Bernardo Figuerêdo Domingues

3D-Pharma: Uma Ferramenta para Triagem Virtual Baseada em Fingerprints de Farmacóforos

Belo Horizonte

28 de outubro de 2013

Bernardo Figuerêdo Domingues

3D-Pharma: Uma Ferramenta para Triagem Virtual Baseada em Fingerprints de Farmacóforos

Orientador:
Prof. Dr. Júlio César Dias Lopes

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Belo Horizonte

28 de outubro de 2013

BERNARDO FIGUERÊDO DOMINGUES

3D-PHARMA: UMA FERRAMENTA PARA TRIAGEM VIRTUAL BASEADA EM FINGERPRINTS DE FARMACÓFOROS

Tese de Doutorado apresentada à Universidade Federal de Minas Gerais, como parte da exigência do Programa de Pós-Graduação em Bioinformática, do Departamento de Bioquímica e Imunologia, para obtenção do título de Doutor

Aprovado em: / /2012

Prof. Dr. Júlio César Dias Lopes
Departamento de Química - ICEx - UFMG
Orientador

Prof. Dr. Carlos Henrique da Silveira
Campus Avançado Itabim - UNIFEI

Prof. Dr. Alex Guterres Taranto
Campus Centro-Oeste Dona Lindu - UFSJ

Prof. Dr. Antônio Flávio de Carvalho Alcântara
Departamento de Química - ICEx - UFMG

Prof. Dr. Wagner Meira Junior
Departamento de Ciência da Computação - ICEx - UFMG

Dedicatória

Dedico este trabalho a todos cuja companhia tive de sacrificar por causa do mesmo. Principalmente, dedico aos meus pais, Paulo e Minervina, por todo o apoio. À Priscila, pela paciência, companheirismo e ao enorme carinho e amor. E aos amigos do DCC, da Grad022 e dos Cotonetes, a quem estou devendo horas e horas de companhia, cerveja e papos jogado fora.

Agradecimentos

Primeiramente devo agradecer a agora Professora Dra. Raquel Cardoso que, a época uma aluna deste mesmo programa, me mostrou como aplicar a Ciência da Computação em áreas aparentemente não-afins. Ao Prof. Dr. Júlio Lopes, que me acolheu no Departamento de Química e no seu laboratório, e teve a paciência de me reensinar e ampliar os conceitos que residiam no meu Ensino Médio. Aos professores Glória, Miguel, Glaura, Wagner, Marcelo e Gisele, pelos conhecimentos extras adquiridos durante o curso. Aos colegas do NEQUIM: Andrelly, Ramon, Fábio, Moema, Sérgio, Julio Rodriguez e Eduardo; e aos alunos que por lá passaram: Henrique, Kellen, Pedro e Marcos. A todos agradeço as discussões, insights e distrações durante esses anos. Agradeço aos colegas da Bioinformática: Ricardo, Wagner, Valdete, Douglas, Bráulio, Calouro, Raquel, Cris e Chico e principalmente ao meu grande amigo Deive, a quem devo várias e várias discussões (e cafés) sobre o trabalho de ambos. Aos amigos e colegas do DCC, especialmente ao Fabiano Botelho, por ter me "emprestado"a CMPH e ao Pedro Calais, pelas inúmeras revisões, cafés e insights.

“Every year is getting shorter
Never seem to find the time
Plans that either come to naught
Or half a page of scribbled lines”
Pink Floyd - Time (D. Gilmour / R. Waters / N. Mason / R. Wright)

Resumo

A indústria farmacêutica vive uma crise sem precedentes, cuja causa pode ser atribuída à queda vertiginosa de descobertas e registros de novas entidades moleculares, agravada pela proximidade da expiração de patentes altamente lucrativas e aos custos crescentes associados à pesquisa e desenvolvimento. Uma possível saída encontra-se em inovações no processo de descoberta de fármacos, nas quais a Bioinformática e a Quimioinformática têm um papel essencial na seleção e desenvolvimento racional de compostos candidatos a fármaco antes dos custosos testes clínicos. Neste contexto, uma das técnicas computacionais centrais ao Projeto Racional de Fármacos é a Triagem Virtual, que aplica técnicas *in silico* para a descoberta de novos compostos bioativos. A triagem virtual baseada em ligantes, que usa apenas a estrutura de compostos que possuem atividade previamente conhecida, é uma das mais utilizadas. Este trabalho visa apresentar o 3D-Pharma, um método de triagem virtual baseado na estrutura de ligantes ativos, que usa a informação derivada de características farmacofóricas dos átomos das moléculas, codificadas em vetores binários, para construir modelos simples e preditivos. No 3D-Pharma, as moléculas são descritas por representações que consideram múltiplos tautômeros com seus respectivos estados de protonação e múltiplas conformações (abordagem Multi-Espécie Multi-Modo). Propriedades farmacofóricas são atribuídas a pontos que correspondem à interação potencial dos átomos pesados de cada representação. Estes pontos são agrupados em arranjos de três pontos pelo 3D-Pharma. O conjunto destes tripletes de farmacóforo é codificado em uma *fingerprint*, um vetor que determina a presença ou ausência de cada configuração de três pontos em uma molécula. Todas as representações são unificadas em um único vetor, e o conjunto de vetores de compostos ativos são usados para a construção os modelos, os quais são validados internamente através de um novo protocolo exaustivo de validação cruzada. Os estudos de validação externa mostram o alto poder de predição dos modelos produzidos pelo 3D-Pharma para dez conjuntos de compostos ativos e inativos oriundos do DUD (*Directory of Useful Decoys*). Uma análise de performance entre o 3D-Pharma e outros sete métodos disponíveis na literatura mostrou que este apresenta melhores resultados quanto à acurácia, reconhecimento precoce e recuperação de

diversidade estrutural.

Abstract

The pharmaceutical industry is going through a crisis of unheard proportions. Its cause can be related to the abrupt fall of new molecular entities approval by regulatory agencies, aggravated by the proximity of the expiration date of highly profitable classes of patented compounds and the increasing aggregated cost of drug design. The innovation on the drug discovery process may be one way out of this situation, and Bioinformatics and Chemoinformatics have a pivotal role on selection and rational design of drug candidates, looking for elimination of non-promising substances from the discovery pipeline before the highly costly clinical trials. In this context, Virtual Screening is regarded as a invaluable asset in rational drug design. Ligand-Based Virtual Screening is one of the oldest and most utilized techniques used in computer-aided molecular design, since chemical data is readily and widely available. Nevertheless, this work presents 3D-Pharma, a new Ligand-Based Virtual Screening method that uses fingerprints of pharmacophore triplets at atomic resolutions to build very simple and predictive models. Within 3D-Pharma the molecules are described by multiple representations that comprehend several prototropic species and conformations (multiple species, multiple mode approach). Pharmacophoric features are assigned to points that share spacial coordinates and interaction properties to heavy atoms of each molecular representation. All possible three-point pharmacophores of each representation of a moloceule are indexed in a fingerprint, and the multiple representations of a compound are concatenated into a unique fingerprint that accounts for most of its chemical and conformational diversity. The biological activity of an ensemble of active molecules are represented by a single modal fingerprint or model, validated through a new exhaustive 10-fold cross-validation scheme, which improves robustness and internal consistency of the models, as well as its predictive power. Retrospective validation studies were made with 10 datasets of active compounds and decoys gathered from the DUD database. They show the high predictive power of the models built by 3D-Pharma from three external and independent datasets of bioactive compounds (Drugs, PDB Ligands and WOMBAT), which was compared against seven state-of-the-art LBVS methods. We concluded that 3D-Pharma overperforms all other state-of-the-art LBVS tools analyzed, in

terms of global accuracy as well as scaffold hopping and early recovery capacities. Furthermore, the models produced by 3D-Pharma are simple, robust, consistent and predictive.

Lista de Figuras

Lista de Tabelas

Lista de Símbolos e Abreviações

2SHA	*Two-Step Hierarchical Assignment*
4D FAP$_{OA}$	*4D Flexible Atom Pairs (Optimal Assignment)*
ADMET	Absorção, Distribuição, Metabolismo, Excreção e Toxicidade
ALR2	Aldose Redutase
AMPc	Adenosina Monofosfato Cíclico
ANN	*Arificial Neural Networks*
ANVISA	Agência Nacional de Vigilância Sanitária
AR	Receptor de Androgênio
AT III	Anti-Trombina III
ATP	Adenosina Tri-Fosfato
AUC	*Area Under Curve*
awROC	*Arithmetic Weighted ROC*
BEDROC	*Boltzmann-Enhanced Discrimination of ROC*
CCG	*Chemical Computing Group*
CDK2	Ciclina Dependente de Cinase 2
cDNA	DNA Complementar
CMPH	*C Minimal Perfect Hash*
CoMFA	*Comparative Molecular Field Analysis*
COX-2	Ciclooxigenase-2
Da	Dalton
DCOP	Doença Pulmonar Obstrutiva Crônica
DNA	Ácido Desoxirribonucléico
dp	Desvio Padrão
DUD	*Directory of Useful Decoys*
EF	*Enrichment Factor*

EGFR Receptor de Fator de Crescimento Epidérmico
EM *Expectation-Maximization*
FDA *Food and Drug Administration*
FLAP *Fingerprints for Ligands And Proteins*
FV Fonte de Variância
FX$_\alpha$ Fator de Coagulação X α
g.l. Graus de Liberdade
GB Gigabyte
GHz Gigahertz
GMM Modelo de Mistura de Gaussianas
GMPc Guanosina Monofosfato Cíclico
HIV Vírus da Imunodeficiência Humana
HIVRT Transcriptase Reversa de HIV-1
HTS *High Throughput Screening*
hwROC *Harmonic Weighted ROC*
IUPAC *International Union of Pure and Applied Chemistry*
JPS *Joint Pharmacophore Space*
kcal/mol Quilocalorias por mol
kNN *K-Nearest Neighbours*
LBVS *Ligand-Based Virtual Screening*
LMCS *Low-mode Conformational Sampling*
logP Coeficiente de Partição (Logaritmo)
LOO *Leave One Out*
MAPK14 Cinase Protéica Ativada por Mitogênio 14
MDDR *MDL Drug Data Report*
MIF *Molecular Interaction Field*
MLR *Multiple Linear Regression*
MOE *Molecular Operating Environmentmoe*
MS-MM *Multi-Species Multi-Mode*
MUV *Maximum Unbiased Validation*
OA *Optimal Assignment*

OAAP	*Optimal Local Atom Pair Environment Assignment*
OAK	*Optimal-Assignment Kernel*
P38	Cinase Protéica Ativada por Mitogênio 14
PDB	*Protein Data Bank*
PDE5	Fosfodiesterase V
PLS	*Partial Least Squares*
PPAR$_\gamma$	Receptor Ativado por Proliferador de Peroxissomo γ
PPP	*Potential Pharmacophore Points*
QSAR	*Quantitative Structure-Activity Relationship*
REA	Relação Estrutura-Atividade
REF	*Relative Enrichment Factor*
RES	Relação Estrutura-Seletividade
RMSD	*Root Mean Square Distance*
RNA	Ácido Ribonucléico
ROC	*Receiver-Operator Characteristic*
s	Segundo
SAR	*Structure-Activity Relation*
SARA	Síndrome da Angústia Respiratória do Adulto
SOS	*Self-Organizing Superimposition*
SPE	*Stochastic Proximity Embedding*
SQ	Soma dos Quadrados
SVM	*Support Vector Machines*
TBVS	*Target-Based Virtual Screening*
TCG	*TrixX Conformer Generator*
TTD	*Therapeutic Targets Database*
VS	*Virtual Screening*
WOMBAT	*World Of Molecular BioAcTivity*

Sumário

1 Introdução

1.1 Descoberta Racional de Fármacos e a Quimioinformática

A indústria farmacêutica passa por um momento de mudanças. Ao se analisar o período entre 2001 e 2011, vê-se que, apesar do aumento da receita das maiores empresas, a contribuição da pesquisa e desenvolvimento para esta receita diminuiu (1). Um dos grandes motivos para esta queda é a alta taxa de rejeição de novos candidatos a fármacos nos testes clínicos, principalmente na Fase II, onde pelo menos 50% dos candidatos são descartados por falta de eficiência (2). Além disso, a probabilidade de um dado composto que acabou de entrar na fase clínica conseguir chegar ao mercado passou de 10% (em 2001) para apenas 5% (Figura 1.1). Esta queda da taxa de sucesso do desenvolvimento de terapêuticas inovadoras está mudando o foco da indústria, que passa a restringir as áreas de atuação terapêutica e a questionar os prospectos de candidatos com alto risco de rejeição (3). Ainda assim, o investimento destinado ao desenvolvimento de fármacos cresceu, apesar de outros fatores depreciarem os orçamentos das empresas, como a aproximação da expiração de patentes de classes de fármacos lucrativos, problemas relacionados com a determinação de preços de fármacos e uma regulamentação mais rígida para a aprovação de novas entidades moleculares. Neste contexto, as companhias farmacêuticas buscam adotar novas maneiras de aumentar a aprovação e comercialização de compostos através de inovação e não através de iterações sobre os já existentes, através, por exemplo, de reposicionamento de fármacos ou através de pequenas modificações nas estruturas moleculares.

Uma das estratégias para aumentar a probabilidade de um composto candidato a fármaco chegar ao mercado consiste em diminuir o nível de rejeição de candidatos a fármacos nos testes com humanos (4). A rejeição de um composto nas fases finais do fluxo de desenvolvimento acarreta em um grande prejuízo financeiro, já que o custo agregado de descoberta de um novo

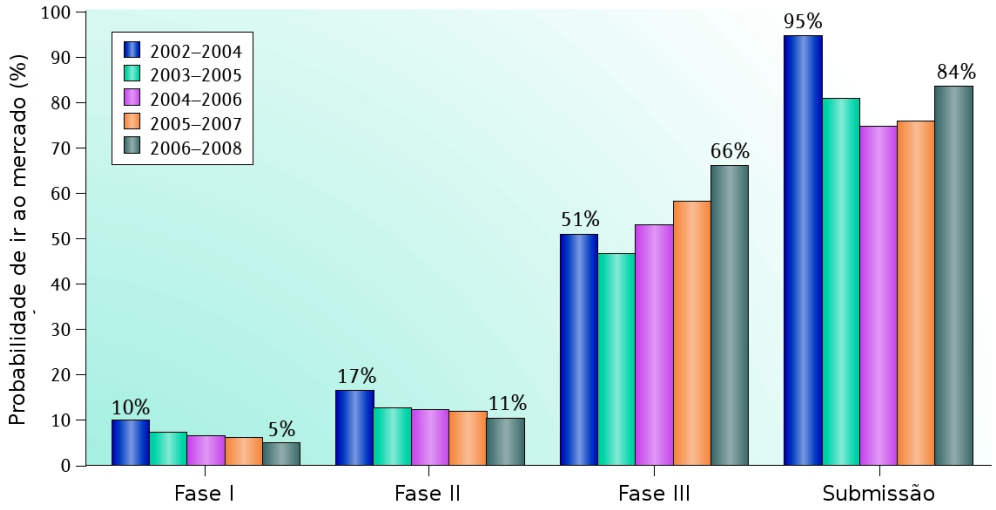


Figura 1.1: Probabilidade de um candidato a fármaco chegar ao mercado dada a fase do desenvolvimento em que o mesmo se encontra. Nota-se a queda constante através dos anos nas Fases I (Toxicologia) e II (Eficiência). Adaptada de (1)

fármaco cresce à medida que passam as etapas do fluxo de desenvolvimento (Figura 1.2). Visto que a ineficiência ao tratar a condição para a qual o composto foi desenvolvido é a maior causa de rejeição de candidatos em fases clínicas, a etapa mais indicada para se realizar um controle rígido de qualidade seria na fase de identificação de moléculas bioativas, onde o custo agregado é ainda baixo (4).

As moléculas bioativas são um dos pontos iniciais de um processo de descoberta racional de novos fármacos. Elas são substâncias que mostram alguma atividade biológica quanto a um alvo de interesse. Caso não haja informação *a priori* de bioatividade, tais moléculas podem ser identificadas através de testes biológicos de larga escala (*High Throughput Screening* - Triagem de alto desempenho). A informação de bioatividade de moléculas ou a informação estrutural dos alvos terapêuticos podem ser usadas em métodos *in silico*, que auxiliam o processo de desenvolvimento racional. Uma vez identificadas, as moléculas bioativas passam por ensaios biológicos confirmatórios de atividade, otimização de características farmacocinéticas (ADMET - Absorção, Distribuição, Metabolismo, Excreção e Toxicidade), otimização de seletividade através de estudos de REA (Relação Estrutura-Atividade, também conhecido pela sua sigla em inglês: SAR - *Structure-Activity Relationship*), para finalmente gerar um ou mais compostos-protótipo. Os

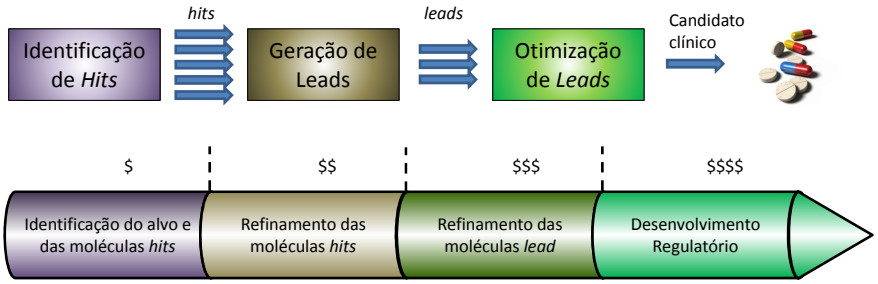


Figura 1.2: As etapas de desenvolvimento e otimização de um novo candidato a fármaco antes das fases clínicas. À medida que se completam as etapas, o número de compostos candidatos diminui, até chegar em uma única substância a ser submetida a testes *in vivo*. Entretanto, o custo agregado de desenvolvimento cresce ao longo do fluxo de trabalho. Adaptada de (4)

compostos-protótipo passam ainda por outro ciclo de estudos de otimização de características ADMET e de REA, antes de serem testados *in vivo* em animais, nos estudos pré-clínicos. Sendo aprovados nestes testes, os compostos passam a ser candidatos a fármaco, sendo admitidos em estudos com humanos, nas fases de estudos clínicos (5). Na primeira fase, o candidato a fármaco é administrado em um pequeno número de humanos sadios, para o estudo dos efeitos de toxicidade e segurança. A Fase II consiste na administração do fármaco experimental em humanos doentes, para avaliar sua eficiência no tratamento das condições patológicas, bem como estudos de dose-resposta. Finalmente, na Fase III, um grupo ainda maior de pacientes é utilizado em um estudo controlado. Os resultados da Fase III devem conter os dados definitivos de eficiência e a comparação com outros tratamentos existentes, caso seja possível, dados que devem ser submetidos às agências reguladoras para aprovação. Somente após a aprovação um fármaco poderia ser comercializado.

Ao focar na validação do alvo biológico e na investigação do modo de atividade da molécula, a probabilidade de se evitar testes clínicos desnecessários diminui, assim como o custo agregado do desenvolvimento. Assim, para aumentar a probabilidade de aprovação de um novo composto, Bleicher et al (4) sugeriram em 2003 adotar uma estratégia de desenvolvimento racional de fármacos com uma validação extensiva do alvo terapêutico e da atividade dos compostos. Dentro deste contexto, a Bioinformática e a Quimioinformática podem atuar em várias das fases que precedem os estudos clínicos. Validação de alvos biológicos, desenvolvimento de bibliotecas de substâncias para experimentos e otimização dos compostos-protótipo são exemplos de tarefas em que um bio/quimioinformata poderia atuar (6). Dentre estas, a identificação de moléculas bioativas ainda é um dos maiores desafios no processo de descoberta de fármacos. O HTS é o padrão da indústria para identificação de compostos bioativos em larga escala. A tecnologia atual permite realizar milhões de experimentos *in vitro* em relativamente pouco tempo (7). Mas a taxa de identificação de moléculas bioativas é extremamente baixa, o que gera um aumento significativo no custo final de desenvolvimento de um novo fármaco. É nesse ponto que ferramentas de triagem virtual podem atuar, ao selecionar dentro de um banco de moléculas aquelas com maior probabilidade de mostrar algum tipo de atividade biológica contra um alvo de interesse (8).

1.2 Triagem Virtual

O processo de triagem *in silico* de bibliotecas de compostos para uma certa atividade é conhecida como triagem virtual (VS, do inglês *Virtual Screening*). A triagem virtual é geralmente definida como um processo em que grandes bibliotecas de compostos são automaticamente avaliadas usando técnicas computacionais (9). Seu objetivo é descobrir moléculas potencialmente bioativas em grandes bancos de dados de compostos químicos (geralmente ligantes para alvos biológicos) e remover moléculas identificadas como tóxicas ou que possuam propriedades farmacodinâmicas e farmacocinéticas desfavoráveis (10) (Figura 1.3).

De acordo com os dados usados como referência, as técnicas de triagem virtual podem se dividir em duas grandes categorias. As técnicas que incorporam dados de estrutura de alvos terapêuticos na sua abordagem são conhecidas como técnicas de Triagem Virtual baseadas em alvos biológicos (TBVS - *Target-Based Virtual Screening*). Apesar de métodos computacionais serem usados em processos de descoberta de fármacos desde a década de 70, a triagem virtual começou a se tornar popular na indústria somente a partir da década de 90, impulsionada pelo aumento da disponibilidade de estruturas de raios-X de alvos biológicos, o que também passou a popularizar aplicações computacionais baseadas em estruturas de alvos, denominadas ancoragem proteína-ligante ou *Docking* molecular automático (11). O *Docking* (12–14) consiste em procedimentos computacionais que predizem o modo de ligação e a afinidade de um ligante com um alvo receptor, e suas origens remotam à década de 80 (15). Cada ligante de um banco de dados é ancorado a um sítio ativo de um receptor, e os algoritmos fazem uma busca extensiva das possibilidades de conformação e orientação do ligante dentro do sítio. Cada possibilidade é avaliada quantitativamente por uma função de escore, e os ligantes podem ser ordenados através desta afinidade com o receptor (16). Outras técnicas TBVS incluem Projetos *De Novo* (17–19), que envolve o projeto de inibidores usando apenas a informação do sítio ativo do alvo terapêutico como referência inicial; e Projetos de Fármaco baseado em Fragmentos (20–22), que usa moléculas pequenas, com massa molécular até 300 Da, para mapear pontos de interação no sítio ativo e servirem como “peças de um quebra-cabeça” para a montagem de um novo inibidor.

Apesar do aumento contínuo da disponibilidade de estruturas de alvos biológicos, a quantidade de dados disponíveis sobre a atividade biológica de moléculas pequenas ainda é muito maior. Uma rápida consulta às bases de dados públicas mais populares mostra a diferença de

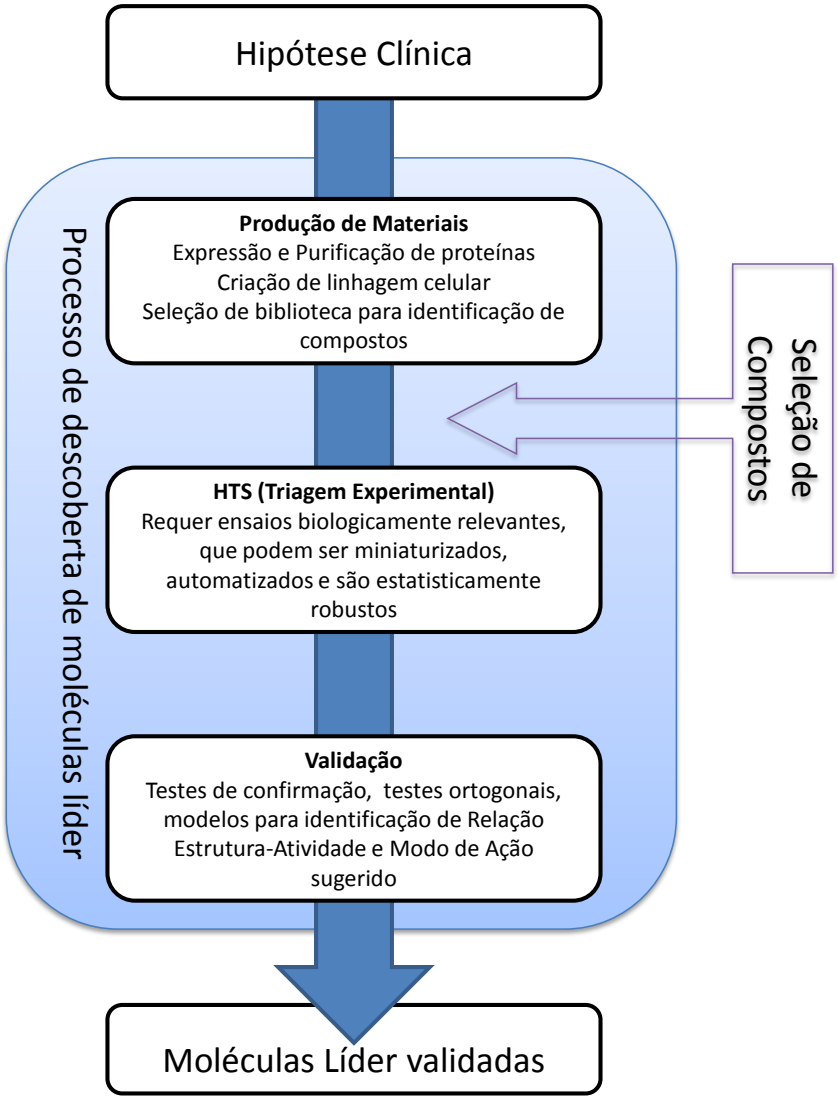


Figura 1.3: O papel da triagem virtual para a identificação de moléculas bioativas no fluxo de desenvolvimento de compostos-protótipo. Adaptada de (6)

grandeza entre as duas categorias. O PDB (23) (`www.pdb.org`) é a base de dados pública onde estruturas cristalográficas tridimensionais são depositadas desde a década de 70. O banco de dados possuia aproximadamente 79.500 estruturas depositadas até o dia 28 de Fevereiro de 2012. Já para os compostos, o PubChem (24) (`http://pubchem.ncbi.nlm.nih.gov/`) é a maior base de dados pública, já que agrega informações de outros bancos de dados menores e mais específicos. Na mesma data, o PubChem possuia cerca de 32,3 milhões de compostos diferentes. Destes, 1,7 milhão possuem dados de ensaios biológicos e 11.307 possuem dados anotados de ação farmacológica. Além do PubChem, existem outras bases públicas de compostos como o ChEMBL (25) (aproximadamente 1,14 milhão de compostos únicos e 5,4 milhões de medições de bioatividade) e o ChemSpider (`www.chemspider.com`, com aproximadamente 26 milhões de compostos), além de bases comerciais como o MDDR (*MDL Drug Data Report*), construído a partir de dados de patentes, e o *World Of Molecular BioAcTivity* (WOMBAT), da Sunset Molecular (26). Tal disparidade na disponibilidade de informação, que encontrou nas técnicas de HTS desenvolvidas a partir da década de 90 um grande fomentador (11), é um dos atrativos para as técnicas de Triagem Virtual baseadas em Ligante (LBVS - *Ligant-Based Virtual Screening*), que usam somente dados de ligantes ativos e pequenas moléculas. De acordo com Ripphausen (27), existem aproximadamente três vezes mais técnicas TBVS do que LBVS disponíveis, mas, em estudos prospectivos, as técnicas baseadas em ligante identificam moléculas bioativas em média mais potentes que as identificadas pelas abordagens baseadas em alvos biológicos.

1.2.1 Classificação de Técnicas de Triagem Virtual Baseadas em Ligante

Como visto anteriormente, as técnicas em Triagem Virtual se dividem em duas categorias de acordo com a natureza dos dados de referência: LBVS e TBVS. As técnicas em Triagem Virtual baseadas em ligante ainda podem se subdividir de acordo com a dimensão da representação dos dados utilizada (1D, 2D ou 3D, como mostra a Figura 1.4). Descritores 1D são números que representam diferentes propriedades moleculares do composto como um todo, como a massa molecular, número de heteroátomos ou ligações rotacionáveis; ou ainda parâmetros fisico-químicos computados, como logP. Descritores 2D podem existir em duas formas: índices topológicos ou descritores estruturais. Um índice topológico é um descritor composto por um único número que tipicamente caracteriza a estrutura de acordo com sua forma e tamanho. Os índices mais simples caracterizam moléculas de acordo com seu tamanho, forma e grau de ramificação, en-

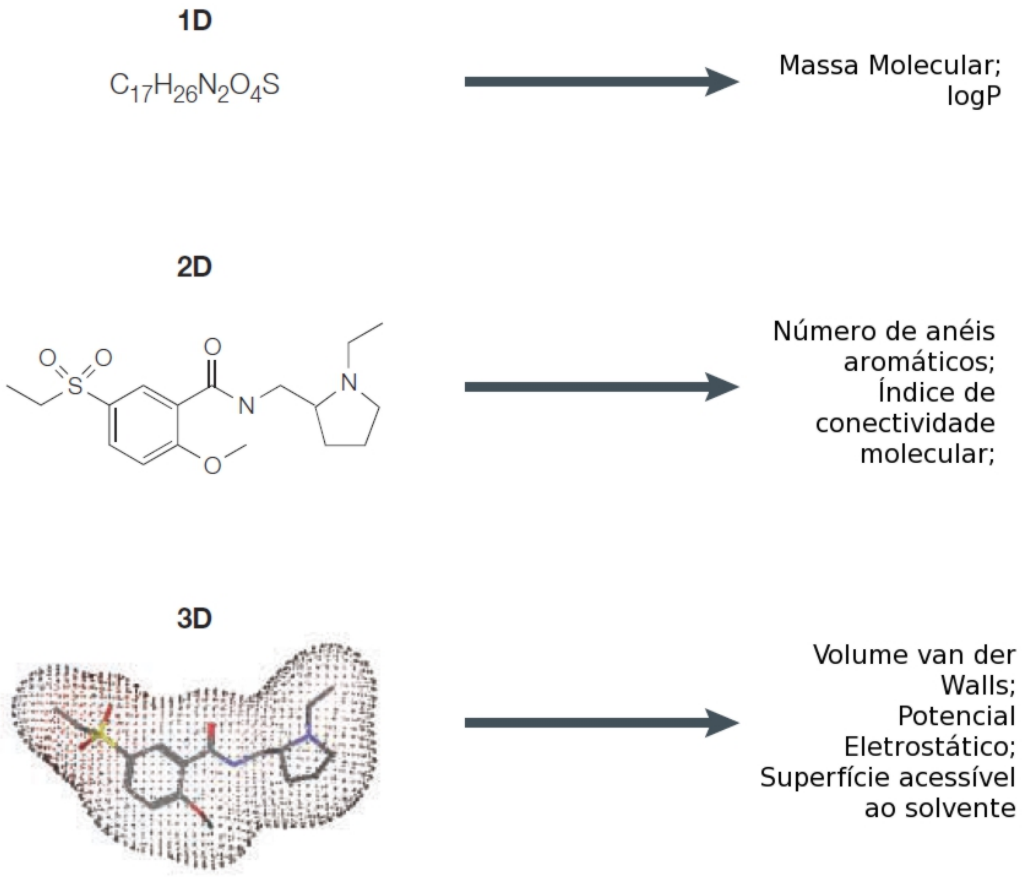


Figura 1.4: Classificação dos descritores usados para representar compostos em métodos de triagem virtual de acordo com o número de dimensões da estrutura molecular. Adaptada de (8)

quanto índices mais complexos levam em consideração tanto as propriedades dos átomos quanto suas conectividades. Já os descritores estruturais caracterizam a molécula pela suas sub-estrutura química, e podem ser encontrados em forma de grafos químicos 2D ou vetores binários que codificam os fragmentos presentes (28). Descritores 3D possuem uma maior complexidade inerente, pois precisam considerar as conformações das moléculas. Dentre suas aplicações, destacam-se o uso de distâncias inter-atômicas, farmacóforos 3D, superfícies moleculares, campos eletrostáticos, dentre outros. Vários estudos comparativos revelam que não há um único descritor que tenha desempenho melhor do que outros descritores em qualquer aplicação de triagem virtual (29). Logo, moléculas são normalmente representadas por um conjunto de descritores (28).

A diversidade conceitual e algorítmica das técnicas de LBVS permite uma terceira categorização

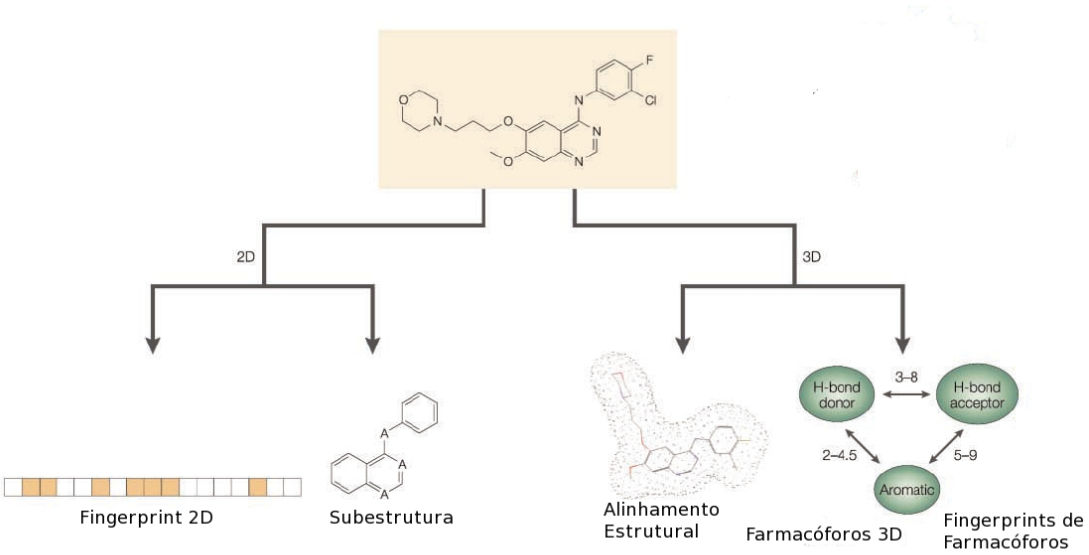


Figura 1.5: Diferentes representações de uma molécula que podem ser usadas por algoritmos baseados em similaridade. Adaptada de (8)

de acordo com a abordagem do procedimento. Aqui se diferenciam os algoritmos baseados em similaridade, algoritmos quantitativos e os baseados em aprendizado de máquina.

Algoritmos baseados em busca por similaridade

Busca por similaridade (Figura 1.5) é um tipo particular de triagem virtual em que as moléculas cujas estruturas sejam mais semelhantes a uma ou mais estruturas de substâncias ativas, denominadas compostos de referência, teriam maior probabilidade de compartilhar a mesma atividade biológica (28). Em uma busca por similaridade, uma ou mais estruturas são usadas como referência. No contexto de descoberta de fármacos, tais estruturas normalmente possuem um nível de atividade biológica de interesse. Cada estrutura do banco de dados é representada usando o mesmo esquema das estruturas de referência, de modo a compará-las para avaliar o nível de semelhança estrutural. A similaridade é quantificada por um coeficiente calculado para todas as moléculas da base de dados que está sendo usada na busca.

Esta abordagem para acesso a bancos de dados químicos foi primeiramente descrita por Carhart et al. (30) em 1985 e por Willett et al. (31) no ano seguinte e sua base lógica é deri-

vada de um paradigma da química medicinal, que diz que moléculas que apresentam estruturas similares deveriam compartilhar atividades biológicas similares (28). Tal hipótese atribuída originalmente a Johnson e Maggiora (32), tem sido notada em uma série de observações e continua sendo aceita pela comunidade de química medicinal (33). Entretanto, o processo de otimização de moléculas bioativas e de compostos-protótipo pode se beneficiar da situação contrária: compostos ativos sujeitos a pequenas modificações químicas podem apresentar uma significativa diferença de potência e/ou seletividade. Estas "descontinuidades" nos relacionamentos estrutura-atividade são conhecidas como "fendas de atividade" (*activity cliffs*) (34, 35).

A técnica em VS baseada em similaridade 2D mais intuitiva e direta é baseada em subestruturas (ou fragmentos moleculares) (36–39). Indo além da bidimensionalidade, buscas em bancos de dados de compostos também podem ser realizadas através de representações moleculares tridimensionais, particularmente através de modelos farmacofóricos (40–43), superposição de volumes (44, 45), campos de interação molecular (MIFs - *Molecular Interaction Fields*) (46), bem como combinações destas abordagens em metodologias híbridas (47–52). Estes métodos geralmente requerem uma geração eficiente de conformações de baixa energia (estáveis) ou que se aproximem do arranjo bioativo para o alinhamento tridimensional das estruturas (8).

Algoritmos Quantitativos

Uma análise QSAR (*Quantitative Structure-Activity Relationship*, Relação Estrutura-Atividade Quantitativa) pode ser definida como a aplicação de métodos matemáticos e estatísticos ao problema de encontrar equações empíricas na forma $Y_i = F(X_1, X_2, \ldots, X_n)$, onde a variável Y_i é a atividade biológica de moléculas, e $X_1, X_2, \ldots, X_n$ são propriedades estruturais (descritores moleculares) experimentais ou calculadas de compostos (53). Cada composto pode ser representado por um ponto em um espaço n-dimensional, onde cada X_i é uma coordenada independente da substância. O objetivo de um modelo QSAR é estabelecer uma tendência dos valores dos descritores moleculares que correlacione com os valores de atividade biológica. Tais modelos podem ser aplicados em várias etapas de um projeto racional de fármacos, desde a triagem virtual até a otimização de características ADMET de compostos-protótipo (54). Para tanto, os modelos devem passar por uma validação estatística rigorosa para que possam ser considerados modelos robustos e preditivos (55).

De acordo com os descritores moleculares utilizados nos cálculos, métodos de análise QSAR podem ser divididos em três grupos (53). O primeiro é baseado em um número relativamente pequeno de propriedades físico-químicas e parâmetros que descrevem efeitos hidrofóbicos, estéricos e eletrostáticos, dentre outros. Estes descritores são usados como variáveis independentes em abordagens de regressão múltipla e são conhecidos como análise Hansch (56, 57). O segundo grupo abrange métodos baseados em características quantitativas da estrutura molecular (descritores topológicos) como índices de conectividade molecular (58–60), índices topológicos (61), eletrotopológicos (62–65), descritores de pares de átomos (66, 67), etc. Alguns descritores topológicos ainda podem ser combinados com descritores físico-químicos. Como a fórmula estrutural dos compostos é bidimensional, estes métodos são conhecidos como QSAR bidimensional (2D). Diferentes métodos de correlação são usados em estudos de QSAR 2D, sejam eles lineares, como a Regressão Linear Múltipla (MLR - *Multiple Linear Regression*) com seleção de variáveis (68) ou Mínimos Quadrados Parciais (PLS - *Partial Least Squares*) (69), ou não-lineares, como kNN (*k-Nearest Neighbours*) (70, 71) ou Redes Neuronais Artificiais (ANN - *Artificial Neural Networks*) (72). O terceiro grupo compreende métodos que utilizam descritores baseados em representação espaciais (tridimensionais) de estruturas moleculares, os métodos de QSAR 3D. A maioria dos métodos de QSAR 3D requerem um alinhamento tridimensional das moléculas de acordo com um modelo farmacofórico ou baseado em ancoragem molecular no sítio do receptor. Na metodologia CoMFA (*Comparative Molecular Field Analysis* - Análise Comparativa de Campos Moleculares) (73, 74), os descritores são calculados a partir de campos de força moleculares, utilizando sondas para calcular a energia eletrostática e estereoquímica posicionadas em uma grade de pontos que envolve a estrutura molecular (53).

Algoritmos baseados em técnicas de Aprendizado de Máquina

Os métodos baseados em busca por similaridade possuem uma tendência a necessitar de pouca informação de referência para criar modelos preditivos. Muitos deles precisam apenas de uma única estrutura a ser comparada, o que fez com que fossem amplamente adotados (28). Mas, a medida que a informação sobre novas substâncias e suas atividades biológicas começaram a se tornar publicamente disponíveis, métodos capazes de inferir modelos preditivos a partir de uma grande quantidade de dados moleculares de compostos ativos e inativos puderam ser aplicados. Tais métodos são classificados como métodos de aprendizado de máquina (75).

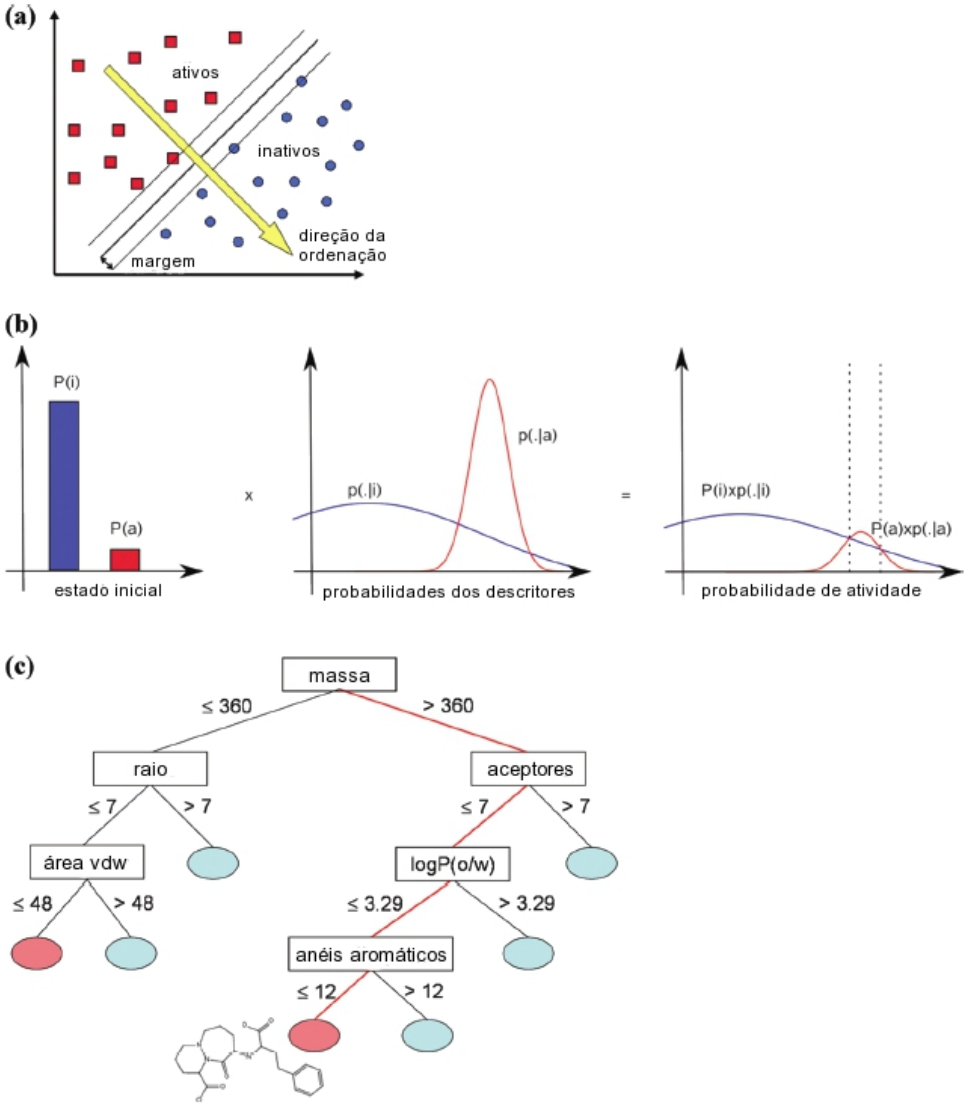


Figura 1.6: Principais abordagens em triagem virtual baseadas em aprendizado de máquina: **a)** Esquematização do funcionamento de um algoritmo baseado em SVM. As margens do hiperplano (aqui representado em duas dimensões) otimiza a separação entre ativos e inativos. **b)** Métodos Bayesianos: o Teorema de Bayes é usado para obter a probabilidade de atividade condicionadas aos descritores **c)** Árvores de Decisão. Folhas vermelhas identificam ativos, ao contrário de folhas azuis, que identificam inativos. O caminho destacado em vermelho foi calculado para a molécula exemplo. Adaptada de (75)

Os métodos baseados em aprendizado de máquina comumente usados para Triagem Virtual podem se dividir em três abordagens (Figura 1.6). Técnicas baseadas em SVM (*Support Vector Machines*) (76, 77) trabalham com a idéia de construir um hiperplano em um espaço *n*-dimensional que separe as duas classes de compostos (ativos e inativos) (Figura 1.6 a). A metodologia do SVM tem uma forte fundação teórica na Teoria de Aprendizado Estatístico (78) por considerar dois objetivos geralmente contraditórios em aprendizado de máquina: representar bem os dados de treinamento e ser suficientemente genérico para classificar corretamente o conjunto externo de teste. Isto é, quanto mais complexa é a estrutura de um modelo a ser treinado, melhor ele se encaixa nos dados treino, mas menor é a probabilidade do modelo conseguir classificar corretamente dados previamente desconhecidos, o que é conhecido como *overfitting*. Ao contrário das outras técnicas de aprendizado de máquina, o SVM consegue controlar este risco de *overfitting* estrutural, porque constrói um hiperplano que maximiza a distância entre este e os dados treino mais próximos (75).

As técnicas Bayesianas são assim denominadas por se basearem no Teorema de Bayes, base da Estatística Bayesiana. Apesar de diferirem algoritimicamente, as técnicas Bayesianas para Triagem Virtual acabam por derivar a probabilidade de um dado composto ser ativo para um alvo ou atividade biológica de interesse (Figura 1.6 b). Devido ao fato dos classificadores Bayesianos tipicamente gerarem estimativas numéricas para probabilidades de atividade, eles são capazes de ordenar compostos em bases de dados de acordo com estas probabilidades. Métodos Bayesianos geralmente baseam-se na estimação de distribuições de probabilidade de representações numéricas de substâncias, representações estas baseadas em descritores de propriedades moleculares. Os métodos mais usados incluem classificadores Bayesianos *naïve* (79, 80) e discriminação binária de *kernels* (81, 82), os quais podem ser aplicados para vetores binários ou de frequência (75).

Finalmente, Árvores de Decisão (Figura 1.6 c) são modelos preditivos que mapeam observações sobre um item em conclusões sobre o valor alvo do item. Dependendo da natureza deste valor alvo, as árvores de decisão podem ser especificamente denominadas como Árvores de Classificação (caso o valor alvo seja uma classe, como ativo ou inativo) ou Árvores de Regressão (caso o valor alvo seja um valor contínuo, como nível de atividade biológica). As folhas de uma Árvore de Descisão correspondem a uma classificação do item, enquanto seus ramos representam o conjunto de características que definem tal classificação (83). Estes classificadores ainda são muito

usados em aplicações de Quimioinformática, especialmente na forma de florestas aleatórias, com o intuito de melhorar a taxa de classificação (75). Aplicações de Árvores de Decisão abrangem análises de QSAR (84), predição de solubilidade em água de compostos (85) e Triagem Virtual (86).

1.2.2 Critérios de Avaliação de Desempenho em Triagem Virtual

A triagem virtual consiste em técnicas e modelos preditivos que tentam classificar compostos de acordo com sua atividade biológica contra certo alvo terapêutico. Existem duas variáveis centrais a qualquer método de avaliação de desempenho preditivo: a Cobertura e a Precisão (28). Considere que uma lista de indivíduos, cuja classificação verdadeira é conhecida *a priori*, está ordenada por um critério determinante para sua classificação. Por exemplo, uma lista de moléculas ordenadas pela similaridade a um modelo de referência que tenta predizer sua bioatividade. Ao selecionar-se um subconjunto do topo desta lista, digamos 1%, a precisão seria a razão entre o número de verdadeiros positivos e o número de elementos na seleção. Já a cobertura seria a razão entre o número de verdadeiros positivos e o número total de positivos em todo o conjunto de dados. Logo o problema de uma busca se torna a maximização destas duas variáveis. Em uma seleção ideal, todos os verdadeiros positivos estariam no subconjunto selecionado.

Existem diversas métricas derivadas destas duas variáveis que podem ser aplicadas ao mensurar o desempenho de sistemas de triagem virtual de compostos. A mais intuitiva, e talvez por isso amplamente aplicada, é o *Fator de Enriquecimento* (EF - *Enrichment Factor*), que consiste no número de moléculas ativas (a) recuperadas até um certo ponto de corte ($\chi\%$, o que corresponde a n compostos) em relação ao número de moléculas ativas que seriam recuperadas por uma busca aleatória (A/N, onde A é o número total de moléculas ativas e N é o número total de moléculas), ou seja:

$$\mathrm{EF}_{\chi\%} = \frac{a/n}{A/N}$$

Apesar da simplicidade, o fator de enriquecimento tem uma série de problemas, pois depende do número de verdadeiros positivos e verdadeiros negativos, além de depender do valor de corte, o que o torna mais uma medida de desempenho do experimento ao invés de medir o desempenho do método (87). Korff et al. (88) propuseram em 2009 uma normalização ao fator de enriquecimento para eliminar a dependência do número de estruturas ativas, o chamado Fator

de Enriquecimento relativo (REF - *Relative Enrichment Factor*):

$$\mathrm{REF}_{\chi\%} = \frac{100a}{\min(n,A)}$$

.

Uma métrica muito usada em vários campos de pesquisa cujos problemas envolvem recuperar um pequeno conjunto de "positivos" de um grande conjunto de "negativos" é a curva ROC (*Receiver-Operator Characteristic*) e a área sob a curva ROC ($\mathrm{AUC_{ROC}}$ - *Area Under Curve*). A $\mathrm{AUC_{ROC}}$ tem a grande vantagem de não depender do número de ativos e de inativos, ou da razão entre eles (87). A curva ROC é uma linha que representa a fração de verdadeiros positivos (valor no eixo y) em sucessivos pontos de corte, com a fração dos falsos positivos (valor no eixo x) nos mesmos pontos. A área sobre a curva ROC ($\mathrm{AUC_{ROC}}$) representa a probabilidade de um composto ativo escolhido aleatoriamente estar melhor colocado na lista ordenada do que um composto inativo também escolhido aleatoriamente.

Entretanto, a curva ROC é criticada ao ser usada para mensurar o desempenho de métodos em triagem virtual, devido ao fato de não dar a devida atenção aos compostos melhor colocados na lista ordenada, os quais seriam considerados para experimentos *in vitro* (89), o chamado "reconhecimento precoce" (*early recognition*). Pensando neste quesito, Truchon e Bayly (89) desenvolveram a métrica BEDROC_α (*Boltzmann-Enhanced Discrimination of ROC*), que usa uma ponderação exponencial, regulada pelo parâmetro α, para dar maior valor aos verdadeiros positivos nas primeiras posições da lista ordenada. A BEDROC_α não tem as mesmas limitações quanto ao número de moléculas e a razão entre ativos e inativos, característico do fator de enriquecimento, além de ser limitada entre zero e um, assim como a $\mathrm{AUC_{ROC}}$. Entretanto, Nicholls (90) argumenta que há uma grande correlação entre $\mathrm{AUC_{ROC}}$ e BEDROC, considerando o perfil das aplicações de triagem virtual. De acordo com o autor, a curva ROC é uma métrica suficiente para medidas de desempenho, além de não requerer um parâmetro livre, como é o caso da BEDROC. Além disso, Clark e Webster-Clark (91) argumentam que a presença do parâmetro α no BEDROC é um fator de complicação a mais na análise de desempenho e, por isso, propuseram a pROC, uma tranformação logarítmica da curva ROC que busca dar maior ênfase ao primeiros valores da curva.

Uma terceira característica esperada de sistemas de triagem virtual é a capacidade de amos-

trar a diversidade estrutural (*scaffold hopping*). O objetivo da identificação de diferentes estruturas é conseguir substituir a estrutura química central de um composto bioativo por um outro padrão molecular e ainda manter o nível de atividade biológica da molécula (29), o que é considerado um desafio na área, já que séries de ativos em bases de dados costumam sofrer o que Good e Oprea chamam de "viés do análogo" (*analogue bias*) (92). Devido a fatores intrínsicos à própria descoberta de fármacos (por exemplo, detalhes de patentes que tentam cobrir uma classe de compostos ativos, as chamadas patentes guarda-chuva - *umbrella patents*), séries de ativos tendem a favorecer um ou outro motivo estrutural, o que acaba por enviesar os resultados em direção às classes estruturais mais populosas. Neste caso, métodos que priorizam os compostos de classes mais populosas e relegam classes menores tendem a "se sair melhor" que métodos que identificam alguns ativos de cada classe uniformemente. Este viés é ainda mais punitivo quando considera-se que, na prática, o tamanho da classe é inversamente proporcional à oportunidade de inovação que ela representa (91).

Ao tentar normalizar a contribuição de cada classe estrutural para o desempenho geral do método, pode-se contar o número de classes encontradas até um certo ponto da busca, ao invés de se considerar simplesmente o número de ativos, de maneira análoga ao fator de enriquecimento. Mas tal métrica sofre dos mesmos contratempos do EF mencionados anteriormente. Clark e Webster-Clark (91) propuseram dois esquemas de ponderação da curva ROC: a ponderação aritmética (awROC - *arithmetic weighted ROC*) e a ponderação harmônica (hwROC - *harmonic weighted ROC*), sendo que a primeira tem sido mais usada para avaliar a capacidade de diversificação estrutural de métodos em triagem virtual. Na curva ROC tradicional, cada verdadeiro positivo contribui de maneira uniforme no valor da AUC, independente de classe estrutural. Na curva awROC, a contribuição de cada verdadeiro positivo é inversamente proporcional ao número de moléculas da classe estrutural ao qual o composto pertence. Na curva hwROC, a contribuição do verdadeiro positivo é inversamente proporcional à sua colocação na lista ordenada dentro da classe a que pertence.

1.3 Farmacóforos

O termo *Farmacóforo* tem sido usado na Química Medicinal a muitos anos (42). Ele foi definido pela IUPAC (*International Union of Pure and Applied Chemistry* - União Internacional

da Química Pura e Aplicada) como “um arranjo de características estéricas e eletrônicas que é necessário para assegurar as interações supramoleculares ótimas com a estrutura de um alvo biológico específico, e para ativar (ou bloquear) a resposta biológica deste alvo. Entretanto, um farmacóforo não representa uma mólecula real ou uma associação real de grupos funcionais, mas um conceito puramente abstrato que descreve as capacidades de interações moleculares comuns a um grupo de compostos para um mesmo alvo. O farmacóforo pode ser considerado como o maior denominador comum compartilhado por um conjunto de moléculas bioativas.” (93)

1.3.1 Conceito e Perspectiva Histórica

Existem duas vertentes sobre a origem do conceito de farmacóforos. A primeira, baseada no trabalho de Ariëns (94) e propagada por Gund (95), considera Paul Ehrlich o criador do conceito de farmacóforo. Esta vertente é a mais popular, presente em livros especializados (40, 41) e em várias publicações. Mas em 2007, um artigo publicado por John H. Van Drie, contestou esta linha: de acordo com o autor, foi Lemont Kier, numa série de artigos publicados entre 1967-1971, quem cunhou o conceito moderno de farmacóforo (96). Esta segunda versão, por ser melhor documentada, vem sendo gradualmente aceita na comunidade e já consta em algumas publicações recentes (por exemplo, no trabalho de Caporuscio e Tafi (97)). A seção seguinte contém a evolução do conceito de farmacóforo segundo Ariëns, seguida da versão de Van Drier.

Paul Ehrlich

A história dos farmacóforos se confunde com a própria história da busca de curas para doenças e enfermidades. Quando Louis Pasteur demostrou no século XIX que muitas doenças eram causadas por micróbios e parasitas, pesquisadores da época reconheceram que substâncias poderiam ser testadas nestes microrganismos, identificando os compostos que muito provavelmente ajudariam no tratamento das respectivas doenças. Na mesma época, químicos orgânicos começaram a estudar a síntese de corantes e pigmentos, desde que várias tinturas feitas por Perkins e outros viraram um grande sucesso comercial. Estes pesquisadores descobriram que era frequentemente possível manter o esqueleto molecular destas tinturas intacto e variar apenas as estruturas conectadas para obter uma gama de cores e propriedades relacionadas. A parte da molécula essencial para a definição da cor era chamada *cromóforo*.

No final do século XIX, Paul Ehrlich descobriu que diferentes tinturas coloriam diferentes tecidos humanos. Assim descobriu os granulócitos eosinófilos e basófilos (leucócitos nomeados pela afinidade às tinturas eosina e básica). Descobriu tinturas que coloriam especificamente o bacilo *Mycobacterium tuberculosis* e os plasmódios causadores de malária, demonstrando ainda que esta tintura (Azul de metileno) poderia tratar a malária. Mais tarde, sua pesquisa o levou a realizar o primeiro estudo em larga escala de síntese e teste de agentes quimioterapêuticos, dos quais o 606º composto era um arsênico que mostrou ser muito eficiente no tratamento de sífilis, o qual ele denominou *Salvartan*. Ele ainda criou o termo *quimioterapia* para descrever o uso de tais agentes sintéticos no tratamento de doenças.

Ehrlich ainda estudou o efeito de toxinas, sugerindo que tais compostos possuíam grupos *haptofóricos*, que os permitiam permanecer ligados à celula, e grupos *toxofóricos* separados, que causavam o efeito tóxico. Inicialmente, Ehrlich não aplicava esta idéia para fármacos, porque muitos pareciam não combinar irreversivelmente com a célula, como fazem as toxinas. Mas logo ele começou a argumentar que células possuíam *quimiorreceptores* que deveriam se ligar aos fármacos antes destas iniciarem seu efeito terapêutico. Então Ehrlich supostamente teria criado o termo *farmacóforo* para descrever os grupos moleculares de um fármaco que são essenciais para a definição da atividade biológica. Ele ainda assumiu que os quimiorreceptores estavam envolvidos nos processos fisiológicos, e logo existiam em todas as células. Portanto a busca pela chamada "*bala mágica*", que mataria o agente patológico sem afetar o hospedeiro, estava fadada ao fracasso; o melhor que os pesquisadores poderiam esperar era maximizar a eficácia e minimizar a toxicidade (40).

Lemont Kier

Segundo Van Drie, Ehrlich não foi quem criou o termo "farmacóforo" (96). De acordo com o autor, não se vê na publicação original de Ehrlich (98) ou mesmo nos seus trabalhos seguintes a palavra *pharmakophor*. A fonte de tal equívoco foi um artigo de Peter Gund (95), mas o próprio admite não ter investigado a fonte original e ter seguido apenas a citação errônea do livro de Ariëns (94). Ainda de acordo com o autor, o verdadeiro criador do conceito de farmacóforo foi Lemont Kier, em uma série de publicações entre 1967 e 1971, onde estudava agonistas muscarínicos. Kier (99) fez o primeiro cálculo de um modelo de farmacóforo que se tem registro, o denominando *proposta de padrão de receptor*. Entretanto, em 1971 (100), ele publicou a mesma

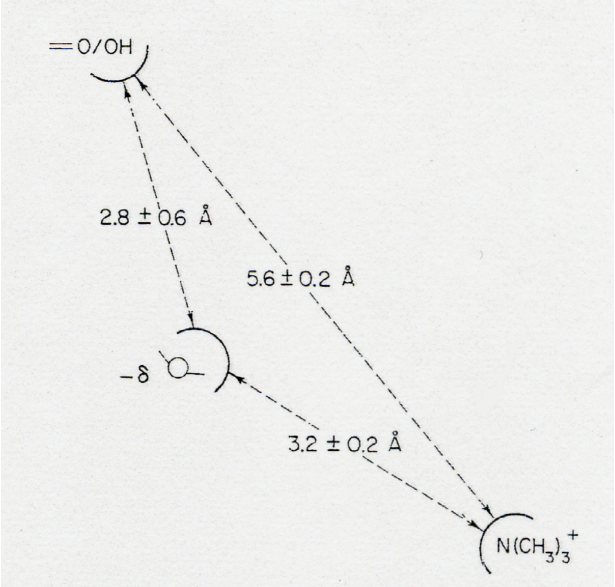


Figura 1.7: O Farmacóforo Muscarínico, segundo Kier (99, 100) - o primeiro modelo farmacofórico proposto.

figura, renomeando-a *farmacóforo muscarínico* (Figura 1.7). Estes trabalhos pioneiros inspiraram uma série de outras publicações que contribuiram não só para a solidificação do conceito de farmacóforo, mas para as primeiras ferramentas de busca e de triagem virtual. O mesmo Peter Gund citado anteriormente foi o autor do primeiro programa capaz de realizar uma busca em uma base de dados moleculares usando um farmacóforo como entrada, o MOLPAT (95).

Farmacóforos e a Seletividade

O conceito de alvos e receptores específicos, apesar de ser amplamente aceito hoje, não era muito claro até o começo do século XX. John Langley, ao observar como a nicotina e o curare pareciam interagir e competir pela mesma substância envolvida em processos de resposta muscular, denominou-a "*substância receptiva*" e especulou que muitos fármacos e venenos agiriam de maneira similar. Mas a palavra *receptor* só foi introduzida por Ehrlich alguns anos depois, nomeando o fator que explicaria a potência da ação dos fármacos em pequenas concentrações, além de sua alta especificidade química (a ponto de isômeros terem diferentes efeitos farmacológicos) e biológica (*e.g.* a adrenalina tem um poderoso efeito sobre o músculo cardíaco, mas não o mantém em músculos estriados).

Com o desenvolvimento dos conceitos sobre estruturas moleculares, percebeu-se que somente a presença dos farmacóforos simultaneamente no composto e no alvo não seria suficiente para a atividade biológica. Seria preciso então um arranjo geométrico tridimensional aceitável pelo receptor. A confiança na teoria de fármaco-alvo cresceu ainda mais quando as primeiras estruturas de enzimas cristalizadas foram resolvidas, mostrando alguns compostos unidos à cadeia. Hoje, os farmacóforos são um paradigma que sumarizam os efeitos da estrutura química na atividade biológica, permitindo um maior nível de abstração (40, 41).

1.3.2 Farmacóforos em Quimioinformática

Um modelo farmacofórico satisfatório deve primeiramente destacar os grupos funcionais envolvidos na interação com o alvo, a natureza das interações não-covalentes e as diferentes distâncias entre as cargas (101). O modelo também deve mostrar algum poder preditivo e apontar possibilidades de novos compostos, mais potentes e estruturalmente diferentes dos já conhecidos. Entretanto, a simplicidade das representações farmacofóricas também significa que, inevitavelmente, as representações não conseguem explicar a natureza completa das interações farmacológicas. Portanto, é preciso saber os limites do poder do modelo farmacofórico para aplicá-lo com sucesso em processos de triagem virtual e desenvolvimento racional de fármacos (42, 102). Assim, é importante ressaltar que farmacóforos não representam as reais interações entre o ligante e a macromolécula, mas são apenas abstrações e simplificações das interações possíveis entre um ligante putativo e o sítio do alvo.

Modelos farmacofóricos são usados em várias etapas de um projeto racional de fármacos, desde a identificação de moléculas bioativas e compostos-protótipo, passando pela racionalização dos dados de REA (Relação Estrutura-Atividade) e RES (Relação Estrutura-Seletividade), otimização de compostos-protótipo, predição de sítios de metabolismo, interações medicamentosas, efeitos colaterais e toxicidade (97). Os modelos são usualmente construídos a partir de um conjunto de moléculas ativas, mas podem usar a estrutura do alvo biológico como única referência ou ainda como informação adicional para melhor posicionar os farmacóforos do modelo. A construção de um modelo farmacofórico passa por três etapas básicas: caracterização das interações, alinhamento conformacional e geração das hipóteses (103).

O primeiro passo ao construir um modelo de farmacóforos é a classificação dos átomos de

acordo com sua natureza e o ambiente químico à sua volta em categorias pré-definidas associadas a um comportamento de interação específico (104). Usualmente, a caracterização farmacofórica não vai além de uma classificação fisico-química simplista dos átomos / grupos funcionais, classificando-os entre "hidrofóbicos" (que podem incluir ou não os anéis aromáticos, já que estes podem ser classificados separadamente), "polares positivos" (que podem ser subdivididos entre doadores de hidrogênio e cátions) e "polares negativos" (aceptores de hidrogênio e ânions). Alguns grupamentos químicos permitem que dois tipos diferentes de farmacóforo possam ser atribuídos ao mesmo átomo/fragmento, como por exemplo, hidroxilas alcoólicas ou aminas, que podem participar como doadores ou receptores de ligações de hidrogênio.

O segundo passo, alinhamento das moléculas, requer um algoritmo eficiente de amostragem conformacional. Esta etapa é crítica para a construção de um bom modelo farmacofórico, já que uma amostragem deficiente pode diminuir seu poder de predição (105–107). Além disso, ao usar o modelo para triagem virtual, o banco de dados de compostos a ser buscado também deve passar pelo mesmo processo de amostragem conformacional, o que pode aumentar proibitivamente o tempo de processamento.

No terceiro passo, a construção da hipótese farmacofórica, um conjunto de regiões no espaço tridimensional (geralmente esféricas) é delimitado. Estas regiões supostamente deveriam abrigar os grupos funcionais do tipo especificado, baseados em um conjunto alinhado de moléculas ativas, de um estudo do sítio ativo ou uma mistura de ambos (104) (Figura 1.8). Tais regiões funcionam como restrições geométricas e de afinidade físico-químicas para o alinhamento de moléculas, e deveriam representar um mapa de interações favoráveis com o sítio de atividade. Moléculas que possuam ao menos uma conformação que satisfaça tais restrições de alinhamento são consideradas candidatas para um estudo experimental *in vitro*.

A modelagem farmacofórica é uma técnica muito difundida na indústria farmacêutica, principalmente devido à sua disponibilidade em suítes comerciais de software voltadas para o desenvolvimento racional de fármacos (42). Tais pacotes incluem CATALYST (109) (HipHop (110) e HypoGen (111)) da Accelrys, GASP (112) e GALAHAD (113, 114) da Tripos, o módulo de farmacóforos do MOE (115) da CCG (*Chemical Computing Group*), Phase (116) da Schrödinger e LigandScout (108) da Inte:Ligand. Outras iniciativas não-comerciais para modelagem farmacofórica incluem o PharmaGist (117, 118) e o PharmID (119). A triagem virtual através de modelos farmacofóricos 3D é uma técnica poderosa, mas Hert et al. (120) sugerem que estes

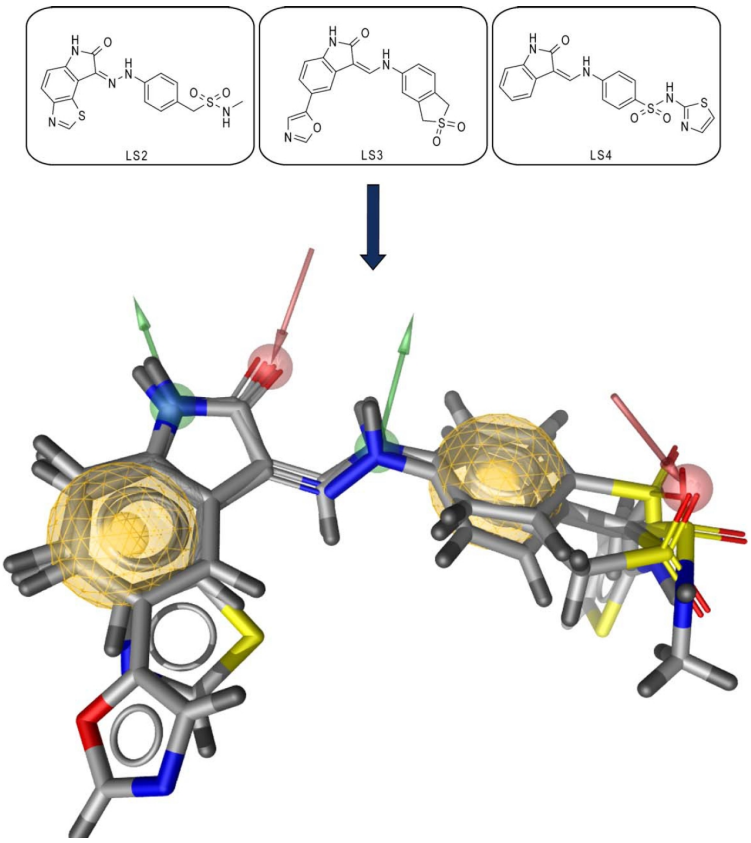


Figura 1.8: Um exemplo de uma hipótese farmacofórica construída pelo LigandScout (108), da Inte:Ligand. Três inibidores de Cinase Dependente de Ciclina 2 (CDK2) (identificados no PDB como LS2, LS3 e LS4), na conformação apresentada em estruturas cristalizadas depositadas no PDB (identificadores 1KE6, 1KE7 e 1KE8, respectivamente) foram usados para gerar um modelo de farmacóforos, contendo seis pontos de interação potencial, sendo duas regiões aromáticas (esferas amarelas), dois receptores de pontes de hidrogênio (esferas e vetores vermelhos) e dois doadores de hidrogênio (esferas e vetores verdes). Retirado de (102)

modelos podem não ser aplicáveis a moléculas bioativas estruturalmente heterogêneas, características de experimentos HTS. Neste caso, é apropriado considerar abordagens baseadas em busca por similaridade usando *fingerprints* de farmacóforo, que também são mais eficientes ao realizar a triagem em grandes bases de dados.

1.3.3 Farmacóforos codificados em vetores binários

Descritores moleculares poder ser codificados em vetores ou *fingerprints*, na maioria das vezes binários (75). As primeiras representações em vetores binários foram desenvolvidas para otimizar a busca por subestruturas químicas, que antes era feita através de casamento de grafos moleculares, um problema conhecidamente exponencial e que inviabilizava a busca em grandes bases de dados de compostos (28). Uma vez que as estruturas moleculares foram codificadas nos vetores binários, foi possível construir filtros preliminares que descartavam as substâncias que não continham os fragmentos presentes na referência. A partir da década de 80, as *fingerprints* passaram a também ser usadas em buscas por similaridade (30, 31), não só representando subestruturas e fragmentos moleculares, mas também o arranjo estrutural dos átomos e suas características fisico-químicas, visando o *bioisosterismo*, ou seja, identificar estruturas com propriedades similares à referência, mas com um conjunto diferente de átomos. Em 2002, Raymond e Willett (121) relataram que, ao comparar aplicações de triagem virtual, as que usavam *fingerprints* para representar as moléculas eram no mínimo tão eficazes quanto os métodos que representavam seus compostos através de grafos químicos, apesar do fato que *fingerprints* contêm uma representação muito menos precisa da estrutura molecular quando comparadas aos grafos, que contêm a descrição completa da topologia da substância.

O tipo de *fingerprint* mais simples é o vetor binário, onde cada posição é associada univocamente a uma ocorrência de subestrutura, fragmento ou arranjo de átomos ou farmacóforos, e o estado do bit designa a presença ou ausência da característica na molécula representada. Caso deseje-se agregar informações à *fingerprint* que extrapole sua natureza binária, o vetor pode armazenar o número de ocorrências de cada representação, se tornando uma *fingerprint* de frequência, que ainda pode ser normalizada, transformando os contadores em pesos que variam entre zero e um. Além de representar uma única molécula, *fingerprints* podem ser usadas para representar um conjunto de moléculas diferentes. Ao incorporar a informação de

múltiplas estruturas, sistemas baseados em *fingerprints* podem condensar os dados em uma única representação (120) através de *fingerprints* modais, onde um bit é aceso se sua frequência nas moléculas está acima de um determinado corte, ou ponderadas, onde cada bit recebe um peso de acordo com sua frequência. Pode-se usar também métodos de fusão de dados, incorporando a informação das múltiplas fontes no modelo final.

Para calcular a similaridade entre duas moléculas representadas por suas *fingerprints*, deve-se adotar um coeficiente que quantifique a medida. Em 2002, Holliday et al. (122) realizou um estudo compreensivo entre vários coeficientes de similaridade entre *fingerprints* 2D. Os resultados encontrados sugeriram que não há um coeficiente que se sobressaia como sendo o mais indicado. Dependendo da aplicação e da implementação das *fingerprints*, alguma métrica pode ser mais indicada que outra. Entretanto, o índice de Tanimoto mostrou-se o mais genérico dentre os coeficientes testados, tendo um desempenho satisfatório em todos os experimentos. O coeficiente (S_{TAN}) é calculado a partir do número de bits acesos em comum entre duas *fingerprints* (c), cada uma com a e b bits acesos no total:

$$S_{\mathrm{TAN}} = \frac{c}{a+b-c}$$

Fingerprints de farmacóforo codificam a informação sobre a presença ou ausência de pontos de farmacóforos potenciais (PPP - *Potential Pharmacophore Points*) e as distância entre eles em um composto ou conjunto de compostos (103). De modo geral, as *fingerprints* de farmacóforo codificam agrupamentos de dois, três ou até quatro PPPs (Figura 1.9), mas existem casos com agrupamentos maiores, como no trabalho de Martin e Hoeffel (123). Tripletes de PPPs são os mais usados, já que tradicionamente são considerados os mais efetivos em termos de conteúdo informacional versus complexidade (103). As distâncias euclidianas são discretizadas em intervalos e a escolha destes intervalos tem um impacto significativo no desempenho do método. Ao romper com as restrições geométricas presentes nos modelos farmacofóricos clássicos, as *fingerprints* conseguem descrever com mais facilidade compostos ativos com diferentes modos de ligação com o mesmo sítio ativo (42).

Assim como os programas de geração de hipóteses de farmacóforos estão presentes em pacotes de programas comerciais, os mesmos contêm módulos que implementam *fingerprints* de farmacóforos. Um dos pioneiros e um dos mais populares métodos de fingerprints 3D de far-

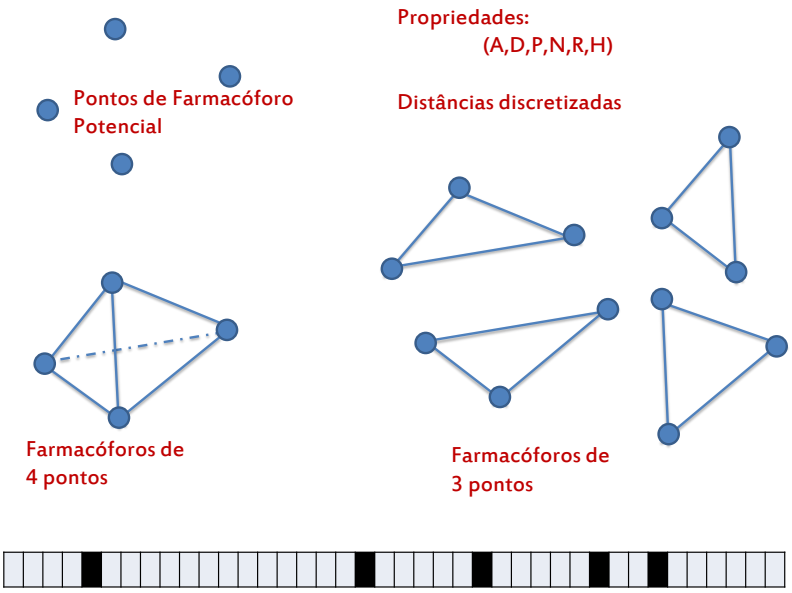


Figura 1.9: Como calcular *fingerprints* de farmacóforos: cada ponto potencial de interação é associado a um tipo físico-químico (**A**ceptor de hidrogênio, **D**oador de Hidrogênio, **P**ositivamente ou **N**egativamente carregado, a**R**omático e **H**idrofóbico) . As tuplas são formadas dependendo da aplicação (pares, triângulos ou tetraedros). Cada representação é então identificada no vetor binário, que armazena a informação de presença (1) ou ausência (0) daquela configuração específica

macóforo é o ChemX/ChemDiverse (124), da Chemical Design/Oxford Molecular (hoje incorporada à Accelrys). Outros exemplos comerciais incluem o PharmPrint (125, 126), da Affymax; a extensão do pacote MOE, da CCG, OSPPREYS (123) (*Oriented Substituent Pharmacophore PRopErtY Space*); o 3D-Keys (127), da Accelrys, baseado nas definições de farmacóforo do CATALYST e incluído na suíte Cerius^2; e o Tuplets (128), da Tripos. Dentre as iniciativas não-comerciais mais recentes, destacam-se o Pharmer (129) e o JPS (130) (*Joint Pharmacophore Space*).

1.4 Validação de Modelos

Ao produzir modelos que visam a predição de alguma propriedade ou classe, é preciso que ele passe por algum processo de validação para garantir que o modelo não só represente os dados que o geraram, mas seja genérico o suficiente para conseguir classificar dados não utilizados em sua geração. Um modelo que se baseie apenas em seus dados de treinamento sem nenhum tipo de validação externa não pode ser considerado confiável (131).

Uma das técnicas usadas para estimar o poder de predição de modelos é a validação cruzada. O conceito central das técnicas de validação cruzada é o particionamento do conjunto de dados em subconjuntos mutualmente exclusivos, e posteriormente, utiliza-se alguns destes subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento) e o restante dos subconjuntos (dados de validação ou de teste) são empregados na validação do modelo. O método k-particionado (*k-fold*) consiste em dividir o conjunto total de dados em k subconjuntos mutualmente exclusivos aproximadamente do mesmo tamanho e, a partir disto, utilizar um dos subconjuntos para ser o grupo teste, usando as $k-1$ partições restantes para a construção do modelo. O poder preditivo do modelo é avaliado com o grupo teste, medição esta que é repetida para os outros k pares partição-modelo. O conjunto destas medições é usado para estimar o poder de predição do modelo. Não existe consenso sobre qual deve ser o valor ideal de k, apesar que um dos valores mais usados seja $k = 10$. Existe um compromisso entre o valor de k e o tempo de processamento da validação, já que k testes de predição devem ser realizados. Para o caso degenerado de k ser igual ao número de amostras usadas no treinamento, a técnica de validação recebe o nome de LOO (*Leave One Out*), muito usada em estudos de QSAR, mas não recomendada para experimentos com número elevado de amostras, já que o tempo de processamento

cresce proporcionalmente com o valor de k.

Caso não haja dados independentes para um estudo de validação externa propriamente dito, técnicas de reamostragem podem ser usadas para estimar o poder de generalização do *método*. A técnica de *bootstrap* (132) é especialmente interessante, pois permite realizar um número suficientemente grande de experimentos independentes, já que retira, de modo aleatório, uma fração dos dados de treinamento para ser usada como dados de validação externa. O restante dos dados disponíveis são usados como dados de treinamento normalmente e a adequação do modelo gerado com os dados retirados é mensurada. O processo é repetido, recolocando as amostras previamente retiradas e realizando um novo sorteio. O *bootstrap* também pode ser usado para testar o impacto de parâmetros livres na modelagem.

Ao estudar a validação de modelos de QSAR, Golbraikh, Tropsha e outros (53, 55, 131, 133, 134) perceberam que a maioria das técnicas em QSAR alegavam produzir modelos com alto poder de predição por ter um coeficiente de correlação cruzada de LOO (q^2) acima de 0.5 para os dados de treinamento do modelo (55). O coeficiente de correlação cruzada é calculado tomando o nível de bioatividade de um composto (y_i), a média do nível de atividade de todos os compostos ($\bar{y}$) e a bioatividade prevista pelo modelo ($\hat{y}_i$) e aplicando estes valores na equação:

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)}{\sum (y_i - \bar{y})}$$

Apesar de Golbraikh e Tropsha concordarem que um valor de $q^2 < 0.5$ para as moléculas usadas para construir o modelo é um sinal que a habilidade de predição do modelo é ruim (55), eles discutem que um alto valor de q^2 não é critério suficiente para assegurar o poder de predição de um modelo QSAR, justamente por se basear apenas nos dados de treinamento do modelo (53). Os autores sugerem que a única maneira de se obter modelos realmente preditivos seria dividir os dados experimentais em dois grupos, um grupo treino e um grupo teste suficientemente grande. Assim, um modelo preditivo deve apresentar um valor de $q^2 > 0.5$ e ter um alto coeficiente de correlação entre a atividade prevista e a atividade biológica observada experimentalmente das moléculas do grupo teste ($r^2_{\text{pred}} > 0.6$). Posteriormente, Tropsha et al. (131, 134) estabelecem um fluxo de trabalho para produzir modelos de QSAR validados estatisticamente, robustos e com alto poder de predição, que pode ser generalizado para qualquer aplicação preditiva (Figura 1.10).

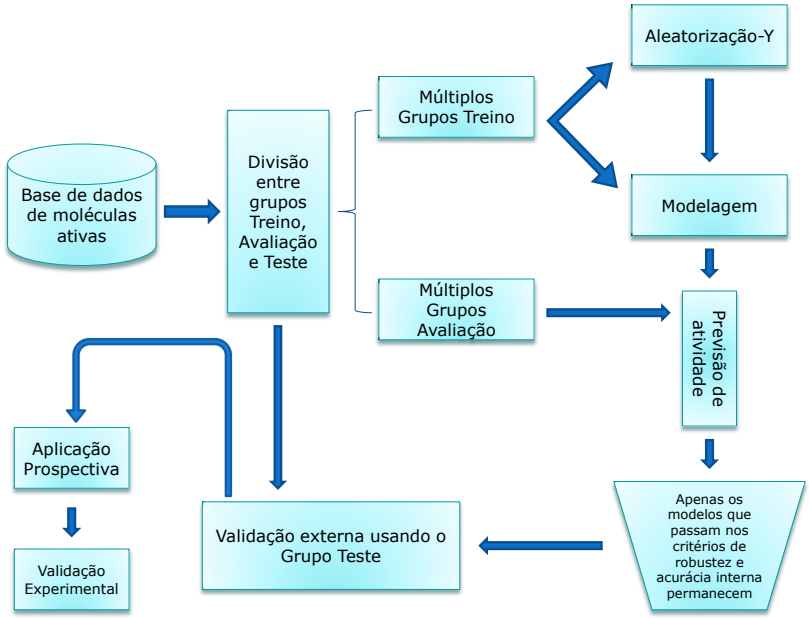


Figura 1.10: Fluxo de trabalho proposto por Tropsha et al. (131, 134) para produção de modelos validados em QSAR, generalizado para qualquer aplicação que produza modelos preditivos usados em triagem virtual.

A composição do grupo teste também foi discutida em publicações (53, 133, 135, 136) que sugerem que o grupo teste não deva ser escolhido aleatoriamente, mas tenha uma distribuição no espaço dos descritores próxima a do grupo treino, incluindo algoritmos para fazer a seleção dos grupos de maneira racional (53, 133, 135). Filimonov e Poroikov, ao discutir sobre abordagens probabilísticas em predição de atividade biológica (136), argumentam que a distribuição dos dados entre os grupos treino e teste para aplicações de triagem virtual sofre de um pequeno paradoxo: o grupo teste deve ser o maior possível para testar o desempenho do sistema, ao mesmo tempo que o grupo treino precisa da maior quantidade de dados possíveis para que o modelo construído seja aplicável. Para resolver esta contradição, o método de validação mais indicado para a construção de grupos testes é a validação cruzada k-particionada.

1.5 Tratamento de Estruturas Moleculares

Os equilíbrios tautoméricos e de ionização nas estruturas de macromoléculas biológicas e seus ligantes são fundamentais em muitos processos de reconhecimento molecular. Em estudos

de triagem virtual é crescente o reconhecimento de que a inclusão de múltiplas espécies pode gerar modelos mais preditivos. Por exemplo, de acordo com Scior et al (137), a inclusão de um tautômero incorreto na análise de características farmacofóricas pode influenciar erroneamente o processo de determinação de doadores e aceptores de ligações de hidrogênio. No entanto, ainda não existe um protocolo definido tanto no modo como são geradas e selecionadas as espécies mais relevantes, quanto na maneira como seriam utilizadas na geração de modelos. Os trabalhos que investigaram o efeito da tautomeria e da protonação em protocolos de triagem virtual baseados em alvos biológicos (138–140) não conseguiram mostrar uma influência positiva da enumeração de espécies na acurácia do *docking* molecular, mas, de maneira geral, recomendam considerar estas variações químicas nos estudos de triagem virtual, o que provavelmente pode ser aplicado nas iniciativas de LBVS.

Mais recentemente, nota-se uma preocupação em incluir múltiplas espécies e múltiplas conformações em uma abordagem mais ampla do conceito de atividade e da dinâmica inerente do processo de ligação entre moléculas pequenas e seus alvos biológicos. Tanto as macromoléculas biológicas quanto seus metabólitos e fármacos são moléculas adaptativas e flexíveis e, portanto, sujeitas a exibir propriedades emergentes que podem estar associadas tanto à resposta biológica quanto a efeitos colaterais ou indesejados, como a inibição ou o estímulo alostérico, ou efeitos agonistas e antagonistas provocados por pequenas moléculas ligadas ao mesmo sítio. Métodos que consigam reter a maior parte da informação da flexibilidade molecular poderão descrever com melhor precisão os eventos dinâmicos que governam a maioria dos processos biológicos, já que uma vez ligados a uma proteína, compostos ativos podem sofrer alterações conformacionais significativas em relação à conformação em solução aquosa ou no vácuo (141, 142).

O uso de abordagens Multi-Espécies Multi-Modo (MS-MM - *Multi-Species Multi-Mode*) em estudos baseados na estrutura 3D dos ligantes não é muito difundido, exceto em estudos de QSAR. O mais comum é o uso de uma espécie dominante no pH do experimento usado e de uma única conformação, que pode ser uma conformação otimizada (mínimo global) produzida por métodos de química teórica, uma conformação apresentada em uma estrutura cristalográfica de um complexo ligante-receptor, ou uma conformação provinda de um alinhamento com outros compostos ativos, visando reproduzir um possível modo de interação entre o composto e a macromolécula. Esta abordagem, apesar de simples e direta, assume a condição que uma única conformação é responsável pela atividade do composto. É comum o pressuposto que esta

conformação "bioativa" corresponda à conformação detectada em complexos cristalográficos. Entretanto, os cristais usados para a obtenção da estrutura não são obtidos em ambiente fisiológico, o que pode gerar dúvidas sobre a validade desta suposição (143). Idealmente, deveria se obter o máximo de conformações possíveis para cada molécula para garantir uma cobertura adequada dos possíveis arranjos bioativos. Entretanto, uma busca conformacional extensiva em um grande conjunto de dados moleculares iria acarretar um custo computacional proibitivo. Logo, é necessário estabelecer um compromisso entre custo computacional e a diversidade da amostragem conformacional (137).

Existe uma grande variedade de métodos de amostragem conformacional, muitos disponíveis nos pacotes de programas comerciais usados para a descoberta racional de fármacos. De modo geral, estes métodos se dividem em dois grupos: determinísticos, que calculam as conformações a partir dos ângulos torsionais de maneira sistemática; e estocásticos, que usam um elemento aleatório para explorar o espaço conformacional de uma molécula (144). O objetivo principal destes métodos é gerar e identificar uma série de conformações com um custo computacional baixo que tenha uma alta probabilidade de conter os arranjos bioativos (107).

O DiscoveryStudio (109), da Accelrys, contém três algoritmos diferentes para a amostragem conformacional. O CatConf (ou ConFirm) era previamente disponível através do CATALYST e possui dois modos de busca conformacional: *fast* e *best*. O primeiro faz uma busca semi-extensiva nas partes mais flexíveis da estrutura, enquanto usa uma série de conformações pré-definidas para ciclos e anéis. Já o modo *best* produz conformações de melhor qualidade, mas com um maior custo computacional quando comparado ao modo *fast*, usando uma matriz de coordenadas internas com limites superiores e inferiores para cada átomo para gerar arranjos tridimensionais (107). Ambos os modos do CatConf passam por uma etapa de exploração da diversidade conformacional usando o campo de força CHARMM (145) para cálculo das energias, mas o modo *fast* realiza um "relaxamento" das conformações encontradas dentro de uma janela de energia definida pelo usuário, enquanto o modo *best* utiliza uma técnica chamada *poling* (146), onde barreiras de energia potencial são colocadas entre mínimos locais para aumentar a diversidade conformacional. O outro método contido no DiscoveryStudio é o CAESAR (147), que realiza uma busca conformacional sistemática ao representar a molécula através de uma árvore, na qual as folhas são as menores unidades conformacionais em que a molécula pode ser dividida, enquanto as arestas representam as ligações rotacionáveis que ligam dois nós da árvore.

O OMEGA (148, 149), da OpenEye Scientific Software, é um método sistemático baseado em regras pré-definidas que descrevem as características torsionais de fragmentos. O algoritmo divide a molécula em fragmentos conectados, os quais podem conter entre uma e cinco ligações rotacionáveis. Conformações para cada fragmento são geradas através de uma biblioteca de ângulos torsionais pré-definidos, os quais são recolocados na estrutura inicial, visando construir as conformações da molécula como um todo. A amostragem conformacional da molécula inteira é feita usando um algoritmo de busca em profundidade, combinando os conjuntos de fragmentos de menor energia. Conformações duplicadas ou com sérias restrições estéricas são removidas, e as restantes são base para uma amostragem torsional mais refinada, gerando um número fixo de conformações (definido pelo usuário) dentro de um intervalo de energia.

O Macromodel (150), da Schrödinger, contém vários métodos para amostragem conformacional, muitos deles usando uma estrutura híbrida contendo elementos estocásticos e sistemáticos. Dentre eles, se destaca o LMCS (*Low-mode Conformational Sampling*) (151), que combina um método Monte-Carlo de amostragem do espaço torsional com um método baseado em autovetores para amostrar a vizinhança das conformações provindas do método anterior. Além disso, o Macromodel usa modelos de solvatação para enviesar a amostragem em direção a prováveis conformações bioativas. Entretanto, o custo computacional destes métodos os inviabilizam para estudos de triagem virtual em larga escala (107).

O MOE (115) contém métodos estocásticos e determinísticos para realizar a amostragem conformacional. O método estocástico é similar ao algoritmo RIPS (*Random Incremental Pulse Search*) (152), onde novas conformações são geradas ao se alterar repetida e aleatoriamente o valor de ângulos torsionais das ligações rotacionáveis na conformação anterior. Ao fim do processo as conformações são minimizadas e apenas as que estejam dentro de uma janela de energia e que não estejam duplicadas são mantidas. Já o algoritmo determinístico pode ser sistemático, incrementando ângulos torsionais por valores fixos, ou baseado em fragmentos com conformações pré-computadas.

Outros métodos de amostragem conformacional incluem o CONFORT (153), da Tripos, que realiza buscas sistemáticas cujos resultados são minimizados por campos de força incluído no próprio pacote da Tripos. O pacote QXP (154) realiza uma amostragem conformacional aleatória, onde as conformações encontradas são minimizadas sob um campo de força; o MIMUMBA (155, 156), da BASF, que usa regras e dados derivados de estruturas cristalográficas

para realizar a amostragem conformacional. Agrafiotis et al (157) desenvolveram um gerador de conformações baseados nas técnicas SPE (*Stochastic Proximity Embedding*) e SOS (*Self-Organizing Superimposition*). Já Griewel et al, da Universidade de Hamburgo, desenvolveram o TCG(*TrixX Conformer Generator*) (158), que quebra a molécula em fragmentos e começa a construir a conformação a partir do componente mais central, usando conformações de fragmento e ângulos diedros pré-calculados.

1.6 O Estado da Arte em Triagem Virtual Baseada na Estrutura de Ligantes

Apesar do uso de processos *in silico* no desenvolvimento racional de fármacos remontar da década de 60 e 70 (16), o desenvolvimento de novos métodos e algoritmos aplicados ao processo de desenvolvimento de fármacos é uma área muito ativa de pesquisa. A triagem virtual de alto desempenho é uma das áreas com um grande número de publicações de novas abordagens e metodologias e esta seção visa fornecer uma visão geral do estado da arte em triagem virtual, focando especialmente nos métodos que usam somente dados de compostos ativos como referência.

1.6.1 Metodologias LBVS baseadas em similaridade e/ou alinhamento

FLAP

O FLAP (*Fingerprints for Ligands And Proteins*) (159–162) é um programa que realiza triagem virtual através de buscas por similaridade usando *fingerprints* baseadas em farmacóforos de ligantes e/ou estruturas de proteínas. Para construir os PPPs, o FLAP usa uma abordagem similar à metodologia CoMFA, posicionando uma grade tridimensional gerada pelo programa GRID (163) ao redor de estruturas de proteínas, ligantes ou complexos proteína-ligante para mensurar as energias de interação em cada ponto da grade, através de sondas de diversos tipos como, por exemplo, sondas hidrofóbicas, polares ou aromáticas. A partir dos campos de interação molecular (MIFs - *Molecular Interaction Fields*) gerados pelo GRID, o FLAP condensa regiões favoráveis a interações de um determinado tipo em um conjunto de pontos de mínimos locais de energia, os quais passam a representar características farmacofóricas (Figura

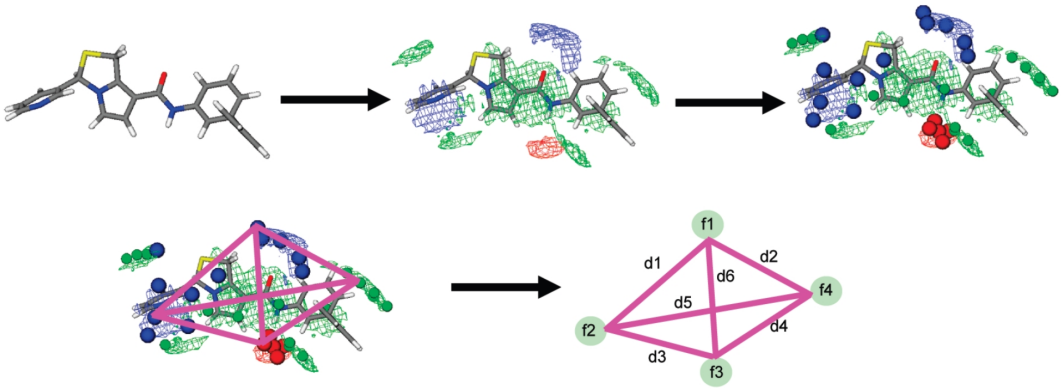


Figura 1.11: Um exemplo da aplicação do FLAP para geração de PPPs a partir da estrutura de um ligante: MIFs são calculados ao redor da estrutura usando o GRID e são condensados em PPPs. Todos os arranjos de quatro pontos são considerados, gerando grupos codificados em *fingerprints*. O procedimento é repetido para todas as conformações previamente geradas por um método estocástico. Retirado de (162)

1.11). Estes PPPs então são usados para gerar todos os arranjos de quatro pontos, formando tetraedros, os quais são codificados em *fingerprints*.

Para realizar a triagem virtual, o FLAP gera um conjunto de farmacóforos de quatro pontos de referência a partir de uma molécula ativa, um ligante co-cristalizado à estrutura do receptor ou a partir da estrutura do sítio ativo do receptor. Uma busca por similaridade é realizada entre a molécula de referência e os compostos de um conjunto de dados. Os tetraedros dos compostos mais similares que coincidem com um tetraedro do conjunto de referência são usados para alinhar a conformação selecionada da molécula à referência e este alinhamento é base das funções de similaridade e escore implementadas no programa. Opcionalmente, técnicas de fusão de dados (*Data Fusion*) podem ser empregadas para incorporar informações de múltiplas referências.

4D FAP$_{OA}$

O 4D FAP (*Flexible Atom Pairs*) (164–166) é um método de busca por similaridade no espaço conformacional das moléculas. Ao analisar a distribuição de distância entre átomos para múltiplas conformações, o método gera modelos de mistura de gaussianas (GMMs - *Gaussian Mixture Models*) para descrever de forma probabilística a distribuição das distâncias. GMMs são modelos probabilísticos que representam distribuições complexas baseados em combinações li-

neares de sub-distribuições individuais. Ao aplicar esta modelagem sobre conjuntos de conformações, o método consegue generalizar os dados discretos provindos de uma amostragem conformacional para um modelo contínuo, que pode ser armazenado apenas em função dos parâmetros da curva gaussiana (média e variância) e das combinações lineares. Os parâmetros das gaussianas são estimados através de um algoritmo de otimização de parâmetros chamado Maximização de Esperança (EM - *Expectation-Maximization*) (167).

Para o cálculo da similaridade entre duas moléculas, uma matriz de similaridade é construída tomando uma função mista entre a natureza dos pares de átomos e a sobreposição dos GMM de cada par. A similaridade final pode ser calculada através da soma normalizada dos valores dentro da matriz, como feito pelo 4D FAP original, aplicado em estudos QSAR/QSPR (164, 165). Para os estudos em triagem virtual, a similaridade foi calculada através de um algoritmo que implementa a resolução ótima do Problema de Atribuição[1] (OA - *Optimal Assignment*), que maximiza a valor da similaridade entre pares de átomos das duas moléculas sendo comparadas, resultando na variante 4D FAP_{OA} (166). Esta mudança foi inspirada nos trabalhos de Fröhlich (168, 169), que aplicou métodos de atribuição ótima a problemas de similaridade molecular.

FieldScreen

FieldScreen (48) é um método de busca por similaridade através do alinhamento tridimensional de pontos de mínimos locais de energia de interação calculados através de MIFs. O processo de comparação implementado pelo FieldScreen pode ser visualizado na Figura 1.12. Cada molécula passa por uma amostragem conformacional realizado por um gerador estocástico chamado XedeX (170), desenvolvido pelo mesmo grupo. As conformações são descritas por quatro campos de força moleculares definidos pela interação entre a molécula e uma sonda positiva, neutra, negativa ou hidrofóbica. Os pontos de mínimo local dos MIFs, juntamente com os valores destes campos, são usados como os descritores moleculares. A partir da comparação entre matrizes de distância dos pontos de mínimo da molécula usada como referência e da molécula a ser comparada, são gerados cliques[2] coloridos, ou seja, cliques cujos vértices são do mesmo tipo.

[1] Dado um grafo ponderado bipartido completo $G = (U \cup V, U \times V)$, uma atribuição ótima é um acoplamento $M \subset U \times V$ onde a soma dos pesos de cada aresta é maximizada

[2] *Cliques* são subgrafos completos. Dado um grafo $G(V, E \subset V \times V)$, um clique $C(U \subseteq V, U \times U)$ é um subconjunto de vértices do grafo G que são totalmente conectados, ou seja, cada vértice de um clique possui uma aresta ligando-o a todos os outros vértices do clique.

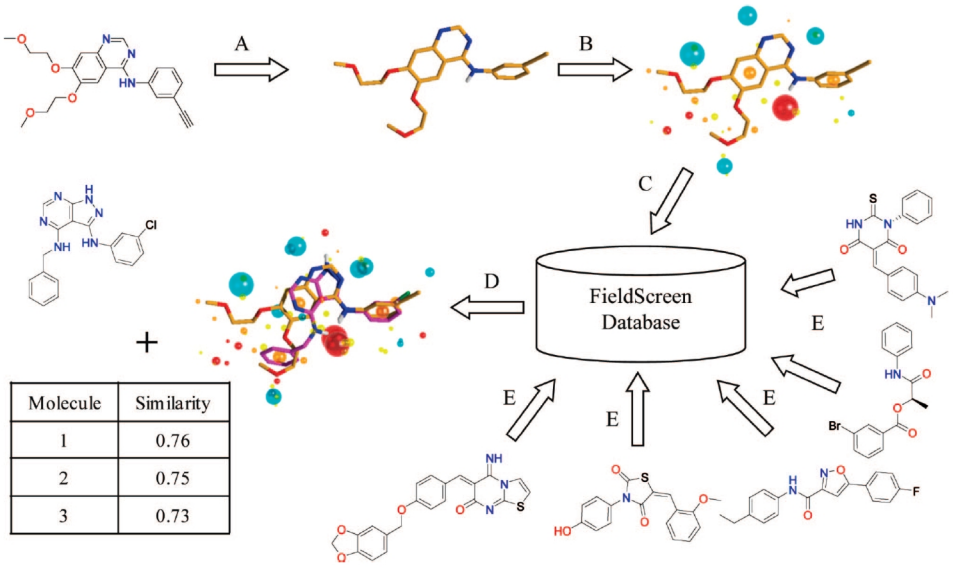


Figura 1.12: Representação esquemática os passos envolvidos na busca por similaridade em uma base de dados tratada pelo FieldScreen: **a)** Uma molécula ativa é selecionada como referência e a sua conformação "relevante" é calculada. **b)** Pontos de mínimo local são calculados e utilizados na representação da molécula referência. **c)** Uma busca em uma base de compostos que receberam o mesmo tratamento é realizada, usando os pontos de mínimo para alinhar moléculas similares **d)** Recuperação dos compostos com melhor alinhamento, quantificado através de um escore de similaridade molecular. A base de dados do FieldScreen é populada (**e**) pela exploração conformacional de todas as moléculas, com os pontos de mínimo adicionados e armazenados junto com as conformações. Retirada de (48)

Os cliques são avaliados através do seu tamanho e do valor de energia dos pontos de mínimo local armazenados, e o clique de maior escore é usado para alinhar as duas moléculas. O alinhamento é finalmente avaliado usando uma métrica de similaridade (índice de Dice).

LigMatch

LigMatch (171) é um método de busca por similaridade que usa tripletes de átomos para alinhar duas moléculas, usando um algoritmo conhecido como *hashing* geométrico (172). Cada molécula têm sua diversidade conformacional amostrada através do programa OMEGA (149) e todos os arranjos de três átomos são calculados, com as distâncias interatômicas armazenadas. Os triângulos de cada molécula sendo comparada cujos vértices são formados pelos mesmos elementos e com distâncias parecidas são considerados iguais, e são usados para alinhar as duas moléculas. O alinhamento é numericamente avaliado usando o número de átomos coincidentes

como fator de escore, o qual é calculado para todos os confôrmeros. O maior escore ou o escore médio são usados como critério de ordenamento da lista de moléculas.

PharmaGist

PharmaGist (117, 118, 173) é um método que alia uma detecção de farmacóforos em comum a um conjunto de compostos ativos com um alinhamento par a par entre o modelo farmacofórico construído e um conjunto de moléculas. A análise conformacional é feita de maneira explícita e determinística durante o processo de alinhamento. A elucidação do farmacóforo comum a um conjunto de ligantes é realizada pelo PharmaGist através de um algoritmo iterativo que define um ligante como molécula pivô e realiza múltiplos alinhamentos par a par entre o pivô e múltiplas conformações dos outros compostos. As melhores superposições pareadas são consideradas para gerar um alinhamento múltiplo, onde subconjuntos significativos de PPPs do ligante pivô se alinham ao maior número possível de compostos (Figura 1.13). O processo se repete, selecionando cada composto como pivô por vez. Entretanto, um pivô fixo pode ser selecionado pelo usuário. Os farmacóforos são ponderados de acordo com o número de compostos que apresentam a característica. Uma busca por similaridade pode ser feita ao alinhar os farmacóforos da molécula pivô com os presentes em uma molécula a ser buscada em um banco de dados. O programa está disponível através de um servidor (`http://bioinfo3d.cs.tau.ac.il/pharma/index.html`).

1.6.2 Uma Metodologia LBVS baseada em Aprendizado de Máquina

O JPS (*Joint Pharmacophore Space*) (130) é um método de predição de atividade biológica que usa técnicas de mineração de dados e aprendizado de máquina. Dado um conjunto de compostos ativos e inativos, PPPs são calculados a partir de conformações das estruturas todas as combinações de três pontos de farmacóforos são geradas. Os farmacóforos de três pontos são divididos em classes de acordo com os tipos envolvidos, e dois arranjos são mapeáveis se e somente se forem da mesma classe (i.e. seus vértices são do mesmo tipo farmacofórico.) Os farmacóforos são divididos em conjuntos formados por arranjos mapeáveis, que por sua vez são aglomerados através de um algoritmo de agrupamento k-medóides[3].

[3]Um medóide é análogo a um centróide de um aglomerado. A diferença é que o centróide é uma entidade que não necessariamente faz parte do conjunto de dados, enquanto o medóide é necessariamente um elemento do *cluster*

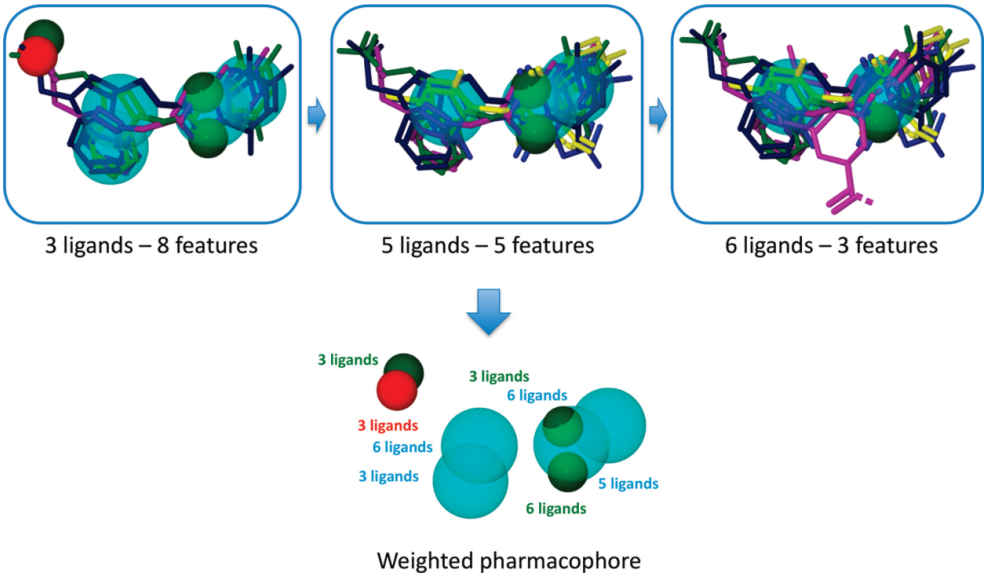


Figura 1.13: Um conjunto de alinhamentos múltiplos gerados pelo PharmaGist para um conjunto de ligantes da Aldose Redutase (ALR2). Os farmacóforos em comum apresentados por cada alinhamento múltiplo são ponderados de acordo com o número de compostos que apresentam os mesmos em cada alinahmento. Retirado de (118).

Os *clusters* são classificados de acordo com o seu poder de discriminação estatística em três classes: positivos, negativos e não-discriminativos. *Clusters* positivos possuem uma razão maior do que a razão esperada entre o número de farmacóforos de três pontos que foram gerados a partir de compostos ativos e o número de farmacóforos de três pontos no grupo em análise. Analogamente, *clusters* negativos contém uma razão maior do que a esperada entre farmacóforos provindos de compostos inativos e o número total de farmacóforos. Cada agrupamento discriminativo tem um farmacóforo central, chamado "farmacóforo significativo". Estes são usados para alimentar um classificador baseado em SVM, que por sua vez classifica novas moléculas de acordo com a atividade biológica pesquisada (Figura 1.14).

1.6.3 Metodologias LBVS baseadas em superposição de volume

SHAFTS

SHAFTS (50) é uma abordagem híbrida de busca por similaridade molecular, a qual é calculada através da soma das similaridades normalizadas de *fingerprints* de farmacóforo e de superposição de volumes. A abordagem do SHAFTS para triagem virtual (Figura 1.15) pode

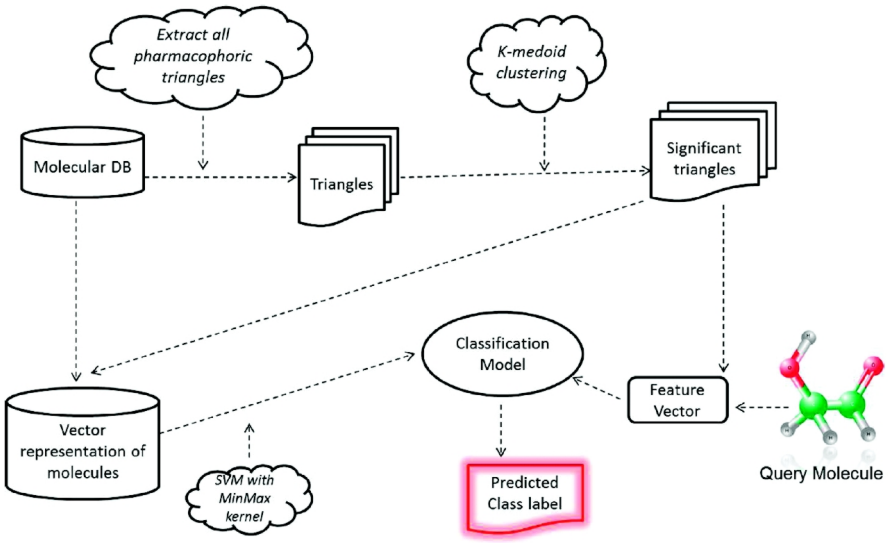


Figura 1.14: Fluxo do algoritmo usado pelo JPS para realizar a classificação molecular de atividade biológica. Os triângulos formados por PPPs são extraídos dos dados de treinamento (contendo compostos ativos e inativos) e são agrupados através do algoritmo *k*-medóide. Os aglomerados têm sua significância estatística determinada e seus farmacóforos centrais são usados para gerar um modelo de classificação baseado em SVM, que podem ser usados para determinar a atividade de uma nova molécula. Retirado de (130)

ser dividida em três estágios. Primeiramente, os PPPs, baseados em regras de grupos funcionais, são determinados na molécula de referência, que pode ter uma conformação pré-definida por uma estrutura cristalográfica ou ter seu espaço conformacional amostrado (Figura 1.15 A) através do Cindy (174), um método baseado em um algoritmo evolucionário desenvolvido pelo próprio grupo que desenvolveu o SHAFTS. A seguir, a base de moléculas a ser buscada passa pelo mesmo protocolo de amostragem conformacional e determinação de PPPs, com o passo adicional opcional de otimização da conformação através do alinhamento entre as moléculas da base e a molécula de referência (Figura 1.15 B). Finalmente, é realizada uma busca por similaridade usando os tripletes de PPPs da referência indexados em uma tabela *hash*. Para cada coincidência, as conformações das moléculas que apresentam o mesmo triplete são alinhadas com a referência. Cada alinhamento é avaliado de acordo com uma medida de similaridade que leva em conta o alinhamento dos PPPs e a superposição dos volumes (Figura 1.15 C). Apenas a conformação com melhor índice de similaridade é armazenada.

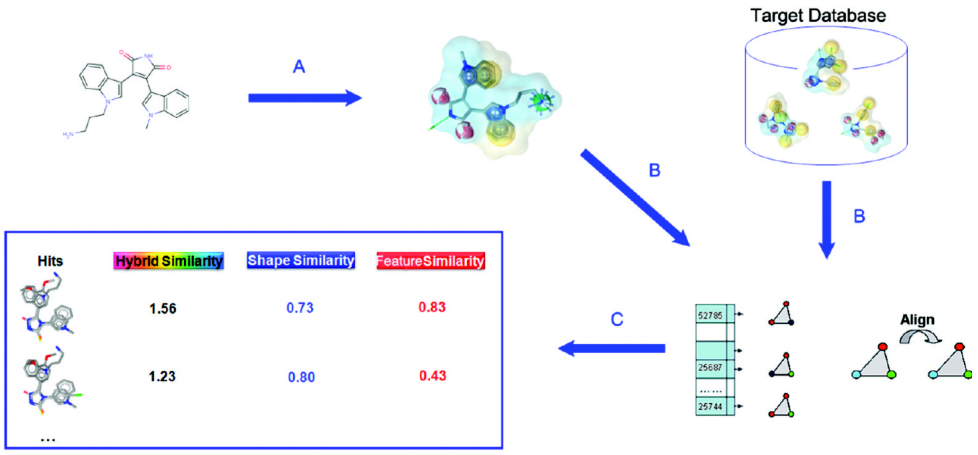


Figura 1.15: Representação esquemática do fluxo de trabalho do SHAFTS para triagem virtual: **(A)** Seleção de um composto ativo covertido a uma conformação específica e adição dos PPPs como referência. **(B)** Busca na base de dados sobrepondo cada estrutura à referência usando a indexação por tabela *hash*. **(C)** O resultado da busca é ordenado de acordo com a similaridade híbrida, e os alinhamentos resultantes são fornecidos como saída. Retirado de (50).

Phase Shape

Phase Shape (51) é um novo método de superposição de volume da Schrödinger, que pode considerar tipagem de unidades de volume para o cálculo da superposição. O método consiste em uma rápida busca por similaridade usando a distribuição das distâncias radiais entre átomos vizinhos como filtro, seguida de um processo de maximização da superposição dos volumes entre os melhores resultados. O escore da superposição pode ser avaliado tomando apenas os volumes de van der Walls dos átomos pesados e hidrogênios carregados, ou pode ser ponderado de acordo com o elemento considerado (Figura 1.16). Um terceiro modo usa características farmacofóricas mapeadas aos átomos para a avaliação do alinhamento, o que gera os melhores resultados.

1.6.4 Bases de Dados para Estudos Retrospectivos Comparativos

Apesar da grande diversidade de métodos disponíveis para triagem virtual, a comparação entre eles é uma tarefa árdua, já que cada grupo de pesquisadores publica métricas diferentes para bases de dados diferentes (90). Com a finalidade de reverter este panorama caótico, alguns grupos publicaram bases de dados específicas para *benchmarking* comparativo de ferramentas em virtual screening. Dentre estas, o DUD (175) e o MUV (176) se destacam como as mais

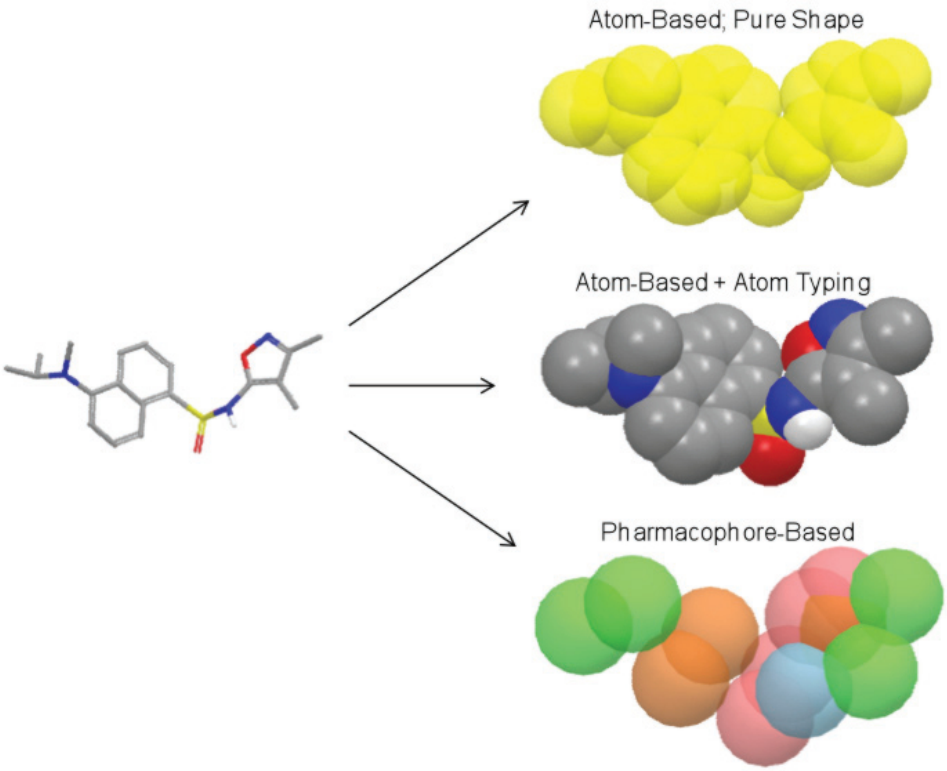


Figura 1.16: Os três modelos de volume suportados pelo Phase Shape. Retirado de (51).

usadas em recentes estudos retrospectivos comparativos em triagem virtual. Outra iniciativa recente para a uniformização de *benchmarks* é o REPROVIS-DB (177), que usou dados de estudos prospectivos de triagem virtual para gerar um conjunto de compostos ativos e inativos para 25 sistemas relacionados a alvos terapêuticos, sistenas que se assemelham a um cenário prático de aplicação de triagem virtual prospectiva.

DUD

O DUD (175) (*Directory of Useful Decoys* - `www.dud.docking.org`) originalmente surgiu como uma base de dados químicos para uniformizar estudos comparativos de ancoragem molecular. 40 alvos protéicos de interesse terapêutico foram selecionados de acordo com a disponibilidade de ligantes anotados, estruturas cristalizadas e estudos prévios de *docking* molecular. Para cada um dos alvos selecionados, um conjunto de ligantes foi retirado da base de dados ZINC (178), em conjuntos que variavam entre 12 e 416 compostos, totalizando 2950 moléculas na publicação original. Para cada um destes conjuntos de compostos ativos, um conjunto de compostos supostamente inativos foi gerado, buscando moléculas topologicamente dissimilares mas com propriedades fisico-químicas parecidas com os ativos, os chamados *decoys*. Este protocolo foi realizado com o intuito de se evitar que classificadores triviais conseguissem diferenciar os ativos do conjunto de inativos usando descritores simples, como massa molecular, número de grupos funcionais, dentre outros. Uma média de 36 compostos inativos foi gerado para cada composto ativo.

Apesar dos dados serem inicialmente planejados para estudos comparativos de desempenho de ferramentas de ancoragem molecular, muitos estudos retrospectivos em triagem virtual baseados na estrutura dos compostos ativos utilizaram os dados do DUD como base de *benchmark* ((48, 88, 118, 162, 166, 171, 179–181) para citar alguns poucos). Para tanto, é recomendado usar não a versão original do DUD, mas uma versão que passou por um filtro de características de compostos-protótipo (182), sugerido por Good e Oprea (92) e aplicado por Jahn et al (179) nos estudos de métodos baseados no Problema da Atribuição. Estes dados também estão disponíveis através da página do DUD. Duas meta-análises independentes (87, 180) consideraram esta versão filtrada do DUD adequada para estudos LBVS retrospectivos. Além disso, Good e Oprea (92) agruparam os compostos ativos pela sua similaridade estrutural, gerando agrupamentos de moléculas estruturalmente parecidas dentre compostos ativos. Para cada um desses

agrupamentos, foi selecionada uma molécula medóide que representaria todo o *cluster*, gerando um conjunto de ligantes estruturalmente únicos, os chamados DUD-Parents, cujo tamanho é igual ao número de classes estruturais dos compostos ativos de cada alvo. O resultado deste agrupamento também está disponível na página do DUD.

MUV

O conjunto de dados MUV(*Maximum Unbiased Validation*) (176) foi projetado para evitar enriquecimentos artificiais em estudos LBVS. Uma base de dados de compostos usada para estudos retrospectivos pode sofrer tipicamente dois vieses que influeciam a avaliação de seus desempenhos: o "viés do análogo" (92) (os compostos ativos são muitos similares entre si) e o "enriquecimento artificial" (183) (os compostos ativos são muito dissimilares aos compostos inativos). Com o intuito de se medir estes vieses em bases de dados de compostos, os autores aplicaram uma técnica de estatística espacial, chamada Análise Refinada do Vizinho mais Próximo (*Refined Nearest Neighbour Analysis*) (184, 185), para mensurar estes vieses de conjuntos de dados de *benchmark* e projetar um novo conjunto de compostos baseado em dados de ensaios *in vitro* disponíveis no PubChem BioAssay (186). A coleção de conjuntos de dados resultante deste estudo compreende ligantes associados a 17 alvos terapêuticos, sendo que cada conjunto é formado por 30 compostos ativos (com pelo menos 21 classes estruturais) e 15000 compostos inativos. A dimensão destes conjuntos foi recentemente criticada por Nicholls (87) no capítulo de sua autoria presente no livro editado por Bajorath (187), que sugere que uma razão de 40 compostos inativos para cada ativos é mais que suficiente para minimizar o erro experimental em estudos comparativos de triagem virtual. Coincidentemente, esta razão é bem próxima da apresentada pelo DUD (36:1).

2 Objetivos

2.1 Objetivo Geral

Construir e validar uma ferramenta que possibilite estudos de triagem virtual baseada na estrutura dos ligantes ativos, codificada através de *fingerprints* de farmacóforos

2.2 Objetivos Específicos

- Propor um novo método para construção e validação de modelos robustos e preditivos, baseados em extensa validação cruzada interna.
- Propor uma nova representação das estruturas moleculares na forma de farmacóforos potenciais utilizando múltiplas espécies e conformações.
- Realizar estudos retrospectivos para validar o bom funcionamento da ferramenta, e comparar seu desempenho contra outras aplicações semelhantes.

3 *Metodologia*

3D-Pharma é uma aplicação de Triagem Virtual baseada na estrutura de substâncias com atividade biológica previamente conhecida. Um protocolo de tratamento das estruturas moleculares foi desenvolvido e aplicado sobre todos os compostos (Seção 3.1), com o objetivo de uniformizar as moléculas e amostrar o espaço de possibilidades de configurações químicas e conformacionais das substâncias. Regras para o mapeamento farmacofórico serão definidas na Seção 3.2, mapeamento este que será usado na construção das *fingerprints* (Seção 3.3), vetores binários usados pelo 3D-Pharma. Os vetores correspondentes às substâncias ativas são submetidos a um processo de construção e validação de modelos (Seção 3.4) que, por sua vez, podem ser comparados a outras substâncias e usados para predizer a atividade das mesmas.

3.1 Tratamento das Estruturas Moleculares

Para evitar introdução de viéses nos modelos produzidos, é necessária uma padronização dos compostos envolvidos como discutido na Seção 1.5. Para tanto, um pré-tratamento foi realizado através de uma dessanilização manual, que consistiu em uma inspeção manual dos arquivos obtidos das diferentes bases de dados de compostos, quando os mesmos eram formados por múltiplas estruturas (sais, misturas, etc.). Além disso, cada molécula é representada pelo seu tautômero mais provável. Estes procedimentos foram realizados pelos programas Instant JChem e Standardizer, ambos da ChemAxon (188), que também foram usados para transformar todas as estruturas para o formato SMILES.

O protocolo de tratamento molecular desenvolvido para amostrar o espaço de possibilidades químicas e conformacionais de substâncias inicia-se com a determinação dos tautômeros dominantes, ou seja, as variantes tautoméricas com maior concentração nas faixas de pH entre 0 e 14.

O cálculo é feito pelo módulo ‘tautomers’ do programa CXCalc, da ChemAxon (188). Para cada tautômero, calcula-se seu estado de protonação em pH 7. Tal cálculo é feito pelo módulo ’majorms’ do programa CXCalc (188). Na Figura 3.1 a. pode se ver um exemplo do cálculo de tautômeros aplicado à Histidina.

Para a amostragem conformacional, escolheu-se um protocolo que visa otimizar a relação custo computacional / precisão, como discutido na Seção 1.5. Tal protocolo consiste em três passos: cálculo da conformação mais estável, cálculo de cargas parciais e amostragem conformacional. Usando o programa Omega2, da OpenEye (149) e o campo de força MMFF94s (189), calculou-se a conformação mais estável de cada representação da molécula. A seguir, aplicou-se o programa molcharge, parte integrante do pacote QuacPac, também da OpenEye (149), sobre cada conformação, a fim de calcular as cargas parciais de cada átomo com o método semi-empírico AM1-BCC (190). Finalmente, amostrou-se o espaço conformacional de cada representação molecular, novamente através do Omega2 e usando o campo de força MMFF94s. As moléculas que possuem mais de 25 ligações rotacionáveis foram descartadas. Cada representação (espécie) de cada molécula que foi mantida gerou conformações que obedecem às seguintes restrições

- Apenas conformações com energia menor que a soma da energia da conformação mais estável (mínimo global de energia) e uma janela de cinco kcal/mol (para moléculas com menos de cinco ligações livres) a dez kcal/mol (para moléculas com cinco ou mais ligações rotacionáveis).
- A seleção é limitada a 50 diferentes conformações para moléculas com até quatro ligações rotacionáveis, 100 diferentes conformações para moléculas que possuem de cinco a nove, 150 conformações para moléculas que possuam de 10 a 14, e a 200 conformações para moléculas com 15 ou mais ligações livres.
- Uma conformação é considerada uma duplicata de uma outra conformação se ela possui uma distância quadrática média (RMSD - *Root Mean Square Distance*) menor que 0,5 Å(para moléculas com até quatro ligações rotacionáveis) ou menor que 1,0 Å(para moléculas com cinco ou mais ligações livres).

Ao final do processo de tratamento, cada molécula é descrita por uma miríade de representações

e conformações que representam os aspectos dinâmico e químico da substância. Diferentemente dos métodos mais tradicionais, nenhum tautômero ou conformação é descartado *a priori*, exceto se apresentar baixa distribuição em ampla faixa de pH ou elevada energia, respectivamente. Assim, os aspectos dinâmicos inerentes à estrutura molecular são conservados e mesmo nos casos onde existe mais de uma conformação ativa ou diferentes mecanismos de ação, é possível a análise e a geração de modelos preditivos de atividades biológicas das moléculas. Mesmo no caso de uma mesma molécula interagir com diferentes alvos por meio de diferentes conformações, diferentes modelos podem aflorar a partir da mesma descrição da estrutura molecular (Figura 3.1).

3.2 Mapeamento Farmacofórico

Ao atribuir características farmacofóricas aos átomos pesados de uma molécula, utilizou-se o programa PMapper, da ChemAxon (188), que usa uma série de regras para determinar a(s) classe(s) designada(s) a cada átomo. O programa analisa se um dado átomo é capaz de realizar ligações de hidrogênio e/ou se faz parte de um anel aromático. Tais informações combinadas com a informação de cargas parciais são suficientes para o programa atribuir uma ou mais características farmacofóricas, tranformando o átomo em um (ou mais) ponto(s) de farmacóforo potencial, ou PPP (*Potential Pharmacophore Point*). Cada PPP possui uma coordenada espacial para localização, correspondente à coordenada tridimensional do átomo, e um tipo associado.

PPPs podem assumir uma das seguintes características:

- Positivamente carregado (P), se um átomo possui uma carga parcial acima de +0.4
- Negativamente carregado (N), se um átomo possui uma carga parcial abaixo de -0.4
- Doador de Hidrogênio (D), se um átomo pode doar hidrogênios para estabelecer uma Ligação de Hidrogênio
- Aceptor de Hidrogênio (A), se um átomo pode receber hidrogênios para estabelecer uma Ligação de Hidrogênio
- Aromático (R), se um átomo faz parte de um anel aromático
- Hidrofóbico (H), se um átomo não se encaixa em nenhuma das características anteriores.

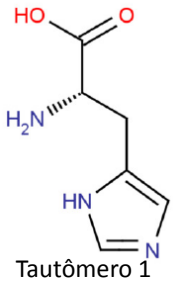
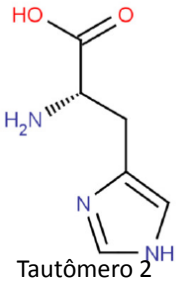
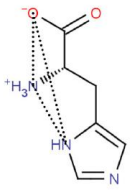
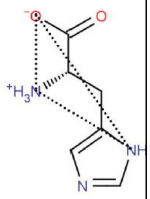
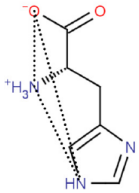
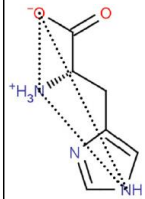
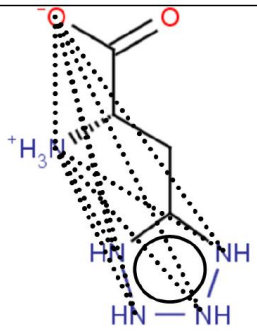
a	SMILES	N[C@@H](CC1=CN=CN1)C(O)=O			
b	Tautômeros	 Tautômero 1		 Tautômero 2	
c	Protômeros, Conformações e Farmacóforos	 Conf. 1a	 Conf. 1b	 Conf. 2a	 Conf. 2b
d	Farmacóforo triplete alfanumérico	PND111	PND123	PND123	PND134
e	Farmacóforo indexado pela função *hash*	25421	356924		12382
f	Estrutura híbrida com o farmacóforo dinâmico do 3D-Pharma				

Figura 3.1: Efeito do tratamento das estruturas moleculares na geração de farmacóforos no 3D-Pharma. O aminoácido Histidina foi selecionado como exemplo e todas as estruturas estão representadas em 2D para melhor visualização. **a)** A representação em formato SMILES da Histidina. **b)** As estruturas dos dois tautômeros dominantes da Histidina, mostrando a troca do hidrogênio entre os dois átomos de nitrogênio no anel imidazólico. **c)** As estruturas das microespécies dominantes (protômeros) de cada tautômero da Histidina em pH 7. Duas conformações hipotéticas são representadas para cada protômero, considerando apenas a rotação do anel imidazólico. O triângulo de PPPs formado por um dos átomos de oxigênio carregados negativamente no grupo carboxila (N), o nitrogênio-α carregado positivamente (P), e o átomo de nitrogênio no anel imidazólico, o qual faz o papel de doador de hidrogênio (D). O mesmo triângulo é representado nas outras três conformações. **d)** Os triângulos farmacofóricos de cada conformação são convertidos em caracteres alfanuméricos. "PND" representa a trinca formada pelos farmacóforos: Positivamente carregado, Negativamente carregado e Doador de hidrogênio. Os números após os caracteres representam as distâncias discretizadas entre os átomos (ver Figura 3.2). **e)** O farmacóforo triplete indexado pela função *hash*. **f)** A representação híbrida hipotética do farmacóforo PND da Histidina codificado pela *fingerprint* do 3D-Pharma. Todos os farmacóforos triplete detectados em todas as conformações são igualmente considerados.

Ao final do mapeamento, cada átomo pesado dará lugar a um ou mais PPPs com a mesma coordenada tridimensional, mas com um tipo diferente de farmacóforo associado (por exemplo, Negativo e Aceptor de hidrogênio). Esta informação será usada para a geração das *Fingerprints*.

3.3 Construção de *Fingerprints*

Para gerar as *fingerprints*, calcula-se todos os agrupamentos possíveis de três PPPs da molécula. Agrupamentos que possuam PPPs derivadas do mesmo átomo (quando a distância euclidiana entre dois PPPs é zero) são eliminados. Estes tripletes têm as distâncias entre seus vértices discretizadas de acordo com os intervalos de distância (*bins*) presentes na Tabela 3.1.

Bins	0	1	2	3	4	5	6	7	8	9
Intervalos	$< 3\mathring{A}$	$3-4,5\mathring{A}$	$4,5-6\mathring{A}$	$6-8\mathring{A}$	$8-10\mathring{A}$	$10-12,5\mathring{A}$	$12,5-15\mathring{A}$	$15-18\mathring{A}$	$18-21\mathring{A}$	$> 21\mathring{A}$

Tabela 3.1: Intervalos de discretização de distâncias entre pontos

Os tripletes são convertidos em *strings* de seis caracteres, de modo que cada uma esteja associada univocamente ao trio de PPPs. Os tipos farmacofóricos estão representados nos primeiros três caracteres, enquanto os números finais correspondem às distâncias discretizadas entre cada ponto, tal como mostra a Figura 3.2. Essa string é indexada através de uma função *hash* previamente calculada pela biblioteca CMPH (191) e apenas seu índice é armazenado. Quando todas as combinações de três PPPs passarem por este processo, a molécula será representada por uma série de índices numéricos que codificam todas as informações provenientes de cálculos de tautomeria, de protonação, conformacionais e de interação potencial (na forma de características farmacofóricas). Tal representação permite um cálculo quantitativo de similaridade entre moléculas, através do Índice de Tanimoto, que pode ser calculado pela Equação 3.1, onde A representa o conjunto de índices de uma molécula a ser comparada com os índices de outra molécula B.

$$S_{\text{Tanimoto}} = \frac{|A \cap B|}{|A \cup B|} \tag{3.1}$$

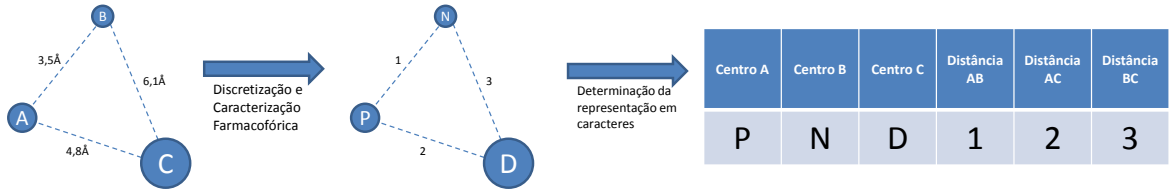


Figura 3.2: Exemplo de conversão de um triângulo formado por PPPs em uma string de seis caracteres. Cada par de PPP tem suas distâncias discretizadas e os centros são ordenados de acordo com essas distâncias de maneira a formar sempre uma string que identifica univocamente o triplete

3.4 Geração e Validação dos Modelos

Um modelo é uma abstração dos dados provenientes das substâncias ativas e deve conter os descritores que são significativos para a caracterização da atividade biológica para um certo alvo terapêutico. Quanto mais similar for o modelo às moléculas que o originaram, maior a probabilidade de que ele contenha a informação que descreva a atividade (28). Um modelo do 3D-Pharma é formado pelos índices dos tripletes de farmacóforos mais comuns entre as moléculas do grupo treino. Formalmente, o modelo M é um conjunto de configurações de PPP (x) que estão presentes nas moléculas integrantes do grupo treino (T) acima de um valor de corte τ, ou seja:

$$x \in M \Leftrightarrow \frac{\sum_{i=1}^{m} f(x,T_i)}{m} \geq \tau, f(x,T_i) = \begin{cases} 1 & \text{se } x \in T_i \\ 0 & \text{caso contrário} \end{cases} \tag{3.2}$$

onde m é o tamanho do grupo treino T, e T_i é a i-ésima molécula do grupo. Experimentos preliminares (Seção 4.2.1) sugeriram que um valor padrão de $\tau = 0.7$ é genérico o suficiente para manter o poder preditivo dos modelos.

Considerando as sugestões contidas nos trabalhos do Prof. Tropsha sobre a validação de modelos de QSAR (53, 55, 131, 133, 134), o modelo deve ser extensivamente validado para assegurar seu poder de predição. Para tanto, foi desenvolvido um protocolo de construção de modelos com alto poder preditivo baseado em um esquema de validação cruzada *10-fold*.

Inicialmente, o grupo de moléculas ativas que farão parte do modelo deve ser separada em dez partições. Esta separação pode ser tanto aleatória quanto sistemática, ou seja, distribui-se as moléculas através das partições de acordo com um critério, buscando maior homogeneidade entre elas, o que torna o processo de validação uma Validação Cruzada Estratificada. O critério usado foi a similaridade média entre as moléculas. Uma matriz de similaridade é calculada, contendo a similaridade entre todos os pares de moléculas. As moléculas são ordenadas pela similaridade média e distribuídas pelas partições. A primeira partição recebe a molécula centróide (com a maior média de similaridade), a segunda partição recebe o vizinho mais próximo ao centróide e assim por diante, até que todas as moléculas tenham sido distribuídas.

Durante o processo de construção do modelo (Figura 3.3), as dez partições podem assumir três papéis. O grupo teste é usado para validação interna e é formado por uma partição. As nove partições restantes são divididas em dois grupos. O grupo de avaliação, formado por três partições, é usado para avaliar a similaridade do modelo com um conjunto diferente de moléculas. Já o grupo treino, formado por seis partições, é responsável pela construção do modelo. Os tripletes de PPP que obterem uma frequência acima de um certo valor de corte (único parâmetro de entrada do programa) entre as moléculas integrantes das partições que fazem parte do grupo treino farão parte do modelo (Equação 3.2). Todas as 840 possibilidades de distribuição das partições entre os grupos Treino, Teste e Avaliação são consideradas.

Durante a validação cruzada, uma partição assume o papel de Grupo Teste, enquanto as outras nove são usadas em todas as combinações possíveis para gerar os grupos Avaliação e Treino[1]. A similaridade média entre cada um dos 84 modelos e seu respectivo Grupo Avaliação é medida, e apenas os dez modelos com a maior similaridade são selecionados para a validação interna.

Na validação interna, o Grupo Teste é então inserido em um conjunto de moléculas inativas e os modelos são usados para recuperar as moléculas do Grupo Teste. O desempenho dessa busca é medido através da área sobre a curva ROC (AUC_{ROC} - *Area Under the Receiver-Operator Characteristic Curve*) e apenas o modelo com a maior AUC_{ROC} é mantido. Este processo é repetido, de forma que todas as partições assumam o papel de Grupo Teste, totalizando dez modelos finais.

[1]Como estamos distribuindo nove grupos entre seis posições no Grupo Treino, temos $C_9^6 = \binom{9}{6} = 84$ possíveis pares de Grupos Treino e Avaliação

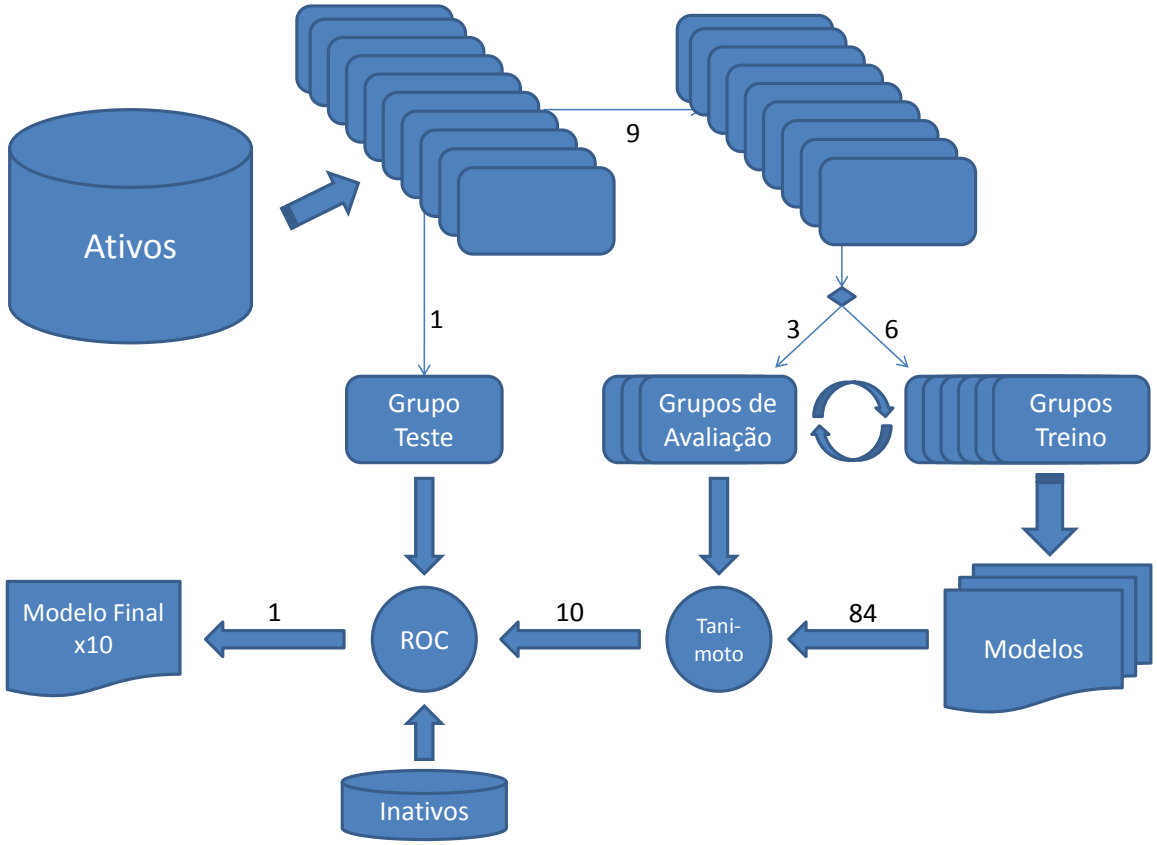


Figura 3.3: Fluxograma do processo de construção e validação interna de modelos usando Validação Cruzada *10-fold*. Primeiramente, uma partição (correspondente a aproximadamente 10% das moléculas ativas) é separada como grupo Teste, enquanto as outras nove partições são distribuídas entre grupos Treino e Avaliação. Dos 84 modelos produzidos, apenas dez são selecionados, de acordo com o valor da similaridade média entre o modelo e seu respectivo Grupo de Avaliação. Finalmente, estes dez modelos são validados, tentando recuperar as moléculas do grupo Teste dentre um conjunto de moléculas inativas. Apenas o modelo com o melhor desempenho é mantido. O processo então se repete para as outras nove partições, totalizando dez modelos finais.

Todo esse processo requer ao menos dez moléculas ativas para a execução. Caso o conjunto de ativos não seja maior ou igual a dez, apenas um modelo simples é produzido, de acordo com a Equação 3.2. No caso, ao invés do grupo T_i ser um subconjunto das moléculas ativas, todas elas fazem parte do grupo, o que inviabiliza a validação interna do modelo, já que não existem grupos Avaliação e Teste. Este modelo simples também pode ser gerado para os datasets maiores, caso seja a opção do usuário.

A aplicação do modelos deve ser feita sobre um conjunto de moléculas que passou pelo mesmo tratamento descrito na Seção 3.1. O modelo deve ser comparado par a par com cada uma das moléculas, que por sua vez devem ser ordenadas de acordo com a similaridade com o modelo. Como iremos demostrar nos resultados (Seção 4.3.2), as moléculas no topo da lista ordenada tem uma alta probabilidade de interagirem com o mesmo alvo terapêutico para o qual as substâncias usadas para a construção do modelo são ativas.

4 Resultados e Discussão

Neste capítulo encontram-se os dados dos experimentos usados para ajuste de parâmetros (Seção 4.2) e validação do 3D-Pharma (Seção 4.3). Os experimentos foram realizados sobre os dados apresentados na Seção 4.1.

4.1 Bases de Dados

Os estudos de validação do 3D-Pharma foram feitos usando conjuntos de compostos ativos e inativos associados a alvos constantes na base de dados do DUD (175) (*Directory of Useful Decoys* - `www.dud.docking.org`). O conjunto de compostos inativos foram filtrados de acordo com o filtro aplicado por Jahn et al (179). Três conjuntos de dados externos e independentes de moléculas ativas foram usados nos experimentos para gerar os modelos de predição de atividade. O conjunto "Fármacos" contém moléculas aprovadas pela FDA[1] ou ainda em fase clínica de desenvolvimento. Tais substâncias foram obtidas de três bases de dados: DrugBank (192, 193) (`www.drugbank.ca`), KEGG DRUG (194) (`http://www.genome.jp/kegg/drug/`) e TTD (195) (`http://bidd.nus.edu.sg/group/cjttd/`). O conjunto "Ligantes-PDB" contém as moléculas que se ligam às estruturas cristalográficas dos respectivos alvos do DUD depositadas no PDB (23) (`www.pdb.org`). A lista dos ligantes associados a cada alvo proveniente do PDB foi analisada manualmente a procura de íons, moléculas que fazem ligações covalentes com a cadeia protéica ou aminoácidos modificados. Tais compostos foram retirados da análise. O terceiro conjunto de dados é baseado no banco de moléculas WOMBAT (26, 92). Apesar do banco de dados ser comercial, foram disponibilizados no site do DUD alguns conjuntos de

[1] *Food and Drug Administration* - Órgão norte-americano que regulamenta a aprovação e venda de fármacos e artigos alimentícios nos EUA, a mesma função realizada no Brasil pela ANVISA (Agência Nacional de Vigilância Sanitária)

compostos ativos para 13 alvos (92), sendo que 11 deles também fazem parte do DUD. Destes onze, o Receptor de Estrogênio foi o único que não foi incluído nos estudos de validação, pois ao contrário do DUD, as bases externas não diferenciam agonistas de antagonistas no seu conjunto de moléculas ativas. Ao final, dentre as 40 atividades do DUD, foram selecionados dez alvos representantes para os experimentos, devido à disponibilidade de dados do WOMBAT.

Todas as substâncias passaram pelo mesmo protocolo de tratamento de estruturas moleculares descrito na Seção 3.1. Os conjuntos externos de compostos ativos também passaram por um filtro 2D para eliminar redundâncias entre os dados utilizados na construção de modelos e os dados de compostos ativos do DUD. O número de compostos de cada conjunto de dados pode ser visto na Tabela 4.1.

4.2 Experimentos Preliminares

4.2.1 Valores de Corte da Modelagem

Utilizando uma versão anterior do 3D-Pharma, foram testados dois valores de corte que parametrizavam a construção dos modelos, com o intuito de gerar um único modelo final ao unir os dez modelos gerados pelo protocolo de validação. Além do valor de corte τ, presente na Equação 3.2, um segundo valor de corte (υ) era utilizado de maneira análoga sobre os modelos para unir os tripletes de PPP sob uma única *fingerprint* modal. Com o intuito de encontrar os melhores valores de corte, um estudo foi realizado sobre os dez conjuntos de dados de compostos ativos associados aos alvos da base de dados do DUD presentes na Tabela 4.1, usando o conjunto de dados DUD-Parents como conjunto de compostos ativos, e o conjunto DUD-Decoys como conjunto de compostos inativos.

Os modelos construídos a partir das moléculas dos conjuntos DUD-Parents foram avaliados quanto à sua capacidade de priorizar os compostos pertencentes aos DUD-Ativos quando misturados aos compostos inativos do DUD-Decoys através da área sob a curva ROC. Os resultados (Tabela 4.2.1) sugerem que o valor ótimo de τ é 0,6, seguido por um corte υ que poderia ser tanto 0,2 ou 1,0, já que a diferença entre as médias de AUC_{ROC} dos mesmos não é significativa estatisticamente (confiança acima de 95% no teste t-Student). Ao evoluir o algoritmo de validação, decidiu-se que um modelo único não mais seria gerado, mas sim um conjunto de dez

Alvo	Conjunto de Dados	Número de Compostos	Número de Classes Estruturais no DUD-Ativos
ALR2	DUD-Ativos DUD-Decoys Fármacos Ligantes-PDB WOMBAT	26 910 11 26 41	14
AR	DUD-Ativos DUD-Decoys Fármacos Ligantes-PDB WOMBAT	68 2616 45 13 36	10
$PPAR_{\gamma}$	DUD-Ativos DUD-Decoys Fármacos Ligantes-PDB WOMBAT	6 38 12 63 27	6
CDK2	DUD-Ativos DUD-Decoys Fármacos Ligantes-PDB WOMBAT	47 1702 37 135 148	32
COX-2	DUD-Ativos DUD-Decoys Fármacos Ligantes-PDB WOMBAT	212 11577 77 4 66	44
EGFR	DUD-Ativos DUD-Decoys Fármacos Ligantes-PDB WOMBAT	365 14516 10 12 62	40
FX_{α}	DUD-Ativos DUD-Decoys Fármacos Ligantes-PDB WOMBAT	64 1888 5 85 105	19
HIVRT	DUD-Ativos DUD-Decoys Fármacos Ligantes-PDB WOMBAT	34 1370 4 29 97	17
P38	DUD-Ativos DUD-Decoys Fármacos Ligantes-PDB WOMBAT	135 5416 16 77 52	20
PDE5	DUD-Ativos DUD-Decoys Fármacos Ligantes-PDB WOMBAT	26 1561 12 10 85	22

Tabela 4.1: Número de substâncias dos conjuntos de dados associados aos alvos do DUD selecionados para o estudo de validação externa do 3D-Pharma, além do número de classes estruturais dos compostos ativos do DUD para cada conjunto.

τ	0,0	0,1	0,2	0,3	0,4	0,5	**0,6**	0,7	0,8	0,9	1,0
AUC_{ROC}	0,645	0,692	0,787	0,833	0,857	0,870	**0,877**	0,871	0,858	0,843	0,837
υ	0,0	0,1	**0,2**	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
AUC_{ROC}	0,803	0,805	**0,832**	0,807	0,810	0,811	0,815	0,816	0,820	0,823	0,829

Tabela 4.2: Resultados da avaliação do impacto dos valores de corte τ e υ sobre o poder preditivo dos modelos em uma versão anterior do 3D-Pharma. Na versão atual, um único corte é usado. Considerando que $\upsilon = 0,0$ rendeu o pior resultado, foi decidido que o valor padrão do corte único (τ=0,7) seria mais restritivo do que o ótimo encontrado anteriormente.

modelos que poderiam, por exemplo, ser usados para gerar uma lista de consenso. Para tanto, a estratégia de dois cortes sucessivos não era mais adequada, fazendo que o corte υ não fosse mais necessário. Entretanto, a sua contribuição para a melhora do poder preditivo dos modelos finais foi considerada, já que foi decidido que um valor mais restritivo do valor de τ (0,7) fosse adotado como padrão.

4.2.2 Tempo de processamento das comparações moleculares

Para se definir o tempo de processamento de um experimento em triagem virtual usando o 3D-Pharma, é preciso somar a contribuição dos vários módulos que compõe o programa, desde o tratamento molecular até a geração de modelos e busca por similaridade. O algoritmo de geração de farmacóforos de três pontos para o cálculo das *fingerprints* é exaustivo, com complexidade $O(n^3)$, onde n é o número de PPPs em uma molécula. Entretanto, n é relativamente pequeno, tornando a execução deste módulo extremamente rápida para uma única conformação, embora a contribuição do tempo necessário para se processar todas as conformações de todas as espécies de uma molécula seja considerável. Apesar disso, o tratamento molecular e a geração das *fingerprints* são executados apenas uma vez para cada composto de uma base de dados, possibilitando a realização de vários experimentos de triagem virtual com as mesmas susbstâncias, sem necessidade de reexecutar os mesmos protocolos. Portanto, a contribuição destes módulos não será considerada na análise de tempo de processamento.

Por sua vez, a geração de modelos também usa um algoritmo exaustivo para calcular todas as 840 combinações entre grupos treino, avaliação e teste para enfim construir os dez modelos finais. Mas, apesar de ser um algoritmo "força bruta", o tempo total de processamento da geração de modelos é dominado pela avaliação dos mesmos, especialmente quando os modelos são testados

Alvo	Tempo médio de comparação Molécula x Molécula (ms)	Número médio de grupos de 3 PPPs por molécula	Tempo médio de comparação Modelo x Molécula (ms)	Numero médio de grupos de 3 PPPs por por modelo
ALR2	17,42	2950,0	8,82	602,0
PDE5	51,49	7258,8	30,16	1729,9
AR	40,94	6058,6	24,09	268,07
P38	N/A	5908,3	28,55	1966,03

Tabela 4.3: Tempo de processamento para a comparação entre *fingerprints* de moléculas e modelos usando o 3D-Pharma, detalhado para quatro alvos do DUD. Os experimentos foram realizados em um servidor Linux Ubuntu 9.04 com dois processadores Intel Xeon 2,33 GHz e 2GB de memória, com quatro núcleos cada.

sobre um grupo de moléculas inativas misturadas com as moléculas do grupo teste. Dado que cada grupo teste é comparado com dez modelos a fim de se escolher um modelo final e que a validação cruzada implementada usa dez grupos, são feitos 100 testes, avaliados através de curvas ROC, por execução do protocolo de validação cruzada. Logo, a busca por similaridade se justifica como o principal componente computacional do 3D-Pharma.

Com o intuito de avaliar o tempo de processamento das buscas por similaridade, além de comparar o desempenho computacional do 3D-Pharma com dados de outras aplicações que também realizam comparações moleculares, um estudo foi realizado usando as moléculas provenientes do DUD-Decoys sem considerar o filtro *drug-like* aplicado por Jahn et al. (179). Foram escolhidos quatro alvos dentre os alvos do DUD (Tabela 4.1) que possuem um número diverso de compostos inativos: ALR2 (918 compostos), PDE5 (1808 compostos), AR (2618 compostos) e P38 (7312 compostos). Dois experimentos distintos foram realizados: comparação DUD-Decoys x DUD-Decoys[2] e comparação Modelos x DUD-Decoys. O primeiro visa avaliar o tempo de processamento médio necessário para se comparar duas moléculas, enquanto o segundo visa avaliar a busca por similaridade como seria usada na prática em experimentos de triagem virtual, comparando-se o modelo com um grande conjunto de substâncias. Os experimentos foram realizados em um servidor Linux Ubuntu 9.04 com dois processadores Intel Xeon 2,33 GHz e 2GB de memória, com quatro núcleos cada. Apesar do número de núcleos, o processamento do código de comparação não é distribuído, ou seja, a comparação usa apenas um núcleo por vez.

[2]O tempo de comparação molécula x molécula para o conjunto P38 DUD-Decoys não pôde ser calculado. A dimensão do número de comparações ($7312^2 \cong 2,67 \times 10^7$) não permitiu a medição em tempo viável.

Método	3D-Pharma Mol x Mol	3D-Pharma Modelo x Mol	ChemAxon CXN-H (196)	ROCS (197, 198)	FRED (199)	ICMsim (200)	Surflex-sim (201)	FlexS (202)
Tempo médio por composto (s)	$0,027 \pm 0,020$	$0,023 \pm 0,002$	0,07	0,5	1,0	2,4	6,7	6,9
Processador	Intel Xeon 2,33 GHz	Intel Xeon 2,33 GHz	Intel Core 2 Quad 2,40 GHz	Intel Xeon 2,40 GHz	Intel Xeon 2,40 GHz	Intel Xeon 2,40 GHz	Intel Xeon 2,40 GHz	Intel Xeon 2,40 GHz

Tabela 4.4: Tabela comparativa entre o tempo de processamento de comparação e busca por similaridade por composto. A média dos tempos mensurados para o 3D-Pharma, tanto em comparações molécula versus molécula quanto as comparações modelo versus molécula, está bem abaixo dos tempos apresentados por outras técnicas em LBVS. Além disso, a variância dos tempos mensurados para as comparações modelo versus molécula é muito menor quando comparada à variância das comparações entre moléculas, o que é esperado devido à grande variabilidade de tamanho entre diferentes moléculas. Entretanto, como o 3D-Pharma usa dez modelos na busca, o tempo de comparação médio em uma aplicação típica de triagem virtual deve ser multiplicada por dez, equiparando o tempo médio de comparação, em termos de ordem de grandeza, com o tempo apresentado pelo ROCS.

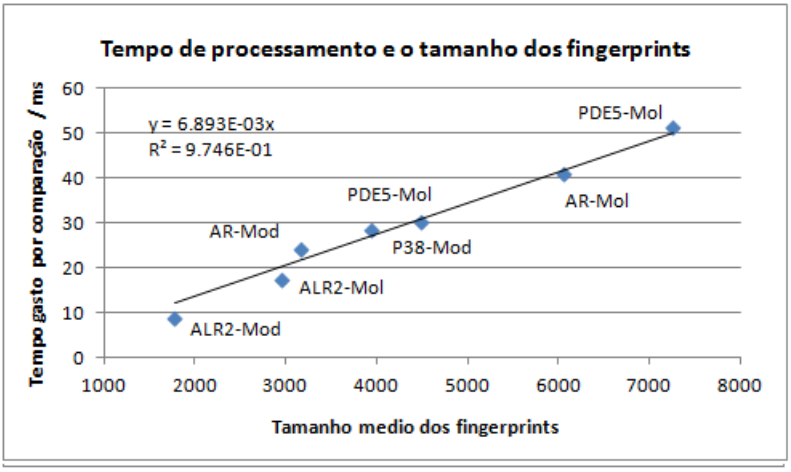


Figura 4.1: A relação linear ($r^2 = 0,975$) entre o tamanho dos vetores envolvidos na comparação e o tempo de processamento da mesma.

Os dados gerais e comparativos estão apresentados na Tabela 4.4. Os dados do método CXN-H foram obtidos de uma apresentação em um simpósio de usuários dos programas da ChemAxon, em Maio de 2011 (196), enquanto que os dados das outras técnicas foram retirados do trabalho de Giganti et al. (203). O 3D-Pharma é capaz de realizar em média 37 comparações entre duas moléculas por segundo. Quando usado para triagem virtual, ao realizar a comparação entre um conjunto de dez modelos e uma molécula, o 3D-Pharma é capaz de avaliar 4,3 compostos por segundo. A comparação molecular é 2,6 vezes mais rápida que a apresentada pelo método CXN-H, da ChemAxon (14 moléculas/s), o método mais rápido dentre os investigados. Entretanto, a avaliação da similaridade de uma molécula contra os dez modelos produzidos pelo 3D-Pharma, o tempo cai para 0,23s por comparação, 3,25 vezes mais lento, mas ainda duas vezes mais rápido que o segundo método, o ROCS (2 moléculas/s).

Ao analisar em detalhe os tempos obtidos pelo 3D-Pharma (Tabela 4.3), podemos perceber que o tempo de processamento do algoritmo de comparação entre duas *fingerprints* implementado é linearmente proporcional ao total de farmacóforos de três pontos detectados nas duas moléculas (ou modelo) representadas pelas *fingerprints* (Figura 4.1). Uma forte correlação linear ($r^2 = 0,975$) pode ser observada. Esta relação era esperada, já que a complexidade do algoritmo de comparação é $O(m+n)$, onde n e m é o tamanho dos vetores das duas moléculas (ou modelo) sendo comparadas.

4.2.3 Análise do impacto do número de conformações no poder preditivo do modelo

O 3D-Pharma usa uma amostragem conformacional compreensiva, com cada molécula podendo ter até 200 conformações diferentes por representação (espécie), todas contribuindo para sua representação binária final. Desejou-se saber se caso as conformações menos parecidas com os modelos finais fossem eliminadas e um novo modelo fosse construído a partir do novo conjunto de conformações, o poder de predição do modelo poderia aumentar, indicando uma convergência da amostragem conformacional a um pequeno conjunto significativo. Um experimento então foi realizado usando os conjuntos de dados de compostos associados à Cinase dependente de Ciclina 2 e os dados dos quatro conjuntos de substâncias ativas: DUD-Ativos Representativos (DUD-Parents), Fármacos, Ligantes-PDB e WOMBAT.

Cada conjunto de compostos ativos gerou dez modelos, de acordo com o protocolo descrito na Seção 3.4. Cada modelo teve sua acurácia mensurada através da AUC_{ROC} contra os conjuntos DUD-Ativos e DUD-Decoys. Cada uma das conformações de cada molécula ativa do conjunto de dados usado na construção dos modelos foi comparada com os modelos gerados. Todas as conformações foram ordenadas em ordem decrescente por similaridade com cada modelo e a sua posição na lista ordenada de cada modelo é somada. Esta soma é usada como critério de ordenação, gerando uma lista das conformações ordenadas de acordo com o consenso entre os dez modelos. Um quinto das conformações piores colocadas no consenso é retirado da análise, assegurando-se que pelo menos uma conformação por molécula seja mantida. Usando as conformações restantes de cada composto, calcula-se uma nova *fingerprint*, a qual é usada para construir um novo conjunto de modelos. Os passos descritos acima são repetidos, sendo retiradas 20% das conformações a cada iteração.

O resultado pode ser visualizado na Figura 4.2. De modo geral, a retirada das conformações mais dissimilares aos modelos não altera significamente o poder preditivo dos modelos, mas tende a piorá-lo à medida que a amostra conformacional diminui. Entretanto, a consistência interna dos modelos melhora, como mostra os valores crescentes de AUC_{ROC} da validação interna dos modelos, exceto para o DUD-Parents, onde a AUC interna praticamente se manteve inalterada. Há que se ressaltar que apenas as conformações dos compostos ativos usados nos modelos foram alteradas, mas os modelos foram comparados com as moléculas do DUD-

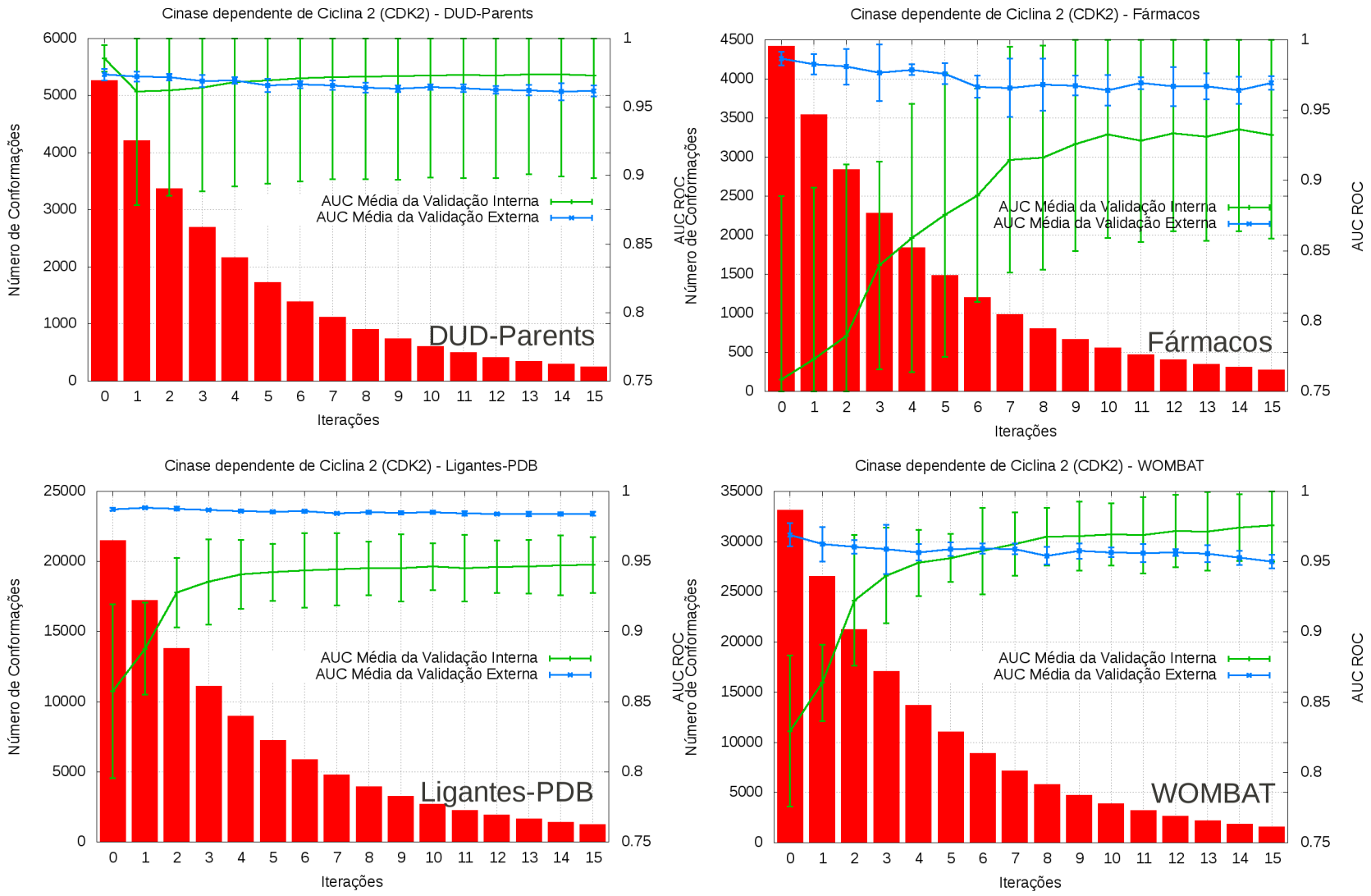


Figura 4.2: Variação dos valores de AUC_{ROC} média da validação interna (usando o grupo teste) e externa (usando o DUD-Ativos) à medida que as conformações mais dissimilares aos modelos são retiradas da análise. Em sentido horário, começando do canto superior esquerdo: Parents DUD, Fármacos, Ligantes-PDB e WOMBAT. Nota-se que a AUC média da validação externa praticamente se mantém inalterada, com uma fraca tendência a diminuir, enquanto que a AUC média da validação interna tende a melhorar (exceto para o conjunto Parents DUD, onde os valores de AUC da validação interna também se mantêm).

Ativos e DUD-Decoys com sua amostragem conformacional completa. O resultado sugere que os descritores mais importantes para a seleção dos compostos ativos encontram-se codificados em poucas conformações, já que a seleção das conformações mais parecidas com modelo influenciam pouco o resultado final. Entretanto, não parece ser possível saber *a priori* em qual conformação específica estes descritores se encontram, se é que se encontram em uma única conformação. Logo, uma seleção de poucas conformações pode perder arranjos críticos para o reconhecimento de outros compostos ativos, sendo recomendado manter a amostragem extensiva das conformações.

4.3 Validação

4.3.1 Validação por *bootstrap*

Foi realizado um estudo de validação pelo método de *bootstrap*, que consiste em retirar da base de compostos ativos uma pequena fração (no caso 20%) das moléculas e utilizá-las como conjunto de validação externa. Tal seleção deve ser feita aleatoriamente e repetida diversas vezes. Para essa validação, foram feitos 60 testes por base externa de substâncias ativas presentes na tabela 4.1, sendo 30 testes com distribuição aleatória das moléculas pelas partições, e 30 através da distribuição sistemática dos compostos pelas partições, as quais são usadas para gerar normalmente os dez modelos, como descrito pela Seção 3.4. Os dez modelos são então usados como referência em uma busca por similaridade para recuperar os compostos separados anteriormente (para a validação externa) de um conjunto de compostos inativos. Cada busca foi avaliada através da área sobre a curva ROC (AUC_{ROC}). Os resultados também foram usados para avaliar o desempenho de cada método de distribuição dos compostos ativos entre os grupos da validação.

O resultado pode ser visualizado na Figura 4.3. O resultado geral atesta o bom funcionamento do 3D-Pharma, já que a AUC_{ROC} média da distribuição aleatória foi 0.892, enquanto que a AUC_{ROC} média da distribuição sistemática foi um pouco menor: 0.879. Nos histogramas da Figura 4.3, pode-se visualizar a distribuição das AUCs nos dois métodos. Não é possível chegar a alguma conclusão sobre qual dos métodos é mais indicado baseado apenas na avaliação visual da distribuição dos AUCs ou baseado nos valores das médias, já que ambos os métodos apresentaram perfis parecidos e médias dentro dos intervalos de desvio padrão. Ao realizar um teste

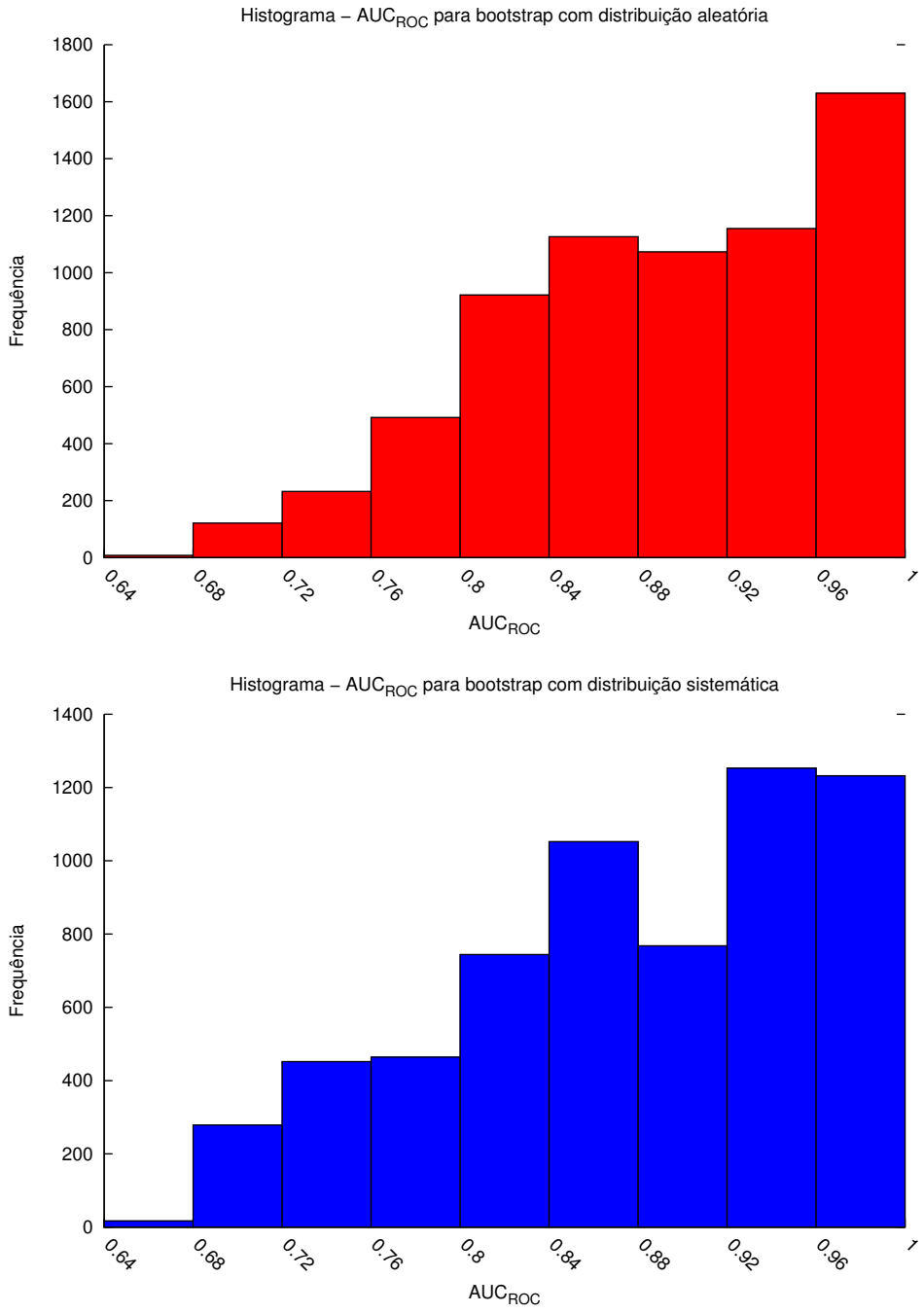


Figura 4.3: Histogramas de frequência de AUC_{ROC} para a validação por *bootstrap* do 3D-Pharma separados pelo método de distribuição dos compostos pelos dez grupos do protocolo de validação cruzada *10-fold*. Acima, AUCs referentes à distribuição aleatória (em vermelho) e abaixo AUCs referentes à distribuição sistemática (Validação cruzada estratificada, em azul). Os dois métodos têm uma distribuição de AUC parecida, entretanto o método aleatório apresenta um pico de AUC entre 0,96 e 1,00, enquanto o método sistemático apresenta dois picos nos intervalos $0,84-0,88$ e $0,92-0,96$. Além disso, um método aleatório apresentou uma distribuição de AUC um pouco mais uniforme do que o método sistemático.

Fonte de Variância	g.l.	SQ
Total	$n-1$	$\mathrm{SQ}_{\mathrm{TOTAL}} = 82,99$
Método de Distribuição	1	$\mathrm{SQ}_{\mathrm{Dist}} = 1,19$
Alvos	9	$\mathrm{SQ}_{\mathrm{Alvos}} = 25,93$
Conjuntos de Dados	2	$\mathrm{SQ}_{\mathrm{Conjuntos}} = 13,59$
Erro Experimental	$n-13$	$\mathrm{SQ}_{\mathrm{Erro}} = 42,28$

Tabela 4.5: Resultados da análise de variância dos dados do *bootstrap*. Observa-se que boa parte da variância dos valores de AUC no *bootstrap* provém dos diferentes alvos e diferentes conjuntos de dados utilizados, sendo necessário uma análise mais específica.

t-Student (usando o programa R) para ver se as médias são estatisticamente diferentes, se obtém um valor de $T = 9,56$. A hipótese nula (as duas médias são iguais) foi rejeitada com $p < 2,2^{-16}$. A análise de variância dos dados mostra que há uma grande contribuição da dispersão dos alvos e dos conjuntos de dados para a variância geral do método (Tabela 4.5), sendo necessária uma análise mais detalhada da série de dados de cada par alvo/conjunto de dados. A contribuição de cada fonte de variância para a soma dos quadrados (SQ) relativa à variação total do experimento foi calculada através da fórmula $\mathrm{SQ}_{\mathrm{FV}} = \sum \frac{T_{\mathrm{FV}}^2}{r_{\mathrm{FV}}} - \frac{(\sum X)^2}{n}$, onde T_{FV} é a soma das r_{FV} observações relativas à fonte de variância (FV) investigada, enquanto o termo $\frac{(\sum X)^2}{n}$ é um fator de correção considerando todas as observações. Com o valor de $\mathrm{SQ}_{\mathrm{Erro}}$, podemos calcular a variância do erro: $s^2 = \frac{\mathrm{SQ}_{\mathrm{Erro}}}{\mathrm{g.l.}_{\mathrm{Erro}}} = \frac{42,28}{n-13} \approx 0,0033$, com $n = 13020$.

Ao realizar os testes T-Student sobre os dados do *bootstrap* separados por alvo e conjunto de referência (Tabela 4.6), pode-se verificar que o método aleatório obteve as melhores médias de $\mathrm{AUC}_{\mathrm{ROC}}$ em 14 dos 24 pares alvo/conjunto de dados. Em três casos, a média dos dois métodos é estatisticamente igual e em sete a média do método sistemático foi melhor. Considerando a média de 300 observações por par alvo/conjunto de dados e um intervalo de confiança de 99%, valores de $|T| > 2,6$ na Tabela 4.6 significam que as médias de AUC entre os métodos aleatório e sistemático são diferentes. Valores positivos sinalizam que a média do modo aleatório é maior e, analogamente, valores negativos sinalizam que a média do modo sistemático é maior. Quanto maior o valor absoluto de T, maior a diferença significativa entre as médias. É interessante perceber que o método sistemático obteve a melhor média em seis dentre os dez alvos quando analisados somentes os dados do WOMBAT, enquanto que o método aleatório obteve melhores médias nos outros conjuntos de dados.

Também foram analisadas as médias de AUC entre os dois métodos quando um dos métodos

Alvo	Global	Fármacos	Ligantes-PDB	WOMBAT
ALR2	6,6	N/A	32,7	-23,0
AR	-3,2	0,6	0,7	-24,1
$PPAR_{\gamma}$	5,7	6,7	8,6	-13,0
CDK2	18,3	17,7	8,9	3,1
COX2	69,4	58,0	N/A	43,6
EGFR	6,4	N/A	20,7	-7,3
FX_{α}	-34,6	N/A	198,3	2,4
HIVRT	5,1	N/A	6,7	-4,0
P38	28,3	24,6	5,8	-6,1
PDE5	-18,9	-37,5	N/A	35,4

Tabela 4.6: Valores de testes T-Student entre as séries de $AUCs_{ROC}$ geradas pelo *bootstrap* para cada conjunto de dados associados aos alvos da Tabela 4.1 sem diferenciação dos conjuntos de dados de compostos ativos (Global) e considerando cada conjunto de dados separadamente. Os conjuntos de dados que contém menos de 12 compostos ativos não foram considerados e são marcados na tabela pelo valor "N/A".

AUC_{ROC} Média		Melhor Método	
		Aleatório	Sistemático
Método Aplicado	Aleatório	0,893	0,903
	Sistemático	0,860	0,936

Tabela 4.7: Valores de AUC_{ROC} média ao se aplicar um dos métodos de distribuição de compostos entre os grupos de validação cruzada nas situações em que o mesmo método tem o melhor desempenho e nas situações em que o método não é o mais indicado.

é o melhor para um dado par alvo/conjunto de dados (Tabela 4.7). O método aleatório tem a melhor média geral (0,892) e o maior número de pares alvo/conjunto de dados em que ele é o método de melhor desempenho. Além disso, ele tem um desempenho mais robusto (diferença de 0,010 unidades de AUC entre as duas médias) quando o mesmo é aplicado nos sete conjunto de dados em que o sistemático é o melhor método (média = 0,902). Já o método sistemático (Média Geral = 0,874) tem um desempenho superior (0,936) quando o mesmo é o melhor método, mas sofre uma degradação substancial (0,076 unidades de AUC) quando aplicado nos treze conjuntos de dados em que o melhor método é o aleatório (média=0,860).

Com base nas análises aqui apresentadas, podemos concluir que o método aleatório tem o desempenho mais robusta, independente dos dados sobre os quais ele é aplicado. Já o método sistemático tem um desempenho melhor sobre dados de compostos do WOMBAT, uma base de dados mais completa que as outras utilizadas neste estudo, como indicam os resultados da

validação externa (Seção 4.3.2). Entretanto, a distribuição sistemática por similaridade média não é robusta para a maioria dos conjuntos de dados de moléculas, pois seu desempenho é inesperadamente pior que a do método aleatório. Isto pode ser revertido com uma melhor política de divisão dos grupos da validação cruzada, que garanta a homogeneidade entre os grupos mas também garanta uma heterogeneidade interna. Apesar disso, a distribuição sistemática foi a escolhida para a realização do estudo de validação externa na seção a seguir por ser a distribuição determinística, não dependendo de múltiplas execuções para a avaliação do desempenho do 3D-Pharma.

4.3.2 Validação Externa

Usando as fontes de dados descritas na seção 4.1 como referência para a construção de modelos e o conjunto de compostos ativos e *decoys* do DUD para o teste de desempenho, foi feito um estudo de validação externa do 3D-Pharma, bem como um comparativo de desempenho com outras técnicas que também usaram a base de dados do DUD como *benchmark*. Foram selecionados dez sistemas associados a alvos dentre os 40 presentes no DUD. Três contêm substâncias ativas com baixa variedade estrutural (menos de 15 classes estruturais): Aldose Redutase (ALR2), Receptor de Androgênio (AR) e Receptor Ativado por Proliferador de Peroxissomo γ (PPAR$_\gamma$). Os outros sete possuem moléculas ativas com alta diversidade estrutural (mais de 15 classes estruturais). São eles: Cinase dependente de Ciclina 2 (CDK2), Ciclooxigenase-2 (COX-2), Receptor de Fator de Crescimento Epidérmico (EGFR), Fator de Coagulação X α (FX$_\alpha$), Transcriptase Reversa de HIV-1 (HIVRT), Cinase Protéica Ativada por Mitogênio 14 (P38) e Fosfodiesterase V (PDE5). Estes dez sistemas foram escolhidos por terem dados de compostos ativos provindos do WOMBAT disponíveis através do site do DUD (`http://dud.docking.org`). A tabela 4.1 contém o tamanho de cada conjunto de dados associado aos alvos, bem como o número de classes estruturais para os compostos ativos do DUD.

Com o intuito de realizar uma validação verdadeiramente externa, dez modelos foram gerados e validados de acordo com o protocolo de validação da Seção 3.4 para cada conjunto externo de compostos ativos (Fármacos, Ligantes-PDB e WOMBAT) usando o valor de corte (τ) igual a 0,7. Três conjuntos de dados contêm menos de dez moléculas e não puderam passar pelo protocolo de validação: COX-2 Ligantes-PDB (4 moléculas), FX$_\alpha$ Fármacos (5 moléculas) e HIVRT

Fármacos (5 moléculas). Além disso, todos os conjuntos de dados de substâncias ativas para cada alvo tiveram um modelo simples (que não passou pelo protocolo de validação) gerado, usando $\tau = 0,7$. Cada modelo foi usado como referência de busca em um grupo de moléculas formado pelo DUD-Ativos e DUD-Decoys do respectivo alvo. Cada molécula teve sua similaridade com o modelo calculada através do Coeficiente de Tanimoto (Equação 3.1) e foi colocada em uma lista ordenada pelo valor de tal similaridade.

A acurácia de cada modelo foi medida através da área sob a curva ROC (AUC_{ROC}), tendo como conjunto positivo o conjunto DUD-Ativos e como conjunto negativo o conjunto DUD-Decoys. A capacidade de reconhecimento precoce foi medida através do $BEDROC_{\alpha}$. O parâmetro α assumiu os valores de $160,9$; $32,2$ e 20, que correspondem a seleções de 1%, 5% e 8% da lista ordenada, respectivamente (89). A capacidade de amostragem de diversidade estrutural (*scaffold hopping*) foi medida através da área sobre a curva ROC ponderada aritmeticamente (AUC_{awROC}), usando a informação de agrupamento estrutural dos DUD-Ativos disponível nos dados provindos do próprio DUD. Os valores das métricas apresentados nas sessões seguintes correspondem à média entre os dez modelos e é acompanhada pelo desvio padrão, exceto quando o conjunto de compostos ativos gerou apenas um modelo.

Ao mesmo tempo em que se realizou a validação externa, uma comparação foi feita com outras técnicas que usaram o DUD como uma base de medida de desempenho e disponibilizaram todos os dados para fins de comparação: o 4D FAP_{OA} (166); os métodos de atribuição ótima OAK e OAK_{FLEX} (168, 169), 2SHA e OAAP (179); o FieldScreen (48) e o FLAP (159–162). Com exceção do 4D FAP_{OA}, que realizou seus estudos de validação com todos os 40 sistemas do DUD, os autores das técnicas citadas apenas realizaram estudos de desempenho sobre o conjunto de sistemas com alta diversidade estrutural de compostos ativos. Treze dentre os 40 conjuntos de DUD-Ativos são considerados de alta diversidade estrutural, ou seja, as moléculas ativas se distribuem em mais de 15 classes estruturais (ou *scaffolds*) (48). A interseção entre os treze sistemas com alta diversidade estrutural de compostos ativos e os dez que possuem dados do WOMBAT disponíveis contém sete dos dez sistemas selecionados para este estudo (CDK2, COX-2, EGFR, FX_{α}, HIVRT, P38 e PDE5). Os dados de validação externa para os outros três sistemas (ALR2, AR e $PPAR_{\gamma}$) foram comparados apenas aos dados provenientes do 4D FAP_{OA}.

Aldose Redutase

A Aldose Redutase (ALR2) é uma oxirredutase envolvida no metabolismo de açúcares, especialmente conhecida por seu papel na degradação da glicose através da produção de sorbitol. Por ter menor afinidade com a glicose quando esta está em concentrações normais, a alr2 não participa ativamente do metabolismo do açúcar. Mas em alta concentração, como no caso da diabetes, parte da glicose em meio celular passa a ser degradada pela via do poliol, que envolve a enzima, o que pode estar associado a uma série de complicações da doença, como neuropatias, cegueira e falência renal. A Aldose Redutase não é um alvo terapêutico validado, mas possui alguns candidatos a fármaco em testes clínicos, como o Epalrestat e o Sulindac (Figura 4.4).

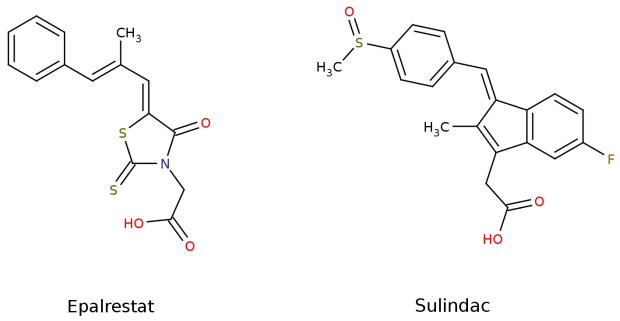


Figura 4.4: Fármacos em testes clínicos inibidores da Aldose Redutase: Epalrestat e Sulindac

A Figura 4.5 contém as curvas ROC para os três conjuntos de dados usados pelo 3D-Pharma e para o 4D FAP_{OA}. A Tabela 4.8 contém os dados numéricos do experimento. O 3D-Pharma usando os ligantes PDB e as moléculas do WOMBAT como referências produziram bons modelos, com AUC_{ROC} acima de 0.8, respectivamente, enquanto o 3D-Pharma usando os fármacos conhecidos e o 4D FAP_{OA} tiveram um desempenho razoável, com $AUCs_{ROC}$ acima de 0.6. Entretanto os modelos baseados nos Fármacos tiveram um desempenho melhor no reconhecimento precoce, atrás apenas dos ligantes PDB. Ao analisar o enriquecimento da amostragem de diversidade estrutural, o desempenho do 3D-Pharma destaca-se, principalmente ao observar o aumento das $AUCs_{awROC}$ em comparação às $AUCs_{ROC}$ da base de Fármacos. Os modelos simples têm desempenho igual aos modelos validados, com seus escores dentro da faixa de desvio padrão.

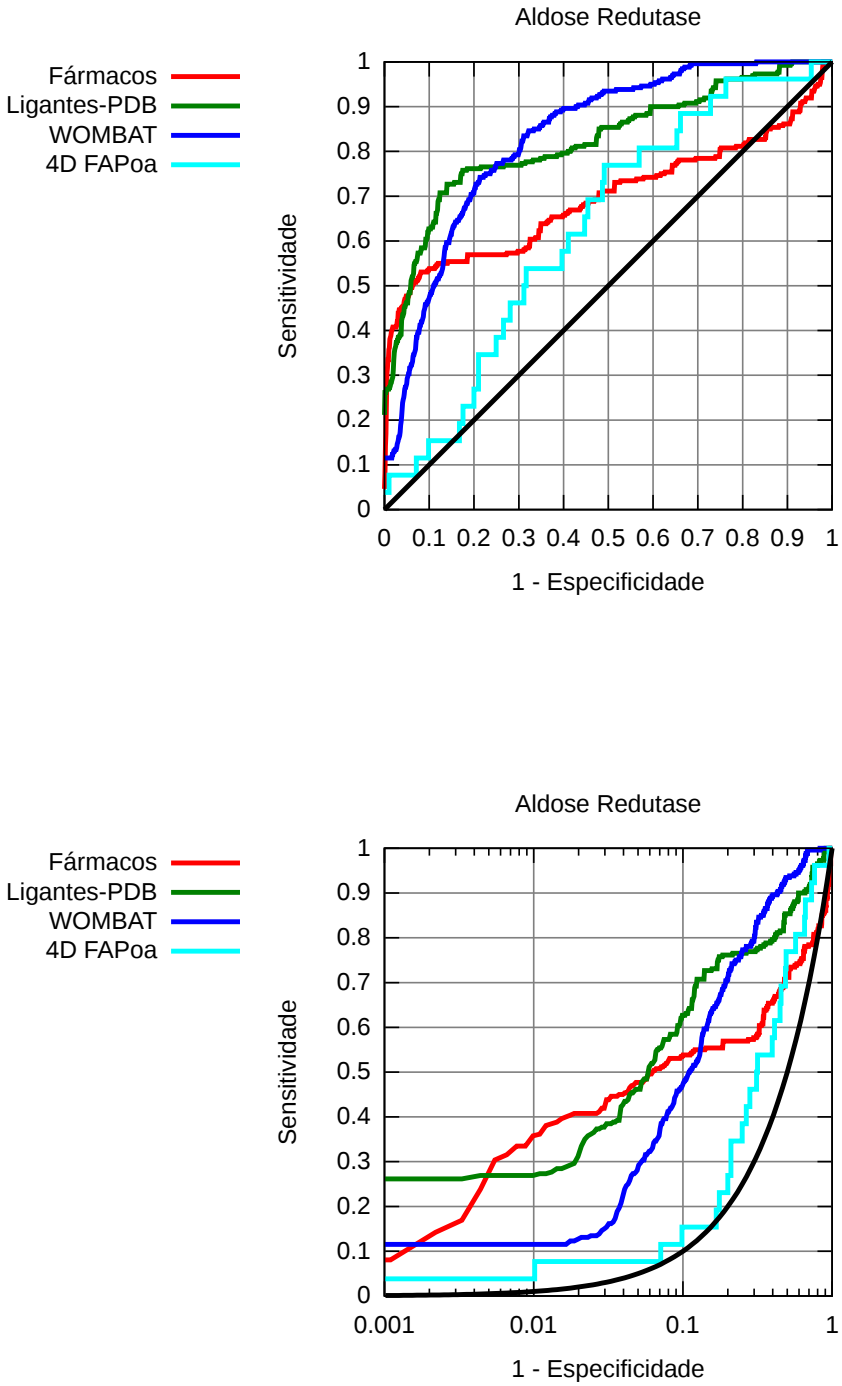


Figura 4.5: Curvas ROC linear e em escala logarítmica comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho do 4D FAP_{OA} para a Aldose Redutase.

	3D-Pharma		3D-Pharma Modelo Simples	4D FAP_{OA}
Desempenho Geral				
$AUC_{ROC} \pm$ dp.	Fármacos	0.69 ± 0.05	0.66	0.63
	Ligantes-PDB	0.82 ± 0.02	0.81	
	WOMBAT	$\mathbf{0.83 \pm 0.03}$	0.80	
Reconhecimento Precoce				
$BEDROC_{160.9} \pm$ dp.	Fármacos	0.54 ± 0.09	0.54	0.18
	Ligantes-PDB	$\mathbf{0.70 \pm 0.01}$	0.65	
	WOMBAT	0.41 ± 0.00	0.37	
$BEDROC_{32.2} \pm$ dp.	Fármacos	$\mathbf{0.48 \pm 0.09}$	0.44	0.10
	Ligantes-PDB	0.47 ± 0.02	0.46	
	WOMBAT	0.26 ± 0.03	0.25	
$BEDROC_{20} \pm$ dp.	Fármacos	0.48 ± 0.07	0.45	0.11
	Ligantes-PDB	$\mathbf{0.49 \pm 0.02}$	0.48	
	WOMBAT	0.31 ± 0.03	0.29	
Diversidade Estrutural				
$AUC_{awROC} \pm$ dp.	Fármacos	0.76 ± 0.04	0.74	0.59
	Ligantes-PDB	$\mathbf{0.86 \pm 0.02}$	0.84	
	WOMBAT	0.85 ± 0.02	0.83	

Tabela 4.8: Desempenho dos métodos em triagem virtual para a Aldose Redutase, usando os dados do DUD para validação Externa. O 3D-Pharma usando os Ligantes-PDB como referência teve melhor reconhecimento precoce, apesar dos modelos provindos das moléculas do WOMBAT terem a melhor taxa de recuperação em geral.

Receptor de Androgênio

O Receptor de Androgênio (AR - *Androgen Receptor*) é um receptor nuclear ligado ao DNA e regulador de fatores de transcrição gênica, ativado através da ligação dos hormônios androgênicos testosterona e di-hidrotestosterona. Os genes regulados pelo receptor estão envolvidos no desenvolvimento dos fenótipos sexuais masculinos, mas o receptor também é associado a fenótipos tipicamente femininos. O receptor de androgênios é um alvo terapêutico validado de uma série de fármacos, a maioria envolvida no tratamento de câncer de próstata com ação antagonista (Figura 4.6). Outras indicações terapêuticas de moduladores do Receptor de Androgênio abrangem desde reposição hormonal, aumento de massa muscular, tratamento de câncer de mama e dos sintomas da menopausa (agonistas) a contraceptivo feminino oral (antagonistas).

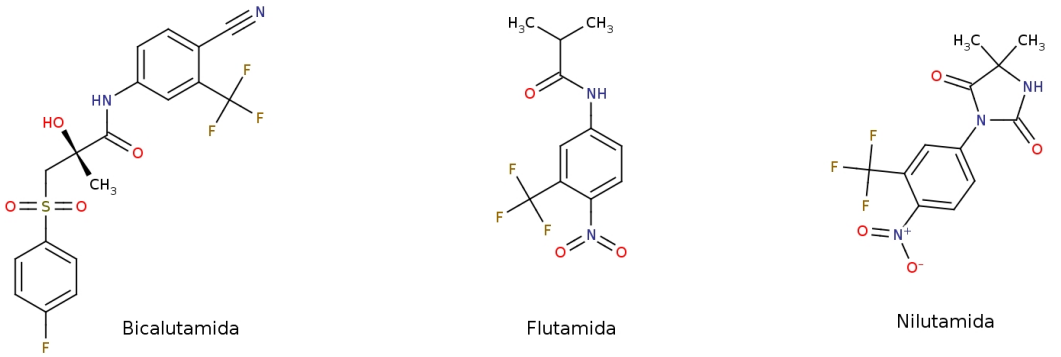


Figura 4.6: Fármacos com ação antagonista para o Receptor de Androgênio, voltados ao tratamento de câncer de próstata: Bicalutamida (Casodex), Flutamida (Eulexin) e Nilutamida (Anandron)

A Figura 4.7 contém as curvas ROC para os três conjuntos de dados usados pelo 3D-Pharma e para o 4D FAP_{OA}. A Tabela 4.9 contém os dados numéricos do experimento. Todas as técnicas e bases de dados produzem excelentes modelos, com AUC_{ROC} acima de 0.9. Porém, o 3D-Pharma tem um desempenho superior, com modelos acima de 0.98, enquanto o 4D FAP_{OA} chega a 0.92. Entretanto, ao analisar o reconhecimento precoce e a amostragem de diversidade estrutural, o 3D-Pharma se destaca, especialmente quando analisados os dados dos conjuntos Ligantes-PDB e WOMBAT. O 4D FAP_{OA} tem um desempenho muito inferior nos dois quesitos, o que

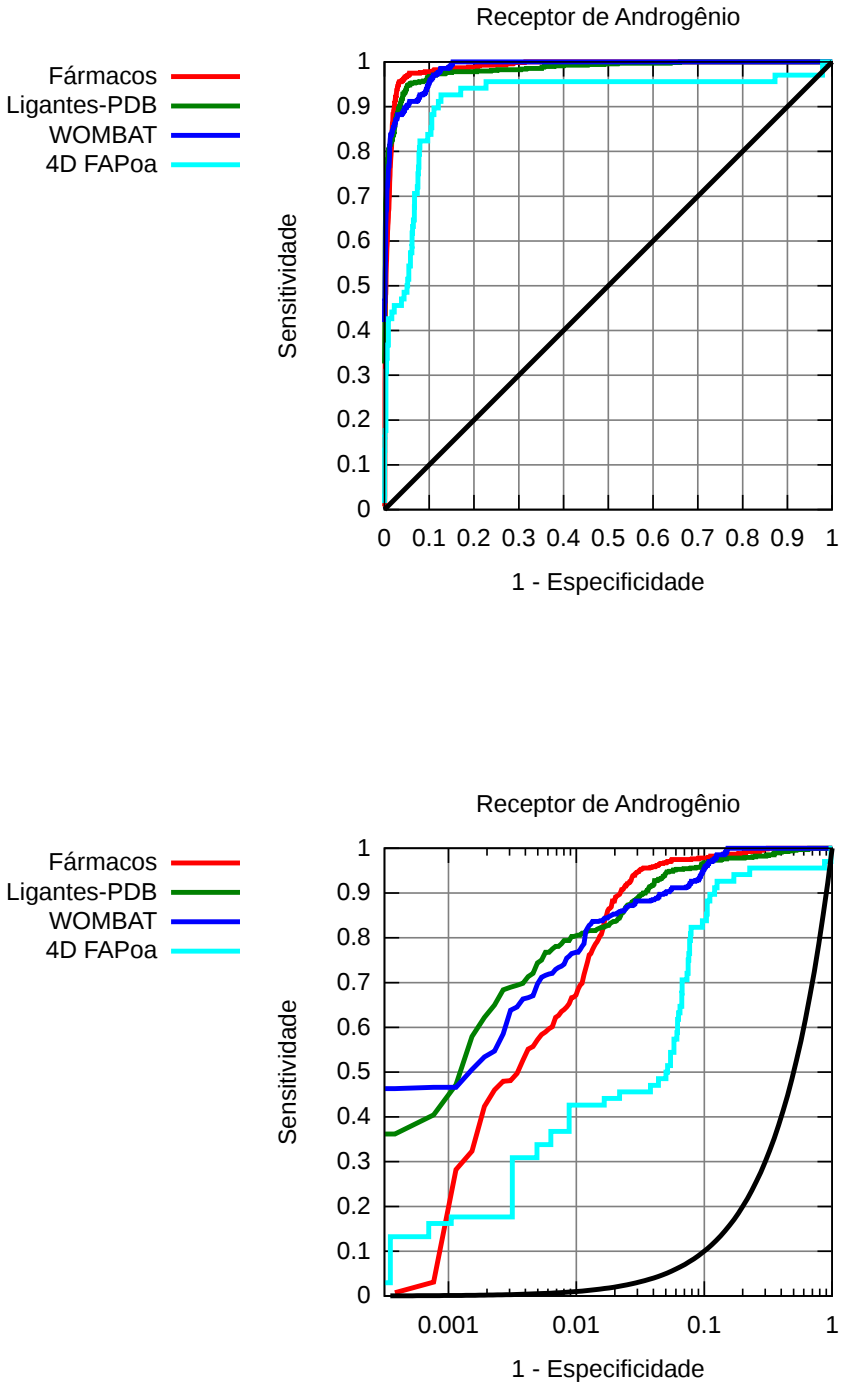


Figura 4.7: Curvas ROC linear e em escala logarítmica comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho do 4D FAP_{OA} para o Receptor de Androgênio.

	3D-Pharma		3D-Pharma Modelo Simples	4D FAP_{OA}
Desempenho Geral				
$AUC_{ROC} \pm$ dp.	Fármacos	$\mathbf{0.99 \pm 0.00}$	**0.99**	0.92
	Ligantes-PDB	0.98 ± 0.00	0.98	
	WOMBAT	$\mathbf{0.99 \pm 0.00}$	**0.99**	
Reconhecimento Precoce				
$BEDROC_{160.9} \pm$ dp.	Fármacos	0.76 ± 0.00	0.76	0.64
	Ligantes-PDB	0.93 ± 0.00	**0.94**	
	WOMBAT	0.93 ± 0.01	**0.94**	
$BEDROC_{32.2} \pm$ dp.	Fármacos	0.83 ± 0.02	0.82	0.54
	Ligantes-PDB	$\mathbf{0.88 \pm 0.01}$	0.87	
	WOMBAT	0.86 ± 0.00	0.87	
$BEDROC_{20} \pm$ dp.	Fármacos	0.87 ± 0.02	0.86	0.59
	Ligantes-PDB	$\mathbf{0.89 \pm 0.00}$	**0.89**	
	WOMBAT	0.88 ± 0.00	**0.89**	
Diversidade Estrutural				
$AUC_{awROC} \pm$ dp.	Fármacos	$\mathbf{0.99 \pm 0.00}$	**0.99**	0.77
	Ligantes-PDB	0.98 ± 0.00	0.98	
	WOMBAT	0.95 ± 0.00	0.96	

Tabela 4.9: Desempenho dos métodos em triagem virtual para o Receptor de Androgênio, usando os dados do DUD para validação Externa. 3D-Pharma e 4D FAP_{OA} tiveram desempenho próximos quando se analisa somente os dados de AUC_{ROC}, mas o método de Jahn et al. tem capacidade reduzida de reconhecimento precoce e de amostragem de diversidade estrutural.

não transparece quando analisada somente a curva ROC. Os modelos simples têm desempenho muito próximo da média dos modelos produzidos pelo protocolo de validação cruzada, mas em alguns casos eles se sobressaem acima da faixa de desvio padrão, mas ainda assim apenas um centésimo de unidade de AUC, como nos casos de $BEDROC_{32.2}$ e awROC usando as substâncias do WOMBAT.

Receptor γ Ativado por Proliferador de Peroxissomo

O Receptor γ Ativado por Proliferador de Peroxissomo (PPAR$_\gamma$ - *Peroxisome Proliferator-Activated Receptor* γ) é um receptor nuclear ligante-dependente que regula a transcrição de genes, como a Acetilcolina Oxidase, ligados ao metabolismo de ácidos graxos e da glicose, e à criação de células adiposas (adipogênese). Sua modulação tem sido associada à redução do nível de glicose no sangue sem aumentar a produção de insulina no pâncreas, o que torna o PPAR$_\gamma$ um alvo para fármacos relacionados ao tratamento de Diabetes tipo II (Figura 4.8). O papel do receptor na adipogênese também o torna alvo de fármacos relacionados ao tratamento de obesidade, mas tal ação ainda é experimental.

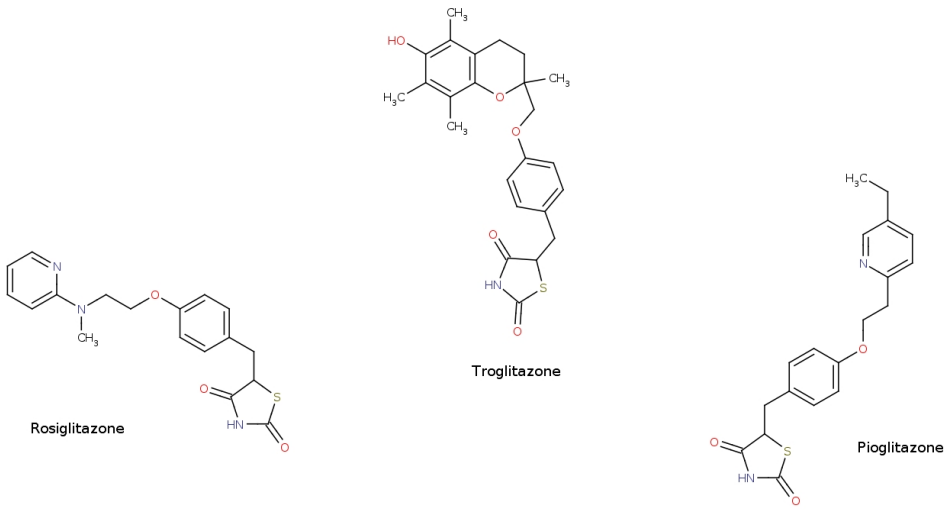


Figura 4.8: Fármacos usados no controle de Diabetes tipo II que têm como alvo o Receptor γ Ativado por Proliferador de Peroxissomo. Sua ação consiste em aumentar a sensibilidade à insulina nos tecidos muscular esquelético e adiposo: Troglitazone, retirado do mercado pelo risco de hepatotoxicidade e substituído por Rosiglitazone (Avandia) e Pioglitazone (Actos)

A Figura 4.9 contém as curvas ROC comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho do 4D FAP$_{OA}$ e a Tabela 4.10 contém os dados numéricos do experimento. Os dados do DUD para o PPAR$_\gamma$ têm uma peculiaridade: o baixo número de compostos inativos em relação ao número de compostos ativos, já que a maioria foi eliminada no filtro *lead-like* de Oprea (92) (6,33 substâncias inativas por substância ativa, contra uma relação de 521,17 compostos inativos por composto ativo na versão sem filtro). Isso causa uma anormalidade nos cálculos, especialmente nos números da BEDROC, onde os cortes iniciais podem considerar

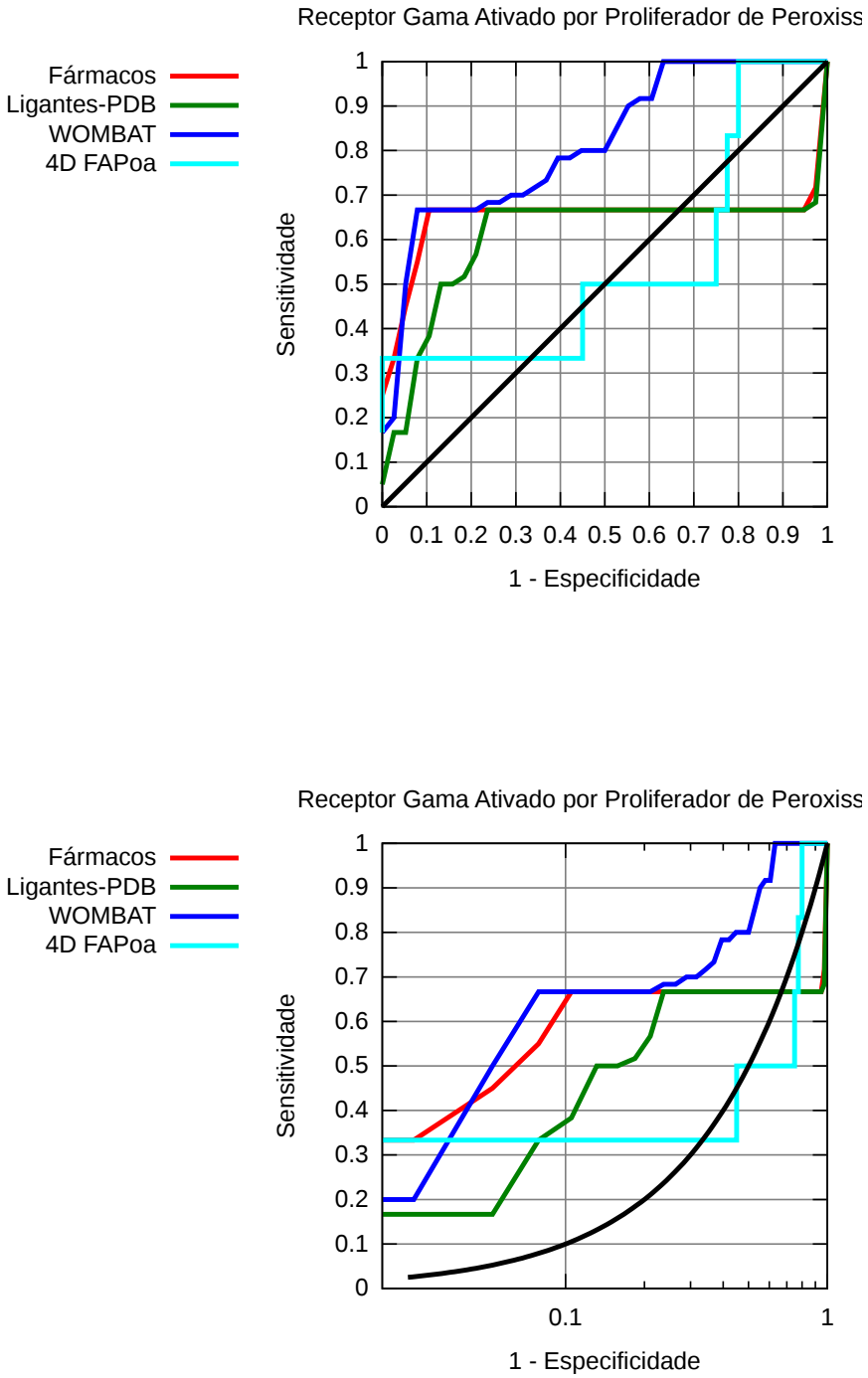


Figura 4.9: Curvas ROC linear e em escala logarítmica comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho do 4D FAP_{OA} para o Receptor γ Ativado por Proliferador de Peroxissomo

	3D-Pharma		3D-Pharma Modelo Simples	4D FAP_{OA}
Desempenho Geral				
$AUC_{ROC} \pm$ dp.	Fármacos	0.64 ± 0.01	0.63	0.54
	Ligantes-PDB	0.59 ± 0.00	0.59	
	WOMBAT	$\mathbf{0.81 \pm 0.03}$	0.75	
Reconhecimento Precoce				
$BEDROC_{160.9} \pm$ dp.	Fármacos	0.98 ± 0.01	0.97	**0.99**
	Ligantes-PDB	0.31 ± 0.43	0.03	
	WOMBAT	0.97 ± 0.00	0.97	
$BEDROC_{32.2} \pm$ dp.	Fármacos	0.73 ± 0.10	0.62	**0.76**
	Ligantes-PDB	0.37 ± 0.13	0.28	
	WOMBAT	0.63 ± 0.03	0.62	
$BEDROC_{20} \pm$ dp.	Fármacos	$\mathbf{0.66 \pm 0.09}$	0.57	0.63
	Ligantes-PDB	0.38 ± 0.07	0.33	
	WOMBAT	0.59 ± 0.02	0.58	
Diversidade Estrutural				
$AUC_{awROC} \pm$ dp.	Fármacos	0.64 ± 0.01	0.63	0.54
	Ligantes-PDB	0.59 ± 0.00	0.59	
	WOMBAT	$\mathbf{0.81 \pm 0.03}$	0.75	

Tabela 4.10: Desempenho dos métodos em triagem virtual para o Receptor γ Ativado por Proliferador de Peroxissomo, usando os dados do DUD para validação Externa. Devido à disparidade entre os números de DUD-Ativos e DUD-Decoys, os dados de BEDROC não são confiáveis para determinar o desempenho do modelo. Assim mesmo, o 3D-Pharma usando o WOMBAT obteve os melhores resultados.

menos de uma molécula (no caso do $BEDROC_{160.9}$). Apesar disso, a AUC_{ROC} é robusta o suficiente para ser considerada nesta análise. Assim, o 3D-Pharma usando WOMBAT foi o único a produzir bons modelos, com AUC_{ROC} acima de 0.8. A análise de amostragem de diversidade estrutural não tem significado com os dados do DUD para o $PPAR_{\gamma}$, já que o número de classes estruturais é igual ao número de compostos no conjunto DUD-Ativos, o que torna os valores de awROC iguais aos da curva ROC. Os modelos simples tiveram desempenho estatisticamente iguais aos modelos validados internamente, já que obtiveram marcas dentro da faixa de desvio padrão, exceto quando analisado os dados do WOMBAT, onde o modelo simples obteve uma acurácia menor.

Cinase dependente de Ciclina 2

A Cinase dependente de Ciclina 2 (CDK2 - *Cyclin-Dependent Kinase 2*) é uma transferase serina/treonina envolvida no controle do ciclo celular, essencial para a meiose mas dispensável na mitose. O ápice de sua expressão é na transição entre as fases G1 e S do ciclo celular, onde a duplicação do DNA é regulada. Acredita-se que a CDK2 tenha influência em diversos tipos de neoplasias: Leucemia Mielóide Aguda, Carcinomas Hepatocelulares, Câncer Nasofaríngeo, dentre outros. Não existem fármacos regulamentados que tenham a CDK2 como alvo, mas compostos como o Flavopiridol, Purvalanol e Staurosporina (Figura 4.10) estão sendo testados clinicamente.

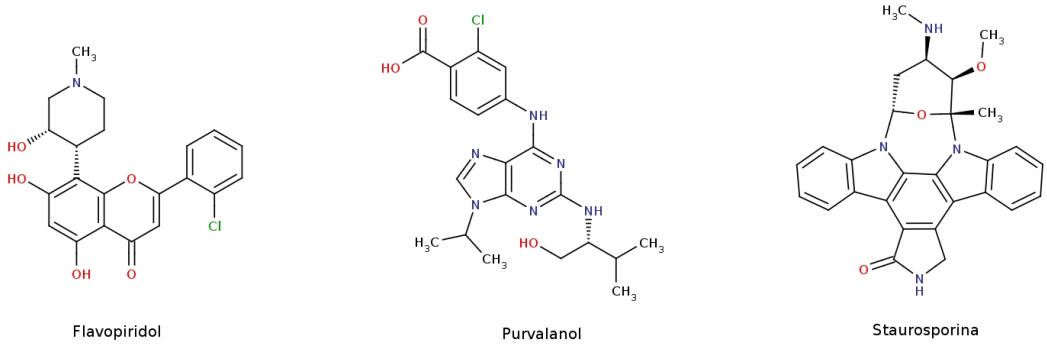
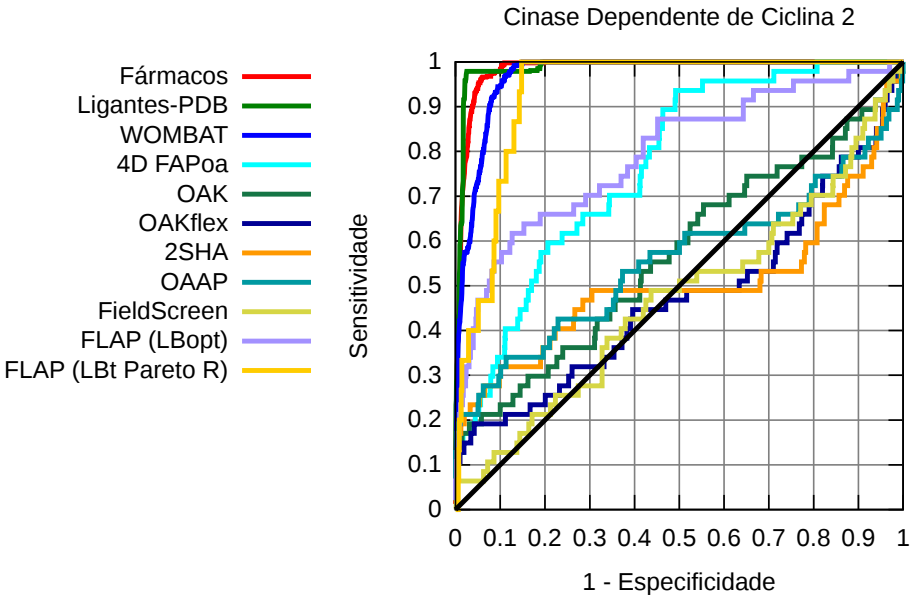


Figura 4.10: Compostos em testes clínicos que têm a CDK2 como alvo terapêutico: Flavopiridol, Purvalanol e Staurosporina

A Figura 4.11 contém as curvas ROC comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho das técnicas em triagem virtual incluídas neste estudo. A Tabela 4.11 contém os dados numéricos do experimento. O 3D-Pharma e o FLAP usando o modo LBt Pareto R produzem os melhores modelos, com $AUCs_{ROC}$ acima de 0.9. Entretanto, o desempenho do FLAP LBt Pareto R no reconhecimento imediato é bem aquém do esperado, inconsistente ao bom valor de AUC_{ROC} obtido pelo método. Apesar dos valores de AUC_{ROC} serem próximos, os valores de BEDROC das duas técnicas são bem diferentes, destacando a melhor capacidade do 3D-Pharma de posicionar os compostos ativos no topo da busca. Entre os diferentes conjuntos de dados, os Ligantes-PDB obtiveram consistentemente a melhor desempenho. Os modelos simples



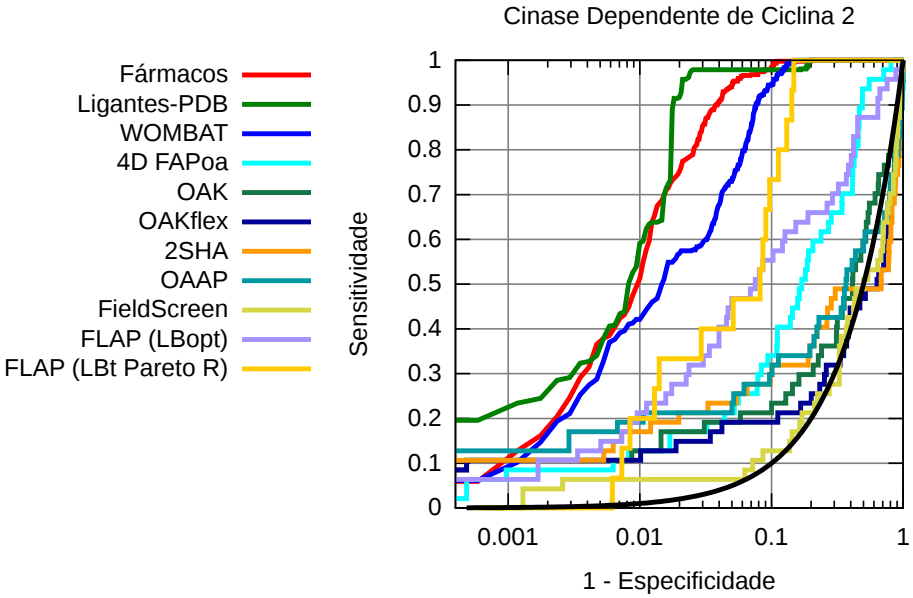


Figura 4.11: Curvas ROC linear e em escala logarítmica comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho dos outros métodos de triagem virtual para a Cinase dependente de Ciclina 2

	3D-Pharma		3D-Pharma Modelo Simples	OAK	OAK_{FLEX}	2SHA	OAAP	FieldScreen	FLAP (LBopt)	FLAP (LBt Pareto R)	4D FAP_{OA}
Desempenho Geral											
$AUC_{ROC} \pm$ dp.	Fármacos	0.98 ± 0.00	**0.99**	0.57	0.47	0.50	0.55	0.47	0.79	0.93	0.77
	Ligantes-PDB	$\mathbf{0.99 \pm 0.00}$	**0.99**								
	WOMBAT	0.97 ± 0.00	0.97								
Reconhecimento Precoce											
$BEDROC_{160.9} \pm$ dp.	Fármacos	0.66 ± 0.02	0.67	0.34	0.33	0.39	0.46	0.20	0.44	0.12	0.29
	Ligantes-PDB	$\mathbf{0.74 \pm 0.01}$	0.73								
	WOMBAT	0.63 ± 0.01	0.64								
$BEDROC_{32.2} \pm$ dp.	Fármacos	0.75 ± 0.02	0.75	0.21	0.19	0.26	0.28	0.10	0.38	0.33	0.22
	Ligantes-PDB	$\mathbf{0.80 \pm 0.01}$	**0.80**								
	WOMBAT	0.63 ± 0.01	0.64								
$BEDROC_{20} \pm$ dp.	Fármacos	0.80 ± 0.03	0.81	0.22	0.19	0.27	0.28	0.11	0.42	0.42	0.25
	Ligantes-PDB	$\mathbf{0.85 \pm 0.00}$	**0.85**								
	WOMBAT	0.69 ± 0.02	0.69								
Diversidade Estrutural											
$AUC_{awROC} \pm$ dp.	Fármacos	0.98 ± 0.01	0.98	0.55	0.46	0.48	0.53	0.44	0.82	0.94	0.80
	Ligantes-PDB	$\mathbf{0.99 \pm 0.00}$	**0.99**								
	WOMBAT	0.96 ± 0.00	0.97								

Tabela 4.11: Desempenho dos métodos em triagem virtual para a Cinase dependente de Ciclina 2, usando os dados do DUD para validação Externa. O FLAP LBt Pareto R teve desempenho similar ao 3D-Pharma quando se analisa a acurácia e a amostragem de diversidade estrutural, mas a técnica têm uma capacidade de reconhecimento imediato mediana.

tiveram desempenhos dentro da faixa de desvio padrão dos modelos validados internamente.

Ciclooxigenase-2

A Ciclooxigenase-2 (COX-2) ou Prostaglandina G/H Sintase 2 é uma enzima envolvida nas respostas inflamatórias, ao produzir prostaglandinas a partir do ácido araquidônico. As ciclooxigenases estão envolvidas em uma série de processos bioquímicos, e sua inibição farmacológica tem efeitos analgésicos, anti-inflamatórios e antipiréticos. Historicamente, anti-inflamatórios não-esteroidais inibiam tanto a Ciclooxigenase-1 quanto a COX-2, entretanto, devido ao fato da COX-1 estar presente em praticamente todos os tecidos, sua inibição não seletiva acarretava uma série de efeitos colaterais, notadamente no trato gastro-intestinal. Com a comercialização dos fármacos seletivos para COX-2, os chamados *coxib* (Figura 4.12), a maior parte dos efeitos adversos relacionados à inibição da COX-1 deixaram de ser observados em larga escala. Entretanto, esta classe de medicamentos tem sido associada ao aumento do risco de infarto do miocárdio e de acidente vascular cerebral, com alguns representantes sendo retirados do mercado.

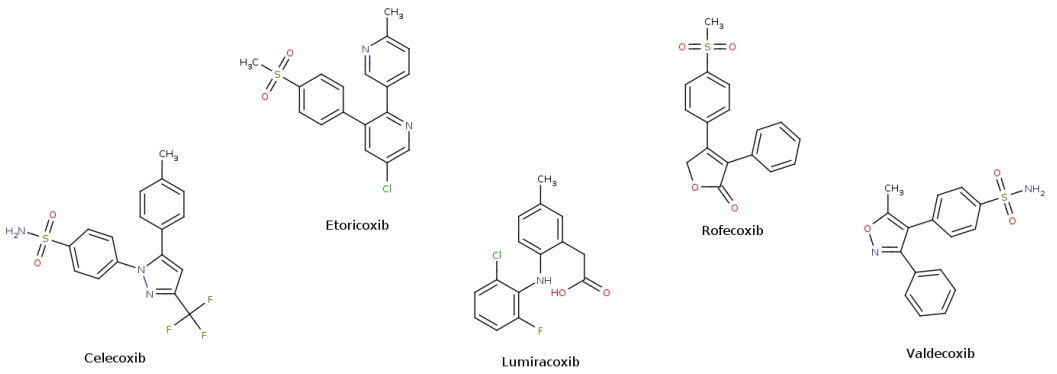


Figura 4.12: Compostos *coxib*, que apresentam inibição seletiva para COX-2: Celecoxib (Celebra), Etoricoxib (Arcoxia), Lumiracoxib (Prexige), Rofecoxib (Vioxx) e Valdecoxib (Bextra)

A Figura 4.13 contém as curvas ROC comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho das técnicas em triagem virtual incluídas neste estudo. Já a Tabela 4.12 contém os dados numéricos do experimento. Todas as técnicas construíram bons modelos, com $AUCs_{ROC}$ acima de 0.8, exceto o 3D-Pharma usando os Ligantes-PDB. Isto pode ser explicado pela escassez de dados de ligantes provindos do PDB, já que a Ciclooxigenase-2 humana possui apenas nove estruturas depositadas (até 28 de Fevereiro de 2012), sendo que apenas

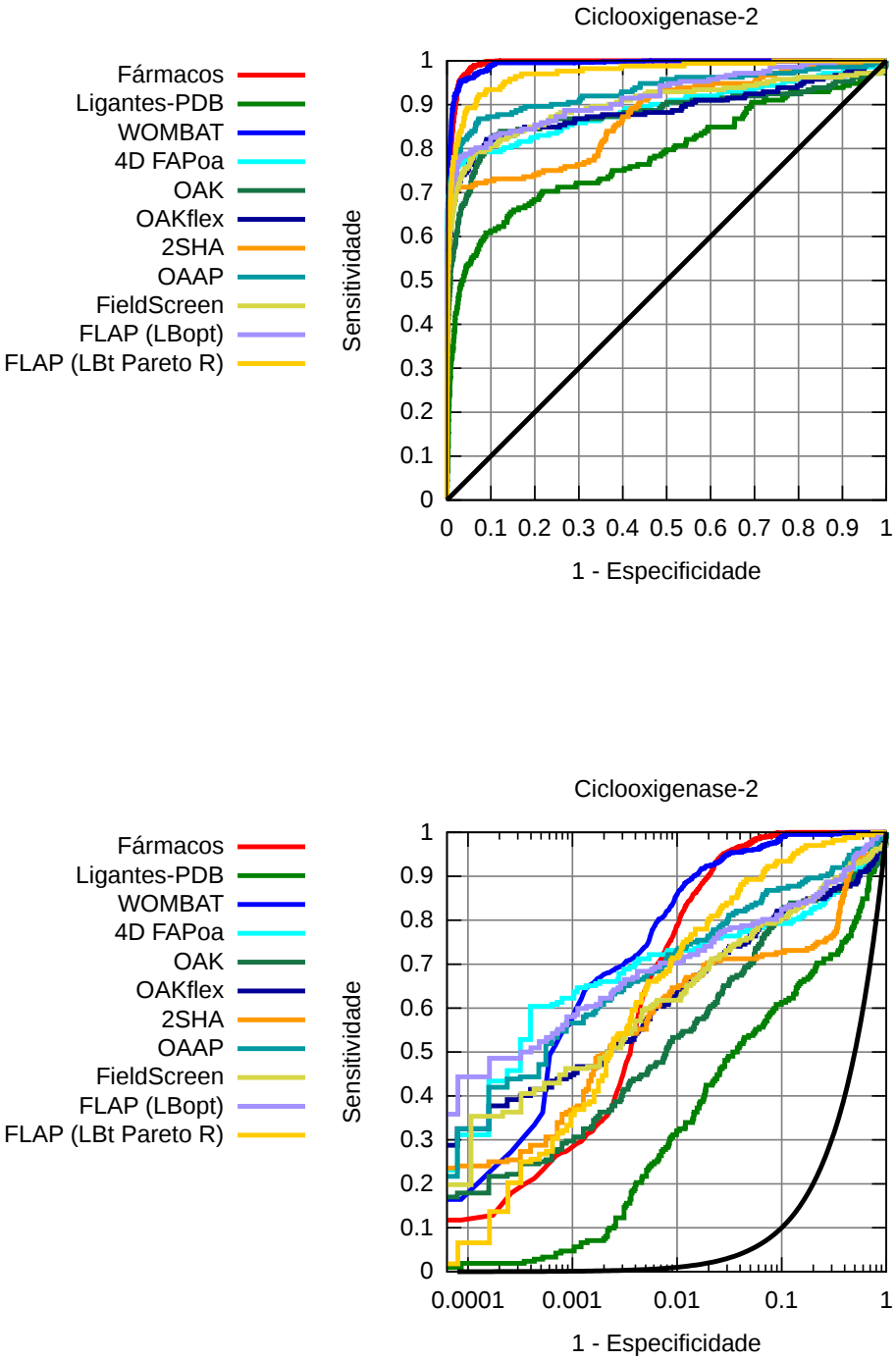


Figura 4.13: Curvas ROC linear e em escala logarítmica comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho dos outros métodos de triagem virtual para a Ciclooxigenase-2

	3D-Pharma		3D-Pharma Modelo Simples	OAK	OAK_{FLEX}	2SHA	OAAP	FieldScreen	FLAP (LBopt)	FLAP (LBt Pareto R)	4D FAP_{OA}
Desempenho Geral											
$AUC_{ROC} \pm$ dp.	Fármacos	$\mathbf{0.99 \pm 0.00}$	**0.99**	0.88	0.89	0.87	0.93	0.90	0.92	0.97	0.89
	Ligantes-PDB	0.78	0.78								
	WOMBAT	$\mathbf{0.99 \pm 0.00}$	**0.99**								
Reconhecimento Precoce											
$BEDROC_{160.9} \pm$ dp.	Fármacos	0.79 ± 0.02	0.77	0.72	0.82	0.80	0.87	0.88	0.88	0.76	0.89
	Ligantes-PDB	0.42	0.42								
	WOMBAT	$\mathbf{0.90 \pm 0.00}$	0.89								
$BEDROC_{32.2} \pm$ dp.	Fármacos	0.86 ± 0.01	0.83	0.65	0.73	0.71	0.80	0.75	0.79	0.80	0.79
	Ligantes-PDB	0.46	0.46								
	WOMBAT	$\mathbf{0.90 \pm 0.00}$	**0.90**								
$BEDROC_{20} \pm$ dp.	Fármacos	0.89 ± 0.01	0.87	0.69	0.75	0.71	0.82	0.76	0.79	0.83	0.79
	Ligantes-PDB	0.50	0.50								
	WOMBAT	$\mathbf{0.92 \pm 0.00}$	**0.92**								
Diversidade Estrutural											
$AUC_{awROC} \pm$ dp.	Fármacos	$\mathbf{0.99 \pm 0.00}$	0.98	0.77	0.77	0.79	0.87	0.79	0.82	0.94	0.80
	Ligantes-PDB	0.75	0.75								
	WOMBAT	0.98 ± 0.00	0.98								

Tabela 4.12: Desempenho dos métodos em triagem virtual para a Ciclooxigenase-2, usando os dados do DUD para validação Externa. Todas as técnicas têm desempenhos parecidos, com exceção do 3D-Pharma PDB. Entretanto, o melhor desempenho é atingida pelo 3D-Pharma usando os Fármacos e os compostos ativos do WOMBAT.

quatro ligantes foram co-cristalizados. 3D-Pharma (Fármacos e WOMBAT), OAAP, FieldScreen e FLAP (LBOpt e LBt Pareto R) construíram modelos com AUCs acima de 0.9, sendo que os modelos do 3D-Pharma tiveram o melhor desempenho. Quando se analisa a capacidade de amostragem de diversidade estrutural, apenas o 3D-Pharma e o FLAP LBt Pareto R obtiveram $AUCs_{awROC}$ acima de 0.9. Ao analisarmos o reconhecimento imediato, com uma taxa de falso positivos de 1%, as técnicas 3D-Pharma WOMBAT, OAAP, OAK_{FLEX}, FieldScreen, FLAP LBopt e 4D FAP_{OA} têm valores de $BEDROC_{160.9}$ bem próximos. Mas a partir daí, o 3D-Pharma Fármacos e WOMBAT começam a se destacar, mantendo seus altos valores de BEDROC enquanto que os das outras técnicas diminuem. Os modelos simples estão sempre dentro da faixa de desvio padrão dos modelos validados, exceto no caso do conjunto Fármacos, com valores de $BEDROC_{32.2}$ e $BEDROC_{20}$, onde o desempenho dos modelos simples é relativamente pior.

Receptor de Fator de Crescimento Epidérmico

O Receptor de Fator de Crescimento Epidérmico (EGFR - *Epidermal Growth Factor Receptor*) é uma Tirosina-Cinase com um domínio extracelular que se liga ao Fator de Crescimento Epidérmico (EGF). Quando o fator se liga, a cinase dispara uma cascata de sinalizações intracelulares através de fosforilações, que resulta no aumento da atividade mitótica e da proliferação das células. Esta função biológica é importante, por exemplo, para a recuperação de úlceras nas mucosas gastrointestinais. A EGFR é também considerada um oncogene, ou seja, mutações, superexpressão ou superatividade deste receptor são considerados fatores precursores de processos cancerígenos, especialmente neoplasias pulmonares, intestinais e epiteliais. Assim, o EGFR é alvo de quimioterapias anticâncer, que incluem anticorpos monoclonais que visam o bloqueio do sítio receptor do EGF, ou pequenas moléculas (Figura 4.14) que visam o bloqueio do sítio do ATP no domínio citoplasmático (responsável pela atividade de Cinase).

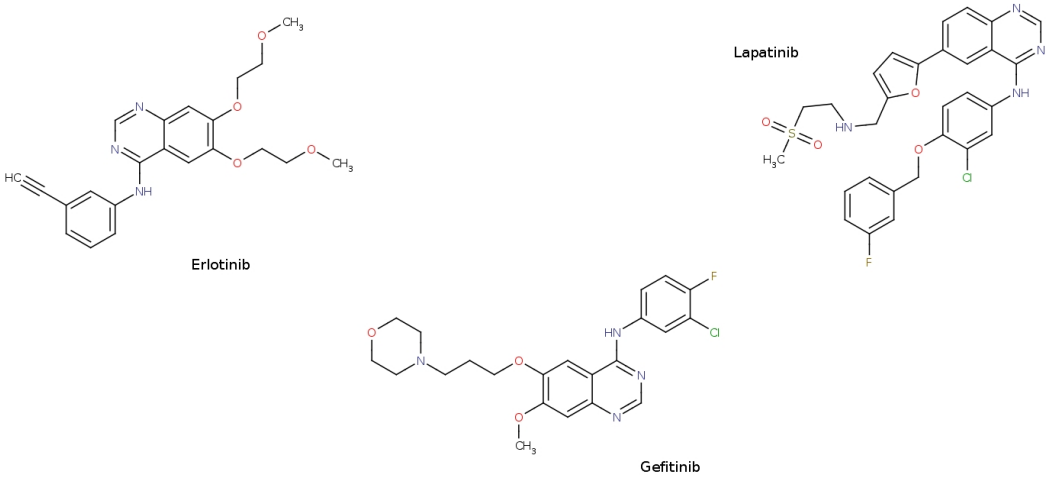
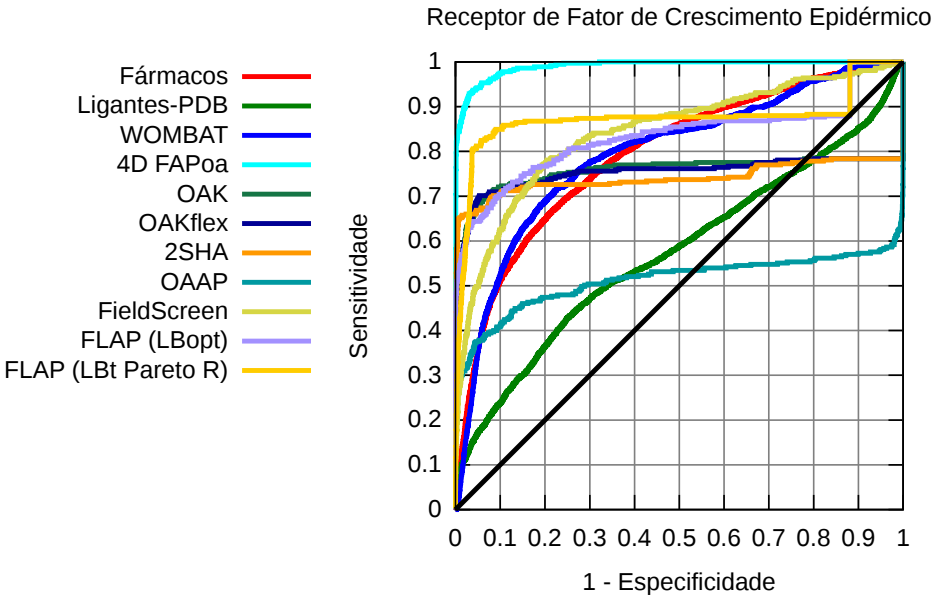


Figura 4.14: Fármacos usados em quimioterapias para tratamento de cânceres sólidos, como os de pulmão, pâncreas e mamário, que têm como alvo o Receptor de Fator de Crescimento Epidérmico: Erlotinib (Tarceva), Gefitinib (Iressa) e Lapatinib (Tycerb)

A Figura 4.15 contém as curvas ROC comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho das técnicas em triagem virtual incluídas neste estudo. Já a Tabela 4.13 contém os dados numéricos do experimento. Este é um sistema problemático para o 3D-Pharma, que obteve desempenho abaixo de técnicas como o FLAP, FieldScreen e o 4D FAP_{OA}. Este último, inclusive, obteve o melhor desempenho em todas as métricas. A se



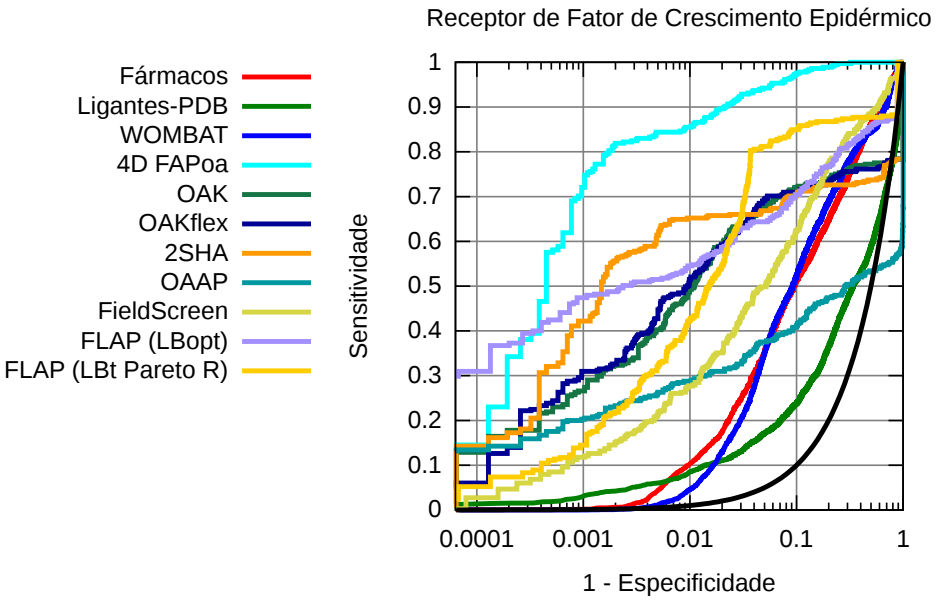


Figura 4.15: Curvas ROC linear e em escala logarítmica comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho dos outros métodos de triagem virtual para o Receptor de Fator de Crescimento Epidérmico

	3D-Pharma		3D-Pharma Modelo Simples	OAK	OAK_{FLEX}	2SHA	OAAP	FieldScreen	FLAP (LBopt)	FLAP (LBt Pareto R)	4D FAP_{OA}
Desempenho Geral											
$AUC_{ROC} \pm$ dp.	Fármacos	0.79 ± 0.02	0.78	0.76	0.75	0.74	0.51	0.84	0.83	0.87	**0.99**
	Ligantes-PDB	0.57 ± 0.03	0.59								
	WOMBAT	0.79 ± 0.01	0.79								
Reconhecimento Precoce											
$BEDROC_{160.9} \pm$ dp.	Fármacos	0.18 ± 0.03	0.16	0.76	0.78	0.87	0.61	0.58	0.87	0.63	**0.95**
	Ligantes-PDB	0.21 ± 0.02	0.25								
	WOMBAT	0.09 ± 0.01	0.08								
$BEDROC_{32.2} \pm$ dp.	Fármacos	0.28 ± 0.03	0.22	0.65	0.65	0.72	0.40	0.48	0.68	0.64	**0.92**
	Ligantes-PDB	0.17 ± 0.01	0.18								
	WOMBAT	0.24 ± 0.02	0.23								
$BEDROC_{20} \pm$ dp.	Fármacos	0.33 ± 0.03	0.28	0.68	0.66	0.71	0.40	0.51	0.67	0.69	**0.93**
	Ligantes-PDB	0.19 ± 0.01	0.20								
	WOMBAT	0.31 ± 0.02	0.30								
Diversidade Estrutural											
$AUC_{awROC} \pm$ dp.	Fármacos	0.72 ± 0.03	0.75	0.72	0.71	0.72	0.49	0.82	0.78	0.85	**0.98**
	Ligantes-PDB	0.76 ± 0.01	0.77								
	WOMBAT	0.66 ± 0.00	0.67								

Tabela 4.13: Desempenho dos métodos em triagem virtual para o Receptor de Fator de Crescimento Epidérmico, usando os dados do DUD para validação Externa. O 3D-Pharma não teve bom desempenho para este conjunto de dados, sendo ofuscado pelo excelente desempenho do 4D FAP_{OA}

destacar sobre o desempenho do 3D-Pharma é o salto na amostragem de diversidade estrutural dos modelos baseados nos Ligantes-PDB quando comparado à AUC_{ROC}. Os modelos simples tiveram o mesmo desempenho ruim dos modelos validados, com valores dentro das faixas de desvio padrão, exceto nas análises $BEDROC_{32.2}$ e $BEDROC_{20}$, onde o modelo simples baseado no conjunto Fármacos teve uma taxa de reconhecimento precoce pior do que os modelos que passaram pelo protocolo de validação cruzada.

Fator de Coagulação Xα

O Fator de Coagulação Xα (FX$_\alpha$) é uma serino-endopeptidase envolvida na cadeia de coagulação sanguínea, sintetizada no fígado com a presença obrigatória da Vitamina K. Quando ativada, a enzima cliva a protrombina em dois pontos, transformando-a em trombina e dando sequência ao processo de coagulação. O Fator Xα pode ser desativado pela Antitrombina III (AT III). A inibição da enzima é de interesse farmacêutico para o tratamento de condições trombóticas ou para se evitar a coagulação em procedimentos cirúrgicos, e pode ser feita de três maneiras: compostos que mimetizam a Vitamina K e impedem a sintetização do fator, inibição indireta através de compostos como a heparina e derivados (Enoxaparina e Fondaparinux) que potencializam a ação da AT III e compostos que inibem diretamente a enzima (Figura 4.16).

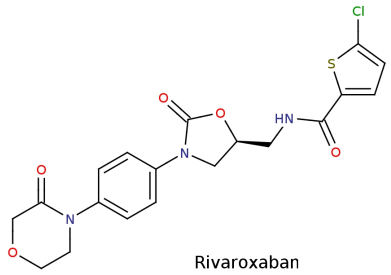


Figura 4.16: A estrutura do Rivaroxaban, o primeiro fármaco comercializado que inibe diretamente o Fator de Coagulação Xα.

A Figura 4.17 contém as curvas ROC comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho das técnicas em triagem virtual incluídas neste estudo. Já a Tabela 4.14 contém os dados numéricos do experimento. O modelo produzido pelo 3D-Pharma usando as moléculas do WOMBAT teve um desempenho praticamente perfeito, posicionando todos os compostos ativos antes do primeiro inativo na busca. O FLAP e os outros modelos do 3D-Pharma vieram logo depois, produzindo modelos com AUC$_{\text{ROC}}$ entre 0.88 e 0.94, mas o 3D-Pharma Ligantes-PDB e o FLAP LBopt apresentaram melhor reconhecimento precoce, com ligeira vantagem para o 3D-Pharma. As outras técnicas não produziram modelos com boa acurácia. Mais uma vez, as bases de Fármacos e Ligantes-PDB tiveram um grande salto ao priorizar o reconhecimento das classes estrurais, como mostra o salto da AUC$_{\text{awROC}}$ em relação a AUC$_{\text{ROC}}$. Um destaque deve ser dado ao desempenho do 3D-Pharma Fármacos, que conseguiu AUC$_{\text{ROC}} = 0.88$ e a impressionante marca de AUC$_{\text{awROC}} = 0.96$ com apenas cinco moléculas de

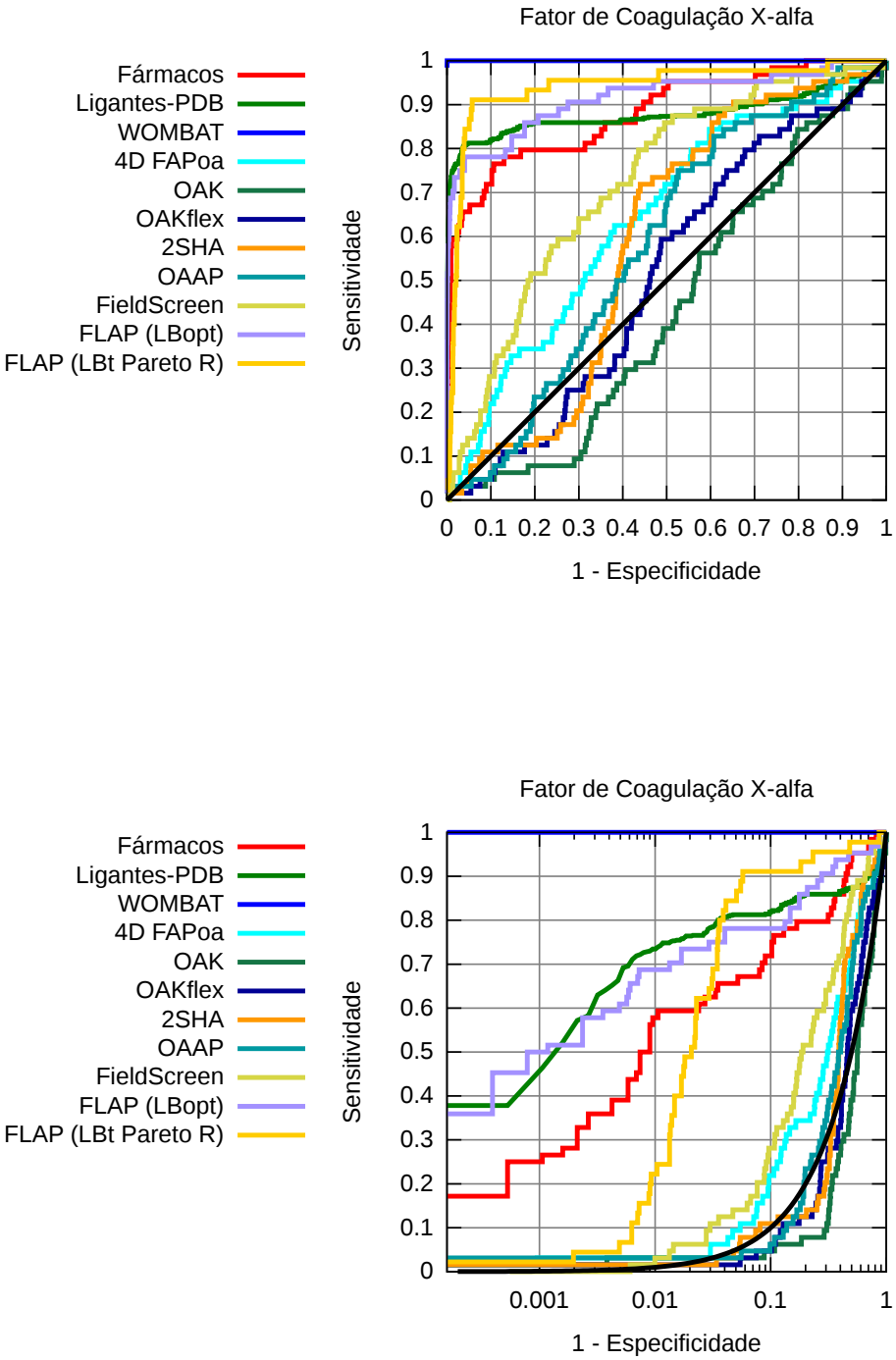


Figura 4.17: Curvas ROC linear e em escala logarítmica comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho dos outros métodos de triagem virtual para o Fator de Coagulação Xα

	3D-Pharma		3D-Pharma Modelo Simples	OAK	OAK_{FLEX}	2SHA	OAAP	FieldScreen	FLAP (LBopt)	FLAP (LBt Pareto R)	4D FAP_{OA}
Desempenho Geral											
$AUC_{ROC} \pm$ dp.	Fármacos Ligantes-PDB WOMBAT	0.88 0.88 ± 0.01 $\mathbf{1.00 \pm 0.00}$	0.88 0.86 **1.00**	0.43	0.51	0.59	0.58	0.72	0.92	0.94	0.64
Reconhecimento Precoce											
$BEDROC_{160.9} \pm$ dp.	Fármacos Ligantes-PDB WOMBAT	0.85 0.95 ± 0.00 $\mathbf{1.00 \pm 0.00}$	0.85 0.94 **1.00**	0.11	0.07	0.07	0.14	0.07	0.80	0.19	0.13
$BEDROC_{32.2} \pm$ dp.	Fármacos Ligantes-PDB WOMBAT	0.69 0.83 ± 0.01 $\mathbf{1.00 \pm 0.00}$	0.69 0.81 **1.00**	0.05	0.03	0.05	0.06	0.13	0.74	0.52	0.09
$BEDROC_{20} \pm$ dp.	Fármacos Ligantes-PDB WOMBAT	0.69 0.82 ± 0.00 $\mathbf{1.00 \pm 0.00}$	0.69 0.80 **1.00**	0.05	0.04	0.07	0.06	0.17	0.76	0.63	0.13
Diversidade Estrutural											
$AUC_{awROC} \pm$ dp.	Fármacos Ligantes-PDB WOMBAT	0.96 0.95 ± 0.00 $\mathbf{1.00 \pm 0.00}$	0.96 0.94 **1.00**	0.46	0.50	0.56	0.57	0.74	0.82	0.92	0.62

Tabela 4.14: Desempenho dos métodos em triagem virtual para o Fator de Coagulação Xα, usando os dados do DUD para validação Externa. O 3D-Pharma WOMBAT teve uma recuperação de compostos ativos perfeita, com $AUC_{ROC} = 1.00$. Outro destaque deve ser feito ao 3D-Pharma Fármacos, que com somente cinco moléculas de referência, conseguiu uma $AUC_{awROC} = 0.96$.

referência. Os modelos simples obtiveram desempenhos dentro das faixas de desvio padrão dos modelos validados, com a única exceção sendo a análise $BEDROC_{20}$ para o modelo baseado nos Ligantes-PDB.

Transcriptase Reversa de HIV-1

A transcriptase Reversa é uma das enzimas principais no processo de infecção do HIV, já que transforma o RNA viral em cDNA, que virá a ser integrado ao DNA da célula hospedeira, e por isso é alvo de inibidores presentes nos coquetéis anti-HIV. Os inibidores de Transcriptase Reversa se dividem em três classes: análogos de nucleosídeos, análogos de nucletídeos e os não-nucleosídeos. As duas primeiras classes competem com os nucleotídeos, impedindo a formação da cadeia de DNA pela enzima. A terceira classe consiste de inibidores não-competitivos, que impedem o funcionamento mecânico da proteína (Figura 4.18).

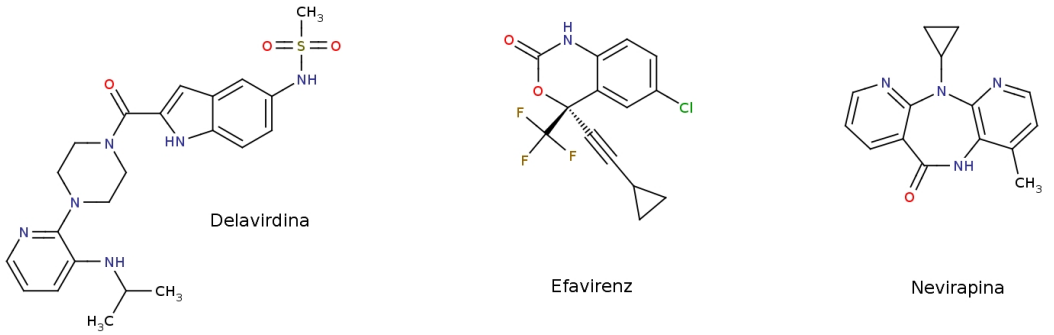


Figura 4.18: Inibidores não-competitivos (Não-nucleosídeos) de Transcriptase Reversa de HIV-1, usados no tratamento de infecções por HIV-1: Delavirdina, Efavirenz e Nevirapina.

A Figura 4.19 contém as curvas ROC comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho das técnicas em triagem virtual incluídas neste estudo. Já a Tabela 4.15 contém os dados numéricos do experimento. O 3D-Pharma apresenta um desempenho muito superior a das outras técnicas, inclusive para o conjunto de dados de Fármacos, que com apenas quatro compostos de referência conseguiu um $\mathrm{AUC}_{\mathrm{ROC}} = 0.87$, mas teve uma queda quando analisada a $\mathrm{AUC}_{\mathrm{awROC}}$. Os dados do WOMBAT geraram o melhor modelo, mas os Ligantes-PDB obtiveram um desempenho bem próximo. Os modelos simples apresentaram desempenho equivalente (dentro da faixa de desvio padrão) aos modelos validados.

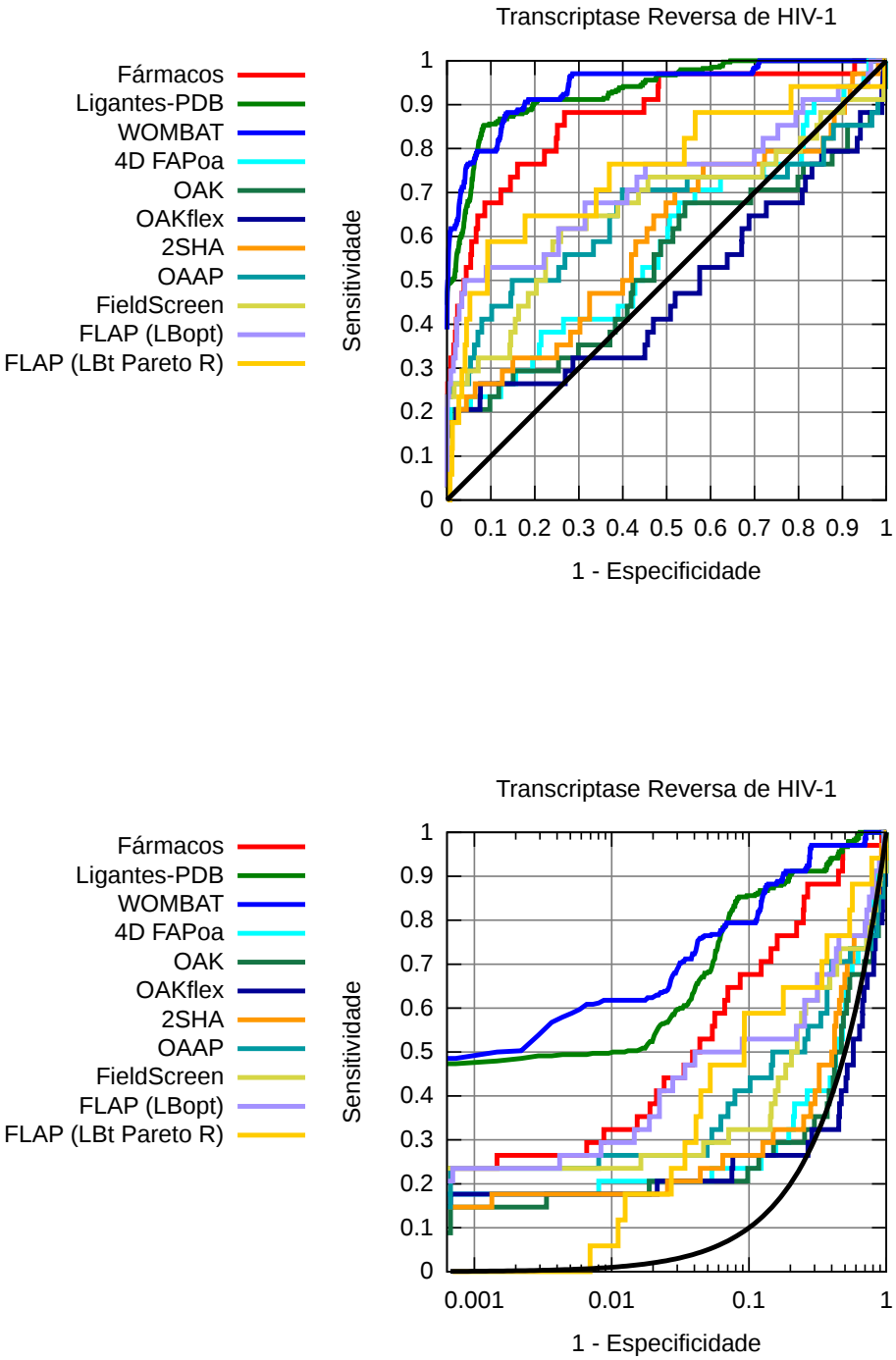


Figura 4.19: Curvas ROC linear e em escala logarítmica comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho dos outros métodos de triagem virtual para a Transcriptase Reversa de HIV-1

	3D-Pharma		3D-Pharma Modelo Simples	OAK	OAK_{FLEX}	2SHA	OAAP	FieldScreen	FLAP (LBopt)	FLAP (LBt Pareto R)	4D FAP_{OA}
Desempenho Geral											
$AUC_{ROC} \pm$ dp.	Fármacos	0.87	0.87	0.53	0.48	0.60	0.65	0.66	0.72	0.75	0.58
	Ligantes-PDB	0.93 ± 0.00	0.93								
	WOMBAT	$\mathbf{0.94 \pm 0.00}$	**0.94**								
Reconhecimento Precoce											
$BEDROC_{160.9} \pm$ dp.	Fármacos	0.68	0.68	0.45	0.48	0.47	0.58	0.62	0.63	0.07	0.50
	Ligantes-PDB	0.87 ± 0.00	0.87								
	WOMBAT	$\mathbf{0.89 \pm 0.00}$	**0.89**								
$BEDROC_{32.2} \pm$ dp.	Fármacos	0.51	0.51	0.25	0.26	0.26	0.35	0.34	0.46	0.25	0.27
	Ligantes-PDB	0.68 ± 0.01	0.68								
	WOMBAT	$\mathbf{0.74 \pm 0.01}$	**0.74**								
$BEDROC_{20} \pm$ dp.	Fármacos	0.54	0.54	0.24	0.25	0.26	0.36	0.34	0.47	0.32	0.26
	Ligantes-PDB	0.70 ± 0.01	0.71								
	WOMBAT	$\mathbf{0.75 \pm 0.00}$	**0.75**								
Diversidade Estrutural											
$AUC_{awROC} \pm$ dp.	Fármacos	0.82	0.82	0.53	0.47	0.53	0.65	0.56	0.69	0.62	0.62
	Ligantes-PDB	0.90 ± 0.01	0.90								
	WOMBAT	0.92 ± 0.00	**0.93**								

Tabela 4.15: Desempenho dos métodos em triagem virtual para o Transcriptase Reversa de HIV-1, usando os dados do DUD para validação Externa. Mesmo com apenas quatro compostos, o conjunto Fármacos conseguiu bons modelos, mas os modelos baseados nos Ligantes-PDB e nas moléculas do WOMBAT tiveram um desempenho superior às outras técnicas.

Cinase Protéica 14 Ativada por Mitogênio

A Cinase Protéica 14 Ativada por Mitogênio (MAPK14 - *Mitogen Activated Protein Kinase 14*, ou simplesmente p38) é uma cinase envolvida na via de transdução de sinal de mitogênio, usualmente envolvida em resposta a estímulos de estresse externos, como citocinas, irradiação ultra-violeta, choque térmico ou choque osmótico. A MAPK14 fosforila um grande conjunto de proteínas e fatores transcricionais e, apesar de ainda ser um alvo terapêutico experimental, sua modulação está associada ao tratamento de diversas disfunções, como Doença Pulmonar Obstrutiva Crônica (DCOP), depressão, doenças cardiovasculares e Síndrome da Angústia Respiratória do Adulto (SARA) (Figura 4.20).

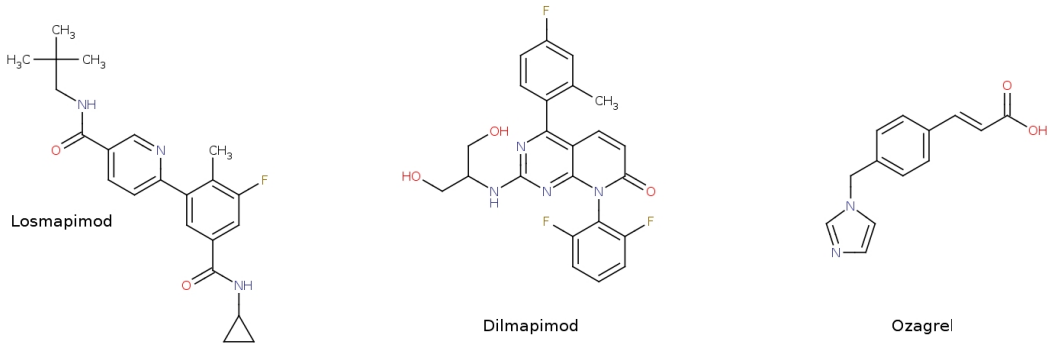


Figura 4.20: Candidatos a fármacos para a modulação da Cinase Protéica 14 Ativada por Mitogênio: Losmapimod (Fase Clínica II para tratamento de DCOP, Depressão e doenças cardiovasculares), Dilmapimod (Fase Clínica I para SARA) e Ozagrel (Fase Clínica I para DCOP)

A Figura 4.21 contém as curvas ROC comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho das técnicas em triagem virtual incluídas neste estudo. Já a Tabela 4.16 contém os dados numéricos do experimento. Com exceção dos modelos gerados à partir dos Ligantes-PDB, os modelos do 3D-Pharma obtiveram a melhor acurácia em geral, conseguindo $AUCs_{ROC}$ acima de 0.9, especialmente o 3D-Pharma WOMBAT, que obteve um $AUC_{ROC} = 0.97$. Entretanto, o FLAP LBopt teve o melhor escore $BEDROC_{160.9}$, mas o reconhecimento precoce do mesmo perde força quando os maiores cortes são considerados. Observa-se que para todos os métodos (exceto OAAP, FieldScreen e 4D FAP_{OA}), há uma melhora na AUC_{awROC},

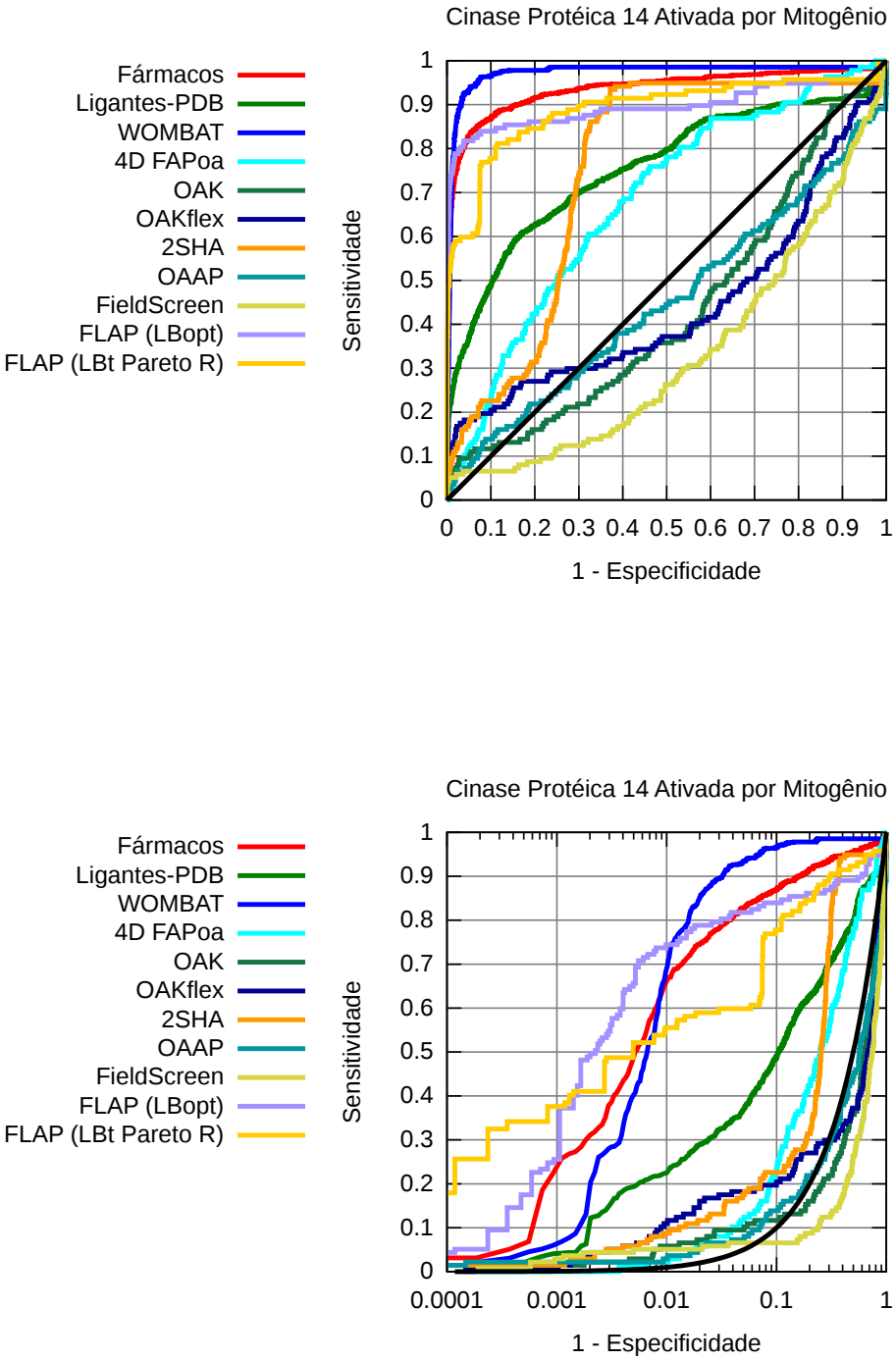


Figura 4.21: Curvas ROC linear e em escala logarítmica comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho dos outros métodos de triagem virtual para a Cinase Protéica 14 Ativada por Mitogênio

	3D-Pharma		3D-Pharma Modelo Simples	OAK	OAK_{FLEX}	2SHA	OAAP	FieldScreen	FLAP (LBopt)	FLAP (LBt Pareto R)	4D FAP_{OA}
Desempenho Geral											
$AUC_{ROC} \pm$ dp.	Fármacos	0.94 ± 0.02	0.95	0.43	0.44	0.75	0.45	0.33	0.89	0.89	0.68
	Ligantes-PDB	0.75 ± 0.01	0.75								
	WOMBAT	$\mathbf{0.97 \pm 0.00}$	**0.97**								
Reconhecimento Precoce											
$BEDROC_{160.9} \pm$ dp.	Fármacos	0.76 ± 0.05	0.78	0.10	0.19	0.17	0.08	0.16	**0.79**	0.69	0.05
	Ligantes-PDB	0.44 ± 0.01	0.44								
	WOMBAT	0.65 ± 0.08	0.61								
$BEDROC_{32.2} \pm$ dp.	Fármacos	0.75 ± 0.06	0.80	0.10	0.18	0.16	0.09	0.08	0.66	0.54	0.10
	Ligantes-PDB	0.37 ± 0.01	0.37								
	WOMBAT	$\mathbf{0.79 \pm 0.02}$	0.78								
$BEDROC_{20} \pm$ dp.	Fármacos	0.78 ± 0.06	**0.83**	0.10	0.19	0.18	0.09	0.07	0.65	0.54	0.14
	Ligantes-PDB	0.40 ± 0.02	0.40								
	WOMBAT	$\mathbf{0.83 \pm 0.02}$	**0.83**								
Diversidade Estrutural											
$AUC_{awROC} \pm$ dp.	Fármacos	0.98 ± 0.01	**0.99**	0.47	0.49	0.76	0.43	0.27	0.95	0.91	0.68
	Ligantes-PDB	0.80 ± 0.01	0.80								
	WOMBAT	$\mathbf{0.99 \pm 0.00}$	**0.99**								

Tabela 4.16: Desempenho dos métodos em triagem virtual para a Cinase Protéica 14 Ativada por Mitogênio, usando os dados do DUD para validação Externa. O 3D-Pharma obterve o melhor desempenho, exceto para os Ligantes-PDB. Entretanto, o FLAP LBopt obteve o melhor reconhecimento precoce a 1%, como mostra seu escore $BEDROC_{160.9}$.

indicando melhor distribuição das classes estruturais entre os compostos ativos do DUD. Os modelos simples apresentaram desempenho equivalente (dentro da faixa de desvio padrão) aos modelos validados.

Fosfodiesterase V

Fosfodiesterases são um grupo de enzimas que têm como substrato nucleotídeos cíclicos, como o AMPc (Adenosina Monofosfato Cíclico) e o GMPc (Guanosina Monofosfato Cíclico), e tem um papel importante em transdução de sinais que envolvem os nucelotídeos cíclicos. Dentre a família de PDEs (*PhosphoDiEsterase*), a PDE5 (ou Fosfodiesterase GMPc-específica 3',5'-cíclica) tem um interesse clínico relevante devido à sua maior concentração no tecido do corpo cavernoso peniano, sendo sua modulação o alvo de fármacos que visam reverter quadros de disfunção erétil. Fármacos com maior seletividade para a PDE5 (Figura 4.22, como o Sildenafil (Viagra), Vardenafil (Levitra) e Tadalafil (Cialis), obtêm os efeitos desejados contra a disfunção erétil com efeitos adversos minimizados quando comparados aos inibidores menos seletivos.

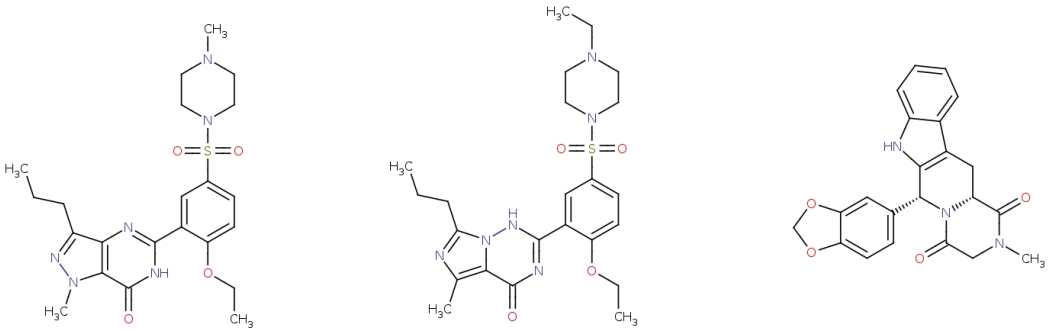


Figura 4.22: Fármacos com alta seletividade para a Fosfodiesterase V, indicados para homens com disfunção erétil: Sildenafil (Viagra), Vardenafil (Levitra) e Tadalafil (Cialis)

A Figura 4.23 contém as curvas ROC comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho das técnicas em triagem virtual incluídas neste estudo. Já a Tabela 4.17 contém os dados numéricos do experimento. Dos modelos produzidos pelo 3D-Pharma, apenas os que usaram as moléculas do WOMBAT como referência tiveram um bom desempenho. As substâncias presentes nas bases de Fármacos e dos Ligantes-PDB parecem não cobrir alguma classe de substâncias ativas, o que pode explicar a recuperação mais lenta dos verdadeiros positivos. Já as moléculas do WOMBAT geram um modelo quase ideal: apenas quatro compostos inativos foram recuperados antes de todos os 26 compostos ativos para a PDE5.

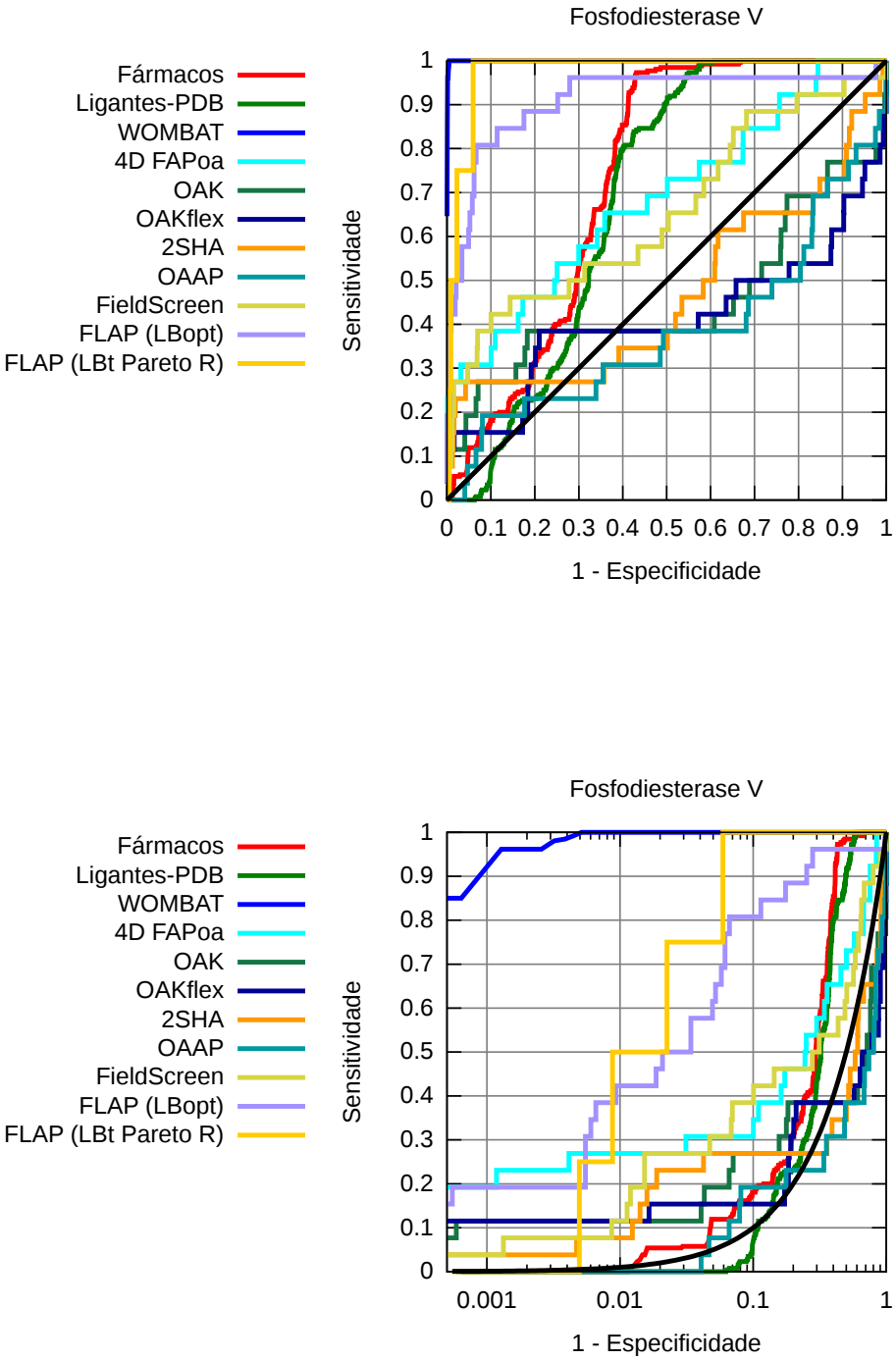


Figura 4.23: Curvas ROC linear e em escala logarítmica comparativas entre as três bases de dados usadas pelo 3D-Pharma e o desempenho dos outros métodos de triagem virtual para a Fosfodiesterase V

	3D-Pharma		3D-Pharma Modelo Simples	OAK	OAK_{FLEX}	2SHA	OAAP	FieldScreen	FLAP (LBopt)	FLAP (LBt Pareto R)	4D FAP_{OA}
Desempenho Geral											
$AUC_{ROC} \pm$ dp.	Fármacos	0.73 ± 0.04	0.71	0.46	0.41	0.47	0.38	0.66	0.91	0.98	0.69
	Ligantes-PDB	0.69 ± 0.01	0.57								
	WOMBAT	$\mathbf{1.00 \pm 0.00}$	**1.00**								
Reconhecimento Precoce											
$BEDROC_{160.9} \pm$ dp.	Fármacos	0.02 ± 0.01	0.00	0.26	0.27	0.17	0.00	0.24	0.41	0.20	0.49
	Ligantes-PDB	0.00 ± 0.00	0.00								
	WOMBAT	$\mathbf{0.98 \pm 0.00}$	0.97								
$BEDROC_{32.2} \pm$ dp.	Fármacos	0.07 ± 0.03	0.04	0.18	0.17	0.21	0.04	0.27	0.41	0.57	0.33
	Ligantes-PDB	0.01 ± 0.00	0.00								
	WOMBAT	$\mathbf{0.99 \pm 0.00}$	0.98								
$BEDROC_{20} \pm$ dp.	Fármacos	0.10 ± 0.04	0.07	0.19	0.17	0.23	0.07	0.30	0.45	0.68	0.33
	Ligantes-PDB	0.03 ± 0.01	0.00								
	WOMBAT	$\mathbf{0.99 \pm 0.00}$	**0.99**								
Diversidade Estrutural											
$AUC_{awROC} \pm$ dp.	Fármacos	0.75 ± 0.04	0.73	0.37	0.32	0.38	0.35	0.62	0.94	0.96	0.63
	Ligantes-PDB	0.71 ± 0.01	0.60								
	WOMBAT	$\mathbf{1.00 \pm 0.00}$	**1.00**								

Tabela 4.17: Desempenho dos métodos em triagem virtual para a Fosfodiesterase V, usando os dados do DUD para validação Externa. O 3D-Pharma WOMBAT obteve o melhor desempenho, com $AUC_{ROC} \approx 1.00$, enquanto os outros conjuntos de dados obtiveram desempenhos bem piores.

O excelente desempenho do FLAP também deve ser destacado, com $AUCs_{ROC}$ acima de 0.9 e bons valores de BEDROC. Como os 26 compostos ativos do DUD se distribuem em 22 classes estruturais, o desempenho mensurado pela AUC_{awROC} não é muito diferente da AUC_{ROC}. Os modelos simples tiveram um desempenho próximo dos modelos validados, exceto o produzido pelos Ligantes-PDB, que obteve um desempenho pior.

Análise Geral do desempenho do 3D-Pharma sobre a base de dados do DUD

Observando as $AUCs_{ROC}$ médias de todos os métodos (Figura 4.24) para os dez sistemas (no caso do 3D-Pharma e do 4D FAP_{OA}) ou os sete sistemas com alta diversidade estrutural de compostos ativos do DUD (técnicas restantes), pode-se ver que o desempenho do 3D-Pharma usando os dados do WOMBAT tem a melhor média de AUC entre todos os métodos (0.93), seguida pelo FLAP LBt Pareto R (0.91), FLAP LBopt (0.86), 3D-Pharma Fármacos (0.85), 3D-Pharma-PDB (0.80) e 4D FAP_{OA} (0.73). As técnicas restantes têm médias inferiores a 0.7. Apesar do melhor resultado, não se pode afirmar com significância estatística que o 3D-Pharma é melhor que o FLAP, por exemplo. De acordo com Nicholls (90), é preciso realizar pelo menos 100 experimentos (no caso, 100 conjuntos diferentes de dados) para se afirmar que um método é melhor que outro com uma confiança de 95%. Infelizmente, não há bases de dados de *benchmark* que possuam tantos conjuntos de compostos ativos/inativos de alvos terapêuticos, e fazer tal afirmação baseado em sete experimentos, como foi apresentado aqui, é subestimar a contribuição do erro em cada experimento. De fato, o desvio padrão de cada método (não mostrado nos gráficos das Figuras 4.24 e 4.25) é, em média, 0,1 unidades de AUC, tornando qualquer tentativa de se determinar o melhor método uma afirmação sem significância estatística. Entretanto, ao observar que o 3D-Pharma obtém sempre o melhor desempenho (exceto no EGFR), pode-se esperar que esta tendência se confirme em outros experimentos. Além disso, há de ser considerado o fato que o FLAP usa os dados dos compostos presentes no conjunto DUD-Parents como referência, o qual é um subconjunto do DUD-Ativos. Esta escolha pode melhorar artificialmente os resultados.

O mesmo pode-se dizer ao analisar a capacidade de diversificação estrutural (ou *scaffold hopping*), já que os valores de AUC_{awROC} (Figura 4.24) são proporcionais aos valores de AUC_{ROC}. Entretanto, uma observação interessante pode ser feita em relação às duas métricas. Em todos os métodos, há uma pequena mas constante queda dos valores médios de AUC_{awROC} em relação aos de AUC_{ROC}, exceto para os conjuntos de dados Fármacos e Ligantes-PDB, em que o valor de AUC_{awROC} aumenta. Isso pode ser explicado pela natureza dos dados provindos das bases de dados de fármacos e do PDB, ou seja, as substâncias ativas destas fontes de dados provavelmente não sofrem do "viés do análogo", ao contrário da série de dados do DUD e do WOMBAT, as quais aparentam sofrer o viés.

Já ao analisar o reconhecimento precoce (Figura 4.25), percebe-se claramente a disparidade

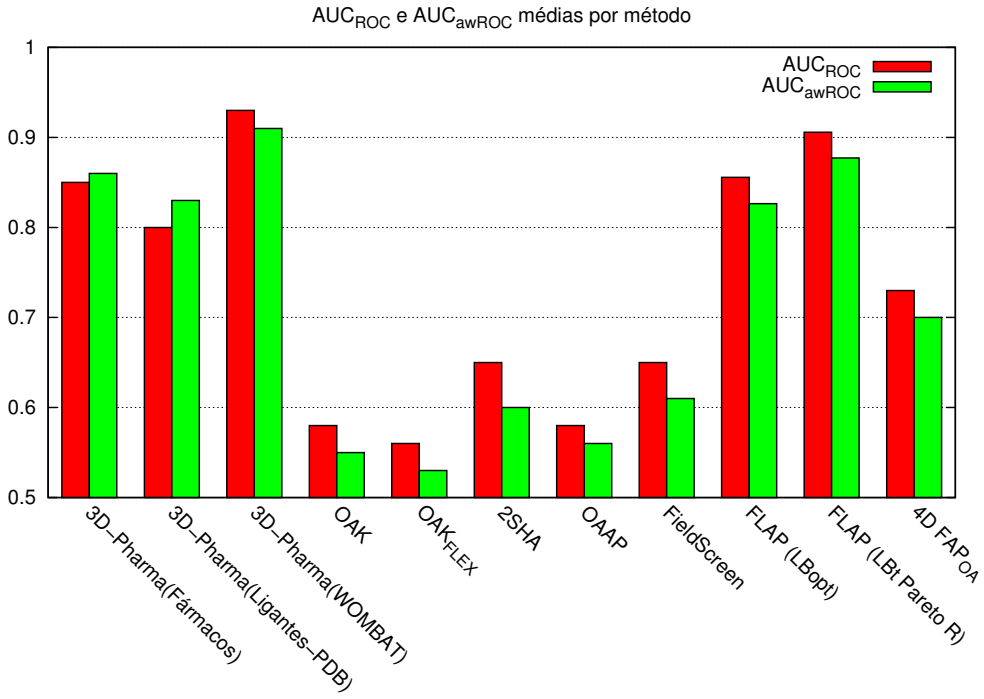


Figura 4.24: AUC$_{ROC}$ e AUC$_{awROC}$ médias para cada método. As médias do 3D-Pharma e do 4D FAP$_{OA}$ foram calculadas sobre os dez conjuntos de dados usados na validação externa, enquanto que a média das demais técnicas foram calculadas sobre os sete sistemas com alta diversidade estrutural. É possível perceber que os métodos 3D-Pharma e FLAP têm um alto desempenho, com AUCs$_{ROC}$ médias acima de 0.8, seguidos pelo 4D FAP$_{OA}$, com média acima de 0.7. As demais técnicas obtiveram médias inferiores.

dos valores obtidos pelo FLAP LBt Pareto R. Apesar de ter a segunda melhor AUC_{ROC} média dentre os métodos investigados (perdendo apenas para o 3D-Pharma WOMBAT), os valores de $BEDROC_{\alpha}$ estão muito aquém dos apresentados pelo 3D-Pharma e pelo FLAP LBopt, especialmente ao considerar os cortes mais restritivos ($\alpha = 160.9$ e 32.2, correspondentes a enriquecimentos em 1% e 5% respectivamente). Novamente considerando as análises de Nicholls (90), não podemos afirmar se um método é melhor que outro com significância estatística. Mas pode-se observar que o 3D-Pharma WOMBAT tem os melhores valores médios de $BEDROC_{\alpha}$ para todos os cortes (0.74/0.70/0.72), seguido pelo 3D-Pharma Fármacos (0.62/0.59/0.61) e pelo FLAP LBopt (0.69/0.59/0.60). Uma observação interessante: os outros métodos têm um valor médio de $BEDROC_{160.9}$ maior que os outros cortes ($BEDROC_{32.2}$ e $BEDROC_{20}$), mas o 3D-Pharma tende a manter os três valores razoavelmente iguais, o que contribui para o melhor desempenho do método quando analisados os valores de AUC_{ROC}, ou seja, o 3D-Pharma continua a posicionar melhor os compostos ativos em relação aos inativos mesmo após o primeiro corte de 1%, ao contrário das outras técnicas, que tendem a misturar moléculas ativas e inativas após este corte.

Vale frisar que os modelos produzidos pelo 3D-Pharma foram construídos a partir de três conjuntos de dados completamente externos ao DUD, ao contrário do FLAP, que usou os compostos do DUD-Parents (incluídos no DUD-Ativos) como refeência. As outras técnicas usaram dados de ligantes do PDB, os quais não são necessariamente os mesmos contidos no conjunto Ligantes-PDB usado pelo 3D-Pharma.

Um fator determinante para o bom desempenho do 3D-Pharma é o tamanho dos conjuntos de compostos ativos usados para a geração do modelo, como mostra a Figura 4.26. Há uma percepção imediata que quanto maior o conjunto de dados usado como referência pelo 3D-Pharma, maior é a probabilidade dos modelos construídos a partir dos dados obter um bom poder preditivo. De fato, dos 30 grupos de modelos construídos a partir dos pares alvo/conjunto de moléculas ativas, 14 (46,6%) obtiveram AUC_{ROC} acima de 0.9 e todos eles foram construídos a partir de conjuntos com pelo menos 13 compostos. Já dentre os conjuntos de dados com menos de 13 compostos, nenhum obteve modelos com AUC_{ROC} acima de 0,9, mas dois obtiveram um bom desempenho (AUC$> 0,8$): FX_{α} Fármacos (0,88) e HIVRT Fármacos (0,87), coincidentemente modelos simples construídos a partir de cinco e quatro compostos ativos, respectivamente.

O único par alvo/conjunto de dados formado por 13 ou mais compostos ativos que produziu modelos com AUC abaixo de 0.7 foi o $PPAR_{\gamma}$/Ligantes-PDB. Entretanto, a análise deste

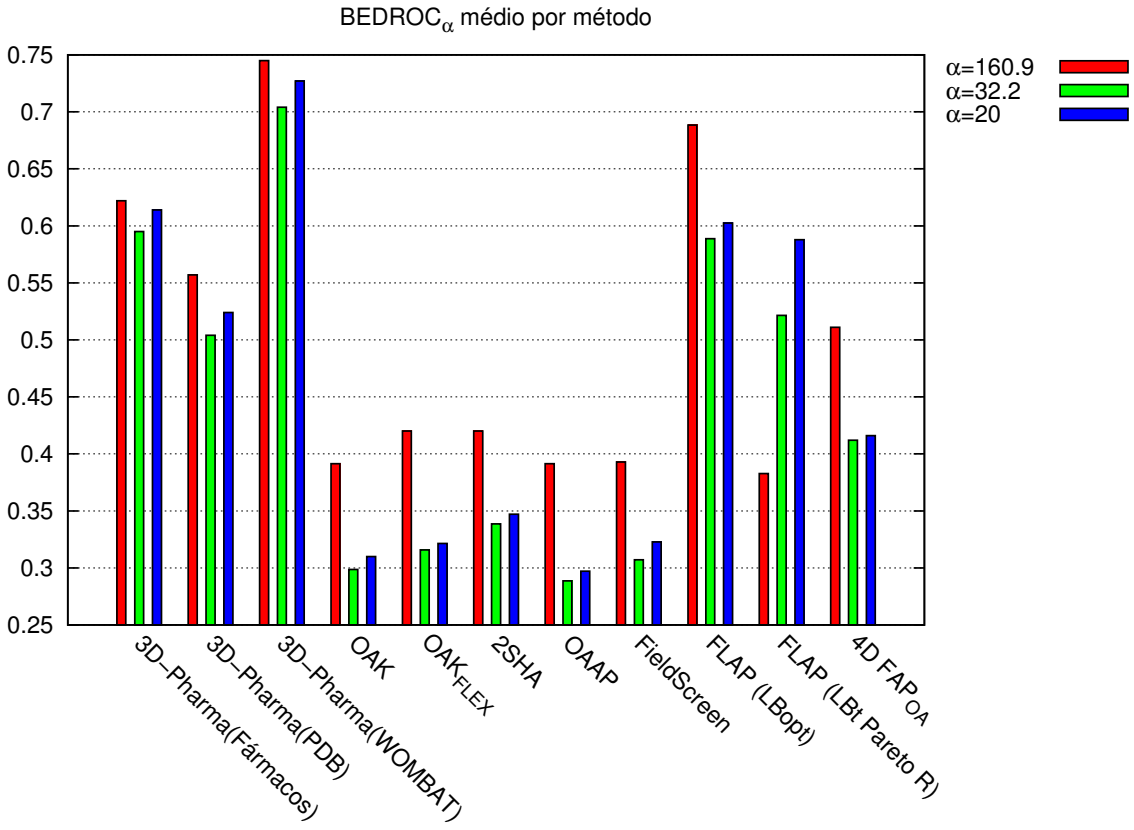


Figura 4.25: BEDROC$_\alpha$ médio por método, para os três valores de α (160.9, 32.2 e 20). O 3D-Pharma WOMBAT tem o melhor desempenho médio ao se considerar o reconhecimento precoce, seguido pelo FLAP LBopt e os outros conjuntos de compostos ativos usados pelo 3D-Pharma. O FLAP LBt Pareto R, apesar do excelente desempenho geral, tem um decepcionante reconhecimento imediato, principalmente nos cortes mais restritivos.

sistema foi comprometida devido ao pequeno número de negativos (DUD-Decoys), como explicitado na Seção 4.3.2. Dois modelos provindos de conjuntos com pelo menos 13 compostos ativos obtiveram AUCs entre 0.7 e 0.8: EGFR/WOMBAT e P38/Ligantes-PDB. O 3D-Pharma não conseguiu gerar bons modelos para o EGFR com nenhum dos conjuntos de compostos ativos, mesmo usando os dados do WOMBAT, com 62 substâncias. Já o caso do P38/MAPK14 requer uma análise mais profunda dos seus ligantes e suas interações com o alvo biológico que possa explicar a diferença no resultado. Peréz-Nueno e outros (181), usando os programas de superposição de volume PARAFIT e PARASURF para investigar os alvos do DUD quanto a múltiplos modos de ligação com os ligantes do mesmo, obtiveram resultados que sugerem que os ligantes da Cinase P38 podem interagir com a proteína em múltiplos sítios, o que também foi observado na literatura (204). Por outro lado, no caso dos modelos construídos a partir de conjuntos que possuiam menos de 13 compostos, dois se destacam: HIVRT/Fármacos (4 substâncias) e FX_{α}/Fármacos (5 substâncias), os quais geraram modelos que obtiveram uma AUC_{ROC} acima de 0.8 e, no caso do modelo FX_{α}/Fármacos, $AUC_{awROC} = 0.96$.

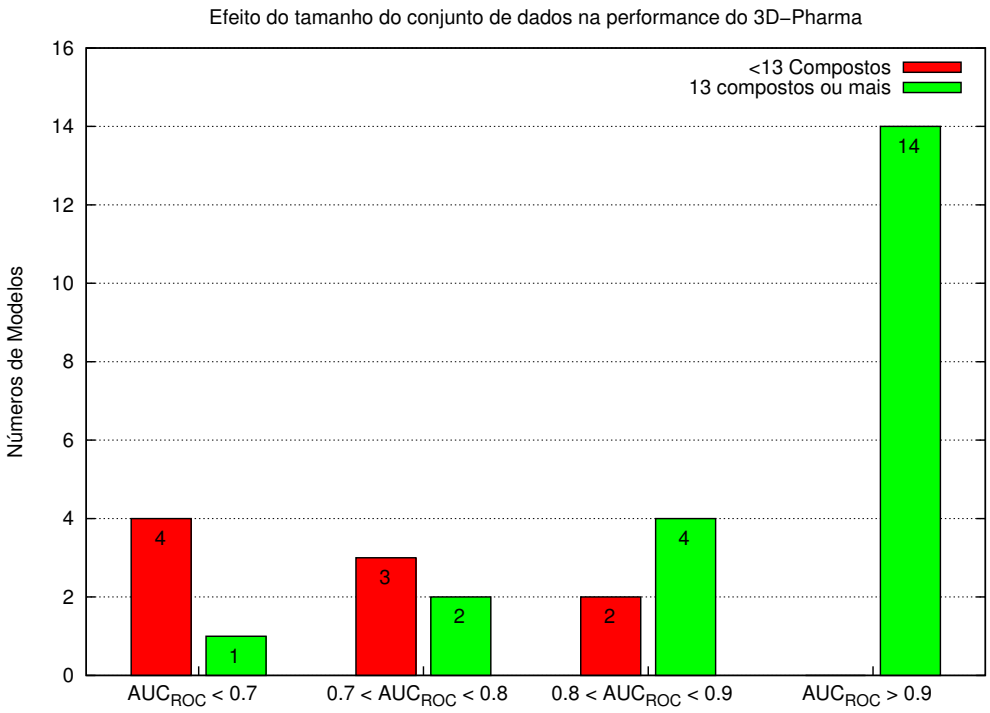


Figura 4.26: O efeito do tamanho do conjunto de substâncias ativas no desempenho do 3D-Pharma. Dos nove conjuntos de modelos provenientes de grupos formados por menos de 13 compostos ativos, nenhum obteve AUC_{ROC} acima de 0.9. Já dos 21 conjuntos de modelos gerados a partir de grupo de dados formados por 13 compostos ou mais, dois terços obtiveram valores de AUC_{ROC} acima de 0.9.

5 *Conclusões*

Esta tese apresentou o 3D-Pharma, uma nova ferramenta para triagem virtual baseada na estrutura de compostos ativos que usa uma abordagem baseada em múltiplas espécies e múltiplos modos (MS-MM). As características principais do 3D-Pharma consistem no uso de vetores binários indexados que codificam farmacóforos de três pontos em resolução atômica; o uso de múltiplas representações moleculares que incluem tautômeros e seus estados de protonação e múltiplos confôrmeros; o extensivo protocolo de validação dos modelos; e a abordagem baseada em múltiplas referências para a produção de modelos, os quais são representados por *fingerprints* modais baseadas na frequência dos farmacóforos de três pontos sobre os compostos ativos.

Esta abordagem foi aplicada com sucesso em um estudo retrospectivo de triagem virtual, produzindo modelos simples, robustos, consistentes e com alto poder preditivo. O 3D-Pharma também teve seu desempenho avaliado em um estudo comparativo contendo oito técnicas LBVS, sendo três delas (FLAP LBopt, FLAP LB Pareto R e 4D FAP_{OA}) técnicas de alto desempenho (AUC_{ROC} média acima de 0,7). Os resultados nos levam a crer que o 3D-Pharma supera todos os outros métodos LBVS investigados, em termos de acurácia geral, reconhecimento precoce e de diversidade estrutural, principalmente quando os dados provindos do WOMBAT foram usados como referência.

Logo, o 3D-Pharma se qualifica como um método *in silico* muito promissor, que pode auxiliar no processo de descoberta racional de fármacos, apesar de seu poder preditivo ainda precisar ser testado em cenários prospectivos e experimentais.

Referências Bibliográficas

1 ARROWSMITH, J. A decade of change. *Nat Rev Drug Discov*, v. 11, n. 1, p. 17–18, 2012.

2 ARROWSMITH, J. Trial watch: Phase II failures: 2008-2010. v. 10, n. 5, p. 328–329, 2011.

3 BENNANI, Y. L. Drug discovery in the next decade: innovation needed ASAP. *Drug Discovery Today*, v. 16, n. 17-18, p. 779 – 792, 2011.

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

5 KESERű, G. M.; MAKARA, G. M. Hit discovery and hit-to-lead approaches. *Drug Discovery Today*, v. 11, n. 15-16, p. 741 – 748, 2006.

6 KÜMMEL, A.; PARKER, C. N. The interweaving off chemoinformatics and hts. In: BAJORATH, J. (Ed.). *Chemoinformatics and Computational Chemical Biology*. [S.l.]: Springer Science+Business Media, LLC 2011, 2011. cap. 17, p. 435–457.

7 MAYR, L. M.; FUERST, P. The future of high-throughput screening. *Journal of Biomolecular Screening*, v. 13, n. 6, p. 443–448, 2008.

8 BAJORATH, J. Integration of virtual and high-throughput screening. *Nature Reviews Drug Discovery*, v. 1, p. 882–894, 2002.

9 WALTERS, W. P.; STAHL, M. T.; MURCKO, M. A. Virtual screening – an overview. *Drug Discovery Today*, v. 3, n. 4, p. 160–178, 1998.

10 VARNEK, A. Fragment descriptors in structure-property modeling and virtual screening. In: BAJORATH, J. (Ed.). *Chemoinformatics and Computational Chemical Biology*. [S.l.]: Springer Science+Business Media, LLC 2011, 2011. cap. 9, p. 213–243.

11 VOGT, M.; BAJORATH, J. Virtual screening methods based on bayesian statistics. In: LODHI, H.; YAMANISHI, Y. (Ed.). *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*. [S.l.]: IGI Global, 2011. cap. 10, p. 190–211.

12 LEACH, A. R.; SHOICHET, B. K.; PEISHOFF, C. E. Prediction of protein-ligand interactions. docking and scoring: Successes and gaps. *Journal of Medicinal Chemistry*, v. 49, n. 20, p. 5851–5855, 2006.

13 IRWIN, J. Community benchmarks for virtual screening. *Journal of Computer-Aided Molecular Design*, v. 22, p. 193–199, 2008.

14 LIEBESCHUETZ, J. Evaluating docking programs: keeping the playing field level. *Journal of Computer-Aided Molecular Design*, v. 22, p. 229–238, 2008.

15 KUNTZ, I. D. et al. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, v. 161, n. 2, p. 269 – 288, 1982.

16 WARR, W. A. Some trends in chem(o)informatics. In: BAJORATH, J. (Ed.). *Chemoinformatics and Computational Chemical Biology*. [S.l.]: Springer Science+Business Media, LLC 2011, 2011. cap. 1, p. 1–38.

17 BABER, J. C.; FEHER, M. Predicting synthetic accessibility: Application in drug discovery and development. *Mini Reviews in Medicinal Chemistry*, v. 4, n. 6, p. 681–692, 2004.

18 ZALIANI, A. et al. Second-generation de novo design: a view from a medicinal chemist perspective. *Journal of Computer-Aided Molecular Design*, v. 23, p. 593–602, 2009.

19 PATEL, H. et al. Knowledge-based approach to de novo design using reaction vectors. *Journal of Chemical Information and Modeling*, v. 49, n. 5, p. 1163–1184, 2009.

20 CARR, R. A. et al. Fragment-based lead discovery: leads by design. *Drug Discovery Today*, v. 10, n. 14, p. 987 – 992, 2005.

21 WARR, W. Fragment-based drug discovery. *Journal of Computer-Aided Molecular Design*, v. 23, p. 453–458, 2009.

22 JOSEPH-MCCARTHY, D. Challenges of fragment screening. *Journal of Computer-Aided Molecular Design*, v. 23, p. 449–451, 2009.

23 BERMAN, H. M. et al. The Protein Data Bank. *Nucl. Acids Res.*, v. 28, n. 1, p. 235–242, 2000.

24 BOLTON, E. E. et al. Pubchem: Integrated platform of small molecules and biological activities. In: WHEELER, R. A.; SPELLMEYER, D. C. (Ed.). [S.l.]: Elsevier, 2008, (Annual Reports in Computational Chemistry, v. 4). p. 217 – 241.

25 GAULTON, A. et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, v. 40, n. D1, p. D1100–D1107, 2012.

26 OLAH, M. et al. WOMBAT: World of Molecular Bioactivity. In: OPREA, T. I. (Ed.). *Chemoinformatics in Drug Discovery*. New York: Wiley-VCH, 2004. cap. 9, p. 223–239.

27 RIPPHAUSEN, P. et al. Quo vadis, virtual screening? a comprehensive survey of prospective applications. *Journal of Medicinal Chemistry*, v. 53, n. 24, p. 8461–8467, 2010.

28 WILLETT, P. Similarity searching using 2d structural fingerprints. In: BAJORATH, J. (Ed.). *Chemoinformatics and Computational Chemical Biology*. [S.l.]: Springer Science+Business Media, LLC 2011, 2011. cap. 5, p. 133–158.

29 HESSLER, G.; BARINGHAUS, K.-H. The scaffold hopping potential of pharmacophores. *Drug Discovery Today: Technologies*, v. 7, n. 4, p. e263 – e269, 2010.

30 CARHART, R. E.; SMITH, D. H.; VENKATARAGHAVAN, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, v. 25, n. 2, p. 64–73, 1985.

31 WILLETT, P.; WINTERMAN, V.; BAWDEN, D. Implementation of nearest-neighbor searching in an online chemical structure search system. *Journal of Chemical Information and Computer Sciences*, v. 26, n. 1, p. 36–41, 1986.

32 JOHNSON, M. A.; MAGGIORA, G. M. *Concepts and applications of molecular similarity*. New York: John Wiley, 1990.

33 PELTASON, L.; BAJORATH, J. Computational analysis of activity and selectivity cliffs. In: BAJORATH, J. (Ed.). *Chemoinformatics and Computational Chemical Biology*. [S.l.]: Springer Science+Business Media, LLC 2011, 2011. cap. 4, p. 119–132.

34 MAGGIORA, G. M. On outliers and activity cliffs: Why QSAR often disappoints. *Journal of Chemical Information and Modeling*, v. 46, n. 4, p. 1535–1535, 2006.

35 PELTASON, L.; BAJORATH, J. Sar index: Quantifying the nature of structure-activity relationships. *Journal of Medicinal Chemistry*, v. 50, n. 23, p. 5571–5578, 2007.

36 CRAMER, R. D.; REDL, G.; BERKOFF, C. E. Substructural analysis. novel approach to the problem of drug design. *Journal of Medicinal Chemistry*, v. 17, n. 5, p. 533–535, 1974.

37 VARNEK, A. et al. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *Journal of Computer-Aided Molecular Design*, v. 19, p. 693–703, 2005.

38 HUGHES, S. J. et al. Fragment based discovery of a novel and selective pi3 kinase inhibitor. *Bioorganic & Medicinal Chemistry Letters*, v. 21, n. 21, p. 6586 – 6590, 2011.

39 HORST, E. van der et al. Substructure-based virtual screening for adenosine a2a receptor ligands. *ChemMedChem*, WILEY-VCH Verlag, v. 6, n. 12, p. 2302–2311, 2011.

40 GÜNER, O. *Pharmacophore perception, development, and use in drug design*. La Jolla, CA: International University Line, 1999. (IUL biotechnology series). ISBN 9780963681768.

41 LANGER, T.; HOFFMANN, R. D. *Pharmacophores and Pharmacophore Searches*. Weinheim, Germany: Wiley-VCH, 2006. (Methods and Principles in Medicinal Chemistry, v. 32).

42 LEACH, A. R. et al. Three-dimensional pharmacophore methods in drug discovery. *Journal of Medicinal Chemistry*, v. 53, n. 2, p. 539–558, 2010.

43 SHENG-YONG; YANG. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discovery Today*, v. 15, n. 11-12, p. 444 – 450, 2010.

44 BEMIS, G.; KUNTZ, I. A fast and efficient method for 2D and 3D molecular shape description. *Journal of Computer-Aided Molecular Design*, v. 6, n. 6, p. 607–628, 1992.

45 GRANT, J.; PICKUP, B. A gaussian description of molecular shape. *The Journal of Physical Chemistry*, v. 99, n. 11, p. 3503–3510, 1995.

46 CRUCIANI, G. et al. *Molecular Interaction Fields: Applications in Drug Discovery and ADME Prediction*. [S.l.]: John Wiley & Sons, 2006. (Methods and Principles in Medicinal Chemistry). ISBN 9783527607136.

47 NICHOLLS, A.; GRANT, J. Molecular shape and electrostatics in the encoding of relevant chemical information. *Journal of computer-aided molecular design*, v. 19, n. 9, p. 661–686, 2005.

48 CHEESERIGHT, T. J. et al. Fieldscreen: Virtual screening using molecular fields. application to the dud data set. *Journal of Chemical Information and Modeling*, v. 48, n. 11, p. 2108–2117, 2008. Disponível em: <http://pubs.acs.org/doi/abs/10.1021/ci800110p>.

49 VAINIO, M.; PURANEN, J.; JOHNSON, M. ShaEP: molecular overlay based on shape and electrostatic potential. *Journal of chemical information and modeling*, v. 49, n. 2, p. 492–502, 2009.

50 LIU, X.; JIANG, H.; LI, H. SHAFTS: A Hybrid Approach for 3D Molecular Similarity Calculation. 1. Method and Assessment of Virtual Screening. *Journal of Chemical Information and Modeling*, v. 51, n. 9, p. 2372–2385, 2011.

51 SASTRY, G. M.; DIXON, S. L.; SHERMAN, W. Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring. *Journal of Chemical Information and Modeling*, v. 51, n. 10, p. 2455–2466, 2011.

52 CAI, C. et al. A novel, customizable and optimizable parameter method using spherical harmonics for molecular shape similarity comparisons. *Journal of Molecular Modeling*, v. 18, p. 1597–1610, 2012.

53 GOLBRAIKH, A.; TROPSHA, A. Predictive qsar modeling based on diversity sampling of experimental datasets for the training and test set selection. *Journal of Computer-Aided Molecular Design*, Springer Netherlands, v. 16, p. 357–369, 2002. ISSN 0920-654X.

54 OPREA, T. I. et al. Computational systems chemical biology. In: BAJORATH, J. (Ed.). *Chemoinformatics and Computational Chemical Biology*. [S.l.]: Springer Science+Business Media, LLC 2011, 2011. cap. 18, p. 459–488.

55 GOLBRAIKH, A.; TROPSHA, A. Beware of q2! *Journal of Molecular Graphics and Modelling*, v. 20, n. 4, p. 269 – 276, 2002.

56 HANSCH, C.; FUJITA, T. $\rho - \sigma - \pi$ Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *Journal of the American Chemical Society*, v. 86, n. 8, p. 1616–1626, 1964.

57 KUBINYI, H. *QSAR : Hansch analysis and related approaches*. New York, USA: VCH, 1993. (Methods and principles in medicinal chemistry). ISBN 9781560817680.

58 RANDIĆ, M. Characterization of molecular branching. *Journal of the American Chemical Society*, v. 97, n. 23, p. 6609–6615, 1975.

59 KIER, L.; HALL, L. *Molecular connectivity in chemistry and drug research*. New York, USA: Academic Press, 1976. (Medicinal chemistry). ISBN 9780124065604.

60 KIER, L.; HALL, L. *Molecular connectivity in structure-activity analysis*. New York, USA: Research Studies Press, 1986. (Chemometrics series). ISBN 9780471909835.

61 HALL, L. H.; KIER, L. B. Determination of topological equivalence in molecular graphs from the topological state. *Quantitative Structure-Activity Relationships*, WILEY-VCH Verlag, v. 9, n. 2, p. 115–131, 1990.

62 HALL, L. H.; MOHNEY, B.; KIER, L. B. The electrotopological state: An atom index for qsar. *Quantitative Structure-Activity Relationships*, v. 10, n. 1, p. 43–51, 1991.

63 HALL, L. H.; MOHNEY, B.; KIER, L. B. The electrotopological state: structure information at the atomic level for molecular graphs. *Journal of Chemical Information and Computer Sciences*, v. 31, n. 1, p. 76–82, 1991.

64 KIER, L.; HALL, L. *Molecular Structure Description: The Electrotopological State*. New York, USA: Academic Press, 1999. ISBN 9780124065550.

65 KELLOGG, G. E. et al. E-state fields: Applications to 3D QSAR. *Journal of Computer-Aided Molecular Design*, v. 10, p. 513–520, 1996.

66 SHERIDAN, R. P.; NACHBAR, R. B.; BUSH, B. L. Extending the trend vector: The trend matrix and sample-based partial least squares. *Journal of Computer-Aided Molecular Design*, v. 8, p. 323–340, 1994.

67 MATTER, H. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *Journal of Medicinal Chemistry*, v. 40, n. 8, p. 1219–1229, 1997.

68 CLEMENTI, S.; WOLD, S. How to choose the proper statistical method. In: WATERBEEMD, H. van de (Ed.). *Chemometric methods in molecular design*. New York, USA: VCH, 1995, (Methods and principles in medicinal chemistry). cap. 5.2, p. 319–338.

69 WOLD, S. PLS for multivariate linear modeling. In: WATERBEEMD, H. van de (Ed.). *Chemometric methods in molecular design*. New York, USA: VCH, 1995, (Methods and principles in medicinal chemistry). cap. 4.4, p. 195–218.

70 HOFFMAN, B. et al. Quantitative structure-activity relationship modeling of dopamine d1 antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and k nearest neighbor methods. *Journal of Medicinal Chemistry*, v. 42, n. 17, p. 3217–3226, 1999.

71 ZHENG, W.; TROPSHA, A. Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle. *Journal of Chemical Information and Computer Sciences*, v. 40, n. 1, p. 185–194, 2000.

72 AJAY. A unified framework for using neural networks to build qsars. *Journal of Medicinal Chemistry*, v. 36, n. 23, p. 3565–3571, 1993.

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

74 MARSHALL, G. R.; CRAMER, R. D. Three-dimensional structure-activity relationships. *Trends in Pharmacological Sciences*, v. 9, n. 8, p. 285 – 289, 1988.

75 GEPPERT, H.; VOGT, M.; BAJORATH, J. Current trends in ligand-based virtual screening: Molecular representations, data mining methods, new application areas, and performance evaluation. *Journal of Chemical Information and Modeling*, v. 50, n. 2, p. 205–216, 2010.

76 BURBIDGE, R. et al. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers & Chemistry*, v. 26, n. 1, p. 5 – 14, 2001.

77 WARMUTH, M. K. et al. Active learning with support vector machines in the drug discovery process? *Journal of Chemical Information and Computer Sciences*, v. 43, n. 2, p. 667–673, 2003.

78 VAPNIK, V. *The nature of statistical learning theory*. [S.l.]: Springer, 2000. (Statistics for engineering and information science). ISBN 9780387987804.

79 KLON, A. E.; DILLER, D. J. Library fingerprints: A novel approach to the screening of virtual libraries. *Journal of Chemical Information and Modeling*, v. 47, n. 4, p. 1354–1365, 2007.

80 WATSON, P. Naive bayes classification using 2d pharmacophore feature triplet vectors. *Journal of Chemical Information and Modeling*, v. 48, n. 1, p. 166–178, 2008.

81 HARPER, G. et al. Prediction of biological activity for high-throughput screening using binary kernel discrimination. *Journal of Chemical Information and Computer Sciences*, v. 41, n. 5, p. 1295–1300, 2001.

82 WILLETT, P. et al. Prediction of ion channel activity using binary kernel discrimination. *Journal of Chemical Information and Modeling*, v. 47, n. 5, p. 1961–1966, 2007.

83 ZHANG, S. Application of machine learning in drug discovery and development. In: LODHI, H.; YAMANISHI, Y. (Ed.). *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*. [S.l.]: IGI Global, 2011. cap. 12, p. 235–256.

84 ZHOU, Y.-P. et al. Modified particle swarm optimization algorithm for adaptively configuring globally optimal classification and regression trees. *Journal of Chemical Information and Modeling*, v. 49, n. 5, p. 1144–1153, 2009.

85 PALMER, D. S. et al. Random forest models to predict aqueous solubility. *Journal of Chemical Information and Modeling*, v. 47, n. 1, p. 150–158, 2007.

86 EHRMAN, T. M.; BARLOW, D. J.; HYLANDS, P. J. Virtual screening of chinese herbs with random forest. *Journal of Chemical Information and Modeling*, v. 47, n. 2, p. 264–278, 2007.

87 NICHOLLS, A. What do we know?: Simple statistical techniques that help. In: BAJORATH, J. (Ed.). *Chemoinformatics and Computational Chemical Biology*. [S.l.]: Springer Science+Business Media, LLC 2011, 2011. cap. 22, p. 531–581.

88 KORFF, M. von; FREYSS, J.; SANDER, T. Comparison of ligand- and structure-based virtual screening on the dud data set. *Journal of Chemical Information and Modeling*, v. 49, n. 2, p. 209–231, 2009.

89 TRUCHON, J.-F.; BAYLY, C. I. Evaluating virtual screening methods: Good and bad metrics for the ”early recognition” problem. *Journal of Chemical Information and Modeling*, v. 47, n. 2, p. 488–508, 2007.

90 NICHOLLS, A. What do we know and when do we know it? *Journal of Computer-Aided Molecular Design*, Springer Netherlands, v. 22, p. 239–255, 2008.

91 CLARK, R.; WEBSTER-CLARK, D. Managing bias in roc curves. *Journal of Computer-Aided Molecular Design*, Springer Netherlands, v. 22, p. 141–146, 2008.

92 GOOD, A.; OPREA, T. Optimization of camd techniques 3. virtual screening enrichment studies: a help or hindrance in tool selection? *Journal of Computer-Aided Molecular Design*, Springer Netherlands, v. 22, p. 169–178, 2008.

93 WERMUTH, C. G. et al. Glossary of terms used in medicinal chemistry (IUPAC recommendations 1998). *Pure and Applied Chemistry*, v. 70, n. 5, p. 1129–1143, 1998.

94 ARIËNS, E. J. *Molecular Pharmacology*. New York, USA: Academic Press, 1964.

95 PETER; GUND. Chapter 29 pharmacophoric pattern searching and receptor mapping. In: HESS, H.-J. (Ed.). [S.l.]: Academic Press, 1979, (Annual Reports in Medicinal Chemistry, v. 14). p. 299 – 308.

96 DRIE, J. H. V. Monty kier and the origin of the pharmacophore concept. *Internet Electronic Journal of Molecular Design*, v. 6, n. 9, p. 271–279, 2007.

97 CAPORUSCIO, F.; TAFI, A. Pharmacophore modelling: A forty year old approach and its modern synergies. *Current Medicinal Chemistry*, v. 18, n. 17, p. 2543–2553, 2011.

98 EHRLICH, P. Über den jetzigen Stand der Chemotherapie. *Chem Ber*, v. 42, p. 17, 1909.

99 KIER, L. B. Molecular orbital calculation of preferred conformations of acetylcholine, muscarine, and muscarone. *Molecular Pharmacology*, v. 3, n. 5, p. 487–494, 1967.

100 KIER, L. B. *Molecular orbital theory in drug research*. New York, USA: Academic Press, 1971.

101 WERMUTH, C. G. Pharmacophore: Historical perspective and viewpoint from a medicinal chemist. In: LANGER, T.; HOFFMANN, R. D. (Ed.). *Pharmacophores and Pharmacophore Searches*. Weinheim, Germany: Wiley-VCH, 2006. cap. 1.

102 SEIDEL, T. et al. Strategies for 3d pharmacophore-based virtual screening. *Drug Discovery Today: Technologies*, v. 7, n. 4, p. e221 – e228, 2010.

103 POPTODOROV, K.; LUU, T.; HOFFMANN, R. D. Pharmacophore model generation software tools. In: LANGER, T.; HOFFMANN, R. D. (Ed.). *Pharmacophores and Pharmacophore Searches*. Weinheim, Germany: Wiley-VCH, 2006. cap. 2.

104 HORVATH, D. Pharmacophore-based virtual screening. In: BAJORATH, J. (Ed.). *Chemoinformatics and Computational Chemical Biology*. [S.l.]: Springer Science+Business Media, LLC 2011, 2011. cap. 11, p. 261–298.

105 COTTRELL, S. J. et al. Generation of multiple pharmacophore hypotheses using multiobjective optimisation techniques. *Journal of Computer-Aided Molecular Design*, Springer Netherlands, v. 18, p. 665–682, 2004.

106 KRISTAM, R. et al. Comparison of conformational analysis techniques to generate pharmacophore hypotheses using catalyst. *Journal of Chemical Information and Modeling*, v. 45, n. 2, p. 461–476, 2005.

107 SCHWAB, C. H. Conformations and 3d pharmacophore searching. *Drug Discovery Today: Technologies*, v. 7, n. 4, p. e245 – e253, 2010.

108 WOLBER, G.; LANGER, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J Chem Info Model*, v. 45, n. 1, p. 160–169, 2005.

109 Accelrys Software. *DiscoveryStudio*®. 2001–2012. www.accelrys.com.

110 BARNUM, D. et al. Identification of common functional configurations among molecules. *Journal of Chemical Information and Computer Sciences*, v. 36, n. 3, p. 563–571, 1996.

111 LI, H.; SUTTER, J.; HOFFMANN, R. HypoGen: An automated system for generating 3d predictive pharmacophore models. In: GÜNER, O. (Ed.). *Pharmacophore perception, development, and use in drug design*. [S.l.]: International University Line, 1999, (IUL biotechnology series). cap. 10, p. 173–189.

112 JONES, G.; WILLETT, P.; GLEN, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *Journal of Computer-Aided Molecular Design*, Springer Netherlands, v. 9, p. 532–549, 1995.

113 RICHMOND, N. et al. GALAHAD: 1. pharmacophore identification by hypermolecular alignment of ligands in 3D. *Journal of Computer-Aided Molecular Design*, Springer Netherlands, v. 20, p. 567–587, 2006.

114 Tripos Inc. *GALAHAD*™. 2006–2012. www.tripos.com.

115 Chemical Computing Group. *Molecular Operating Environment*. Montreal, Canada: [s.n.]. www.chemcomp.com.

116 DIXON, S. et al. Phase: a new engine for pharmacophore perception, 3d qsar model development, and 3d database screening: 1. methodology and preliminary results. *Journal of Computer-Aided Molecular Design*, Springer Netherlands, v. 20, p. 647–671, 2006.

117 SCHNEIDMAN-DUHOVNY, D. et al. Deterministic pharmacophore detection via multiple flexible alignment of drug-like molecules. *Journal of Computational Biology*, v. 15, n. 7, p. 737–754, 2008.

118 DROR, O. et al. Novel approach for efficient pharmacophore-based virtual screening: Method and applications. *Journal of Chemical Information and Modeling*, v. 49, n. 10, p. 2333–2343, 2009. PMID: 19803502.

119 FENG, J.; SANIL, A.; YOUNG, S. S. PharmID: Pharmacophore Identification Using Gibbs Sampling. *J Chem Info Model*, v. 46, n. 3, p. 1352–1359, 2006.

120 HERT, J. et al. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *Journal of Chemical Information and Computer Sciences*, v. 44, n. 3, p. 1177–1185, 2004.

121 RAYMOND, J. W.; WILLETT, P. Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2d chemical structure databases. *Journal of Computer-Aided Molecular Design*, v. 16, p. 59–71, 2002.

122 HOLLIDAY, J.; HU, C.-Y.; WILLETT, P. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2d fragment bit-strings. *Combinatorial Chemistry & High Throughput Screening*, v. 5, n. 2, p. 155–166, 2002.

123 MARTIN, E. J.; HOEFFEL, T. J. Oriented Substituent Pharmacophore PRopErtY Space (OSPPREYS): A substituent-based calculation that describes combinatorial library products better than the corresponding product-based calculation. *Journal of Molecular Graphics and Modelling*, v. 18, n. 4-5, p. 383 – 403, 2000.

124 CATO, S. J. Exploring phrarmacophores with Chem-X. In: GÜNER, O. (Ed.). *Pharmacophore perception, development, and use in drug design*. La Jolla, CA: International University Line, 1999, (IUL biotechnology series). p. 107–125.

125 MCGREGOR, M. J.; MUSKAL, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *Journal of Chemical Information and Computer Sciences*, v. 39, n. 3, p. 569–574, 1999.

126 MCGREGOR, M. J.; MUSKAL, S. M. Pharmacophore fingerprinting. 2. application to primary library design. *Journal of Chemical Information and Computer Sciences*, v. 40, n. 1, p. 117–125, 2000.

127 Accelrys Software. *Cerius*^{2}. San Diego, CA: [s.n.]. `www.accelrys.com`.

128 Tripos, L.P. *Tuplets*. St. Louis, MO: [s.n.]. `www.tripos.com`.

129 KOES, D. R.; CAMACHO, C. J. Pharmer: Efficient and exact pharmacophore search. *Journal of Chemical Information and Modeling*, v. 51, n. 6, p. 1307–1314, 2011.

130 RANU, S.; SINGH, A. K. Novel method for pharmacophore analysis by examining the joint pharmacophore space. *Journal of Chemical Information and Modeling*, v. 51, n. 5, p. 1106–1121, 2011. Disponível em: <http://pubs.acs.org/doi/abs/10.1021/ci100503y>.

131 TROPSHA, A. Best practices for qsar model development, validation, and exploitation. *Molecular Informatics*, WILEY-VCH Verlag, v. 29, n. 6-7, p. 476–488, 2010.

132 EFRON, B.; TIBSHIRANI, R. J. *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1993.

133 TROPSHA, A.; GRAMATICA, P.; GOMBAR, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of qspr models. *QSAR & Combinatorial Science*, WILEY-VCH Verlag, v. 22, n. 1, p. 69–77, 2003.

134 TROPSHA, A.; GOLBRAIKH, A. Predictive qsar modeling workflow, model applicability domains, and virtual screening. *Current Pharmaceutical Design*, v. 13, n. 34, p. 3494–3504, 2007.

135 GOLBRAIKH, A. et al. Rational selection of training and test sets for the development of validated qsar models. *Journal of Computer-Aided Molecular Design*, Springer Netherlands, v. 17, p. 241–253, 2003.

136 FILIMONOV, D.; POROIKOV, V. Probabilistic approaches in activity prediction. In: VARNEK, A.; TROPSHA, A. (Ed.). *Chemoinformatics Approaches to Virtual Screening*. [S.l.]: RSC Publishing, 2008. cap. 6, p. 182–216.

137 SCIOR, T. et al. Recognizing pitfalls in virtual screening: A critical review. *Journal of Chemical Information and Modeling*, v. 52, n. 4, p. 867–881, 2012.

138 BRINK, T. ten; EXNER, T. E. Influence of protonation, tautomeric, and stereoisomeric states on protein-ligand docking results. *Journal of Chemical Information and Modeling*, v. 49, n. 6, p. 1535–1546, 2009.

139 KALLIOKOSKI, T. et al. The effect of ligand-based tautomer and protomer prediction on structure-based virtual screening. *Journal of Chemical Information and Modeling*, v. 49, n. 12, p. 2742–2748, 2009.

140 MILLETTI, F.; VULPETTI, A. Tautomer preference in pdb complexes and its impact on structure-based drug discovery. *Journal of Chemical Information and Modeling*, v. 50, n. 6, p. 1062–1074, 2010.

141 VIETH, M.; HIRST, J. D.; BROOKS, C. L. Do active site conformations of small ligands correspond to low free-energy solution structures? *Journal of Computer-Aided Molecular Design*, v. 12, p. 563–572, 1998.

142 SITZMANN, M. et al. Pdb ligand conformational energies calculated quantum-mechanically. *Journal of Chemical Information and Modeling*, v. 52, n. 3, p. 739–756, 2012.

143 DEPRISTO, M. A.; BAKKER, P. I. W. de; BLUNDELL, T. L. Heterogeneity and inaccuracy in protein structures solved by x-ray crystallography. *Structure*, v. 12, p. 831 – 838, 2004.

144 AGRAFIOTIS, D. K. et al. Conformational sampling of bioactive molecules: A comparative study. *Journal of Chemical Information and Modeling*, v. 47, n. 3, p. 1067–1086, 2007.

145 BROOKS, B. R. et al. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, v. 4, n. 2, p. 187–217, 1983.

146 SMELLIE, A.; TEIG, S.; TOWBIN, P. Poling: promoting conformational variation. *Journal of Computational Chemistry*, v. 16, n. 2, p. 171–187, 1995.

147 LI, J. et al. CAESAR: a new conformer generation algorithm based on recursive buildup and local rotational symmetry consideration. *Journal of Chemical Information and Modeling*, v. 47, n. 5, p. 1923–1932, 2007.

148 HAWKINS, P. et al. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *Journal of Chemical Information and Modeling*, v. 50, n. 4, p. 572–584, 2010.

149 OpenEye Scientific Software. *OMEGA*. 1997–2012. `www.eyesopen.com`.

150 MOHAMADI, F. et al. MacroModel–an integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *Journal of Computational Chemistry*, v. 11, n. 4, p. 440–467, 1990.

151 KOLOSSVÁRY, I.; GUIDA, W. Low mode search. an efficient, automated computational method for conformational analysis: Application to cyclic and acyclic alkanes and cyclic peptides. *Journal of the American Chemical Society*, v. 118, n. 21, p. 5011–5019, 1996.

152 FERGUSON, D.; RABER, D. A new approach to probing conformational space with molecular mechanics: random incremental pulse search. *Journal of the American Chemical Society*, v. 111, n. 12, p. 4371–4378, 1989.

153 PEARLMAN, R.; BALDUCCI, R. CONFORT: a novel algorithm for conformational analysis. In: *National Meeting of the American Chemical Society, New Orleans*. New Orleans, USA: [s.n.], 1998.

154 MCMARTIN, C.; BOHACEK, R. QXP: powerful, rapid computer algorithms for structure-based drug design. *Journal of Computer-Aided Molecular Design*, v. 11, n. 4, p. 333–344, 1997.

155 KLEBE, G.; MIETZNER, T. A fast and efficient method to generate biologically relevant conformations. *Journal of Computer-Aided Molecular Design*, v. 8, n. 5, p. 583–606, 1994.

156 KLEBE, G.; MIETZNER, T.; WEBER, F. Methodological developments and strategies for a fast flexible superposition of drug-size molecules. *Journal of Computer-Aided Molecular Design*, v. 13, n. 1, p. 35–49, 1999.

157 TRESADERN, G.; AGRAFIOTIS, D. K. Conformational sampling with stochastic proximity embedding and self-organizing superimposition: Establishing reasonable parameters for their practical use. *Journal of Chemical Information and Modeling*, v. 49, n. 12, p. 2786–2800, 2009.

158 GRIEWEL, A. et al. Conformational sampling for large-scale virtual screening: accuracy versus ensemble size. *Journal of Chemical Information and Modeling*, v. 49, n. 10, p. 2303–2311, 2009.

159 PERRUCCIO, F. et al. FLAP: 4-point pharmacophore fingerprints from GRID. In: CRUCIANI, G. (Ed.). *Molecular Interaction Fields: Applications in Drug Discovery and ADME Prediction*. Weinheim, Germany: Wiley-VCH, 2006. cap. 4.

160 BARONI, M. et al. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins (FLAP): Theory and Application. *J Chem Info Model*, v. 47, n. 2, p. 279–294, 2007.

161 CAROSATI, E. et al. Ligand-based virtual screening and adme-tox guided approach to identify triazolo-quinoxalines as folate cycle inhibitors. *Bioorganic & Medicinal Chemistry*, v. 18, n. 22, p. 7773 – 7785, 2010.

162 CROSS, S. et al. Flap: Grid molecular interaction fields in virtual screening. validation using the dud data set. *Journal of Chemical Information and Modeling*, v. 50, n. 8, p. 1442–1450, 2010.

163 GOODFORD, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem*, v. 28, n. 7, p. 849–857, 1985.

164 JAHN, A. et al. Probabilistic modeling of conformational space for 3D machine learning approaches. *Molecular Informatics*, v. 29, n. 5, p. 441–455, 2010.

165 JAHN, A. et al. Boltzmann-Enhanced Flexible Atom-Pair Kernel with Dynamic Dimension Reduction. *Molecular Informatics*, v. 30, n. 4, p. 307–315, 2011.

166 JAHN, A. et al. 4d flexible atom-pairs: An efficient probabilistic conformational space comparison for ligand-based virtual screening. *Journal of Cheminformatics*, v. 3, n. 1, p. 23, 2011.

167 DEMPSTER, A.; LAIRD, N.; RUBIN, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 39, n. 1, p. 1–38, 1977.

168 FRÖHLICH, H. et al. Optimal assignment kernels for attributed molecular graphs. In: *ICML*. [S.l.: s.n.], 2005. p. 225–232.

169 FRöHLICH, H. et al. Kernel functions for attributed molecular graphs - a new similarity-based approach to adme prediction in classification and regression. *QSAR & Combinatorial Science*, WILEY-VCH Verlag, v. 25, n. 4, p. 317–326, 2006. ISSN 1611-0218. Disponível em: <http://dx.doi.org/10.1002/qsar.200510135>.

170 Cresset Group. *XedeX - XED Tools*. Hertfordshire, UK: [s.n.], 2012. `http://www.cresset-group.com/products/xedtools/`.

171 KINNINGS, S. L.; JACKSON, R. M. Ligmatch: A multiple structure-based ligand matching method for 3d virtual screening. *Journal of Chemical Information and Modeling*, v. 49, n. 9, p. 2056–2066, 2009.

172 BRAKOULIAS, A.; JACKSON, R. Towards a structural classification of phosphate binding sites in protein–nucleotide complexes: An automated all-against-all structural comparison using geometric matching. *Proteins: Structure, Function, and Bioinformatics*, v. 56, n. 2, p. 250–260, 2004.

173 INBAR, Y. et al. Deterministic pharmacophore detection via multiple flexible alignment of drug-like molecules. In: *Research in Computational Molecular Biology (RECOMB) - 11th Annual International Conference*. Oakland, CA, USA: [s.n.], 2007. v. 4453, p. 412–429.

174 LIU, X. et al. Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation. *BMC Bioinformatics*, v. 10, n. 1, p. 101, 2009.

175 HUANG, N.; SHOICHET, B. K.; IRWIN, J. J. Benchmarking sets for molecular docking. *J Med Chem*, v. 49, n. 23, p. 6789–6801, 2006.

176 ROHRER, S.; BAUMANN, K. Maximum unbiased validation (MUV) data sets for virtual screening based on pubchem bioactivity data. *Journal of Chemical Information and Modeling*, v. 49, n. 2, p. 169–184, 2009.

177 RIPPHAUSEN, P.; WASSERMANN, A. M.; BAJORATH, J. REPROVIS-DB: A benchmark system for ligand-based virtual screening derived from reproducible prospective applications. *Journal of Chemical Information and Modeling*, v. 51, n. 10, p. 2467–2473, 2011.

178 IRWIN, J.; SHOICHET, B. ZINC-a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, v. 45, n. 1, p. 177–182, 2005.

179 JAHN, A. et al. Optimal assignment methods for ligand-based virtual screening. *Journal of Cheminformatics*, v. 1, n. 1, p. 14, 2009. ISSN 1758-2946. Disponível em: <http://www.jcheminf.com/content/1/1/14>.

180 VENKATRAMAN, V. et al. Comprehensive comparison of ligand-based virtual screening tools against the dud data set reveals limitations of current 3d methods. *Journal of Chemical Information and Modeling*, v. 50, n. 12, p. 2079–2093, 2010.

181 PÉREZ-NUENO, V. I.; RITCHIE, D. W. Using consensus-shape clustering to identify promiscuous ligands and protein targets and to choose the right query for shape-based virtual screening. *Journal of Chemical Information and Modeling*, v. 51, n. 6, p. 1233–1248, 2011.

182 OPREA, T. I. et al. Is there a difference between leads and drugs? a historical perspective. *Journal of Chemical Information and Computer Sciences*, v. 41, n. 5, p. 1308–1315, 2001.

183 VERDONK, M. et al. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *Journal of Chemical Information and Computer Sciences*, v. 44, n. 3, p. 793–806, 2004.

184 UPTON, G.; FINGLETON, B. *Spatial Data Analysis by Example*. New York: John Wiley & Sons Ltd., 1985. (Point Pattern and Quantitative Data).

185 FORTIN, M.; DALE, M. *Spatial Analysis: A Guide For Ecologists*. [S.l.]: Cambridge University Press, 2005.

186 WANG, Y. et al. PubChem's BioAssay Database. *Nucleic Acids Research*, v. 40, n. D1, p. D400–D412, 2012.

187 BAJORATH, J. *Chemoinformatics and Computational Chemical Biology*. Humana Press, 2011. (Methods in Molecular Biology). ISBN 9781607618386. Disponível em: <http://books.google.com.br/books?id=shHJSAAACAAJ>.

188 ChemAxon. 1999–2012. `www.chemaxon.com`.

189 HALGREN, T. A. MMFF VI. MMFF94s option for energy minimization studies. *Journal of Computational Chemistry*, v. 20, n. 7, p. 720–729, 1999.

190 JAKALIAN, A.; JACK, D. B.; BAYLY, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. parameterization and validation. *Journal of Computational Chemistry*, v. 23, n. 16, p. 1623–1641, 2002.

191 BOTELHO, F. C.; ZIVIANI, N. External perfect hashing for very large key sets. In: *CIKM '07: Proceedings of the sixteenth ACM Conference on information and knowledge management*. Lisbon, Portugal: ACM, 2007. p. 653–662.

192 WISHART, D. S. et al. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, v. 34, n. suppl 1, p. D668–D672, 2006.

193 WISHART, D. S. et al. DrugBank: a knowledge base for drugs, drug actions and drug targets. *Nucl. Acids Res.*, v. 36, n. Database Issue, p. D901–906, 2008.

194 KANEHISA, M.; GOTO, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acids Res.*, v. 28, n. 1, p. 27–30, 2000.

195 CHEN, X.; JI, Z. L.; CHEN, Y. Z. TTD: Therapeutic Target Database. *Nucl. Acids Res.*, v. 30, n. 1, p. 412–415, 2002.

196 KALASZI, A.; POLGAR, T.; VARGYAS, M. *Ligand Based Virtual Screening using Screen3D*. Budapest, Hungary: [s.n.], 2011. ChemAxon User Group Meeting.

197 GRANT, J.; GALLARDO, M.; PICKUP, B. A fast method of molecular shape comparison: A simple application of a gaussian description of molecular shape. *Journal of Computational Chemistry*, v. 17, n. 14, p. 1653–1666, 1996.

198 OpenEye Scientific Software. *ROCS*. 2005–2012. `http://www.eyesopen.com/rocs`.

199 OpenEye Scientific Software. *FRED - Fast Exaustive Docking*. 2003–2012. `http://www.eyesopen.com/oedocking`.

200 ABAGYAN, R.; TOTROV, M.; KUZNETSOV, D. ICM–a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry*, v. 15, n. 5, p. 488–506, 1994.

201 JAIN, A. Ligand-based structural hypotheses for virtual screening. *Journal of Medicinal Chemistry*, v. 47, n. 4, p. 947–961, 2004.

202 LEMMEN, C.; LENGAUER, T.; KLEBE, G. FLEXS: a method for fast flexible ligand superposition. *Journal of Medicinal Chemistry*, v. 41, n. 23, p. 4502–4520, 1998.

203 GIGANTI, D. et al. Comparative evaluation of 3D virtual ligand screening methods: impact of the molecular alignment on enrichment. *Journal of Chemical Information and Modeling*, v. 50, n. 6, p. 992–1004, 2010.

204 PARGELLIS, C. et al. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nature Structural & Molecular Biology*, v. 9, n. 4, p. 268–272, 2002.

Artigo: Journal of Chemical Information and Modeling

Submission Confirmation

Thank you for submitting your manuscript to *Journal of Chemical Information and Modeling*.

Manuscript ID:	ci-2012-00217a
Title:	3D-Pharma, A Ligand-based Virtual Screening tool using 3D Pharmacophore Fingerprints
Authors:	Domingues, Bernardo Lopes, Júlio César Martins-José, Andrelly
Date Submitted:	04-May-2012

Print Return to Dashboard

3D-Pharma, A Ligand-based Virtual Screening tool using 3D Pharmacophore Fingerprints

Bernardo F. Domingues,[*,†,‡] Andrelly Martins-José,[†] and Júlio C. D. Lopes[*,†]

Chemoinformatics Group NEQUIM, Departamento de Química, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

E-mail: bernardofd@gmail.com; jlopes.ufmg@gmail.com

Abstract

In this work, we introduced 3D-Pharma, a new Ligand-Based Virtual Screening method that uses fingerprints of pharmacophore triplets at atomic resolutions to build very simple and predictive models. Within 3D-Pharma the molecules are described by multiple representations that comprehend several prototropic species and conformations (multiple species, multiple mode approach). All the multiple representations of a compound are concatenated into a unique fingerprint that accounts for most of its chemical and conformational diversity. The biological activity of an ensemble of active molecules are represented by a single modal fingerprint or model, validated through a new exhaustive 10-fold cross-validation scheme, which improves robustness and internal consistency of the models, as well as its predictive power. We benchmark our method with 10 datasets of active compounds and decoys gathered from DUD database and compare its performance against seven state-of-the-art LBVS methods. To generate the models we used three external and independent datasets of bioactive compounds (Drugs, PDB Ligands and WOMBAT). We concluded that 3D-Pharma overperforms all other

[*]To whom correspondence should be addressed
[†]NEQUIM-UFMG
[‡]Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

state-of-the-art LBVS tools analyzed, in terms of global accuracy as well as scaffold hopping and early recovery capacities. Furthermore, the models produced by 3D-Pharma are simple, robust, consistent and predictive.

Introduction

Since 1996, the pharmaceutical industry has experienced a long period of low profits whilst experiencing a continuous increase in the expenditures. The productivity, as measured by the number of approved NME's, achieved its lowest number since the 80's.[1] As a consequence, there is an impressive number of blockbuster drugs reaching its patents covering in the next few years (a phenomena known as "patent cliff")[2–4] and the major players in the pharmaceutical industry are pushed to abandon less profitable branches and to concentrate the efforts on those whose financial return is more likely.[5] Within this scenario, *in silico* methods became even more important in order to speed up the process of discovery at lower aggregated cost.[6]

A major problem in drug discovery is the identification of novel compounds that show binding properties to protein targets of pharmaceutical interest together with appropriate pharmacokinetic properties. In the lack of previous information regarding a target of interest nor known active compounds, there is no choice but to rely on a brute force hit identification process, such as HTS (High Throughput Screening), over a whole chemical compound library.[7,8] But as data about known ligands or crystallographic structures from protein targets become available, such informations can be used for selecting a small portion of some virtual compound database that shows a better likelihood to interact with a given target than a randomly chosen compound.[9] Such enrichment is the goal of Virtual Screening (VS) applications.[10] VS methods can be broadly classified by the origin of the data used in the screening process. If it uses only data derived from previously known ligands, the application is known as *Ligand-Based VS*; if the application uses data derived from the structure of the protein target, the method is known as *Target-Based VS*. There is a plethora of methods available within each class, and a handful of hybrid methods,[11–13] most of them summarized in

several recent reviews.[10,14–16]

The methods of discovery and development of new drugs based on knowledge of the structures of biological targets are unavoidably dependent on their validation as potential therapeutic targets, as well as the specific mode of action through which the new candidate drug will exert its putative/predicted effect. It is worth mentioning that approximately 50% of the compounds rejected at Clinical Phase II is due to insufficient efficacy,[17] suggesting deficiencies in target or endpoint biomarkers validations.[18] A TBVS study is typically carried out through molecular docking approaches. However, although they are able to reproduce with reasonable precision the conformations observed in the crystallographic complexes, the scoring functions cannot consistently differentiate active from inactive compounds.[19] Moreover, the methods of drug discovery and development based on the knowledge of the active ligands structures, e.g. those substances previously known to cause a biological, pharmacological or therapeutic effect of interest, has a great potential to generate predictive models aiming not only for a specific target, but also for biological effects which can be associated with *in vitro* or *in vivo* biological tests.[20–22] Conversely, the ligand-based tools do not depend of information regarding a biological target nor a mechanism of action, hence enabling them to create predictions over effects that should be multifactorial.[23–25] The methods based on the structures of the active ligands, although not necessarily dependent on a specific mode of action, lack of a sufficient knowledge of the molecular structure and its role in the interactions with the components of the biological environment.[26] Hence, there is not a consistent set of molecular descriptors capable of foreseeing the biological properties of small molecules with reasonable confidence.

The most popular ligand-based methods are those derived from 2D molecular structures, possibly due to its availability, quickness, easiness of use and relatively well established protocols.[27] In general, they present a good performance in retrospective studies,[28] but the results of prospective studies generally are disappointing, although propitiate a relative enrichment of actives.[29] The inclusion of conformational data in 3D LBVS methods increases the complexity of the algorithms and computational costs. Therefore, most of the 3D methods selects only one conforma-

tion for each compound. The options are the lowest energy conformation, a conformation from a co-crystallized complex, or a conformation based on alignment with other active compounds. However, the binding event could change dramatically the energy of conformers, and the active conformation could not be same of the global minimum calculated in water or in vacuum, as usually done by most molecular mechanics force-fields used in chemoinformatics and medicinal chemistry. In fact, recently Nicklaus and co-workers[30] analyzed the conformations of ligands found in PDB database at Density Functional level of theory. The difference of energy between the conformations found in PDB structures and the fully optimized conformer stay in the range of 0 to 25 kcal/mol, distributed quite evenly and independently of the crystallographic resolution. The ligands deposited on the Protein Data Bank[31] mostly present only one conformation bound to the active site of the biological target. However, there is an increasing amount of data showing that is not uncommon for a ligand to exhibit multiple conformations or multiple binding modes.[32,33] In addiction, the fact that crystallographic structures are not obtained in physiological conditions casts doubts over which conformation should be considered as the bioactive.[34] Even the soaking method used to produce the crystal could influence the conformation observed in models obtained from the X-ray diffraction maps.[35]

Some recent studies suggest that 2D methods outperform 3D approaches in terms of accuracy.[28] These can be due to deficiencies in the quality of the 3D descriptors or problems with the conformational sampling. However, some 3D methods based on shape, electrostatics, or pharmacophore features have been shown to perform better than 2D LBVS methods when considering scaffold hopping.[36] Pharmacophore-based VS approaches are numerous. The classical use of pharmacophore elucidation through the alignment of active compounds is the approach taken by the majority of commercially available tools like CATALYST,[37] GALAHAD,[38] GASP,[39] the pharmacophore module of MOE,[40] PHASE[41] and many others non-commercial tools, like PharmaGist.[42,43] Although the traditional pharmacophore mapping approach is well established and performs satisfactorily, it is very sensitive to the size of the dataset to be screened[14] and biased by the conformation selection.[44,45] Pharmacophore keys (or fingerprints) encode the pharmacophoric

features in binary vectors. They lack the intuitive nature of the classical pharmacophore elucidation, but they have a considerable higher throughput when dealing with large molecular databases, making them a very popular approach to VS. There are several implementations of pharmacophore searches powered by fingerprints, such as FLAP[46–49] and Pharmer.[50] Other examples found are commercially available suites like Tripos' Tuplets[51] and Accelrys' 3DKeys.[52] Despite the number of available techniques, any comparison between them is difficult due to the lack of benchmarking standards both in databases and metrics of evaluation.[15]

Currently there is a search for new molecular descriptors that associated with strong information modeling techniques and robust statistical methods that could make predictions in levels beyond of those produced by the traditional methods.[28] Furthermore, considering that biological systems are dynamic in their very nature, both ligands and biological targets are adaptive flexible molecules and the emergent properties that arise from the binding event can trigger cascade effects that ultimately produce the observed biological effects.[53] To clarify such a chain of events and to develop suitable tools for practical application is far beyond the knowledge of the structures of the binding partners.[26] Thus, we are currently at a crossroads in which the available methods are not up to the challenge of predicting accurately the biological activities of small organic molecules.[54]

Molecular recognition events are strongly dependent on conformational and proton-exchange equilibria but they are often neglected in virtual screening studies.[55,56] Different microspecies of biological macromolecules and their ligands are characterized by different conformations and stereo-electronic characteristics. Therefore, including multiple species and multiple conformational data (MultiSpecies-MultiMode approach, or MS-MM)[57,58] in VS applications could result in a comprehensive approach of biological activity and of inherent dynamics of the process of binding for small molecules to its biological targets. Considering those LBVS methods that use DUD datasets as benchmark, there are a few that have as a feature a full multimode approach,[59–61] being more common to start with multiple conformations but the final results are computed over a single conformation.[46–49,62,63] To the best of our knowledge there is no published LBVS method that use the MS-MM approach with DUD datasets.

This work aims to introduce 3D-Pharma, a new method for LBVS based on pharmacophore fingerprints. The prediction power of 3D-Pharma is compared against other state-of-the-art LBVS methods available elsewhere that used the Directory of Useful Decoys (DUD) database[64] as a benchmarking data set. The 3D-Pharma method applies a multispecie-multimode (MS-MM) approach for the generation of atom-centered potentials pharmacophore triplets. Within 3D-Pharma, all the multiple representations of a compound are concatenated into a unique binary vector or fingerprint that accounts for most of the chemical and conformational diversity of the compound. 3D-Pharma uses a single modal fingerprint[65] to represent the biological activity of an ensemble of active molecules, each one represented by its unique fingerprint. These models were built from three external and independent datasets (Drugs, PDB-Ligands and Wombat) and were validated by an exhaustive 10-fold cross validation scheme where each external dataset was divided in three subsets comprising training, evaluation and test sets. The final models were selected after extensive internal validation against the evaluation and testing sets.

Material and Methods

Data

A retrospective virtual screening study was made against ten protein targets chosen among the 40 targets available in the Directory of Useful Decoys (DUD) dataset.[64] The DUD Dataset is a benchmarking dataset for docking tools, but is commonly used to assess performance of ligand-based methods and is well suited to do so.[28] The selected targets are: Aldose Reductase (ALR2), Androgen Receptor (AR), Cyclin-dependent Kinase 2 (CDK2), Cyclooxygenase-2 (COX-2), Epidermal Growth Factor Receptor Kinase (EGFR), Factor X-α (FX$_\alpha$), Mitogen activated Protein Kinase 14 (P38), HIV-1 Reverse Transcriptase (HIVRT), Phosphodiesterase V (PDE5) and Peroxisome Proliferator Activated Receptor γ (PPAR$_\gamma$). Those targets were selected based on availability of WOMBAT datasets[66,67] molecules through the DUD website (`http://dud.docking.org`).

For each active molecule in the original release of DUD, there are approximately 36 other

inactive molecules with similar topological features. Jahn et al[68] performed a lead-like filter[69] over the molecules, as suggested by Good and Oprea,[67] in order to make the benchmark set more suitable for LBVS applications. For each of the aforementioned targets, sets of actives and inactives molecules were obtained from DUD according to Jahn's filter. Aiming to have true and reliable external validation of 3D-Pharma approach, three independent datasets were used to build the models for each target. The first dataset, called "Drugs", consists of all available approved and experimental drugs for each target, gathered from the public databases DrugBank,[70,71] KEGG Drugs[72] and Therapeutic Targets Database (TTD).[73] The second dataset, called "PDB-Ligands", contains the ligands bound to any crystallographic structure of the target deposited on the Protein Data Bank (PDB).[31] Finally, the third dataset, as mentioned above, was the WOMBAT datasets available through the DUD website. All datasets used to build models for each target (Drugs, PDB Ligands and WOMBAT) were previously filtered using a 2D comparison within ChemAxon's Instant JChem[74] against the corresponding DUD Actives subset. Any redundancies between them were excluded from the external datasets. A complete list of the molecules gathered for this study is available in the Supporting Information. Table 1 shows the datasets sizes as well as the number of different chemotypes found in DUD Actives datasets for each target. The numbers may vary from the original DUD release and from other publications since some of the molecules generated errors throughout the molecular treatment protocol and did not entered in the final subsets.

Treatment of Molecular Structures

All molecules used in this work were preprocessed following a pre-treatment protocol of manual dessalting, succeeded by standardization and dominant tautomer calculation with Standardizer program by ChemAxon.[74] These steps ensure that all structures are in the same initial state. All datasets were submitted to the same protocol of molecular treatment, which starts by determining all dominant tautomer between pH 0 and 14, followed by a major microspecie calculation at pH 7 for each tautomer. These steps are important to be sure that a relevant sample of the chemical variability of the compound is taken into account when computing the potential pharmacophore

Table 1: Number of unique compounds for each target in the DUD and external datasets (Drugs, PDB-Ligands and WOMBAT), as well as the number of chemotypes for each DUD Active dataset.

Target	Dataset	Number of Compounds	Number of Chemotypes in DUD Actives
ALR2	DUD Actives	26	14
	DUD Decoys	910	
	Drugs	11	
	PDB Ligands	26	
	WOMBAT	41	
AR	DUD Actives	68	10
	DUD Decoys	2616	
	Drugs	45	
	PDB Ligands	13	
	WOMBAT	36	
CDK2	DUD Actives	47	32
	DUD Decoys	1702	
	Drugs	37	
	PDB Ligands	132	
	WOMBAT	148	
COX-2	DUD Actives	212	44
	DUD Decoys	11577	
	Drugs	77	
	PDB Ligands	4	
	WOMBAT	66	
EGFR	DUD Actives	365	40
	DUD Decoys	14516	
	Drugs	10	
	PDB Ligands	12	
	WOMBAT	62	
FX_{α}	DUD Actives	64	19
	DUD Decoys	1888	
	Drugs	5	
	PDB Ligands	85	
	WOMBAT	105	
HIVRT	DUD Actives	34	17
	DUD Decoys	1370	
	Drugs	4	
	PDB Ligands	29	
	WOMBAT	97	
P38	DUD Actives	135	20
	DUD Decoys	5416	
	Drugs	16	
	PDB Ligands	76	
	WOMBAT	52	
PDE5	DUD Actives	26	22
	DUD Decoys	1561	
	Drugs	12	
	PDB Ligands	10	
	WOMBAT	85	
$PPAR_{\gamma}$	DUD Actives	6	6
	DUD Decoys	38	
	Drugs	12	
	PDB Ligands	60	
	WOMBAT	27	

points (PPP) (as shown in Figure 1). To accomplish these tasks the ChemAxon's JChem[74] suite was used.

The next step is a conformational sampling along with partial charge calculations for each representation of the initial molecule. As the partial charge distribution affect the conformation and vice-versa, an iterative process would be the most accurate. But this approach is not viable when dealing with large datasets with thousands of molecules, due to its huge computational cost. Hence, a better approach was devised, trying to optimize CPU time without significant loss of precision.[75–77] Using OpenEye's QuacPac and Omega2[78] suites, the lowest-energy conformation is computed using the MMFF94s[79] force field, and its partial charges are determined using the semi-empirical AM1BCC[80] method. The last step is a conformational sampling, limited up to 200 conformers (for the maximum of 25 rotatable bonds), within an energy window of $5-10$ kcal/mol and an RMSD between 0.5-1.0 Å.

Pharmacophore Fingerprint Build

After the molecular treatment, each compound is represented by a large ensemble of several conformations, associated to a small number of tautomers and protonation states. The next step is a pharmacophoric mapping, performed by ChemAxon's PMapper.[74] This process is done atom-wise and assign at least one out of the following six PPP types to all heavy atoms in a molecule. The Aromatic (R) feature is assigned to any atom that is part of an aromatic ring. Hydrogen Bond Donor (D) and/or Hydrogen Bond Acceptor (A) are assigned to atoms able to establish hydrogen bonds with a potential target. Positively Charged (P) and Negatively Charged (N) are assigned to atoms with partial charges above $+0.4$ or below -0.4, respectively. Any other heavy atom that fails to fit in the classes above is assigned as Hydrophobic(H). All triplets formed by PPP's in 3D space are generated for each conformer. The Euclidean distance between each pair of points is discretized in ten distance bins (in Å): 0–3, 3–4.5, 4.5–6, 6–8, 8–10, 10–12.5, 12.5–15, 15–18, 18–21 and 21–∞. Each triplet is a putative pharmacophore formed by three heavy atoms with a PPP type assigned and a defined distance bin for each edge. This triplet is represented by a 6 character string

(3 characters for the feature on vertices and 3 for the distance bins on the edges) that identifies it univocally. Figure 1 illustrates how is represented within 3D-Pharma the pharmacophore triplet formed by a negatively charged Carboxyl group, a positively charged Amine group and Hydrogen donor Nitrogen in the structure of histidine.

Each conformation of a single molecule could have hundreds of three-point potential pharmacophores, so 3D-Pharma transforms the strings in indexes of a binary fingerprint using a hash function provided by the CMPH[81] library. These numeric fingerprints are analog to standard binary fingerprints, but instead of storing a very sparse array of bits, only the indexes of lit bits are stored. This decision was made due to the non-scalability of the binary vector size when increasing the number of nodes used in each tuplet (for example, when using tetrahedrons instead of triangles). Therefore, new operations are needed to substitute the binary operators AND and OR, since they are not applicable to a non-binary representation. Using Set Theory, the analogs to the two binary operations can be redefined: using Intersection ($\cap$) for AND and Union ($\cup$) for OR. The Tanimoto coefficient was used for similarity computation between two vectors using Set Theory operations. Given the fingerprints of two molecules A and B, the Tanimoto coefficient is given by:

$$T = \frac{|A \cap B|}{|A \cup B|}$$

.

Model Construction

In some LBVS applications, it is necessary to generate comprehensive queries that represent a pursued activity profile. This query should hold enough information to search and retrieve molecules with a potential activity from a large compound database. In 3D-Pharma, the query is a model built from a set of molecules previously known to be actives. A model (M) is formally defined by a set of pharmacophore triplets (x) that are present in the molecules that form the training set (T) in a frequency above a given threshold (τ). It can be formally defined as:

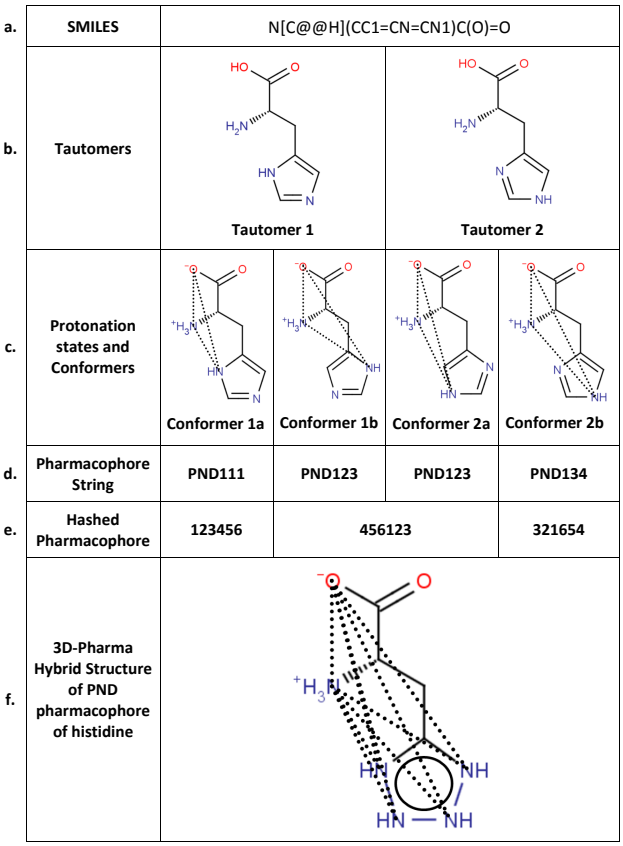


Figure 1: The effects of the proposed molecular treatment and generation of the 3D-Pharma pharmacophore triplets fingerprint. The amino acid histidine were selected as an example and all structures are depicted in 2D for clarity. **a)** The SMILES representation of the neutral form of histidine. **b)** The structures of the two dominant tautomers of histidine (pH between 0 and 14), showing the hydrogen exchange between the two atoms of nitrogen at imidazole ring. **c)** The structures of the major microspecies (protomers) of each tautomer of histidine at pH 7. For each protomer two hypothetical conformations are presented (solely the imidazole ring flip are considered). The dotted lines represent the pharmacophore triplet formed by one of the negative charged (N) oxygen atom at carboxyl group, the positive charged (P) alpha-nitrogen atom, and the hydrogen-bond donor (D) nitrogen atom at imidazole ring. The same pharmacophore triplet is monitored over all conformations **d)** the pharmacophore triplets of each conformation are converted to a string. PND stands for a triplet formed by a positive, a negative and hydrogen-bond donor pharmacophores. The numbers after the alphabetic string are indicatives of the distance between the atoms (see text). **e)** The hashed pharmacophore form derived from alphanumeric string. **f)** The hypothetical hybrid representation of the PND pharmacophore of histidine encoded by 3D-Pharma fingerprint. All pharmacophore triplets detected over all conformations, represented by the dotted lines, are equally considered

$$x \in M \Leftrightarrow \frac{\sum_{i=1}^{m} f(x,T_i)}{m} \geq \tau, f(x,T_i) = \begin{cases} 1 & \text{if } x \in T_i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

m being the size of the training set T, and T_i is the i$^{\text{th}}$ molecule in the set. We tried to optimize the performance by incrementing the value of τ by 0.1 ($0 \leq \tau \leq 1$). The gathered data (not shown here) suggested that a value of 0.7 is a generally optimal cutoff when treating single molecular target datasets.

In order to perform a virtual screening study, is necessary to have at least two datasets of active molecules: a training set and a test set. The former is used for model construction, which should be able to retrieve the latter among a set of inactive molecules. Although this is a widely used approach, it presents a major problem: how to split the active data between these two groups? The query construction is strongly affected by the training set selection. Besides, test and training groups should not be too similar to turn the classification problem into a trivial one, nor be too different, so molecules could show different profiles of activity or action mode.[82,83]

To build and validate the models, 3D-Pharma uses a new protocol inspired by the work of Tropsha[84,85] on validation problems of QSAR models. In his work, Tropsha argues that the model must be first exhaustively tested and validated internally before being used on external comparisons. In order to accomplish this, one should generate multiple training and test groups, and only the most internally consistent models should be considered to an external validation. 3D-Pharma splits the active molecules among ten groups, using the average 3D similarity between them to create homogeneous groups. These groups are used to build models in a stratified 10-fold cross-validation scheme where each group plays the role of test group once, and the remaining nine groups are recursively split between six training groups and three evaluation groups. Each training group is used to build a model, which is compared against the molecules in its correspondent evaluation group. Since 84 combinations of training/evaluation sets can be formed from nine groups, 84 models are generated. Of these, only the ten models with the highest average similarity to the molecules of its respective evaluation group are selected. These models are then used as a query to recover the

test group among a set of inactive molecules, and only the one which have the best recovery rate, measured by the area under the Receiver Operating Characteristic (ROC) curve is retained (Figure 2). Subsequently, the next group assumes the role of test group and the process is repeated. At the end, the full protocol generated a total of 840 models and produced 10 final models, one per test group. Henceforth, all results from 3D-Pharma presented here were averaged over these 10 models. When a dataset contains less than 10 molecules, is not possible to do a stratified 10-fold cross-validation and, in this case, a simple model is built without internal validation, considering all active molecules to be part of the Training Set.

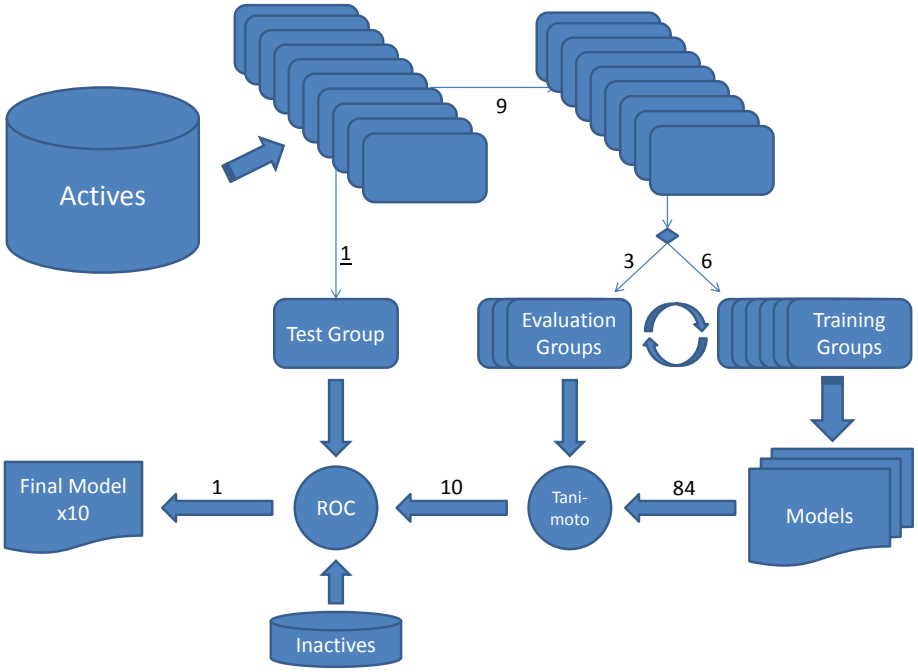


Figure 2: The 3D-Pharma exhaustive 10-fold cross validation scheme used to model construction and validation. Each group out of ten assumes the role of test group once, and the remaining groups are systematically split between training (six groups) and evaluation (three groups) sets. Each possible distribution generates a model from its training set to be compared against its correspondent evaluation set. There are 84 possible distributions and 84 models are generated. The 10 models most similar to its evaluation set are selected to a final validation against the test set. Only the highest predictive model for each test set is selected, resulting in 10 final models built by 3D-Pharma.

Metrics of Evaluation

To assess the performance of 3D-Pharma and also compare it against other available tools, a set of metrics were defined for measuring the performance of 3D-Pharma in VS applications. Three key features should be addressed when designing a VS method, and each one has its own set of metrics:

- *Accuracy* - The overall performance of a VS method. Can be easily quantified using the area under the Receiver-Operating Characteristic curve (AUC_{ROC}), which is a statically relevant and unbiased metric for classification performance assessment.[86]

- *Early Recognition* - The capacity of the VS method to recover active compounds at early cuts. The AUC_{ROC}'s capacity to assess the "early recognition" has been criticized,[87] since whenever a true positive is found, it's contribution to the final score is proportionally the same regardless of ranking position. Usually results published elsewhere rely on Enrichment Factor (EF) of selected cuts to assess the "Early Recognition" problem. However this metric is not suitable as it depends on the size of the database and on the actives/inactives ratio. Even on standardized databases, unbiased metrics are preferable. In search of a better metric, Truchon and Bayly [87] generalized the Receiver-Operator Characteristic and designed the parametric Boltzmann-Enhanced Discrimination ROC curve ($BEDROC_{\alpha}$). The α parameter is used to specify the range of the ranked list that would contribute the most to the overall score. In their work, Truchon and Bayly formalized this relation and suggested some α values. Within these suggestions, values of $\alpha = 160.9$, 32.2 and 20, which corresponds respectively to an EF at 1%, 5% and 8% of the selection, were chosen.

- *Scaffold Hoping* - The ability of the VS method to find novel (or diverse) molecular scaffolds. To account for scaffold hoping in retrospective studies, one would need to cluster the actives into groups of similar molecular structures. Since the DUD database has already clustered its active molecules, it is straightforward to apply the metrics. The arithmetic weighting of the ROC curve (awROC)[88] was used to assess scaffold hopping capabilities, which weights

the ROC curve to take into account cluster information. A true positive influence in the score is inversely proportional to the size of the cluster that it is inserted into, so early recognition of low represented clusters contributes more to the final score.

Methods in LBVS used for comparison

The performance of 3D-Pharma was compared to those of other LBVS methods that also used DUD as a benchmarking dataset and its authors supplies enough data to support the full comparison.

Optimal Assignment methods

Optimal Assignment is a graph theory optimization problem. Given a bipartite graph where each node is linked to another node by a weighted edge, an optimal assignment is a graph-matching where the sum of the edges are maximized. This was first applied in molecular similarity by Fröhlich and co-workers,[89,90] when they created the OAK (Optimal Assignment Kernel) method. OAK_{FLEX}[91] was a modification of OAK made by Fechner et al that included conformational space similarity into the calculations. Other implementations of algorithms which mapped the optimal assignment problem into molecular similarity measures include 2SHA (Two-Step Hierarchical Assignment)[68] and OAAP(Optimal Local Atom Pair Environment Assignment).[68]

4D FAP_{OA}

4D FAP_{OA}[59] generates a very large ensemble of conformations whose atom-pair distance profile are encoded in a series of Gaussian Mixed Models (GMM) generating a single probabilistic model. The energy of conformations is used as a weight factor of each measured atom-pair distance in the GMM generation. The complete information of the conformational space of a molecule is encoded into a list of Gaussian mixture models that could be used to compare different molecules without the need of original conformational ensemble. The final similarity value is computed through an optimal assignment algorithm over atom-pairs in a distance matrix.

FLAP

FLAP (Fingerprints for Ligands and Proteins)[46–49] is a well known 3D fingerprint tool that utilizes molecular interaction fields (MIFs) from both ligand and target structures, generated by the program GRID.[92] Each grid point with a local maximal value of the MIF generates a pharmacophoric point of the type of the probe used to generate the MIF and all tetrahedrons formed by these points are stored in a fingerprint. A similarity search is made considering the fingerprints and the alignment of the query molecule to the template. In the recent work by Cross et al,[49] data fusion techniques were used over several data sources to improve recall rates. Of these, LBtParetoR and LBOpt were the best ligand-based techniques reported. LBtParetoR uses a recursive Pareto sum ranking of the alignments, using the DUD cluster representatives (DUD-Parents) as templates. The LBOpt mode uses information of inactive compounds to choose among the DUD-Parents the best template.

FieldScreen

FieldScreen[93] computes molecular field points around a "relevant" conformation of the query molecule and search a multiconformer database for matching patterns, using maximal colored cliques of field points for the alignment.

Results and discussion

All compounds included in this study (DUD Actives, DUD Decoys, Drugs, PDB Ligands and WOMBAT entries) were submitted to the same protocol of molecular treatment. A set of models was built from the molecules of each active dataset (except COX-2 PDB Ligands, FX_{α} Drugs and HIVRT Drugs, which had less than 10 molecules in the dataset). Each model was used as query in a similarity search against a pool of molecules formed by DUD Actives and DUD Decoys. The resulting ranking was evaluated through the metrics aforementioned and compared to the LBVS methods mentioned previously.

As shown in Figure 4, only 3D-Pharma and 4D FAP_{OA} had data comprising all 10 targets included in this study. The other techniques mentioned above only made data available for 13 targets with high chemotype diversity (greater than 15 classes). Within these, there are seven targets in common with our selection : CDK2, COX-2, EGFR, FX_{α}, HIVRT, P38 and PDE5. Hence, all averaged data on the LBVS methods depicted in Figure 5 and Figure 6 (except for 3D-Pharma and 4D FAP_{OA}) are averaged across these seven targets.

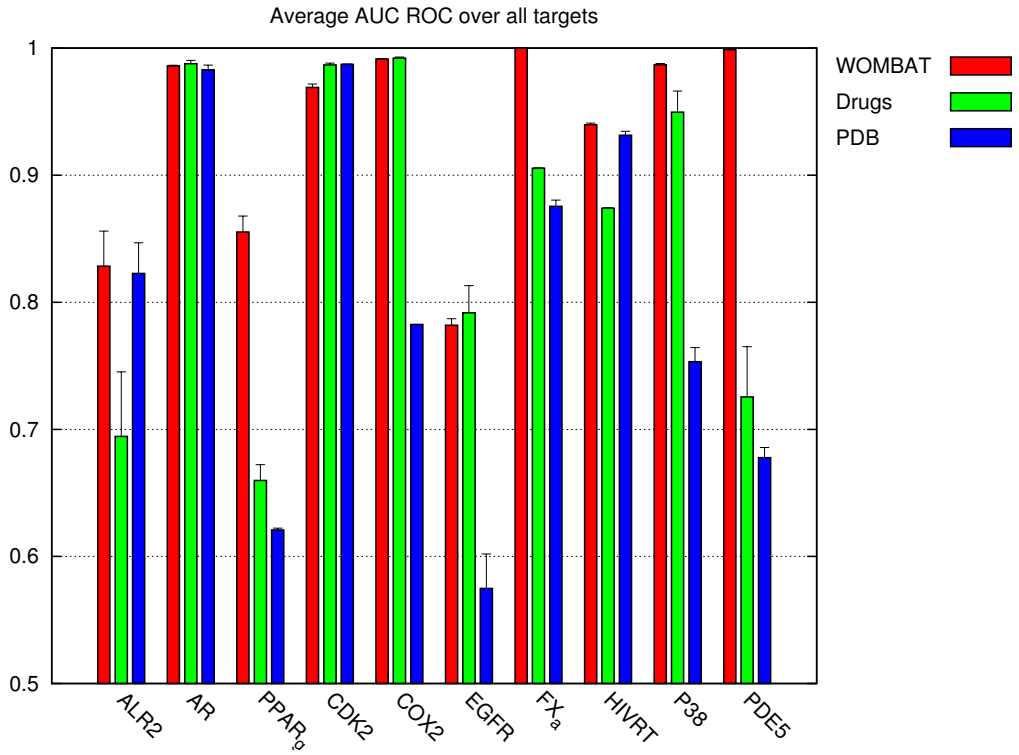


Figure 3: Average AUC_{ROC} produced by 3D-Pharma for all targets. For each dataset, the AUC_{ROC} value was averaged over the 10 models derived from the 10-fold Cross-validation protocol. The error bars correspond to the calculated standard deviation for each dataset.

As seen in Figure 3, 3D-Pharma had an excellent overall accuracy, with 20 out of 30 models with AUC_{ROC} above 0.8. Of these, 14 had AUC_{ROC} above 0.9. As for the targets, seven out of ten had at least one dataset with AUC_{ROC} above 0.9, with nine out of ten targets with at least one model with AUC_{ROC} above 0.8. The models constructed from the WOMBAT database presented the best accuracy, with average AUC_{ROC} of 0.93 ± 0.08, compared to the other datasets (Drugs $AUC_{ROC} = 0.85 \pm 0.12$ and PDB $AUC_{ROC} = 0.80 \pm 0.14$). This might be due to the fact that

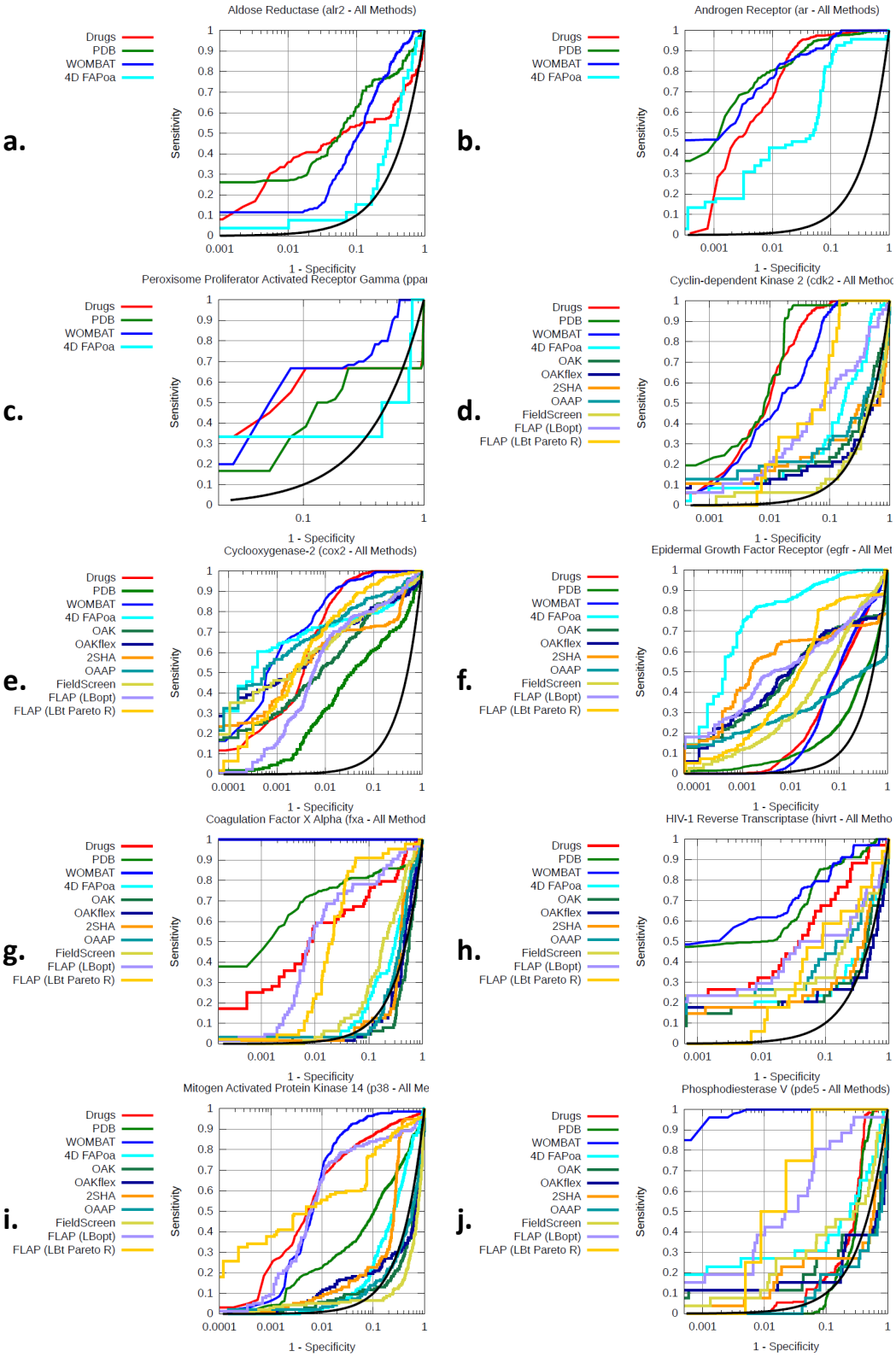


Figure 4: Logarithmic ROC plots of all analysed Ligand-Based Virtual Screening tools against the DUD datasets: **a)** AR **b)** ALR2 **c)** PPAR$_\gamma$ **d)** CDK2 **e)** COX2 **f)** EGFR **g)** FX$_\alpha$ **h)** HIVRT **i)** P38 **j)** PDE5

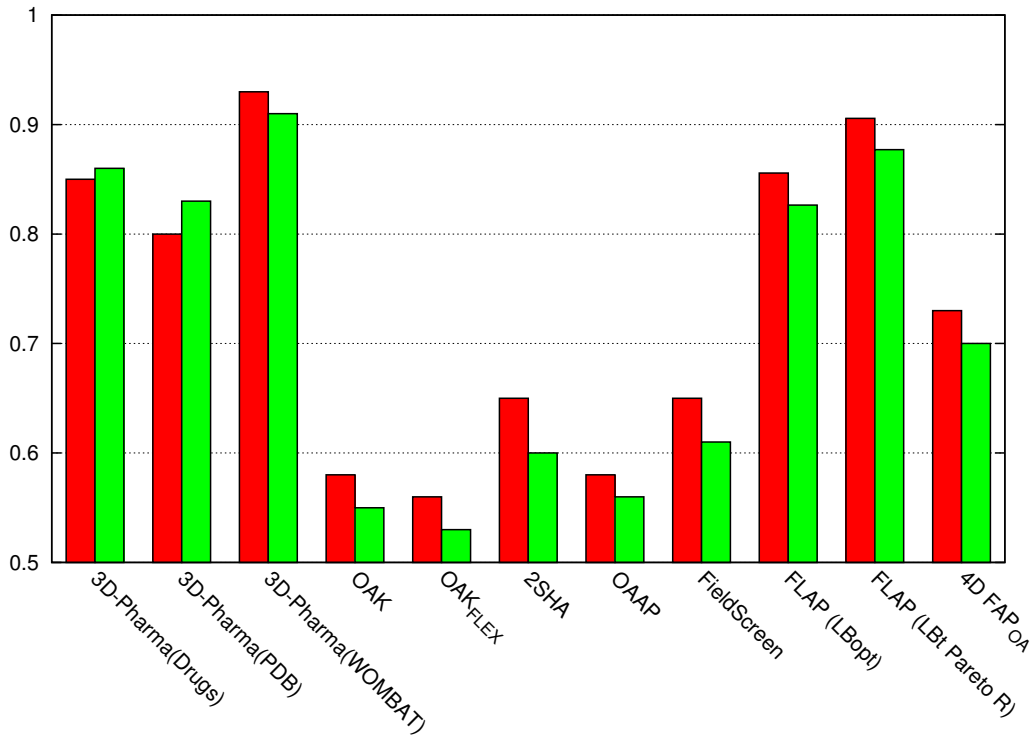


Figure 5: Average AUC_{ROC} and AUC_{awROC} over 10 selected targets for 3D-Pharma (datasets WOMBAT, Drugs and PDB Ligands) and 4D FAP_{OA}. The results of the remaining six Ligand-Based Virtual Screening tools were averaged over seven targets. It is worth of note that all methods, except 3D-Pharma with PDB ligands and Drugs datasets, show an decrease the area under the curve (AUC) from ROC to awROC.

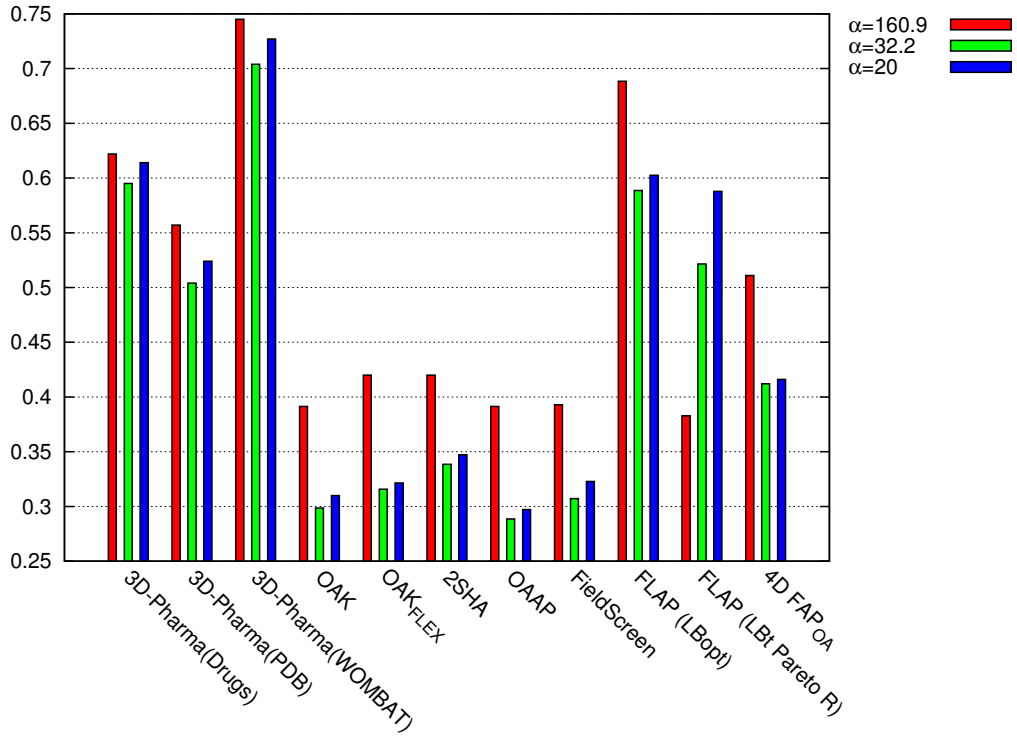


Figure 6: Average BEDROC$_\alpha$ scores over 10 selected targets for 3D-Pharma (datasets WOMBAT, Drugs and PDB Ligands) and 4D FAP$_{\mathrm{OA}}$. The results of the remaining six Ligand-Based Virtual Screening tools were averaged over seven targets. The values for the α parameter were 160.9, 32.2 and 20, which correspond to Enrichment Fators (EF) at 1%, 5% and 8% of selection, respectively.

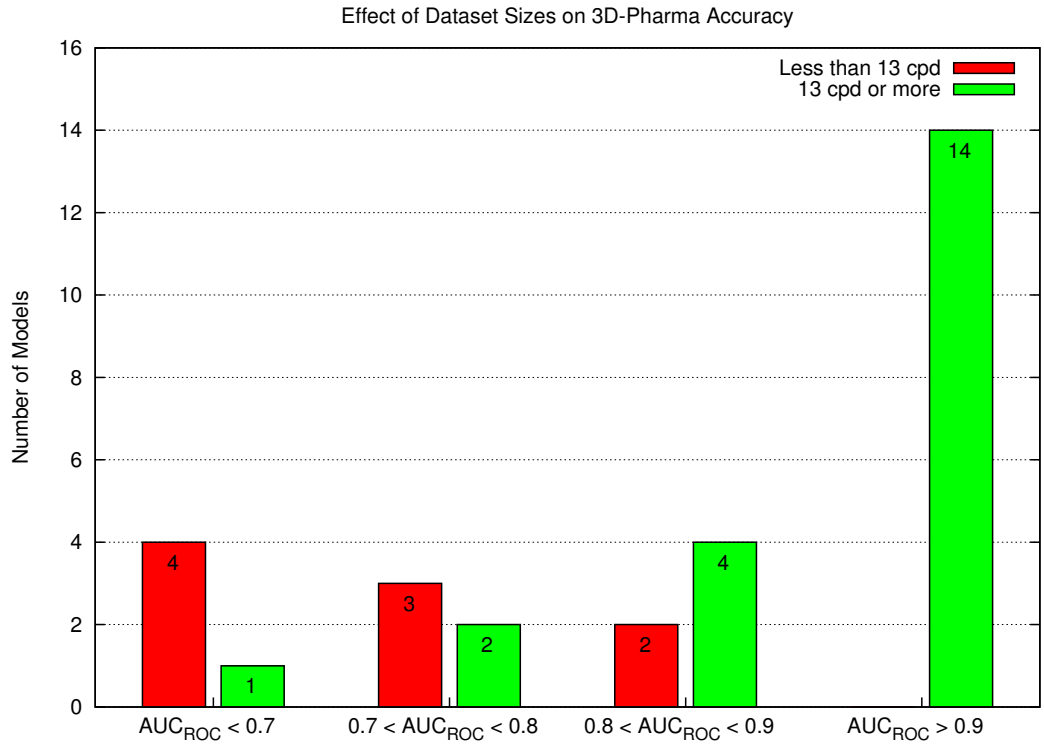


Figure 7: The effect of dataset sizes on 3D-Pharma accuracy. For datasets with 13 or more compounds most datasets (14 out of 21) produced high quality models (AUC$>$ 0.9). For those datasets with 12 or less compounds, two out of nine datasets produced good models (AUC$>$ 0.8).

WOMBAT is a solid well stablished database, whereas the Drugs and PDB are still a growing set of gathered data. Beyond the nature and quality of datasets used to build the models, it seems that the size of the datasets shows the major impact on the performance of 3D-Pharma (Figure 7). For those datasets with at least 13 compounds (21 datasets), not less than 14 (67%) produced high quality models with AUC_{ROC} above 0.9, and 18 (86%) produced very good models with AUC_{ROC} above 0.8. On the contrary, for those datasets with less than 13 compounds, none was able to generate models with AUC_{ROC} above 0.9 and only two out of nine (22%) generated models with AUC_{ROC} above 0.8.

In general, 3D-Pharma outperforms the other methods regardless of the dataset chosen to build the models, with FLAP as a close second. However, despite of results achieved, both FLAP ligand-based approaches are subject of one drawback: the use of DUD cluster-parent actives as query or templates. Therefore, the analysis is subject of *analogue bias*[67] that could potentially increment artificially the results. Another interesting observation arises when BEDROC scores are analyzed. When one looks over the bigger cuts ($\alpha = 32.2$ and 20), all methods analyzed have an abrupt fall in the early recovery rate, as seen in Figure 6, but 3D-Pharma fairly maintains its scores as α diminishes. It seems that the higher $BEDROC_{160.9}$ score with a subsequent substantial decrease on $BEDOC_{32.2}$ is an evidence of some kind of *analog bias* that puts a few actives compounds very high in the rank (before 1% of selection) and leaves many other active molecules spread over the rank positions. In 3D-Pharma we used three really external public datasets, used "as is", that is, without any kind of filter, except for redundant molecules between the DUD actives and external datasets. As a consequence the BEDROC score is sustained over all α values used in the benchmark.

LBOpt and LBParetoR FLAP scenarios present very good overall results and perform better than the other methods surveyed, except for 3D-Pharma. LBParetoR uses all DUD chemotype cluster parents (DUD-own dataset) for each target as query in an ensemble approach with consensus analysis of the individual template similarity results. Consequently, the number of DUD actives in the ROC analysis is smaller than all other methods under comparison in this paper, and

the conclusions must be seen carefully. When we look at the ROC and awROC from LBParetoR, the results are impressive, with average AUC above 0.9 for the seven targets under scrutiny. However, the analysis of BEDROC results discloses a disappointing "early recovery", mainly for $BEDROC_{160.9}$ and $BEDROC_{32.2}$. The LBOpt approach produces smaller AUC in the ROC analysis than LBParetoR, but the "early recovery" is much better. LBopt also uses the DUD-Parents as query, but they are subject to a previous *optimal template selection* that optimize proportions of false positives and false negatives in order to select the single template to be used as query.

When analyzing scaffold hopping capabilities, one can note that all techniques, except 3D-Pharma, have a significant drop in their average AUCs when considering awROC over the standard ROC (Figure 5). On the contrary, 3D-Pharma using the Drugs and PDB-Ligands datasets have an increase in the score. The other techniques tend to rank higher the most populated scaffolds, hence lowering AUC_{awROC} scores in relation to AUC_{ROC}.

Conclusions

The main characteristics of 3D-Pharma are the use of pharmacophore triplets fingerprint based on atom-centered potential pharmacophores, the use of several representations of the compound that include tautomers, protonation states and conformers and the *ensemble template* approach producing a single modal fingerprint based on frequency of pharmacophore triplets over the active compounds. This dynamic pharmacophore fingerprint encodes all chemical and conformational variability of the compound in a single fingerprint representation. Thus, the 3D-Pharma approach adopts a paradigm where the full ensemble of conformers is taken into account at the same time in a single modal fingerprint, similarly to the approach implemented by Ranu and Singh.[60] In our study, 30 sets of models were generated for 10 selected targets from the DUD database, using three external and independent datasets as reference. Is seems that the size of the modeling dataset exerts the major impact on the model quality. For those datasets with at least 13 compounds (21 datasets) not less than 18 (or 86%) were able to produce very good models with AUC_{ROC} above 0.8, and 14

datasets (67%) produced very high quality models with an AUC_{ROC} above 0.9.

The analysis of the scaffold hopping and early recovery capabilities of 3D-Pharma has shown two distinguishing behaviors. The AUC_{awROC}, used to estimate scaffold hopping capabilities, shows evidences that 3D-Pharma with Drugs and PDB Ligands datasets has a better performance in detecting rarer scaffolds than other LBVS tools analyzed. The second 3D-Pharma dicerning behavior can be seen in the BEDROC plot (Figure 6) were 3D-Pharma datasets sustain high scores over the three values chosen for the α parameter. All other LBVS tools (except FLAP LBParetoR) presented a higher score at lower cuts ($BEDROC_{160.9}$) than those at higher cuts, with a significant decrease at the $BEDROC_{32.2}$ and $BEDROC_{20}$ scores.

Thus, the data shown here lead us to strongly believe that 3D-Pharma overperforms all other state-of-the-art LBVS tools analyzed, in terms of global accuracy as well as scaffold hopping and early recovery capacities. The fact that three really external datasets were used to generate the models that are at the same time simple, robust, consistent and predictive should be highlighted. It remains to be seen its predictive power in prospective virtual screening cases, but as far as the results shown here can assess, 3D-Pharma is a promising method that can effectively contribute to the success of any drug discovery process.

Acknowledgement

We would like to acknowledge the following agencies for their support: CNPq, CAPES and FAPEMIG. We would also like to thanks ChemAxon and OpenEye Scientific Software for providing us academic licenses to their products.

Supporting Information Available

Supporting Information contains the ROC plots and complete tables comparing all methods for each target. It also contains the molecular files (in SMILES) for Drugs and PDB-Ligands datasets for each target, and the WOMBAT IDs considered on this study. The full ranking od DUD molecules from each model is also available within the files. Please refer to the `README.txt`

file for more details.

References

(1) Paul, S.; Mytelka, D.; Dunwiddie, C.; Persinger, C.; Munos, B.; Lindborg, S.; Schacht, A. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery* **2010**, *9*, 203–214.

(2) IMAP's Pharma & Biotech Industry Global Report 2011. `http://www.imap.com/imap/media/resources/IMAP_PharmaReport_8_272B8752E0FB3.pdf`, 2011; Accessed 05/01/2012.

(3) Koenig, J. Does process excellence handcuff drug development? *Drug Discovery Today* **2011**, *16*, 377 – 381.

(4) Harrison, C. Patent Watch: The patent cliff steepens. *Nature Reviews Drug Discovery* **2011**, *10*, 12–13.

(5) Arrowsmith, J. A decade of change. *Nat Rev Drug Discov* **2012**, *11*, 17–18.

(6) Mucsi, Z.; Csizmadia, I. The Future of the Drug Discovery Process and the Fate of the Pharmaceutical Industry: An economical and scientific study. *Philosophic Nature* **2009**, *1*.

(7) Mayr, L. M.; Fuerst, P. The Future of High-Throughput Screening. *Journal of Biomolecular Screening* **2008**, *13*, 443–448.

(8) Kümmel, A.; Parker, C. N. In *Chemoinformatics and Computational Chemical Biology*; Bajorath, J., Ed.; Springer Science+Business Media, LLC 2011, 2011; Chapter 17, pp 435–457.

(9) Bajorath, J. Integration of virtual and high-throughput screening. *Nature Reviews Drug Discovery* **2002**, *1*, 882–894.

(10) Ripphausen, P.; Nisius, B.; Bajorath, J. State-of-the-art in ligand-based virtual screening. *Drug Discovery Today* **2011**, *16*, 372 – 376.

(11) Cortés-Cabrera, Á.; Gago, F.; Morreale, A. A reverse combination of structure-based and ligand-based strategies for virtual screening. *Journal of Computer-Aided Molecular Design* **2012**, *26*, 319–327.

(12) Svensson, F.; Karlén, A.; Sköld, C. Virtual Screening Data Fusion Using Both Structure- and Ligand-Based Methods. *Journal of Chemical Information and Modeling* **2012**, *52*, 225–232.

(13) Swann, S. L.; Brown, S. P.; Muchmore, S. W.; Patel, H.; Merta, P.; Locklear, J.; Hajduk, P. J. A Unified, Probabilistic Framework for Structure- and Ligand-Based Virtual Screening. *Journal of Medicinal Chemistry* **2011**, *54*, 1223–1232.

(14) Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *Journal of Medicinal Chemistry* **2010**, *53*, 539–558.

(15) Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *Journal of Chemical Information and Modeling* **2010**, *50*, 205–216.

(16) Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo Vadis, Virtual Screening? A Comprehensive Survey of Prospective Applications. *Journal of Medicinal Chemistry* **2010**, *53*, 8461–8467.

(17) Arrowsmith, J. Trial watch: Phase II failures: 2008-2010. **2011**, *10*, 328–329.

(18) Colburn, W. Biomarkers in drug discovery and development: from target identification through drug marketing. *The Journal of Clinical Pharmacology* **2003**, *43*, 329–341.

(19) Virtanen, S.; Pentikäinen, O. Efficient virtual screening using multiple protein conformations described as negative images of the ligand-binding site. *Journal of Chemical Information and Modeling* **2010**, *50*, 1005–1011.

(20) Koutsoukas, A.; Simms, B.; Kirchmair, J.; Bond, P.; Whitmore, A.; Zimmer, S.; Young, M.; Jenkins, J.; Glick, M.; Glen, R.; Bender, A. From in silico target prediction to multi-target

drug design: Current databases, methods and applications. *Journal of Proteomics* **2011**, *74*, 2554–2574.

(21) Marrero-Ponce, Y.; Siverio-Mota, D.; Gálvez-Llompart, M.; Recio, M. C.; Giner, R. M.; García-Domènech, R.; Torrens, F.; Arán, V. J.; Cordero-Maldonado, M. L.; Esguera, C. V.; de Witte, P. A.; Crawford, A. D. Discovery of novel anti-inflammatory drug-like compounds by aligning in silico and inăvivo screening: The nitroindazolinone chemotype. *European Journal of Medicinal Chemistry* **2011**, *46*, 5736 – 5753.

(22) Bottegoni, G.; Favia, A. D.; Recanatini, M.; Cavalli, A. The role of fragment-based and computational methods in polypharmacology. *Drug Discovery Today* **2012**, *17*, 23 – 34.

(23) Ekins, S.; Mestres, J.; Testa, B. *In silico* pharmacology for drug discovery: methods for virtual ligand screening and profiling. *British Journal of Pharmacology* **2007**, *152*, 9–20.

(24) Bajorath, J. Computational analysis of ligand relationships within target families. *Current Opinion in Chemical Biology* **2008**, *12*, 352–358.

(25) Luis G. Valerio, J. *In silico* toxicology for the pharmaceutical sciences. *Toxicology and applied pharmacology* **2009**, *241*, 356–370.

(26) Maggiora, G. The reductionist paradox: are the laws of chemistry and physics sufficient for the discovery of new drugs? *Journal of Computer-Aided Molecular Design* **2011**, *25*, 699–708.

(27) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug discovery today* **2006**, *11*, 1046–1053.

(28) Venkatraman, V.; Pérez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data set Reveals Limitations of Current 3D Methods. *Journal of Chemical Information and Modeling* **2010**, *50*, 2079–2093.

(29) Stumpfe, D.; Bajorath, J. In *Virtual Screening*; Mannhold, R., Kubinyi, H., Folkers, G., Eds.; Wiley-VCH Verlag GmbH & Co. KGaA, 2011; Chapter 11, pp 291–318.

(30) Sitzmann, M.; Weidlich, I. E.; Filippov, I. V.; Liao, C.; Peach, M. L.; Ihlenfeldt, W.-D.; Karki, R. G.; Borodina, Y. V.; Cachau, R. E.; Nicklaus, M. C. PDB Ligand Conformational Energies Calculated Quantum-Mechanically. *Journal of Chemical Information and Modeling* **2012**, *52*, 739–756.

(31) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. i. E. The Protein Data Bank. *Nucl. Acids Res.* **2000**, *28*, 235–242.

(32) Mattos, C.; Ringe, D. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; Escom, 1993; pp 226–256.

(33) Lewis, P. et al. On the detection of multiple-binding modes of ligands to proteins, from biological, structural, and modeling data. *Journal of Computer-Aided Molecular Design* **2003**, *17*, 129–134.

(34) DePristo, M. A.; de Bakker, P. I. W.; Blundell, T. L. Heterogeneity and Inaccuracy in Protein Structures Solved by X-Ray Crystallography. *Structure* **2004**, *12*, 831 – 838.

(35) Steuber, H.; Zentgraf, M.; Gerlach, C.; Sotriffer, C.; Heine, A.; Klebe, G. Expect the unexpected or caveat for drug designers: multiple structure determinations using aldose reductase crystals treated under varying soaking and co-crystallisation conditions. *Journal of Molecular Biology* **2006**, *363*, 174–187.

(36) Tresadern, G.; Bemporad, D.; Howe, T. A comparison of ligand based virtual screening methods and application to corticotropin releasing factor 1 receptor. *Journal of Molecular Graphics and Modelling* **2009**, *27*, 860–870.

(37) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of Common Functional Con-

figurations Among Molecules. *Journal of Chemical Information and Computer Sciences* **1996**, *36*, 563–571.

(38) Richmond, N.; Abrams, C.; Wolohan, P.; Abrahamian, E.; Willett, P.; Clark, R. GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D. *Journal of Computer-Aided Molecular Design* **2006**, *20*, 567–587.

(39) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *Journal of Computer-Aided Molecular Design* **1995**, *9*, 532–549.

(40) Chemical Computing Group, Molecular Operating Environment. `www.chemcomp.com`.

(41) Dixon, S.; Smondyrev, A.; Knoll, E.; Rao, S.; Shaw, D.; Friesner, R. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *Journal of Computer-Aided Molecular Design* **2006**, *20*, 647–671.

(42) Schneidman-Duhovny, D.; Dror, O.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. Deterministic Pharmacophore Detection via Multiple Flexible Alignment of Drug-Like Molecules. *Journal of Computational Biology* **2008**, *15*, 737–754.

(43) Dror, O.; Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. Novel Approach for Efficient Pharmacophore-Based Virtual Screening: Method and Applications. *Journal of Chemical Information and Modeling* **2009**, *49*, 2333–2343, PMID: 19803502.

(44) Cottrell, S. J.; Gillet, V. J.; Taylor, R.; Wilton, D. J. Generation of multiple pharmacophore hypotheses using multiobjective optimisation techniques. *Journal of Computer-Aided Molecular Design* **2004**, *18*, 665–682.

(45) Kristam, R.; Gillet, V. J.; Lewis, R. A.; Thorner, D. Comparison of Conformational Analysis Techniques To Generate Pharmacophore Hypotheses Using Catalyst. *Journal of Chemical Information and Modeling* **2005**, *45*, 461–476.

(46) Perruccio, F.; Mason, J. S.; Sciabola, S.; Baroni, M. In *Molecular Interaction Fields: Applications in Drug Discovery and ADME Prediction*; Cruciani, G., Ed.; Wiley-VCH: Weinheim, Germany, 2006; Chapter 4.

(47) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins (FLAP): Theory and Application. *J Chem Info Model* **2007**, *47*, 279–294.

(48) Carosati, E.; Sforna, G.; Pippi, M.; Marverti, G.; Ligabue, A.; Guerrieri, D.; Piras, S.; Guaitoli, G.; Luciani, R.; Costi, M. P.; Cruciani, G. Ligand-based virtual screening and ADME-tox guided approach to identify triazolo-quinoxalines as folate cycle inhibitors. *Bioorganic & Medicinal Chemistry* **2010**, *18*, 7773 – 7785.

(49) Cross, S.; Baroni, M.; Carosati, E.; Benedetti, P.; Clementi, S. FLAP: GRID Molecular Interaction Fields in Virtual Screening. Validation using the DUD Data Set. *Journal of Chemical Information and Modeling* **2010**, *50*, 1442–1450.

(50) Koes, D. R.; Camacho, C. J. Pharmer: Efficient and Exact Pharmacophore Search. *Journal of Chemical Information and Modeling* **2011**, *51*, 1307–1314.

(51) Tripos, L.P., Tuplets. `www.tripos.com`.

(52) Accelrys Software, Cerius^2. `www.accelrys.com`.

(53) de Benedetti, P.; Fanelli, F. Computational quantum chemistry and adaptive ligand modeling in mechanistic QSAR. *Drug Discovery Today* **2010**, *15*, 859–866.

(54) Schneider, G. Virtual screening: an endless staircase? *Nat Rev Drug Discov* **2010**, *9*, 273–276.

(55) Kubinyi, H. Drug research: myths, hype and reality. *Nature Reviews Drug Discovery* **2003**, *2*, 665–667.

(56) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *Journal of Chemical Information and Modeling* **2012**, *52*, 867–881.

(57) Natesan, S.; Wang, T.; Lukacova, V.; Bartus, V.; Khandelwal, A.; Balaz, S. Rigorous Treatment of Multispecies Multimode Ligand-Receptor Interactions in 3D-QSAR: CoMFA Analysis of Thyroxine Analogs Binding to Transthyretin. *Journal of Chemical Information and Modeling* **2011**, *51*, 1132–1150.

(58) Natesan, S.; Subramaniam, R.; Bergeron, C.; Balaz, S. Binding Affinity Prediction for Ligands and Receptors Forming Tautomers and Ionization Species: Inhibition of Mitogen-Activated Protein Kinase-Activated Protein Kinase 2 (MK2). *Journal of Medicinal Chemistry* **2012**, *55*, 2035–2047.

(59) Jahn, A.; Rosenbaum, L.; Hinselmann, G.; Zell, A. 4D Flexible Atom-Pairs: An efficient probabilistic conformational space comparison for ligand-based virtual screening. *Journal of Cheminformatics* **2011**, *3*, 23.

(60) Ranu, S.; Singh, A. K. Novel Method for Pharmacophore Analysis by Examining the Joint Pharmacophore Space. *Journal of Chemical Information and Modeling* **2011**, *51*, 1106–1121.

(61) Pérez-Nueno, V. I.; Ritchie, D. W. Using Consensus-Shape Clustering To Identify Promiscuous Ligands and Protein Targets and To Choose the Right Query for Shape-Based Virtual Screening. *Journal of Chemical Information and Modeling* **2011**, *51*, 1233–1248.

(62) Sastry, G. M.; Dixon, S. L.; Sherman, W. Rapid Shape-Based Ligand Alignment and Virtual Screening Method Based on Atom/Feature-Pair Similarities and Volume Overlap Scoring. *Journal of Chemical Information and Modeling* **2011**, *51*, 2455–2466.

(63) Cai, C.; Gong, J.; Liu, X.; Jiang, H.; Gao, D.; Li, H. A novel, customizable and optimizable parameter method using spherical harmonics for molecular shape similarity comparisons. *Journal of Molecular Modeling* **2012**, *18*, 1597–1610.

(64) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J Med Chem* **2006**, *49*, 6789–6801.

(65) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 1177–1185.

(66) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: New York, 2004; Chapter 9, pp 223–239.

(67) Good, A.; Oprea, T. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *Journal of Computer-Aided Molecular Design* **2008**, *22*, 169–178.

(68) Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal assignment methods for ligand-based virtual screening. *Journal of Cheminformatics* **2009**, *1*, 14.

(69) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 1308–1315.

(70) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* **2006**, *34*, D668–D672.

(71) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledge base for drugs, drug actions and drug targets. *Nucl. Acids Res.* **2008**, *36*, D901–906.

(72) Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acids Res.* **2000**, *28*, 27–30.

(73) Chen, X.; Ji, Z. L.; Chen, Y. Z. TTD: Therapeutic Target Database. *Nucl. Acids Res.* **2002**, *30*, 412–415.

(74) ChemAxon, `www.chemaxon.com`, 1999–2012.

(75) Mobley, D. L.; Dumont, É.; Chodera, J. D.; Dill, K. A. Comparison of Charge Models for Fixed-Charge Force Fields: Small-Molecule Hydration Free Energies in Explicit Solvent. *The Journal of Physical Chemistry B* **2007**, *111*, 2242–2254.

(76) Tsai, K.-C.; Wang, S.-H.; Hsiao, N.-W.; Li, M.; Wang, B. The effect of different electrostatic potentials on docking accuracy: A case study using DOCK5.4. *Bioorganic & Medicinal Chemistry Letters* **2008**, *18*, 3509 – 3512.

(77) Wang, J.-C.; Lin, J.-H.; Chen, C.-M.; Perryman, A. L.; Olson, A. J. Robust Scoring Functions for Protein-Ligand Interactions with Quantum Chemical Charge Models. *Journal of Chemical Information and Modeling* **2011**, *51*, 2528–2537.

(78) OpenEye Scientific Software, OMEGA. `www.eyesopen.com`, 1997–2012.

(79) Halgren, T. A. MMFF VI. MMFF94s option for energy minimization studies. *Journal of Computational Chemistry* **1999**, *20*, 720–729.

(80) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry* **2002**, *23*, 1623–1641.

(81) Botelho, F. C.; Ziviani, N. External perfect hashing for very large key sets. CIKM '07: Proceedings of the sixteenth ACM Conference on information and knowledge management. Lisbon, Portugal, 2007; pp 653–662.

(82) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Journal of Computer-Aided Molecular Design* **2002**, *16*, 357–369.

(83) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.-D.; Lee, K.-H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *Journal of Computer-Aided Molecular Design* **2003**, *17*, 241–253.

(84) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR & Combinatorial Science* **2003**, *22*, 69–77.

(85) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **2010**, *29*, 476–488.

(86) Nicholls, A. In *Chemoinformatics and Computational Chemical Biology*; Bajorath, J., Ed.; Springer Science+Business Media, LLC 2011, 2011; Chapter 22, pp 531–581.

(87) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *Journal of Chemical Information and Modeling* **2007**, *47*, 488–508.

(88) Clark, R.; Webster-Clark, D. Managing bias in ROC curves. *Journal of Computer-Aided Molecular Design* **2008**, *22*, 141–146.

(89) Fröhlich, H.; Wegner, J. K.; Sieker, F.; Zell, A. Optimal assignment kernels for attributed molecular graphs. ICML. 2005; pp 225–232.

(90) Fröhlich, H.; Wegner, J. K.; Sieker, F.; Zell, A. Kernel Functions for Attributed Molecular Graphs - A New Similarity-Based Approach to ADME Prediction in Classification and Regression. *QSAR & Combinatorial Science* **2006**, *25*, 317–326.

(91) Fechner, N.; Jahn, A.; Hinselmann, G.; Zell, A. Atomic Local Neighborhood Flexibility Incorporation into a Structured Similarity Measure for QSAR. *Journal of Chemical Information and Modeling* **2009**, *49*, 549–560.

(92) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* **1985**, *28*, 849–857.

(93) Cheeseright, T. J.; Mackey, M. D.; Melville, J. L.; Vinter, J. G. FieldScreen: Virtual Screening Using Molecular Fields. Application to the DUD Data Set. *Journal of Chemical Information and Modeling* **2008**, *48*, 2108–2117.

This material is available free of charge via the Internet at `http://pubs.acs.org/`.