UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

Fernando Ribeiro Caires

Estudos e desenvolvimento de métodos baseados em harmônicos esféricos para análise de
similaridade estrutural entre ligantes

São Carlos

2016

Fernando Ribeiro Caires

Estudos e desenvolvimento de métodos baseados em harmônicos esféricos para análise de similaridade estrutural entre ligantes

Dissertação apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de Mestre em Ciências.

Área de concentração: Física Aplicada

Opção: Física Biomolecular

Orientador: Prof. Dr. Rinaldo Wander Montalvão

Versão Corrigida
(Versão original disponível na Unidade que aloja o Programa)

São Carlos
2016

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Ficha catalográfica revisada pelo Serviço de Biblioteca e Informação do IFSC,
com os dados fornecidos pelo(a) autor(a)

Caires, Fernando Ribeiro

Estudos e desenvolvimento de métodos baseados em harmônicos esféricos para análise de similaridade estrutural entre ligantes / Fernando Ribeiro Caires; orientador Rinaldo Wander Montalvão - versão corrigida -- São Carlos, 2016.
130 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Física Biomolecular) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2016.

1. Similaridade . 2. Harmônicos Esféricos. 3. Ligantes. I. Montalvão, Rinaldo Wander, orient. II. Título.

Aos meus pais e minha namorada com amor,
gratidão e admiração pelo apoio, compreensão,
carinho e a presença durante a elaboração desse
trabalho.

AGRADECIMENTOS

Ao corpo docente do curso de Ciências Físicas e Biomoleculares e do programa de pós-graduação em Física Biomolecular que contribuíram de forma essencial para meu crescimento científico, intelectual, pessoal e profissional.

Ao Dr. Rinaldo Wander Montalvão, pelo apoio, atenção, direcionamento e orientação durante o mestrado, sempre disponível em ajudar, indicando os melhores caminhos a serem trilhados na elaboração do trabalho além da ajuda na parte teórica como na parte prática do meu trabalho.

Aos meus pais, Ângela e José Lúcio, que sempre me deram amor, educação, carinho e me apoiaram em minhas decisões com a tentativa de me guiar pelo melhor caminho e na busca pela minha realização pessoal e profissional.

À minha namorada, Giovana, que sempre esteve ao meu lado em todos os momentos dessa etapa de minha vida me ajudando e fornecendo amor, carinho e companheirismo.

Aos meus amigos de Araraquara, Lauand, André, Pedro, João, Antônio, Vitor, Mayra e Isabela que trilharam comigo o caminho até a faculdade e o mestrado. E mesmo cada um seguindo um rumo diferente ainda encontramos tempo para celebrar a amizade.

Aos meus amigos de São Carlos, B2, Nardão, Tahzib, Link, Frutily, Barretos, Anderson, Tio Clóvis, Coxinha, Jorge, Banana, Café, Letícia, Renata, Antônio, Amanda, Natália, Paula, Zé entre outros que estiveram comigo durante o período de graduação e uma parte do mestrado seja em horas de estudos, futebol, Gaperia, Interfísicas, Tuscas e outras festas.

“Tenho a impressão de ter sido uma criança brincando à beira-mar, divertindo-me em descobrir uma pedrinha mais lisa ou uma concha mais bonita que as outras, enquanto o imenso oceano da verdade continua misterioso diante de meus olhos”.

Isaac Newton

RESUMO

CAIRES, F. **Estudos e desenvolvimento de métodos baseados em harmônicos esféricos para análise de similaridade estrutural entre ligantes**. 2016. 130p. Dissertação (Mestrado em Ciências) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2016.

Descritores moleculares são essenciais em muitas aplicações de física e química computacional, como na análise de similaridade entre ligantes baseada em sua estrutura. Harmônicos esféricos têm sido utilizados como descritores da superfície molecular por serem uma forma compacta de descrição geométrica e por possuírem um descritor invariante por rotação. Assim, este trabalho propõe um método de análise de similaridade estrutural entre ligantes no qual se modela a superfície de uma molécula através de uma expansão em harmônicos esféricos realizada pelo programa LIRA. Os coeficientes encontrados são utilizados para percorrer o banco de dados DUD-E, com descritores previamente calculados, utilizando Distância Euclidiana e diversos valores de corte para selecionar compostos mais semelhantes. O potencial do método é avaliado usando o Ultrafast Shape Recognition (USR) como método padrão, pelo fato de ser uma excelente e rápida métrica para análise da similaridade de ligantes. Foram selecionadas 50 moléculas de diferentes tamanhos e composição de forma a representar todos os grupos moleculares presentes na DUD-E. Em seguida, cada molécula foi submetida à busca de similares variando-se valores de corte para o LIRA em que o conjunto de moléculas selecionadas foi comparado com as selecionadas pelo USR através de um processo de classificação binária e criação e interpretação de curvas ROC. Além do benchmarking, foi realizada a análise das componentes principais para determinar quais descritores são os mais importantes e carregam as melhores informações utilizadas na descrição da superfície da molécula. A partir das componentes principais, foi realizado um estudo do uso de funções peso, associando mais importância aos descritores adequados, e a redução da dimensionalidade do banco de dados, seleção de um novo conjunto de autovetores que formam as bases do espaço vetorial e uma nova descrição das moléculas para o novo espaço, no qual cada variação foi avaliada através de um novo benchmarking. O LIRA se mostrou tão rápido quanto o USR e apresentou grande potencial de seleção de moléculas similares, para a maioria das moléculas testadas, pois as curvas ROC apresentaram pontos acima da linha do aleatório. Tanto a redução da dimensionalidade quanto o uso de funções de ponderação agregaram valor à métrica deixando-a mais veloz, no caso da redução da quantidade de descritores, e seletiva, em ambos os casos. Dessa forma, o método proposto se mostrou eficiente em mensurar a similaridade

entre ligantes de forma seletiva e rápida utilizando somente informações a respeito da superfície molecular.

Palavras-chave: Similaridade. Harmônicos esféricos. Ligantes.

ABSTRACT

CAIRES, F. **Study and development of spherical harmonics based methods for similarity ligand analysis**. 2016. 130p. Dissertação (Mestrado em Ciências) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2016.

Molecular descriptors are essential for many applications in computational chemistry and physics, such as ligand-based similarity searching. Spherical harmonics have previously been suggested as comprehensive descriptors of molecular structure due to their properties, orthonormality and rotationally invariant. Here we proposed a ligand similarity analysis method where molecule's surface is modeled by an expansion in Spherical Harmonics, called LIRA, whose coefficient are used to perform a search in DUD-E database, with all descriptors previously calculated, measured by Euclidian Distance and different cutoff's values to select similar compounds. Method's potential is evaluated against Ultrafast Shape Recognition (USR), due to it is an excellent a fast metric to ligand similarity analysis, in a benchmarking. Fifty molecules are selected varying chemical composition and size to represent all molecular groups of DUD-E. After that, which one was submitted in a search with different values of cutoff for LIRA and the subset selected was compared with the ones selected by USR through binary classification and ROC curves analysis. Beyond benchmarking, it was performed a principal component analysis to identify which are the most valuable coefficient for shape description. Using principal components two other studies are made, weight functions are applied to descriptors, providing more value for those carry more information, and dimensionality reduction, where a subset of eigenvectors are select to form the new basis of the vector space and the new molecule's description was made in the new space, which variation was tested in a new benchmarking. Lira showed to be as fast as USR and a big potential to select similar molecules, for the majority of the molecules tested, because ROC curves had points over the random line. Dimensionality reduction and weight functions improved LIRA results raising velocity, due to the use of less descriptors to model molecule's surface, and the selection power, for both cases. In summary, the proposed method showed to be an efficient and fast tool for measure similarity between ligands based in molecular shape.

Key Words: Similarity. Spherical harmonics. Ligands.

LISTA DE FIGURAS

Figura 1 -	Esquema de desenvolvimento de um novo fármaco através do processo de desenho racional de fármacos.....	25
Figura 2 -	Métodos computacionais utilizados no Desenho Racional de Fármacos. Podem ser baseados na estrutura do ligante ou do sítio receptor. O conhecimento ou não da estrutura determina o tipo de abordagem a ser utilizada.....	26
Figura 3 -	Representação estrutural da interação proteína-ligante segundo modelo chave-fechadura. Ligante e sítio receptor apresentam estruturas complementares essenciais para o processo de interação. De acordo com o tipo de ligante pode-se ter ou não atividade biológica.....	28
Figura 4 -	Representação esquemática da interação proteína-ligante segundo Teoria do encaixe induzido. Nessa teoria o sítio ativo e o ligante têm suas conformações alteradas para que a ligação possa ocorrer.....	29
Figura 5 -	Campo de força utilizado por funções de pontuação em experimentos de Docagem Molecular.....	34
Figura 6 -	Esquema representando a construção de descritores moleculares, mostrando de onde são extraídas as informações, onde podem ser aplicados e por quem podem ser processados.....	42
Figura 7 -	Representação geométrica das bandas dos harmônicos esféricos. Para cada banda são encontrados 2l+1 componentes.....	46
Figura 8 -	Reconstrução da forma de um objeto usando harmônicos esféricos. Conforme aumenta-se a quantidade de bandas mais detalhes são atribuídos a superfície reconstruída deixando-a cada vez mais próxima da original.....	49
Figura 9 -	Erro da expansão em harmônicos esféricos. Estudo conduzido por Morris revela que no processo de descrição molecular baseada em harmônicos esféricos o valor de RMSD diminui conforme aumenta-se a quantidade de bandas porém somente até o valor $l_{\text{máx}}=14$, no qual o erro é mínimo e em seguida voltar a crescer.....	50
Figura 10 -	Representação explícita dos valores para a matriz de rotação calculada por Ritchie e Kemp.....	52
Figura 11 -	Coordenadas polares esféricas.....	58
Figura 12 -	Método PDBParser usado para extrair as informações atômicas de um arquivo PDB. Construído em Python de forma a extrair as informações das moléculas que compõe o banco de dados DUD-E.....	59
Figura 13 -	Representação da superfície de Van der Waals e das superfícies acessíveis ao solvente e não acessíveis determinadas pela esfera de prova.....	63
Figura 14 -	Descritores rotacionalmente invariantes calculados pelo Lira para $l_{\text{máx}} = 16$	64
Figura 15 -	Superfície reconstruída usando o método baseado em Harmônicos Esféricos e a biblioteca de visualização 3D Mayavi.....	64
Figura 16 -	Representação das distâncias utilizadas para descrever a superfície molecular a partir dos quatro momentos atômicos usados pelo USR.....	68
Figura 17 -	Representação dos momentos descritores de uma molécula descrita pelo USR. Os quatro momentos utilizados na descrição são: o centro da molécula, o átomo mais próximo do centro da molécula, o átomo mais longe do centro da molécula e o átomo mais longe do átomo mais longe do centro da molécula.....	70
Figura 18 -	Momento atômicos calculados para duas moléculas pelo método USR. No centro encontra-se a métrica própria utilizada pelo USR para classificar as moléculas similares.....	71
Figura 19 -	Comparação entre a quantidade de moléculas analisadas por segundo pelos métodos Shape Signature, USR e ROC, mostrando que o USR consegue varrer uma quantidade maior de arquivos no mesmo período de tempo quando comparado aos métodos comerciais.....	72

Figura 20 -	Comparação das moléculas selecionadas para a molécula query 35kith pelo Lira e USR usando como valores de corte 0.5 e 0.95, respectivamente. As moléculas A, B e C apresentam a basicamente a mesma forma tendo apenas a mudança de alguns átomos. A segunda e terceira linhas mostram a superfície atômicas de moléculas selecionadas pelo Lira e sua representação 3D. Em ambos os casos as moléculas 12kith, 14kith e 33kith foram selecionadas e revelam que o Lira aparenta ter mais flexibilidade conformacional.....	78
Figura 21 -	Diagrama utilizado no processo de classificação binária usado no benchmarking. Determinação de propriedades estatísticas como acurácia, especificidade e taxas de verdadeiro e falso positivo para análise de curvas ROC.....	81
Figura 22 -	Representação de curva ROC. A linha tracejada descrita pela equação $y=x$ representa um classificador aleatório. Pontos acima da reta apresentam status de bom classificador, pontos abaixo apresentam status de mau classificar e pontos sobre a reta são tidos como classificadores aleatório.....	83
Figura 23 -	Valores médios de calculado para cada componente dos descritores moleculares utilizados pela métrica L^2	85
Figura 24 -	Representação geométrica dos autovalores. U é autovetor de T pois $\exists \lambda \in \mathbf{R} / T(U) = \lambda u$. V não é autovetor de T pois $\nexists \lambda \in \mathbf{R} / T(v) = \lambda v$	88
Figura 25 -	Autovalores calculados para métrica L^2	88
Figura 26 -	Matriz gerada pelos autovetores selecionados na análise de componentes principais para criação do novo subespaço vetorial.	89
Figura 27 -	Dados obtidos para experimento de PCA métrica L^1 . Representação de todos os coeficientes usados para descrição das moléculas. Valores médios calculados para cada componente. Autovalores calculados durante análise. Análise de curvas de utilizada para determinação de uma função ponderação.....	94
Figura 28 -	Dados obtidos para experimento de PCA métrica L^2 . Representação de todos os coeficientes usados para descrição das moléculas. Valores médios calculados para cada componente. Autovalores calculados durante análise. Análise de curvas de utilizada para determinação de uma função de ponderação.....	95
Figura 29 -	Comparação entre moléculas selecionadas pela métrica L^2 utilizando a distância euclidiana original, ponderada por um polinômio de quinta ordem e pela função gerada pelo software Eureqa.....	96
Figura 30 -	Curvas ROC calculadas para a métrica L^2 utilizando como função de ponderação polinômio de quinta ordem. Análise revela que uso de função melhora status de classificação para programa Lira.....	97
Figura 31 -	Análise de curva ROC gerados para espaço vetorial original e novo espaço vetorial fruto do processo de redução de dimensionalidade gerado pelo experimento de PCA. Para molécula 305urok não há alterações entre gráficos originais e novos evidenciando manutenção de características principais.....	98
Figura 32 -	Análise de curva ROC gerados para espaço vetorial original e novo espaço vetorial fruto do processo de redução de dimensionalidade gerado pelo experimento de PCA. Para molécula 35kith não há alterações entre gráficos originais e novos evidenciando manutenção de características principais.....	99
Figura 33 -	Análise de curva ROC gerados para espaço vetorial original e novo espaço vetorial fruto do processo de redução de dimensionalidade gerado pelo experimento de PCA. Para molécula 313aces novo espaço vetorial mantém informações essenciais para análise de similaridade mostrando ligeira melhora em relação à experimento inicial.....	100
Figura 34 -	Análise de curva ROC gerados para espaço vetorial original e novo espaço vetorial fruto do processo de redução de dimensionalidade gerado pelo experimento de PCA. Para molécula 74bace1 apesar de ambos os gráficos permanecerem abaixo da curva pode se notar melhora na descrição com 90 componentes pois há aproxima dos pontos a curva.....	101
Figura 35 -	Análise de curva ROC gerados para espaço vetorial original e novo espaço vetorial fruto do processo de redução de dimensionalidade gerado pelo experimento de PCA. Para molécula 35kith não há alterações entre gráficos originais e novos evidenciando manutenção de características principais para a descrição com 6 componentes.....	102

Figura 36 -	Análise de curva ROC gerados para espaço vetorial original e novo espaço vetorial fruto do processo de redução de dimensionalidade gerado pelo experimento de PCA. Para molécula 74bace1 a diminuição de descritores favorece a seleção de mais molécula que alteram o formato da curva ROC.....	103
Figura 37 -	Análise de curva ROC gerados para espaço vetorial original e novo espaço vetorial fruto do processo de redução de dimensionalidade gerado pelo experimento de PCA. Para molécula 305urok não há alterações entre gráficos originais e novos evidenciando manutenção de características principais.....	104
Figura 38 -	Análise de curva ROC gerados para espaço vetorial original e novo espaço vetorial fruto do processo de redução de dimensionalidade gerado pelo experimento de PCA. Para molécula 313aces há bom aumento na quantidade de moléculas selecionadas pela descrição com 6 componentes originando aumento dos pontos de curva e na status de bom classificador da métrica.....	105
Figura 39 -	Comparação entre as classificações para cada métrica e variação utilizada no benchmarking.....	108
Figura 40 -	Comparação entre os tempos de processamento para cada métrica e variação utilizada no benchmarking.....	109
Figura 41 -	Página inicial do Web site Lira no qual uma molécula é informada ao site em busca das moléculas mais similares.....	109
Figura 42 -	Visualização 2D e 3D das moléculas selecionadas pelo Lira e pelo USR. Para a estrutura 3D utiliza-se a ferramenta JsMol.....	111
Figura 43 -	Cálculo dos polinômios associados de Legendre.....	124

LISTA DE TABELAS

Tabela 1 - Representação dos tamanhos dos bancos de dados gerados para cada métrica. O tamanho é dependente da quantidade de componentes sendo o banco da métrica L^1 o maior.....	74
Tabela 2 - Intervalos utilizados para classificar as moléculas de acordo com a quantidade de átomos.....	75
Tabela 3 - Combinações de 3 átomos utilizados para a seleção das moléculas a serem utilizados no benchmarking.....	76
Tabela 4 - Combinações de 2 átomos utilizados para a seleção das moléculas a serem utilizados no benchmarking	76
Tabela 5 - 50 moléculas selecionadas para análise de benchmarking tentando representar a diversidade do banco de dados DUD-E.....	76
Tabela 6 - Tempo gasto para leitura, cálculo, varredura do banco de dados e tempo total gasto para cada métrica utilizada nos experimentos de benchmarking.....	79
Tabela 7 - Relações, fórmulas e descrição utilizadas no processo de classificação binária e geração de curvas ROC.....	82
Tabela 8 - Medidas estatísticas utilizadas durante a análise de componente principal.....	86
Tabela 9 - Funções utilizadas na tentativa de descrever os autovalores encontrados a fim de determinar uma função de ponderação a ser usada para métrica Lira.....	91
Tabela 10 - Funções obtidas pelo software Eureqa e as definições utilizadas em sua definição.....	92
Tabela 11- Porcentagem de classificação para cada variação da métrica L^2	106
Tabela 12- Tempo de processamento para as métricas L^2	107
Tabela 13 - Porcentagem de classificação para cada variação da métrica L^1	107
Tabela 14 - Tempo de processamento para as métrica L^1	108
Tabela 15 - Classificação das moléculas métrica L^2	127
Tabela 16 - Classificação das moléculas métrica L^1	129

LISTA DE ABREVIATURAS E SIGLAS

2D	Bidimensional
3D	Tridimensional
6D	Seis dimensões
90D	Noventa dimensões
ADMET	Absorção, Distribuição, Metabolismo, Excreção e Toxicidade
CADD	Computer aided drug design
CoMFA	Comparative Molecular Field Analysis
DM	Medida de dissimilaridade
DUD-E	Database of Useful Decoys: Enhanced
IUPAC	União Internacional de Química Pura e Aplicada
MLR	Análise de Regressão Linear Multivariáveis
NUMPY	Numeric Python
PCA	Análise das Componentes Principais
PDB	Protein Data Bank
PLS	Análise de Mínimos Quadrados Parciais
QSAR	Quantitative structure–activity relationship
RDF	Funções de Distribuição Radial
RMN	Ressonância Magnética Nuclear
RMSD	Root Mean Square Deviation
ROC	Receiver Operating Characteristic
SM	Medida de similaridade
USR	Ultrafast Shape Recognition

SUMÁRIO

1 INTRODUÇÃO.....	25
1.1 Desenho Racional de Fármacos.....	25
1.2 Reconhecimento Molecular.....	26
1.2.1 Fármacos estruturalmente inespecíficos.....	27
1.2.2 Fármacos estruturalmente específicos.....	27
1.2.3 Encaixe induzido.....	28
1.3 Desenho de drogas baseado na estrutura do receptor (Structure Based Drug Design).....	29
1.3.1 Docagem Molecular Proteína Ligante.....	29
1.3.2 Criação do Ensemble de conformações.....	30
1.3.3 Busca sistemática.....	31
1.3.4 Métodos Estocásticos ou Aleatórios.....	32
1.3.5 Simulações.....	32
1.4 Funções de Pontuação.....	33
1.4.1 Baseadas em Campos de Força.....	33
1.4.2 Funções Empíricas de Pontuação.....	34
1.4.3 Baseadas no conhecimento (<i>knowledge-based</i>).....	34
1.5 Desenho de Fármaco baseado na estrutura do ligante (Ligand-Based Method).....	35
1.5.1 Modelo Farmacofórico.....	35
1.6 QSAR - Conceitos básicos.....	36
1.6.1 Desenvolvimento de um modelo QSAR.....	37
1.6.2 Validação.....	38
1.6.4 QSAR 3D.....	39
1.6.5 CoMFA.....	39
1.7 Similaridade molecular.....	40
1.8 Similaridade estrutural tridimensional – Conceitos Básicos.....	42
1.8.1 Algoritmo Genético.....	42
1.8.2 Algoritmos para similaridade de estruturas tridimensionais.....	43
1.8.3 Similaridade baseada na superfície da molécula.....	43
1.8.4 Curvas de Correlação (Correlation Plots).....	44
1.8.5 Similaridade Morfológica (Morphological similarity).....	44
1.8.6 Harmônicos Esféricos.....	44

1.9 Harmônicos Esféricos	45
1.9.1 Equação de Laplace em coordenadas esféricas.....	46
1.9.2 Dependência do ângulo azimutal	47
1.9.3 Dependência do ângulo polar.....	47
1.9.4 Teorema da Expansão em Harmônicos Esféricos (Séries de Laplace)	48
1.9.5 Rotação dos Harmônicos Esféricos.....	50
1.9.6 Harmônicos esféricos rotacionalmente invariantes.....	52
2 OBJETIVOS	55
3 MATERIAL E MÉTODOS	57
3.1 Material	57
3.2 Métodos.....	57
3.2.1 Implementação Lira	57
3.2.1.1 PDBParser.....	58
3.2.1.2 Integração de Monte Carlo.....	59
3.2.1.3 Cálculo dos Harmônicos Esféricos	61
3.2.1.4 Expansão em Harmônicos Esféricos.....	62
3.2.2 Medidas de Similaridade e Dissimilaridade.....	65
3.2.2.1 Definição de dissimilaridade.....	65
3.2.2.2 Definição de Similaridade.....	65
3.2.2.3 Métrica Lp Ponderada.....	66
3.2.2.4 Distância Euclidiana	66
3.2.2.5 Distância de Manhattan.....	67
3.2.3 Implementação Ultrafast Shape Recognition (USR)	67
3.2.3.1 Descritores moleculares usados pelo USR.....	68
3.2.3.2 Medida de Similaridade	70
3.2.3.4 Tempo de execução.....	71
3.2.4 Benchmarking	72
3.2.4.1 Montagem do Banco de Dados	73
3.2.4.2 Determinação das moléculas do benchmarking.....	74
3.2.4.3 Determinação dos ligantes similares.....	77
3.2.4.4 Validação das moléculas selecionadas pelas métricas	79
3.2.5 Classificação Binária.....	80
3.2.6 Curvas ROC	81

3.2.7.1 Análise de Componente Principal (PCA).....	83
3.2.7.2 Autovalores e autovetores	87
3.2.7.3 Redução de Dimensionalidade	88
3.2.7.4 Criação do novo subespaço	89
3.2.7.5 Projeção dos dados no novo subespaço	89
3.2.7.6 Funções de Ponderação	90
4 RESULTADOS	93
4.1 Novos Benchmarkings.....	97
4.2 Análise estatística do benchmarking	105
4.3 Website Lira	109
4.3.1 Método de busca.....	109
4.3.2 Display das moléculas	110
4.3.3 Melhorias futuras	111
5 CONSIDERAÇÕES FINAIS	113
5.1 Melhorias Futuras	114
REFERÊNCIAS	117
APÊNDICE A – Cálculo estável dos harmônicos esféricos.....	123
APÊNDICE B – Classificação das moléculas utilizadas no benchmarking.....	127

1 INTRODUÇÃO

1.1 Desenho Racional de Fármacos

Drug design ou desenho racional de fármacos (Figura 1) trata-se do processo de criação de novos fármacos baseando-se no conhecimento de um determinado alvo biológico. Define-se como droga todo e qualquer composto, geralmente pequenas moléculas, que em contato com o alvo biológico é capaz de gerar efeitos terapêuticos promovendo benefícios para a saúde do paciente.¹ Com a premissa de que um fármaco deve apresentar complementaridade geométrica e propriedades físico-químicas adequadas para ligar-se a seu alvo molecular, o desenho de drogas tornou-se uma importante ferramenta na predição de afinidade, seletividade e estabilidade usados no processo de criação e/ou melhoramento de fármacos.²⁻³

Figura 1- Esquema de desenvolvimento de um novo fármaco através do processo de desenho racional de fármacos.

Fonte: Adaptada de SCHEMATIC...⁴

Proteínas e ácidos nucleicos são biomoléculas comumente exploradas como alvos biológicos. Moléculas que desempenham um papel central em determinada função, processo de sinalização ou via metabólica associado a uma doença específica ou disfunção tornam-se potenciais alvos de ação de fármacos. Pequenas moléculas podem ser criadas para inibir ou potencializar a via metabólica de interesse, por exemplo, atuando como inibidores competitivos, agonistas, antagonistas, moduladores, ativadores ou inibidores de enzimas e/ou canais iônicos. Uma das estratégias para desenvolvimento de um novo fármaco é baseada no

em informações da estrutura do alvo molecular (Structure-Based Drug Design)⁵⁻⁶ ou do ligante (Ligand-Based Drug Design)^{5, 7} que interage com o alvo de interesse, as duas alternativas normalmente se classificam como pertencentes ao grupo de Desenho de Drogas auxiliado por Computadores (CADD - Computer aided drug design) conforme mostrado na Figura 2.

Desenho de drogas auxiliado por computadores		
	Ligantes Conhecidos	Ligantes Desconhecidos
Estrutura proteica conhecida	Desenho de drogas baseado na estrutura proteica Docagem Molecular proteína-ligante Modelagem por Homologia	<i>De novo Design</i>
Estrutura proteica desconhecida	Desenho de drogas baseadas no ligante 1 ou mais ligantes •Busca por similaridade Vários ligantes •Busca farmacofórica Muitos ligantes(20+) QSAR	Necessita de informações experimentais Podem ser aplicados filtros ADMET
ADMET: absorção, distribuição, metabolismo, excreção, toxicidade		

Figura 2- Métodos computacionais utilizados no Desenho Racional de Fármacos. Podem ser baseados na estrutura do ligante ou do sítio receptor. O conhecimento ou não da estrutura determina o tipo de abordagem a ser utilizada.

Fonte: Adaptada de MARTEL⁸

1.2 Reconhecimento Molecular

O processo de reconhecimento molecular é essencial para toda e qualquer atividade biológica⁹⁻¹⁰, dessa forma o entendimento de como um ligante interage com seu alvo é fundamental para a criação de drogas cada vez mais específicas e eficientes. O reconhecimento entre moléculas é mediado pela complementaridade geométrica entre alvo e receptor e por interações intermoleculares, as quais definem o grau de afinidade e a especificidade da ligação. O complexo alvo-ligante é estabilizado por um a combinação de forças eletrostáticas, dispersão

hidrofóbica, ligações de hidrogênio e ligações covalentes.¹¹⁻¹⁵ O modo de interação entre fármaco e receptor permite a classificação dos fármacos em dois grupos genéricos: fármacos estruturalmente inespecíficos e fármacos estruturalmente específicos.¹³

1.2.1 Fármacos estruturalmente inespecíficos

São fármacos que dependem única e exclusivamente das propriedades físico-químicas (coeficiente de partição, pKa) para produzir efeito biológico.¹³ O principal exemplo de fármaco dessa classe são os anestésicos, cujo mecanismo de ação é a depressão inespecífica de biomembranas lipoproteicas, elevando o limiar de excitabilidade celular ou a interação inespecífica com sítios hidrofóbicos de proteínas do sistema nervoso central, provocando a perda de consciência. O reconhecimento molecular nesse caso é determinado principalmente por interações de Van der Waals e sua potência está extremamente ligada a lipossolubilidade do fármaco. Como exemplo pode-se citar a morfina e a heroína, ambos derivados do ópio, agem no sistema nervoso central provocando sensação de euforia, anestesia e depressão. A heroína é sintetizada a partir da morfina, no qual é realizada uma reação de esterificação dando origem a diacetilmorfina, a adição dos grupos acetila aumenta a lipossolubilidade da molécula o que favorece sua passagem na barreira hematoencefálica potencializando o efeito da droga.

1.2.2 Fármacos estruturalmente específicos

São fármacos que dependem das propriedades físico-químicas, a orientação espacial dos grupamentos funcionais da molécula e das propriedades da superfície do ligante que devem ser complementares ao sítio de ligação localizado no receptor. O processo de reconhecimento molecular pode ser descrito de duas formas: Chave-Fechadura (Figura 3) ou Encaixe Induzido (Figura 4).

No modelo chave-fechadura, assume-se um paralelo a relação entre a chave e a fechadura encontradas no cotidiano ao o sítio de ligação e diversos ligantes, respectivamente. A interação entre a chave e fechadura permite que uma ação seja realizada pelo sistema da fechadura a qual não é acessível na ausência da chave. Esse modelo pressupõe que ambas as partes envolvidas possuem uma forma rígida e pouco mutável. A resposta biológica dessa interação normalmente consiste em alteração na estrutura do alvo molecular que pode promover ou inibir alguma atividade associada à biomolécula, o equivalente a abertura ou não da fechadura. O conceito de que as interações entre enzima e seu substrato seriam complementares

foram introduzidas inicialmente por Emil Fischer em 1894.¹² O mesmo conceito foi aplicado para as reações antígeno-anticorpo e posteriormente para a interação entre fármacos e receptores. Em nível molecular, a complementaridade entre ligante e receptor é mediada por interações hidrofóbicas, ligações de hidrogênio, interações eletrostáticas entre outras, capazes de diminuir a energia livre do complexo fármaco-receptor e estabilizar a ligação.

Figura 3- Representação estrutural da interação proteína-ligante segundo modelo chave-fechadura. Ligante e sítio receptor apresentam estruturas complementares essenciais para o processo de interação. De acordo com o tipo de ligante pode-se ter ou não atividade biológica.

Fonte: FRAGA¹³

1.2.3 Encaixe induzido

Teoria descrita por Koshland sugere que para complexar com seu alvo, um substrato induz alterações conformacionais na subunidade com a qual interage no sítio receptor, essa mudança pode ser transmitida às subunidades vizinhas induzindo no receptor a conformação responsável pela atividade biológica.¹² Esse processo também pode ser aplicado em interações de inibidores com enzimas ou agonistas e antagonistas como receptores. A teoria do encaixe induzido considera que o ligante apresenta capacidade de selecionar e induzir modificações estruturais tridimensionais em seu receptor biológico. Além disso, sugere que o receptor apresente uma estrutura flexível que pode ser alterada para reconhecer mais de uma conformação do ligante. Ao induzir a alteração estrutural do sítio ativo do receptor o ligante está criando a melhor conformação para seu reconhecimento. As mudanças conformacionais observadas decorrentes da interação do fármaco com seu alvo biológico podem apresentar diferentes graus de amplitude e complexidade, podem envolver a reorientação da cadeia lateral de um aminoácido de forma a maximizar as interações com o ligante, reduzindo a energia livre do complexo e estabilizando-o.

Figura 4- Representação esquemática da interação proteína-ligante segundo Teoria do encaixe induzido. Nessa teoria o sítio ativo e o ligante têm suas conformações alteradas para que a ligação possa ocorrer.

Fonte: VERLI; BARREIRO¹²

1.3 Desenho de drogas baseado na estrutura do receptor (Structure Based Drug Design)

Desenho de drogas baseado na estrutura do alvo molecular é uma abordagem dependente do conhecimento da estrutura tridimensional do alvo, a qual pode ser obtida principalmente através de experimentos de Cristalografia de raios-X ou Ressonância Magnética Nuclear (RMN).⁵ Em casos que estrutura experimental não exista ou apresente problemas, pode-se criar um modelo homólogo baseando em estruturas experimentais relacionadas a proteína. A partir do conhecimento do sítio de ativo do alvo, os fármacos são desenhados otimizar a interação com ele, aliando complementaridade geométrica e carga, propiciando grande afinidade e seletividade. Atualmente são utilizadas três abordagens para o desenvolvimento de novos fármacos baseados na estrutura do alvo biológico: modelagem por homologia, docagem molecular proteína-ligante e *de novo* design.

1.3.1 Docagem Molecular Proteína Ligante

Processo de docagem molecular é dependente de duas informações: Conformação e Orientação (posição) das moléculas envolvidas.¹⁶ Tem como meta descrever com grande precisão o reconhecimento entre alvo e ligante, a afinidade com que ocorre interação e predizer corretamente a atividade fornecida pela ligação.

Devido ao alto grau de complexidade, proveniente da combinação de efeitos entálpicos e entrópicos que controlam a ligação, o processo é dividido em várias etapas no qual em cada

passo é adicionado um ou mais graus de complexidade.¹⁷ Inicia-se pela colocada da molécula no sítio de ativo, em seguida várias conformações de moléculas são geradas a partir de variações nos graus de liberdade da molécula através de modificações nos ângulos entre as ligações atômicas. Todo o ensemble de conformações deve ser gerado rapidamente e classificado de acordo com as funções de pontuação (Scoring Functions), criadas para predizer a atividade biológica entre o ligante e o alvo, selecionando as melhores.

Funções de pontuações podem ser utilizadas em todo o processo de desenho de fármacos desde a análise de complementaridade geométrica e eletrostática entre o ligante e o sítio de ligação, a análise do potencial de Van der Waals, solvatação das moléculas ou efeitos entrópicos. Deve-se ressaltar que tanto efeitos entálpicos como entrópicos podem dominar o processo de reconhecimento molecular o que dificulta a definição de uma função de pontuação adequada, grande parte delas está focada em capturar os efeitos energéticos do que os entrópicos. Outros fatores limitantes incluem a baixa resolução cristalográfica dos alvos, mudanças na estrutura do sítio molecular devido ao processo de ligação (ajuste induzido) e participação de moléculas de água nas interações proteínas-ligantes.¹⁷

Para determinar qual algoritmo e função de pontuação utilizar em um experimento de docagem molecular é importante levar em consideração as formas de representação da proteína e do ligante definidas por estes. Há três tipos básicos de representação do receptor: atômica, superfície e grid.^{5, 13} A abordagem atômica é utilizada em conjunto com funções de potenciais de energia e frequentemente na parte final do processo de ranking devido à alta complexidade e custo computacional nos cálculos das interações dos pares atômicos.¹³⁻¹⁴ Representações de superfície, que são comumente utilizadas em docagens de proteína-proteína, tem como característica principal promover o alinhamento dos pontos das superfícies das proteínas através da minimização dos ângulos das moléculas opostas. Grid com potenciais de energias foram criados para representar o cálculo de energia, na qual a ideia básica é armazenar os potenciais eletrostáticos e de Van der Waals nos pontos do grid informações a respeito das contribuições energéticas do receptor.

1.3.2 Criação do Ensemble de conformações.

Um ensemble de conformações robusto é essencial para um bom desempenho e resultado em um experimento de docagem molecular, para isso são usadas três abordagens distintas nos algoritmos responsáveis por criar a flexibilidade do ligante e da proteína, divididos

em três grandes categorias básicas : Métodos Sistemáticos (Construção Incremental, Busca Conformacional e Banco de Dados), Métodos Aleatórios ou Estocásticos (Monte Carlo e Algoritmo Genético) e Simulações (Dinâmica molecular e Minimização de Energia).

1.3.3 Busca sistemática

Essa classe de algoritmos talvez seja a de mais fácil entendimento no qual a maioria das mudanças conformacionais são realizadas a partir da rotação de uma única ligação considerando uma pequena porção de ângulos tentando explorar todos os graus de liberdade da molécula.¹⁸ Entretanto, o número de conformações possíveis pode ser muito grande dependendo da quantidade de ligações que passíveis de serem rotacionadas gerando o chamado problema de explosão conformacional.¹⁶ Para reduzir a quantidade de graus de liberdade utiliza-se que os comprimentos e ângulos da ligação são fixos em conjunto com a seleção de uma série de valores de ângulos torcionais demonstrando as conformações permitidas.

$$N_{conformações} = \prod_{i=1}^N \prod_{j=1}^{n_{inc}} \frac{360}{\theta_{i,j}} \quad (1)$$

No qual N é o número de ligações rotacionáveis e $\theta_{i,j}$ é o passo incremental com a rotação j para a ligação i . Podem ser implementados de várias maneiras: acoplamento de fragmentos moleculares dentro do sítio ativo e conectando-os por meio de ligações covalentes, abordagem conhecida como *de novo*, ou fracionando-se os ligantes em uma parte rígida (núcleo do ligantes) e partes flexíveis (cadeias laterais). A parte rígida do ligante é inserida dentro do sítio ativo da proteína e as cadeias flexíveis são adicionadas de forma incremental. Algumas conformações podem ser descartadas caso violem critérios energéticos e/ou geométricos para aumentar a eficiência. O controle das conformações possíveis pode ser demonstrado através de uma árvore de conformações, nos quais os nodos representam uma variação incremental formando uma nova estrutura. A prévia verificação das estruturas pré-estabelecidas pode melhorar o desempenho do algoritmo, pois caso um problema seja detectado toda a árvore abaixo dela pode ser descartada diminuindo o ensemble de conformações e aumentando a performance do método.¹⁸

O algoritmo *DOCK₄₀*¹⁶ posiciona a parte rígida do ligante no sítio ativo de acordo com a complementaridade estérica e em seguida adiciona as partes flexíveis, ligação por ligação, vasculhando as possibilidades possíveis, previamente selecionadas por um algoritmo que elimina as conformações mais desfavoráveis simplificando a complexidade da busca. Por sua

vez o *FLEX*¹⁶ posiciona o corpo rígido no interior do sítio baseando-se na interação geométrica. As interações são avaliadas seguindo a ordem: aceptoras e doadoras de ligação de hidrogênio e posteriormente os grupos hidrofóbicos. O *Heammer Head*¹⁶ realiza o docking de cada fragmento no sítio ativo executando um processo de minimização de energia a cada passo incremental e por fim reconstrói o ligante utilizando aqueles que apresentam melhor pontuação.

1.3.4 Métodos Estocásticos ou Aleatórios

Realizam alterações translacionais e/ou rotacionais aleatórias em um único ligante ou em toda a população, os novos ligantes são testados e classificados de acordo com as funções pré-definidas.^{6, 16-17, 19} Os algoritmos mais utilizados nessa abordagem são Monte Carlo e Algoritmo Genético. O método de Monte Carlo gera uma conformação inicial de um ligante dentro do sítio ativo consistido de alterações conformacionais, rotacionais e translacionais aleatórias e determina-se a pontuação da estrutura. Em seguida, uma nova estrutura é gerada e tem sua pontuação calculada. Para verificar se a nova conformação será mantida, aplica o critério de Metrópolis. Caso a nova estrutura apresente pontuação melhor, seja um mínimo global de energia, do que a anterior, ela é imediatamente aceita. Caso não seja um mínimo de energia, uma função de probabilidade de Boltzmann é aplicada no novo composto como um teste e se a molécula passar pelo teste ela é aceita e caso contrário é descartada. O algoritmo genético utiliza conceitos da teoria da evolução e seleção natural de Darwin para selecionar os compostos com energia de um mínimo global, podendo ser aplicado em diversos tipos de problemas apesar de apresentar alto custo computacional.

1.3.5 Simulações

Dinâmica molecular é a ferramenta mais comum para esse tipo de abordagem, costuma não conseguir ultrapassar barreiras energéticas muito elevadas em períodos de tempos viáveis. Além disso, aloca os ligantes somente em pontos de mínimos de energia na superfície da proteína.¹³ É comum realizar simulações em diferentes partes do sistema proteína-ligante com diferentes valores de temperatura. Outra estratégia para abordar o problema de mínimos locais é começar os cálculos da dinâmica molecular utilizando diferentes posições para os ligantes. Ao contrário do que acontece com a dinâmica molecular, os processos de minimização de energia nunca são utilizados como única técnica busca e sim associado a outra abordagem, como Monte Carlo.

1.4 Funções de Pontuação

As funções de pontuação e ordenação das conformações dos ligantes tem papel crucial no processo de virtual screening baseado na estrutura, o uso de uma função não apropriada, incapaz de diferenciar as conformações corretas das incorretas pode acarretar no fracasso de todo um experimento, dessa forma desenvolver funções de pontuação confiáveis e robustas é de extrema importância.¹⁶

Técnicas para simulação de energia livre foram criadas para quantizar as interações no sistema proteína-ligante. Entretanto, devido à alta complexidade dos cálculos e elevado custo computacional, o método ainda é útil para ser usados em complexos grandes, pois o custo computacional cresce na proporção em que aumenta o tamanho do complexo. Dessa forma o grande desafio no processo de criação de funções de pontuação é conciliar acurácia e velocidade. Funções de pontuação criadas para experimentos de docagem molecular apresentam associações e simplificações, como desconsiderar o fenômeno de reconhecimento molecular dirigido principalmente por efeitos entrópicos, usados para diminuir a complexidade e o custo computacional na análise dos modelos. São divididas em três classes: Baseadas em Campos de Força (*Force Field based*), empíricas e baseadas no conhecimento (*knowledge-based*).

1.4.1 Baseadas em Campos de Força

Quantifica a soma de duas energias, a energia proveniente da interação do receptor com o ligante e a energia interna do ligante.¹³ A maioria das funções dessa classe consideram apenas uma única conformação da proteína, o que facilita os cálculos, pois não é necessário calcular a energia interna da proteína. Cada função pode ser baseada em diferentes parâmetros do campo de força, criando dessa forma uma coleção de funções para essa classe. Entretanto, algumas propriedades são sempre analisadas da mesma forma, tanto interações entre ligante e receptor, como a energia interna do ligante, são descritas através de termos de energias Van der Waals e eletrostática. O termo de Van der Waals é descrito a partir de um potencial de Lennard-Jones, termos altos como 12-6 resultam em um potencial repulsivo, termos mais baixos como 8-4 resultam em um potencial mais suave. O termo eletrostático é descrito por um potencial de Coulomb, dependente da constante dielétrica do meio, as cargas dos átomos e inversamente proporcional ao quadrado da distância entre eles.^{13, 20}

As funções de campo de força padrão apresentam uma série de limitações, foram desenvolvida para considerar contribuições entálpicas e energéticas e não considerar termos de

solvatação e entrópicos. Todavia algumas funções apresentam termos relacionados às ligações de hidrogênio, ângulo de curvatura e ângulo diedro. As ligações de hidrogênio podem ser descritas de duas formas distintas, dependentes da natureza e geometria da ligação ou representada em um potencial de Lennard-Jones. (Figura 5)

Figura 5- Campo de força utilizado por funções de pontuação em experimentos de Docagem Molecular.
Fonte: GUIDO²⁰

1.4.2 Funções Empíricas de Pontuação

Essa classe de funções utiliza dados experimentais para criar um modelo no qual a soma de funções diversas derivadas das contribuições energéticas das ligações e das conformações consigam reproduzir com grande eficiência os dados experimentais. Baseia-se na ideia de que as energias de ligação podem ser aproximadas por uma soma de termos individuais não relacionados entre si, os coeficientes das várias contribuições são obtidos através de análises de regressão linear usando informações de energias de ligação e potenciais extraída de uma estrutura determinada por Raios-X.¹³ Consideram em seus cálculos ligações de hidrogênio, interações polares e apolares, efeitos de solvatação e entrópicos, porém são mais simples que as funções de campo, tornando-as mais rápidas. Além da determinação dos coeficientes, é comum o uso de pesos para cada um deles, dificultando que termos obtidos em outras análises sejam facilmente combinados para gerar uma nova função. A desvantagem dessa metodologia é a dependência de dados experimentais confiáveis para serem usados no processo de regressão.

1.4.3 Baseadas no conhecimento (*knowledge-based*)

Metodologia desenvolvida para reproduzir a estrutura experimental em vez da energia de ligação, onde complexos proteína-ligantes são descritos a partir do potencial de interação de um simples par atômico.¹³ A quantidade de pares atômicos utilizados é definida a partir do tipo de ambiente, dessa forma são capazes de descrever efeitos de ligação difíceis de se modelar. Os

potencias são construídos considerando a frequência com que dois átomos diferentes são encontrados dentro do conjunto de dados estrutural. Os diferentes tipos de interações encontrados são classificados e ponderados de acordo com a sua frequência. A pontuação final é dada pela soma de todas as contribuições. São funções extremamente simples, permitindo uma eficiente e rápida análise de grandes bancos de dados. A desvantagem desse método é que a informação implícita está limitada ao conjunto de dados dos complexos estruturais proteína-ligante.

1.5 Desenho de Fármaco baseado na estrutura do ligante (Ligand-Based Method)

Processo de desenho de fármaco e otimização de compostos, amplamente utilizado em Drug Discovery, baseado no conhecimento da estrutura 3D das moléculas que se ligam a determinado receptor de interesse.¹³ A partir das estruturas é possível determinar quais são as principais características que um fármaco deve apresentar para que ocorra a interação com o receptor, formando o complexo receptor-ligante. Experimentos quantitativos de relações entre estrutura e atividade (QSAR 3D) e modelos farmacofóricos são os mais utilizados nesse tipo de metodologia fornecendo informações essenciais em relação a natureza das ligações entre o ligante e receptor e bons modelos de predição para otimização de compostos.

1.5.1 Modelo Farmacofórico

O conceito de farmacóforo foi introduzido em 1909 por Ehrlich, definindo como estrutura molecular que contenha todas as características essenciais (phoros) responsáveis pela atividade biológica (pharmacon) de uma droga. Após mais de cem anos uma nova definição, mantendo o mesmo conceito, foi divulgada pela União Internacional de Química Pura e Aplicada (IUPAC)²¹ no qual modelo farmacofórico é um ensemble de propriedades estéricas e eletrônicas necessárias para assegurar as interações supramoleculares ideais com um alvo biológico específico para acionar sua resposta biológica.

Modelos farmacofóricos podem ser criados usando-se tanto conhecimentos baseados na estrutura do ligante, por superposição de fragmentos de moléculas essenciais para a bioatividade como conhecimento baseados na estrutura de sítio de ligação considerando os possíveis pontos de interação entre o ligante e o alvo. São amplamente usados no processo de desenho racional de fármacos em experimentos de virtual screening, *de novo* e otimização de compostos. Devido aos avanços nas áreas de física, química e computação o número de

ferramentas, processos e softwares desenvolvidos que utilizam essa metodologia tem aumento constantemente, evidenciando seu grande potencial.

1.6 QSAR - Conceitos básicos

Método computacional criado para quantificar a correlação entre uma série de compostos químicos de um conjunto de dados e o processo biológico ou químico envolvido através de modelos matemáticos, baseando-se na hipótese de que similaridade estrutural ou físico-química proporciona uma atividade similar.²²

O primeiro passo para a criação de um modelo de QSAR é selecionar e criar um conjunto de dados de moléculas que tenham a atividade biológica desejada, uma vez escolhidas elas são modeladas *in silico* e minimizadas por meios de mecânica molecular ou mecânica quântica. Em seguida são gerados descritores moleculares para o set de moléculas para descrever as características estruturais e físico-químicas necessárias para que haja atividade biológica, a principal meta é criar "fingerprint" molecular, digital única, para cada molécula relacionado à atividade biológica. A partir dos descritores moleculares é desenvolvido um modelo matemático capaz de explicar a variação da atividade biológica para o conjunto de moléculas.¹⁵⁻

¹⁶ No passo final, o modelo obtido é submetido a vários processos de validação para testar seu valor estatístico e poder de predição.

Com o modelo QSAR desenvolvido, e devidamente validado, é realizado o processo de otimização dos compostos ativos, com finalidade de maximizar a atividade biológica. Em seguida esses compostos são testados experimentalmente em busca da atividade desejada. O modelo pode ser usado como ferramenta guia para identificar modificações que aumentam a atividade biológica.

A metodologia geral do processo de QSAR pode ser descrita da com os passos abaixo:

- Identificar e definir uma ou grupo de moléculas que apresentam a atividade biológica desejada, geralmente os compostos apresentam uma correlação entre si, porém grupos envolvendo diferentes tipos de molécula é uma abordagem relevante.
- Identificar e determinar descritores moleculares associados com variações estruturais e propriedades físico-químicas para as moléculas de estudo.
- Encontrar e quantificar a relação estabelecida entre os descritores moleculares e a atividade biológica, capaz de explicar as variações encontradas no conjunto de dados. Esse processo é a criação do modelo QSAR, responsável por otimizar os compostos.

- Testar experimentalmente os compostos com a finalidade de validar a estabilidade e o poder de predição do modelo criado.

1.6.1 Desenvolvimento de um modelo QSAR

Uma escolha adequada dos descritores moleculares com habilidade de gerar modelos matemáticos apropriados para descrever as relações entre os descritores e a atividade biológica de interesse, é essencial para o desenvolvimento de um modelo QSAR eficiente e robusto.¹⁶ Atualmente existem diversos softwares capazes de criar um grande número de descritores moleculares que podem ser usados experimentalmente, ainda assim o ponto crucial é escolha do melhor conjunto de descritores que expliquem os dados de atividade. Comumente nos métodos de QSAR são usadas três abordagens no processo de determinação de propriedades moleculares importantes para atividade biológica: (i) Análise de Regressão Linear Multivariáveis (MLR) (ii) Análise das Componentes Principais (PCA) (iii) Análise de Mínimos Quadrados Parciais (PLS). Também são usadas outras abordagens como Algoritmo Genético e Métodos Bayesianos

Análise de regressão linear multivariável é o método mais simples para quantificar os descritores moleculares e apresenta uma boa correlação entre os descritores e as variações da atividade, pode ser feito adicionando-se ou retirando-se descritores de forma a criar o modelo ideal. Para conjunto de dados muito grandes esse tipo de análise pode consumir muito tempo, sendo necessária a diminuição do tamanho do set dos descritores excluindo-se de algumas variações cuidadosamente para não selecionar aquelas que apresentam alta correlação. O uso de softwares estatísticos como R e MATLAB podem automatizar esse processo com a escolha adequada das condições.

Análise de PCA trata-se de uma diminuição do espaço analisado selecionando-se as variáveis que mais contribuem e melhor descrevem o conjunto de informações. Dessa forma, consegue ultrapassar os problemas encontrados pela MLR, pois extrai informações de um conjunto de dados menor e já tratado. Fornece uma ferramenta eficiente para redução do número de variáveis independentes utilizadas no modelo QSAR, sendo muito utilizado em sistemas com grande número de descritores moleculares. A desvantagem desse método é que os resultados são difíceis de serem analisados, pois nem sempre possibilitam a identificação de uma estrutura particular ou propriedades físico-químicas importantes para atividade biológica.

Análise de Mínimos Quadrados Parciais é uma combinação de PCA e MLR no qual a atividade biológica, variável dependente, também é extraída para novas componentes que otimizam a correlação.¹⁶

Sistemas biológicos frequentemente apresentam relações não lineares, por esse motivo as redes neurais são o principal método de análise de regressão não linear para a construção de um modelo QSAR.¹⁵⁻¹⁶ Nesse método, a rede neural aprende a associação entre os descritores moleculares e a atividade biológica relacionada a ele baseado no treinamento de um conjunto de moléculas. Além disso, uma rede neural treinada pode realizar o processo inverso, ou seja, prever a atividade, propriedades físico-químicas e a estrutura para uma nova molécula. Como qualquer outro processo de regressão, as redes neurais estão suscetíveis ao "overfitting" o que pode gerar modelos com baixo poder de predição. Redes neurais regularizadas por métodos Bayesianos são uma variação das redes neurais, podem ser utilizadas para desenvolver um modelo de QSAR não lineares e tem como vantagem eliminar o processo de "overfitting" e otimizar a rede neural.¹⁶

1.6.2 Validação

Há dois principais tipos de validação de um modelo QSAR: (i) Validação interna e (ii) validação externa. No caso de validação interna a abordagem mais utilizada é a *leave-one-out cross validation*, no qual um composto observado é deixado como dado para validação enquanto que o resto dos dados forma o conjunto de teste usados para determinar os coeficientes do modelo, em seguida a atividade dos compostos testados é predita usando-se o modelo baseado no conjunto de dados treinados, processo repetido para todos os compostos até que todos tenham sido usados como compostos de teste pelo menos uma vez. O poder de predição do modelo criado é então determinado através da medida de Q^2 .

$$Q^2 = 1 - \frac{\sum (y_{pred} - y_{obs})^2}{\sum (y_{obs} - y_{mean})^2} \quad (2)$$

A principal desvantagem desse método é que o tempo de processamento aumenta com o quadrado do tamanho do conjunto de dados, dessa forma sendo impróprio para conjuntos de dados muito grandes. Uma variação desse método é o *k-fold cross validation*, na qual um subconjunto de tamanho k , definido pelo usuário, de dados é deixado como validação, apresenta melhora da performance de processamento devido à redução na quantidade de dados do conjunto de treinamento.¹⁶ Ambas as metodologias são falhas quando todos os dados são usados ao mesmo tempo para validação do modelo. Métodos de validação externa para predição da

atividade envolvem o uso de moléculas externas não utilizadas no desenvolvimento do modelo QSAR, pode utilizar todos os dados disponíveis no conjunto de treinamento para a validação.

1.6.4 QSAR 3D

QSAR 3D envolve descritores que descrevem as características tridimensionais de uma molécula para o desenvolvimento de um modelo QSAR.¹⁶ Propriedades geométricas, físicas e descritores químicos quânticos podem ser utilizados para descrever as características do ligante em um QSAR 3D. Em seguida, os descritores são combinados para formar um modelo capaz de explicar a atividade biológica do ligante, definido pela orientação espacial 3D de várias características, como doadores e aceptores de ligação de hidrogênio, que são essenciais para a atividade biológica desejada. O modelo criado é então testado analisando-se a estabilidade e os valores estatísticos para determinar o modelo QSAR 3D final. Há diversos tipos de técnicas para modelagem de QSAR 3D, falaremos brevemente sobre o CoMFA.

1.6.5 CoMFA

Análise comparativa de campos moleculares (Comparative Molecular Field Analysis)¹⁶ é o método de QSAR 3D mais utilizado, foi o primeiro a relacionar as propriedades estéricas e eletrostáticas dependentes da forma 3D de uma molécula a sua atividade biológica. As moléculas são alinhadas em um grid 3D baseado na sua estrutura tridimensional e os valores das energias potenciais, estéricos e eletrostáticos, são calculados para cada ponto do grid para cada molécula e então são associados à atividade biológica.

CoMFA costuma associar a conformação com menor energia livre como sendo a bioativa, quando a estrutura cristalográfica é conhecida as coordenadas do cristal podem ser utilizadas para definir a conformação bioativa. PCA e PLS são ferramentas úteis para o processo de desenvolvimento do modelo. Após definido, o modelo é testado e são verificadas suas propriedades estatísticas e a sua robustez. O sucesso e a habilidade de prever dos modelos CoMFA são extremamente sensíveis ao alinhamento das conformações bioativas.¹⁶ Pelo fato da conformação bioativa não ser necessária na ausência do receptor, a associação feita pelo CoMFA na escolha da conformação bioativa e no alinhamento correspondente podem produzir modelos problemáticos e errôneos. Apresenta limitações em sua aplicabilidade, pois não considera a natureza dinâmica dos ligantes. Outra limitação é a forma da função de energia, pois não considera hidrofobicidade e as interações de hidrogênio.¹⁶ Para descrever as interações

estéricas e eletrostáticos utiliza um potencial de Lennard-Jones e potencial Coulombico, respectivamente. Entretanto, o uso dessas funções pode fornecer valores de energias extremamente altos e irreais, a escolha de um valor de corte (Cutoff Value) para os potenciais pode evitar esse tipo de comportamento.

1.7 Similaridade molecular

O Princípio da Similaridade Molecular afirma que moléculas estruturalmente semelhantes devem apresentar propriedades físico-químicas e biológicas semelhantes mais do que moléculas dissimilares.²³ No contexto biológico, para uma molécula se ligar ao sítio ativo de seu alvo, ambos devem ter a capacidade de se reconhecer e interagir, propriedades como superfície molecular e eletronegatividade desempenham papel importante nesse processo. Ao se conhecer o sítio ativo e, pelo menos, um composto que se liga a esse sítio a busca por moléculas que apresentem atividade e funcionalidades semelhantes ao composto se torna de grande importância, pois moléculas similares devem ter alta probabilidade de se ligar ao mesmo sítio de ligação enquanto que moléculas dissimilares apresentam baixa probabilidade. Dessa forma, a busca por moléculas similares se tornou uma ferramenta importante e amplamente utilizado no processo de Desenho de Drogas. Entretanto, similaridade não é uma propriedade intrínseca das moléculas e sim definida segundo método e métricas que quantificam quão similar ou dissimilar uma molécula é da outra.²³⁻²⁴

Para mensurar a similaridade entre moléculas podem ser utilizadas diversas abordagens, geralmente divididas de acordo com os descritores moleculares adotados. Conforme definido por Todeschini e Consonni: "Descritores moleculares são o resultado final de um processo matemático e lógico que transforma informações químicas e físicas em uma representação simbólica em um número útil ou resultado que possa ser utilizado em algum experimento"²⁴ (Figura 6). Bons descritores moleculares devem apresentar algumas características:

- Ser invariantes aos tipos de átomos e as quantidades;
- Ter uma definição única fornecida por um algoritmo;
- Valores plausíveis para as moléculas no intervalo em que será usado;
- Ter interpretação estrutural;
- Ter boa correlação com pelo menos uma propriedade;
- Sem correlações triviais com outros descritores;
- Mudanças graduais em seus valores associadas com mudanças graduais na estrutura da molécula;

- Abranger o maior número de estruturas moleculares possíveis;
- Diferenciar moléculas isômeras;
- Ser simples;
- Preferencialmente, permitir reversão, ou seja, a partir dos valores dos descritores encontrar a estrutura;

São classificados em duas classes: Medidas Experimentais e Descritores Moleculares Teóricos. Dentro da primeira classe se encontram:

- Coeficiente de Partição (*Log P*), definido como o logaritmo da razão entre a solubilidade de um composto em um solvente orgânico e a mesma solubilidade em água;
- Momento de dipolo;
- Polarização;
- Peso molecular

Dentro da segunda classe, há subdivisão em relação a dimensão dos descritores: ^{23, 25}

0D - Adimensional - Constituintes

São descritores que utilizam informações referentes aos elementos que compõe a estrutura da molécula, como massa molecular, número de átomos, número de átomos de carbono, número de ligações simples, duplas, triplas, aromáticas e rotáveis e número de anéis aromáticos.

1D - Unidimensional - Contadores de fragmentos e fingerprints

São descritores que utilizam informações a respeito da contagem de fragmentos, como número de carbonos sp³, amidas aromáticas, aminas alifáticas, grupos nitro, ésteres e doadores e aceptores de ligação de hidrogênio.

2D - Bidimensional - Descritores Gráficos, topológicos, fingerprints 2D

São descritores que utilizam informações a respeito da topologia da molécula como vetores de correlação 2D, topologia estrutural, topologia química, subestruturas comuns, índice de Zagreb, índice de Wiener entre outros.

3D - Tridimensional - Descritores geométricos, superfície, campos moleculares

São descritores que utilizam informações sobre a estrutura 3D dos ligantes, coordenadas atômicas, como cálculo de campos a partir de valores de eletronegatividade atômica, propriedades da superfície, descrição da superfície através do uso de harmônicos esféricos, método CoMFA, vetores de autocorrelação 3D, índice de Wiener 3D, funções de distribuição radial (RDF), entre outros.

Figura 6- Esquema representando a construção de descritores moleculares, mostrando de onde são extraídas as informações, onde podem ser aplicados e por quem podem ser processados.

Fonte: Adaptada de TODESCHINI²⁴

1.8 Similaridade estrutural tridimensional – Conceitos Básicos

1.8.1 Algoritmo Genético

Algoritmo pertencente a classe dos algoritmos evolutivos, ou seja, utiliza ideias a respeito de seleção natural e genética e seguindo o princípio da sobrevivência do mais apto, foi proposto inicialmente por John Holland na década de 70.^{23, 26} Pode ser usada em diversos campos, principalmente o de otimização e buscas, dessa forma não sendo específico para análise estrutural. Simulam a evolução dos elementos das populações aptos a estar no espaço de dados analisado. Os elementos da população apresentam “cromossomos”, responsável por conter suas principais características e que estão sujeitos à ação dos operadores genéticos, seleção mutação e crossing-over. Operador seleção é responsável por filtrar os elementos mais aptos a sobreviver no espaço analisado e dessa forma transmitir seus "genes" para os novos elementos, seleção feita de acordo com análise de seus resultados perante a uma determinada métrica ou análise subjetiva. O operador mutação introduz alterações randômicas modifica nos cromossomos de um elemento, apresenta baixa incidência para promover alterações em apenas uma parte da população, mantendo assim a diversidade. Para o operador crossover, é feita a recombinação de partes dos cromossomos de dois elementos distintos, selecionados pelo

operador seleção, simulando o processo de reprodução sexuada, e em ambos os casos novos indivíduos são gerados e adicionado a população, tem como objetivo a busca por elementos com maior chance de sobreviver. O processo é dividido em etapas evolucionárias composta de 3 passos: Seleção dos indivíduos, aplicação de operadores genéticos e avaliação da população.²³

26

Ao final de cada etapa evolucionária, todos os indivíduos são avaliados de acordo com a função de aptidão, que fornece a probabilidade de cada indivíduo sobreviver, de acordo com essa avaliação as modificações genéticas são ou não carregadas para a próxima geração. Em estudos sobre estrutura moleculares, os cromossomos contêm informações sobre ângulos torcionais e os graus de liberdade das ligações.

Apresentam grande potencial de uso no campo de análise de estruturas moleculares, pelo fato de serem flexíveis e de conseguirem lidar com objetos complexos, entretanto apresentam alto custo computacional, sendo essa sua principal limitação de uso.

1.8.2 Algoritmos para similaridade de estruturas tridimensionais

Algoritmos utilizados na determinação da similaridade entre estruturas tridimensionais costumam ser divididos em duas principais classes: abordagens invariantes a transformações e abordagens que precisam de alinhamento para determinar a similaridade. Além dessa primeira divisão, ainda podem ser subdivididos de acordo com a flexibilidade molecular: Rígidos, no qual a estrutura molecular é mantida fixa, e flexível, no qual a estrutura pode ser alterada durante o processamento. A similaridade pode ser avaliada de acordo com posição atômica, baseado em um farmacóforo, volume da molecular e superfície molecular²³, nessa dissertação focaremos na análise da superfície molecular através do uso de harmônicos esféricos para descrever a superfície molecular.

1.8.3 Similaridade baseada na superfície da molécula

Tem como objetivo encontrar similaridades moleculares capazes de explicar a funcionalidade de uma molécula. Para isso, utiliza informações a respeito das partes das moléculas mais envolvidas no processo de ligação e reconhecimento molecular.²³ Essas partes podem ser eficientemente descritas através de conceitos de superfícies moleculares. Abordagem tem ganhado adeptos nas duas últimas décadas propiciando a criação e desenvolvimento de novas técnicas de análise. Dentre os principais métodos de análise destacam-se Análise de Correlação, Similaridade morfológica e Harmônicos Esféricos.

1.8.4 Curvas de Correlação (Correlation Plots)

Tem como estratégia descrever a superfície molecular através de representações discretas. Katchalski-Katzir utilizam a representação da molécula em um grid, no qual cada nó pode apresentar dois status: dentro ou fora da superfície molecular. Por sua vez, Barequet e Sharir utilizam pontos da superfície molecular.²³ Assumindo que as moléculas estão centradas na origem, várias rotações em volta da origem são realizadas utilizando-se ângulos de Euler com variações angulares pré-definidas e para cada rotação calcula-se a correlação entre as superfícies para os diferentes vetores translacionais. Katchalski-Katzir utilizam transformações de Fourier para analisar as funções de correlação, enquanto que Barequet e Sharir calculam os vetores translacionais para cada par de pontos representando as moléculas. Para todas as rotações que apresentam altos picos nas curvas de correlação, a vizinhança em torno dessa rotação é analisada com mais detalhes. O processo é feito de forma recursiva até que a melhor correlação seja encontrada. Para a rotação final, o melhor vetor translacional pode ser facilmente utilizado na busca do melhor alinhamento das superfícies moleculares. A principal desvantagem dessa implementação é a necessidade de um grande número de rotações para garantir transformações importantes.

1.8.5 Similaridade Morfológica (Morphological similarity)

O conceito de similaridade morfológica é baseado em um conjunto de observadores das superfícies moleculares e das propriedades presentes.²⁷ Toda superfície é posicionada em um grid 3D e cada ponto do grid fora da superfície molecular é considerado um observador. Para cada observador é calcula-se e armazena-se a distância mínima a superfície, a distância do doador de ligação de hidrogênio mais próximo ou do átomo carregado positivamente e a distância do acceptor de ligação de hidrogênio mais próximo ou átomo carregado negativamente. Para observadores de grupamentos polares também são usadas informações direcionais. A similaridade entre duas moléculas é calculada como uma soma ponderada e normalizada de funções Gaussianas referentes às diferenças nas distancias dos observadores. Pelo fato da função ser contínua e diferenciável, otimização baseada em gradiente pode ser aplicada.

1.8.6 Harmônicos Esféricos

Devido ao fato de muitas moléculas apresentarem ou quase apresentarem forma globular, foi proposto que sua superfície poderia ser descrita através de uma expansão de

harmônicos esféricos.²³ Max e Getzloff demonstraram que o uso de baixas ordens de harmônicos esféricos é capaz de capturar as características principais da superfície. O refinamento da superfície molecular pelo uso de ordens mais elevadas foi demonstrado por Duncan e Olson e por Leicester. Além da visualização e classificação de moléculas através de harmônicos esféricos, Ritchie e Kemp propuseram o alinhamento das superfícies baseadas na rotação dos coeficientes da expansão e otimizaram o alinhamento pela aplicação de um algoritmo quasi-Newton.²⁸

1.9 Harmônicos Esféricos

Harmônicos esféricos são a parte angular da solução da equação de Laplace.²⁹ Apresentam esse nome pois são definidos sobre a superfície da esfera e pelo fato das soluções da equação de Laplace serem chamadas de funções harmônicas.³⁰ São definidos em termos dos parâmetros angulares (θ, φ) e podem ser descritos por:

$$Y_l^m(\theta, \varphi) = \sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}} P_l^m(\cos \theta) e^{im\varphi} \quad (3)$$

No qual os números quânticos l , momento quântico angular, e m , momento quântico azimutal, pertencem aos intervalos $l \geq 0$ e $|l| \geq m$, respectivamente e o termo P_l^m representa os polinômios associadas de Legendre. Apresentam diversas aplicações como descrição de superfícies 3D^{10, 29, 31-35}, configurações dos orbitais atômicos em mecânica quântica e iluminação em modelos 3D.³¹ Na Figura 7 tem-se a representação geométrica das bandas dos

harmônicos

esféricos.

Figura 7- Representação geométrica das bandas dos harmônicos esféricos. Para cada banda são encontradas $2l + 1$ componentes.

Fonte: GREEN³¹

1.9.1 Equação de Laplace em coordenadas esféricas

A equação de Laplace propõe que um o divergente do gradiente de um campo escalar f apresenta valor zero, conforme equação abaixo:

$$\nabla^2 f = 0 \quad (4)$$

Em coordenadas esféricas pode ser representada por:

$$\nabla^2 f = \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial f}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial f}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 f}{\partial \varphi^2} = 0 \quad (5)$$

Apresenta como solução da equação em coordenadas esféricas funções do tipo

$$f(r, \theta, \varphi) = R(r)Y(\theta, \varphi) \quad (6)$$

obtidas pelo uso do método de separação de variáveis. Dessa forma, a equação pode ser reescrita em função somente de r , parte radial, e em função dos ângulos θ e φ , a parte angular.

$$\frac{1}{R} \frac{d}{dr} \left(r^2 \frac{dR}{dr} \right) = \lambda \quad (7)$$

$$\frac{1}{Y} \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial Y}{\partial \theta} \right) + \frac{1}{Y} \frac{1}{\sin^2 \theta} \frac{\partial^2 Y}{\partial \varphi^2} = -\lambda \quad (8)$$

Entretanto a parte angular pode ser novamente separada em uma função

$$Y(\theta, \varphi) = \Theta(\theta)\Phi(\varphi) \quad (9)$$

soluções das equações abaixo pela dependência dos ângulo azimutal(Φ) e ângulo polar(θ).

1.9.2 Dependência do ângulo azimutal

A dependência do ângulo azimutal é representada pela equação abaixo:

$$\frac{1}{\Phi} \frac{d^2 \Phi}{d\varphi^2} = -m^2 \quad (10)$$

Com soluções

$$\Phi(\varphi) = e^{im\varphi} \quad (11)$$

$$\Phi(\varphi) = e^{-im\varphi} \quad (12)$$

e m inteiro, que satisfaça a equação de ortogonalidade

$$\int_0^{2\pi} e^{im_1\varphi} \cdot e^{-im_2\varphi} d\varphi = 2\pi \delta_{m_1 m_2} \quad (13)$$

no qual $\delta_{m_1 m_2}$ é chamado de delta de Kronecker, função de duas variáveis, geralmente dois valores positivos inteiros definida por:

$$\delta_{m_1 m_2} = \begin{cases} 0 & \text{se } m_1 \neq m_2 \\ 1 & \text{se } m_1 = m_2 \end{cases} \quad (14)$$

Como solução da equação de ortogonalidade tem-se que a dependência azimutal pode ser descrita pela equação

$$\Phi_m = \frac{1}{\sqrt{2\pi}} e^{im\varphi} \quad (15)$$

ortonormal, ortogonal e normalizada, em relação ao ângulo azimutal.

1.9.3 Dependência do ângulo polar

A dependência do ângulo polar é descrita pela equação associada de Legendre abaixo

$$\lambda \sin^2 \theta + \frac{\sin \theta}{\theta} \frac{d}{d\theta} \left(\sin \theta \frac{d\Theta}{d\theta} \right) = m^2 \quad (16)$$

satisfeita pelas funções associadas de Legendre

$$\Theta(\theta) = P_n^m(\cos \theta) \quad (17)$$

Entretanto, podem-se incluir valores negativos para m e para definir as funções associadas com valores negativos usa-se a fórmula de Rodrigues.

$$P_n^m(\cos \theta) = \frac{1}{2^n n!} (1 - x^2)^{m/2} \frac{d^{m+n}}{dx^{m+n}} (x^2 - 1)^n, -n \leq m \leq n \quad (18)$$

A ortogonalidade do ângulo polar é definida pela equação:

$$\int_0^\pi P_p^m(\cos \theta) P_q^m(\cos \theta) \sin \theta d\theta = \frac{2}{2q+1} \frac{(q+m)!}{(q-m)!} \delta_{pq} \quad (19)$$

Cuja solução são funções ortonormais

$$\Theta(\theta) = \sqrt{\frac{2n+1}{2} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos \theta), -n \leq m \leq n \quad (20)$$

Pela multiplicação das funções ortonormais Φ_m e $\Theta(\theta)$ definem-se os Harmônicos Esféricos.

$$Y_n^m(\theta, \varphi) \equiv (-1)^m \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos \theta) e^{im\varphi} \quad (21)$$

A ortogonalidade deles é descrita segundo a equação.

$$\int_{\varphi=0}^{2\pi} \int_{\theta=0}^{\pi} Y_{n_1}^{m_1*}(\theta, \varphi) \cdot Y_{n_2}^{m_2}(\theta, \varphi) \sin \theta d\theta d\varphi = \delta_{m_1 m_2} \delta_{n_1 n_2} \quad (22)$$

1.9.4 Teorema da Expansão em Harmônicos Esféricos (Séries de Laplace)

Os harmônicos esféricos representam um conjunto completo de funções ortonormais responsáveis por formar um espaço vetorial análogo aos vetores unitários da base. Vetores da base são vetores linearmente independentes que por meio de combinações lineares geram um espaço vetorial. Dessa forma, qualquer função de harmônicos esféricos pode ser descrita como combinação linear de funções que formam a base do espaço vetorial. As projeções nos eixos cartesianos, pelo uso de produto escalar entre dois vetores, também podem ser aplicadas para funções de harmônicos esféricos.³⁴

Pelo fato de formarem uma base completa de um espaço vetorial, os harmônicos esféricos apresentam a propriedade da completude, consequência da forma de Sturm-Liouville da equação de Laplace. Essa propriedade indica que qualquer função $f(\theta, \varphi)$ que apresente propriedades de continuidade suficientes avaliada sobre a superfície de uma esfera pode ser expandida em uma série duplamente convergente de harmônicos esféricos (Série de Laplace).³⁰

$$f(\theta, \varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^l c_{lm} Y_{lm}(\theta, \varphi) \quad (23)$$

Figura 8- Reconstrução da forma de um objeto usando harmônicos esféricos. Conforme aumenta-se a quantidade de bandas mais detalhes são atribuídos a superfície reconstruída deixando-a cada vez mais próxima do original.

Fonte: GREEN³¹

A expansão de harmônicos esféricos é exata e não uma aproximação, os erros introduzidos são fruto da limitação dos valores de l .³⁴ (Figura 8) Os valores de l são inteiros, começam em zero e não há limite superior e os valores de m estão compreendidos entre $-l$ e $+l$, a quantidade de componentes é dada em função do número de bandas utilizadas na descrição sendo igual a $(l_{max} + 1)^2$ coeficientes. Para a representação de uma superfície molecular, séries com coeficientes entre $l = 0$ até $l = 6$ são suficientes para representá-la com robustez. O aumento do número de termos em uma série tende a diminuir o erro associado (Figura 9), entretanto segundo Morris há um limite superior, sendo $l_{máx} = 14$, no qual o erro proveniente da reconstrução da superfície molecular diminui até esse valor e em seguida volta a aumentar conforme a quantidade de termos que formam a série aumenta.¹⁰

Figura 9 - Erro da expansão em harmônicos esféricos. Estudo conduzido por Morris revela que no processo de descrição molecular baseada em harmônicos esféricos o valor de RMSD diminui conforme aumenta-se a quantidade bandas porém somente até o valor $l_{\max} = 14$, no qual o erro é mínimo e em seguida voltar a crescer.

Fonte: KAHRAMAN¹⁰

Os valores c_{lm} são os coeficientes da expansão, são únicos, podem ser utilizados diretamente como descritores da superfície molecular e são definidos pela multiplicação da expansão pelo complexo conjugado do harmônico esférico (Y_{lm}^*) e integrando-se sobre o ângulo sólido Ω .³⁴

$$c_{lm} = \int_{S_2} f(\theta, \varphi) Y_{lm}^*(\theta, \varphi) d\Omega = \int_0^{2\pi} \int_0^{\pi} f(\theta, \varphi) Y_{lm}^*(\theta, \varphi) \sin\theta d\theta d\varphi \quad (24)$$

1.9.5 Rotação dos Harmônicos Esféricos

Durante análise de similaridade de superfícies moleculares, a orientação na qual as moléculas se encontram é de extrema importância, sendo um grande desafio posicioná-las de forma adequada. O processo de rotação molecular pode ser feito através de uma transformação de rotação na qual todas as coordenadas espaciais são rotacionadas em relação aos eixos da base. Para o caso dos coeficientes da expansão em harmônicos esféricos as transformações de rotação podem ser feitas diretamente nos coeficientes, conforme demonstrado por Ritchie e Kemp.²⁸ Uma transformação de rotação pode ser feita através do uso de ângulos de Euler ($\alpha\beta\gamma$) e matrizes de Wigner, conforme equação abaixo:

$$\hat{R}(\alpha\beta\gamma) Y_{lm}(\theta, \varphi) = \sum_{m'} Y_{lm'}(\theta, \varphi) D_{m'm}^{(l)}(\alpha\beta\gamma) \quad (25)$$

No qual $\hat{R}(\alpha\beta\gamma)$ representa uma operação de rotação em termos dos ângulos de Euler e $D_{m'm}^{(l)}$ são as matrizes de rotação unitárias de Wigner, nos quais os elementos são definidos por

$$D_{m'm}^{(l)}(\alpha\beta\gamma) = e^{-im'\alpha} d_{m'm}^{(l)}(\beta) e^{-im\gamma} \quad (26)$$

onde

$$d_{m'm}^{(l)}(\beta) = \left[\frac{(1+m')!(1-m')!}{(l+m')!(l-m)!} \right]^{1/2} x \sum_{k=\max(0, m-m')}^{\min(l-m', l+m)} \left[(-1)^{k+m'-m} \binom{l+m}{k} x \binom{l-m}{l-m'-k} \right] x \cos(\beta/2)^{2l+m-m'-2k} \sin(\beta/2)^{2k+m'-m} \quad (27)$$

A operação de rotação $\hat{R}(\alpha\beta\gamma)$ aplicada a uma função $r(\theta, \varphi)$ cria uma nova representação $r'(\theta, \varphi)$ dos mesmos coeficientes, porém em uma orientação diferente.

$$r'(\theta, \varphi) = \hat{R}(\alpha\beta\gamma) r(\theta, \varphi) \quad (28)$$

Substituindo corretamente as funções harmônicas da expansão e multiplicando-se pelo complexo conjugado $Y_{lm}^*(\theta, \varphi)$ em ambos os lados da equação de rotação e integrando-se sobre toda a superfície da esfera, temos que os novos coeficientes são definidos por:

$$a'_{lm} = \sum_{m'} a_{lm'} D_{m'm}^{(l)}(\alpha\beta\gamma) \quad (29)$$

Para representações de superfícies moleculares o uso de harmônicos esféricos complexos não é adequada, pois estamos querendo descrever uma superfície real. Dessa forma, o uso de harmônicos esféricos reais é mais adequada e podem ser definidos por meio de combinações lineares de harmônicos esféricos complexos.

$$y_{lm}(\theta, \varphi) = \begin{cases} \frac{1}{\sqrt{2}} (Y_{lm}(\theta, \varphi) + Y_{lm}^*(\theta, \varphi)), & \text{para } m > 0 \\ Y_{l,0}(\theta, \varphi) & , \text{para } m = 0 \\ \frac{\sqrt{-i}}{2} (Y_{l\bar{m}}(\theta, \varphi) - Y_{l\bar{m}}^*(\theta, \varphi)), & \text{para } m < 0 \end{cases} \quad (30)$$

As propriedades de rotação são preservadas para a nova definição e dessa forma os coeficientes são definidos por

$$a'_{lm} = \sum_{m'} a_{lm'} R_{m'm}^{(l)}(\alpha\beta\gamma) \quad (31)$$

nos quais os valores de $R_{m'm}^{(l)}(\alpha\beta\gamma)$ foram definidos explicitamente (Figura 10) por Ritchie e Kemp.²⁸

$$R_{m'm}^{(l)} = \begin{cases} d_{m'm}^{(l)}(\beta) \cos(m\gamma + m'\alpha) + (-1)^{m'} d_{\bar{m}'m}^{(l)}(\beta) \cos(m\gamma - m'\alpha); & m' > 0, m > 0; \\ d_{0m}^{(l)}(\beta) \sqrt{2} \cos(m\gamma); & m' = 0, m > 0; \\ (-1)^{m'+1} d_{m'm}^{(l)}(\beta) \sin(m\gamma + m'\alpha) + d_{\bar{m}'m}^{(l)}(\beta) \sin(m\gamma - m'\alpha); & m' < 0, m > 0; \\ d_{m'0}^{(l)}(\beta) \sqrt{2} \cos(m'\alpha); & m' > 0, m = 0; \\ d_{00}^{(l)}(\beta); & m' = 0, m = 0; \\ (-1)^{m'+1} d_{m'0}^{(l)}(\beta) \sqrt{2} \sin(m'\alpha); & m' < 0, m = 0; \\ (-1)^m d_{m'm}^{(l)}(\beta) \sin(m\gamma + m'\alpha) + (-1)^{m+m'} d_{\bar{m}'m}^{(l)}(\beta) \sin(m\gamma - m'\alpha); & m' > 0, m < 0; \\ (-1)^m d_{0m}^{(l)}(\beta) \sqrt{2} \sin(m\gamma); & m' = 0, m < 0; \\ (-1)^{m+m'} d_{m'm}^{(l)}(\beta) \cos(m\gamma + m'\alpha) + (-1)^{m+1} d_{\bar{m}'m}^{(l)}(\beta) \cos(m\gamma - m'\alpha); & m' < 0, m < 0. \end{cases}$$

Figura 10- Representação explícita dos valores para a matriz de rotação calculada por Ritchie e Kemp.
Fonte: RITCHIE; KEMP²⁸

1.9.6 Harmônicos esféricos rotacionalmente invariantes

Coeficientes da expansão dos harmônicos esféricos podem ser feitos invariantes, ou seja, apresentarem sempre o mesmo valor independente da orientação espacial. Entretanto essa propriedade só é válida para a norma L^2 , pois é a representação da energia da função esférica³⁵⁻³⁶, para qualquer valor de l .

$$EI(l) = \sum_{|m| \leq l} \|C_l^m\|^2 \quad (32)$$

Contudo a forma aplicada por Kazhdan apresenta problemas na discriminação de formas, pelo fato de moléculas dissimilares poderem apresentar coeficientes similares e moléculas similares poderem apresentar coeficientes distintos.³⁶ Além disso, há perda de informação nessa representação, pois para norma L^1 existem $(l_{max} + 1)^2$ coeficientes, enquanto que para a norma L^2 há apenas $(l_{max} + 1)$ coeficientes. A perda de informação pode ocorrer em dois momentos.³⁵

- Sejam $f = \sum_{l=0}^b f_l$ e $g = \sum_{l=0}^b R_l(f_l)$ onde R_l são as rotações aplicadas a cada frequência dos harmônicos esféricos, tem-se que os valores de f e g apresentam o mesmo valor, ou seja, independente da rotação aplicada o valor da representação não será alterado.

- A superfície original pode não ser reconstruída completamente pois para cada banda dos harmônicos esféricos (l) é armazenado somente a energia descrita por ela. Para bandas em que $l \geq 2$ não é verdadeiro que $||f_l|| = ||g_l||$ sendo assim há uma rotação R que satisfaz a equação $R(f_l) = g_l$.

Um conjunto de harmônicos esféricos invariantes completos é proposto por *Hongdong*. Os descritores de um conjunto completo não sofrem ambiguidade no processo de classificação e reconhecimento de formas semelhantes. O princípio utilizado é de que harmônicos esféricos de para qualquer banda $l, l \geq 1$ formam uma representação irredutível de um grupo $SO(3)$ ³⁶, grupo de todas as rotações no espaço Euclidiano R^3 sobre a origem através de uma operação de composição. Ao aplicar-se uma rotação à função original, os coeficientes resultantes apresentarão os mesmos valores que os coeficientes da função original.

Os novos valores dos descritores podem ser obtidos através de uma rotação usando-se ângulos de Euler ($\alpha\beta\gamma$), ao se especificar alguns harmônicos esféricos com valores canônicos é possível determinar uma rotação canônica $R(\alpha^*\beta^*\gamma^*)$, de forma que os coeficientes resultantes sejam rotacionalmente invariantes, conforme definição dos coeficientes abaixo.³⁶

$$C_l^m = e^{-im(\alpha+\pi/2)} \cdot \sum_{|m'| < l} e^{-im'(\gamma+\pi/2)} C_l^{m'} \cdot \sum_{|k| < l} P_l^{m'k}(0) P_l^{mk}(0) e^{-ik(\beta+\pi)} \quad (33)$$

A similaridade entre duas funções harmônicas pode ser determinada pela diferença entre as energias armazenadas em cada banda, norma L^2 , usando-se distância euclidiana.

$$\begin{aligned} & ||SH(f) - SH(g)||^2 \\ &= \sum_{l=0}^{\infty} (||f_l - g_l||)^2 \leq \sum_{l=0}^{\infty} (||f_l - g_l||)^2 = ||f(\theta, \phi) - g(\theta, \phi)||^2 \end{aligned} \quad (34)$$

Pelo fato dos harmônicos esféricos serem rotacionalmente invariantes tem-se que a diferença entre as bandas pode ser descrita:

$$||SH(f) - SH(g)|| \leq \min_{R \in SO(3)} ||f - R(g)|| \quad (35)$$

2 OBJETIVOS

Desenvolver método de análise de similaridade entre ligantes utilizando harmônicos esféricos como descritores da superfície molecular.

Para isso, foram propostas as seguintes metas:

- Adquirir conhecimento a respeito do processo de desenho racional de fármacos;
- Adquirir conhecimento sobre descritores moleculares e medidas de similaridade;
- Utilizar harmônicos esféricos como descritores da superfície molecular;
- Desenvolver ferramenta utilizando um método baseado em harmônicos esféricos capaz de auxiliar o processo de desenho racional de fármacos;
- Validar ferramenta através de um benchmarking utilizando a métrica do Ultrafast Shape Recognition (USR) como padrão de comparação
- Criar Website para divulgação e utilização do método desenvolvido como ferramenta para o processo de desenho racional de fármacos;

3 MATERIAL E MÉTODOS

3.1 Material

Esse projeto foi desenvolvido no grupo de Cristalografia do Instituto de Física de São Carlos, Universidade de São Paulo (IFSC-USP). Todos os métodos computacionais foram realizados utilizando um computador Lenovo com processador Intel® Core™ i5-3470S CPU @2,90Ghz com memória de 8Gb DDR3 rodando uma máquina virtual criada com VMware Player Workstation 9.0 Virtual Machine utilizando o sistema operacional Xubuntu 12.04 Xface 64bits contando com 1 processado, 2 Gb de memória e HD de 20Gb.

3.2 Métodos

3.2.1 Implementação Lira

O programa Lira é utilizado para determinar os descritores da superfície molecular de pequenas moléculas usando uma expansão de harmônicos esféricos e suas propriedades e a partir desses descritores determinar os ligantes mais similares a ela encontrados um banco de dados de moléculas. O programa é escrito em Python e Fortran 90, sendo que a parte gráfica e de montagem de banco de dados feita em Python e os cálculos matemáticos mais complexos realizado em Fortran. As funções escritas em Fortran são executadas dentro dos scripts em Python através da ferramenta f2py, responsável por criar uma interface capaz de estabelecer interação entre ambas moléculas.

Os descritores da superfície molecular são determinados através de um arquivo no formato PDB³⁷, contendo as coordenadas atômicas tridimensionais (x,y,z) de todos os átomos e as conexões existentes entre eles, determinadas por experimentos de Cristalografia e Ressonância Magnética Nuclear. Para o cálculo dos harmônicos esféricos as coordenadas devem estar descritas no sistema ISO (Figura 11), também conhecidas como coordenadas esféricas, dependentes das do raio, ângulo polar e ângulo azimutal (r, θ, φ). O raio ou distância radial é o cálculo da distância euclidiana da origem até um determinado ponto P, a inclinação ou ângulo polar é o ângulo entre a direção $\hat{z}$ e o segmento de linha $\overline{OP}$ e o ângulo azimutal é medido entre a projeção ortogonal do segmento $\overline{OP}$ no plano XY e a direção $\hat{x}$. As coordenadas são extraídas através de um PDBParser, transformadas para o sistema ISO e armazenadas em vetores. Em seguida, as bases do espaço vetorial de harmônicos esféricos são calculadas, as coordenadas são expandidas em função das novas bases harmônicas, calcula-se os valores dos

coeficientes da expansão, os chamados descritores moleculares e a partir dos descritores a superfície molecular é reconstruída.

Figura 11 - Coordenadas polares esféricas
Fonte: ARFKEN³⁰

A metodologia utilizada para implementação do Lira é baseada no paper "Spherical Harmonic Lighting: The Gritty Details" de Robin Green³¹, no qual os harmônicos esféricos são utilizados para descrever o sombreamento, iluminação e efeitos de luz em modelos 3D para jogos. Utiliza integração de Monte Carlo para calcular as integrações sobre os ângulos sólidos e assim mapear todos os pontos de uma superfície esférica que será utilizada na determinação das bases harmônicas do Espaço Vetorial. Além do processo de integração, propõe como calcular os valores dos harmônicos esféricos através dos Polinômios Associados de Legendre e a fórmula de Rodrigues.

3.2.1.1 PDBParser

Método utilizado para ler e extrair as informações atômicas a partir de um arquivo do Banco de Dados de Proteínas (PDB). Existem bibliotecas implementadas para tal uso como o PDBParser encontrado dentro da Biopython, porém o Lira utiliza uma versão própria.

O arquivo PDB é lido linha a linha e verifica-se a presença das literais de heteroátomos ("HETATM") ou átomos ("ATOM") nas substrings da linha a fim de verificar se aquela linha representa um átomo ou uma ligação atômica (Figura 12). Em caso de representar um átomo, são extraídas informações a respeito do número serial do átomo, coordenadas atômicas e o elemento químico. Cada informação é atribuída a uma variável, para as coordenadas cartesianas é utilizado um vetor real de três posições, em seguida essas variáveis são armazenadas em uma lista. Em caso de a linha representar uma ligação, são extraídas informações a respeito do número serial do átomo e as quais átomos o átomo representado se liga.

Todas as informações coletadas são atribuídas a um objeto do tipo molécula, que será usado pelo programa para os cálculos das coordenadas atômicas esféricas e dos harmônicos esféricos.

```
class Molecule:
    def __init__(self):
        self.xyz = {}
        self.group = {}
        self.coords = []
        self.bonds = set()

    def load_atoms(self, id):
        print "id =",id
        in_file = open(id,"r")
        lines = in_file.readlines()
        in_file.close()
        for line in lines:
            if line[:6] == "HETATM" or line[:4]=="ATOM":
                ser_num = int(line[6:11])
                coords = [0.0, 0.0, 0.0]
                coords[0] = float(line[30:38])
                coords[1] = float(line[38:46])
                coords[2] = float(line[46:54])
                self.xyz[ser_num] = coords
                self.group[ser_num] = line[76:78].strip()
                self.coords.append(coords)

            if line[:6] == "CONNECT":
                ser_nums = line[7:]
                tokens = ser_nums.split()
                tokens = map(int,tokens)
                for i in xrange(1,len(tokens)):
                    if self.xyz.has_key(tokens[0]) and self.xyz.has_key(tokens[i]):
                        bond = frozenset([tokens[0],tokens[i]])
                        self.bonds.add(bond)

        return
```

Figura 12- Método PDBParser usado para extrair as informações atômicas de um arquivo PDB. Construído em Python de forma a extrair as informações das moléculas que compõe o banco de dados DUD-E.
Fonte: Elaborada pelo autor.

3.2.1.2 Integração de Monte Carlo

Técnica matemática para integração numérica de funções sobre domínios complicados, D , no qual são escolhidos pontos aleatórios criando um superdomínio D' e verifica se cada ponto escolhido se encontra dentro de D e estima-se a área de D como sendo a área de D' multiplicada pela probabilidade de o ponto pertencer a D .³⁸

A estimativa de uma função é calculada a partir da média de um grande número de valores aleatórios aplicados à função, conforme se aumenta a quantidade de amostras o valor da estimativa se aproxima cada vez mais do valor esperado.

$$E[f(x)] \approx \frac{1}{N} \sum_{i=1}^N f(x_i) \quad (36)$$

A função da densidade de probabilidade $p(x)$ é responsável por determinar a probabilidade de uma amostra aleatória pertencer a um determinado intervalo. Para intervalos que entre $[-\infty, +\infty]$ a probabilidade é de 100% ou 1, para intervalos menores a probabilidade é definida por

$$P(x \in [a, b]) = \int_a^b p(x) dx \quad (37)$$

Combinando-se a estimativa da integral com a função densidade de probabilidade $p(x)$ a fim de obter uma estimativa do valor de uma integral, estimador de Monte Carlo, determina-se a seguinte relação.

$$\int f(x) dx = \int \frac{f(x)}{p(x)} p(x) dx \approx \frac{1}{N} \sum_{i=1}^N \frac{f(x)}{p(x)} \quad (38)$$

O estimador de Monte Carlo pode ser escrito de outra forma, na qual ao invés de dividirmos pela função densidade de probabilidade, multiplica-se pela função peso, definida por:

$$w(x) = 1/p(x) \quad (39)$$

Sendo assim, tem-se:

$$\int f(x) \approx \frac{1}{N} \sum_{i=1}^N f(x_i) w(x_i) \quad (40)$$

Caso $p(x)$ seja uma distribuição uniforme no espaço o que se faz é simplesmente calcular a somatória de todos os valores multiplicados pela distribuição e dividir pelo número de amostras. Para o caso do cálculo dos harmônicos esféricos deve-se mapear uma esfera em função do raio e dos ângulos θ e ϕ . Para simplificar os cálculos usa-se uma esfera unitária e o par de ângulos pode ser obtido da seguinte forma:

$$(2\arccos(\sqrt{1-x}), 2\pi y) \rightarrow (\theta, \phi) \quad (41)$$

Além disso, a probabilidade de um ponto se encontrar na superfície da esfera é igual para todos os pontos, ou seja, há uma distribuição uniforme das probabilidades, dada por

$$p(x) = 1/4\pi \quad (42)$$

Logo

$$w(x) = 4\pi \quad (43)$$

Para simular a superfície da esfera são calculados cem valores para x e para cada valor de x são calculados cem valores de y , gerando um total de 10 mil pontos. Nos quais os valores de a e b pertencem ao intervalo $[0,100]$.

$$x = (a + \text{valor aleatório}) * 1/\text{quantidade total de pontos} \quad (44)$$

$$y = (b + \text{valor aleatório}) * 1/\text{quantidade total de pontos} \quad (45)$$

3.2.1.3 Cálculo dos Harmônicos Esféricos

Para cada ponto determinado aleatoriamente é calculado os valores de θ e φ , após a determinação desses valores e o armazenamento nas variáveis adequadas é feito o cálculo dos harmônicos esféricos através de uma combinação de funções trigonométricas, $\sin m\varphi$ ou $\cos m\varphi$, e fórmulas de recorrência dos Polinômios Associados de Legendre, no qual são informados os valores de l, m e θ . Ambas as funções estão compreendidas entre o intervalo $[-1,1]$ e, além disso, os Polinômios Associados de Legendre retornam valores reais, ao invés de valores complexos, como os retornados pelos polinômios ordinários³⁴, sendo mais úteis para a análise de superfícies moleculares.

$$y_l^m(\theta, \varphi) = \begin{cases} \sqrt{2}K_l^m \cos(m\varphi) P_l^m(\cos \theta), & m > 0 \\ \sqrt{2}K_l^m \sin(-m\varphi) P_l^{-m}(\cos \theta), & m < 0 \\ K_l^0 P_l^0(\cos \theta), & m = 0 \end{cases} \quad (46)$$

As variáveis l e m são responsáveis por determinar as bandas dos harmônicos esféricos, nos quais $l \in [0, l_{max}]$ e $m \in [-l, l]$. O cálculo dos Polinômios de Legendre envolve o uso de variáveis complexas e cancelamentos de termos que podem tornar o processo custoso demais. A fim de evitar esse problema, Green sugere o uso de fórmulas de recorrência para a determinação dos valores, partindo de resultados anteriores descritas por apenas três relações.

$$P_m^m(x) = (-1)^m (2m-1)!! (1-x^2)^{m/2} \quad (47)$$

$$(l - m)P_l^m = x(2l - 1)P_{l-1}^m - (l + m - 1)P_{l-2}^m \quad (48)$$

$$P_{m+1}^m = x(2m + 1)P_m^m \quad (49)$$

Para $l = 0$ e $m = 0$, para esses valores determina-se que $P_0^0(x) = 1$. Para o cálculo de valores nos quais $l = m$ utilizamos a primeira equação e o sinal !! significa o duplo fatorial. Para o cálculo de bandas em que o valor de $l = m + 1$ utiliza-se a terceira equação, que parte do valor de P_m^m . Por fim, os valores de P_l^m são calculados a partir de valores obtidos para $l - 1$ e $l - 2$.

O método de cálculo inicia-se pela tentativa de determinar o maior valor de P_m^m através do uso da primeira regra, no qual o maior valor será quando $l = m$. Dessa forma, todos os valores que sobram para m são menores que o valor de l . O próximo passo é aumentar o valor de l uma vez, de forma que $l = m + 1$ e realizar o cálculo segundo a terceira fórmula e por fim iterar pela segunda regra até que o valor correto seja encontrado.

O último passo do cálculo dos harmônicos esféricos é a determinação da constante de normalização K_l^m .

$$K_l^m = \sqrt{\frac{(2l + 1)}{4\pi} \frac{(l - |m|)!}{(l + |m|)!}} \quad (50)$$

Todos os valores são calculados em uma rotina Fortran e os valores encontrados são armazenados para que sejam calculados os coeficientes da expansão da função na nova base. A fim de facilitar o armazenamento, pensa-se que os harmônicos esféricos ocorrem em uma determinada ordem e devido a esse fato, podem ser armazenado em um vetor unidimensional, de forma que:

$$y_l^m(\theta, \varphi) = y_i(\theta, \varphi) \text{ no qual } i = l(l + 1) + m \quad (51)$$

O método proposto o Green obtém bons resultados para o cálculo dos harmônicos esférico, porém apresenta problemas de estabilidade devido a perda de acurácia para grandes valores de l . Dessa forma, uma abordagem complementar foi utilizada a fim de garantir a estabilidade do cálculo e sua descrição detalhada pode ser encontrada no apêndice A.

3.2.1.4 Expansão em Harmônicos Esféricos

O primeiro passo para a determinação dos coeficientes da expansão, que serão os descritores da superfície molecular, é o cálculo dos harmônicos esféricos. A segunda parte de processo é calcular os pontos que serão utilizados durante a expansão, sua definição é fruto do

mapeamento da superfície da molécula, nos quais os pontos determinados aleatoriamente e as coordenadas atômicas e os nomes de cada átomo é passado para a função surface responsável por alterar o centro geométrico da molécula e considerando envelope Van der Waals.

A superfície molecular é uma representação abstrata ou modelo de uma molécula no qual a superfície molecular a partir de uma estimativa matemática definida pelos raios atômicos de Van der Waals. O cálculo do envelope de Vander Waals pode ser definido utilizando cálculos computacionais considerando a interações interatômicas e os raios atômicos individuais. A partir da superfície de Van der Waals é possível obter outras duas superfícies através de informações obtidas rolando uma esfera de prova, geralmente água de raio $1.4\text{\AA}$, superfície acessível para o solvente definida pela nova superfície traçada pelo centro da esfera de prova sobre ao rolá-la sobre a superfície da molécula e a superfície não acessível para o solvente, definida pela região na qual a esfera de prova não consegue acessar devido ser um côncavo ou ponto de encontro entre a superfície de dois átomos, conforme na Figura 13.

Figura 13- Representação da superfície de Van der Waals e das superfícies acessíveis ao solvente e não acessíveis determinadas pela esfera de prova.

Fonte: RITCHIE³⁹

Com a superfície mapeada pelo envelope de Van der Waals o próximo passo é o cálculo dos coeficientes definidos segundo as Séries de Laplace. Cada ponto mapeado é projetado na base de harmônicos esféricos para as diferentes bandas formadas por combinações dos números quânticos l e m e somado. Ao final da soma, o coeficiente obtido é multiplicado pelo elemento de sólido, definido por 4π , e dividido pela quantidade de pontos utilizados na soma.

$$c_{lm} = \int_0^{2\pi} \int_0^{\pi} f(\theta, \varphi) Y_{lm}^*(\theta, \varphi) \sin\theta d\theta d\varphi = \frac{4\pi}{N} \sum_{l=0}^{\infty} \sum_{m=-l}^l f(\theta, \varphi) Y_l^m(\theta, \varphi) \quad (52)$$

Para cada banda dos harmônicos esféricos tem-se $2 * l + 1$ coeficientes, devido ao fato de $m \in [-l, +l]$. Com isso o número total de coeficientes é $(l_{max} + 1)^2$. Conforme já descrito anteriormente, os coeficientes podem ser armazenados em um vetor unidimensional. Para determinação dos descritores rotacionalmente invariantes (Figura 14) é feita a soma de cada coeficiente presente na banda elevado ao quadrado e extraída a raiz quadrada da soma obtida, dessa forma a quantidade de coeficientes torna-se $l_{max} + 1$. O programa Lira utiliza o $l_{max} = 15$ como tentativa de minimizar ao máximo os erros durante a reconstrução da superfície original.

RIDs				
[14.11964317	3.14644836	7.64602445	2.98995343	2.58362594
2.0586789	1.66175375	1.20229668	0.78043585	0.5290073
0.6594225	0.60683006	0.46869021	0.53090028	0.44694869
0.48603064	0.39409871]			

Figura 14- Descritores rotacionalmente invariantes calculados pelo Lira para $l_{max} = 16$.

Fonte: Elaborada pelo autor.

A reconstrução da superfície da molécula é feita pela soma dos coeficientes obtidos multiplicados pelas bases dos harmônicos esféricos. O programa Lira utiliza a biblioteca Mayavi para a visualização da superfície reconstruída, conforme Figura 15.

Figura 15- Superfície reconstruída usando o método baseado em Harmônicos Esféricos e a biblioteca de visualização 3D Mayavi.

Fonte: Elaborada pelo autor.

3.2.2 Medidas de Similaridade e Dissimilaridade

Determinar as propriedades que tornam duas moléculas semelhantes é fundamental para a análise de similaridade, entretanto medir o quão similar ou dissimilar essas informações são, é essencial para um resultado robusto e eficiente. Essa quantificação é feita através de medidas de similaridade. Cada medida de similaridade apresenta particularidades, pode ser descrita de diferentes formas e analisar diferentes tipos de informação. Dessa forma, utilizar a medida mais adequada, aquela que melhor se avalia o conjunto de dados analisado é uma tarefa tão árdua quanto definir quais informações moleculares utilizar. Diversas medidas de similaridade foram desenvolvidas ao longo dos anos e nesse tópico falaremos sobre as mais utilizadas nos estudos de similaridade entre ligantes.

3.2.2.1 Definição de dissimilaridade

Medidas de dissimilaridade são responsáveis por determinar o quão diferentes duas moléculas ou conjunto de dados são.⁴⁰ Podem ser definidas matematicamente da seguinte forma.

A medida de dissimilaridade (DM) de d é uma função de X definida por:

$d: X \times X \rightarrow R$ onde R é um conjunto de números reais que satisfaça as condições abaixo:

$$I: \exists d_0 \in R : -\infty \leq d_0 \leq d(x, y) \leq +\infty, \forall x, y \in X$$

$$II: d(x, x) = d_0, \forall x \in X$$

$$III: d(x, y) = d(y, x), \forall x, y \in X$$

$$IV: d(x, y) = d_0, \text{ se e somente se } x = y$$

$$V: d(x, z) \leq d(x, y) + d(y, z), \forall x, y, z \in X$$

Exemplos de medidas de dissimilaridade são a Métrica L_p ponderada, Métrica L_2 ponderada (Mahalanobis), Distancia de Manhattan, Distância de Hamming e Distância Euclidiana.

3.2.2.2 Definição de Similaridade

Medidas de similaridade são responsáveis por determinar o quão similar duas moléculas ou conjunto de dados são.⁴⁰ Podem ser definidas matematicamente da seguinte forma:

A medida de similaridade (SM) de s é uma função de X definida por:

Seja $s: X \times X \rightarrow R$ onde R é um conjunto de números reais que satisfaça as condições abaixo:

I: $\exists s_0 \in R : -\infty \leq s(x, y) \leq s_0 \leq +\infty, \forall x, y \in X$

II: $s(x, x) = s_0, \forall x \in X$

III: $s(x, y) = s(y, x), \forall x, y \in X$

IV: $s(x, y) = s_0$, se e somente se $x = y$

V: $s(x, y)s(y, z) \leq [s(x, y) + s(y, z)]s(x, z), \forall x, y, z \in X$

Exemplos de medidas de similaridade são Medida de Tanimoto, Produto Interno e Medida de Similaridade Cosseno.

3.2.2.3 Métrica L_p Ponderada

Medida de dissimilaridade usada principalmente para vetores reais, responsável por determinar as diferenças absolutas entre os elementos dos vetores atribuindo um peso para cada posição do vetor.³⁰⁻³¹ Pode ser definida matematicamente por:

Sejam $p = (p_1, p_2, p_3, \dots, p_n)$ e $q = (q_1, q_2, q_3, \dots, q_n)$ dois vetores reais pertencentes a um espaço n-dimensional e um peso $w = (w_1, w_2, w_3, \dots, w_n)$ a distância entre os vetores é definida por:

$$d_p(p, q) = \left(\sum_{i=1}^n w_i |p_i - q_i|^p \right)^{1/p} \quad (53)$$

Nos casos em que $w_i = 1$ para todas as coordenadas tem-se uma métrica não ponderada. Para métricas não ponderadas, para os valores de $p = 1$ e $p = 2$ temos as distâncias de Manhattan e Euclidiana, respectivamente.

3.2.2.4 Distância Euclidiana

Medida responsável por determinar a distância entre dois pontos no espaço Euclidiano, ou seja, o comprimento do segmento de linha que liga os dois pontos. Também conhecida como métrica pitagórica, devido ao fato de ser semelhante a fórmula de Pitágoras, ou generalizada por distância L^2 ou norma L^2 .⁴⁰⁻⁴¹ Trata-se de uma medida de dissimilaridade, conforme descritos pelas definições anteriores, apresenta limite inferior igual a 0 quando o mesmo ponto é analisado, indicando máxima similaridade e não há limite superior, indicando que quanto maior for a distância mais dissimilares os conjuntos de pontos são. É definida matematicamente por:

Sejam dois conjuntos de pontos pertencentes a um n-espaço Euclidiano definidos por $p = (p_1, p_2, p_3, \dots, p_n)$ e $q = (q_1, q_2, q_3, \dots, q_n)$ a distância entre os pontos é dada por:

$$d(x, y) = d(y, x) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (54)$$

3.2.2.5 Distância de Manhattan

Medida de dissimilaridade responsável por determinar a soma dos valores absolutos entre dois pontos no espaço vetorial em relação a suas coordenadas cartesianas. Também conhecida por métrica dos taxistas, pois representam o caminho que um taxi deve percorrer pelas ruas, que formam uma grade, de Manhattan, "city block distance" ou métrica L_1 .⁴⁰⁻⁴¹ Pode ser definida matematicamente por:

Sejam dois vetores reais pertencentes a um n-espaço vetorial definidos com coordenadas cartesianas fixas definidos por $p = (p_1, p_2, p_3, \dots, p_n)$ e $q = (q_1, q_2, q_3, \dots, q_n)$ a distância entre os dois pontos é dado pela soma de todas as projeções dos segmentos de linha entre os dois pontos nos eixos coordenados, conforme a fórmula:

$$d(p, q) = d(q, p) = \sum_{i=1}^n |p_i - q_i| \quad (55)$$

3.2.3 Implementação Ultrafast Shape Recognition (USR)

Técnica desenvolvida por Pedro Ballester é caracterizada por utilizar as coordenadas atômicas das moléculas para determinação de descritores de forma e ter alta velocidade de operação.⁴² Apresentam grande eficiência e bons resultados quando comparadas às metodologias tradicionais encontradas no mercado, como EasyShape 3D, ROCS e OpenEye, em estudos de similaridade de moléculas e virtual screening.

A superfície de uma molécula pode ser determinada unicamente pelas posições relativas dos átomos. Dessa forma, o arranjo espacial tridimensional fornece um ótimo conjunto de dados para a medição de distâncias interatômicas em relação a pontos de referências estrategicamente definidos. O uso das coordenadas atômicas elimina a necessidade da superposição das moléculas, pois as posições relativas dos átomos são invariantes, ou seja, independente da orientação espacial as distâncias interatômicas serão as mesmas, sendo assim uma métrica

rotacionalmente invariante como o Lira. A descrição da superfície se torna completa e robusta pela descrição dos três momentos estatísticos das quatro distâncias interatômicas utilizadas, resultando em doze descritores moleculares, e dessa forma elimina o problema de comparação de moléculas com quantidade de átomos diferentes.

A ideia chave do USR é a introdução de um conjunto de dados representativo para o centro geométrico e os extremos da molécula capazes de caracterizar a superfície molecular (Figura 16). Moléculas similares apresentam geometrias similares e, portanto, devem apresentar pontos de referência em pontos de localização similar nas moléculas, por exemplo, o átomo mais afastado do centro geométrico da molécula deve ter localização similar em todas as moléculas similares. Conforme aumenta a quantidade de pontos de referência dissimilares entre duas moléculas, aumenta a dissimilaridade das superfícies delas tornando-se cada vez mais distintas.

Figura 16- Representação das distâncias utilizadas para descrever a superfície molecular a partir dos quatro momentos atômicos usados pelo USR.

Fonte: BALLESTER⁴²

3.2.3.1 Descritores moleculares usados pelo USR

Devido à grande quantidade de dados fornecidos pelas coordenadas atômicas, mais que suficiente para a descrição da superfície molecular, pode-se reduzir a quantidade de pontos utilizados sem perda de informação. Para isso, o USR mede as distâncias em relação a quatro pontos de referência: o centro da molécula (ctd), o átomo mais próximo ao centro da molécula (cst), o átomo mais longe do centro da molécula (fct) e o átomo mais longe do átomo mais longe do centro da molécula (fft). Dessa forma, a superfície molecular é descrita por quatro distribuições de distâncias interatômicas.

No primeiro passo na determinação dos descritores calcula-se o centro geométrico da molécula a partir das coordenadas atômicas da molécula. O segundo passo é determinação de um conjunto das distâncias euclidianas de todos os átomos em relação ao centro geométrico da

molécula $\{d_j^{ctd}\}_{i=1}^n$, no qual n é o número de átomos da molécula. Em seguida calcula-se os momentos dessa distribuição, o primeiro momento em relação ao centro (m_1) apresenta informações sobre a primeira potência e é definido por:

$$m_l \equiv E[d^l] \equiv \sum_{j=1}^N w_j d_j^l = \frac{1}{N} \sum_{j=1}^N d_j^l \quad (56)$$

Considera-se que cada átomo contribui igualmente para a determinação da superfície molecular, dessa forma a função de ponderação é definida por

$$w_j = 1/N \quad (57)$$

Para a obtenção dos descritores moleculares com interpretação geométrica em relação à média utiliza-se a seguinte definição:

$$u_l \equiv E[(d - m_l)^l] = \sum_{n=0}^l \binom{l}{n} (-1)^n m_1^n m_{l-n} \quad (58)$$

Os três primeiros momentos da distribuição são:

$$\mu_1 = 0 \quad (59)$$

$$\mu_2 = m_2 - m_1^2 \quad (60)$$

$$\mu_3 = m_3 - 3m_1 m_2 - 2m_1^3 \quad (61)$$

O primeiro momento não carrega nenhuma informação por definição e por isso o primeiro momento é utilizado como uma indicação do tamanho da molécula. Para a determinação dos outros nove descritores, o processo é repetido para os conjuntos de dados das distâncias: $\{d_j^{cst}\}_{j=1}^n$, $\{d_j^{fct}\}_{j=1}^n$ e $\{d_j^{fft}\}_{j=1}^n$. Outros pontos de referência podem ser adicionados para criar mais descritores e uma descrição de superfície mais precisa. Em nossos experimentos foram utilizados os três primeiros momentos pelo fato de considerarmos que contemplam informações suficientes para a descrição e para a efetividade do processo. Dessa forma, uma molécula pode ser descrita por um vetor de 12 valores reais (Figura 17). Raízes são utilizadas para que cada descritor tenha dimensão de distância.

$$\vec{M} = (m_1^{ctd}, (\mu_1^{ctd})^{\frac{1}{2}}, (\mu_2^{ctd})^{\frac{1}{3}}, m_1^{cst}, (\mu_1^{cst})^{\frac{1}{2}}, (\mu_2^{cst})^{\frac{1}{3}}, m_1^{fct}, (\mu_1^{fct})^{\frac{1}{2}}, (\mu_2^{fct})^{\frac{1}{3}}, m_1^{fft}, (\mu_1^{fft})^{\frac{1}{2}},) \quad (62)$$

Figura 17- Representação dos momentos descritores de uma molécula descrita pelo USR. Os quatro momentos utilizados na descrição são: o centro da molécula, o átomo mais próximo do centro da molécula, o átomo mais longe do centro da molécula e o átomo mais longe do átomo mais longe do centro da molécula.

Fonte: BALLESTER⁴²

3.2.3.2 Medida de Similaridade

Com os momentos de duas moléculas definidos é necessário determinar a similaridade entre ambas. O primeiro passo é o cálculo da distância de Manhattan dos vetores, em seguida o valor obtido é dividido pela quantidade de descritores utilizados. A medida de dissimilaridade obtida é acrescentada de um e invertida, resultando em uma função de pontuação normalizada.

$$S_{qi} = (1 + \frac{1}{12} \sum_{l=1}^{12} |M_l^q - M_l^i|)^{-1} \quad (63)$$

Os valores das medidas pertencem ao intervalo [0,1], no qual 0 representa superfícies dissimilares e 1 superfícies similares. A Figura 18 demonstra os coeficientes moleculares e o valor da medida de similaridade determinada pelo USR.

Figura 18- Momento atômicos calculados para duas moléculas pelo método USR. No centro encontra-se a métrica própria utilizada pelo USR para classificar as moléculas similares.

Fonte: BALLESTER⁴²

3.2.3.4 Tempo de execução

O tempo de execução leva em consideração duas medidas principais, o tempo da leitura da base de dados de descritores moleculares e o carregamento em memória e o tempo de realizar as comparações entre a molécula alvo e todo o banco de dados. Para melhorar a performance do USR, foi executado um pré-processamento e a criação de um banco de dados de moléculas contendo os doze descritores para cada uma. Em um experimento no qual 100 moléculas foram analisadas frente a um banco de dados de 690 milhões de conformações o processo levou cerca de 83 minutos para ser executado completamente, o que corresponde a cerca de 13,7 milhões de moléculas por segundo⁴² considerando o tempo de carregamento do banco de dados e a varredura das moléculas. O tempo total é totalmente dependente do tamanho do banco de dados a ser varrido e da quantidade de moléculas a serem analisadas podendo ser descrito pela equação:

$$t_{USR}(q, n) = nt_{leitura} + qnt_{varredura} \quad (64)$$

Onde q é a quantidade de moléculas a serem analisadas, n é o tamanho do banco de dados utilizado e $t_{leitura}$ e $t_{varredura}$, são os tempos de leitura e varredura, respectivamente. Em casos, no qual somente uma molécula é analisada o tempo de leitura domina a equação, mostrando que o processo de carregamento do banco de dados em memória é mais oneroso. O processo foi executado em um processador dual core. Em comparação com métodos

tradicionais, o USR se mostra mais eficiente e rápido na determinação da similaridade e na varredura de bancos de dados de grande extensão (Figura 19). Devido aos ótimos resultados obtidos pelo método, novos experimentos têm sido propostos, como o uso de características farmacofóricas⁴³, em busca de conciliar a velocidade do USR com outras propriedades importantes no processo de desenho racional de fármacos e similaridade de moléculas.

Figura 19- Comparação entre a quantidade de moléculas analisadas por segundo pelos métodos Shape Signature, USR e ROC, mostrando que o USR consegue varrer uma quantidade maior de arquivos no mesmo período de tempo quando comparado aos métodos comerciais.

Fonte: BALLESTER⁴²

3.2.4 Benchmarking

Para validar o método Lira, desenvolvido durante o projeto, foi criado um benchmarking baseando-se no processo descrito por Ballester^{42, 44} e alterado de acordo com os propósitos desse trabalho. Sendo assim, foram executados os seguintes passos:

- Montagem de banco de dados com descritores pré-calculados;
- Determinação das moléculas a serem utilizadas no experimento;
- Determinação do algoritmo para seleção de ligantes similares;
- Validação utilizando classificação binária e curvas ROC;
- Redução da dimensionalidade do espaço vetorial a partir de análise de componentes principais (PCA);
- Utilização de funções de ponderação no algoritmo de seleção de ligantes similares;

3.2.4.1 Montagem do Banco de Dados

A fim de executar os experimentos de validação do LIRA, verificar se realmente representa uma métrica capaz de determinar moléculas similares, foi montado um banco de dados conforme descrito por Ballester.^{42, 44} No qual os descritores moleculares são pré-calculados e armazenados para uso futuro e dessa forma melhorar a performance do experimento e da busca dos ligantes similares.

O primeiro passo para a criação do banco de dados é a escolha do conjunto de moléculas a serem utilizadas. O banco de dados escolhido para o LIRA foi a Database of Useful Decoys: Enhanced, DUD-E. A escolha se deve ao fato do banco de dados DUD derivado do ZINC e ter sido desenhado para ser usado em benchmarkings de docagem molecular e desenho racional de fármacos fornecendo compostos desafiadores⁴³. Todas as moléculas presentes no DUD-E podem ser adquiridas comercialmente e são agrupadas de acordo com os alvos moleculares, sendo que existem 50 compostos para cada composto ativo e apresentando propriedades físico-químicas semelhantes, porém com diferente topologia 2D. A DUD-E é fornecida pela Universidade da Califórnia, apresenta 102 alvos moleculares dos quais, 38 são provenientes dos 40 originais da DUD. Suas moléculas estão armazenadas em arquivos com diferentes extensões como. SDF e MOL2 e o download pode ser feito de acordo com o alvo a ser estudado ou do banco de dados completo, no LIRA adotamos a segunda opção. As moléculas baixadas da DUD-E apresentam formato MOL2, sendo necessária a conversão do formato para PDB para serem utilizadas tanto para o Lira quanto para o USR. Esse procedimento é realizado pela biblioteca da Open Babel utilizando o método de conversão adequado⁴⁵ e é aplicada para todos os 102 alvos, gerando arquivos individuais para cada molécula de cada alvo e são nomeadas de forma sequencial, como 9aa2ar no qual 9 significa ser a nona molécula do alvo aa2ar, resultando num total de 40490 moléculas.

O segundo passo é a escolha do banco de dados a ser utilizado e tanto para o Lira quanto para o USR será usado o SQLite pois é um banco de dados com boa performance, que atende os requisitos necessários para nossa aplicação⁴⁶ e por estar conectado ao Python. Para cada banco de dados é criada uma tabela contendo dois campos de texto, o nome da molécula e os descritores moleculares.⁴⁷⁻⁴⁸

Após a definição do banco de dados de moléculas e do tipo de banco de dados a ser utilizado é executado um script responsável por varrer todas as moléculas de todos os alvos, calcular os descritores moleculares do Lira para cada molécula e insere seu nome sequencial e

coeficientes no banco de dados. Para a determinação dos descritores do Lira é utilizado um valor de $l_{m\acute{a}x} = 15$, resultando em 16 coeficientes para a métrica L^2 e 256 coeficientes para métrica L^1 , cada um armazenado em um banco separado. De forma semelhante é gerado um banco de dados para o USR, no qual são calculados 12 descritores moleculares para cada molécula.

Os tempos para montagem do banco de dados é praticamente o mesmo para todas as métricas, entretanto o tamanho de cada banco é dependente da quantidade de descritores que serão armazenados por molécula, quanto mais descritores mais informação a ser armazenada o que resulta em um banco de dados maior. Dessa forma, o banco criado para métrica L^1 apresenta o maior tamanho, pois armazena 256 coeficientes, enquanto que a métrica L^2 e USR tem praticamente o mesmo tamanho armazenando 16 e 12 coeficientes para cada molécula, respectivamente. (Tabela 1)

Tabela 1- Representação dos tamanhos dos bancos de dados gerados para cada métrica

Banco de Dados	Tamanho (MB)
Lira L^2 (LIRA.db)	8,1
Lira L^1 (COEFFS.db)	163,89
USR (USR.db)	8,0

Fonte: Elaborada pelo autor

3.2.4.2 Determinação das moléculas do benchmarking

Para o estudo de benchmarking foram selecionadas 50 moléculas da DUD-E de forma a tentar representar o banco por completo, sendo uma amostra bem com moléculas variadas. Dessa forma, foram escolhidas moléculas de quantidade e composição atômica diferentes e propriedades físico química diversas.

Foi criado um script para seleção das moléculas do estudo, na primeira parte realizou-se a leitura dos arquivos PDB de todas as moléculas e nas linhas que continham a string "HETAM" ou "ATOM" verificava-se o átomo que se encontrava na posição especificada e somava-se um a um contador de átomos. Para cada átomo encontrado, era verificado se o mesmo já não estava armazenado na string responsável por determinar os diferentes elementos que compunham a molécula, caso não estivesse o mesmo era adicionado a string caso contrário era descartado e seu valor era somente incrementado no contador de átomos. Ao final da leitura

de cada arquivo avalia-se a quantidade de átomos presente na molécula e classificava-a de acordo com a Tabela 2.

Tabela 2 - Intervalos utilizados para classificar as moléculas de acordo com a quantidade de átomos

Tipos	Intervalo
Top 100	≥ 100
Top 90	$90 \leq x < 100$
Top 80	$80 \leq x < 90$
Top 70	$70 \leq x < 80$
Top 60	$60 \leq x < 70$
Top 50	$50 \leq x < 60$
Top 40	$40 \leq x < 50$
Top 30	$30 \leq x < 40$
Top 20	$20 \leq x < 30$
Top 10	$10 \leq x < 20$
Top 1	$1 \leq x < 10$

Fonte: Elaborada pelo autor

Em seguida, foi realizada a análise da composição atômica das moléculas de acordo com as strings criadas para cada uma, todas continham os elementos carbono (C) e hidrogênio (H) e outros elementos comuns eram oxigênio (O) e nitrogênio (N). Dessa forma, buscou-se átomos diferentes desses sendo encontrados: Flúor (F), Fósforo (P), Boro (B), Iodo (I), Enxofre (S) e Lítio (Li). Para buscar mais diversidade nas moléculas foram criadas combinações de 3 (Tabela 3) e 2 (Tabela 4) elementos, combinações com maiores números não foram encontradas. Para cada quantidade de moléculas classificadas acima foi realizada a varredura em busca das combinações dos elementos desejados e para cada classificação foi criado um arquivo contendo a molécula que continha a estrutura selecionada e quantidade de átomos que ela apresentava.

Tabela 3 - Combinações de 3 átomos utilizados para a seleção das moléculas a serem utilizados no benchmarking

Combinação de 3 elementos			
LiSP	LiFI	BIS	FPI
LiSF	FSP	LiPF	SPB
LiSB	FSB	LiFB	FIB
LiSI	FPB	LiBI	BIP
LiPI	SPI	FSI	LiPB

Fonte: Elaborada pelo autor

Tabela 4- Combinações de 2 átomos utilizados para a seleção das moléculas a serem utilizados no benchmarking

Combinação de 2 elementos		
LiB	BI	IP
LiI	BF	IS
LiF	BS	FS
LiS	BP	FP
LP	IF	SP

Fonte: Elaborada pelo autor

Ao final de ambas análises foram encontradas cerca de 500 moléculas das quais as 50 (Tabela 5) foram selecionadas pelo fato de acreditarmos que compunham o conjunto de moléculas que melhor representava todo o banco de dados.

Tabela 5- 50 moléculas selecionadas para análise de benchmarking tentando representar a diversidade do banco de dados DUD-E.

Moléculas selecionadas para Benchmarking				
313aces	45esr1	34tysy	98hxx4	93hivrt
10aces	129reni	342vgfr2	4andr	13fpps
44ptn1	261hmdh	724mmp13	474ada17	252urok
314aces	52csf1r	305urok	229mp2k1	35kith
1266hivpr	407aa2ar	129ptn1	174gria2	539esr2
237reni	123src	106hs90a	172mp2k1	640lck
74bace1	347casp3	566drd3	112pgh2	120xiap
7def	416egfr	6ital	314src	166try1
536aces	599fnta	208ptn1	781cah2	488hivpr
332cdk2	811hivpr	325vgfr2	785dpp4	618cah2

Fonte: Elaborada pelo autor

3.2.4.3 Determinação dos ligantes similares

A determinação dos compostos mais semelhantes é dividida em três passos: cálculo dos descritores de forma da molécula escolhida, leitura do banco de dados para a memória do computador e comparação dos descritores moleculares da molécula utilizada com os dados extraídos do banco de dados.⁴² Todo esse processo é realizado por um script que executado de forma similar para as três métricas analisadas, L^2 , L^1 e USR .

No primeiro passo é feita a leitura do arquivo PDB e os descritores da forma da molécula são calculados de acordo com a métrica a ser utilizada e são armazenados em um vetor unidimensional. No segundo passo todas as moléculas são carregadas em memória, abordagem utilizada para evitar o tráfego excessivo de informações entre a aplicação e o banco de dados que onera o processo, através da execução da query "Select * from [Tabela]" não são aplicados filtros com a finalidade de trazer os dados da tabela em sua totalidade. Os coeficientes e o nome de cada molécula são armazenados em uma linha de uma matriz de dimensão 40490 x 2. Como os coeficientes de cada molécula foram armazenados na forma de uma string é necessário isolar cada valor e convertê-los para valores reais para que seja feita a comparação dos descritores. No passo final é realizada a comparação dos descritores calculados para a molécula escolhida e para cada molécula extraída do banco de dados, ou seja, são realizadas 40490 comparações, a pontuação e o nome da molécula testada são armazenados em um dicionário, estrutura de dados presente em diversas linguagens de programação no qual são armazenadas chaves e valores e cada valor está associado a uma única chave, tendo o nome da molécula como chave e a pontuação como valor. Após o processo ser executado para todas as moléculas o dicionário é ordenado e os nomes das moléculas mais similares são informados ao operador. A Figura 20 representa o resultado encontrado para a molécula 35kith utilizando a métrica L^2 e o USR .

Figura 20- Comparação das moléculas selecionadas para a molécula query 35kith pelo Lira e USR usando como valores de corte 0.5 e 0.95, respectivamente. As moléculas A, B e C apresentam a basicamente a mesma forma tendo apenas a mudança de alguns átomos. A segunda e terceira linhas mostram a superfície atômica de moléculas selecionadas pelo Lira e sua representação 3D. Em ambos os casos as moléculas 12kith, 14kith e 33kith foram selecionadas e revelam que o Lira aparenta ter mais flexibilidade conformacional.

Fonte: Elaborada pelo autor.

O tempo de execução é dependente da quantidade de moléculas encontradas no banco de dados e da quantidade de descritores que cada molécula apresenta. Para bancos muito grandes a leitura de todos os compostos de uma única vez pode representar um problema demandando grande quantidade de memória, no qual o particionamento do banco de dados pode ser utilizado como solução e o resultado de cada experimento é analisado e tendo como resultado final a soma das informações de todas as análises.⁴² No caso do LIRA, tanto para métrica L^1 quanto para métrica L^2 e do USR, não é necessário executar o particionamento do banco de dados devido a quantidade de 40490 compostos. Para o processo apresentado o tempo total gasto na análise de uma única molécula é dado pela soma do tempo de leitura e cálculo da molécula escolhida, o de leitura dos coeficientes das moléculas presentes no banco de dados e o tempo da varredura (*screening*) no qual é feita a comparação dos descritores.

$$t_{total} = t_{calculo} + t_{leituraBD} + t_{varredura} \quad (65)$$

Onde

$$t_{leituraBD} = n * t_{molécula} \quad (66)$$

$$t_{varredura} = n * t_{individual} \quad (67)$$

sendo n a quantidade de moléculas armazenadas no banco de dados, $t_{molécula}$ o tempo necessário para ler os coeficientes de uma molécula e $t_{individual}$ tempo de comparação dos descritores da molécula inicial com a presente no banco de dados.

Tabela 6- Tempo gasto para leitura, cálculo, varredura do banco de dados e tempo total gasto para cada métrica utilizada nos experimentos de benchmarking

Banco de Dados	Tempo de Leitura (s)	Tempo de Cálculo (s)	Tempo de Varredura (s)	Tempo Total (s)
LIRA L^2	0.776181171	0.488794340	0.200461483	1.465436994
LIRA L^1	18.779040623	0.488770221	2.355925202	21.623736046
USR	0.734834063	0.000883206	0.184565067	0.920282336

Fonte: Elaborada pelo autor

Devido a quantidade de coeficientes semelhantes entre as métricas L^2 e USR observa-se tempos de leitura e varredura do banco de dados muito próximo, evidenciando que a busca de ligantes semelhante através da análise dos coeficientes rotacionalmente invariantes do LIRA é tão rápida quanto a mesma análise realizada pelo USR, um método considerado extremamente rápido. (Tabela 6) Mesmo para a métrica L^1 , com 256 descritores ao invés de 12 ou 16, tem-se um resultado satisfatório no processo de busca de ligantes similares levando cerca de 2,4 segundos na varredura de uma base de dados de 40490.

3.2.4.4 Validação das moléculas selecionadas pelas métricas

Para validação, as moléculas encontradas pelo Lira foram comparadas com as moléculas selecionadas pelo USR, sempre o considerando como resultado ideal. Dessa forma, todas as 50 moléculas escolhidas para a realização do benchmarking foram submetidas ao processo de determinação de ligantes. Para determinar se uma molécula é semelhante a outra foram determinados valores de corte ("Cutoff"). Adotou-se a estratégia de variação do valor o incrementando a uma taxa constante para determinação dos ligantes similares. Sendo assim para a métrica L^2 do Lira os valores foram variados de 0 a 2 com incremento de 0.25 e no caso da métrica L^1 os valores foram variados de 7 a 10 variando-se de 0.5. Para ambos os casos foi utilizado a distância euclidiana para determinação da similaridade entre os ligantes, ou seja, quanto mais próximo de zero o valor da comparação, mais similares as moléculas são. Para o USR, fui utilizado o valor de corte de 0.75, o qual se mostrou capaz de trazer a quantidade adequada de ligante para a validação dos dados do Lira e pelo fato dos valores de comparação pertencerem ao intervalo[0,1] no qual 1 indica máxima similaridade. Todos os valores de corte e medidas de similaridade considerados, tanto para o Lira quanto para o USR, apresentam dimensão de distância e são comumente medidas em Å.

Nesse processo um novo script foi criado para cada molécula seguindo o seguinte roteiro:

1. Busca descritores da superfície para moléculas selecionada no banco de dados LIRA
2. Varredura do banco de moléculas do LIRA e a comparação entre os coeficientes armazenados em lista
3. Busca dos descritores da superfície para moléculas selecionada no USR
4. Varredura do banco USR e comparação molécula a molécula entre os coeficientes a armazenamento em lista
5. Variações dos valores de corte e armazenamento das moléculas selecionadas
6. Comparação dos resultados obtidos através de Classificação Binária
7. Criação de Curvas ROC (Receiver Operating Characteristic)

3.2.5 Classificação Binária

O sistema de classificação binária de informação pode ser utilizado em diferentes áreas de conhecimento como área da saúde na tomada de decisões, mineração de dados e aprendizado de máquina. Para realizar a classificação dos dados construiu-se um modelo binário representado pelos dados do USR, visto como dados corretos, e os dados do Lira, a hipótese a ser testada⁴⁹, nos quais os dados do Lira são comparados ao do USR a fim de verificar se uma determinada molécula selecionada pelo Lira também foi selecionada pelo USR. Dessa forma, há quatro situações possíveis para essa classificação: verdadeiro positivo, falso positivo, verdadeiro negativo e falso negativo.

Classifica-se uma molécula como verdadeiro positivo no caso em que a molécula selecionada pelo Lira é pertence ao conjunto de moléculas selecionadas pelo USR, ou seja, ambas as métricas classificaram a mesma molécula como similar. No caso de um falso positivo, a molécula analisada foi selecionada somente pelo Lira, ou seja, não se encontra no conjunto selecionado pelo USR.

Quando uma molécula é selecionada pelo USR e não se encontra no conjunto de moléculas selecionadas pelo Lira, ou seja, somente o USR classificou a molécula como similar e pelo fato de considerarmos o USR como a métrica correta, classifica-se essa situação como falso negativo. Por fim, tem-se o caso do verdadeiro negativo que é a situação no qual consideramos as moléculas que não foram selecionadas por nenhuma das duas métricas. As relações entre as classificações formam uma matriz de confusão ou matriz de classificação responsável por auxiliar a análise e a tomada de decisões. (Figura 21)

		LIRA	
		V	F
US R	V	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	F	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 21- Diagrama utilizado no processo de classificação binária usado no benchmarking. Determinação de propriedades estatísticas como acurácia, especificidade e taxas de verdadeiro e falso positivo para análise de curvas ROC.

Fonte: Elaborada pelo autor

A partir da matriz de confusão diversas relações podem ser derivadas como Sensibilidade ou taxa de verdadeiros positivos, Especificidade ou taxa de verdadeiros negativos, Acurácia, Precisão, Taxa de falsos negativos dentre outros.⁴⁹ Em nossos estudos foram utilizadas as medidas de Especificidade e Acurácia e as Taxa de Verdadeiro Positivo e Taxa de Falso Positivo. (Tabela 7)

3.2.6 Curvas ROC

Curvas ROC são gráficos bidimensionais no qual em que o eixo Y reflete a taxa de verdadeiro positivos e no eixo X a taxa de falso positivo, ambos os eixos estão definidos para o intervalo [0,1] e o gráfico demonstra a relação entre os benefícios (Verdadeiros Positivos) e os custos (Falsos Positivos). São utilizadas em diferentes campos de estudos, sendo importantes na área da saúde para tomada de decisões.

Tabela 7- Relações, fórmulas e descrição utilizadas no processo de classificação binária e geração de curvas ROC

Relação	Fórmula	Descrição
Taxa de Verdadeiro Positivo	$\frac{VP}{VP + FP}$	Determina taxa de moléculas que foram selecionadas por ambas as métricas
Especificidade	$\frac{VN}{FP + VN}$	Determina Taxa de moléculas que não foram selecionadas por nenhuma das duas métricas
Acurácia	$\frac{VP + VN}{VP + FP + VN + FN}$	Determina a taxa de moléculas acerto do LIRA em relação ao USR
Taxa de Falso Positivo	$\frac{FP}{FP + VN}$	Determina taxa de moléculas classificadas pelo LIRA e não classificadas e pelo USR

Fonte: Elaborada pelo autor

Cada classificador é determinado por um único ponto no gráfico gerado a partir da combinação das taxas de falso positivo, eixo das abscissas, e falso positivo, eixo das ordenadas, formando o par ordenado⁴⁹ (Taxa de Falso Positivo, Taxa de Verdadeiro Positivo). Entretanto, existem diversos pontos que devem ser considerados durante a análise e montagem das curvas, o ponto (0,0) representa o fato do classificador não apresentar falsos positivos e nem verdadeiros positivos, ou seja, não agrega valor à análise. Por sua vez, o ponto (1,1) representa a o classificador que para cada verdadeiro positivo há um falso positivo, ou seja, atua de forma aleatória. O ponto (0,1) representa a classificação perfeita, ou seja, todas as informações foram classificadas corretamente. Apesar se considerado como o resultado ideal, pode evidenciar um conjunto de dados viciados e não apresentar informações relevantes para o estudo. Por fim, o ponto (1,0) representa a pior classificação, no qual todas as informações foram classificadas erroneamente.

Figura 22- Representação de curva ROC. A linha tracejada descrita pela equação $y = x$ representa um classificador aleatório. Pontos acima da reta apresentam status de bom classificador, pontos abaixo apresentam status de mau classificador e pontos sobre a reta são tidos como classificadores aleatório.

Fonte: FAWCETT⁴⁹

O gráfico apresenta uma linha divisória determinada pela equação $y = x$ que representa a classificação aleatória, ou seja, classifica uma determinada amostra utilizando um classificador aleatório.⁴⁸ (Figura 22) Um bom classificador apresenta taxa de verdadeiro positivo alta e taxa de falso positivo baixa, sendo assim localizado acima do classificador aleatório, na imagem tem-se que A e B são classificadores melhores do que o aleatório e D representa o classificador perfeito. C se encontra sobre a linha do classificador aleatório indicando que não apresenta ganhos quando comparado a um ele e E está localizado abaixo da reta baixa performance frente a um classificador aleatório.

3.2.7.1 Análise de Componente Principal (PCA)

Tem sua origem não definida podendo ser atribuída a Pearson (1901) ou Cauchy (1829) ou Jordan (1874), porém sua inicialização moderna foi formalizada por Hotelling em 1933 que cunhou o termo componente principal e trata-se de uma técnica estatística amplamente utilizada por diversas áreas do conhecimento como reconhecimento de e compressão de imagens⁵⁰⁻⁵², neurociência e computação gráfica devido ao seu poder em extrair informações relevantes de um conjunto de dados através da identificação de padrões. Tem como objetivos extrair as informações mais importantes de um conjunto de dados, reduzir sua dimensão sem que haja perda de informação, simplificar o conjunto de dados e analisar a estrutura das observações e variáveis.⁴⁹⁻⁵² Para determinação das componentes principais é feito a partir de combinações lineares e transformações ortogonais a partir das variáveis originais. As componentes são feitas

de forma que a primeira componente apresente a maior variância possível. A segunda componente é calculada de forma a ter a maior inércia possível e ser ortogonal a primeira componente. Após a determinação das componentes principais é possível criar um novo conjunto de dados a partir da projeção geométrica das observações originais em relação às novas componentes.⁵²

Um experimento de PCA pode ser decomposto em 6 passos essenciais^{51, 53}

1. Seleção do conjunto de dados de tamanho D dimensional a ser analisado.
2. Determinar a média das variáveis do conjunto de dados e armazená-lo em um vetor.
3. Calcular a matriz de espalhamento ou matriz de confusão do conjunto de dados inteiro
4. Calcular os autovalores e autovetores.
5. Ordenar os autovalores de forma decrescente e selecionar os K autovetores com maiores autovalores associados para formação da nova base do espaço vetorial ($W = D \times k$)
6. Utilizar a matriz W em uma transformação para criação de um novo subespaço a partir do original.

Para o Lira a análise de PCA foi feita para ambas as métricas no qual o conjunto de dados analisado foram os conjuntos de descritores moleculares gerados a partir do banco DUD-E. Dessa forma, para a métrica L^2 o conjunto apresentava dimensões 40490 x 16 e para a métrica L^1 o conjunto era apresentava as dimensões 40490 x 256.

O segundo passo foi a determinação dos valores médios de cada componente que descrevem as moléculas. Para isso foi criado-se um vetor nulo de tamanho 16 e 256 para as métricas L^2 e L^1 , respectivamente. Em seguida realizou-se leitura de todas as moléculas para ambos os bancos na memória do computador e para cada molécula seus descritores foram somados nas posições adequadas dos vetores e ao final cada uma foi dividida pelo número de molécula, ou seja, 40490. Dessa forma, foi calculado o valor médio de cada componente através de uma média aritmética. (Figura 23) A biblioteca Numpy apresenta a função *mean*⁵⁴, responsável por calcular o valor médio de uma lista de dados.

$$m = \frac{1}{n} \sum_{k=1}^n x_n \quad (68)$$

Mean Vector				
[17.21942413	1.20288817	5.68271437	2.59833745	2.05891687
1.35594804	0.96925654	0.72941829	0.58879889	0.49506904
0.43693761	0.39425025	0.3642426	0.34320529	0.32966963
0.32270988]				

Figura 23- Valores médios de calculado para cada componente dos descritores moleculares utilizados pela métrica L^2 .

Fonte: Elaborada pelo autor.

O terceiro passo no algoritmo do PCA é a determinação da matriz de espalhamento ou matriz de covariância, são similares sendo a única diferença o uso de um fator no cálculo da segunda matriz. A matriz de espalhamento é definida segundo a equação

$$S = \sum_{k=1}^n (x_k - m)(x_k - m)^T \quad (69)$$

Onde m é o vetor médio, calculado no segundo passo.

Para o cálculo da matriz de covariância é preciso determinar o desvio padrão e variância, duas propriedades estatísticas amplamente utilizadas em análise de dados. (Tabela 8) O desvio padrão mede a dispersão estatística de um conjunto de dados, determina a variação das componentes do conjunto quando comparada ao valor esperado, valor médio.⁵⁰ Conjuntos de dados que apresentam baixo desvio padrão revelam que suas componentes giram em torno do valor médio, aqueles que apresentam altos valores tem dados com valores distantes do valor esperado.

A variância, também conhecida como desvio quadrático médio da média, é a medida da distância de um determinado valor de um conjunto em relação ao valor médio do conjunto. O cálculo de variância pode ser feito de formas distintas, para uma população, variância da população, e para uma amostra populacional, variância amostral, a única mudança entre os métodos é o fator de escala. Quando feito para a população geral usa-se o fator $\frac{1}{n}$, para o caso de uma amostra usa-se o $\frac{1}{n-1}$ para que não seja criado um estimador tendencioso.^{50, 53} Apresenta somente dois resultados possíveis, zero indicando que não há dispersão nos dados da amostra e valores positivos indicando que há dispersão. Quanto mais próximo de zero, menor a dispersão dos dados e conforme a variância aumenta há um aumento da dispersão dos dados.

A partir das informações da variância é possível determinar qual é a relação de interdependência de duas variáveis, verificar o quanto uma variável varia da média em relação a outra variável⁵⁰, essa medida é chamada de covariância ou variância conjunta, sendo calcula

para a combinação de duas variáveis. Caso o valor da obtido seja positivo, tem-se uma relação de aumento combinado, ou seja, ambas variáveis aumentam juntas. Por outro lado, caso o valor obtido seja negativo, tem-se que enquanto uma variável aumenta a segunda diminui, gerando um decréscimo. Caso o valor medida seja zero, indica que as variáveis são independentes, ou seja, não apresentam relação entre si. A medida de covariância é comutativa⁵⁰, ou seja,

$$\text{cov}(x, y) = \text{cov}(y, x) \quad (70)$$

devido a definição da medida envolver a operação de multiplicação que é comutativa.

Tabela 8- Medidas estatísticas utilizadas durante a análise de componente principal

Medidas Estatísticas	
Desvio Padrão	$s = \sqrt{\frac{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}{n}}$
Variância da População	$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2$
Variância Amostral	$s^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2$
Covariância	$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Fonte: Elaborada pelo autor

O experimento de PCA exige uma matriz de covariância definida pelo cálculo das medidas covariância das combinações das variáveis presentes no estudo. A matriz gerada para o estudo é sempre uma matriz quadrada e apresenta dimensões iguais a quantidade de variáveis do estudo, em um sistema contendo apenas 3 variáveis, x , y e z , será construída uma matriz 3 x 3.

$$C^{m \times m} = (c_{i,j}, c_{i,j} = \text{cov}(\text{Dim}_i, \text{Dim}_j)) \quad (71)$$

A diagonal principal da matriz é preenchida com a covariância entre a variável representada na linha e ela mesma. Além disso, o cálculo de grande parte das combinações é reaproveitado através da propriedade comutativa gerando assim uma matriz simétrica. Dentro da biblioteca NumPy há uma implementação para a criação de matrizes de covariância, `numpy.cov`.⁵⁵ No caso do Lira tanto as matrizes de covariância como as de espalhamento apresentaram dimensões 16 x 16 e 256 x 256, para as métricas L^2 e L^1 , respectivamente. Para a construção de nosso modelo usou-se a matriz de espalhamento.

$$C = \begin{pmatrix} cov(x, x) & cov(x, y) & cov(x, z) \\ cov(y, x) & cov(y, y) & cov(y, z) \\ cov(z, x) & cov(z, y) & cov(z, z) \end{pmatrix} \quad (72)$$

Após a definição das matrizes o próximo passo é o cálculo dos autovalores e autovetores da matriz para determinar quais são as componentes mais importantes para o estudo e o novo espaço vetorial que será usado para a projeção dos dados.

3.2.7.2 Autovalores e autovetores

Autovalores e autovetores são um conjunto de números escalares e vetores que resolvem uma equação matricial quando submetida a uma transformação linear.⁵⁶ (Figura 24) Apresentam grande importância em problema físicos e de engenharia, pois fazem parte do processo de diagonalização de matrizes. Para cada autovalor λ há um autovetor associado.^{50, 53} Seja a transformação linear $T: V \rightarrow W$

$$T(v) = \lambda v \quad (73)$$

no qual λ é o autovalor e $v \geq 0$, o autovetor associado a transformação linear. Devido ao fato de que toda transformação linear pode ser escrita como a multiplicação de duas matrizes pode-se substituir

$$T(v) = Av \quad (74)$$

Igualando-se 73 e 74, cria-se o sistema homogêneo,

$$Av = \lambda v \rightarrow (A - \lambda I)v = 0 \quad (75)$$

No qual a solução trivial é dada pelo vetor nulo ($v = 0$), A é uma matriz quadrada $n \times n$ e I é a matriz identidade. Para casos em que $v \neq 0$ há um conjunto de escalares que resolvem o sistema

$$A - \lambda I = 0 \quad (76)$$

A solução do sistema pode ser dada igualando-se o determinante da equação matricial a zero.

$$\det(A - \lambda I) = 0 \quad (77)$$

Como resultado da aplicação do determinante tem-se um polinômio de grau n em λ , conhecido como polinômio característico no qual suas raízes são os autovalores da matriz A . Aplicando-se os autovalores encontrados em 77 determina-se os autovetores associados.

Figura 24- Representação geométrica dos autovalores. U é autovetor de T pois $\exists \lambda \in \mathbf{R} / T(U) = \lambda U$. V não é autovetor de T pois $\nexists \lambda \in \mathbf{R} / T(v) = \lambda v$.

Fonte: CALLIOLI⁵⁷

Após a definição dos da matriz de espalhamento o cálculo dos autovalores e autovetores (Figura 25) é calculado pelo uso da função *linalg.eig* presente na biblioteca NumPy.^{53, 58}

- Caso $\lambda > 1$ diz-se a transformação linear dilata v
- Caso $0 < \lambda < 1$ diz-se a transformação linear contrai v
- Caso $\lambda = 0$ diz-se que a transformação linear anula v
- Caso $\lambda < 0$ diz-se que a transformação linear inverte o sentido de v

```
Eigenvalues
[ 1.75977536e+08  1.74703387e+08  1.74481011e+06  5.23320323e+05
 1.85890743e+05  1.31441717e+05  6.77001411e+04  3.65634970e+04
 1.90698074e+04  1.08179334e+04  6.70267547e+03  4.46120870e+03
 1.60247043e+03  1.89956845e+03  2.41806040e+03  3.20143581e+03]
```

Figura 25 - Autovalores calculados para métrica L^2 .

Fonte: Elaborada pelo autor.

3.2.7.3 Redução de Dimensionalidade

O uso da função *linalg.eig* resolve o sistema linear e os autovalores e autovetores resultantes são armazenados em duas listas possibilitando a redução da dimensionalidade. Os autovetores representam as bases que formarão o novo subespaço, porém todos apresentam o mesmo tamanho. Para decidir quais autovetores serão descartados para a redução, os autovetores são classificados em ordem decrescente de acordo com os autovalores correspondentes. Os autovetores que apresentem autovalores pequenos indicam que há pouca

variância em seus dados tornando-os não essenciais para a análise e sendo passíveis de serem desconsiderados.^{50, 53} A quantidade de componentes que serão usadas para a criação do novo subespaço é determinada pelo usuário, no caso do Lira foram selecionadas 6 e 90 componentes para as métricas L^2 e L^1 , respectivamente.

3.2.7.4 Criação do novo subespaço

Selecionados os autovetores que serão a base do subespaço vetorial é feita a combinação deles com o uso da função *hstack*⁵⁹, responsável por combinar os autovetores e gerar a matriz W , de dimensões 6 x 40490 para métrica L^2 e 90 x 40490 para métrica L^1 . (Figura 26)

```
Matrix W:
[[ 7.89339890e-01  1.50087574e-01  3.30025348e-01  2.05728610e-01
  1.83921416e-01  1.55559189e-01]
 [-5.27889656e-01  2.10968729e-01  6.99305267e-03  1.45405845e-01
  1.70789399e-01  2.02794739e-01]
 [ 2.66280649e-01 -1.71565268e-01 -8.81137025e-01  2.31555788e-01
 -1.16762616e-01 -8.60243074e-03]
 [-1.18524946e-01  4.17350565e-01  3.85966345e-02  8.13118183e-01
 -2.04737661e-01  5.40947040e-02]
 [ 3.69218030e-02 -1.16003124e-01  2.16473201e-01 -6.74782941e-02
 -9.19146100e-01  2.55464805e-02]
 [-9.88146648e-02 -7.47510282e-01  2.54958112e-01  4.61939098e-01
  1.49528769e-01 -3.30376250e-01]
 [-4.59312172e-02 -3.81543309e-01  1.46470271e-02  5.78627517e-02
  1.02772485e-02  8.99347831e-01]
 [-6.82977694e-03 -6.30841301e-02 -2.88161664e-02  5.81784414e-04
 -1.08408640e-01 -6.41904192e-02]
 [ 3.16215254e-03  5.60728956e-02  1.34860318e-02 -1.52827703e-02
  3.69685043e-02  9.16341268e-02]
 [-4.24155683e-03 -5.30199776e-02 -5.58332904e-03  1.34536658e-02
 -1.76124189e-02  1.07864920e-02]
 [ 2.24796380e-03  3.21578759e-02  1.99343122e-03  1.37446802e-04
  1.53827558e-02  4.79796522e-03]
 [ 1.63895583e-03  2.19116889e-02  1.69425693e-03 -2.55433514e-03
  1.93171902e-02  1.16615264e-03]
 [ 9.03012811e-04 -3.02940039e-04 -7.70491048e-04 -2.22206233e-03
 -3.98269345e-03 -4.31883134e-03]
 [ 6.55825741e-04 -1.36757513e-03 -1.14147237e-03 -2.63759253e-03
 -3.47592262e-03 -3.63029632e-03]
 [-7.65474872e-04  1.48715442e-04  2.52016694e-03  4.93433827e-03
  4.79072586e-03  9.83997740e-03]
 [-2.47823820e-04 -9.74714469e-03 -2.20861344e-03 -2.60398859e-03
 -8.29872067e-03 -1.61608751e-02]]
```

Figura 26- Matriz gerada pelos autovetores selecionados na análise de componentes principais para criação do novo subespaço vetorial.

Fonte: Elaborada pelo autor.

3.2.7.5 Projeção dos dados no novo subespaço

O passo final é projetar os dados originais no novo subespaço e dar origem ao novo conjunto de dados. A projeção é realizada através da transformação linear

$$\hat{y} = W^T \cdot a \quad (78)$$

Nesse passo um novo script foi criado no qual todas os descritores de superfície de todas as moléculas são submetidos a transformação dando origem aos 6 novos descritores que serão usados para descrever o banco de dados DUD-E no novo subespaço. A redução de dimensionalidade permite descrever o banco de dados de uma forma sucinta e eficiente sem grande perda de informações e preservando aquelas que mais contribuem para a análise. O fato de usarmos agora apenas 6 componentes para descrever uma molécula reduz o tempo gasto na varredura dos ligantes em busca dos mais similares, deixando dessa forma o programa mais eficiente. O fato de usarmos somente as principais gera redução do valor de RMSD (Root Mean Square Deviation) e para isso é necessário diminuir o valor de corte para classificação de moléculas.

3.2.7.6 Funções de Ponderação

A análise de componentes principais propicia o estudo do uso de funções de ponderação na busca de ligantes similares, de forma que as componentes que apresentam os maiores autovalores apresentam mais informação e, portanto, devem apresentar um peso maior na análise sendo como os principais pontos de comparação e os coeficientes com menores autovalores atuando mais como um refinamento na análise.

Para determinar qual é o peso adequado a ser atribuído a cada componente foi realizado um estudo no qual diversas funções foram utilizadas para tentar descrever da melhor forma possível os autovalores obtidos. Dessa forma, a partir da análise da distribuição dos autovalores foram selecionadas curvas que pudessem apresentar os melhores resultados na descrição dos valores. Foram escolhidos polinômios de 3º, 4º, 5º e 7º grau, Distribuição Rayleigh, Distribuição Gaussiana e função exponencial, conforme descrito na Tabela 9.

A curva descrita por cada função é dependente dos coeficientes usados em seus cálculos, sendo a determinação deles a parte mais complexa e delicada dessa análise. Para determinação dos valores foi usado a função "*curve_fit*" presente na biblioteca *ScyPy.Optimize*⁶⁰, responsável por realizar o "*fit*" da função aos autovalores e retornar os coeficientes obtidos em uma lista. Com a definição dos autovalores foi gerado um gráfico contendo aos dados dos autovalores e a curvas que tentam descrevê-los. A análise da função erro que mede a distância euclidiana entre os valores originais e os valores descritos pelas curvas revela que a que melhor descreve os autovalores é o polinômio de 5º grau para ambas as métricas utilizadas no benchmarking.

Tabela 9 - Funções utilizadas na tentativa de descrever os autovalores encontrados a fim de determinar uma função de ponderação a ser usada para métrica Lira

Funções	Fórmulas
Polinômio 3ª Ordem	$f(x) = ax^3 + bx^2 + cx + d$
Polinômio 4ª Ordem	$f(x) = ax^4 + bx^3 + cx^2 + dx + e$
Polinômio 5ª Ordem	$f(x) = ax^5 + bx^4 + cx^3 + dx^2 + ex$ $+ f$
Polinômio 7ª Ordem	$f(x) = ax^7 + bx^6 + cx^5 + dx^4 + ex^3$ $+ fx^2 + gx + h$
Distribuição Rayleigh	$f(x) = 2\theta x e^{-\theta x^2}$
Distribuição Gaussiana	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
Função Exponencial	$f(x) = A \cdot e^{x/b} + k$
Função Erro	$E = \sum_{i=0}^n (y_i - \bar{y}_i)^2$

Fonte: Elaborada pelo autor

Além dos estudos das funções, foi realizada uma análise de regressão utilizando o software Eureqa⁶¹ para determinação de uma função capaz de descrever os autovalores apresentados pelas métricas L^1 e L^2 , conforme mostrado na Tabela 10.

Eureqa é um software desenvolvido pelo Laboratório de Inteligência Artificial de Cornell e comercializado pela Nutonian. Utiliza algoritmos evolucionários para determinar equações matemáticas capazes de descrever um conjunto de dados em sua forma mais simples e assim criando modelos preditivos em segundos. Enquanto que as técnicas atuais de modelagem exigem que o usuário escolha um número definido de algoritmos predeterminados, o Eureqa constrói modelos numéricos, séries temporais e modelos de classificação a partir zero. O tempo para determinação da função é diretamente proporcional a quantidade de pontos fornecidos, como em nosso experimento foram fornecidos os autovalores de cada métrica temos que o tempo para gasto para o cálculo da função correspondente a métrica L^1 foi maior do que o tempo gasto para métrica L^2 contendo 256 e 16 coeficientes, respectivamente.

Tabela 10 - Funções obtidas pelo software Eureqa e as definições utilizadas em sua definição

Funções Eureqa	
<i>Métrica L1</i>	$3.116e9 + 5.724e4 * X - 7.466e6$ $* \mathit{math.log}(X) - 3.03e9$ $* \mathit{less}(2.775, X)$ $- 5.798e7$ $* \mathit{less}(3.932, 0.524 * X)$
<i>Métrica L2</i>	$1.817e8 + 6.729e7 * \mathit{math.log}(X)^2$ $- 2087 * X^2 - 1.78e8$ $* \mathit{logistic}(X^2$ $* \mathit{math.log}(X)^2$ $- 3.403) - 3.466e6$ $* \mathit{math.log}(X)$ $* \mathit{logistic}(X^2$ $* \mathit{math.log}(X)^2$ $- 3.403) - 6.645e7$ $* \mathit{math.log}(X)^2$ $* \mathit{logistic}(X^2$ $* \mathit{math.log}(X)^2$ $- 3.403)$
<i>Logistic(x)</i>	$1.0/(1 + \mathit{math.exp}(-x))$
<i>Less(x, y)</i>	$\begin{cases} 1, se y > x \\ 0, se x > y \end{cases}$

Fonte: Elaborada pelo autor

4 RESULTADOS

Os descritores da superfície apresentam maior variação de valor para métrica variante do que para a invariante. Na variante há presença de coeficientes negativos e positivos. Por definição, a métrica invariante apresenta somente valores positivos. Para ambas as métricas as primeiras componentes apresentam os maiores valores, olhando em termos absolutos, e há diminuição dos valores para as demais.

Os valores médios para métrica L^1 apresentam em sua maioria valores positivos, o que mostra que mesmo com a presença de coeficientes negativos a grande maioria dos descritores apresenta valores positivos para as primeiras componentes. Há componentes médios com valores negativos mostrando que o comportamento inverso também pode ocorrer. A análise das componentes principais revela que os primeiros termos são os que mais contribuem para descrição das moléculas, o que se deve ao fato de representarem os termos esféricos, as primeiras bandas de harmônicos esféricos. (Figura 27) Após a ordenação dos autovalores e apesar das demais componente apresentarem valores menores do que as iniciais, foram selecionados 90 autovetores para a construção do novo espaço vetorial de forma que não houvesse perda de informação exagerada atribuído e que a ordem de grandeza dos autovalores não variasse muito.

Além da redução da dimensionalidade, os autovalores foram utilizados para tentar descrever uma função de ponderação que atribuísse mais peso as componentes que mais contribuem para a descrição da superfície do ligante. Foi selecionado um polinômio de quinta ordem e uma função gerada através de um método de regressão feito pelo software Eureqa. Ao tentar fazer o benchmarking utilizando a métrica ponderada, há um problema no uso da função Eureqa, pois a mesma contém função logaritmo em sua definição e ao tentarmos aplicar um descritor com valor negativo encontramos uma indefinição matemática que interrompe o cálculo. Dessa forma, no novo benchmarking somente a métrica ponderada pelo uso do polinômio de quinto grau foi levado em consideração.

Figura 27- Dados obtidos para experimento de PCA métrica L^1 . Representação de todos os coeficientes usados para descrição das moléculas. Valores médios calculados para cada componente. Autovalores calculados durante análise. Análise de curvas de utilizada para determinação de uma função ponderação.

Fonte: Elaborada pelo autor.

Para métrica L^2 , observa-se que a segunda componente apresenta valor médio menor do que a terceira e para as demais os valores vão diminuindo. Entretanto o experimento de PCA revela que mesmo tendo menor valor médio, a segunda componente é tão importante quanto a primeira apresentando autovalores maiores, fato atribuído novamente ao fato de serem as primeiras bandas dos harmônicos esféricos. (Figura 28) Para os demais autovalores há grande variação de seus valores em relação aos dois primeiros, porém quando comparados entre si não há grande mudança na ordem de grandeza e devido a esse fato foram selecionados 6 autovetores para a criação do novo subespaço vetorial.

Figura 28- Dados obtidos para experimento de PCA métrica L^2 . Representação de todos os coeficientes usados para descrição das moléculas. Valores médios calculados para cada componente. Autovalores calculados durante análise. Análise de curvas de utilizada para determinação de uma função de ponderação.

Fonte: Elaborada pelo autor.

Os autovalores também foram utilizados para a construção de funções de pontuação, da mesma forma descrita para métrica L^1 . A melhor função foi novamente um polinômio de quinta ordem e a função obtida pelo Eureka. A análise do TOP 10 moléculas selecionadas pela métrica original, ponderada pelo polinômio de quinta ordem e função Eureka, revela que a intersecção de moléculas selecionada pela métrica original e o polinômio é maior do que quando comparada a intersecção proveniente da seleção ponderado pela segunda função. (Figura 29) Dessa forma, um novo benchmarking foi realizado utilizando somente a função polinomial como classificador. Comparando os resultados obtidos nota-se que para o caso da molécula 35kith houve melhora na classificação com o uso da função polinomial, sendo melhor que um classificador aleatório. No caso da molécula 74bace1, em que o classificador era tido como aleatório, houve alteração da curva ROC fazendo com que ora ele classificasse melhor e ora classificasse pior do que o aleatório, as melhores classificações se encontram para os maiores valores de corte. Para a molécula 305urok, o classificador antigo apresentava-se ligeiramente acima do aleatório e com o uso da função polinomial houve melhora considerável na

classificação, fazendo com valores mais elevados fossem atingidos. Também houve melhora na para a molécula 313aces, que apresentava um classificador aleatório e após o uso do polinômio alcançou pontos mais elevados na curva ROC mostrando melhora substancial na classificação de moléculas.

Tanto para o caso da redução da dimensionalidade quanto para o uso de funções de ponderação houve ganho de eficiência gerando melhora na classificação aumentando a acurácia da análise, quando comparado ao USR. Em relação a velocidade de processamento, não houve mudanças significativas ao adicionarmos as funções de ponderação. Já para a redução de dimensionalidade houve diminuição do tempo total gasto, uma vez que se comparam menos descritores do que no benchmarking original 6 e 16 coeficientes para métrica L^2 e 90 e 256 coeficientes para métrica L^1 , respectivamente.

MOLÉCULA 35KITH		
ORIGINAL	PONDERADA	EUREQA
238reni	238reni	238reni
686ada17	120mk01	120mk01
120mk01	686ada17	567drd3
169fpps	169fpps	1131hivpr
5andr	5andr	56aces
341vgfr2	341vgfr2	415egfr
34kith	475ada17	207ptn1
475ada17	12kith	171mp2k1
44esr1	34kith	1134hivpr
53esr2	44esr1	36tgfr1
INTERSECÇÃO	80%	20%

MOLÉCULA 305UOK		
ORIGINAL	PONDERADA	EUREQA
238reni	238reni	238reni
686ada17	686ada17	1131hivpr
341vgfr2	341vgfr2	56aces
44esr1	44esr1	415egfr
53esr2	415egfr	1134hivpr
415egfr	53esr2	595thrb
56aces	56aces	1207hivpr
52esr2	52esr2	1330hivpr
600fnta	600fnta	44esr1
726mmp13	688ada17	177esr2
INTERSECÇÃO	90%	40%

MOLÉCULA 74BACE1		
ORIGINAL	PONDERADA	EUREQA
238reni	238reni	238reni
250bace1	250bace1	369egfr
260bace1	265bace1	368egfr
134bace1	260bace1	654aces
313aces	134bace1	589aces
314aces	313aces	856ada17
11aces	314aces	417adrb1
69bace1	69bace1	100hdac2
23reni	124reni	136pparg
297bace1	246reni	44mk01
INTERSECÇÃO	70%	10%

MOLÉCULA 313ACES		
ORIGINAL	PONDERADA	EUREQA
314aces	314aces	720drd3
10aces	10aces	316drd3
55aces	438adrb2	20csf1r
54aces	77adrb1	314aces
53aces	751mmp13	119egfr
11aces	290aces	120egfr
290aces	103adrb2	352egfr
438adrb2	88adrb1	719drd3
77adrb1	54aces	370aces
103adrb2	55aces	134lck
INTERSECÇÃO	80%	10%

Figura 29- Comparação entre moléculas selecionadas pela métrica L^2 utilizando a distância euclidiana original, ponderada por um polinômio de quinta ordem e pela função gerada pelo software Eureka.

Fonte: Elaborada pelo autor.

4.1 Novos Benchmarkings

Com a criação de um novo subespaço de dados foi possível refazer o benchmarking para as métricas L^1 e L^2 do DUD-E após a redução da dimensionalidade do banco de dados a fim de verificar se a escolha dos autovetores foi adequada e se o novo banco de dados continha as principais informações para descrição das moléculas. A métrica utilizada para comparar os descritores foi a distância euclidiana original, sem adição de pesos as medidas, pois seria necessário realizar uma nova análise de componentes principais para determinar quais são as componentes que detém mais informação no novo subespaço. Dessa forma, o script realizado previamente foi adaptado para a construção dos novos estudos apresentando apenas as alterações necessárias para a busca dos novos bancos de dados.

Para a métrica L^2 , foi realizado um novo estudo analisando somente as curvas ROC utilizando o polinômio de quinta ordem como função de ponderação e observa-se que sua utilização agrega valor ao nosso experimento pois o Lira passou a ser determinado como bom classificador para as moléculas analisadas previamente. (Figura 30)

Figura 30- Curvas ROC calculadas para a métrica L2 utilizando como função de ponderação polinômio de quinta ordem. Análise revela que uso de função melhora status de classificação para programa Lira.

Fonte: Elaborada pelo autor.

Para a métrica L^1 , o novo banco de dados contém 90 descritores para cada molécula, a redução da dimensionalidade diminuiu o tempo gasto para realizar a varredura do banco de

dados deixando-o mais rápido. A comparação entre os resultados da métrica antes e depois da análise de PCA revela que a diminuição da quantidade de descritores não gerou perda considerável de informações, fato comprovado pelo fato das curvas ROC serem parecidas mantendo as características do benchmarking original conforme mostrado nas Figuras 31 e 32.

Figura 31- Análise de curva ROC gerados para espaço vetorial original e novo espaço vetorial fruto do processo de redução de dimensionalidade gerado pelo experimento de PCA. Para molécula 305urok não há alterações entre gráficos originais e novos evidenciando manutenção de características principais.
Fonte: Elaborada pelo autor.

Figura 32- Análise de curva ROC gerados para espaço vetorial original e novo espaço vetorial fruto do processo de redução de dimensionalidade gerado pelo experimento de PCA. Para molécula 35kith não há alterações entre gráficos originais e novos evidenciando manutenção de características principais. Fonte: Elaborada pelo autor.

No caso da molécula 313aces (Figura 33) pois os coeficientes originais apresentavam pontos na curva ROC sobre a linha do classificador aleatório devido ao fato de para os menores valores de corte a quantidade de moléculas ser baixa e somente aumentar quando o valor era alterado de 7.0 para 7.5 e em seguida a taxa de verdadeiro positivo era mantida constante. Para a nova análise a taxa de verdadeiro positivo se mantém constante desde o início do experimento, ou seja, mesmo para os menores valores as moléculas selecionadas pelo USR também são selecionadas pelo Lira usando 90 descritores. Essa alteração revela que a diminuição da dimensionalidade melhora a atuação do Lira sendo agora um classificador melhor do que o original para o mesmo alvo molecular.

Figura 33- Análise de curva ROC gerados para espaço vetorial original e novo espaço vetorial fruto do processo de redução de dimensionalidade gerado pelo experimento de PCA. Para molécula 313aces novo espaço vetorial mantém informações essenciais para análise de similaridade mostrando ligeira melhora em relação à experimento inicial.

Fonte: Elaborada pelo autor.

No caso da molécula 74bace1, mesmo após a redução da dimensionalidade o Lira continua classificando abaixo do aleatório. (Figura 34) Entretanto, a quantidade de moléculas selecionadas aumenta, no qual uma parte se encontra no conjunto de classificado pelo USR, fazendo com que a taxa de verdadeiro positivo aumente enquanto que a taxa de falso positivo permanece igual ao registrado para o benchmarking original fazendo com que o classificador se aproxime cada vez mais de um status de aleatório. Esse fato mostra que a redução da dimensionalidade gerou melhora no classificador, não o suficiente para a alteração de status.

Figura 34- Análise de curva ROC gerados para espaço vetorial original e novo espaço vetorial fruto do processo de redução de dimensionalidade gerado pelo experimento de PCA. Para molécula 74bace1 apesar de ambos os gráficos permanecerem abaixo da curva pode se notar melhora na descrição com 90 componentes pois há aproxima dos pontos a curva.

Fonte: Elaborada pelo autor.

No caso da métrica L^2 , o novo banco de dados utiliza 6 componentes para a descrição de cada molécula, a redução da dimensionalidade deixa o processo ainda mais rápido sendo mais veloz do que o USR. A comparação entre os resultados do experimento original e o novo benchmarking revela que além do aumento de velocidade, houve alteração nas curvas ROC indicando melhora no processo de classificação. No caso da molécula 35kith não houve mudanças significativas de um experimento para o outro, as taxas de verdadeiro positivo e falso positivo de mantiveram constantes e os pontos na curva ROC estão localizados basicamente nas mesmas posições.(Figura 35) Monstando que os autovetores selecionados para criação do novo subespaço não gerou perda de informações significativas e o Lira se manteve como um bom classificador.

Figura 35- Análise de curva ROC gerados para espaço vetorial original e novo espaço vetorial fruto do processo de redução de dimensionalidade gerado pelo experimento de PCA. Para molécula 35kith não há alterações entre gráficos originais e novos evidenciando manutenção de características principais para a descrição com 6 componentes.

Fonte: Elaborada pelo autor.

No caso da molécula 74bace1, o Lira atuava de forma semelhante a um classificador aleatório e no novo experimento fica claro que a quantidade de molécula selecionadas aumenta fazendo com que a taxa de verdadeiro positivo cresça entre os valores de corte de 1.0 e 1.5. (Figura 36) Porém ao mesmo tempo, a taxa de falso positivo também cresce mostrando que uma parte das moléculas selecionadas não se encontra no conjunto de moléculas classificados pelo USR. Apesar dos aumento das taxas de verdadeiro e falso positivos e da quantidade de moléculas o Lira, no geral, continua atuando como um classificador aleatório apresentando alguns pontos ligeiramente abaixo da curva do aleatório.

Figura 36- Análise de curva ROC gerados para espaço vetorial original e novo espaço vetorial fruto do processo de redução de dimensionalidade gerado pelo experimento de PCA. Para molécula 74bace1 a diminuição de descritores favorece a seleção de mais molécula que alteram o formato da curva ROC. Fonte: Elaborada pelo autor.

Para a molécula 305urok, há aumento da quantidade de moléculas selecionadas para valores de corte mais elevados.(Figura 37) Diferente do caso da 74bace1, a taxa de verdadeiro positivo cresce mais do que a taxa de falso positivos, o que gera pontos na curva ROC acima da linha do aleatório mostrando que a redução da dimensionalidade melhorou ligeiramente o potencial do que já era tido como um bom classificador.

Figura 37- Análise de curva ROC gerados para espaço vetorial original e novo espaço vetorial fruto do processo de redução de dimensionalidade gerado pelo experimento de PCA. Para molécula 305urok não há alterações entre gráficos originais e novos evidenciando manutenção de características principais.

Fonte: Elaborada pelo autor.

A molécula 313aces é a que apresenta a principal alteração da utilização do novo espaço, conforme mostram na Figura 38. No intervalo entre 0.5 e 1.0 há aumento da quantidade de molécula selecionadas sendo que sua grande maioria deles se encontra no conjunto de classificado pelo USR. Com isso, há aumento de 20% na taxa de verdadeiro positivo o que faz com que o classificador deixe de ser aleatório e passando para o status de bom.

De modo geral, a análise de componentes principais, diminuição da dimensionalidade e uso de funções de ponderação agregam valor ao método de análise proposto pelo Lira, tornando-o tão bom quanto o USR. Na maioria dos casos apresentados há melhora de seu potencial como classificador, nas situações em que não há melhora o potencial não apresenta queda se mantendo constante, como no caso do 35kith.

Figura 38- Análise de curva ROC gerados para espaço vetorial original e novo espaço vetorial fruto do processo de redução de dimensionalidade gerado pelo experimento de PCA. Para molécula 313aces há bom aumento na quantidade de moléculas selecionadas pela descrição com 6 componentes originando aumento dos pontos de curva e na status de bom classificador da métrica.

Fonte: Elaborada pelo autor.

4.2 Análise estatística do benchmarking

A partir do benchmarking também é possível realizar uma análise estatística para as cinco variações das métrica, L^2 , L^2 ponderada pelo polinômio de quinta ordem, L^2 contendo 6 descritores obtidas a partir do PCA, L^1 e L^1 contendo 90 descritores obtidos a partir do PCA analisadas em nosso experimento utilizando as 50 moléculas. Para esse trabalho as curvas ROC de cada molécula para cada métrica foi analisada e o algoritmo foi classificado de acordo com os pontos gerados para cada valor de corte utilizado. Para as moléculas em que os pontos estavam majoritariamente acima da linha do classificador aleatório, o algoritmo foi tido como um bom classificador. Para aquelas em que os pontos se encontravam majoritariamente abaixo do classificador aleatório, o algoritmo foi determinado como um mau classificador. E para as moléculas cujos pontos se encontravam sobre a linha do classificador aleatório ou que os pontos oscilassem, ora bom classificador, ora mau classificador em uma taxa de pontos semelhantes foi tido como classificador aleatório. A classificação de cada molécula está presente no

Apêndice B e a proporção de classificação para cada métrica L^2 e L^1 e para suas variações estão descritas na Tabela 11 e Tabela 13, respectivamente.

Tabela 11 - Porcentagem de classificação para cada variação da métrica L^2

Classificação	L^2	L^2Ponderada	$L^2$6D
Bom Classificador	80%	88%	86%
Aleatório	18%	4%	12%
Mau Classificador	2%	8%	2%

Fonte: Elaborada pelo autor.

Para a métrica L^2 original, 16 descritores e sem ponderação, o algoritmo apresentasse como um bom classificador tendo taxa de acerto de 80%, ou sejam, em 40 moléculas o algoritmo se comportou melhor do que o USR. É a métrica que apresenta mais moléculas com classificação do algoritmo como aleatório, 18%, e também apresenta a menor quantidade de moléculas em que o algoritmo foi tido como mau classificador, mesma proporção encontrada para a métrica L^2 contendo somente 6 descritores, taxa de 2% e essa classificação foi obtida para a molécula 7def. Dessa forma, a utilização da métrica original para seleção de ligantes similares apresenta alto percentual de acerto quando comparada ao USR.

A comparação da métrica L^2 ponderada com a métrica L^2 original, revela que a utilização do polinômio de quinta ordem como função de ponderação eleva o índice de acerto do algoritmo de 80% para 88%, ou seja, em 44 moléculas analisadas o algoritmo obteve resultados superiores ao USR, diminui a taxa de moléculas em que o algoritmo foi tido como aleatório de 18% para 4%, porém aumenta a proporção de moléculas em que o Lira foi tido como um mau classificador. Sendo assim, pode-se observar que a utilização da função de ponderação agrega valor ao experimento, pois elevada a quantidade de moléculas em que tivemos um bom classificador e faz com que o nosso algoritmo se afaste do classificador aleatório, atuando ora como bom classificador, ora como mau classificador.

A métrica L^2 contendo somente 6 descritores moleculares (L^2 6D) também apresenta melhora na proporção de moléculas em que o Lira foi tido como bom classificador quando comparada com a métrica L^2 original, para essa situação a proporção é incrementado em 6% sendo alterada de 80% para 86%. A taxa de moléculas em que o algoritmo é tido como aleatório é de 12% e esta compreendida entre as taxas encontradas para as métricas L^2 e L^2 ponderada.

Dessa forma, observa-se das três variações utilizadas para que a métrica L^2 , a métrica L^2 ponderada é quem apresenta maior proporção de moléculas em que o Lira foi determinado como bom classificador e as métricas L^2 e L^2 6D são as que apresentam a menor quantidade de

moléculas nas quais o algoritmo foi determinado como mau classificador. Além disso, observa-se que o tempo de processamento para as 50 moléculas é menor para a métrica L^2 6D, são utilizadas menos componentes para descrever a superfície molecular e a métrica L^2 ponderada é aquela que apresenta o maior tempo pois além de apresentar 16 descritores moleculares para cada molécula utiliza o polinômio de quinta ordem para fornecer a ponderação a cada componente. (Tabela 12)

Tabela 12 - Tempo de processamento para as métricas L^2

Métrica	L^2	L^2 Ponderada	L^2 6D
Tempo(s)	1082.042471	2386.363254	746.147301

Fonte: Elaborada pelo autor.

Para a métrica L^1 original, contendo 256 descritores moleculares, a taxa de moléculas em que o algoritmo é determinado como bom classificador é de 56%, apresentando taxa de 6% para classificador aleatório e 38% das moléculas o algoritmo foi tido como um mau classificador. Sendo assim, a métrica L^1 apresenta percentual de acerto mediano e inferior aos encontrados para a métrica L^2 , não sendo indicada como única métrica de análise e sim como um complemento a métrica L^2 .

Tabela 13 - Porcentagem de classificação para cada variação da métrica L^1

Classificação	L^1	L^1 90D
Bom Classificador	56%	58%
Aleatório	6%	6%
Mau Classificador	38%	36%

Fonte: Elaborada pelo autor.

A diminuição da quantidade de descritores moleculares para a métrica L^1 , de 256 para 90 componentes, não fornece grande melhora ao experimento pois o aumento da taxa de moléculas em que o algoritmo foi determinado como bom classificador é de 2%, ou seja, somente uma molécula a mais em relação a quantidade selecionada originalmente. Dessa forma, observa-se que a métrica L^1 apresenta-se melhor que o USR em cerca de 56% dos casos, porém tem resultados baixos quando comparados os obtidos pela métrica L^2 , indicando que seu melhor uso seja de complementar as informações obtidas pela métrica invariante.

Tabela 14 - Tempo de processamento para as métrica L^1

Métrica	L^1	L^1 90D
Tempo(s)	10326.941013	5977.671860

Fonte - Elaborado pelo autor.

Em relação ao tempo de processamento, observa-se que a métrica contendo somente 90 descritores é mais rápida pois o número de comparações e cálculos sendo até 40% mais eficiente.

A comparação entre as proporções encontradas para cada tipo de classificação para as métricas L^2 e L^1 e suas variações está descrita na Figura 39 e podemos observar que os resultados obtidos pela métrica L^2 são mais confiáveis e robustos quando comparado a métrica L^1 .

Figura 39 - Comparação entre as classificações para cada métrica e variação utilizada no benchmarking.

Fonte: Elaborada pelo autor.

Além disso, o tempo de processamento (Figura 40) da métrica invariante é muito menor quando comparado com o da métrica variante, como era esperado devido a diferença na quantidade de descritores utilizados por cada métrica, demonstrando que além de ter maior percentual de acertos, a métrica L^2 é mais rápida sendo a mais adequada para ser usada no site Lira, pois fornecerá resultados robustos e confiáveis em tempos menores.

Figura 40- Comparação entre os tempos de processamento para cada métrica e variação utilizada no benchmarking

Fonte: Elaborada pelo autor.

4.3 Website Lira

Com finalidade de divulgação do método de análise de similaridade de ligantes baseados em harmônicos esféricos foi criado o site Lira (Figura 41), desenvolvido utilizando a plataforma Django⁶² devido ao fato de ser um framework de Python para Web e o método ter sido implementado em Python. É utilizado na busca de ligantes similares a uma determinada molécula alvo inserida pelo usuário.

Figura 41- Página inicial do Web site Lira no qual uma molécula é informada ao site em busca das moléculas mais similares.

Fonte: Elaborada pelo autor.

4.3.1 Método de busca

A busca pelas moléculas semelhantes segue o mesmo algoritmo utilizado para a validação do método nos estudos de benchmarking.

- Leitura de um arquivo contendo as coordenadas e ligações atômicas com extensão .PDB.

- Cálculo dos descritores moleculares da molécula inserida.
- Varredura do banco de dados DUD-E em busca dos ligantes mais similares.

São utilizados os dois métodos para a busca de ligantes, LIRA e USR, cada um sendo avaliado independente do outro e retornado à tela no máximo dez ligantes por métrica. O tempo total gasto para classificação e o retorno na tela leva cerca de dez segundos devido ao fato de ambas as métricas serem rápidas e eficientes.

Para determinar a similaridade das moléculas foram escolhidos dois valores de corte, um para cada métrica, baseados nos estudos realizados anteriormente com a finalidade de selecionar o número as moléculas mais adequadas para os posteriores estudos. Dessa forma, para o Lira escolheu-se o valor de corte de 0.5 Å, sendo 0 menor valor possível ocorrendo no caso da análise da mesma molécula, e para o USR o valor escolhido é de 0.90 Å, sendo 1 o valor máximo. Devido aos valores escolhidos serão classificadas como similares as moléculas com valor de similaridade menor que 0.5 Å para o caso do Lira e valores superiores a 0.90 Å no caso do USR. Após a seleção das moléculas é feita a ordenação do conteúdo partindo da mais similar até a menos similar para o retorno a tela.

4.3.2 Display das moléculas

Após a seleção das moléculas, o resultado é mostrado na tela com representações bidimensionais de cada molécula selecionada e seus respectivos valores de similaridade, ordenadas da mais similar para a menos similar em relação à molécula original. Ao clicar em qualquer molécula presente na tela o site abre uma nova janela na qual é criada uma representação tridimensional da molécula através do uso do JsMOL⁶³⁻⁶⁵, ferramenta semelhante ao Jmol escrita usando se somente JavaScript não sendo necessário a instalação da plataforma Java para sua execução. A nova visualização apresenta funções como aproximar e aumentar o tamanho da molécula, rotação seleção de átomos, seleção do tipo de superfície, salvar uma imagem entre outros tipos que já são empregados em outros visualizadores de moléculas, como PyMol, que se fazem extremamente úteis durante a análise.

Para tornar o processo mais rápido e eficiente todas as representações bidimensionais das moléculas do DUD-E foram geradas previamente e se encontram armazenadas em um banco de dados, que é acessado quando se solicita a imagem no momento da criação da página de resultado da busca. As imagens foram processadas utilizando-se a ferramenta Open Babel. (Figura 42)

Figura 42- Visualização 2D e 3D das moléculas selecionadas pelo Lira e pelo USR. Para a estrutura 3D utiliza-se a ferramenta Jsmol.

Fonte: Elaborada pelo autor.

Além da representação bidimensional e do valor de similaridade, há um link com a literal “Superposition” que ao ser clicado solicita ao servidor que realize a superposição da molécula selecionada com a original. O programa de superposição foi desenvolvido por Samuel Silva, aluno de Doutorado do grupo, e utiliza harmônicos para a descrição da superfície molecular dos ligantes, “simulate annealing” entre outras técnicas para superpor as moléculas.

O cálculo da superposição é feito segundo o algoritmo:

- Cálculo dos descritores da superfície de cada molécula;
- Ordenação da orientação espacial das moléculas, colocando-as na mesma orientação;
- Cálculo dos valores de similaridade das posições atômicas de cada átomo;
- Criação de um arquivo com extensão.PDB contendo as posições e ligações para cada átomo.

Devido ao fato de ser um algoritmo complexo e envolver vários cálculos, a superposição do par de moléculas, a de entrada e a selecionada pelo algoritmo, leva em torno de 300 segundos para ficar pronta e ao final é disponibilizada uma nova guia contendo a representação tridimensional da superposição. Ou seja, o tempo necessário para superpor todas as moléculas encontradas é dado pela multiplicação do tempo de superposição pela quantidade de moléculas selecionadas.

4.3.3 Melhorias futuras

O site ainda se encontra em fase de desenvolvimento e aprimoramento. Atualmente utiliza somente a análise da superfície geométrica como referência para a busca de ligantes, entretanto existem estudos para o uso de descrições de propriedades físico-químicas,

participantes do processo de reconhecimento molecular, tanto para a métrica baseada em harmônicos esféricos quanto para o USR⁴³, capazes de agregar mais valor e robustez ao processo de busca de ligantes similares.

5 CONSIDERAÇÕES FINAIS

Métodos computacionais ganham cada vez mais espaço no processo de desenvolvimento de fármacos e principalmente quando é utilizado o desenho racional. Sua principal vantagem é realizar etapas necessárias e fundamentais para o processo, como a seleção de um conjunto de moléculas que podem ser utilizadas em experimentos de SAR e QSAR. Além da busca de ligantes é possível utilizar métodos computacionais para prever função, atividade, reconhecimento, mapeamento de propriedades físico-químicas e superfície geométrica de compostos biológicos, como moléculas e proteínas, agregando eficiência e valor aos estudos.

O programa Lira utiliza harmônicos esféricos, a parte angular da solução da Equação de Laplace, como método matemático para descrição da superfície geométrica de ligantes. Os harmônicos esféricos formam uma base ortonormal, ou seja, os vetores que formam o espaço vetorial são ortogonais e normalizados, possibilitando que sejam usados como uma base canônica para descrição de moléculas. A quantidade de descritores está associada à quantidade de bandas que serão utilizadas para descrever uma molécula e usados na reconstrução da superfície original, conforme o número de descritores aumenta o erro associado da comparação das superfícies diminui. Porém, a partir de 15 bandas, ou seja $l_{max} = 15$, o erro volta a crescer. Apresenta duas métricas que podem ser utilizadas em experimento de busca de ligantes similares, L^1 e L^2 , cada uma tendo seus pontos positivos e negativos.

A métrica L^2 , também chamada de invariante, apresenta menor quantidade de coeficientes sendo o total de $l_{max} + 1$, e tem como principal vantagem o fato de seus descritores serem rotacionalmente invariantes, ou seja, independente da orientação espacial os valores do coeficiente serão os mesmos, o que facilita a busca por ligantes não sendo necessário o prévio alinhamento das moléculas. Por sua vez, a métrica L^1 , também chamada de métrica variantes, apresenta maior quantidade de descritores moleculares, 256, e é dependente da orientação espacial.

Os resultados apresentados no benchmarking evidenciam que a métrica baseada em harmônicos esféricos proposta pelo Lira é tão boa quanto ao método proposto pelo USR, que já foi validada frente aos principais métodos presentes atualmente. O tempo gasto por cada métrica para realizar a varredura do banco de dados e comparar individualmente as moléculas é dependente da quantidade de descritores para cada uma. No caso da métrica L^2 , há 16 coeficientes, para métrica L^1 há 256 coeficientes e para o USR são 12 componentes divididos em quatro grupos de medidas. Comparando os tempos totais observa-se que a métrica invariante

é tão rápida quanto a USR, havendo variação de um décimo de segundo. Para a métrica variante a maior quantidade de coeficientes resulta em um processo mais lento. Dessa forma, pode se afirmar que há melhor eficiência para métrica invariante do que para a métrica quando comparado à performance do USR.

Em relação as moléculas selecionadas por cada métrica, as curvas ROC mostram que o programa Lira é um bom classificador sendo no mínimo igual ao aleatório para a métrica L^2 , e podendo apresentar resultados melhores e piores no caso da métrica L^1 . A análise de componentes principais proporcional a redução da dimensionalidade e consequente quantidade de descritores moleculares para cada métrica gerando diminuição do tempo total gasto na análise para ambas as métricas. Além disso, demonstrou quais são as componentes que mais contribuem para a descrição molecular e deve ter destaque durante o processo de busca de ligantes similares o qual passou a ter pesos distintos para cada descritor. O uso de funções de ponderação melhorou o benchmarking, fazendo com que moléculas que haviam apresentado um classificador aleatório passassem a ter um bom classificador e moléculas que apresentaram bons classificadores tiveram seu potencial incrementado.

O benchmarking realizado para os novos e reduzidos espaços vetoriais mostram que o processo de PCA é robusto e propicia a redução da dimensionalidade, aumento da eficiência devido a diminuição de componentes a serem comparadas e a manutenção padrões encontrados anteriormente. No caso da métrica L^2 as curvas ROC praticamente foram repetidas evidenciando que os seis autovetores selecionados para criação do novo subespaço continham grande parte da informação necessária para descrever as moléculas.

O site Lira é uma ferramenta simples, porém útil no processo de seleção de moléculas semelhantes no qual pesquisadores podem verificar quais são os ligantes presentes no banco de dados DUD-E que são mais similares à sua molécula analisada. A utilização do JsMol favorece a visualização das moléculas selecionadas permitindo que sejam extraídas fotos e seja feita a manipulação do composto de forma semelhante ao pyMol.

5.1 Melhorias Futuras

Apesar dos bons resultados alcançados pelo programa Lira ainda há pontos que podem ser introduzidos para aumentar seu poder e deixá-lo mais completo. É sabido que além da complementaridade geométrica propriedades físico-químicas também auxiliam no processo de reconhecimento molecular, evento essencial para qualquer processo biológico. No caso do USR, já existem estudos⁴² que utilizam propriedades além da geométrica na busca de ligantes

e os resultados encontrados são animadores. Dessa forma, acredita-se que os usos de novas propriedades moleculares podem agregar valor ao processo de busca deixando-a busca mais seletiva e possibilitando outros usos como docagem molecular. Associado à utilização de novas propriedades se faz necessário o desenvolvimento de uma métrica adequada para mensurar o potencial de cada propriedade a fim de extrair as melhores informações possíveis. A distância euclidiana revela-se eficiente para análise da complementaridade geométrica no processo de seleção de ligantes, porém a adição de pesos aos descritores moleculares favoreceu ainda mais essa busca evidenciando que a métrica pode ser melhorada. Dessa forma, os desenvolvimentos de novas métricas podem agregar valor e refinamento ao processo.

A inclusão de mais moléculas e outros banco de dados pode agregar valor ao método proposto pelo Lira pois o aumento da diversidade molecular favorece uma busca mais completa variando por mais tipos de alvos moleculares o que deixa o experimento mais completo. Entretanto, as moléculas e bancos de dados a serem incluídos devem ter um prévio estudo para validar se seu uso agrega valor ou não aos experimentos.

REFERÊNCIAS

- 1 KROGSGAARD-LARSEN, P.; LILJEFORS, T.; MADSEN, U. **Textbook of drug design and discovery**. 3rd ed. Washington, DC: CRC Press, 2002.
- 2 MERS, K. M.; RINGE, D.; REYNOLDS C. H. **Drug design: structure- and ligand-based approaches**. Cambridge, UK: Cambridge University Press, 2010.
- 3 GLEN, R. C.; BENDER, A. Molecular similarity: a key technique in molecular informatics. **Organic & Biomolecular Chemistry**.v2 n. 22.p 3204-3218, 2 0 0 4.
- 4 SCHEMATIC diagram of drug discovery cycle. Disponível em: <https://commons.wikimedia.org/wiki/File:Drug_discovery_cycle.svg, 2015>. Acesso em: 03 abr. 2016.
- 5 BENDER, A. **Studies on molecular similarity**. 2005. 182p. Ph.D. Thesis (Philosophy) - University of Cambridge, Cambridge, 2005.
- 6 FERREIRA, L. G. et al. Molecular docking and structure-based drug design strategies. **Molecules**, v. 20, n. 7, p. 13384-13421, 2015.
- 7 LEE, C. H.; HUANG, H. C.; JUAN, H. F. Reviewing ligand-based rational drug design: the search for an ATP synthase inhibitor. **International Journal of Molecular Sciences**, v. 12, n. 8, p. 5304-5318, 2011.
- 8 MARTEL, P. **Descoberta, desenho e desenvolvimento de fármacos**. Disponível em: <http://w3.ualg.pt/~pmartel/cadeiras/mmfarm/DD_and_MM.pdf, 2015>. Acesso em: 20 jul. 2016.
- 9 MAK, L.; GRANDISON, S.; MORRIS, R. J. An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison. **Journal of Molecular Graphics & Modelling**, v. 26, n. 7, p. 1035-1045, 2008.
- 10 KAHRAMAN, A. et al. Shape variation in protein binding pockets and their ligands. **Journal of Molecular Biology**, v. 368, n. 1, p. 283-301, 2007.
- 11 ROCHA, W. R. Interações intermoleculares. **Química Nova na Escola**, v. 4, p. 31-36, 2001.
- 12 VERLI, H.; BARREIRO, E. J. Um paradigma da química medicinal: a flexibilidade dos ligantes e receptores. **Química Nova**. v. 28, n. 1.p 95-102. 2005.
- 13 FRAGA, C. A. M. Razões da atividade biológica: interações micro- e biomacro-moléculas. **Química Nova na Escola**, v. 3, p. 33-42, 2001.
- 14 ATIKINS, P; JONES, L. **Princípios de química** - questionando a vida moderna e o meio ambiente. 5a ed. Porto Alegre: Bookman, 2011.
- 15 LIGAÇÃO sigma e ligações pi. Disponível em: <<http://www.colegioweb.com.br/ligacoes-quimicas/ligacao-sigma-e-ligacoes-pi.html>, 2012>. Acesso em: 04 abr. 2016.

- 16 KITCHEN, D. B. et al. Docking and scoring in virtual screening for drug discovery: methods and applications. **Nature Reviews Drug Discovery**, v. 3, n. 11, p. 935-949, 2004.
- 17 SLIWOSKI, G. et al. Computational methods in drug discovery. **Pharmacological Reviews**, v. 66, n. 1, p. 334-395, 2014.
- 18 DEAN, P. M. **Molecular similarity in drug design**. Dordrecht, Netherlands: Springer, 1995.
- 19 VINTER, J. G.; GARDNER, M. **Molecular modelling and drug design**. Washington, DC: CRC Press, 1994.
- 20 DOCAGEM molecular. Disponível em:
<http://www.gradadm.ifsc.usp.br/dados/20102/FFI07761/Aula_06_Docagem_molecular.pdf, 2012> Acesso em: 04 abr. 2016.
- 21 YANG, S. Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. **Drug Discovery Today**, v. 15, n. 11-12, p. 444-450, 2010.
- 22 ACHARYA, C. et al. Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. **Current Computer-Aided Drug Design**, v. 7, n. 1, p. 10-22, 2011.
- 23 BAUM, D. **A point-based algorithm for multiple 3D surface alignment of drug-sized molecules**. 2007. 196p. (Master Thesis) - Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, 2007.
- 24 TODESCHINI, R.; CONSONNI, V. **Molecular descriptors for chemoinformatics**. 2nd ed. Weinheim, Germany: Wiley-VCH, 2009.
- 25 KIAMETIS, A. S. **Modelagem molecular de potenciais candidatos a inibidores da acetilcolinesterase**. 2012. 76p. Tese (Doutorado) - Instituto de Física, Universidade de Brasília. Brasília, 2012.
- 26 GENETIC algorithm. Disponível em:
<http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol1/hmw/article1.html, 1996>. Acesso em: 26 mai. 2016.
- 27 JAIN, A. N. Morphological similarity: a 3D molecular similarity method correlated with protein-ligand recognition. **Journal of Computer-Aided Molecular Design**, v. 14, n. 2, p. 199-213, 2000.
- 28 RITCHIE, D. W.; KEMP, G. J. L. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. **Journal of Computational Chemistry**, v. 20, n. 4, p. 383-395, 1999.
- 29 MORRIS, R. J. An evaluation of spherical designs for molecular-like surfaces. **Journal of Molecular Graphics & Modelling**, v. 24, n. 5, p. 356-361, 2006.
- 30 ARFKEN, G.; WEBER, H. **Física matemática - métodos matemáticos para engenharia e física**. Rio de Janeiro: Campus Elsevier, 2007.

- 31 SPHERICAL harmonic lighting: the gritty details. Disponível em:
<<http://silviojemma.com/public/papers/lighting/spherical-harmonic-lighting.pdf>, 2003>.
Acesso em: 05 jun. 2013.
- 32 CAI, W. S.; SHAO, X. G.; MAIGRET, B. Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. **Journal of Molecular Graphics & Modelling**, v. 20, n. 4, p. 313-328, 2002.
- 33 CAI, W. S.; SHAO, X. G.; MAIGRET, B. Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. **Journal of Molecular Graphics & Modelling**, v. 20, n. 4, p. 313-328, 2002.
- 34 MORRIS, R. J. et al. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. **Bioinformatics**, v. 21, n. 10, p. 2347-2355, 2005.
- 35 KAZHDAN, M.; FUNKHOUSER, T.; RUSINKIEWICZ, S. Rotation invariant spherical harmonic representation of 3D shape descriptors. In: EUROGRAPHICS SYMPOSIUM ON GEOMETRY PROCESSING, 2003. **Proceedings ...**. Aachen, Alemanha: EG Digital Library.2003.p 156-165
- 36 LI, H. D.; HARTLEY, R. Conformal spherical representation of 3D genus-zero meshes. **Pattern Recognition**, v. 40, n. 10, p. 2742-2753, 2007.
- 37 PDB- file format documentation. Disponível em:
<<http://www.wwpdb.org/documentation/file-format>, 2003>. Acesso em: 20 jun. 2016.
- 38 MONTE Carlo integration. Disponível em:
<<http://mathworld.wolfram.com/MonteCarloIntegration.html>, 2016>. Acesso em: 20 jun. 2016.
- 39 RITCHIE, D. W. **Parametric protein shape recognition**. 1998. 149p. (Doctor Thesis) – Department of Computing Science & Molecular and Cell Biology, University of Aberdeen, Scotland, 1998.
- 40 THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern recognition**. 4th ed. Cambridge, Massachusetts: Academic Press, 2008.
- 41 APRENDIZAGEM de máquina: aprendizagem não supervisionada. Disponível em:
<<http://www.ppgia.pucpr.br/~alekoe/AM/2005/6-AprendizagemNaoSupervisionada>, 2005>.
Acesso em: 27 mai. 2016.
- 42 BALLESTER, P. J. Ultrafast shape recognition: method and applications. **Future Medicinal Chemistry**, v. 3, n. 1, p. 65-78, 2011.
- 43 SCHREYER, A. M.; BLUNDELL, T. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. **Journal of Cheminformatics**, v. 4, p. 12, 2012.
- 44 BALLESTER, P. J.; RICHARDS, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. **Journal of Computational Chemistry**, v. 28, n. 10, p. 1711-1723, 2007.

- 45 OPEN babel user documentation. Disponível em: http://openbabel.org/wiki/Tutorial:Basic_Usage, 2009>. Acesso em: 22 jun. 2016.
- 46 SQLITE VS POSTGRESQL: a comparison of relational database management systems. Disponível em: <https://www.digitalocean.com/community/tutorials/sqlite-vs-mysql-vs-postgresql-a-comparison-of-relational-database-management-systems>, 2014>. Acesso em: 22 jun. 2016.
- 47 TUTORIAL passo a passo para SQLite em python. Disponível em: <http://www.carlissongaldino.com.br/post/um-tutorial-passo-passo-para-sqlite-em-python>, 2013>. Acesso em: 23 jun. 2016.
- 48 SQLITE python tutorial. Disponível em: <http://zetcode.com/db/sqlitepythontutorial/>, 2014>. Acesso em: 20 jun. 2013.
- 49 FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861-874, 2006.
- 50 TUTORIAL on principal components analysis. Disponível em: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf, 2012>. Acesso em: 25 jun. 2016.
- 51 TUTORIAL on principal component analysis derivation, discussion and singular value decomposition. Disponível em: https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf, 2003>. Acesso em: 25 jun. 2016.
- 52 PRINCIPAL component analysis. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/wics.101/references,2010>>. Acesso em: 05 jun. 2016
- 53 IMPLEMENTING a principal component analysis (PCA) – in python, step by step. Disponível em: http://sebastianraschka.com/Articles/2014_pca_step_by_step.html, 2014>. Acesso em: 10 out. 2014.
- 54 NUMPY documentation: numpy.mean. Disponível em: <http://docs.scipy.org/doc/numpy-1.10.0/reference/generated/numpy.mean.html>, 2015>. Acesso em: 28 jun. 2016.
- 55 NUMPY documentation: numpy.cov. Disponível em: <http://docs.scipy.org/doc/numpy-1.10.1/reference/generated/numpy.cov.html>, 2015>. Acesso em: 28 jun. 2016.
- 56 EIGENVALUE. Disponível em: <http://mathworld.wolfram.com/Eigenvalue.html>, 2015>. Acesso em: 28 jun. 2016.
- 57 CALLIOLI, C. **Álgebra linear e aplicações**. 7a ed. Rio de Janeiro: Atual Editora. 1995.
- 58 NUMPY documentation: numpy.linalg.eig. Disponível em: <http://docs.scipy.org/doc/numpy1.10.0/reference/generated/numpy.linalg.eig.html>, 2015>. Acesso em: 28 jun. 2016.
- 59 NUMPY documentation: numpy.hstack. Disponível em: <http://docs.scipy.org/doc/numpy/reference/generated/numpy.hstack.html>, 2015>. Acesso em: 28 jul. 2016.

60 SCYPY documentation: `scipy.optimize.curve_fit`. Disponível em: <http://docs.scipy.org/doc/scipy-0.17.0/reference/generated/scipy.optimize.curve_fit.html, 2014>. Acesso em: 28 jul. 2016.

61 NUTONIAN : Eureka. Disponível em: <<http://www.nutonian.com/products/eureka/>, 2016>. Acesso em: 28 mai. 2016.

62 DJANGO documentation. Disponível em: <<https://docs.djangoproject.com/en/1.9/>, 2015>. Acesso em: 28 fev. 2016.

63 DISPLAYING structures using JSmol and GLMol. Disponível em: <<http://www.macinchem.org/reviews/jsmol/jsmol.php>, 2011>. Acesso em: 28 jul. 2014.

64 MOLECULAR modeling basics : getting started with JSmol. Disponível em: <<http://molecularmodelingbasics.blogspot.com.br/2013/02/getting-started-with-jsmol.html>, 2013>. Acesso em: 28 jul. 2014.

65 JSMOL on HTML5. Disponível em: <<http://chemapps.stolaf.edu/jmol/jsmol/test2.htm>, 2013>. Acesso em: 28 jul. 2014.

66 PRESS, W. H.; TEUKOLSKY, S. A.; VETTERLING, W. T.; FLANNERY, B. P. **Numerical recipes:** the art of scientific computing. 3rd ed. Cambridge, Massachusetts: Academic Press, 2007.

APÊNDICE A – Cálculo estável dos harmônicos esféricos

Harmônicos esféricos ocorrem em diversos problemas físicos, como toda vez em que uma equação de onda ou equação de Laplace é resolvida utilizando-se separação de variáveis em coordenadas esféricas. A função $Y_{lm}(\theta, \varphi)$, $-l \leq m \leq l$, é definida pelos ângulos θ e φ sobre a superfície da esfera. Os harmônicos esféricos estão relacionados aos polinômios associados de Legendre pela equação abaixo.

$$Y_{lm}^m(\theta, \varphi) = \sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}} P_l^m(\cos \theta) e^{im\varphi} \quad (1)$$

Os polinômios associados de Legendre são determinados em termos dos polinômios ordinários de Legendre, utilizando a relação $x = \cos(\theta)$ os valores podem ser calculados segundo a equação abaixo.

$$P_l^m(x) = (-1)^m (1-x^2)^{m/2} \frac{d^m}{dx^m} P_l(x) \quad (2)$$

Existem diversas formas de calcular os valores dos polinômios associados de Legendre e eles satisfazem diversas relações de recorrência, somente para l , somente para m e para ambos l e m simultaneamente. A maioria das implementações envolvendo m são instáveis e perigosas para procedimentos numéricos. Entretanto, a relação de recorrência para l abaixo é estável e pode ser aplicada no cálculo para valores moderados de l e m .

$$(l-m)P_l^m = x(2l-1)P_{l-1}^m - (l+m-1)P_{l-2}^m \quad (3)$$

Mesmo sendo um cálculo estável os valores dos harmônicos esféricos crescem rapidamente e por isso a função correta ser utilizada para garantir a estabilidade são os polinômios de associados de Legendre normalizados.

$$\widetilde{P}_l^m = \sqrt{\frac{2l+1(l-m)!}{4\pi(l+m)!}} P_l^m \quad (4)$$

Com uso da equação acima temos que a nova relação de recorrência é:

$$\widetilde{P}_l^m = \sqrt{\frac{4l^2 - 1}{l^2 - m^2}} \left[x \widetilde{P}_{l-1}^m - \sqrt{\frac{(l-1)^2 - m^2}{4(l-1)^2 - 1}} \widetilde{P}_{l-2}^m \right] \quad (5)$$

O primeiro passo é o cálculo do harmônicos esférico para $l = m$

$$\widetilde{P}_m^m = (-1)^m \sqrt{\frac{2l+1}{4\pi(2m)!}} (2m-1)!! (1-x^2)^{m/2} \quad (6)$$

A segunda parte do cálculo é feito utilizando a relação de recorrência par $l = m + 1$ e $\widetilde{P}_{m-1}^m = 0$ com isso temos a segunda informação para o cálculo.

$$\widetilde{P}_{m+1}^m = x \sqrt{2m+3} \widetilde{P}_m^m \quad (7)$$

As equações 6 e 7 fornecem as informações iniciais para o processo de cálculo através da função 5. O livro Numeric Recipes – The art of scientific Computing⁶⁶ traz uma implementação em C que foi adaptada para os propósitos desse trabalho sendo implementada em Fortran90, conforme Figura 43.

```

DOUBLE PRECISION FUNCTION plegendre(l, m, x)
  IMPLICIT NONE

  INTEGER, INTENT(IN) :: l, m
  DOUBLE PRECISION, INTENT(IN) :: x

  INTEGER :: i, ll
  DOUBLE PRECISION :: fact, oldfact, pll, pmm, pmmp1, omx2

  DOUBLE PRECISION, PARAMETER :: pi = 3.1415926535897932384626433832795

  pmm = 1.0
  if (m > 0) then
    omx2 = (1.0-x)*(1.0+x)
    fact = 1.0
    do i = 1, m
      pmm = pmm*omx2*fact/(fact + 1.0)
      fact = fact + 2.0
    end do
  end if

  pmm = sqrt((2*m+1)*pmm/(4.0*pi))

  if (iand(m,1) == 1) pmm = -pmm

  if (l == m) then
    plegendre = pmm
    return
  else
    pmmp1 = x*sqrt(2.0*m+3.0)*pmm
    if (l == m+1) then
      plegendre = pmmp1
      return
    else
      oldfact = sqrt(2.0*m+3.0)
      do ll = m+2, l
        fact = sqrt((4.0*ll**2-1.0)/(ll**2-m**2))
        pll = (x*pmmp1-pmm/oldfact)*fact

        pmm = pmmp1
        pmmp1 = pll

        oldfact = fact
      end do
      plegendre = pll
      return
    end if
  end if

  return
END FUNCTION plegendre

```

Figura 43 - Cálculo dos polinômios associados de Legendre.
 Fonte: Elaborada pelo autor.

APÊNDICE B – Classificação das moléculas utilizadas no benchmarking

A Tabela 15 traz a classificação de todas as moléculas utilizadas no benchmarking considerando a métrica L^2 e suas variações.

Tabela 15 - Classificação das moléculas métrica L^2

Molécula	Métrica L^2	Métrica L^2 Ponderada	Métrica L^2 6D
313aces	Bom Classificador	Bom Classificador	Bom Classificador
106hs90a	Bom Classificador	Bom Classificador	Bom Classificador
10aces	Bom Classificador	Bom Classificador	Bom Classificador
112pgh2	Bom Classificador	Bom Classificador	Bom Classificador
120xiap	Aleatório	Mau Classificador	Aleatório
123src	Bom Classificador	Bom Classificador	Bom Classificador
1266hivpr	Bom Classificador	Bom Classificador	Bom Classificador
129ptn1	Bom Classificador	Bom Classificador	Bom Classificador
129reni	Bom Classificador	Bom Classificador	Bom Classificador
13fpps	Bom Classificador	Bom Classificador	Bom Classificador
166try1	Aleatório	Mau Classificador	Aleatório
172mp2k1	Bom Classificador	Bom Classificador	Bom Classificador
174gria2	Bom Classificador	Bom Classificador	Bom Classificador
208ptn1	Bom Classificador	Bom Classificador	Bom Classificador
229mp2k1	Bom Classificador	Bom Classificador	Bom Classificador
237reni	Bom Classificador	Bom Classificador	Bom Classificador
252urok	Bom Classificador	Bom Classificador	Bom Classificador
261hmdh	Bom Classificador	Bom Classificador	Bom Classificador
305urok	Bom Classificador	Bom Classificador	Bom Classificador
314aces	Bom Classificador	Bom Classificador	Bom Classificador
314src	Bom Classificador	Bom Classificador	Bom Classificador
325vgfr2	Bom Classificador	Bom Classificador	Bom Classificador
332cdk2	Aleatório	Bom Classificador	Bom Classificador
342vgfr2	Bom Classificador	Aleatório	Bom Classificador
347casp3	Bom Classificador	Bom Classificador	Bom Classificador
34tysy	Bom classificador	Bom Classificador	Bom Classificador
35kith	Bom Classificador	Bom Classificador	Bom Classificador

(Continua)

(Continuação)

407aa2ar	Bom Classificador	Bom Classificador	Bom Classificador
416egfr	Bom Classificador	Bom Classificador	Bom Classificador
44ptn1	Aleatório	Mau Classificador	Mau Classificador
45esr1	Bom Classificador	Bom Classificador	Bom Classificador
474ada17	Bom Classificador	Bom Classificador	Bom Classificador
488hivpr	Aleatório	Bom Classificador	Aleatório
4andr	Bom Classificador	Bom Classificador	Bom Classificador
52csf1r	Aleatório	Bom Classificador	Bom Classificador
536aces	Bom Classificador	Bom Classificador	Bom Classificador
539esr2	Bom Classificador	Bom Classificador	Bom Classificador
566drd3	Bom Classificador	Bom Classificador	Bom Classificador
599fnta	Bom Classificador	Bom Classificador	Bom Classificador
618cah2	Aleatório	Mau Classificador	Aleatório
640lck	Aleatório	Bom Classificador	Aleatório
6ital	Bom Classificador	Bom Classificador	Bom Classificador
724mmp13	Bom Classificador	Bom Classificador	Bom Classificador
74bace1	Aleatório	Aleatório	Aleatório
781cah2	Bom Classificador	Bom Classificador	Bom Classificador
785dpp4	Bom Classificador	Bom Classificador	Bom Classificador
7def	Mau Classificador	Bom Classificador	Bom Classificador
811hivpr	Bom Classificador	Bom Classificador	Bom Classificador
93hivrt	Bom Classificador	Bom Classificador	Bom Classificador
98hvk4	Bom Classificador	Bom Classificador	Bom Classificador

 Fonte - Elaborada pelo autor.

A Tabela 16 traz a classificação de todas as moléculas utilizadas no benchmarking considerando a métrica L^1 e suas variações.

Tabela 16 - Classificação das moléculas métrica L^1

Molécula	Métrica L^1	Métrica L^1 90D
313aces	Bom Classificador	Bom Classificador
106hs90a	Mau Classificador	Mau Classificador
10aces	Mau Classificador	Bom Classificador
112pgh2	Bom Classificador	Bom Classificador
120xiap	Mau Classificador	Mau Classificador
123src	Mau Classificador	Mau Classificador
1266hivpr	Bom Classificador	Bom Classificador
129ptn1	Mau Classificador	Mau Classificador
129reni	Mau Classificador	Aleatório
13fpps	Bom Classificador	Bom Classificador
166try1	Mau Classificador	Mau Classificador
172mp2k1	Mau Classificador	Mau Classificador
174gria2	Bom Classificador	Bom Classificador
208ptn1	Bom Classificador	Bom Classificador
229mp2k1	Bom Classificador	Bom Classificador
237reni	Bom Classificador	Bom Classificador
252urok	Bom Classificador	Bom Classificador
261hmdh	Bom Classificador	Bom Classificador
305urok	Bom Classificador	Bom Classificador
314aces	Mau Classificador	Mau Classificador
314src	Bom Classificador	Bom Classificador
325vgfr2	Aleatório	Aleatório
332cdk2	Mau Classificador	Mau Classificador
342vgfr2	Bom Classificador	Bom Classificador
347casp3	Aleatório	Mau Classificador
34tysy	Mau Classificador	Mau Classificador
35kith	Bom Classificador	Bom Classificador
407aa2ar	Mau Classificador	Aleatório
(Continua)		

(Continuação)

416egfr	Bom Classificador	Bom Classificador
44ptn1	Mau Classificador	Mau Classificador
45esr1	Aleatório	Mau Classificador
474ada17	Bom Classificador	Bom Classificador
488hivpr	Bom Classificador	Bom Classificador
4andr	Bom Classificador	Bom Classificador
52csf1r	Bom Classificador	Bom Classificador
536aces	Bom Classificador	Bom Classificador
539esr2	Bom Classificador	Bom Classificador
566drd3	Mau Classificador	Mau Classificador
599fnta	Mau Classificador	Mau Classificador
618cah2	Mau Classificador	Mau Classificador
640lck	Bom Classificador	Bom Classificador
6ital	Bom Classificador	Bom Classificador
724mmp13	Bom Classificador	Bom Classificador
74bace1	Mau Classificador	Mau Classificador
781cah2	Bom Classificador	Bom Classificador
785dpp4	Bom Classificador	Bom Classificador
7def	Mau Classificador	Mau Classificador
811hivpr	Mau Classificador	Mau Classificador
93hivrt	Bom Classificador	Bom Classificador
98hxxk4	Bom Classificador	Bom Classificador

 Fonte - Elaborada pelo autor