou-se a um valor (ainda preliminar) de IC_{50}. A linha celular usada foi a linha HT29 *wild-type*, referente ao cancro colorretal humano. O objetivo seria, com a utilização deste composto, verificar se havia morte ou inibição de crescimento celular. De facto, isso verificou-se, pois em concentrações maiores, além de se verificar um número de células mais diminuído, a absorvância do ensaio também segue nesse sentido. Podemos, por isso, afirmar que o composto tem um efeito na linha celular utilizada, pelo que se chegou a um valor de IC_{50} de 54,69 µM. Convém ressalvar que este é um valor preliminar, pois seriam necessários mais ensaios para obter um valor mais robusto. Ainda assim, este objetivo foi parcialmente cumprido, pois percebemos que existe atividade, mas não temos a certeza se esta é contra a proteína p53.

Em suma, seriam necessários mais estudos para perceber se o composto Dihidroergocristina teria atividade contra a p53. Esta molécula apresenta efeito citotóxico, pois causa morte ou inibição do crescimento celular, mas não podemos afirmar que é por interação (direta ou indireta) com a p53. Para podermos saber se existiria esta interação, poderia ser feito um *Western Blot* para perceber se, com o aumento de composto, existe um aumento da expressão da proteína.

Capítulo 5 – BIBLIOGRAFIA

5. Bibliografia

[1] Gandomani, H. S., Yousefi, S. M., Aghajani, M., Mohammadian-Hafshejani, A., Tarazoj, A. A., Pouyesh, V., & Salehiniya, H. (2017). Colorectal cancer in the world: incidence, mortality and risk factors. Biomedical Research and Therapy, 4(10), 1656–1675.

[2] Yamaguchi, A., Kurosaka, Y., Fushida, S., Kanno, M., Yonemura, Y., Miwa, K., & Miyazaki, I. (1992). Expression of p53 protein in colorectal cancer and its relationship to short-term prognosis. Cancer.

[3] Scott, N., Sagar, P., Stewart, J., Blair, G. E., Dixon, M. F., & Quirke, P. (1991). p53 in Colorectal Cancer : Clinicopathological Correlation and prognostic significance. British Journal of Cancer, 63, 317–319.

[4] Vassilev, L. T., Vu, B. T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., … Liu, E. A. (2004). In Vivo Activation of the p53 Pathway by Small-Molecule Antagonists of MDM2. 303(February), 844–849.

[5] Ryan, K. M., Phillips, A. C., & Vousden, K. H. (2001). Regulation and function of the p53 tumor suppressor protein. 332–337.

[6] Wade, M., Li, Y. C., & Wahl, G. M. (2013). MDM2, MDMX and p53 in oncogenesis and cancer therapy. *Nature Reviews Cancer*, *13*(2), 83-96.

[7] Pereira, F., & Aires-de-Sousa, J. (2018). Computational methodologies in the exploration of marine natural product leads. Marine drugs, 16(7), 236.

[8] Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational methods in drug discovery. Pharmacological reviews, 66(1), 334-395.

[9] Cruz, S. M. D. M. D. (2016). Desenvolvimento de uma abordagem computacional para a descoberta de compostos-líderes para fármacos anticancerígenos (Master dissertation)

[10] Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of computational chemistry, 32(7), 1466-1474.

[11] Xue, L., & Bajorath, J. (2000). Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. Combinatorial chemistry & high throughput screening, 3(5), 363-372.

[12] Shahlaei, M. (2013). Descriptor selection methods in quantitative structure–activity relationship studies: a review study. Chemical reviews, 113(10), 8093-8103.

[13] Todeschini, R., & Consonni, V. (2009). Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references (Vol. 41). John Wiley & Sons.

[14] Wang, L., Le, X., Li, L., Ju, Y., Lin, Z., Gu, Q., & Xu, J. (2014). Discovering new agents active against methicillin-resistant Staphylococcus aureus with ligand-based approaches. Journal of chemical information and modeling, 54(11), 3186-3197.

[15] Gramatica, P., & Sangion, A. (2016). A historical excursus on the statistical validation parameters for QSAR models: a clarification concerning metrics and terminology. Journal of chemical information and modeling, 56(6), 1127-1131.

[16] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences, 43(6), 1947-1958.

[17] Lavecchia, A. (2015). Machine-learning approaches in drug discovery: methods and applications. Drug discovery today, 20(3), 318-331.

[18] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

[19] Yao, X. J., Panaye, A., Doucet, J. P., Zhang, R. S., Chen, H. F., Liu, M. C., ... & Fan, B. T. (2004). Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. Journal of chemical information and computer sciences, 44(4), 1257-1266.

[20] "Quanto mais neurónios uma espécie tem no córtex, mais tempo ela leva para chegar à adolescência". (2019, 27 de março). Retirado de https://www.publico.pt/2019/03/27/ciencia/noticia/suzana-herculanohouzel-numeros-celulas-apenas-primata-cerebro-1866730

[21] Artificial Neural Network. Retirado de http://www.saedsayad.com/artificial_neural_network.htm

[22] Gil, D., & Manuel, D. J. (2009). Diagnosing Parkinson by using artificial neural networks and support vector machines. Global Journal of Computer Science and Technology, 9(4).

[23] Gradient Descent. Retirado de http://www.saedsayad.com/gradient_descent.htm

[24] KNN Classification. Retirado de http://www.saedsayad.com/k_nearest_neighbors.htm

[25] Hanke, J., & Reich, J. G. (1996). Kohonen map as a visualization tool for the analysis of protein sequences: multiple alignments, domains and segments of secondary structures. Bioinformatics, 12(6), 447-454.

[26] JATOON Documentation. Retirado de http://joao.airesdesousa.com/jatoon/v103/doc/index.htm

[27] Reddy, A. S., Pati, S. P., Kumar, P. P., Pradeep, H. N., & Sastry, G. N. (2007). Virtual screening in drug discovery-a computational perspective. Current Protein and Peptide Science, 8(4), 329-351.

[28] ChEMBL Database. Retirado de https://www.ebi.ac.uk/chembl/

[29] Reaxys. Retirado de https://www.reaxys.com/

[30] ZINC. Retirado de https://zinc.docking.org/

[31] Marvin. Retirado de https://chemaxon.com/products/marvin

[32] Standardizer. Retirado de https://chemaxon.com/products/chemical-structure-representation-toolkit

[33] PaDEL-Descriptor. Retirado de http://www.yapcwsoft.com/dd/padeldescriptor/

[34] Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational methods in drug discovery. Pharmacological reviews, 66(1), 334-395.

[35] Zhang, Q., Zheng, F., Fartaria, R., Latino, D. A., Qu, X., Campos, T., ... & Aires-de-Sousa, J. (2014). A QSPR approach for the fast estimation of DFT/NBO partial atomic charges. Chemometrics and Intelligent Laboratory Systems, 134, 158-163.

[36] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[37] Trott, O., & Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry, 31(2), 455-461.

[38] Cygwin. Retirado de https://cygwin.com/index.html

[39] RCSB PDB: Homepage. Retirado de https://www.rcsb.org/

[40] RCSB PDB - 4HG7. Retirado de https://www.rcsb.org/structure/4HG7

[41] RCSB PDB - 3DAB. Retirado de https://www.rcsb.org/structure/3DAB

[42] Xue, X., Wei, J. L., Xu, L. L., Xi, M. Y., Xu, X. L., Liu, F., ... & Lu, M. C. (2013). Effective screening strategy using ensembled pharmacophore models combined with cascade docking: application to p53-MDM2 interaction inhibitors. Journal of chemical information and modeling, 53(10), 2715-2729.

[43] Zhang, H., Gu, L., Liu, T., Chiang, K. Y., & Zhou, M. (2014). Inhibition of MDM2 by nilotinib contributes to cytotoxicity in both Philadelphia-positive and negative acute lymphoblastic leukemia. *PloS one*, *9*(6), e100960.

[44] Carter, B. Z., Mak, P. Y., Mak, D. H., Ruvolo, V. R., Schober, W., McQueen, T., ... & Andreeff, M. (2015). Synergistic effects of p53 activation via MDM2 inhibition in combination with inhibition of Bcl-2 or Bcr-Abl in CD34+ proliferating and quiescent chronic myeloid leukemia blast crisis cells. *Oncotarget*, *6*(31), 30487.

Capítulo 6 – ANEXOS

Anexos

6.1 Composição dos reagentes usados na validação experimental

Meio RPMI completo (RPMIc) (100mL): 1mL de Pen Strep (Gibco)
1mL de Glutamina 200 mM (100X, Gibco)
1mL de Piruvato de Sódio 100 mM (100X, Gibco)
1mL de Aminoácidos não essenciais (100X, Gibco)
10mL de FBS, *Qualified*, *Heat inactivated*
Meio RPMI 1640 (1X, Gibco) até perfazer 100mL

Tripsina (10mL): 1mL 0.5% Tripsina EDTA (10X, Gibco)
Perfazer até aos 10mL com PBS

PBS (*Phosphate buffered saline*): Solução que continha 1.47 mM de KH_2PO_4, 4.29 mM de NA_2HPO_4 e 2.68 mM de KCl, em água destilada (pH=7.4)

6.2 Teste de micoplasma

Objetivo: verificar se existe a presença de micoplasma (um tipo de contaminante neste tipo de estudos)
Procedimento em inglês, baseado no protocolo "PCR Mycoplasma Detection Set" por Takara Bio Inc (http://www.takara.co.kr/file/manual/pdf/6601_e.v0703.pdf):

"1.1. *Sample collection:*

1.1.1. Collect approximately 0.5-1 ml of cell culture supernatant [from a dense culture (80-100% confluent), that has been cultivated for at least 3 days after subculture] into a sterile 1.5 ml tube. Identify the tube with the name of the cell line, date, passage and your initials. Store the tube at -20ºC, in the blue rack identified as "Mycoplasma Test", in the last drawer of freezer B.

1.2. *Preparation of the samples to test and controls:*

1.2.1. Take out all the reagents that are needed to a box with ice and allow them to thaw.

1.2.2. Carefully label 0.2 ml tubes in order to identify the samples that will be tested (ex. A, B, C, D; 1, 2, 3, 4…).

1.2.3. Add 0.5 µl of supernatant of each sample/control to the respective tube and keep samples on ice.

1.2.3.1. Controls:

1.2.3.1.1. At least two positive controls for contamination (supernatant of previously contaminated cells);

1.2.3.1.2. One negative control for contamination (supernatant of previously not-contaminated cells).

2. Reaction mixture preparation:

2.1. Prepare the reaction mixture in a 1.5 ml tube by combining the reagents as shown in Table 2. Always consider 3 samples in excess, to avoid pipetting errors. At the end, each sample should comprise a total volume of 25 µl (24.5 µl of reaction mixture + 0.5 µl of supernatant sample). Keep all reagents on ice.

Note: Care should be taken to prevent cross contamination when adding reagents to the reaction mixture tube.

Table 2 – Reagents and volumes needed for the reaction mixture of 1 sample. For multiple samples, scale-up the values accordingly. Taq Polymerase should be the last reagent to be added

Reagent	Volume	Storage
RNase free H2O (Nzytech)	6,09 µl	Room temperature, sterile, aliquots at -20°C
5x Green Go Taq Flexi Buffer or Transparent Buffer (Promega)	10 µl	-20°C
MgCl2 (Promega)	2 µl	-20°C
dNTP Mix 100 mM (Applied Biosystems)	0,16 µl	-20°C
Primer F1 (Myco F1), 10 µM	1 µl	-20°C
Primer F1* (Myco F1t), 10 µM	1 µl	-20°C
Primer R1 (Myco R1), 10 µM	1 µl	-20°C
Primer R1c (Myco R1ct), 10 µM	1 µl	-20°C
Primer R1T (Myco R1at), 10 µM	1 µl	-20°C
Primer R1TTC (Myco R1ac), 10 µM	1 µl	-20°C
Go Taq DNA Polymerase (Promega)	0,25 µl	-20°C

3. PCR reaction:

3.1. Place all tubes in the thermal cycler. Set the parameters according to Table 3 and perform PCR.

3.1.1. There is currently a program (MYC) inserted in the thermal cycler, with these conditions.

Table 3 – PCR reaction conditions (approximately 2h 40min).

Step	Temperature	Time	
Initial denaturation	95°C	4 min	
x 34 Cycles	**Denaturation**	95°C	30 sec
Annealing	55°C	2 min	
Extension	72°C	1 min	
Final extension	72°C	7 min	
Hold	4°C	∞	

4. Analysis of the PCR products by gel electrophoresis:

4.1 *Preparation of an 1.5% agarose gel:*

4.1.1. Prepare the casting tray by adding yellow tape to the side openings, in order to close all sides of the tray;

4.1.2. Depending on the casting tray, different volumes of TAE 1x are used: 4.1.2.1. Small casting tray (10 lanes) – 40 ml;

4.1.2.2. Medium casting tray (15 lanes) - 120 ml;

4.1.2.3. Big casting tray (50 lanes) – 250 ml

4.1.3. Weight the correct amount of agarose according to the casting tray that is going to be used and add the respective volume of TAE 1x into an Erlenmeyer;

4.1.4. Heat up the mixture on a microwave until the agarose is dissolved and the solution becomes transparent. Let the solution boil for some seconds;

4.1.5. When the gel is ready, cool it by placing the Erlenmeyer under running water;

4.1.6. When the temperature is good, add the correct volume of GreenSafe (NzyTech) to the gel (2 µl/100 ml), mix well and add the gel to the casting tray. Carefully remove the bubbles from the gel with the help of the comb and/or a micropipette tip;

4.1.7. Place the comb on its place and let the gel polymerize (15-30 minutes, depending on the size of the gel);

4.1.8. Prepare enough volume of TAE 1x so that it will cover the gel in the electrophoresis chamber (at least 1.5 liters);

Note: It is better to open the microwave frequently to check if the gel is ready than to wait a longer time; there is a risk that the gel boils and goes out of the Erlenmeyer.
Note: do not let the gel cool down too much, or it will start polymerizing in the Erlenmeyer.

4.1.9. When the gel is polymerized, carefully remove the yellow tape from the casting tray and add the casting tray with the gel to the electrophoresis chamber. Add TAE 1x until the gel is covered and carefully remove the comb.

4.2. *Sample loading:*

4.2.1. With a P10 micropipette, add 5 µl of DNA marker (DNA Ladder V, NZytech) to at least one lane;

4.2.2. Add 10 µl/lane of each PCR product;

4.2.3. Cover the electrophoresis chamber and connect it to the power source. Turn on the power source, program constant voltage at 90-130 V and start the run. Check for bubbles forming in the electrophoresis chamber, as this is an indication of electric current flowing. Depending on the voltage, the run should take between 60 to 120 minutes.

4.3. *Gel analysis:*

4.3.1. Remove the casting tray containing the gel from the electrophoresis chamber.

4.3.2. The Geldoc should be always off, so turn it on to start.

4.3.3. Enter in the PC with the credentials “DVC310” and open the software “Image Lab”.

4.3.4. Place the gel directly in the tray (it should be stored outside the Geldoc, next to it) and then inside the Geldoc.

4.3.5. In the image lab select “New protocol”. In the Application tab select “Ethidium Bromide” – UV tray (purple).

4.3.6. When the gel e placed correctly in the tray and inside the Geldoc, select “Run Protocol” to start acquiring.

4.3.7. To save the image: File -> Export -> Export Pulse Net -> Select the “319 PV” folder inside the documents. The image should be saved as a tiff file.

Note: If the transparent buffer has been used, add 2 µl of loading buffer to each PCR product before loading it into the gel. In this case, the final volume of each sample is 12 µl.

Important notes:

- TAE 10x (1 liter): o Trisbase 0.4 M (48.2 g)
- Acetic acid 1.13% (11.42 ml)
- EDTA 10 mM (20 ml of the solution of EDTA 0.5 M, pH=8.0)
- All components diluted in dH2O
- To use, this solution should be diluted 10 times in dH2O
-
- Cell lines that were under mycoplasma treatment should be cultured in the absence of mycoplasma active antibiotics for 2 passages to maximize test sensitivity and avoid false negatives.
- If cells are contaminated, they should be discarded. If no other cells exist, they should be put in quarantine (manipulated in the flow chamber of the second floor) and treated with Plasmocin/Plasmocure for two weeks.”

6.3 Protocolo do teste de viabilidade utilizando o método da resazurina

O protocolo utilizado dividiu-se em três etapas:

1- Cultura celular

Dados sem controlo → células colocadas em cultura em placas de 24 poços (250 mil células por poço), num volume final de 2mL por poço.

Esquema da placa de 24 poços:

C1	C2	C3	C4	C5	C6
C7	C8	C9	C10	C11	C12
D1	D2	D3	D4	D5	D6
D7	D8	D9	D10	D11	D12

Dados com controlo → células colocadas em cultura em placas de 96 poços (25 mil células por poço), num volume final de 200µL por poço.

Esquema da placa de 96 poços

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12

D	M
D	M
D	M

Nota: Em todos os poços no esquema representado acima foram colocadas células, exceto nos poços com a letra M (controlo de meio)

2- Adição de composto

Após ser retirado o sobrenadante dos poços (células já aderidas), é adicionado o composto (dissolvido em DMSO) com meio de cultura.
O composto foi adicionado nos poços tendo em conta as tabelas:

Placa de 24 poços

	[Composto] µM	Volume composto µL	Volume final
C1	100	10	2mL
C2	90	9	
C3	80	8	
C4	70	7	
C5	60	6	
C6	50	5	
C7	40	4	
C8	30	3	
C9	20	2	
C10	10	1	
C11	0	0	

Nota: Os volumes de DMSO (D1, D2, …) colocados nos poços são iguais aos volumes de composto aí colocados, pois o composto foi dissolvido em DMSO e este podia ser tóxico (controlo).

Placa de 96 poços

	[Composto] µM	Volume composto µL	Volume final
C1	100	10	2mL
C2	90	9	
C3	80	8	
C4	70	7	
C5	60	6	
C6	50	5	
C7	40	4	
C8	30	3	
C9	20	2	
C10	10	1	
C11	0	0	
M	Meio (controlo sem células)		

Nota: Neste caso os poços foram feitos em triplicado. Nas placas de 96 poços foi apenas utilizada uma condição com DMSO (D), correspondente ao volume máximo de composto colocados nos poços. Neste ensaio já existe controlo de meio. Foram esperadas 24 horas até se fazer o ensaio de viabilidade com a resazurina.

3- Incubação com resazurina e leitura de resultados

Após 24h, o sobrenadante foi retirado e os poços lavados com PBS. De seguida, é adicionada a resazurina (10 ou 50%, consoante os ensaios) com meio de cultura. O tempo de incubação foi de 2h.

Dados sem controlo → Após a lavagem, adicionou-se meio com 10% resazurina. Após a incubação, a placa foi lida num leitor de microplacas e os resultados foram analisados.

Dados com controlo → Da mesma forma que o anterior, adicionou-se meio com 50% de resazurina. Após as 2h, a placa foi lida e os resultados analisados.