Universidade Estadual de Campinas

Instituto de Química – Departamento de Físico-Química

TESE DE DOUTORADO

Ferramentas para QSAR-4D dependente de receptores: Aplicação em uma série de inibidores da tripanotiona redutase do *T. Cruzi*

Euzébio Guimarães Barbosa

Orientadora: Profa. Dra. Márcia Miguel Castro Ferreira

Campinas - 2011

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO INSTITUTO DE QUÍMICA DA UNICAMP

B234f

Barbosa, Euzébio Guimarães.
Ferramentas para QSAR-4D dependente de receptores: aplicação em uma série de inibidores da tripanotiona redutase do T. cruzi / Euzébio Guimarães Barbosa. -- Campinas, SP: [s.n], 2011.

Orientador: Profa. Dra. Márcia Miguel Castro Ferreira.

Doutorado - Universidade Estadual de Campinas, Instituto de Química.

1. QSAR-4D. 2. Descritores de campo de interação. 3. Doença de Chagas. 4. LQTA-QSAR. I. Ferreira, Márcia Miguel Castro. II. Universidade Estadual de Campinas. Instituto de Química. III. Título.

Título em inglês: Tool for receptor dependent 4D-QSAR applied to set of T. cruzi trypanothione reductase inhibitors

Palavras-chaves em inglês: 4D-QSAR, Interaction fields descriptors, Chagas' disease, LQTA-QSAR

Área de concentração: Físico-Química

Titulação: Doutor em Ciências

Banca examinadora: Profa. Dra. Márcia Miguel Castro Ferreira (orientador), Profa. Dra. Nelilma Correia Romeiro (NUPEM-UFRJ), Prof. Dr. Eduardo Borges de Melo (CCMF-UNIOESTE), Profa. Dra. Ljubica Tasic (IQ-UNICAMP), Prof. Dr. Nelson Henrique Morgon (IQ-UNICAMP)

Data de defesa: 12/07/2011

"*Who are we? We find that we live on an insignificant planet of a humdrum star lost in a galaxy tucked away in some forgotten corner of a universe in which there are far more galaxies than people.*"

Carl Sagan

"... Pessoas adoram os produtos da ciência, mas odeiam seus métodos ..."

Carl Sagan no último episódio da série Cosmos

Dedico essa tese

A minha amada esposa Dona Lilia Basilio de Caland

Em memória da minha querida segunda mãe, a Naza, que se foi.

Aos meus amados pais Olidia e Waldermar.

Aos meus queridos irmãos, Oswaldo, Waldemar, João e Mariozinho.

E aos grandes amigos do Nordeste

Agradecimentos

Agradeço a Capes pelo auxílio financeiro.

A professora Márcia pelo inestimável conhecimento que me foi repassado.

A todos os colegas do grupo LQTA pela ajuda em todos os momentos de dificuldade.

E a querida amiga Kerly Pasqualolo por todo o apoio que me deu nesses quatro anos.

Currículo Vitae
Euzébio Guimarães Barbosa

Títulos e Formação Acadêmica

2007 **Mestrado em Química Orgânica**

Universidade Federal de Mato Grosso do Sul, UFMS, Brasil.

*Título: "*Dinâmica Molecular de Análogos do Tamoxifeno e Síntese de Análogos da Combretastatina A-4

Orientador: Prof. Dr. Denis Pires de Lima.

Bolsista Fundação de Apoio ao Desenvolvimento do Ensino Ciência e Tecnologia do Estado de Mato Grosso do Sul.

2003 **Graduação em Farmácia**

Universidade Federal de Mato Grosso do Sul, UFMS, Brasil.

Atividades de Pesquisa

Publicação de trabalho científico em periódico com ISSN

Barbosa, E.G.; Ferreira, M.M.C. Digital filters for molecular interaction field descriptors. **Molecular Informatics** (ISSN: 1868-1751), submetido e em fase final de aceitação, 2011.

Martins, J.P.A.; Barbosa, E.G.; Pasqualoto, K. F.M.; Ferreira, M.M.C. LQTA-QSAR: A New QSAR-4D Methodology. **Journal of Chemical Information and Modeling** (*ISSN: 1549-9596)*, v. 49, p. 1428-1436, 2009.

Barbosa, E.G.; Bega, L; Beatriz, A ; Sarkar, T; Hamel, E ; Do Amaral, M; De Lima, D. A diaryl sulfide, sulfoxide, and sulfone bearing structural similarities to combretastatin A-4. **European J. Med. Chem.** (*ISSN:* 0223-5234) p. 2685-2688, 2008.

Resumos publicados em anais de congresso internacional

Barbosa, E.G.; Pasqualoto, K. F.M.; Ferreira, M.M.C. **Exploring protein flexibility in QSAR-4D: Application to a set of Trypanothione Reductase Inhibitors**. Em: 18th EURO-QSAR. 2010, Rhodes, Grécia.

Carvalho-Fresqui, M.A; Barbosa, E.G; Ferreira, M.M.C. **Receptor Dependent 3D-LQTA-QSAR of substituted Amphetamines**. Em: 18th EURO-QSAR. 2010, Rhodes, Grécia.

Martins, J.P.A.; Barbosa, E.G.; Pasqualoto, K. F.M.; Ferreira, M.M.C. **LQTA-QSAR: A New QSAR-4D Methodology**. Em: 50th International Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources. 2009, Istambul, Turquia.

Pereira, F.S.; Barbosa, E.G.; Ferreira, M.M.C. **4D-LQTA-QSAR Analysis of a Set of Antimalarial Compound**. Em: 7th International Congress of Pharmaceutical Science (CIFARP). 2009. Ribeirão Preto, SP, Brasil.

Resumos publicados em anais de congresso nacional

Barbosa, E.G.; Vazquez, P.A.M.; Ferreira, M.M.C. **Desempenho do método TD-M052X no cálculo de energias verticais de excitação eletrônica comparado a outros métodos TD-DFT, CCSD e CASPT2**. Em: 32ª Reunião Anual da SBQ. 2009. Fortaleza – CE.

Barbosa, E.G.; Pasqualoto, K. F.M.; Ferreira, M.M.C. **Aplicação de docking e dinâmica molecular para elucidar o modo de interação de fenotiazínicos derivados de arilaquilamônio com a tripanotiona redutase do Trypanosoma cruzi.** Em: 32ª Reunião Anual da SBQ. 2009. Fortaleza – CE.

Martins, J.P.A ; Barbosa, E.G.; Pasqualoto, K.F.M ; Ferreira, M.M.C. **LQTAgrid: an open source package to generate QSAR-4D descriptors.** Em: 4th Brazilian Symposium on Medicinal Chemistry. 2008. Porto de Galinhas – PE.

Barbosa, E. G.; Sarkar, T.; Bega L.A.S ; Hamel, E.; Amaral, M.S. ; Beatriz, A. ; De Lima, D.P. **Sulfur as a linker in Combretastatin A-4 analogues.** Em: 4th Brazilian Symposium on Medicinal Chemistry. 2008. Porto de Galinhas – PE.

Barbosa, E.G. ; Silva, D.B. ; Amaral, M. S. ; De Lima, D.P.; Miglio, H.S.; Garcez, W.S.; Siqueira, J.M. **Preferências conformacionais do alcalóide duguetina e implicações química.** Em: 29.o Reunião Anual da Sociedade Brasileira de Química. 2006. Águas de Lindóia - SP.

Softwares

Martins, J.P.A ; Barbosa, E.G ; Ferreira, M.M.C. **LQTAgrid All force field (JAVA)**. 2009.

Martins, J.P.A.; Barbosa, E.G.; Pasqualoto, K.F.M.; Ferreira, M.M.C. **LQTA-QSAR**. **(JAVA)** 2008.

Barbosa, E.G ; Ferreira, M.M.C. **CDDA, filter for Molecular Interaction Descriptors (MATLAB)**. 2009.

Atividades de Extensão

Extensão: **Mini-curso: LQTA-QSAR: Pacote de Programas de acesso livre para QSAR-4D.** Carga horária: 10 h. Em: Instituto de Química - UNICAMP. 2009. Campinas - SP.

Extensão: **Mini-curso: Estrutura Tridimensional d Proteínas.** Carga horária: 8 h. Em: II Semana Nacional de Ciência e Tecnologia. 2006. Campo Grande - MS.

Extensão: **Mini-curso: Manipulação de Software de Modelagem Molecular.** Carga horária: 10 h. Em: II Semana Nacional de Ciência e Tecnologia. 2005. Campo Grande - MS.

Atividades de Docência

Prestação de estágio, na função de magistério, em estabelecimento de ensino superior, devidamente autorizado ou reconhecido.

Programa de Estágio de docência – PED, nível B: **Química Orgânica Experimental II.** Supervisor: Prof. Dr. Paulo Cesar Muniz de Lacerda Miranda. Carga horária: 8 h semanais. 2009. Instituto de Química, UNICAMP, Campinas - SP.

Programa de Estágio de docência – PED, nível B: **Química Orgânica II (Engenharia Química).** Supervisor: Prof. Dr. Francisco de Assis Machado Reis. Carga horária: 8 h semanais. 2009. Instituto de Química, UNICAMP, Campinas - SP.

Exercício do magistério em nível superior

Professor Voluntário: **Química Geral (Bacharelado em Física).** Carga horária: 4 h semanais. 2005. Universidade Federal do Mato Grosso do Sul. Campo Grande - SP.

Resumo

LQTA-QSAR é uma metodologia computacional para QSAR-4D desenvolvida pelo Laboratório de Quimiometria Teórica e Aplicada implementada em um software de acesso livre. O método permite considerar simultaneamente as vantagens da representação molecular multiconformacional e os descritores de campos de interação. Esta tese apresenta a evolução da proposta inicial da metodologia LQTA-QSAR independente de receptores para uma abordagem dependente de receptores. Sua aplicação é demonstrada na construção de modelos de QSAR-4D para a previsão da atividade inibitória de compostos fenotiazínicos da enzima tripanotiona redutase. Foi obtido um modelo com bom poder de previsão (Q^2_{prev} = 0,78) e com descritores de fácil interpretação. Tal modelo pode ser usado para a proposição de compostos que poderão vir a ser usados para o tratamento da doença de chagas.

Para a filtragem e seleção de descritores foi necessário o desenvolvimento de um protocolo completamente distinto daquele disponível na literatura. Foi proposto um procedimento automatizado para identificar e eliminar descritores irrelevantes quando a correlação e um algoritmo que elimina descritores com distribuição díspar em relação à atividade biológica. Foram introduzidos também testes de validação de modelos QSAR nunca antes usados para modelos que utilizam descritores de campo de interação. O protocolo completo foi testado em três conjuntos de dados e os modelos obtidos tiveram capacidade de previsão superior aos da literatura. Os modelos mostraram ser bastante simples e robustos quando submetidos aos testes *leave-N-out* e *y-randomization*.

Abstract

The New Receptor-Dependent LQTA-QSAR approach is proposed as a new 4D-QSAR method. The RD-LQTA-QSAR is an evolution to the receptor independent LQTA-QSAR. This approach make use of the simulation package GROMACS to carry out molecular dynamics simulations and generate a conformational ensemble profile for each compound. Such ensemble is used to build molecular interaction field based QSAR models, as in CoMFA. To verify the usefulness of the methodology it was chosen some phenothiazine derivatives that are specific competitive *T. cruzi* trypanothione reductase inhibitors. Using a combination of molecular docking and molecular dynamics simulations the binding mode of 38 phenotiazine derivatives was evaluated in a simulated induced fit approach. The ligands' alignment, necessary to the methodology, was performed using both ligand and binding site atoms hereafter enabling unbiased alignment. The obtained models were extensively validated by Leave-*N*-out cross-validation and **y**-randomization techniques to test robustness and absence of chance correlation. The final model presented Q^2 LOO of 0.87 and R^2 of 0.92 and suitable external prediction = 0.78. It is possible to use the obtained adapted binding site of to perform virtual screening and ligand structures based design, as well as using models descriptors to design new inhibitors.

In the process of QSAR modeling, the relevance of correlation and distribution profiles were tested in order to improve prediction power. A set of tools to filter descriptors prior to variable selection and a protocol for molecular interaction field descriptors selection and models validation are proposed. The algorithms and protocols presents are quite simple to apply and enable a different and powerful way to build LQTA-QSAR models.

Índice

Índice de Figuras

Índice de Tabelas

Índice de Abreviaturas

|*r*|: correlação absoluta de Pearson.

AFF: *All force field.* Opção do LQTAgrid multicampo de força.

AM1-BCC: *Austin Model 1 with simple additive bond charge corrections*. Método de cargas atômicas baseado no potencial eletrostático.

AMBER*:* Nome do campo de força para o programa AMBER.

cc-PVDz: Função de base de *Dunning* consistente com métodos de correlação.

CDDA: *Comparative Distribuition Algoritm,* Algoritmo de distribuição comparativa.

CHELPG: *CHarges from Electrostatic Potentials using a Grid based method,* Método de cargas derivadas do potencial eletrostático usando grades.

CoMFA: *Comparative Molecular Field Analysis,* análise comparativa de campos moleculares.

e_j : Parâmetro distribuição comparativa obtida com CDDA.

GAFF: *General AMBER force field,* Campo de força generalizado para o AMBER.

GLU: Resíduo de ácido glutâmico.

GOLPE: G*enerating Optimal Linear PLS Estimations*. Algoritmo que gera estimativas ótimas para modelos lineares PLS.

GR: Glutationa Redutase.

GROMOS: Nome do campo de força usado no programa GROMACS.

GSH: Glutationa.

GSSG: Glutationa dissulfeto.

HOMO: *Highest Occupied Molecular Orbital*, Orbital Molecular Ocupado mais alto.

LJ: Descritores de van der Waals calculados pelo potencial de Lennard-Jones.

LNO: Validação cruzada *Leave-N-out* (Deixe-*N*-fora)

log P: Logaritmo na base 10 do coeficiente de partição octanol/água.

LOO: Validação cruzada *Leave-one-out* (Deixe-um-fora).

LQTA: Laboratório de Quimiometria Teórica e Aplicada.

LQTAgrid: Programa integrante ao pacote LQTA-QSAR para o cálculo dos descritores de campo de interação.

LQTA-QSAR-DR. Métodologia LQTA-QSAR dependente de receptores.

LQTA-QSAR-IR. Métodologia LQTA-QSAR independente de receptores.

LV: *Latent Variables, Variáveis Latentes.*

M05-2X: Nome do funcional da densidade que leva em conta correlações de energia de médio alcance para moléculas orgânicas.

MET. Resíduo de ácido metionina.

MIF: *Molecular Interaction Fields,* campos de interação molecular.

MLR: *Multiple Linear Regression,* regressão linear múltipla.

OPS: Ordered Predition Selection. Nome do algoritmo para selecionar descritores ordenando variáveis informativas.

PAC: Perfil de Amostragem Conformacional.

PCR: *Principal Component Regression,* Regressão em Componentes Principais.

PHE. Resíduo de ácido fenilalanina.

PLS: *Partial Least Squares*, Quadrados Mínimos Parciais.

PME: *Particle Mesh Ewald*

PSA: *Polar Surface Area,* Área molecular polar superficial.

$\mathbf{Q^2_{LNO}}$: Coeficientes de correlação da validação cruzada *Leave-N-out.*

$\boldsymbol{Q^2_{Loo}}$: Coeficiente de correlação da validação cruzada *Leave-one-out.*

QQ: Descritores calculados com o potencial de Coulomb.

QSAR: *Quantitative Structure-Activity Relationships*, relações quantitativas entre a estrutura química e a atividade.

R^2: Coeficiente de correlação de determinação múltipla.

SEC: *Standard error of calibration,* erro padrão de calibração.

SEV: *Standard error of validation,* erro padrão da validação cruzada.

$T[S]_2$: Tripanotiona oxidada

$T[SH]_2$: Tripanotiona.

TR: Tripanotiona Redutase.

TRP. Resíduo de Triptofano.

TYR. Resíduo de tirosina.

Capítulo I

Introdução

e

Objetivos

I.1 Relações quantitativas entre estrutura e atividade biológica

As relações quantitativas entre a estrutura química e a atividade biológica (QSAR, *Quantitative Structure-Activity Relationships*) têm como principal objetivo a previsão acurada de resultados de testes biológicos usando uma abordagem matemática, automatizada e computacional. Estas relações podem ser usadas para prever a atividade de compostos que nem mesmo foram sintetizados, auxiliando a economia de tempo e recursos financeiros, pois, orienta o planejamento de estruturas químicas para aquelas com maior probabilidade de ter melhor atividade biológica [1].

Os estudos de QSAR começaram a ser amplamente utilizados quando se tornaram disponíveis os computadores e as ferramentas matemáticas necessárias. É atribuído aos pesquisadores Corwin Hansch e Toshio Fujita [2] a introdução dos métodos quantitativos em química medicinal nos anos de 1960, que utilizaram o logaritmo do coeficiente de partição octanol/água (log P) para descrever o processo de transporte de fármacos e sua contribuição crucial para a atividade biológica [3]. Esta propriedade molecular é um dos principais indicativos de lipofilicidade das moléculas, e é frequentemente considerada como um fator determinante para a absorção intestinal, absorção pelo sistema nervoso central, volume de distribuição e associação às proteínas [4, 5, 6] (**Figura 1**).

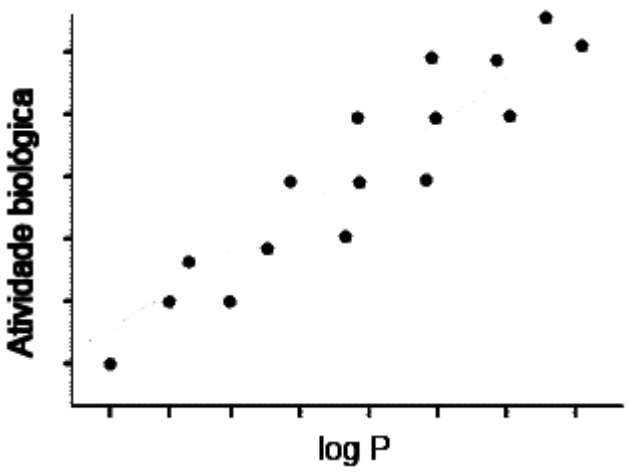


Figura 1. Gráfico mostrando a relação entre os valores de atividade biológica e valores calculados de log P.

Observando a **Figura 1** é fácil perceber que se pode fazer a regressão univariada dos valores de log P para se obter uma expressão que relaciona tais valores à atividade biológica (y) (1):

$$y = a_0 + a_1 \log P \tag{1}$$

onde, a_0 é o intercepto em no eixo y e a_1 é o coeficiente angular da reta. Como a atividade biológica e log P são conhecidos é possível calcular a_0 e a_1 por meio de (2)

$$\mathbf{a} \approx \mathbf{X}^{+} \bullet \mathbf{y}$$
$$\mathbf{a} \approx (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T} \bullet \mathbf{y} \tag{2}$$

onde **y** é o vetor contendo os valores de atividade biológica obtidos experimentalmente, **a** é o vetor de regressão que contém os dois valores a_0 e a_1 e $\mathbf{X}^{+}$ é a pseudoinversa da matriz **X.** A matriz **X** contém a primeira coluna constituída de valores "1" e a segunda coluna com os valores de log P para cada molécula da série. A **Figura 2** ilustra tais matrizes. Os valores de log P para todas as moléculas arranjadas na forma de um vetor numérico o qual é chamado de descritor.

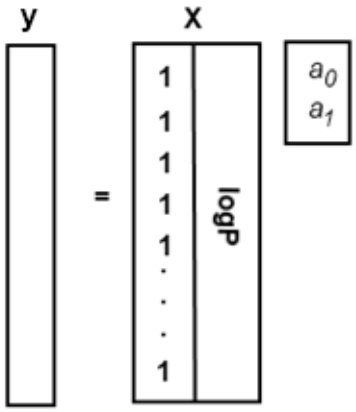


Figura 2. Representação esquemática das matrizes numéricas usadas para regressão univariada entre os valores de atividade biológica (y) e o descritor log P. Os valores a_0 e a_1 são os coeficientes de regressão mencionados na equação (1).

Quando são usados mais de um tipo de descritor é obtido um modelo multivariado. A área polar superficial (PSA), as constantes de substituição de Hammett (σ), refratividade molar, polarizabilidade, energia de orbitais moleculares, dentre outros, são exemplos de descritores comumente utilizados. Porém é necessária agora a utilização da regressão linear múltipla (MLR, *multiple linear regression*) para a calibração da atividade biológica. A matriz **X** terá agora maior número de colunas (**Figura 3**), uma para cada descritor, e o vetor de regressão **a** é obtido também pela expressão (2).

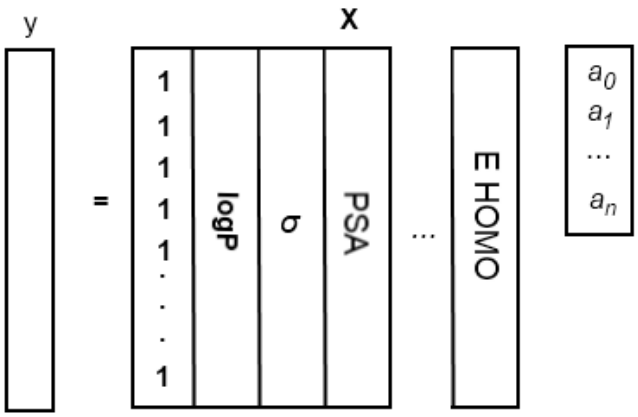


Figura 3. Representação esquemática das matrizes numéricas usadas para regressão linear multivariada (MLR) entre os valores de atividade biológica (y) e os descritores hipotéticos log P, σ, PSA e E HOMO. Os valores a_0 a a_n formam o vetor de regressão, **X** é a matriz que contém todos os descritores.

No final desse processo é obtida uma equação matemática que é uma combinação linear dos diversos descritores, como mostrado na expressão (3), que representa um modelo QSAR.

$$y = a_0 + a_1 \log P + a_2 \sigma + a_3 PSA + ... + a_n E_{HOMO} \quad (3)$$

O método MLR [7] possui problemas quanto à instabilidade numérica dos resultados quando há descritores muito correlacionados entre si e há ainda a impossibilidade de se usar mais descritores do que amostras, e assim vem sendo cada vez menos usado para a construção de modelos QSAR. Atualmente além do MLR, são muito utilizados métodos como a regressão em componentes principais (PCR, *principal component regression*) [8] e os quadrados mínimos parciais (PLS, *partial least squares*) [9]. Tais abordagens utilizam da

transformação da matriz de descritores **X** em uma nova matriz **X'** que pode ser apresentada em um espaço de menor dimensão sem perder muita informação. Tais transformações eliminam os problemas do método MLR por meio da compressão e ortogonalização dos dados, onde novas variáveis, em menor número, são obtidas por combinações lineares das variáveis originais [7].

No método PCR [8] as novas variáveis são chamadas componentes principais (PC, *principal componets*) e no método PLS, são denominadas variáveis latentes (LV: *latent variables*). Enquanto no método PCR as PCs são construídas primeiro, e depois é realizada a regressão de y, no método PLS as LVs são obtidas utilizando-se y no processo de decomposição. O método PLS se tornou muito mais popular, pois, como a covariância do novo sistema de eixos e y é maximizada, espera-se a obtenção de resultados superiores àqueles que seriam obtidos com PCR [7].

Para realizar um estudo de QSAR são necessários dados de atividade biológica de boa qualidade e confiabilidade. São necessários ainda descritores moleculares calculados ou experimentais que, no melhor cenário, possam ser interpretados e ajudar na proposta de novas estruturas. O método de regressão (MLR, PCR, PLS, etc.) deverá relacionar os descritores à atividade biológica. Os dados disponíveis devem ser divididos no mínimo em dois subconjuntos: um denominado conjunto de treinamento, usado para construir o modelo QSAR e outro para certificar se tal modelo realmente terá capacidade de realizar previsões úteis. Este procedimento é chamado de *validação externa*, [10] (**Figura 4).**

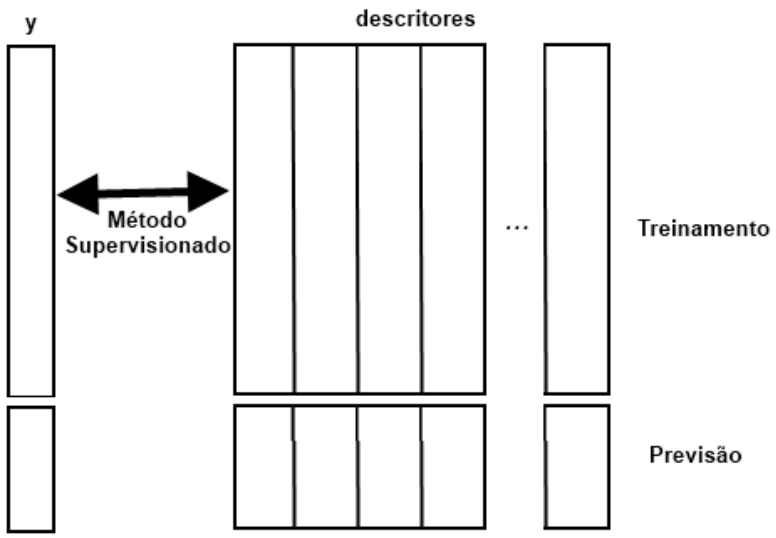


Figura 4. Representação da separação dos dados no vetor de atividade biológica (y) e da matriz de descritores. Uma parte dos dados é usada para construir o modelo QSAR utilizando um método de regressão supervisionado. O modelo obtido é testado para prever a atividade.

É essencial que um modelo de QSAR faça uma previsão confiável da atividade para novas moléculas. Para garantir isso o modelo precisa ser validado. A validação encontra-se atualmente em amplo debate pela comunidade científica e órgãos regulatórios que se utilizam destes modelos, e há uma busca por um protocolo de validação que amplie a aceitação das previsões feitas por tais modelos por parte da comunidade científica [11-13]. As validações que estão sendo propostas não ficariam apenas restritas ao conjunto de amostras separadas para validação externa, mas também para garantir que o modelo por si só tenha qualidade estatística adequada para tal tarefa [12].

Validação cruzada Leave-one-out.

A validação cruzada *leave-one-out* (LOO) é o procedimento mais simples e extremamente necessário para a validação de modelos QSAR. Esta validação, também denominada validação interna, consiste em excluir uma amostra de cada vez do conjunto de treinamento, construir um modelo sem tal amostra, e então realizar a previsão da atividade para esta amostra deixada de fora. Em um conjunto contendo *I* amostras o procedimento é feito *I* vezes. A diferença entre o valor experimental e o estimado para as amostras retiradas são usados para calcular o erro padrão da validação cruzada (*SEV*, *standart error of*

validation) e o coeficiente de correlação da validação LOO (Q^2_{LOO}). O modelo com todas as *I* amostras é expresso em termos do coeficiente de correlação de determinação múltipla (R^2) e o erro padrão de calibração (*SEC*, *standart error of calibration*). Q^2 *LOO* tende a apresentar valores menores que R^2, pois o primeiro é obtido por uma "perturbação" do modelo feita com todas as amostras. Se a diferença entre os valores de Q^2_{LOO} e R^2 for de 0,3 há indícios de sobreajuste do modelo [12, 14] (**Tabela 1**).

Tabela 1. Parâmetros estatísticos básicos para modelos de regressão.

Parâmetro	Definição
Coeficiente de correlação de validação cruzada LOO e LNO	$Q^2 = 1 - \frac{\sum_{i=1}^{I} (y_i - y_{vi})^2}{\sum_{i=1}^{I} (y_i - \langle \mathbf{y} \rangle)^2}$
Coeficiente de correlação de determinação múltipla	$R^2 = 1 - \frac{\sum_{i}^{I} (y_i - y_{ci})^2}{\sum_{i}^{I} (y_i - \langle \mathbf{y} \rangle)^2}$
Coeficiente de correlação de validação externa	$Q^2_{pred} = 1 - \frac{\sum_{i}^{I} (y_i - y_{pi})^2}{\sum_{i} (y_i - \langle \mathbf{y} \rangle)^2}$
Erro padrão de previsão (*SEP**)	$SEP = \sqrt{\frac{\sum_{i}^{I} (y_i - y_{pi})^2}{I - k - 1}}$

I é o número de amostras (conjunto de treinamento ou validação), i é o índice do somatório e também o índice da i-ésima amostra (i = 1,2,...,I); y - valor experimental de $\mathbf{y}$; y_c *- valor de atividade calculado para o conjunto de treinamento;* y_v *- valor de atividade calculado para a validação interna;* y_p *- valor de atividade previsto para o conjunto externo; <y> é o valor médio de atividade para o conjunto de treinamento e k – número de variáveis latentes no modelo.*

** Os erros padrão de validação (SEV) e calibração (SEC) são calculados de forma similar ao SEP [12].*

Validação cruzada Leave-N-out.

A validação cruzada *Leave-N-out* (LNO) [12, 15, 16] conhecida também com *leave-many-out*, é recomendada para testar a robustez do modelo. As *I* amostras do conjunto de treinamento são divididas em blocos consecutivos de *N* amostras, onde as primeiras *N* amostras (arranjadas aleatoriamente) definem o primeiro bloco e assim sucessivamente. Esse teste é baseado no mesmo princípio do LOO, onde cada bloco é excluído uma vez, um novo modelo é construído sem estas amostras, e os valores de atividade biológica são previstos para o bloco de amostras separadas para validação interna. O LNO é feito para N = 2, 3,... , N, e os coeficientes de correlação da validação cruzada LNO (Q^2_{LNO}) são calculados da mesma maneira que para Q^2_{LOO}.

Ao contrário do LOO, o teste LNO é sensível à ordem em que as amostras estão arranjadas na matriz usada para a calibração. Por exemplo, na validação *leave-two-out* apenas uma pequena combinação das possibilidades de duas amostras para validação interna é feita conforme estas amostras são retiradas consecutivamente. Para evitar problemas com essa variação sistemática é necessário rearranjar as linhas da matriz de dados para que outras combinações de duas amostras separadas para previsão sejam contempladas. O valor de N deve representar também uma fração significativa das amostras do conjunto de treinamento (por volta de 20 – 30%). Um bom modelo deve apresentar Q^2_{LNO} muito próximo ao Q^2_{LOO} para todas as N amostras retiradas. A literatura recomenda um valor máximo de desvio de 0,05, ou seja, se o valor de Q^2_{LOO} for 0,71 os valores de Q^2_{LNO} não devem ser menores que 0,66. Para expressar o resultado do teste LNO é feito um gráfico que mostra o desvio máximo entre Q^2_{LOO} e Q^2_{LNO}. Devem também ser expressos dois desvios padrão ao redor do valor médio para a diferença entre o Q^2_{LOO} e Q^2_{LNO} [12].

y-Randomization.

O propósito do teste **y**-*randomization* [12, 14, 15, 17] é detectar e quantificar correlações ao acaso entre a variável dependente e os descritores. Nesse contexto, o termo correlação por acaso (*chance correlation*) significa que o modelo pode conter descritores que são estatisticamente bem correlacionados com a atividade, mas na realidade não decodificam relação casuísticas, pois estes não estão realmente relacionados como o

mecanismo de ação. O teste **y**-*randomization* consiste em construir diversos modelos com os descritores originais, porém trocando a ordem dos valores de y aleatoriamente. Os modelos obtidos nestas condições devem ter baixa qualidade e não ter nenhum sentido real. Os parâmetros R^2 e Q^2_{LOO} podem ser calculados para esses modelos com y aleatório e devem ter valores baixos (R^2_{yrand} e Q^2_{yrand}). Eriksson e Wold [18] propuseram limites para Q^2_{yrand} e R^2_{yrand} onde:

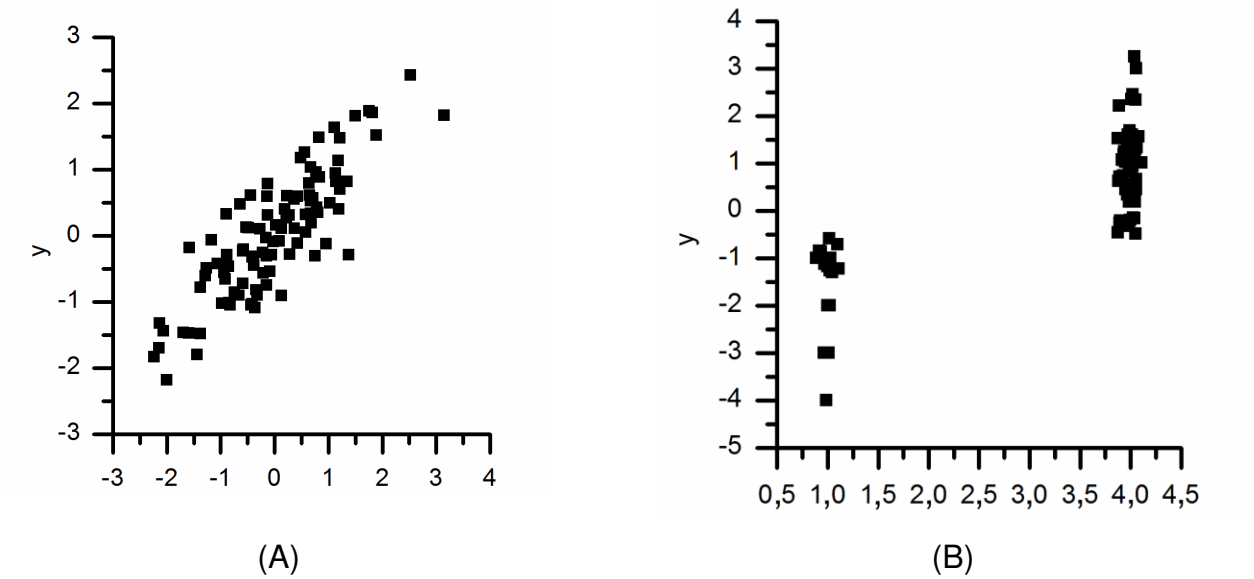
- $Q^2_{yrand} < 0,2$ $R^2_{yrand} < 0,2$ → Não há correlação por acaso
- $Q^2_{yrand} < 0,2$ $R^2_{yrand} < 0,3$ → Correlação por acaso é desprezível
- $Q^2_{yrand} < 0,3$ $R^2_{yrand} < 0,4$ → Correlação por acaso tolerável
- $Q^2_{yrand} > 0,3$ $R^2_{yrand} > 0,4$ → Clara evidencia de correlação por acaso

Existe outra abordagem para expressar os resultados do teste **y**-*randomization*, baseada no coeficiente de correlação absoluta de Pearson, $|r|$, entre o vetor original y e os vetores aleatórios [18]. Podem ser feitos dois gráficos um no eixo y com os valores de Q^2_{yrand} e outro R^2_{yrand} sendo que os dois gráficos contêm no eixo x os valores de $|r|$. No mesmo gráfico também é mostrado o ponto que contém os valores de R^2 e Q^2_{LOO}. É feita então uma regressão linear com os valores e se o modelo foi obtido com descritores apresentam correlação ao acaso, os interceptos desses gráficos são $a_Q > 0,05$ e $a_R > 0,3$. Estes interceptos são medidas da correlação ao acaso. Podem ser realizada de 10 até 1000 randomizações para avaliar se o modelo tem ou não problemas com correlação por chance [12].

Correlação e distribuição dos descritores.

A análise dos descritores que constitui o modelo também pode ser usada como um parâmetro de validação para certificar que o modelo possui significado químico. Descritores que têm baixa correlação com y não fornecem informação relevante para a construção de um modelo e devem ser evitados em QSAR [19]. A **Figura 5** mostra dois gráficos bivariados que ilustram dois descritores distintos expressos graficamente versus y. Quando os descritores são derivados de funções contínuas estes devem apresentar uma resposta uniforme e linear

às variações de y [**Figura 5** (A)]. Em outras palavras deve-se observar uma **distribuição uniforme** [19]. Os descritores derivados da mesma função, mas que se apresentam distribuídos como o descritor (B), devem ser eliminados. Portanto, a maneira com que um descritor se distribui pode ser considerada também um parâmetro de qualidade do mesmo.



(A) (B)

Figura 5. Gráficos bivariados de atividade biológica **y** e dois descritores. O descritor (A) apresenta uma distribuição uniforme ao contrário do descritor (B).

O problema da mudança de sinal.

Quando se obtém uma equação de QSAR é comum basear a interpretação dos descritores pelo sinal do coeficiente de regressão. A expressão (4) mostra valores fictícios para o vetor de regressão (**a**).

$$y = 20 + 5{,}1\log P - 1{,}2PSA \qquad (4)$$

Naturalmente, analisando a equação acima vemos que o valor do coeficiente de regressão para o descritor log P é positivo, ou seja, quanto maior este valor, maior será atividade calculada, se faz uma análise diametralmente oposta para o descritor PSA. O que ocorre muitas vezes é que o pesquisador não recorre à informação da correlação dos descritores log P e PSA com y, que logicamente devem ter o mesmo padrão, ou seja, o sinal da correlação e do correspondente valor de coeficiente de regressão deve ser o mesmo.

Porém, isso muitas vezes não é feito para avaliar os modelos QSAR sendo um fato vastamente negligenciado pela literatura. Kiralj e Ferreira [19] foram os primeiros a questionar a importância do problema da mudança de sinal. Tal problema é de fácil detecção e pode ser eliminado pela seleção de variáveis feita pela verificação da discordância dos sinais manualmente.

I.2 QSAR-3D

Os descritores usados em QSAR podem ser divididos em diversas categorias e podem ser calculados a partir da representação 1D, 2D ou 3D da estrutura molecular. Existem propriedades que podem ser obtidas da molécula inteira ou somente de substituintes em posições específicas. Podem ainda ser derivadas da contagem de, por exemplo, o número de oxigênios, grupos funcionais específicos, grupos que aceitam ou doam ligação de hidrogênio, etc.

O uso de representações tridimensionais das moléculas é a forma mais utilizada em estudo de QSAR para a obtenção de descritores, pois fornecem de uma só vez informações sobre a composição, topologia e formato das moléculas, assim como propriedades eletrônicas. Uma das metodologias interessantes que usam a informação molecular em uma determinada conformação é o chamado método de QSAR-3D.

No método de QSAR-3D de análise comparativa de campos moleculares (**CoMFA**, *Comparative Molecular Field Analisys*) [20] é usado um conjunto de moléculas que podem ser estruturalmente similares ou ter apenas o farmacóforo em comum. Cada molécula, em uma conformação, deve estar alinhada com as demais, e estar posicionada em uma grade virtual tridimensional (**Figura 6**). A energia de interação de um átomo de prova, ou fragmento molecular (sonda), com cada molécula é calculada nos pontos da grade e constituirá os descritores conhecidos como *de campo de interação molecular* (MIF, *molecular interacion fields*) [21, 22].

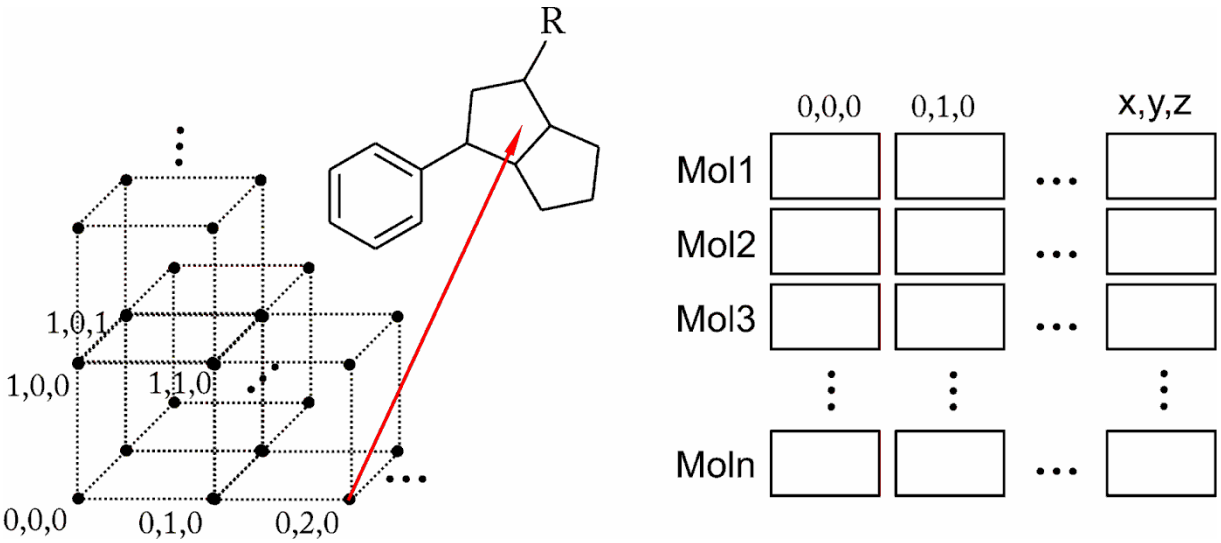


Figura 6. Representação da grade virtual 3D utilizada no método CoMFA e a matriz de descritores MIF. Cada descritor MIF será um vetor de energia de interação em cada ponto da grade para cada molécula do conjunto. Na série hipotética da figura, as moléculas estão alinhadas pelo esqueleto comum tendo diferenciações apenas no grupo R. As coordenadas da grade completa não são totalmente representadas para melhor visualização.

As energias de interação mais empregadas são as de van der Waals e Coulomb, calculadas por funções comuns aos campos de força moleculares, como o potencial de Lennard-Jones (5) e o potencial de Coulomb (6), respectivamente.

$$E_{QQ} = \sum_{k=1}^{K} \frac{q_{sonda} q_k}{D r_{sonda\,k}} \tag{5}$$

$$E_{vdW} = \sum_{k=1}^{K} \left(A r_{sonda\,k}^{-12} - C r_{sonda\,k}^{-6} \right) \tag{6}$$

onde E_{QQ} é a energia de interação que irá compor os descritores eletrostáticos (QQ), q_{sonda} é a carga da sonda, q_k é a carga parcial do átomo k da molécula, D é a constante dielétrica, e $r_{sonda\ k}$ é a distancia da sonda dentro da grade virtual até o átomo k da molécula. E_{vdW} é a energia de interação de van der Waals calculada pelo potencial de Lennard-Jones (6-12) que forma os descritores estéricos (VdW). Os termos A e C são constantes dependentes do raio de van der Waals parametrizados para o campo de força usado no cálculo da energia.

O número de descritores gerados num estudo de QSAR-3D é muito mais elevado em relação ao QSAR tradicional, até centenas de milhares de descritores. Os descritores totais são chamados de *campo*. Porém, nem todos os pontos da grade são igualmente relevantes para explicar a atividade biológica, sendo necessária a seleção de variáveis. Através desta seleção é possível construir um modelo interpretável do campo de interação. Diversos métodos foram desenvolvidos para esse propósito e são amplamente usados em QSAR-3D [23]. O método de seleção de variáveis inspirados no planejamento experimental D-ótimo, GOLPE (*generating optimal linear PLS estimastion*) [24], foi um dos primeiros a serem usados para tratar os descritores CoMFA. Porém existem outras metodologias com o *Smart Region Definition* (SRD) [25] e o *Modified Iterative/Uniformative Variable Elimination-PLS* (IVE/UVE-PLS) [26], que são úteis para produzir um modelo que revela mapas de contorno ao redor das moléculas alinhadas. Tais mapas de contorno revelam regiões do espaço que tem energias repulsivas e atrativas que revelam pontos de modificação estrutural para propor novos compostos (**Figura 7**).

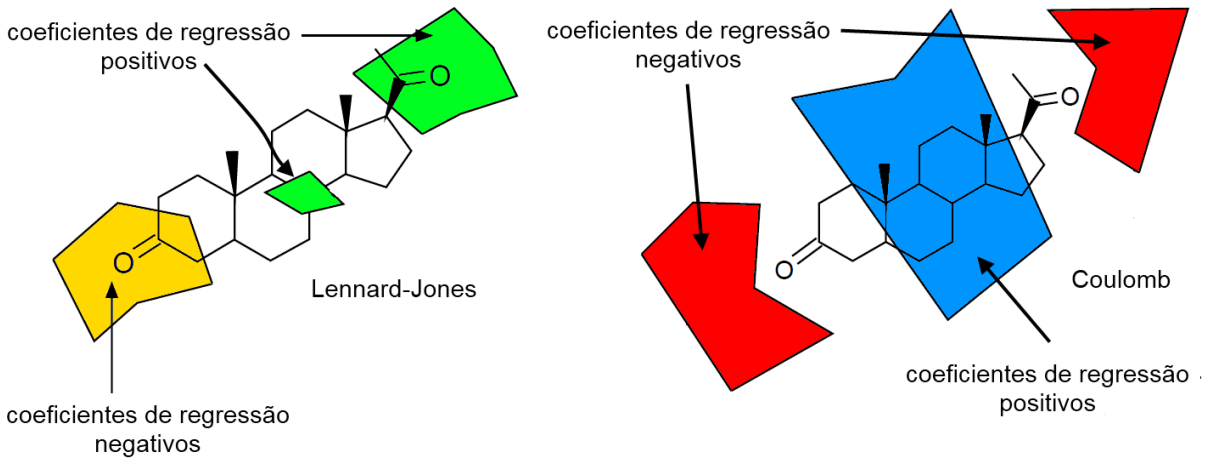


Figura 7. Representações 2D dos mapas de contorno 3D usados para representar o modelo de QSAR-3D com descritores MIF. Nesta representação os descritores calculados com o potencial de Lennard-Jones são mostrados em amarelo e verde, sendo estas cores usadas para diferenciar o sinal do vetor de regressão do método PLS. À direita é mostrado o exemplo para os descritores calculados como o potencial de Coulomb.

A principal vantagem da metodologia de QSAR-3D é que mesmo sem a informação da macromolécula alvo é possível construir modelos que têm grande utilidade para proposição de análogos através do aspecto visual dos descritores. Ao contrário da metodologia de

QSAR tradicional, bons modelos são obtidos com uma série de compostos estruturalmente muito diversa, sendo que a única característica comum deve ser o farmacofóro. As principais desvantagens da abordagem é que as previsões estão limitadas ao espaço 3D definido pelas moléculas do conjunto de treinamento. Assim, a previsão da atividade para compostos muito mais volumosos que aqueles usados para construir o modelo não é confiável [21]. Porém, a maior limitação da metodologia é que a proposta usada para realizar o alinhamento deve refletir o que é observado para o modo de interação no sítio ativo. Algumas vezes moléculas similares ligam-se ao sítio ativo de forma não intuitiva, e se o alinhamento não levar esse fato em consideração, o modelo não será preditivo [27-29].

I.3 QSAR-4D e Simulações de dinâmica molecular

O método de QSAR-3D faz uso de apenas uma conformação molecular (supostamente a conformação bioativa) para gerar modelos. Esta conformação representa a posição média da molécula no sítio de ligação, mas diz pouco a respeito de como é o comportamento dinâmico deste composto.

Hopfinger e colaboradores desenvolveram uma metodologia que considera a dependência temporal em modelos QSAR-3D [30]. Tal metodologia utiliza um perfil de amostragem conformacional (PAC) para cada ligante, e o alinhamento considera tais perfis ao invés de uma única conformação. Na metodologia de QSAR-4D se considera o farmacóforo, a liberdade conformacional e de alinhamento no desenvolvimento de modelos de QSAR-3D para os dados de estrutura-atividade de um determinado conjunto de treinamento, avaliando a média deste conjunto em relação ao tempo, a quarta dimensão [31].

Quando a estrutura 3D da biomacromolécula alvo está disponível e se conhece o modo de interação do conjunto de compostos investigado, a abordagem dependente do receptor (DR) pode ser desenvolvida. No entanto, quando a estrutura da biomacromolécula alvo não está disponível ou, mesmo que estiver, o modo de ação do ligante não está totalmente elucidado, a abordagem QSAR-4D independente do receptor (IR) pode ser realizado.

Ângelo Vedani e Max Dobler [32] propuseram uma metodologia diferente para criar os PACs. As simulações de dinâmica molecular são processadas em ambientes quasi-atômicos adjacentes ao ligante, mimetizando um sítio de ligação. Alega-se que tais restrições formaria uma quinta dimensão nos modelos QSAR. O dito formalismo QSAR-5D utiliza ainda modelos moleculares em muitas orientações e estados de protonação para gerar os descritores empregados na construção dos modelos.

Simulações de dinâmica molecular

Uma das principais ferramentas dos estudos teóricos de moléculas biológicas são as simulações de dinâmica molecular. Esse método computacional calcula o comportamento dependente do tempo para um sistema. Tais simulações são usadas para se obter informações detalhadas sobre as flutuações conformacionais de receptores e seus ligantes.

As simulações de dinâmica molecular foram introduzidas por Alder e Wainwright no fim dos anos de 1950 [33] para estudar as interações de esferas rígidas. Diversas simulações de líquidos utilizaram desse trabalho. O maior avanço veio em 1964, quando Rahman realizou a primeira simulação que usava um potencial realístico para o argônio [34]. A primeira simulação de dinâmica molecular de água líquida foi reportada em 1974 por Rahman e Stillinger [35] e a primeira simulação com proteína foi feita com o inibidor de tripsina pancreática em 1977 [36]. Hoje na literatura é rotineiro encontrar simulações entre proteínas, DNA e até sistemas complexos como membranas biológicas para o desenvolvimento de novos compostos bioativos [37].

Para realizar as simulações de dinâmica molecular com sistemas tão complexos é necessário lançar mão de aproximações da representação atômica trocando o uso da mecânica quântica é substituída pela mecânica clássica. São empregadas as equações da segunda lei de Newton para o movimento das partículas. A partir do conhecimento da força agindo sobre as partículas do sistema, determina-se a aceleração de cada átomo. Pela integração das equações de movimento é possível traçar a trajetória que descreve as posições, velocidades das partículas e a variação desta no decorrer do tempo [38].

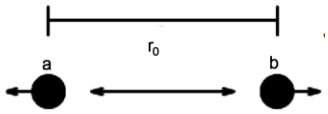
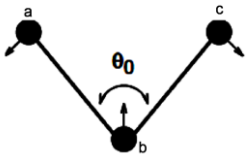
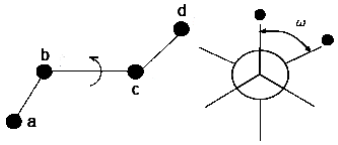
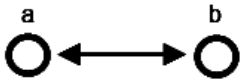
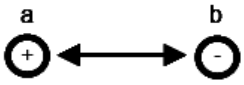
Uma simulação de dinâmica molecular pode gerar as conformações usadas para o formalismo QSAR-4D. Para isso, é necessária uma estrutura contendo a posição inicial para

todos os átomos do sistema incluindo o solvente. O próximo passo é calcular a força que age em cada átomo através da expressão da derivada da energia potencial *V* em relação às posições atômicas (q): $F = -\partial V / \partial q$. Com as informações das forças e velocidades são calculadas as energias potencial e cinética. Utilizando as equações de movimento de Newton calculadas em um intervalo de tempo discreto, o sistema é levado a uma nova configuração. O conjunto destas configurações no decorrer da simulação formará a trajetória dos átomos do sistema fornecendo conformações termodinamicamente acessíveis para o composto que se deseja fazer o estudo de QSAR [38].

Os potenciais de interação usados para calcular V são derivados de um campo de força molecular escolhido para realizar as simulações. Num campo de força de mecânica molecular a energia é assumida como sendo a soma dos termos potenciais que descrevem o estiramento de ligações covalente, deformações angulares, torções de ligações, interações de van de Waals e eletrostáticas entre outras (7). A **Tabela 2** mostra algumas formas de se calcular esses potenciais. Para citar alguns exemplos, no pacote computacional GROMACS [38] são utilizados os campos de força GROMOS [39] que possui parâmetros otimizados para reproduzir resultados experimentais para o comportamento de proteínas. Além do GROMOS, pode ser utilizado o campo de força AMBER [40, 41] que juntamente com o campo de força geral (GAFF) [42] é utilizado para simular sistemas de complexos ligante-receptor de interesse no estudo de QSAR-4D.

$$V_{tot} = \sum V_{estiramento} + \sum V_{def\ angular} + \sum V_{torção} + \sum V_{vdW} + \sum V_{Coulomb} + \ldots \quad (7)$$

Tabela 2. Algumas expressões comumente usadas para calcular a energia potencial dos átomos em um campo de força.

Termo	Expressão	Usado como modelo para
$V_{estiramento}$	$\frac{1}{2}k_{ab}^{\delta}(r_{ab}-r_0)^2$ Descreve o estiramento de dois átomos *a* e *b* ligados covalentemente onde k_{ab}^{δ} é a constante da mola e r_0 é a distância de equilíbrio.	
$V_{def\ angular}$	$\frac{1}{2}k_{abc}^{\theta}(\theta_{abc}-\theta_0)^2$ Descreve a deformação angular de três átomos ligados consecutivamente, onde k_{abc}^{θ} é a constante da mola e θ_0 é o ângulo de equilíbrio.	
$V_{torção}$	$k_{abcd}^{\omega}(1+\cos(n\omega-\omega_s))$ Descreve a energia torcional definida pelo diedro formado por quatro átomos *a*, *b, c* e *d,* onde k_{abcd} n e ω_s são termos ajustados a dados experimentais e/ou de mecânica quântica.	
V_{vdW}	$\frac{C_{ab}^{(12)}}{r_{ab}^{12}}-\frac{C_{ab}^{(6)}}{r_{ab}^{6}}$ Descreve a energia de interação entre dois átomos *a* e *b* não ligados covalentemente separados por mais de três ligações covalentes ou em outra molécula. Os termos $C_{ab}^{(12)}$ e $C_{ab}^{(6)}$ são ajustados para reproduzir o perfil energético de interação.	
$V_{Coulomb}$	$f\frac{q_a q_b}{r_{ab}}$ Descreve a energia de interação entre dois átomos *a* e *b* não ligados covalentemente separados por mais de três ligações covalentes ou em outra molécula. Os termos q_a e q_b são as cargas atômicas e *f* o fator de conversão elétrica.	

I.4 LQTA-QSAR

Num esforço conjunto de membros do grupo de pesquisas LQTA (Laboratório de Quimiometria Teórica e Aplicada) foi possível desenvolver uma metodologia que une as vantagens dos descritores MIF e a dependência temporal descrita no formalismo QSAR-4D, em um único software livre, o LQTA-QSAR [43]. Tal software foi testado em dois conjuntos de dados da literatura, um no qual foi aplicada a metodologia QSAR-4D-IR e outro em que foi aplicado o formalismo CoMFA (QSAR-3D).

O formalismo LQTA-QSAR-IR faz uso de conformações obtidas de simulações de dinâmica molecular com ligantes livres em solvente utilizando o software de livre acesso GROMACS [38]. As trajetórias das simulações de dinâmica molecular são alinhadas considerando informações farmacofóricas ou puramente estruturais, assim formando o PAC para construção de modelos de QSAR-4D. Um dos programas do pacote LQTA-QSAR, o LQTAgrid, cria uma grade virtual 3D de espaçamento de 1 Å ao redor do PAC e, então, são calculados os parâmetros de energia de interação, utilizando a mesma ideia descrita par um estudo de QSAR-3D. A **Figura 8** mostra um esquema para a criação de modelos LQTA-QSAR.

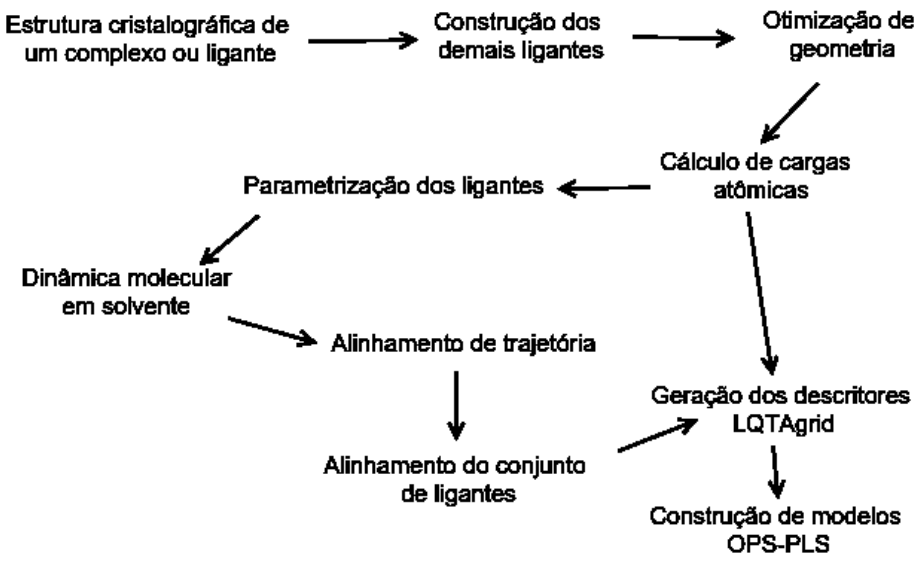


Figura 8. Esquema geral da construção de modelos LQTA-QSAR

As funções potenciais utilizadas para gerar os descritores MIF são mostradas nas expressões (8) e (9). Os parâmetros para a expressão (8) são os mesmos do campo de força GROMOS [39]. Os parâmetros para a expressão (9) são as cargas derivadas do potencial eletrostático CHELPG [44] e RESP [42]. Devido à presença de várias cópias dos compostos em todos os PAC, a energia é normalizada conforme o número de conformações. Os descritores QQ e LJ são dispostos em forma de matriz pelo programa LQTAgrid, onde cada linha corresponde a uma molécula e as colunas correspondem aos pontos da grade onde os descritores foram calculados.

$$E_{LJ} = \frac{1}{n}\sum_{l=1}^{L}\left(\frac{\sqrt{C_{sonda}^{(12)}C_{l}^{(12)}}}{r_{sonda\,l}^{12}} - \frac{\sqrt{C_{sonda}^{(6)}C_{l}^{(6)}}}{r_{sonda\,l}^{6}}\right) \quad (8)$$

$$E_{QQ} = \frac{1}{n}\sum_{l=1}^{L} f\,\frac{q_{sonda}q_{l}}{r_{sonda\,l}} \quad (9)$$

Os termos $C^{(12)}$ e $C^{(6)}$, mostrados em (8) e (9), são os parâmetros para os tipos de átomo segundo o campo de força utilizado, $r_{sonda\,l}$ é a distância da sonda até o átomo *l*, q_{sonda} é a carga da sonda q_l é a carga do átomo *i*, *f* é o fator de conversão elétrica que é igual a 138,935 kJ mol^{-1} nm e^{-2} e *n* é o numero de conformações do perfil.

Em relação a primeira versão [43] LQTAgrid foi aprimorado para utilizar quaisquer parâmetros $C^{(12)}$ e $C^{(6)}$ de qualquer campo de força, sendo agora denominado LQTAgridAFF (AFF, *all force field*). O campo de força GAFF, é de acesso livre através do programa Antechamber [42] e foi implantado com sucesso na nova versão do pacote LQTA-QSAR.

Seleção de descritores para construir modelos LQTA-QSAR

A seleção de descritores é realizada com o módulo QSARModeling (programa ainda sem divulgação) do pacote LQTA-QSAR, que usa o algoritmo OPS [45] (*Ordered Preditions Selection*) para a seleção de variáveis. O método PLS é utilizado para a regressão multivariada. Os descritores são submetidos ao prétratamento autoescalar, onde os valores de cada descritor é subtraída da média e feita a divisão pela variância do mesmo.

Os descritores dos modelos finais são expressos graficamente como esferas **(Figura 9)** e usado um esquema de cores que permite distinguir a natureza do descritor e o sinal do vetor de regressão é utilizado. Nos modelos CoMFA, onde os descritores são bem mais numerosos, eles são mostrados na forma de mapas de contorno pela sua junção na forma de superfícies [43].

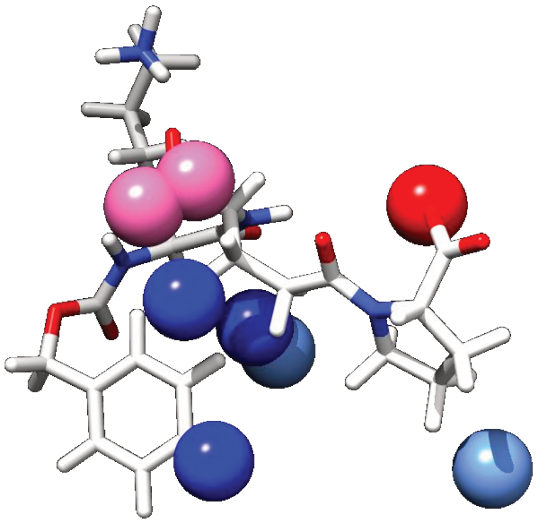


Figura 9. Exemplo genérico da exibição espacial dos descritores em um modelo LQTA-QSAR. Em vermelho e azul escuro são mostrados descritores de Coulomb e róseo e azul claro, os descritores de Lennard-Jones. Os tons azuis denotam correlação positiva com a atividade biológica e os demais o contrário.

O algoritmo OPS explora a capacidade de melhorar a qualidade dos modelos de regressão quando se removem variáveis (descritores) que têm baixa intensidade em um vetor informativo obtido durante a aplicação do método PLS. Tal vetor informativo pode ser composto dos valores absolutos para o vetor de regressão, valores da correlação absoluta da variável com y, ou ainda o produto entre ambos. No algoritmo OPS, os modelos são construídos com um determinado número de descritores (janela inicial) e posteriormente tal janela é aumentada de um incremento fixo até que todos os descritores ou uma porcentagem do total sejam adicionadas a esta. Parâmetros de qualidade dos modelos são obtidos a cada avaliação, para posterior comparação. O conjunto de variáveis que apresentarem os melhores parâmetros de qualidade contém as variáveis que apresentam a melhor

capacidade de previsão para o modelo construído e, portanto, são selecionadas. A **Figura 10** mostra um esquema do funcionamento do algoritmo OPS.

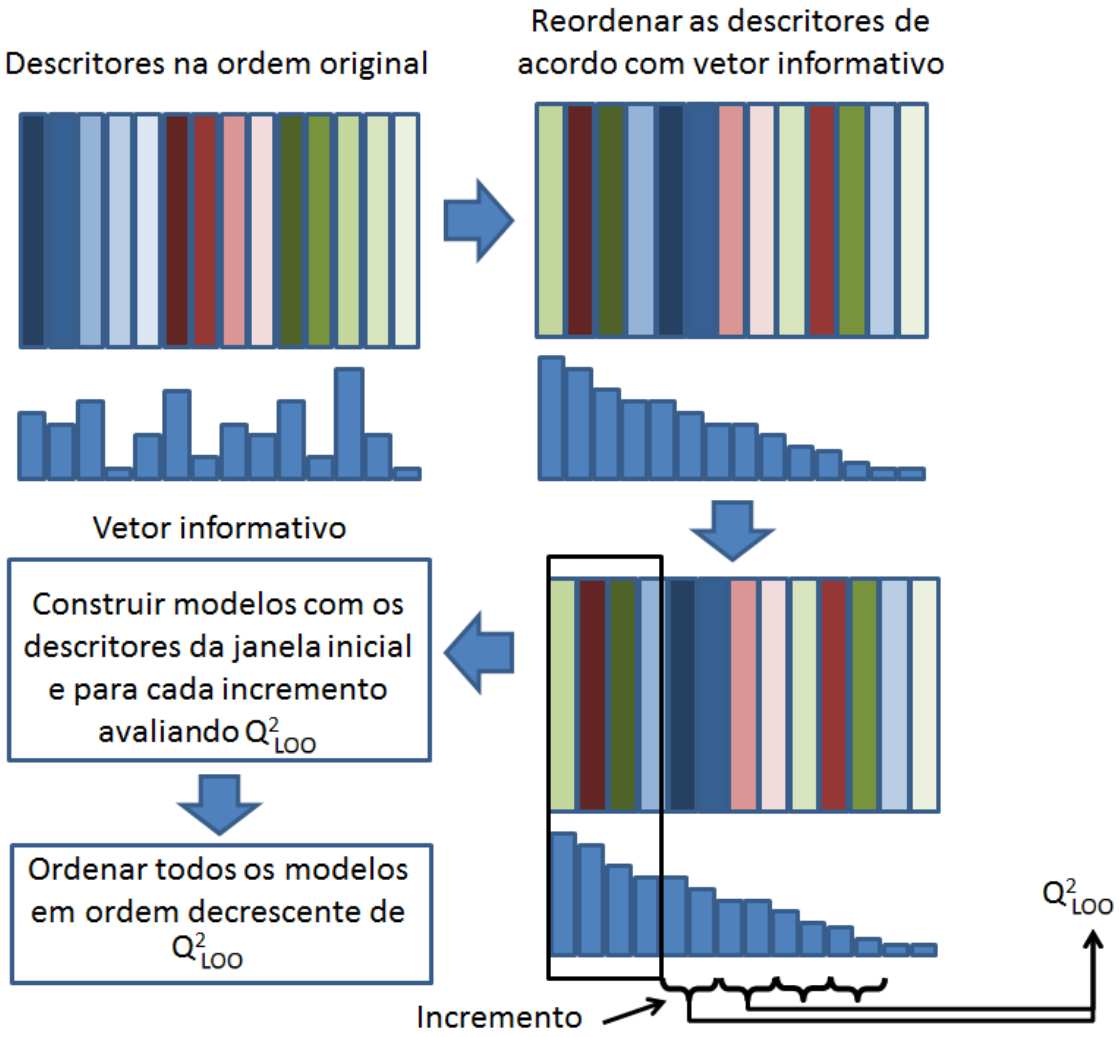


Figura 10. Funcionamento esquemático do algoritmo de seleção de descritores OPS.

I.5 Doença de Chagas e Inibidores da Tripanotiona Redutase

A tripanossomíase americana ou doença de Chagas é causada pelo protozoário *Trypanosoma cruzi*. Tal doença é considerada uma das mais sérias moléstias parasitárias que existem [46] e foi descrita por Carlos Chagas em 1909, sendo considerada uma das descobertas mais bem sucedidas e completas da história da medicina tropical. Carlos Chagas não descobriu somente uma nova patogenia em humanos, mas também o novo parasita e seu vetor. Foram desvendadas também o ciclo completo de infecção e as fases agudas e crônicas da doença [47-49].

Esta doença acomete principalmente seres humanos que vivem na América Latina, particularmente no sul. É estimado que cerca de 10 milhões de pessoas estejam infectadas pelo *T. cruzi* ao redor do planeta. Coura e Viñas [49] recentemente reportaram que populações em movimento das regiões endêmicas para a América do Norte e Europa podem carrear o *T. cruzi*. Essa movimentação do parasita vem impondo novos desafios epidemiológicos, econômicos, políticos e sociais, dado o espalhamento deste ao redor do globo.

A doença é transmitida principalmente por um inseto triatomíneo, o *Triatoma infestans*, que tem hábitos noturnos e se alimenta de sangue humano nesse período do dia. Tais insetos habitam esconderijos fornecidos por habitações precárias, frequentemente encontradas em áreas rurais e suburbanas. Os triatomíneos usam suas antenas (que possuem receptores de calor) para detectar variações de temperatura na pele e localizar vasos sanguíneos. Tais insetos estendem sua probóscide e perfuram a pele do hospedeiro, produzindo uma pequena lesão. Eventualmente, eles podem estar infectados pelo *T. cruzi*, que desenvolve parte da sua vida no intestino dos triatomíneos e é eliminado junto às fezes. Ao coçar a ferida instintivamente, as fezes depositadas acabam por contaminar a ferida ocorrendo introdução do parasita na corrente sanguínea. A **Figura 11** ilustra o ciclo de vida do *T. cruzi*. Existem outras formas de contágio menos frequentes como a ingestão de alimentos contaminados com fezes do inseto, transfusão de sangue, transmissão vertical, transplante de órgãos ou até mesmo acidentes de laboratório [46, 50].

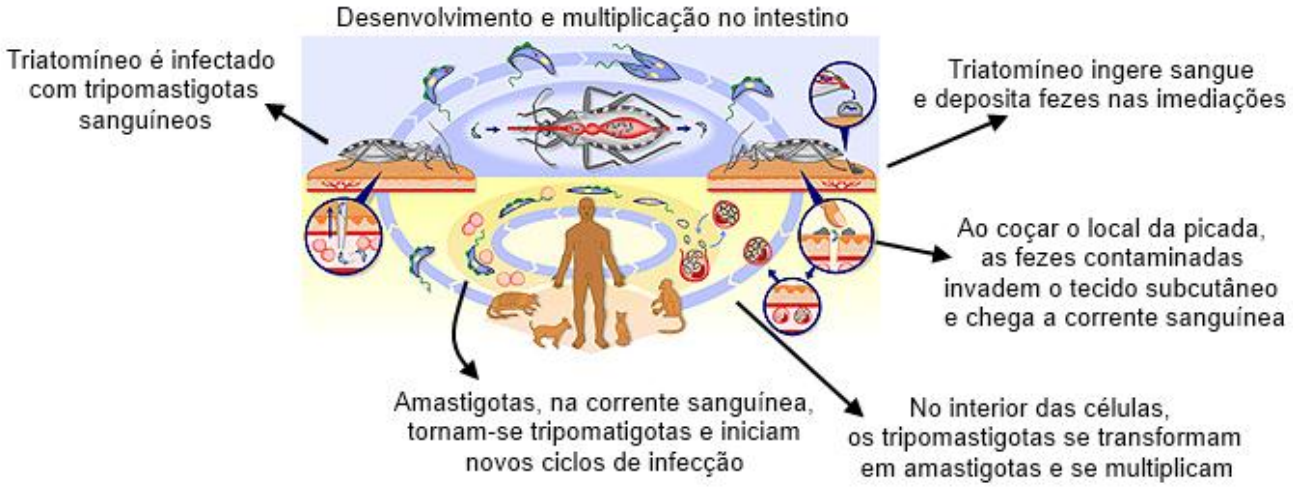
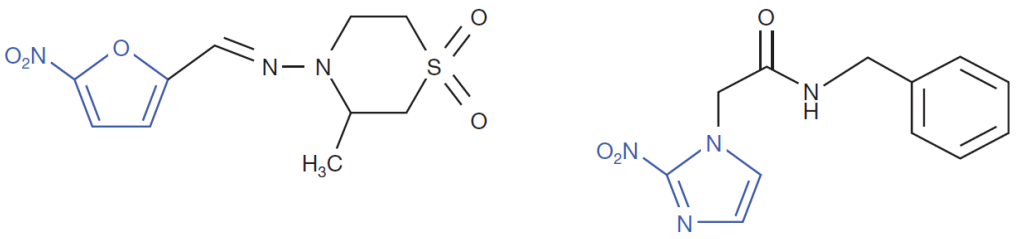


Figura 11. Clico de vida do *T. cruzi* [46].[1]

A doença de Chagas possui duas fases. A fase inicial ou aguda ocorre por volta de dois meses após a infecção, nos quais, um grande número de parasitas pode ser encontrado na corrente sanguínea, e na maioria dos casos há pouca ou nenhuma manifestação de sintomas. Durante esta fase o parasita se aloja principalmente nos músculos cardíacos e intestinais. Cerca de 30% dos pacientes sofrem de disfunções cardíacas e 10% de alterações intestinais, neurológicas ou ambas, devido decorrentes do alojamento do *T. cruzi* nestes órgãos. Após alguns anos de infecção pode ocorrer morte súbita ou parada cardíaca causada pela destruição da musculatura do órgão [46].

Não existem fármacos profiláticos para evitar a contaminação por *T. cruzi*. A quimioterapia atual para a doença de Chagas é baseada nos compostos nitroaromáticos nifurtimox e benznidazol (**Figura 12**). Esses fármacos são úteis para combater o parasita apenas durante a fase aguda da doença, tendo eficácia limitada na fase crônica. O benznidazol e o nifurtimox não apresentam as características ideais para serem usados como tripanossomicidas. Assim, há uma grande necessidade de se buscar novos compostos que tenham baixa toxicidade e alta eficácia para a fase crônica da doença [51]. Foi aberta uma grande campanha para a busca de compostos capazes de agir em novos alvos enzimáticos cruciais para o parasita [52].

[1] Disponível no website da organização mundial da saúde (*www.who.int*).



(A) (B)

Figura 12. Estrutura química dos fármacos nifurtimox (A) e benznidazol (B).

O uso conjunto da genômica, proteômica e da bioinformática propiciaram o inicio de uma nova era da identificação de novos alvos para fármacos nos *tripanossomos*. Enzimas e receptores que são essenciais para o ciclo de vida do *T. cruzi*, e que estão ausentes na maquinaria metabólica humana, ou ainda que tenha correspondentes com baixa similaridade de sequência podem ser usados para o desenvolvimento racional de fármacos. Foram encontrados alvos em potencial são as enzimas da via glicolítica [53], a DNA topoisomerases [54], as enzimas da biossíntese do ergosterol [55], no metabolismo de purinas [56] e a enzima tripanotiona redutase (TR) [57].

Em mamíferos, os danos potenciais que podem ser causados por espécies reativas do oxigênio são minimizados por um sistema de defesa baseado na glutationa (GSH), que se transforma na glutationa dissulfeto (GSSG). A regeneração da glutationa é feita pela glutationa redutase (GR). Em tripanossomas e leishimanias a evolução propiciou o surgimento de um sistema análogo baseado na tripanotiona ($T[SH]_2$) [58]. A tripanotiona oxidada ($T[S]_2$) é recuperada agora pela TR de maneira análoga ao sistema com GR (**Figura 13**). A sensibilidade de tripanossomas ao estresse oxidativo somada à ausência de tripanotiona em mamíferos validam as enzimas envolvidas no metabolismo de tripanotiona como alvos moleculares relevantes ao planejamento de novos agentes anti-*T. cruzi*.

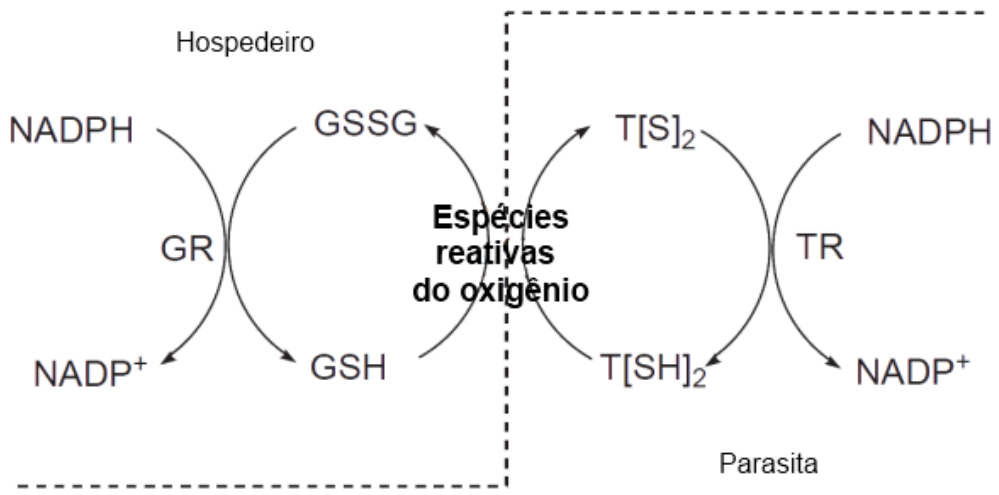


Figura 13. Esquema da proteção bioquímica contra espécies reativas do oxigênio em um hospedeiro mamífero e no *T. cruzi*. No primeiro, o substrato da enzima glutationa redutase (GR) promove a redução do dímero de glutationa (GSSG) em glutationa (GSH), enquanto que no parasita a enzima tripanotiona redutase (TR) é encarregada de reduzir a tripanotiona oxidada ($T[S]_2$) em tripanotiona $T[SH]_2$. O cofator fosfato de nicotinamida adenina dinucleotídeo (NADPH) é a coenzima responsável pela transferência de hidreto [60-61].

Quando se faz o alinhamento de sequência da TR do *T. cruzi* e a GR humana existe uma diferença bastante acentuada de 14 resíduos de aminoácido que participam da cisteína ativa à oxirredução. Apesar do sítio ativo da $T[S]_2$ lembrar o sítio da GR que liga GSSG, a seletividade desses substratos pelos sítios ativos é de cerca de 1000 vezes uma em relação à outra. O sítio da TR é mais hidrofóbico e tem uma carga total negativa que possibilita selecionar ligantes carregados positivamente [59]. O sítio da GR é menor e carregado positivamente repelindo a tripanotiona que é carregado positivamente. Na **Figura 14** é mostrada a estrutura cristalográfica do sítio ativo da TR em complexo com a tripanotiona [60] à região similar da estrutura do complexo GR com GSSG [61].

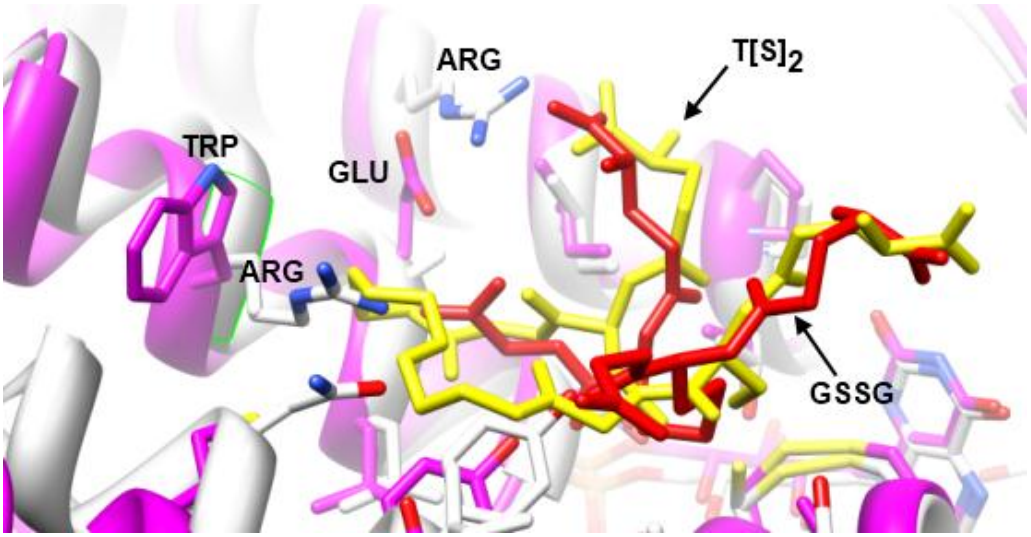
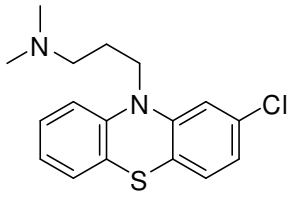
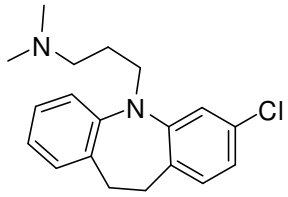
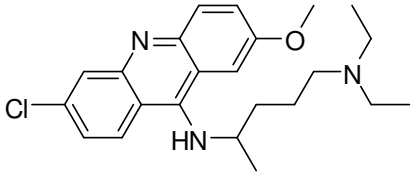
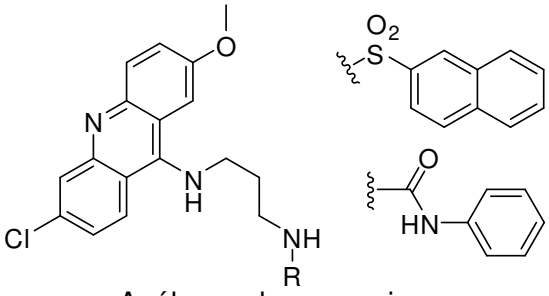
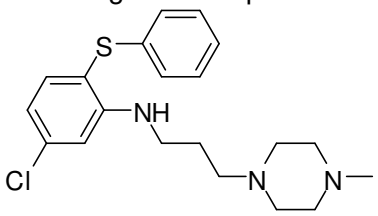
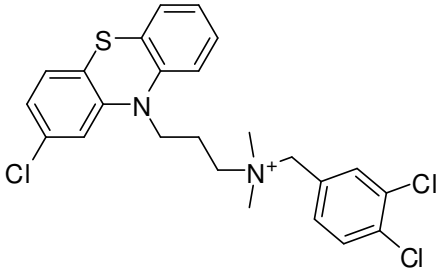


Figura 14. Sobreposição das estruturas da GR (branco) [61] (código PDB: 3DK4) e TR (róseo) [60] (código PDB: 3GRT) com seus respectivos substratos.

As estruturas cristalográficas da TR do *T. cruzi* disponíveis na forma livre e com os substratos naturais e inibidores fornecem subsídios para o desenvolvimento de inibidores seletivos [60, 62, 63]. Estudos de modelagem molecular com tais estruturas foram usados para identificar antidepressivos tricíclicos fenotiazínicos [64, 65], o antimalarial mepacrina [66] e compostos estruturalmente relacionadas à mepacrina [68] como inibidores competitivos da tripanotiona redutase sem inibição da GR. Derivados de sulfonaminas e uréia também se mostraram inibidores competitivos [67]. Modificações do anel central dos fenotiazínicos deram origem às “clorpromazinas de cadeia aberta” [68]. A quaternalização do nitrogênio alquílico da clorpromazina pela substituição por grupos benzílicos forneceu uma nova classe de inibidores competitivos da TR [69] (**Tabela 3**).

Tabela 3. Estrutura com reconhecida capacidade de inibir seletivamente a TR.

Estrutura	Referência
 Clorpromazina	Benson *et al.* [64]
 Clomipramina	Benson *et al.* [64]
 Mepacrina	Jacoby *et al.* [66]
 Análogos da mepacrina	Chibale *et al.* [67]
 clopromazinas de cadeia aberta	Fernandez-Gomes *et al.* [68]
 quaternelização do nitrogênio aliquilico da clorpromazina	Khan *et al.* [69]

É importante frisar que os compostos mostrados na **Tabela 3** não apresentam correlação entre a afinidade pela TR e a atividade antitripanomissida. Tal comportamento mostra que tais compostos ainda necessitam de otimização do perfil farmacocinético para o uso *in vivo* [57].

Justificativa

Quando se faz uma busca na literatura sobre o uso dos métodos de validação mencionados nesta **Introdução** em **QSAR-3D** fica clara a tendência de se encontrar apenas trabalhos muito recentes que as empregam, além do tradicional teste *bootstrapping*. No *bootstrapping* são selecionadas *N* amostras aleatórias para simular diversos conjuntos de validação externa, fazendo isso diversas vezes. Porém ao contrário do teste LNO o número *N* é fixo, geralmente 15% das amostras [20]. O grande problema está no teste **y**-*randomization* que parece ser um pesadelo para os modelos CoMFA. Alguns poucos artigos reportam o uso deste teste, mas na maioria dos casos os resultados são expressos de forma incompleta, ou apenas mencionando que "*tudo deu certo*" [70-75]. Apenas um autor se dignou mostrar a realidade, embora ignorando o fato de que esta era ruim. Segundo Oltulu *et al*. [76] é razoável que um modelo construído para um vetor aleatório de dados tenha valores de Q^2 LOO (chamado R^2CV) e R^2 de ~0,65 e ~0,7, respectivamente.

Aparentemente não há relatos na literatura de QSAR-3D sobre análises explícitas da correlação dos descritores e do perfil de distribuição em relação com y para a construção de modelos. Também não foram encontrados relatos de modelos CoMFA que se preocupam em observar a concordância entre o sinal do vetor de regressão e o de correlação dos descritores.

I.6 Objetivos

- Avaliar o impacto da utilização de uma filtragem dos descritores MIF que possuem baixa correlação e perfil de distribuição ruim em relação à y antes da seleção de descritores.

- Avaliar o impacto da remoção de descritores que oferecem discrepância entre o sinal do vetor de regressão e o coeficiente de correlação dos descritores em relação à y nos modelos QSAR.

- Definir um protocolo diferenciado para a seleção de descritores MIF na construção de modelos de QSAR-3D ou LQTA-QSAR que os utilizam.

- Verificar se os modelos obtidos passam nos testes de validação Leave-N-out e y-randomization, tendo ainda uma capacidade de prever bem a atividade para o conjunto de validação externa.

Como descrito também na **Introdução,** o método LQTA-QSAR independente de receptores pode ser estendido para a abordagem dependente de receptores.

Desta forma, adicionalmente, objetiva-se:

- Propor e testar a nova abordagem para o método LQTA-QSAR designado como o método LQTA-QSAR dependente de receptores.

- Utilizar o protocolo de construção de modelos na matriz de descritores obtida com a abordagem LQTA-QSAR-DR avaliando as figuras de mérito para o modelo, validação externa e interna.

- Aproveitar a característica dinâmica da obtenção dos perfis de amostragem conformacional no contexto do sítio ativo da TR para explicar o modo de interação dos inibidores tricíclicos derivados da fenotiazína.

- Propor análogos baseado nos descritores finais do modelo.

Capítulo II

Metodologia

II.1 Filtros digitais

Notação: Escalares são escritos em itálico, vetores estão em negrito e minúsculo, e matrizes estão em caractere maiúsculo e em negrito. Um descritor retirado da *j*-ésima coluna de uma matriz de descritores **X(*I,J*)** é chamado $\mathbf{x}_j$. *I e J* são as dimensões da matriz nas linhas e colunas, respectivamente.

Corte de variância e tratamento dos descritores LJ

O primeiro passo para filtragem de descritores antes da construção de modelos QSAR com descritores MIF é a eliminação daqueles excessivamente distantes das estruturas 3D alinhadas dentro da grade virtual. É difícil atribuir a estes descritores quais átomos da molécula que contribuem para a energia nestas partes da grade. A maneira mais simples de realizar essa primeira filtragem é eliminar descritores que têm variância reduzida. Kubinyi *et al*. [22] sugerem eliminar descritores que tem variância menor que 0,01 ou 0,02. O valor escolhido para esse corte na variância foi a alternativa mais conservadora de 0,01.

Os descritores LJ foram submetidos a um tratamento especial devido à presença de valores positivos de grande ordem de magnitude. Quando a sonda está numa posição da grade muito próxima a um dos átomos de uma molécula, os valores de energia potencial adquirem valores positivos excessivamente altos devido à expressão do potencial de LJ. Usar concomitantemente valores numéricos pequenos e muito grandes pode causar instabilidade na construção do modelo com o método PLS. Dessa maneira, os descritores de LJ foram submetidos a um prétratamento de acordo com as expressões abaixo (10). Essa transformação garante que a informação dos pontos da grade próximos às moléculas alinhadas não sejam completamente perdidas.

$$\text{se } LJ_{x,y,z} \leq 30\,\text{kcal mol}^{-1} \text{ então } LJ'_{x,y,z} = LJ_{x,y,z}$$
$$\text{se } LJ_{x,y,z} > 30\,\text{kcal mol}^{-1} \text{ então } LJ'_{x,y,z} = 30 + \log_{10}\left(\frac{LJ_{x,y,z}}{\text{kcal mol}^{-1}} - 29\right) \quad (10)$$

O método CoMFA utiliza deste tipo de truncagem dos valores de LJ, mas sem adicionar o logaritmo do restante da energia. Quando todos os valores em um determinado ponto da grade são transformados para 30 kcal mol^{-1} a variância desse descritor se torna zero. Quando o corte de variância mencionado anteriormente é realizado, esses pontos, agora bastante próximos das moléculas, são também eliminados. Para reproduzir esse efeito na metodologia proposta nesta tese, **somente** foram eliminados os pontos da grade que tiveram **todos** os valores transformados pela expressão em (10).

Eliminação de descritores de baixa correlação

A segunda filtragem é feita no intuito de se eliminar os descritores que tem comportamento de um vetor aleatório qualquer. A matriz resultante após o tratamento dos valores de energia de LJ e do corte pela variância foi submetida a um corte de correlação. Esse corte visa eliminar os descritores que têm correlação de Pearson absoluta ($|r|$) com **y** no mesmo nível de ruído aleatório. Para definir o parâmetro de corte pela correlação foi calculado r para um grande número de vetores aleatórios com **y** obtendo o vetor ($\mathbf{r}_{rand}$). O histograma de $\mathbf{r}_{rand}$ deve seguir uma distribuição normal ao redor da média populacional de zero ($\mu = 0$) e o limite de confiança superior de 99% pode ser usado com o valor de corte para $|r|$ ($|r|_{corte}$). Através de testes realizados pode ser observado que o limite de confiança varia conforme o número de amostras do conjunto de treinamento. O valor $|r|_{corte}$ específico para cada conjunto de dados é mostrado na equação (11)

$$|r_{corte}| = Z_{0,99}\sigma_{\mu} \quad (11)$$

onde $Z_{0,99}$ é o número de desvios padrão da média de uma distribuição normal necessária para conter 99% da área e σ_{μ} é o desvio padrão da média.

Ao que parece o valor $|r|_{corte}$ tende a zero para um número infinito de amostras, ou seja, quanto maior o número de amostras menor o valor deste parâmetro. Por experiência do

grupo LQTA, usar descritores com $|r|$ menor que 0,3 pode introduzir problemas nas validações dos modelos. Desta forma, um valor para $|r|_{corte}$ de 0,3 foi utilizado quando este parâmetro fosse menor que 0,3.

Eliminação de descritores mal distribuídos

A terceira etapa de filtragem é feita visando à eliminação de descritores que têm perfil de distribuição muito diferente em relação à **y**. Para isso, foi desenvolvido um algoritmo capaz de quantificar a similaridade de distribuição entre **y** e cada descritor. A esse algoritmo foi dado o nome de *Comparative Distribution Detection Algorithm* (CDDA) é descrito a seguir.

Passo 1: Um descritor ($\mathbf{x}_j$) é escalado pela amplitude, de forma que o valor mínimo seja 0 e máximo 1 obtendo um descritor modificado, $\mathbf{x}'_j$, onde o menor valor será sempre zero e o maior um.

Passo 2: Similarmente à produção de um histograma, cada descritor, $\mathbf{x}'_j(I,1)$ tem o intervalo [0,1] dividido em K subdivisões, $k = 2^n$ (n =1,2,3,..), por exemplo para n=1 e k=2, são formados dois intervalos [0,1/2[e [1/2,1]. O número de valores dentro de cada subdivisão é contado pela rotina a seguir.

Rotina para o cálculo do número de descritores em cada subdivisão

```
for k = 1,…, K
    for i = 1,…, I
       if { x_i ≥ 2^-n (k-1) AND x_i < 2^-n k} OR { x_i = 2^-n k AND x_i = 1};
       f_i,k = 1 else f_i,k = 0
       end
    end
end
```

Os resultados são armazenados em matrizes lógicas **F**(I,K) e então transformadas em um vetor $\mathbf{f}_{k(j)}$ pela soma nas linhas (12).

$$\mathbf{f}_{k(j)} = \sum_{i=1}^{I} f_{i,k} \text{ , k=1,...,K} \quad (12)$$

Um procedimento similar é feito com **y** dando origem ao vetor de frequências $\mathbf{f}_{k(y)}$.

Passo 3. Somando a diferença absoluta entre cada vetor $\mathbf{f}_{k(j)}$ e $\mathbf{f}_{k(y)}$ é obtido um valor ε_j que é dependente do valor de *n* escolhido no **passo 2** (13).

$$\varepsilon_j(n) = \sum_{k=1}^{K} \left| \mathbf{f}_{k(j)} - \mathbf{f}_{k(y)} \right| \quad (13)$$

Se um descritor tem exatamente a mesma distribuição de **y** fará com que qualquer valor $\varepsilon_j(n)$ seja igual a zero. Contudo, $\varepsilon_j(n)$ será maior que zero quando isso não ocorrer, indicando que existem subdivisões sobrecarregadas e, ao mesmo tempo, há outras mais vazias quando se compara $\mathbf{f}_{k(j)}$ e $\mathbf{f}_{k(y)}$. Além disso, os dois valores extremos 0 e 1 estão fixos e dessa forma o valor de $\varepsilon_j(n)$ não poderá superar o valor de número de amostras (*I*) menos 2. Assim, podemos reescrever $\varepsilon_j(n)$ numa forma normalizada, $e_j(n)$, como indicado na equação (14).

$$e_j(n) = 1 - \frac{\varepsilon_j(n)}{2I - 2} \quad (14)$$

Os valores de $e_j(n)$ mudam em função do número de subdivisões escolhido. Para definir um valor padrão para *n*, foi calculado o poder de discriminação entre amostras bem e mal distribuídas. Nesse sentido foi calculado $e_j(n)$ para diversos descritores em vários conjuntos de dados, variando *n* de 1 até 8 (2 a 256 subdivisões). Considerou-se o valor padrão para *n* aquele que maximizava a expressão: $\max(e_j(n)) - \min(e_j(n))$. A **Figura 15** mostra os resultados obtidos com essa abordagem.

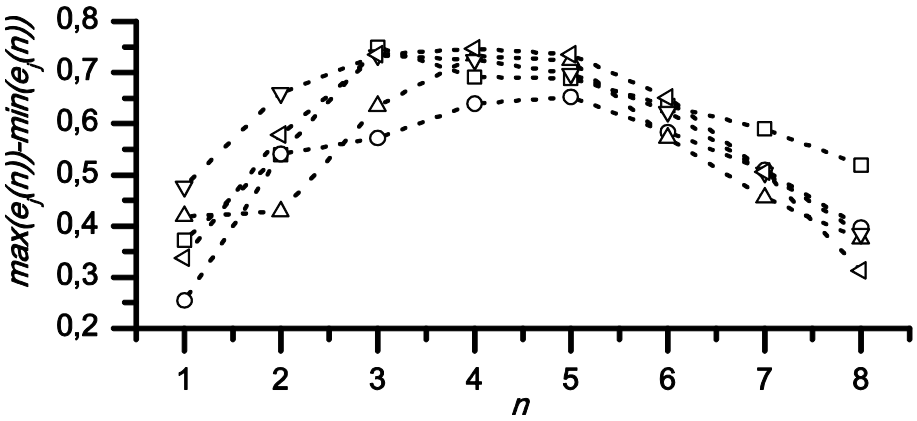


Figura 15. Investigação para definir o valor padrão para *n* utilizando dados com descritores criados a partir do nosso banco de dados (◁) e outros retirados da literatura. Foram usados dados de inibição da diidrofolate redutase (□) [77], receptor benzodiazepínico (○) [78], enzima conversora da angiotensina (△) [79] e da acetilcolinesterase (▽) [80]. A figura permite evidenciar que o valor máximo se apresenta em torno de *n*=4.

Com base nos resultados expostos na **Figura 15** o valor padrão para $e_j(n)$ foi definido como sendo quatro. Portanto, os resultados apresentados no texto subsequente $e_j(4)$ são simplesmente relatados como e_j.

Conjunto de dados para o teste dos filtros digitais

Para testar a aplicabilidade da ferramenta CDDA em descritores do tipo MIF e verificar se o seu uso tem implicações na melhoria dos modelos de QSAR-3D-(4D), foram selecionados 3 conjuntos de dados da literatura. Modelos CoMFA foram aplicados a esses conjuntos de dados e eles foram escolhidos pela disponibilidade de estruturas 3D em formato mol2 com cargas atômicas atribuídas. O primeiro conjunto de moléculas consiste de 40 catalisadores de transferência de fase (30 para o conjunto de treinamento e 10 para previsão) [81]. O segundo conjunto consiste de 114 inibidores da enzima conversora da angiotensina [79], (84 para o conjunto de treinamento e 30 para previsão). O terceiro conjunto de dados consiste de 351 inibidores da diidrofolato redutase (266 para o conjunto de treinamento e 85 para previsão). A **Figura 16** ilustra a variabilidade estrutural dos conjuntos de dados escolhidos mostrando os valores extremos do vetor de atividade.

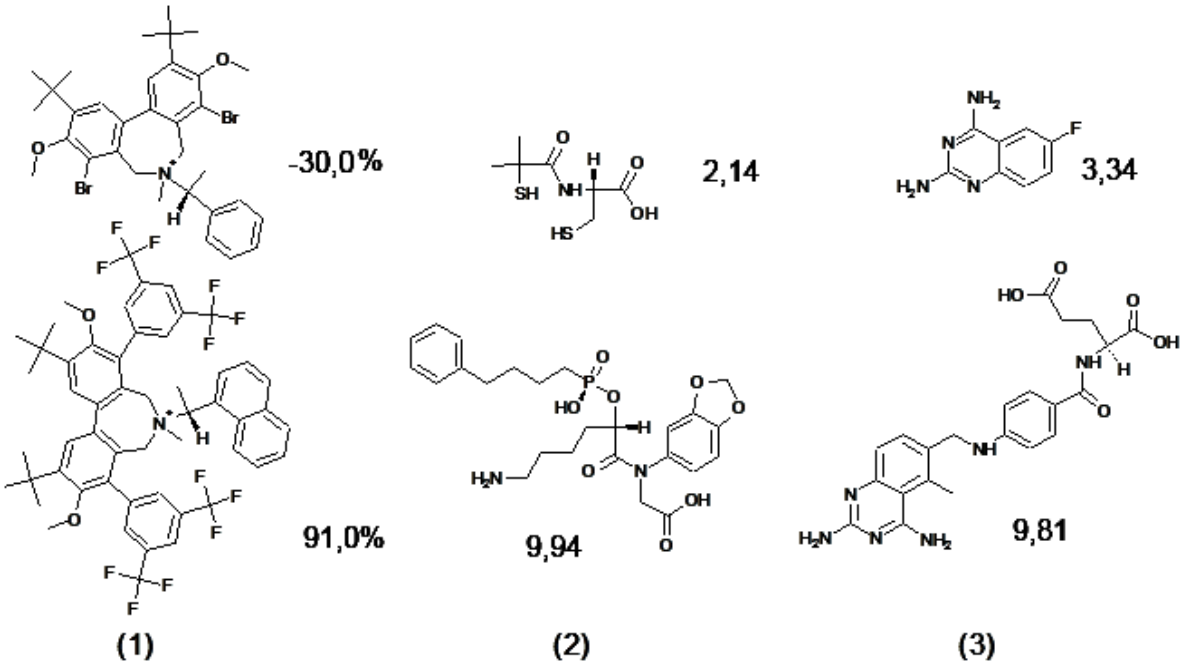


Figura 16. Representantes de menor e maior valor de y para cada conjunto de dados escolhido. Atividades em e/e% para (1) pIC_{50} para (2) e (3)

Os descritores foram calculados em posições fixas determinadas por uma grade virtual 3D visitada pela sonda com parâmetro para o átomo de carbono sp3 com carga +1, em incrementos de 0,5 Å com margem de 5,0 Å de distancias dos átomos das paredes da grade. O número de descritores produzidos foi um total de 158.400 para conjunto de moléculas (1), 162.922 para (2) e 112.800 para (3).

Os conjuntos de validação externa (previsão) foram definidos com o auxílio da Análise de Agrupamentos Hierárquicos (HCA) [82] construídos para um modelo construído para todas as amostras. Os dendrogramas do HCA foram construídos com as matrizes de descritores para modelos construídos com todas as amostras, usando distâncias Euclidianas e método de agrupamento completo. O método HCA foi usado com o objetivo de fornecer um modo visual para ajudar na definição do conjunto externo e evitar a remoção de amostras com características singulares dentro dos conjuntos de dados. Evitou-se selecionar as amostras com os valores de atividade mais baixa e mais alta. O procedimento completo de tratamento dos dados foi reiniciado para a matriz formada pelo conjunto de treinamento

escolhido. Não foram inseridas neste processo as amostras que formaram o conjunto teste [12, 19].

Cálculo dos descritores de campo de interação molecular

O programa LQTAgridAFF foi utilizado para recriar os descritores de MIF. Embora o LQTAgridAFF tenha sido desenvolvido para lidar com um perfil de conformações obtidas de simulações de dinâmica molecular, é possível também criar os mesmos descritores para uma única conformação como no caso do CoMFA.

A energia de Coulomb foi calculada utilizando uma sonda com carga positiva 1 segundo o potencial mostrado em (5), utilizando as cargas atômicas fornecidas pela literatura. A energia de van der Waals foi calculada utilizando o potencial de Lennard-Jones mostrado na equação (6), com parâmetros para o campo de força GAFF [42].

Os descritores foram calculados em posições fixas determinadas por uma grade virtual 3D visitada pela sonda com parâmetro para o átomo de carbono sp^3 com carga +1 em incrementos de 0,5 Å. O número de descritores produzidos foi um total de 158.400 para conjunto de moléculas (1), 162.922 para (2) e 112.800 para (3).

II.2 Seleção de variáveis e validação dos modelos

A matriz resultante após os três passos de filtragem dos descritores foi considerada pronta para a seleção de variáveis com o algoritmo OPS [45] utilizando o pré-processamento autoescalar. A seleção de varáveis foi auxiliada pela remoção de descritores redundantes (coeficiente de correlação interdescritor igual ou maior que 0,9 mantendo os mais bem correlacionados com y). Também formam removidos aqueles que apresentavam discrepância entre o sinal do vetor de regressão e do coeficiente *r* com a atividade biológica.

O método de calibração multivariada PLS, implementado no programa *ad hoc* QSARmodeling foi utilizado para a construção dos modelos QSAR. Os modelos foram considerados prontos para realizar a previsão para o conjunto externo se: i) não houvesse discrepância entre o sinal dos elementos do vetor de regressão e do correlograma, ii) o

modelo apresentasse figuras de mérito razoáveis e iii) passassem nos testes de validação **y**-r*andomization* e *leave-N-out*.

Para o teste *leave-N-out*, os modelos foram considerados robustos se o desvio máximo absoluto do Q^2_{LOO} e o valor de Q^2_{LNO} não fossem maiores do que 0,05 para *N* de cerca de 20-30% separadas para a validação interna [11, 12], os resultados são mostrados graficamente, expressando além dos valores de desvio máximo, o valor médio e dois desvios padrão deste valor.

Para o teste de **y**-r*andomization* os modelos foram considerados livre de correlação ao acaso quando os interceptos para as regressões lineares dos valores de |r| x Q^2_{yrand} e |r| x R^2_{yrand} fossem: $a_Q < 0,05$ e $a_R < 0,3$ [18]. Os resultados desse teste foram expressos em dois gráficos separado que contem a linha para a regressão linear mostrando as equações $Q^2_{yrand} = a_Q + b_Q|r|$ e $R^2_{yrand} = a_R + b_R|r|$.

Visualização dos modelos

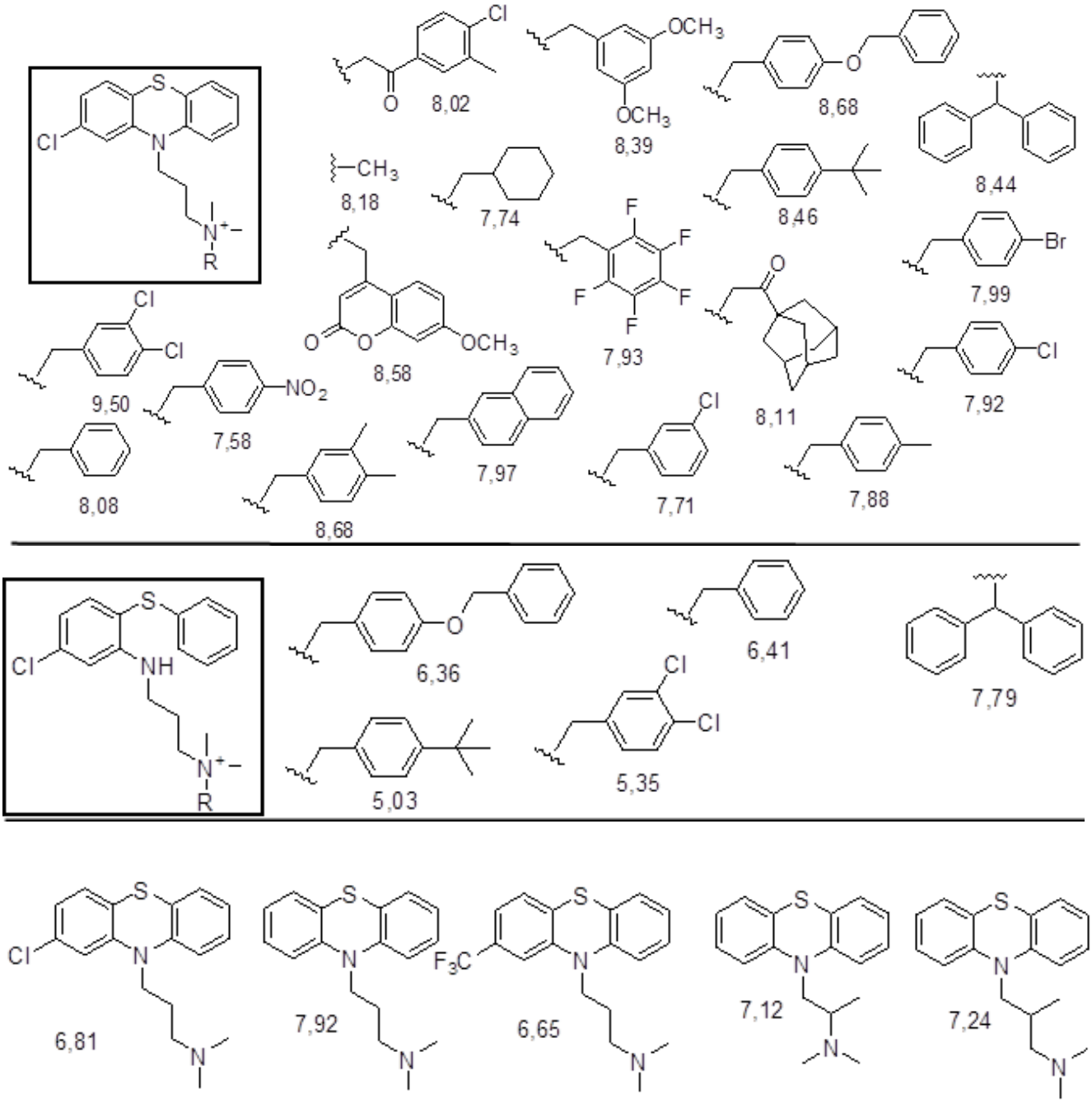
Os descritores dos modelos finais que passaram nas devidas validações são visualizados no espaço 3D ao redor das moléculas alinhadas. As coordenadas dos pontos da grade que definem os descritores são expressas em formato *pdb* [83] e exibidas como esferas utilizando o programa UCSF Chimera [84]. As interpretações dos descritores são feitas com base na disposição espacial, natureza (LJ ou QQ) e no sinal do vetor de regressão. O modelo e seus descritores foram utilizados para a proposição de novos compostos com possível atividade biológica real.

II.3 4D-LQTA-QSAR dependente de receptores

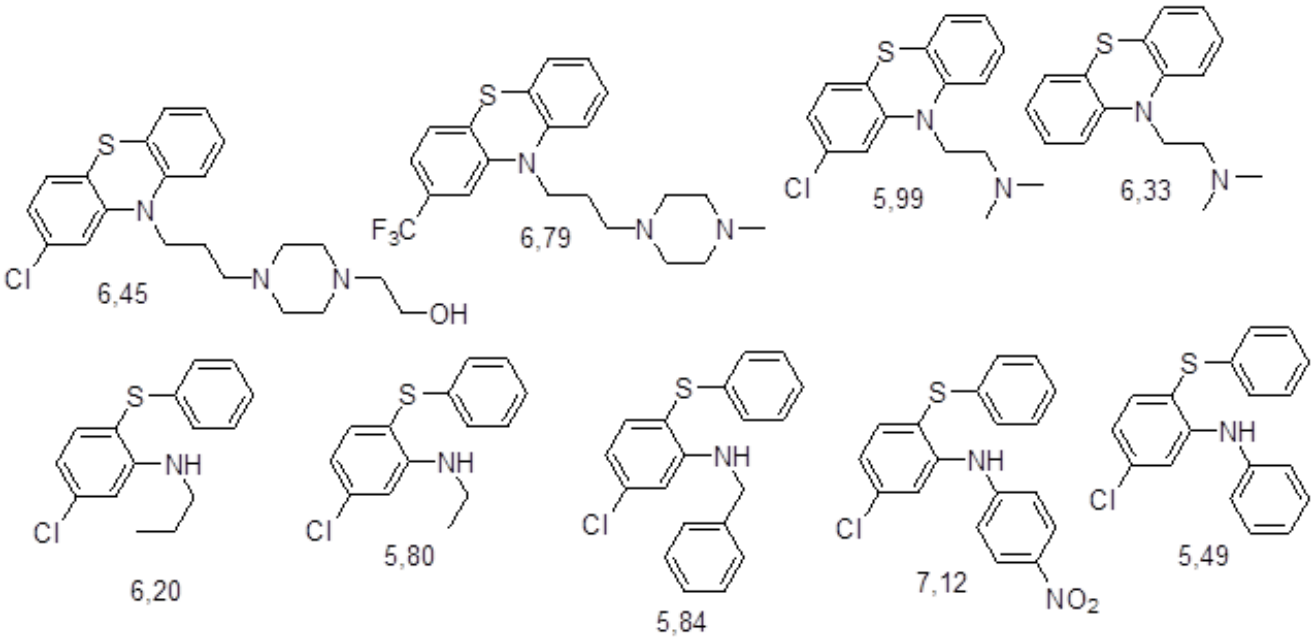
Conjunto de dados

Para a criação de modelos LQTA-QSAR-DR, foram selecionados 38 compostos com constante de inibição (K_i, [64, 65, 69, 85, 86]) conhecida contra a enzima TR [87] do *T. cruzi*. Tal conjunto de dados é inédito na aplicação do QSAR-4D. As estruturas dos compostos estão representadas no **Quadro 1**.

Quadro 1. Estruturas utilizadas no estudo LQTA-QSAR-4D-DR com valores indicativos de energia livre de interação (ΔG_{int}) em kcal mol^{-1}, calculados a partir dos valores de K_i.



Quadro 1. Cont. Estruturas utilizadas no estudo LQTA-QSAR-4D-DR com valores indicativos de energia livre de interação (ΔG_{int}) em kcal mol^{-1}, calculados a partir dos valores de K_i.



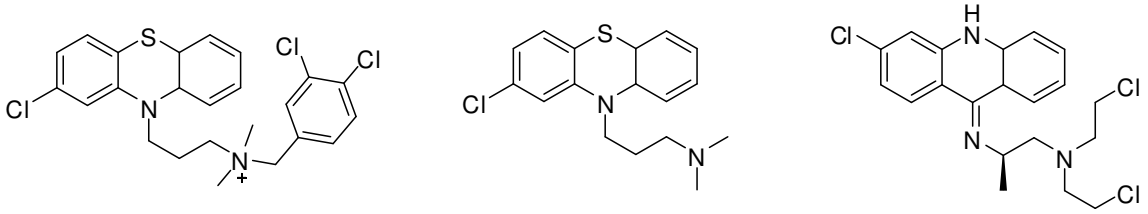
Tratamento dos ligantes

O estado de protonação dos ligantes quando no interior do sítio ativo foi determinado com o servidor PROPKA[2] [88]. Todos os modelos moleculares, nos devidos estados de protonação, tiveram as suas geometrias otimizadas no nível de teoria do funcional de densidade M05-2X [89] com o conjunto de bases cc-pVDZ (Gaussian'03 [90]). A análise populacional de CHELPG [91] foi utilizada para atribuição das cargas atômicas no mesmo nível de teoria. As geometrias otimizadas foram úteis para obter as topologias dos compostos com servidor PRODRG[3] versão beta [92]. Realizou-se a inspeção dos parâmetros para cada ligante para a melhor representação no campo de força GROMOS [39]. As cargas atômicas de Gasteiger, empregadas para as simulações de dinâmica molecular, foram obtidas com o programa AutoDockTools [93]. O funcional M05-2X foi escolhido ao invés do funcional B3LYP, pois o primeiro apresenta desempenho superior para a representação de geometrias de moléculas orgânicas [94].

[2] Disponível em: http://propka.ki.ku.dk/
[3] Disponível em: http://davapc1.bioch.dundee.ac.uk/prodrg/submit.html

Adaptação do sítio ativo aos ligantes estudados

A estrutura cristalográfica do complexo da TR do *T. cruzi* com o inibidor tricíclico, quinacrina (**Figura 17**) (código PDB: 1GXF [95], resolução de 2,7 Å, valor-R de 0,190) foi utilizada como receptor para o LQTA-QSAR-DR. A conformação do anel tricíclico de tal inibidor é bem diferente dos inibidores derivados da clorpromazina, o que implica que o sítio ativo apresenta uma conformação que não permite a inserção direta dos ligantes selecionados para a realização do estudo (**Quadro 1**).



A6 Clorpromazina Quinacrina

Figura 17. Estruturas químicas de inibidores da TR: A6 [69], clorpromazina e quinacrina.

Para realizar a adaptação do sítio ativo da TR aos compostos derivados da clorpromazina, o ligante quinacrina foi removido da estrutura cristalográfica e foi realizado o docagem (*docking*) do inibidor mais potente da série, o A6 (**Figura 17**). Foram obtidas algumas poses de A6 com modos de interação distintos no sítio ativo da TR. Tais estruturas foram submetidas a simulações de dinâmica molecular para propiciar a adaptação do sítio [96]. Estas simulações propiciaram a adaptação do sítio ativo simulando um efeito de encaixe induzido. O novo sítio adaptado para o ligante A6 foi utilizado para aperfeiçoar o modo de interação desse mesmo ligante.

Este processo foi feito iterativamente como apresentado na **Figura 18,** onde o sitio ativo é modificado por dinâmica molecular e aproveitado para potencializar o modo de interação de A6. O sitio modificado é ressubmetido à docagem até que não haja visualmente possibilidades de aperfeiçoar as interações com a TR. A docagem foi realizada com o programa AutoDock [93] e as simulações de dinâmica molecular com o pacote GROMACS [38].

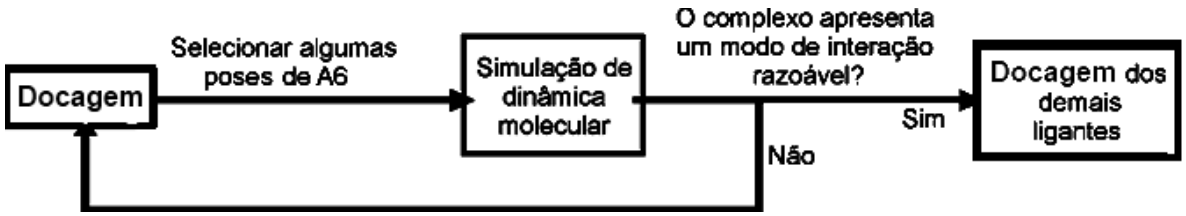


Figura 18. Esquema iterativo de adaptação do sítio ativo para acomodar A6.

Com a obtenção do sítio ativo hipotético adaptado para A6, o modo de interação para tal composto foi utilizado para orientar a docagem dos demais 36 ligantes. Tais complexos foram utilizados para realizar simulações de dinâmica molecular e obter os perfis de amostragem conformacional que foram utilizados para a construção dos modelos LQTA-QSAR-DR.

Parâmetros utilizados para as simulações de dinâmica molecular

As simulações foram realizadas em uma caixa cúbica com o modelo de solvente água SPC/E. As dimensões foram grandes o suficiente para que se tivesse 10 Å de distância do soluto às paredes da caixa. As paredes obedeceram à condições periódicas de contorno. O tratamento eletrostático foi o de *particle mesh Ewald* (PME) e as interações de Lennard-Jones e eletrostáticas foram truncadas em 11 Å. A pressão foi mantida constante pelo barostato de Parrinello-Rahman em 1 bar. A temperatura foi controlada com o termostato de Berendesen [38].

A integração foi realizada passo a passo em intervalos de 1 fs. As posições atômicas foram otimizadas em duas etapas: i) usando *algoritmo de descida mais inclinada* e ii) o *algoritmo de gradientes conjugados*. O critério de convergência empregado nas otimizações foi que a força máxima que agisse sobre os átomos não superasse 50 N. Caso o critério de convergência não fosse atingindo em 40.000 passos, um terceiro algoritmo (LBFGS [97]) foi empregado até a convergência.

A seguir for realizada uma etapa de simulação de posições atômicas restringidas para que o solvente se arranjasse ao redor do complexo. Na etapa de estabilização do sistema foi empregado um esquema de aquecimento gradual. Neste esquema, o sistema foi submetido à simulações em quatro temperaturas, 50 K, 100 K, 200 K e 350 K por 20 ps. Em seguida o sistema foi submetido a uma temperatura 310 K por 1.000 ps. A estabilização do desvio médio quadrático das posições dos átomos da cadeia principal da proteína (RMSd, *root mean square deviation*) foi usado como critério de estabilização do sítio ativo. Se neste período uma simulação não se estabilizasse, o tempo de simulação era estendido por mais 1.000 ps.

Alinhamento molecular e criação dos descritores MIF

Após a estabilização dos valores de RMSd da cadeia principal da proteína, foi dado inicio ao processo de alinhamento molecular. Um sistema não usual para o alinhamento foi utilizado onde, ao invés de considerar somente os átomos do ligante, foram selecionados também átomos dos resíduos de aminoácidos adjacentes no sítio ativo. A **Figura 19** mostra os átomos selecionados para tal procedimento.

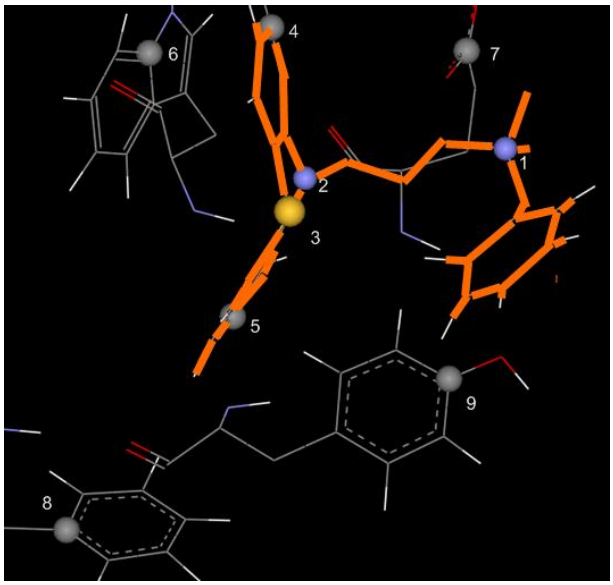


Figura 19. Esquema do sítio ativo mostrando os átomos selecionados do ligante (destacado em laranja) e o sítio ativo para realizar o alinhamento molecular.

A seleção dos átomos usados para o alinhamento foi baseada no modo de interação dos ligantes e os resíduos de aminoácidos no sítio ativo da TR. Foram selecionados para o alinhamento tanto átomos do sítio bem como alguns pertencentes aos ligantes. Este esquema de alinhamento guiado pelo sítio ativo foi proposto para reduzir os efeitos de erros sistemáticos do procedimento tornando os descritores de campo de interação representativos do sítio ativo da TR. Através da **Figura 19** é possível notar que a escolha dos átomos foi feita em pontos extemos dos ligantes e do sítio.

Após os alinhamentos, os perfis de cada ligante foram submetidos ao programa LQTAgrid para gerar os descritores de energia de interação. Uma caixa com dimensões 22 x 22 x 18 Å foi utilizada. Uma sonda com parâmetros idênticos à subunidade *N*-terminal proteica com carga formal +1 ($-NH_3^+$) foi utilizada para varrer toda a extensão da grade 3D virtual, gerando um total de 17.424 descritores com espaçamento de 1 Å.

Capítulo III

Resultados e Discussão

III.1 Ferramentas para QSAR 3D – 4D: Filtros Digitais

A **Figura 20** ilustra os pontos da grade resultantes após a remoção dos descritores que possuíam variância menor que 0,01 [exemplo para o conjunto de dados (3)]. A forma clássica proposta por Kubinyi *et al*. [22] se mostrou útil para a eliminação dos descritores excessivamente distantes das estruturas 3D alinhadas dentro da grade virtual. Os descritores restantes mostram uma clara capacidade dos mesmos em envolver os compostos da série estudada.

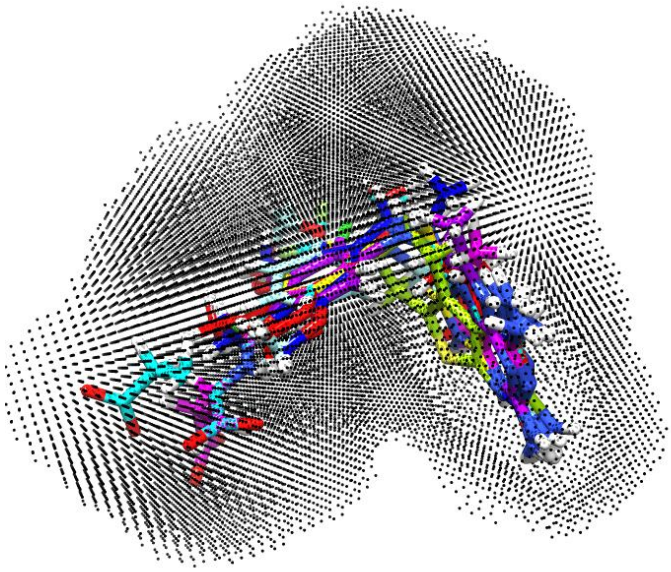


Figura 20. Descritores resultantes após a realização do corte pela variância.

Os descritores muito próximos das conformações alinhadas na grade virtual também são eliminados pelo corte de variância. No método CoMFA isso é feito truncando todos os valores de LJ para maiores que 30 kcal mol^{-1} para o valor 30, o que faz com que a variância deste descritor em particular se torne zero. Para reproduzir esse efeito foram eliminados todos os pontos da grade que tiveram seus valores transformados pelo tratamento proposto para os descritores de LJ. A **Figura 21** [exemplo para o conjunto de dados (1)] mostra que a estratégia foi útil para eliminar tais pontos da grade, que como os descritores muito distantes, são difíceis de interpretar.

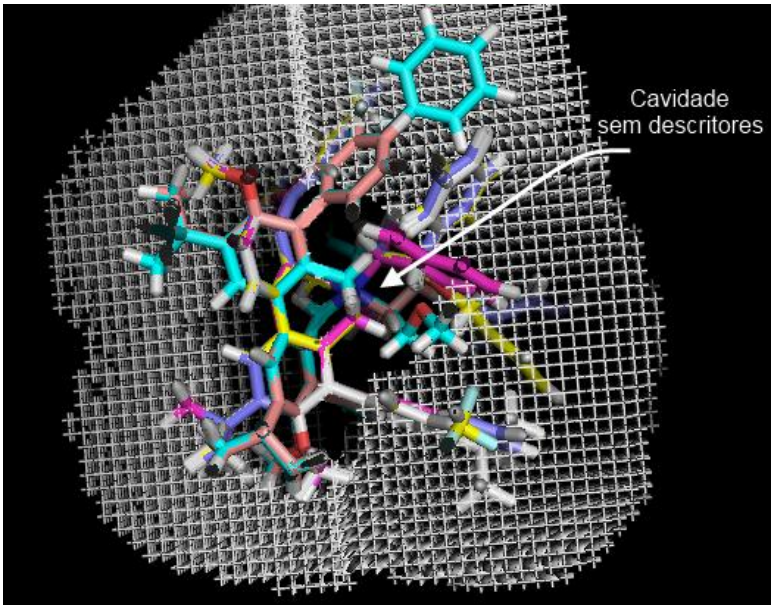
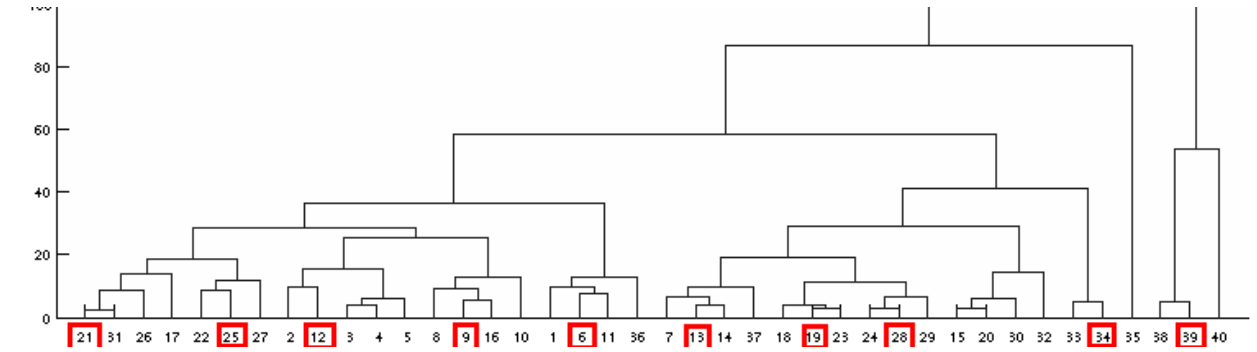


Figura 21. Corte longitudinal dos pontos da grade virtual restantes mostrando a eliminação de descritores muito próximos.

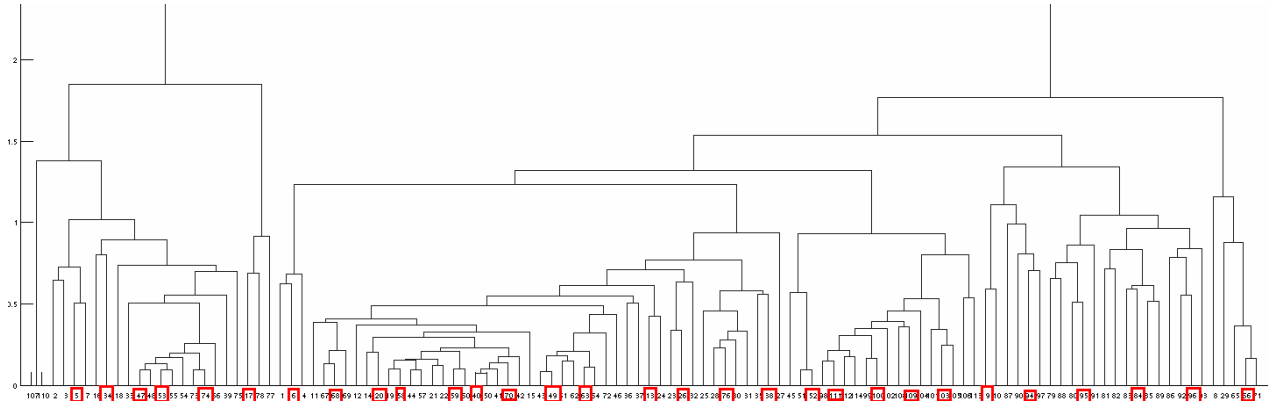
Os modelos construídos com o formalismo CoMFA utilizam, antes de qualquer tratamento nos descritores, o corte pela variância. Desta forma, os procedimentos de filtragem e seleção de variáveis mencionados nesta tese utilizam desta matriz de descritores inicial ($\mathbf{X}_{var}$). Nas seções a seguir são mostrados os resultados para o procedimento completo otimizado de tratamento dos descritores MIF aplicado aos três conjuntos de dados selecionados, que contém um número crescente de amostras.

Seleção do conjunto de dados externo

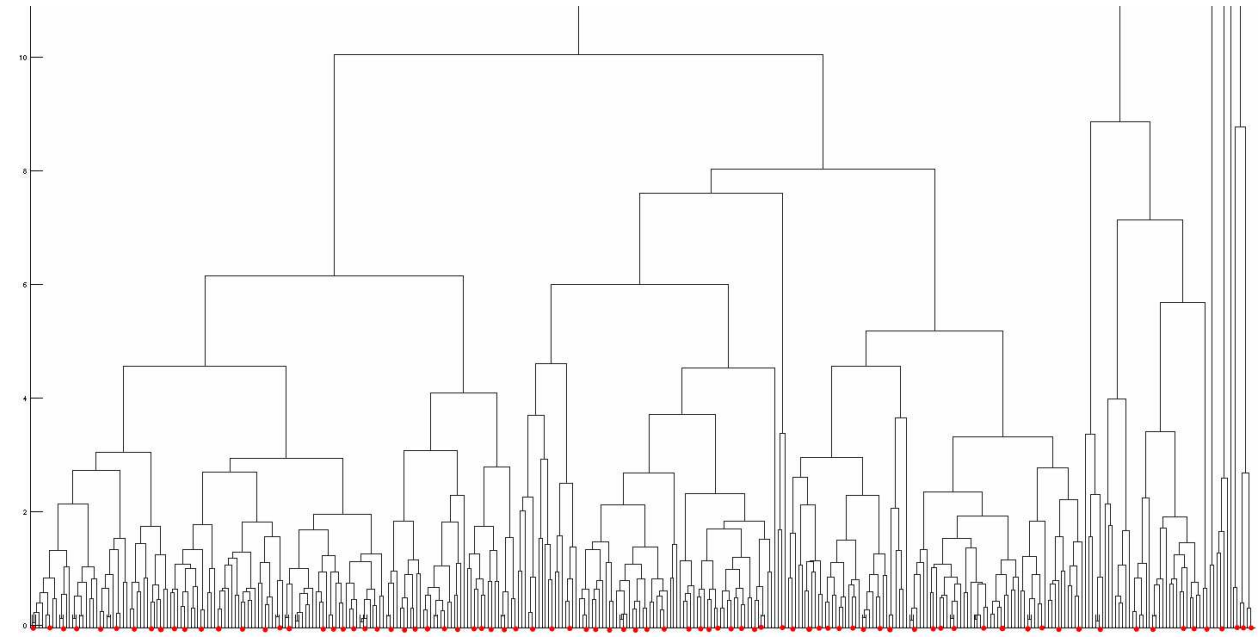
Para definir o conjunto externo foi usado a HCA para construir dendrogramas utilizando um conjunto de descritores de modelos avaliativos que continham todas as amostras. Procurou-se selecionar amostras ao longo de todo dendrograma evitando escolher as que estavam isoladas em agrupamentos. Tal cuidado garante que as amostras escolhidas para compor o conjunto de previsão não retirem do conjunto de treinamento moléculas com características estruturais singulares. Espera-se assim que o conjunto externo esteja dentro do domínio de aplicabilidade do modelo final. A **Figura 22** ilustra as amostras selecionadas.



(1)



(2)



(3)

Figura 22. Dedrogramas obtidos com HCA para os conjuntos de dados (1), (2) e (3) utilizados mostrando a escolha uniforme das amostras que formaram o conjunto de dados externo.

Algoritmo CDDA

O algoritmo CDDA, proposto para fornecer um parâmetro que medisse a semelhança da distribuição de **y** e de um descritor, se mostrou bem sucedido. Quanto mais próximo e_j fosse de 1,0 mais parecidas seriam as distribuições. A **Figura 23** ilustra um exemplo onde são mostrados os valores de e_j para um conjunto de descritores. Nesta figura são destacados dois destes com valores de e_j extremos. Pode se observar claramente a capacidade que o algoritmo CDDA teve para distinguir os descritores. Assim, os descritores que apresentam e_j menor que 0,5 foram eliminados da matriz de descritores após o passo de corte pela correlação.

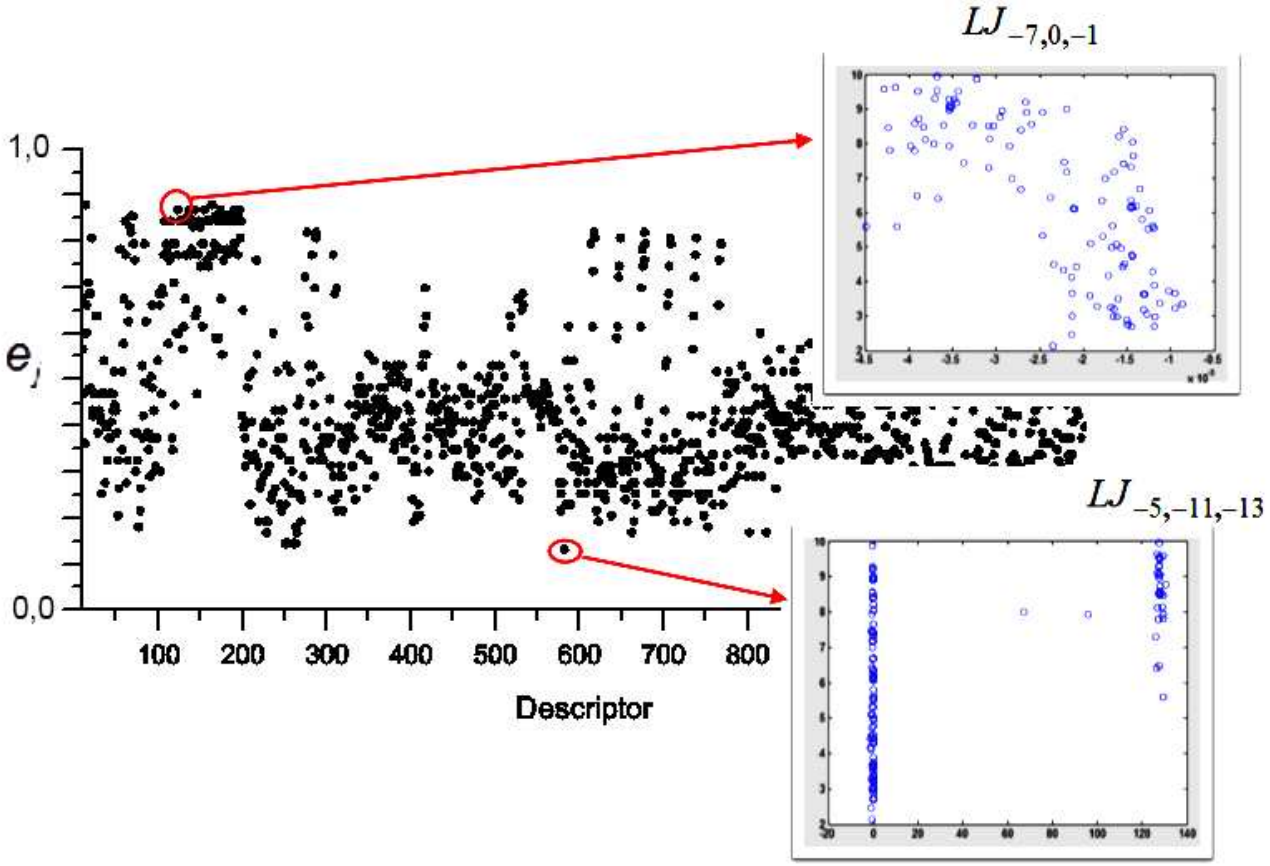


Figura 23. Capacidade do valor e_j em distinguir descritores bem a mal distribuídos.

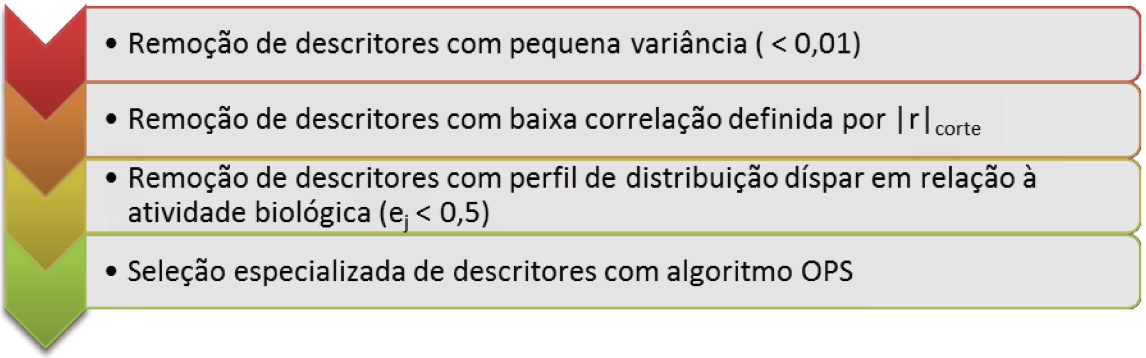
Modelos após filtragem

A **Tabela 4** ilustra a capacidade de eliminação de descritores que cada etapa de filtragem possui. O corte pela variância permite a eliminação de cerca de dois terços dos descritores iniciais, originando a matriz $\mathbf{X}_{var}$. O corte pela correlação é guiado pelo valor de $|r|_{corte}$, que foi estendido para o valor mínimo de 0,3 [19]. Por fim, a filtragem de descritores com perfil de distribuição díspar em relação à atividade biológica (CDDA) proporciona a produção de uma matriz de descritores com tamanho comparável àquelas comuns ao QSAR clássico.

Tabela 4. Número de descritores iniciais e após cada procedimento de filtragem.

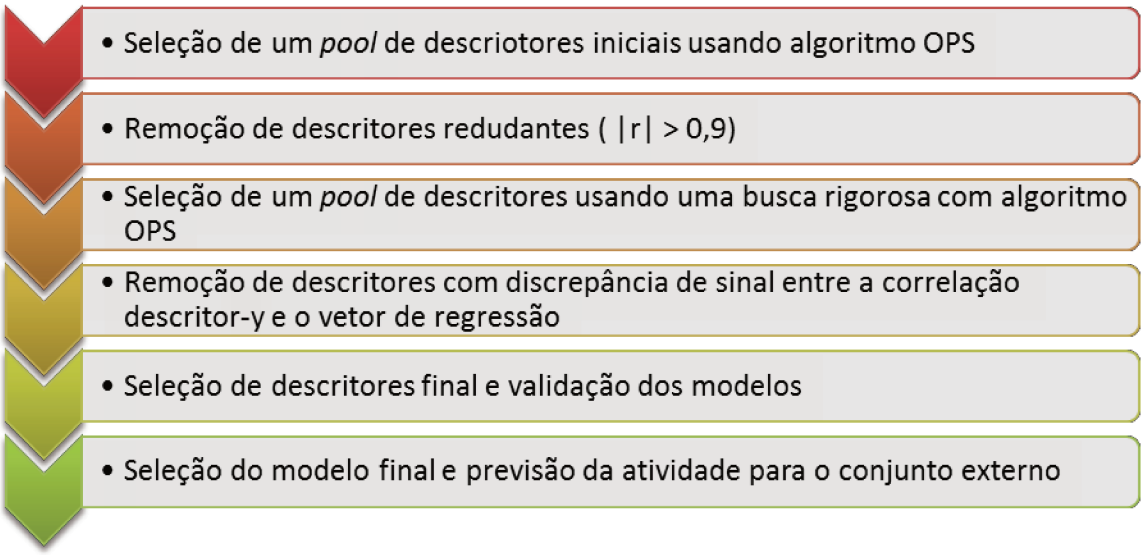
Conjunto de dados/amostras	Descritores iniciais	Corte pela variância (0,01)	$\|r\|_{corte}$	Descritores > $\|r\|_{corte}$	Descritores com $e_j > 0,5$
(1) / 30	162.922	69.518	0,43	14.841	2.825
(2) / 84	158.400	60.512	0,26 (0,3)	29.047	2.156
(3) / 266	112.800	38.516	0,14 (0,3)	2297	277

Após diversos testes realizados com os conjuntos de dados escolhidos foi possível traçar um protocolo para a filtragem dos descritores ilustrada no esquema abaixo.



Também foi traçado um procedimento para a seleção de descritores com o algoritmo OPS para garantir que modelos finais que não apresentassem descritores redundantes. Ou

seja, ao invés de usar a representação usual de modelos CoMFA que expressa mapas de contorno constituído de diversos descritores altamente intercorrelacionados, os modelos com o protocolo otimizado usaram apenas os pontos da grade mais informativos "*hot spots*". Também, durante o processo de seleção de variáveis os descritores com ***problema de sinal*** foram eliminados manualmente.



Os resultados obtidos com o procedimento completo e otimizado é apresentado na **Tabela 5** que fornece uma visão geral comparativa com os modelos da literatura. Nas próximas seções serão mostrados os detalhes de cada modelo, e as porções da grade virtual eliminadas em cada passo de filtragem. Também será reportado o procedimento otimizado para a seleção de descritores com o algoritmo OPS [45].

Tabela 5. Figuras de mérito para os modelos finais obtidos para cada conjunto de dados.

Conjunto de dados	Q^2_{LOO}	R^2	Q^2_{pred}	*SEP*	*ND*[a]	*NVL*[b]
(1)	**0,96**	0,97	**0,73**	5,49	5	4
Melville *et al*. [81]	**0,78**	0,94	**0,64**	-	-	4
(2)	**0,76**	**0,79**	0,68	**1,10**	6	2
Depriest *et al*. [79]	**0,66**	**0,77**	-	**1,31**	-	-
(3)	0,59	0,62	**0,60**	0,88	9	9
Sutherlan *et al*. [98]	0,65	0,76	**0,52**	-	-	7

[a]. Número de descritores incluídos no modelo final. [b] Número ótimo de variáveis latentes no método PLS. Os caracteres em negrito destacam as principais comparações entre os modelos obtidos e os modelos da literatura. Nem todos os valores estão reportados nos artigos originais sendo marcados por um traço.

Na seção anexa desta tese são mostrados todos os erros relativos para as previsões feitas pelos modelos finais obtidos.

Modelo para o conjunto de dados (1)

O conjunto de dados (1) é constituído de 40 catalisadores de fase de amônio quaternário. Tais compostos são usados para promover a alquilação enantiosseletiva de iminas glicinas [99]. Todos os catalisadores dividem a mesma estrutura básica com variações no nitrogênio quaternário. O excesso enantiomérico proporcionado por tais catalisadores compõe a variável dependente.

Esse conjunto de dados em particular apresentou um problema com as estruturas fornecidas na informação suplementar do artigo original. As cargas atômica das estruturas disponibilizadas pelo artigo de Melville *et al*. [81] tiveram que ser alteradas para as do método AM1-BCC [100]. Nenhum modelo com boa capacidade de previsão foi obtido com as cargas originais. Ao que parece, as cargas reportadas pelos autores estão de alguma forma

equivocadamente atribuídas. Portanto, todos os resultados apresentados são para modelos com arquivos de entrada com as cargas atômicas modificadas.

A matriz $\mathbf{X}_{var}$ contendo 69.518 descritores foi submetida ao corte de correlação eliminando os pontos com correlação menor que $|r|_{corte}$ (**Tabela 5**). A **Figura 24** mostra os descritores resultantes para os blocos de potenciais LJ e QQ. É possível observar a permanência de descritores nas mais variadas posições na grade virtual.

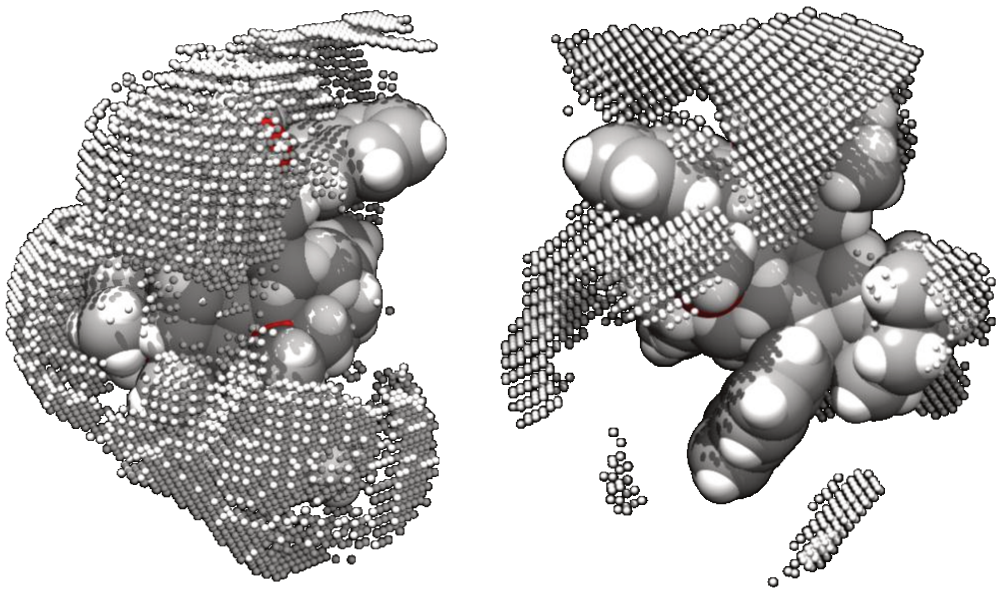


Figura 24. Corte de correlação aplicada à matriz $\mathbf{X}_{var}$ nos descritores de LJ (à esquerda) e QQ (à direita).

O filtro CDDA proporcionou uma grande redução do número de descritores após corte pela correlação (**Tabela 4**). Os descritores restantes evidenciaram regiões ao redor do formato das moléculas alinhadas de uma maneira automatizada. A **Figura 25** ilustra claramente esta tendência.

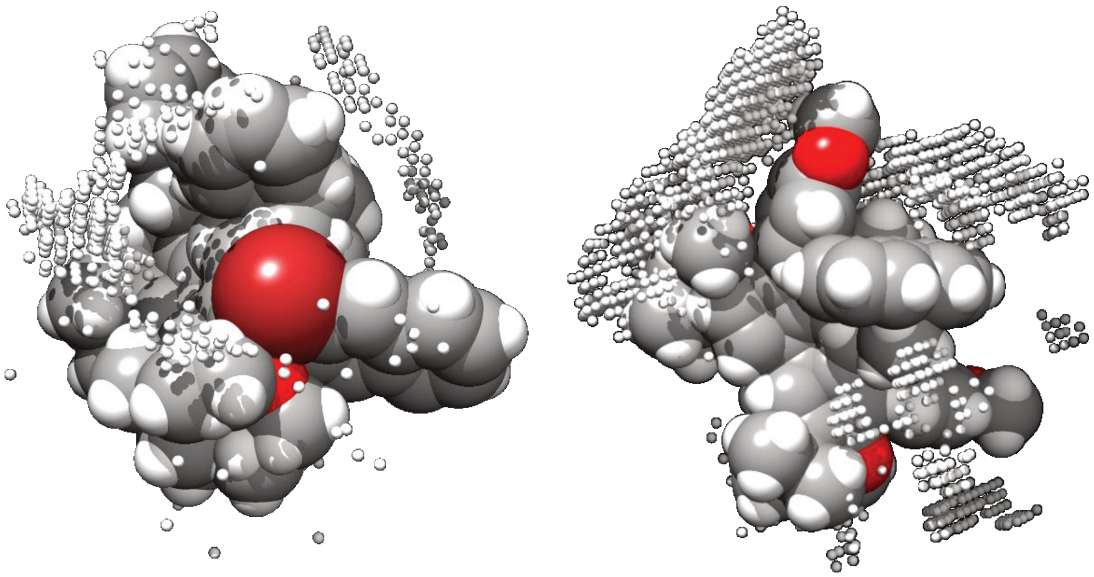


Figura 25. Descritores resultantes após a filtragem CDDA aplicada com descritores LJ (à esquerda) e QQ (à direita).

Após essas duas filtragens a matriz de descritores de LJ e QQ estavam prontas para a seleção de variáveis especializada. Os 2.825 descritores do conjunto de treinamento foram submetidos ao protocolo padrão otimizado de seleção de descritores aplicando o método OPS e, ao final, chegou-se a um modelo que apresentava valores de Q^2_{LOO} 0,96 e R^2 0,97. O modelo mostrou ser bastante simples com apenas 5 variáveis e 4 variáveis latentes. O valor de Q^2_{pred} embora um pouco menor, 0,73, pode ser considerado de qualidade suficiente para prever o excesso enantiomérico proporcionado pelos catalisadores para o conjunto externo com certa exatidão.

Os cinco descritores obtidos são apresentados na **Figura 26**. Pode-se perceber a predominância de descritores eletrostáticos cercando as porções onde foram feitas as modificações estruturais na série. Foi selecionado apenas um descritor de LJ, que apresentou uma correlação positiva com **y**.

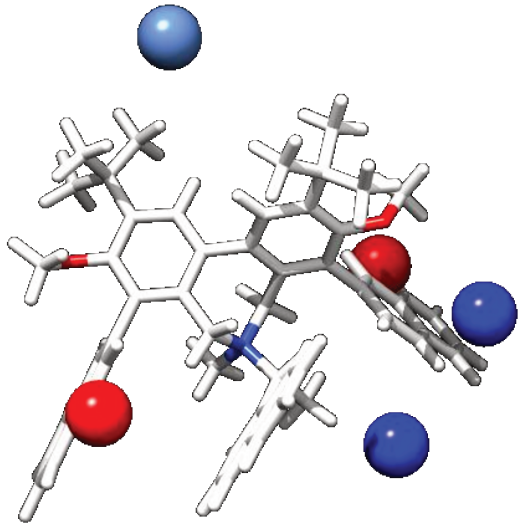


Figura 26. Modelo final para o conjunto de dados (1). As esferas azuis denotam descritores com correlação positiva em relação à atividade biológica e as vermelhas, o contrário. Em vermelho e azul escuro são mostrados os descritores QQ e em azul claro o descritor de LJ.

O modelo final apresentado foi rigorosamente validado para certificar que não fora obtido por acaso e que não apresentava sobreajustes. A **Figura 27** mostra um gráfico dos valores previstos e experimentais para o conjunto de treinamento (azul) e validação (vermelho) no quadrante superior esquerdo. A validação LNO demonstra a robustez da validação interna onde o desvio máximo dos valores Q^2_{LOO} e Q^2_{LNO} não passe de ±0,05. O teste **y**-randomization (quadrante inferior) mostra que os interceptos da regressão linear dos valores de Q^2_{LOO} e R^2 para os modelos obtidos com **y** aleatório e real, como descrito na metodologia, apresentam valores que excluem a possibilidade de o modelo ter sido obtido ao acaso.

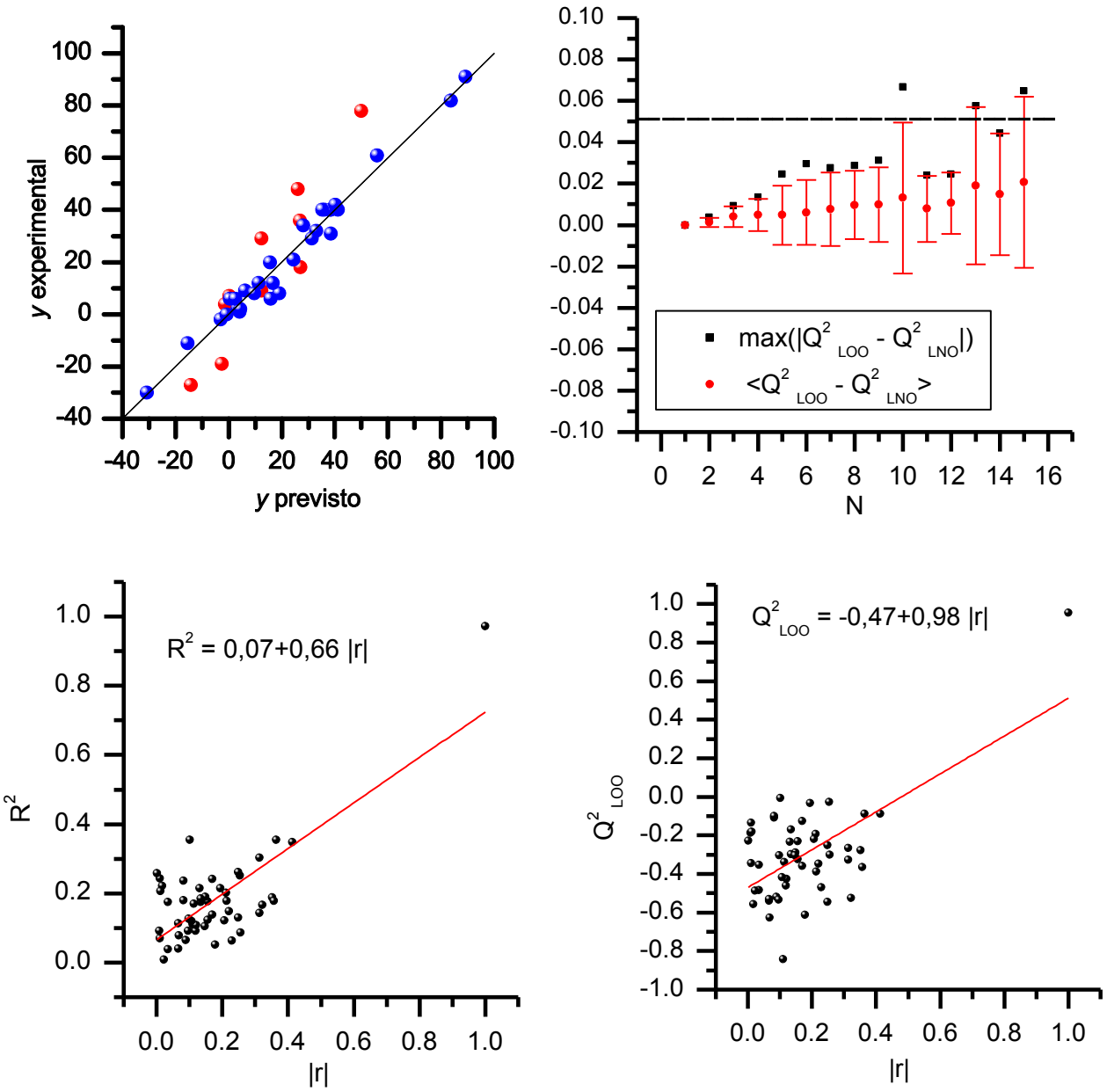


Figura 27. Gráficos apresentando a qualidade de previsão do modelo e os resultados para os testes de validação **y**-randomization e LNO para o modelo para o conjunto de dados (1). As barras de erro denotam dois desvios padrão para 20 rearranjos dos dados.

Para ilustrar a capacidade de obtenção de descritores com boa distribuição usando o CDDA a **Figura 28** mostra gráficos bivariados de y e cada descritor para os modelos do

conjunto de dados (1). Como esperado, as tendências observadas são interessantes, e os descritores se mostraram dispersos ao longo de y.

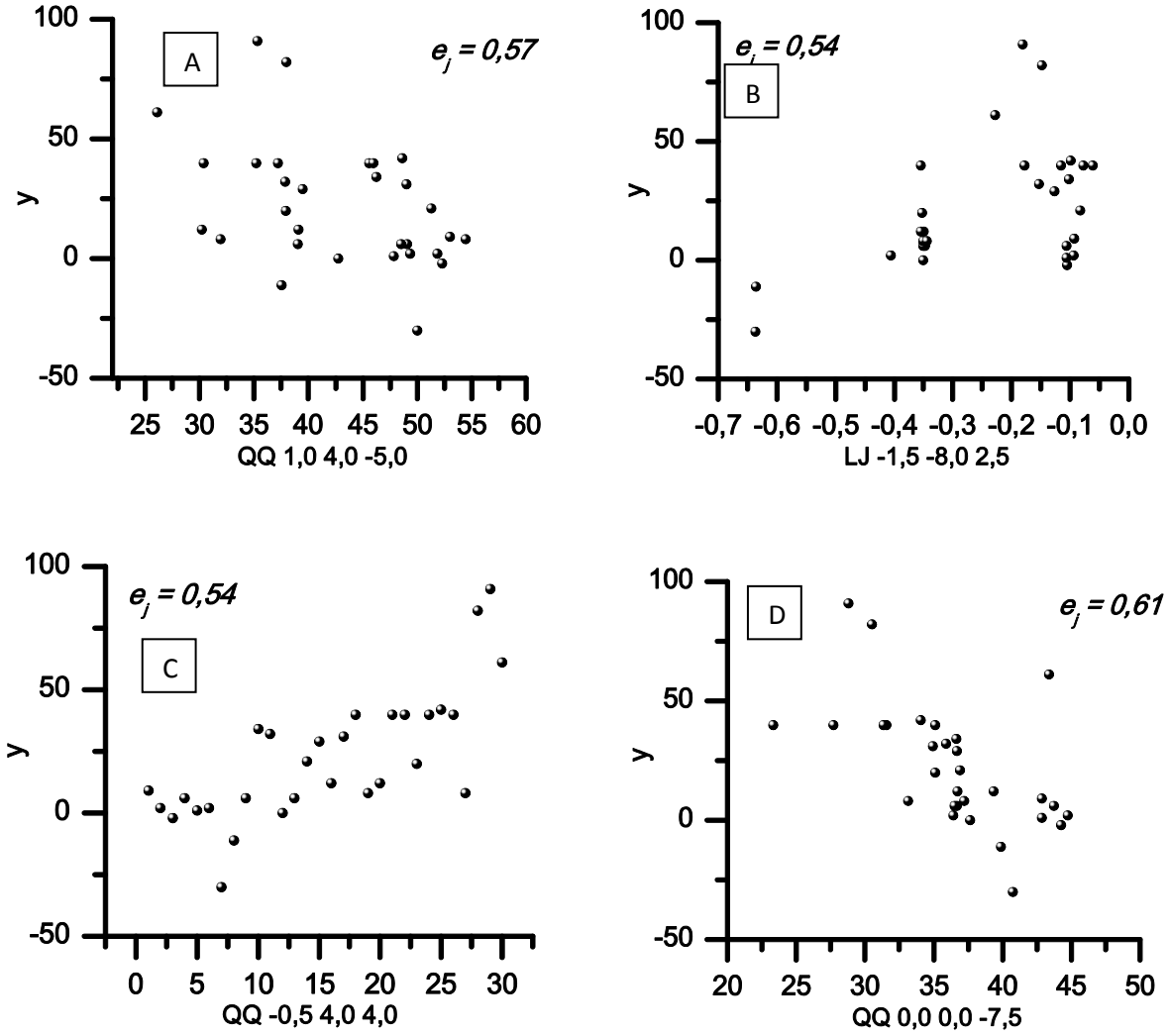


Figura 28. Perfil de distribuição dos descritores (A-D) do modelo final para o conjunto de dados (1). Fica evidente a capacidade do parâmetro e_j em selecionar descritores mais bem distribuídos.

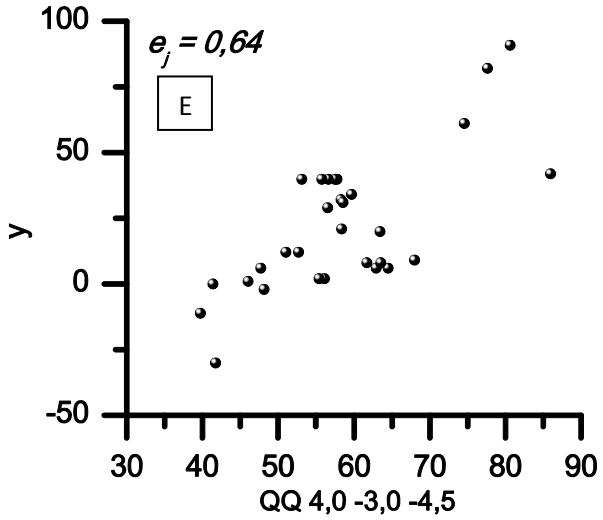


Figura 28. cont. Perfil de distribuição dos descritores (E) do modelo final para o conjunto de dados (1). Fica evidente a capacidade do parâmetro e_j em selecionar descritores mais bem distribuídos.

No artigo original, onde foi publicado o modelo CoMFA [81], consta a utilização de dinâmica molecular para explorar a liberdade conformacional dos compostos e produzir, o que os autores referiram como QSSR-"3,5D" (Quantitative Structure-Selectivity Relationship). Atribuindo um peso para cada conformação segundo a distribuição de Boltzmann, os autores obtiveram descritores MIF ponderados para cada conformação. De modo a evitar conflito entre os conceitos do LQTA-QSAR e a abordagem do artigo, foram selecionados apenas os arquivos que continham apenas uma conformação. Foi feita a comparação somente com o modelo de QSAR-3D.

Não foi possível encontrar uma interpretação para o modelo pela comparação com o artigo original onde nem o mecanismo de ação desses catalisadores nem o modelo CoMFA são discutidos em maior profundidade pelos autores.

Modelo para o conjunto de dados (2)

O conjunto de dados (2) se trata de compostos da classe de fármacos principalmente indicada para o tratamento da hipertensão e prevenção de paradas cardíacas, que são os inibidores da enzima conversora da angiotensina (ECA). Esses compostos exercem os efeitos hemodinâmicos principalmente pela inibição do sistema renina-angiotensina

causando vasodilatação e aumento moderado da eliminação de potássio sem afetar o ritmo e contração cardíaca. O captopril, enalapril e lisinopril são exemplos destes inibidores que previnem a conversão de angiotensina I em angiotensina II, que é um agente vasoconstritor.

O limite mínimo para correlação imposto ao conjunto de dados (1) foi 0,26, que foi aumentado para 0,3, para obter descritores de maior qualidade. Esse corte propiciou a redução para quase metade do número de descritores em $\mathbf{X}_{var}$ (**Tabela 5**). A **Figura 29** mostra os descritores resultantes, destacando que os descritores QQ foram quase completamente eliminados de uma das faces da molécula.

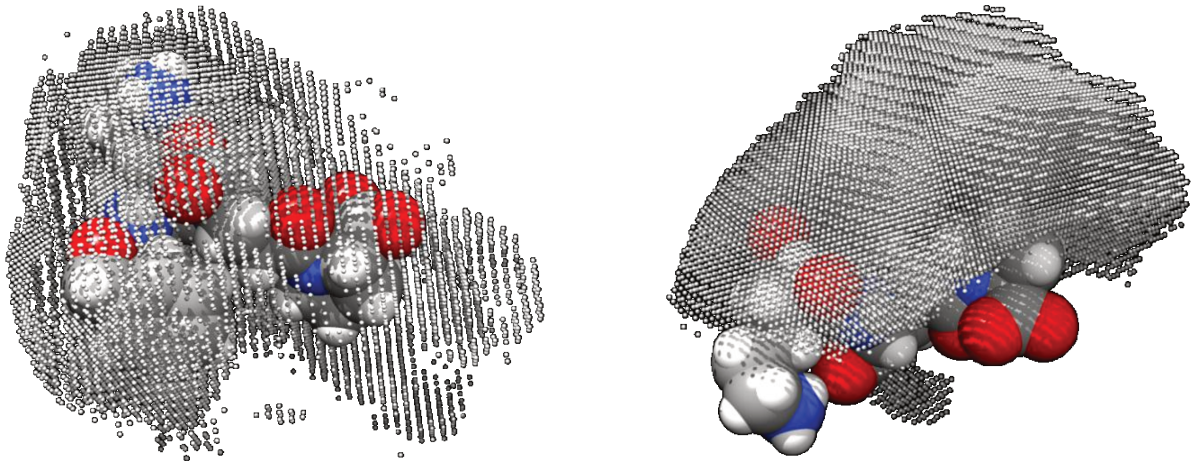


Figura 29. Corte de correlação aplicado à matriz $\mathbf{X}_{var}$ nos descritores de LJ (à esquerda) e QQ (à direita).

O filtro CDDA propiciou a redução bastante acentuada dos descritores. A matriz de entrada teve o número de descritores diminuído em 13 vezes nessa etapa de filtragem (**Tabela 4**). Além disto, tal corte propiciou a evidenciação de regiões ao redor do formato das moléculas alinhadas de uma maneira automatizada. A **Figura 30** ilustra claramente esta tendência.

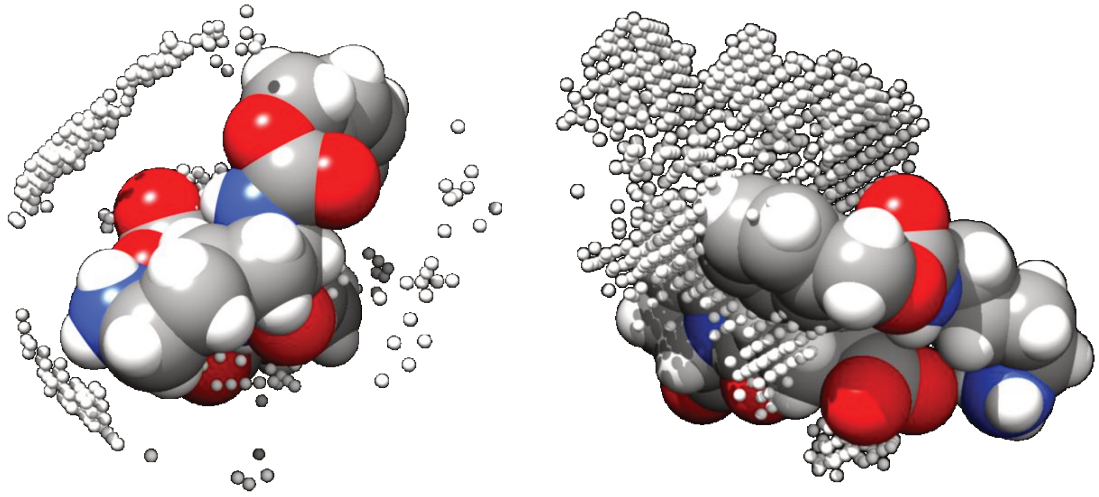


Figura 30. Descritores resultantes após a filtragem CDDA aplicada nos descritores LJ (à esquerda) e QQ (à direita).

Depois de realizadas as filtragens a matriz de descritores QQ e LJ estavam prontas para a seleção de variáveis utilizando o algoritmo OPS [45]. No final chegou-se a um conjunto de descritores que tinha Q^2_{LOO} de 0,76 e se mostrou superior ao valor da literatura [79] de 0,66 (**Tabela 5**). Porém, o resultado mais interessante foi que a capacidade de previsão da atividade para o conjunto externo foi bastante razoável (Q^2_{pred} igual a 0,68). Esses resultados permitem demonstrar claramente a utilidade da filtragem prévia dos descritores MIF antes da construção de modelos QSAR-3D. O modelo final apresentou 6 descritores e 2 variáveis latentes (**Tabela 5**). A **Figura 31** apresenta os descritores finais que ilustram regiões do espaço ao redor das moléculas que são propícias para a modificação estrutural.

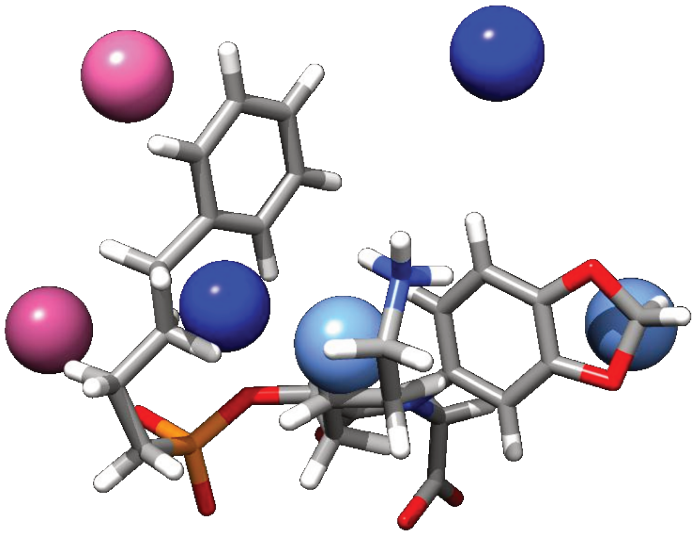


Figura 31. Modelo final para o conjunto de dados (2). As esferas azuis denotam descritores com correlação positiva em relação à atividade biológica e as rosas o contrário. Em azul escuro são mostrados os descritores QQ e em rosa e azul claro os descritores de LJ.

O modelo final apresentado foi rigorosamente validado. A **Figura 32** mostra um gráfico dos valores previstos e experimentais para o conjunto de treinamento (azul) e validação (vermelho) no quadrante superior esquerdo. A validação LNO demonstra a robustez da validação interna onde o desvio máximo dos valores Q^2_{LOO} e Q^2_{LNO} não excede ± 0,05. O teste de **y**-randomization (quadrante inferior) mostra que os interceptos da regressão linear dos valores de Q^2_{LOO} e R^2 para os modelos obtidos com **y** aleatório e real, como descrito na metodologia, apresentam valores que excluem a possibilidade de o modelo ter sido obtido ao acaso.

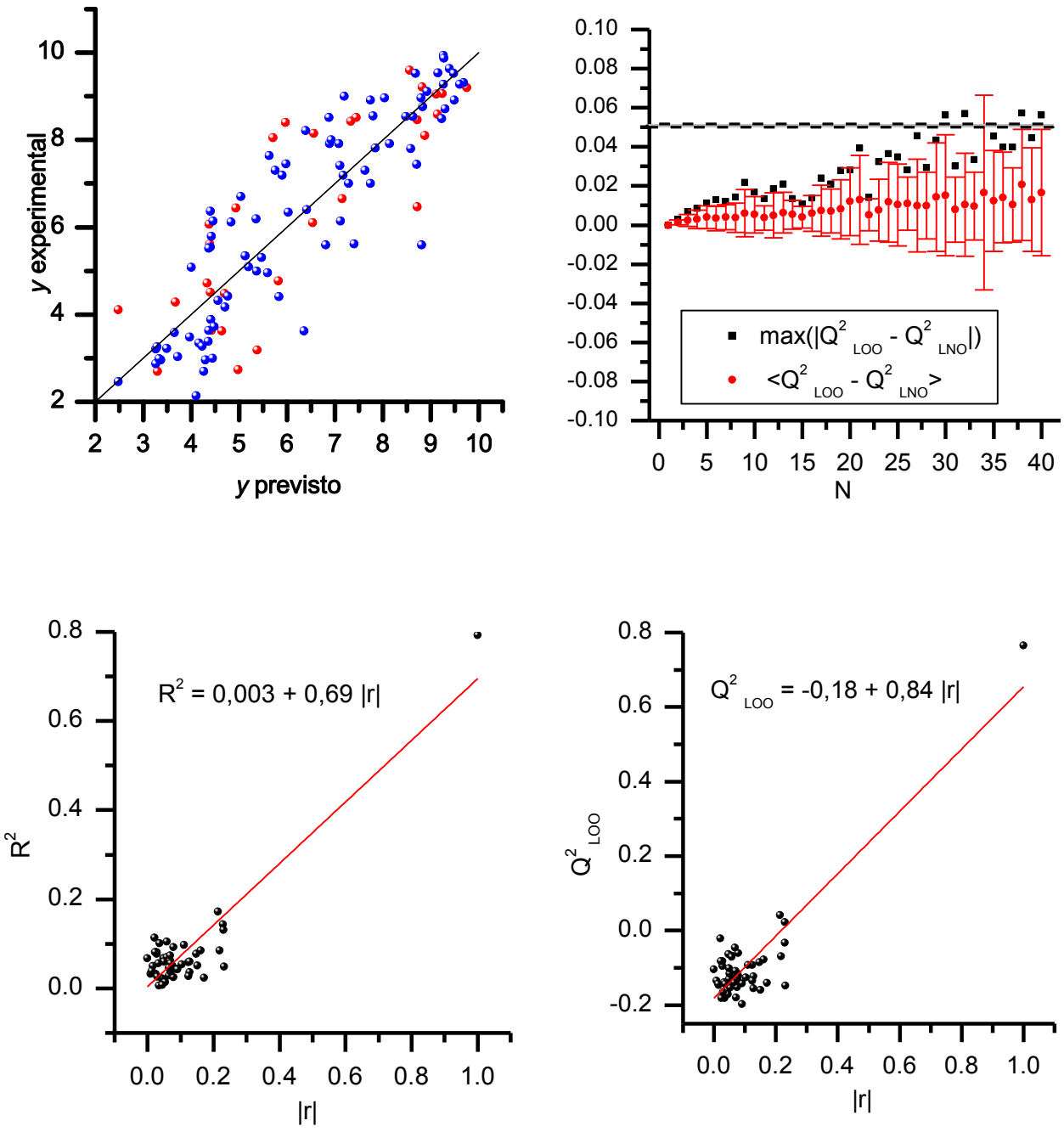


Figura 32.Gráficos apresentando a qualidade de previsão do modelo e os resultados para os testes de validação **y**-randomization e LNO para o modelo para o conjunto de dados (1). As barras de erro denotam dois desvios padrão para 20 rearranjos dos dados.

O artigo de Depriest *et al*. [79] data de um período onde informações cristalográficas a respeito do sítio da ECA eram escassos. Tal informação se mostrou disponível apenas em

2003 com publicação do complexo com o inibidor lisinopril [101]. Se a informação do receptor fosse empregada na construção de modelos os erros sistemáticos do alinhamento poderiam ser minimizados. A literatura apresenta poucos trabalhos empregando essa abordagem [102]. Entretanto, aplicar essa estratégia para a construção dos modelos QSAR-3D foge aos objetivos dessa tese.

Esse conjunto de dados em particular foi utilizado para testar o impacto da não utilização do filtro CDDA. O primeiro grande problema pelo não uso desse filtro é o uso de uma matriz de descritores muito maior para ser submetido à seleção de descritores com OPS (29.047 descritores sem o uso do filtro CDDA, contra 2.156 quando este é empregado). Vale ressaltar que o modelo obtido sem usar o filtro CDDA teve boa qualidade estatística (Q^2_{LOO} 0,80 e R^2 0,85) e a capacidade de previsão, embora menor (Q^2_{pred} = 0,63), ainda era razoável. O pior resultado desse modelo foi no teste LNO. O modelo se mostrou totalmente instável à validação interna com desvios máximos de Q^2_{LOO} e Q^2_{LNO} extrapolando e muito o limite de 0,1. A **Figura 33** mostra estes resultados.

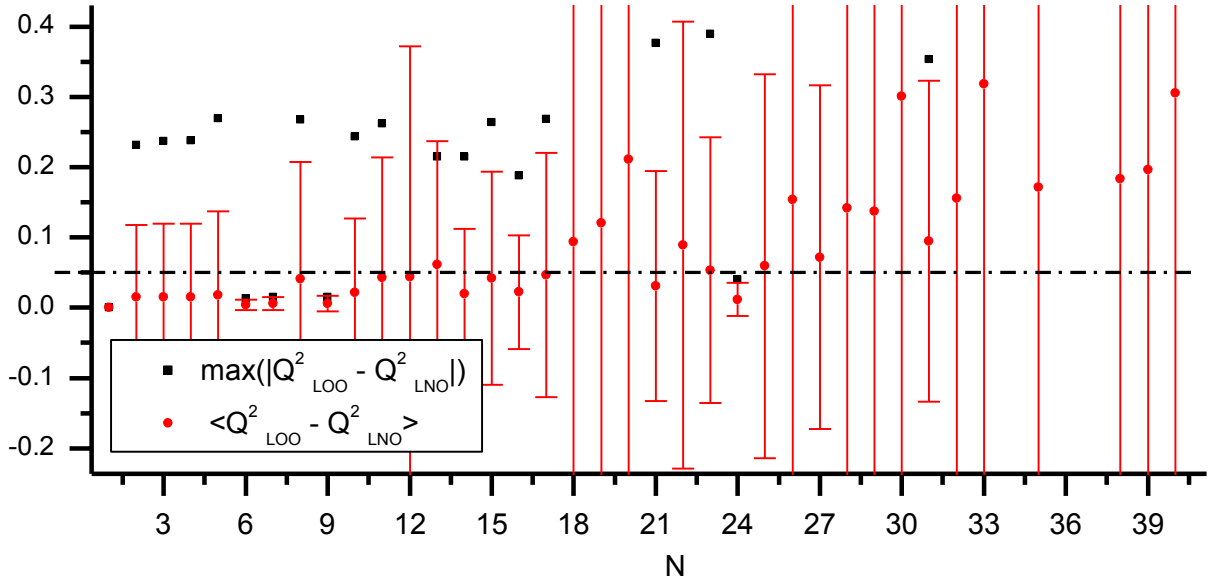


Figura 33. Resultado do teste LNO para um modelo obtido sem o uso do filtro CDDA. As barras de erro denotam dois desvios padrão para 20 rearranjos dos dados.

Ao que parece, o uso de descritores com perfil de distribuição díspar em relação à variável dependente faz com que o modelo fique instável. Isso porque, tais descritores são propensos à deformação do perfil de distribuição conforme um grupo de amostras é deixado de fora. Essa deformação do perfil do descritor pode levar até à inversão da correlação com **y**. Modelos com estes descritores podem ainda gerar uma previsão totalmente equivocada de atividade se um composto produzir um valor para um descritor localizado no espaço vazio da distribuição.

Modelo para o conjunto de dados (3)

A inibição da diidrofolato redutase (DHFR) interrompe a síntese de DNA, levando à morte celular. O inibidor da DHFR metotrexato foi usado como o agente antitumoral, antibacteriano e agente antiprotozoário, mas não é eficiente para o tratamento de patogenias oportunistas causadas, por *Pneumocystis carinii* ou *Toxoplasma gondii,* devido à necessidade de ser transportado ativamente para interior das células. O trimetrexato é mais lipofílico e é absorvido por difusão passiva, porém inibe ativamente a DHFR humana [103]. Uma enorme quantidade de inibidores foi sintetizada para buscar o aumento da seletividade dessa classe de compostos à DHFR exógenas, e constituem o conjunto de dados (3).

Os análogos desse conjunto de dados têm um esqueleto comum muito pequeno ao redor de um núcleo de quinazolina. Os grupos substituintes se espalham amplamente pela grade virtual na matriz de pontos $\mathbf{X}_{var}$. As filtragens de correlação e CDDA propiciaram a concentração dos descritores nas regiões do espaço próximas à porção comum das estruturas (**Figura 34**). Como $|r|_{corte}$ foi aumentado de 0,14 para 0,3, bem menos descritores restaram para a seleção de variáveis com OPS.

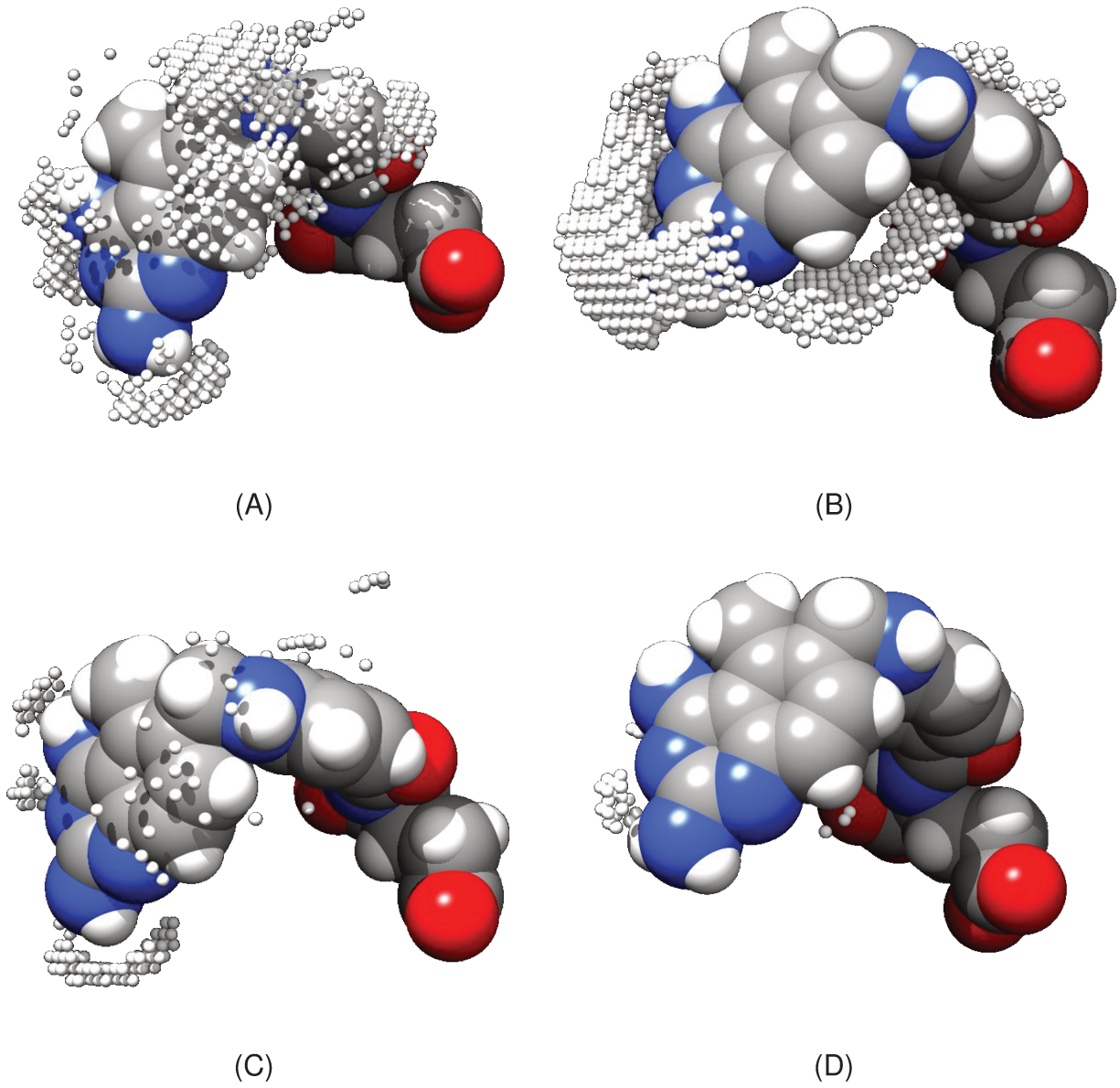


Figura 34. Descritores restantes após o corte de correlação para os descritores LJ (A) e QQ (B) a partir de $\mathbf{X_{var}}$ e após a filtragem CDDA para LJ (C) e QQ(D).

O modelo final apresentou figuras de mérito piores que o artigo original (Q^2_{LOO} de 0,59 e R^2 de 0,62, **Tabela 5**), contudo o poder de previsão do modelo mostrou ser de melhor qualidade (Q^2_{pred} de 0,60 contra 0,52 do artigo original). Como mencionado anteriormente, se o uso da informação do receptor fosse empregado para a realização do alinhamento poder-se-ia obter um modelo com melhor qualidade estatística.

A **Figura 35** mostra os descritores obtidos, que na sua maioria são de LJ com correlação positiva com a atividade biológica. Os descritores eletrostáticos que evidenciam a necessidade da polarização da parte interna da molécula.

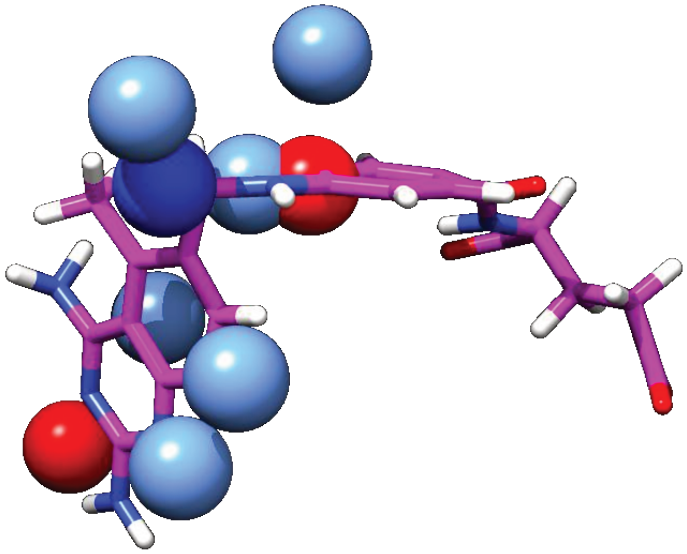


Figura 35. Modelo final para o conjunto de dados (2). As esferas azuis denotam os descritores com correlação positiva em relação à atividade biológica e as vermelhas, o contrário. Em azul escuro e vermelho são mostrados os descritores QQ e em azul claro os descritores de LJ.

O modelo final apresentado foi rigorosamente validado. A **Figura 36** mostra um gráfico dos valores previstos e experimentais para o conjunto de treinamento (azul) e validação (vermelho) no quadrante superior esquerdo. A validação LNO demonstra a robustez da validação interna onde o desvio máximo dos valores Q^2_{LOO} *e* Q^2_{LNO} não passe de ± 0,05. O teste **y**-*randomization* (quadrantes inferiores) mostra que os interceptos da regressão linear dos valores de Q^2_{LOO} e R^2 para os modelos obtidos com **y** aleatório e real, como descrito na metodologia, apresentam valores que excluem a possibilidade de o modelo ter sido obtido ao acaso.

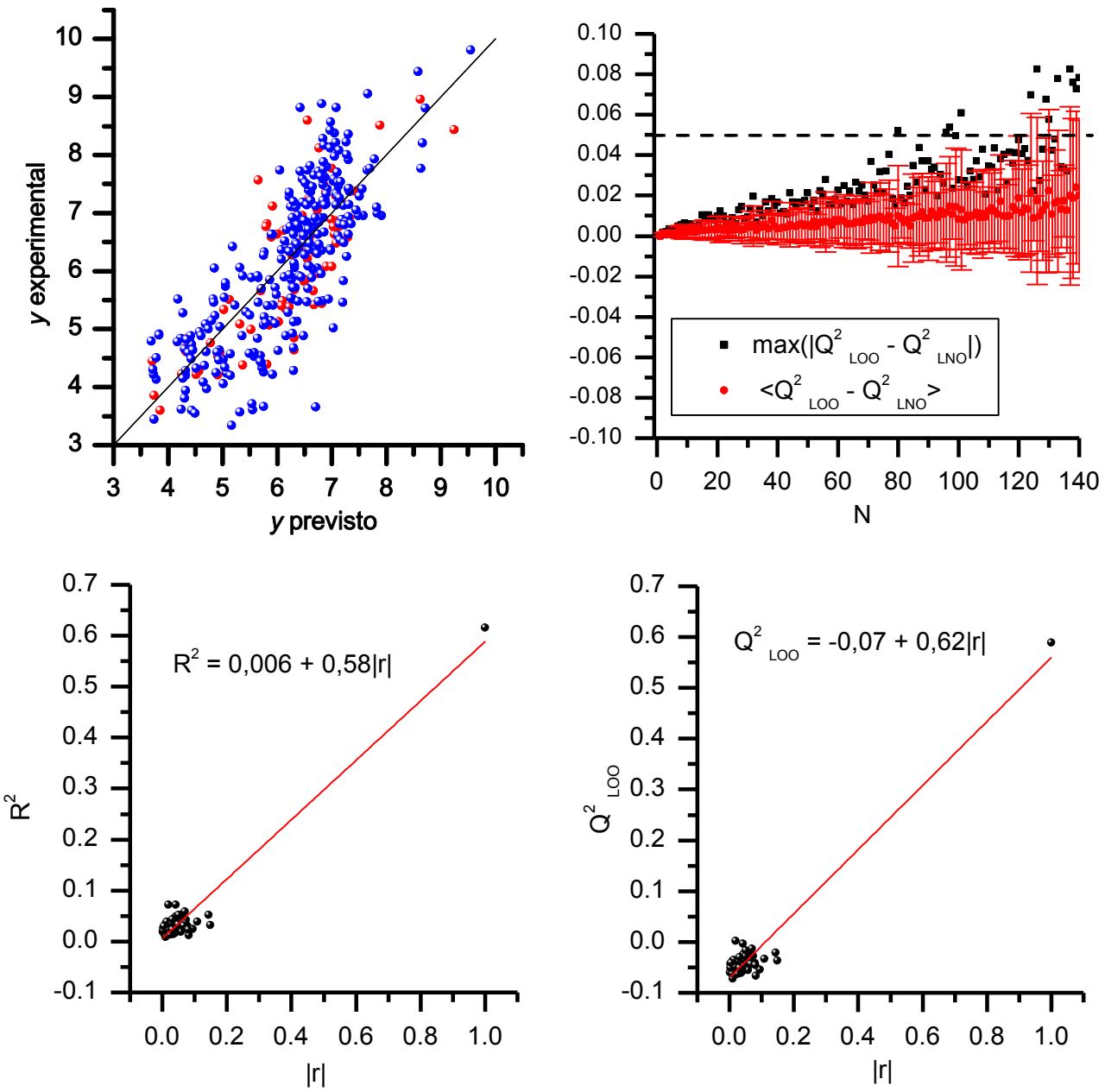


Figura 36. Gráficos apresentando a qualidade de previsão do modelo e os resultados para os testes de validação **y**-randomization e LNO para o modelo para o conjunto de dados (1). As barras de erro denotam dois desvios padrão para 20 rearranjos dos dados.

A interpretação detalhada dos modelos e seus descritores fica além dos objetivos do uso dos conjuntos de dados apresentados nesta tese. Contudo, pode-se afirmar com segurança que o novo procedimento traçado para a criação de modelos com descritores MIF

é bastante útil do ponto de vista da simplicidade em aplicar e o aumento da capacidade de previsão. Esse protocolo otimizado foi estendido para o formalismo LQTA-QSAR-DR onde os resultados são exibidos com seu uso automatizado como se representasse um usuário final da metodologia.

III.2 LQTA-QSAR dependente de Receptores

Como mencionado anteriormente o sítio ativo da TR com o inibidor quinacrina não é apropriado para realizar o procedimento de docagem, dos compostos selecionados para o estudo de QSAR-4D. Assim, foi necessário adaptar o sítio aos ligantes fenotiazinicos. Os próprios autores do artigo onde foi divulgada a estrutura cristalográfica do complexo relatam que o ligante alvo dessa tese se ligaria ao sítio ativo de forma diferenciada, onde o anel tricíclico ficaria voltado para a TRP22 [95] (**Figura 37**).

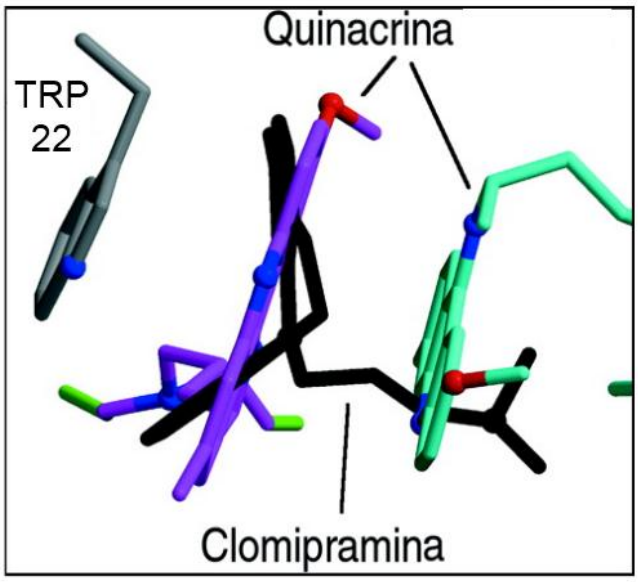


Figura 37. Modo de interação dos ligantes fenotizínicos em comparação com o inibidor quinacrina.

Depois de realizada a primeira etapa de docagem e dinâmica molecular para a adaptação do sítio ativo da TR foram escolhidas quatro poses distintas de A6 no sítio ativo. A avaliação da melhor pose após a simulação foi baseada na energia de interação entre o ligante e o restante do sistema. A melhor pose mostrou uma forte tendência de se enterrar no

sítio ativo rebatendo a MET114 e procurando uma interação aromática do tipo empilhamento "T" com a PHE115 e formação do empilhamento aromático do tipo π-π com TRP22. No entanto esperava-se que a porção do nitrogênio quaternário que concentra boa parte da carga positiva da molécula estivesse mais próxima à GLU19 que é carregada negativamente (**Figura 38**).

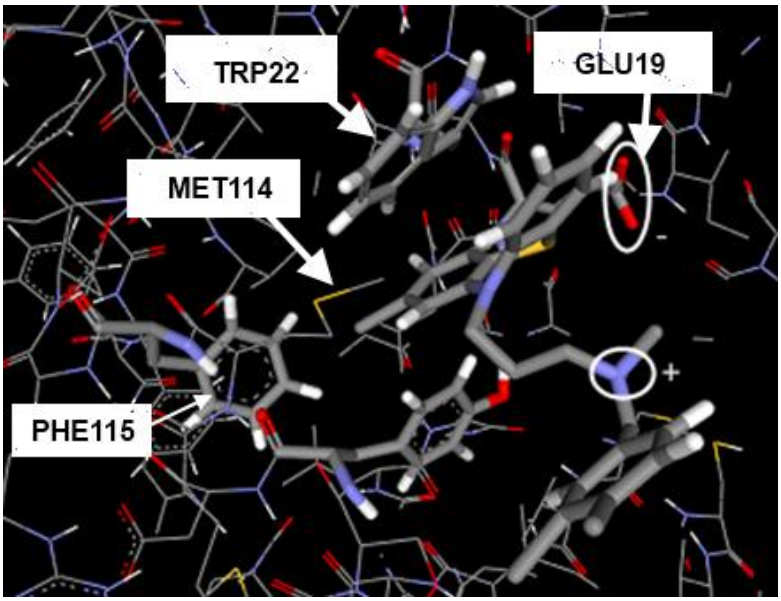


Figura 38. Conformação energeticamente mais favorável no sítio ativo após o primeiro passo de docagem e dinâmica molecular.

Para a segunda etapa de docagem e dinâmica molecular foram escolhidas as conformações iniciais que propiciavam a formação de uma interação íon-íon entre a porção do nitrogênio quaternário e GLU19. Foram escolhidas três poses distintas do ligante A6 (**Figura 39**) com essa aproximação e, subsequentemente, tais conformações foram novamente submetidas à dinâmica molecular.

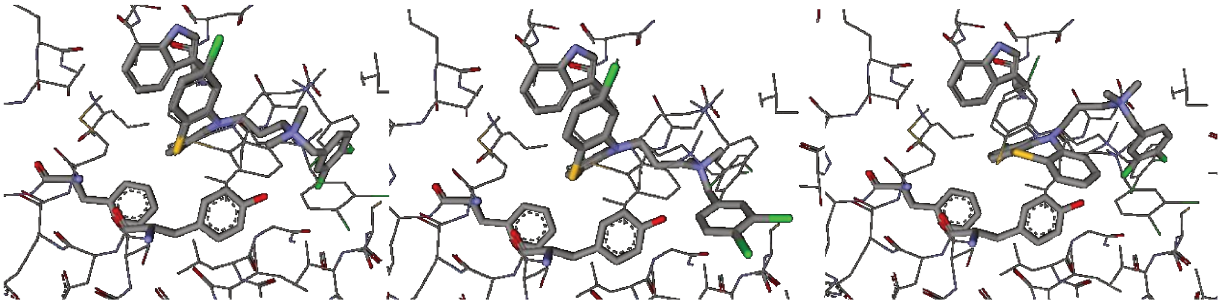


Figura 39. Novas poses escolhidas para uma nova etapa e adaptação do sítio ativo da TR.

Ao final dessa simulação observaram-se interações entre o ligante A6 e o sítio ativo que constituía de: i) Empilhamento - do anel tricíclico com TRP22, ii) Interação aromática em "T" com a PHE115, ii) Uma interação alternante dos empilhamento aromáticos π-π e "T" do anel próximo ao nitrogênio quaternário com TYR111, e iv) Interação íon-íon com GLU19. A **Figura 40** mostra essa rede de interações descrita.

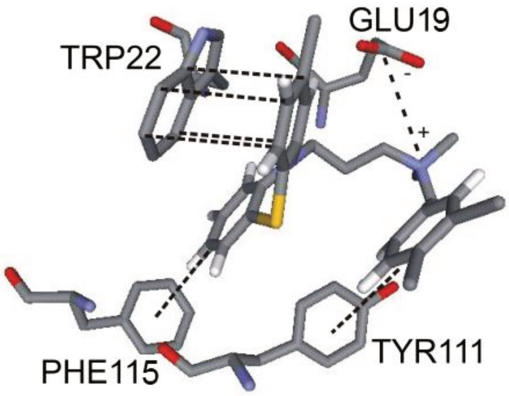


Figura 40. Modo de interação postulado entre A6 e o sítio ativo da TR. Em destaque os resíduos de aminoácidos usados para racionalizar o modo de interação.

Tal rede de interações já havia sido proposta onde o anel tricíclico se aloja contra a parede formada pela TRP22 adjacente à MET114 [67, 89]. Na literatura se observam estruturas de docagem que procuram apenas maximizar a interação com GLU19 [86, 104], contudo, sem levar em conta a flexibilidade do sítio ativo não é possível conseguir essa interação. Os contatos adicionais com a PHE115 (**Figura 39**) ajuda a explicar a diferença de

100 ordens de magnitude a de afinidade por TR de A6 reportados por Khan *et al.* (Ki 0,12 ± 0,01 μM) e a clorpromazina (Ki = 10,8 ± 1,1 μM) [69].

Um estudo recente mostrou uma abordagem similar de docagem e dinâmica molecular que também ajudou a adaptar o sítio ativo e obter um complexo semelhante [104]. Os autores sugeriram um modo de interação semelhante, onde o ligante pode deslocar o substrato $T[S]_2$ que se aloja próximo à ligação dissulfeto no sítio ativo. A concordância entre os dois resultados independentes corrobora a hipótese de obtenção de uma estrutura do sítio ativa apta a ser usada para o LQTA-QSAR-DR.

Foi obtido um complexo TR com todas as moléculas que constituem a série escolhida pela realização da docagem molecular. Os complexos iniciais foram simulados dinamicamente e adaptando a TR para cada ligante estudado. O intervalo de tempo que o ligante levou para se estabilizar dentro do sítio ativo não foi usado para fornecer conformações para o PAC. O desvio médio quadrático (RMSd) dos resíduos de aminoácido do sítio ativo do primeiro frame da simulação em relação aos demais frames foi usado como parâmetro para determinar o intervalo de tempo que cada ligante levou para se acomodar na cavidade. A **Figura 41** mostra três exemplos de RMSd em relação ao tempo destacando que o sítio ativo se estabiliza rapidamente após o inicio da simulação. As conformações acessíveis à dinâmica molecular, quando alinhadas, forneceram os PACs necessários para o LQTAgrid.

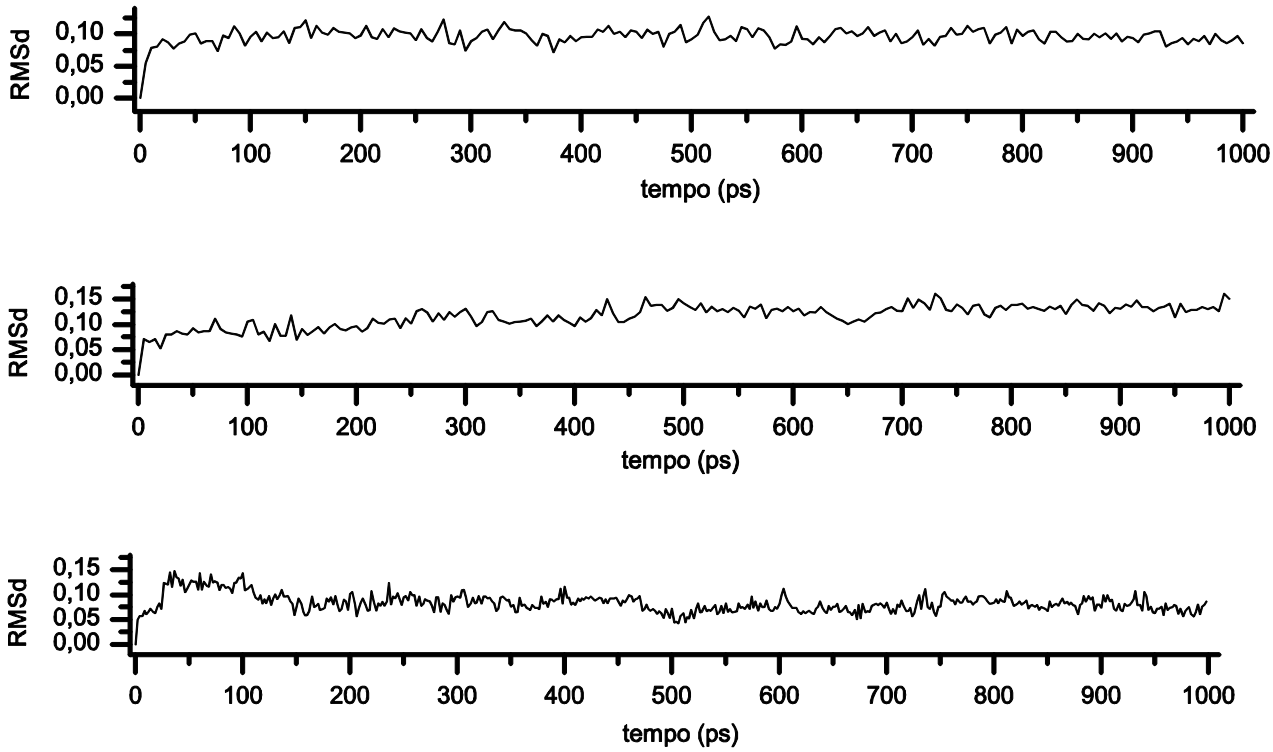


Figura 41. Variação temporal do RMSd do sítio ativo da TR durante simulações de dinâmica molecular para os compostos: (de cima para baixo) A6, clorpormazina e o análogo de A6 com sistema tricíclico extinto.

O alinhamento molecular, feito com átomos tanto dos ligantes como do receptor, resultou em perfis de amostragem conformacional que mantivesse o modo de interação inerente a cada ligante. Tal abordagem permitiu a minimização do ***problema do alinhamento molecular*** muito comum ao QSAR-3D. A **Figura 42** ilustra três perfis de amostragem conformacional obtidos após tal alinhamento. Ao usar a abordagem LQTA-QSAR-DR foram obtidos perfis com variabilidade conformacional mais restrita do que aqueles obtidos na metodologia LQTA-QSAR-IR [43], conforme esperado [43].

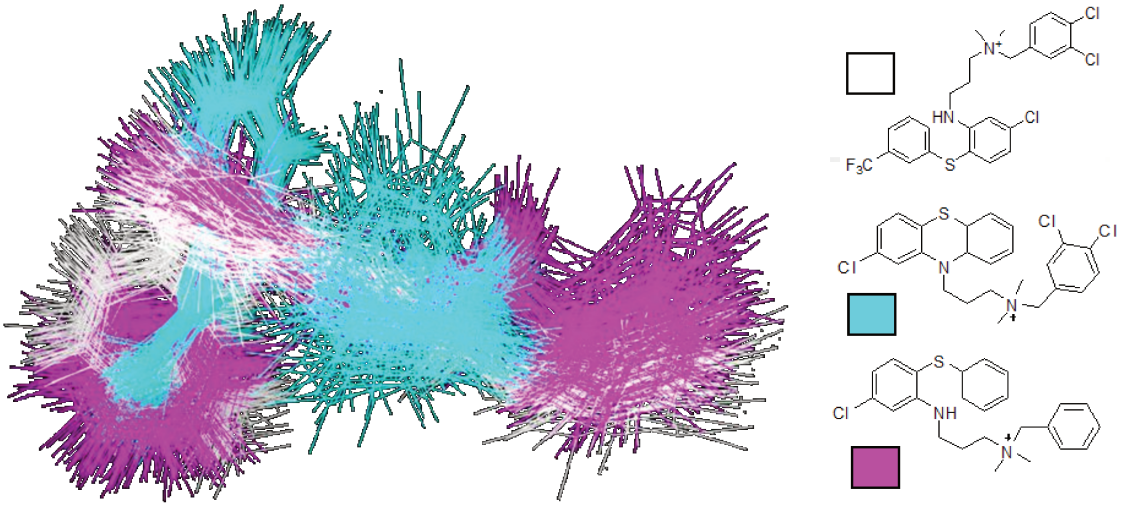


Figura 42. Ilustração de três perfis distintos alinhados utilizando átomos do sítio ativo.

Os descritores obtidos com programa LQTAgrid foram submetidos ao protocolo de filtragem e seleção de variáveis descrito na seção anterior. A matriz de descritores foi tratada como se fosse realizado por um usuário final. O resultado se mostrou bastante satisfatório obtendo-se um modelo com valor de Q^2_{LOO} igual a 0,87 e valor de R^2 igual a 0,92, com apenas 7 descritores e 3 variáveis latentes. A previsão foi também bem satisfatória (Q^2_{pred} = 0,78), apesar de ter sido feita em um número de amostras reduzido (seis), devido a pouca disponibilidade de dados de afinidade para a TR. Outro indicativo de qualidade foi que os erros relativos porcentuais ficaram todos menores que 11% [12].

O modelo final foi rigorosamente validado. A **Figura 43** mostra um gráfico dos valores previstos e experimentais para o conjunto de treinamento (vermelho) e validação (azul) no quadrante superior esquerdo. A validação LNO demonstra a robustez da validação interna onde o desvio máximo dos valores Q^2_{LOO} e Q^2_{LNO} não passam de ± 0,05. O teste **y**-randomization (quadrantes inferiores) mostra que os interceptos da regressão linear dos valores de Q^2_{LOO} e R^2 para os modelos obtidos com **y** aleatório e real, como descrito na metodologia, apresentam valores que excluem a possibilidade de o modelo ter sido obtido ao acaso.

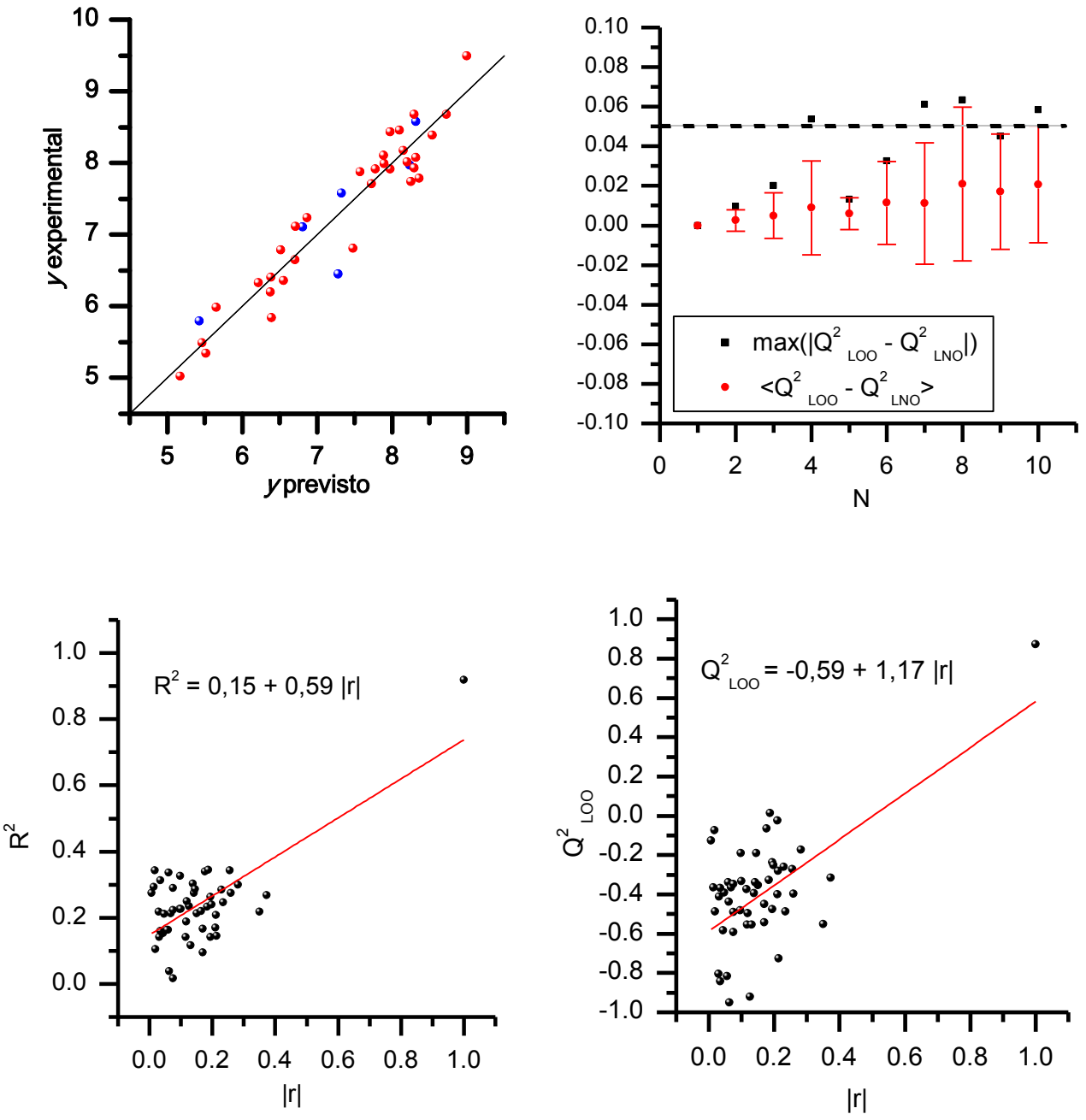


Figura 43. Gráficos apresentando a qualidade de previsão do modelo e os resultados para os testes de validação **y**-randomization e LNO para o modelo para o conjunto de dados (1). As barras de erro denotam dois desvios padrão para 20 rearranjos dos dados [12].

A **Tabela 6** mostra maiores detalhes do modelo final, destacando os valores do vetor de regressão e correlação com a energia livre de interação. Também são exibidas as principais figuras de mérito para o modelo.

Tabela 6. Dados sobre o modelo final do método LQTA-QSAR-DR.

Coordenada do Descritor	**Vetor de regressão**	***r***
39,0 25,0 18,0 LJ	-0,45	-0,30
35,0 23,0 28,0 LJ	-0,33	-0,38
39,0 29,0 27,0 LJ	-0,40	-0,52
36,0 23,0 21,0 LJ	-0,31	-0,47
40,0 28,0 27,0 LJ	0,27	0,42
39,0 28,0 25,0 LJ	-0,35	-0,35
39,0 29,0 22,0 LJ	-0,23	-0,36

Q^2_{LOO}	R^2	Q^2_{pred}	***SEV***	***SEP***	***ND***[a]	***NVL***[b]
0,87	0,92	0,78	0,39	0,37	7	3

[a] ND, número de descritores no modelo e o [b] NVL, número de variáveis latentes.

Os descritores e suas distribuições espaciais são apresentados na **Figura 44**. O modelo final não apresentou contribuições de energia eletrostática. Apenas os descritores de LJ se mostraram capazes de explicar a atividade biológica. Dos setes descritores do modelo final, apenas um apresentou coeficiente de regressão positivo. Os descritores foram classificados em duas categorias, descritores de interação (laranja) e descritores estruturais (vermelho e azul claro). Os descritores estruturais têm valores de energia positiva e se posicionam mais próximos aos átomos dos perfis. Tais descritores variam conforme as estruturas moleculares da série variam.

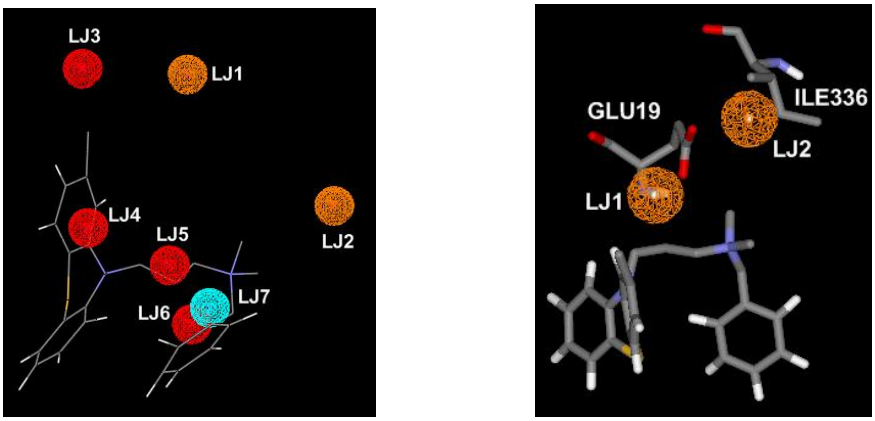


Figura 44. Duas visões esquemáticas da disposição espacial dos descritores do modelo final. À esquerda são mostrados todos os descritores, em azul (*r* positivo), vermelho e laranja (*r* negativo). À direita, um outra disposição ressaltando a relação entre o descritor e os resíduos de aminoácidos dentro do sítio ativo.

Os descritores **LJ1** e **LJ2** podem ser melhor interpretados quando são mostrados simultaneamente com os resíduos adjacentes do sítio ativo da TR. **LJ1** parece estar relacionado à interações com a GLU19. O descritor **LJ2,** que pode ser relacionado com interações com ILE336, e elucida um ponto de interação hidrofóbica que pode ser explorada para guiar a síntese de novos análogos. Tais análogos poderiam, por exemplo, apresentar um maior volume na região do nitrogênio quaternário, aumentando a superfície de contato com a ILE336.

Os descritores **LJ3** à **LJ7** podem ser interpretados como pontos diferenciadores de modificação estrutural dentro da série. **LJ3** está relacionado com as substituições do anel fenotiazínico por grupos Cl e CF_3 (observar **Quadro 1**). **LJ4** está relacionado com a quebra da estrutura tricíclica de alguns análogos. **LJ5** e **LJ6** podem ser relacionados com o tamanho da cadeia que separa o anel tricíclico da porção do nitrogênio quaternário. **LJ7** evidencia as substituições no nitrogênio quaternário por grupos aromáticos que proporcionam interações adicionais com TYR111 no sítio.

Proposta de novos inibidores fenotiazínicos

O método LQTA-QSAR-DR difere da abordagem inicial independente de receptores devido à restrição da liberdade conformacional proporcionada pelas simulações com o

ligante confinado no sítio ativo. O processo de adaptação do sítio por si só já gera um sítio ativo bastante interessante para o desenho de ligantes baseado na estrutura. Além disto, os descritores finais do modelo apresentado se mostraram interessantes para racionalizar as características que inibidores fenotiazínicos devem apresentar para ter boa afinidade pelo sítio da TR. Contudo, a maior dificuldade do método LQTA-QSAR-DR que é oriunda do grande custo computacional para realizar as simulações de dinâmica molecular. O formalismo LQTA-QSAR-DR é particularmente problemático quando na previsão da atividade de um conjunto muito grande de moléculas propostas. Ao contrário da metodologia de QSAR-3D, como o CoMFA, é possível propor centenas de análogos, todos em uma única conformação e assim prever a atividade e selecionar as moléculas mais interessantes com maior atividade prevista e acessibilidade para a síntese. Por isso, apenas alguns poucos análogos foram propostos baseando-se nos descritores do modelo e na estrutura do sítio ativo adaptado.

A **Figura 45** mostra algumas propostas que exploram principalmente os descritores **LJ1** e **LJ2** que elucidam um ponto para extensão do ligante que pode melhorar as características de hidrofobicidade e exploração da interação com a GLU19. Outra característica explorada foi a restrição conformacional do anel aromático que interage com TYR111. A molécula ainda pode ser privada da carga formal +1 já que essa não foi uma característica absolutamente necessária para a atividade, o que foi demonstrado pela ausência de descritores eletrostáticos.

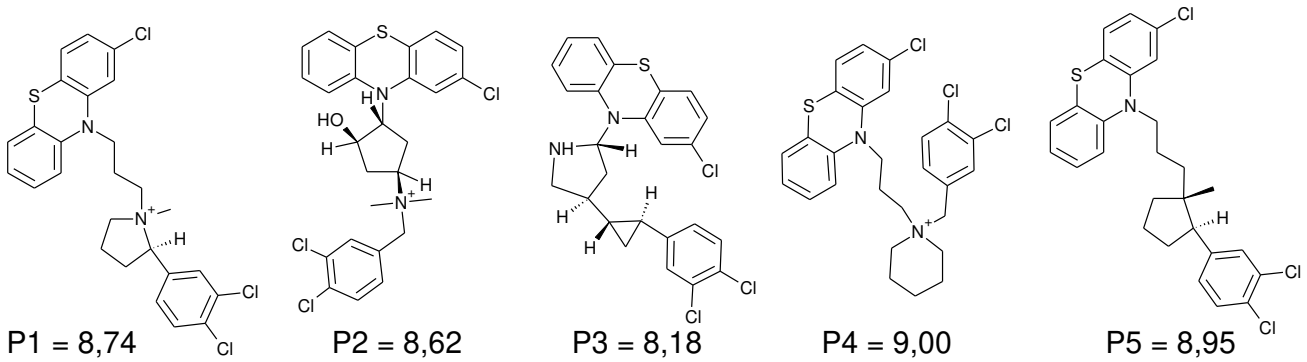


Figura 45. Propostas de estruturas com maior grau de restrição conformacional e energia livre de interação prevista (kcal mol^{-1}).

Foi mencionado na introdução que para esta série de compostos é bastante problemática a falta de correlação entre afinidade por TR e a capacidade tripanossomicida. As propostas feitas podem ser uma alternativa para o aumento da hidrofobicidade que poderia melhorar os perfis farmacocinéticos dessa classe de inibidores.

Capítulo IV

Considerações Finais

IV.1 Filtros digitais para descritores MIF

Com base nos resultados obtidos (**Tabela 5**) com os conjuntos de dados escolhidos da literatura, é possível dizer que os protocolos de filtragem ajudaram na obtenção de modelos QSAR-3D com qualidade superior aos da literatura. A remoção prévia de descritores com o mesmo nível de correlação com o ruído e **y** e de descritores com perfil de distribuição díspar em relação à **y** ajudou na obtenção de modelos robustos e sem correlação ao acaso, de acordo com os testes de validação LNO e **y**-*randomization*.

Foi desenvolvida uma nova forma de tratar os descritores LJ diferente do corte direto usado pelo programa CoMFA. Com tal corte foi possível manter a informação dos descritores próximos à superfície das moléculas alinhadas dentro da grade virtual. Esse tratamento dos descritores LJ, juntamente com a remoção daqueles que possuíam variância reduzida, se mostrou bastante útil para obtenção de descritores interpretáveis que foram úteis para explicar a atividade biológica, e facilitou a proposição de novos compostos. Os modelos QSAR-3D obtidos podem ser vistos e tratados como qualquer modelo de QSAR clássico.

Não obstante a simplicidade do filtro CDDA, tal ferramenta se mostrou bastante útil para a redução do número de variáveis a ser submetido para a seleção de descritores com o algoritmo OPS. O algoritmo CDDA elimina descritores *a priori* fazendo com que os modelos finais tenham uma robustez significativa. Além disto, a junção do filtro CDDA e o corte de correlação foram decisivos para o aumento da estabilidade e do poder de previsão dos modelos para o conjunto de validação externa.

É importante frisar que descritores moleculares derivados de propriedades discretas muito usadas em QSAR, como a contagem do número de átomos na molécula, quantidade de grupos substituintes ou posições destes no esqueleto principal são particularmente sujeitos a eliminação pelo filtro CDDA. Recomenda-se então que o uso deste algoritmo seja restrito a filtragem automatizada de descritores intrinsicamente contínuos como é o caso dos descritores MIF.

Pode acontecer também que **y** não esteja bem distribuído, ou seja, não apresente uma distribuição normal ou quasi-normal. Nestes casos o filtro CDDA tende a selecionar

descritores com o mesmo perfil fornecendo descritores com lacunas na superfície de resposta do modelo. A previsão da atividade para amostras que caem em tais lacunas podem não ser confiáveis. Uma maneira de contornar esse problema é buscar o preenchimento de **y** conferindo variabilidade estrutural às moléculas que constituem a série.

Outra observação interessante foi que a remoção manual de descritores com problema de sinal foi surpreendentemente simples e direta, não conferindo ao modelo final piora nas figuras de mérito do mesmo. Os modelos obtidos têm significado tanto na informação do vetor de regressão com na observação das tendências da atividade, ou seja, pode se afirmar que a confiabilidade dos modelos obtidos é bastante elevada.

Com base nos resultados obtidos foi possível traçar um protocolo totalmente inovador para a construção de modelos que fazem uso de descritores do tipo MIF como QSAR-3D e o LQTA-QSAR. Os modelos originados por com este protocolo são obtidos de forma semiautomática e são explorados em detalhes na seção anexo na forma de dois tutoriais.

IV.2 LQTA-QSAR-DR

Para se aplicar a nova abordagem LQTA-QSAR-DR, introduzida nessa tese, foi necessária a adaptação do sítio ativo da TR para a inserção dos ligantes e obtenção dos PACs por simulações de dinâmica molecular. O sítio ativo otimizado mostrou ser bastante razoável com base no mecanismo de inibição dos compostos tricíclicos derivados da clorpromazina, fornecendo ainda uma estrutura bastante útil para o desenvolvimento de novos ligantes. O método LQTA-QSAR-DR é a evolução natural da metodologia independente de receptores sendo bastante útil para o entendimento do mecanismo de ação dos compostos contidos na série escolhida.

O procedimento utilizado para o alinhamento também se mostrou uma vantagem em relação à abordagem independente do receptor. Ao utilizar os átomos do sítio ativo para orientar os ligantes, foram obtidos PACs que conservavam a maneira característica de associação dos ligantes com o sítio. Isto se refletiu nos descritores e no modelo final que teve excelentes figuras de mérito.

A informação do modelo LQTA-QSAR-DR contido nessa tese se mostrou bastante útil para o desenvolvimento racional de possíveis inibidores da *T. cruzi*. Tais modificações podem conter melhores características de penetração celular como os análogos P3, P4 e P5 (**Figura 44**). Tais estruturas podem propiciar a formação de cooperações científicas futuras a fim de certificar a ação tripanossomicida.

Referências

[1] M. Karelson, V.S. Lobanov, A.R. Katritzky, Chem. Rev., 96 (1996) 1027-1044.
[2] C. Hansch, Drug Development Research, 1 (1981) 267-309.
[3] R.M. Hyde, D.J. Livingstone, J. Comput.-Aid. Mol. Des., 2 (1988) 145-155.
[4] V. Pliška, B. Testa, H. Waterbeemd, Lipophilicity in Drug Action and Toxicology, em: V. Pliska (Ed.) Methods and Principles in Medicinal Chemistry, Zurich, 1996.
[5] A. Leo, C. Hansch, D. Elkins, Chem. Rev., 71 (1971) 525-616.
[6] M.P. Edwards, D.A. Price, E.M. John, Annu. Rep. Med. Chem., Academic Press, 45 (2010) 380-391.
[7] M.M.C. Ferreira, J. Braz. Chem. Soc., 13 (2002) 742-753.
[8] I.T. Jolliffe, J. R. Stat. Soc. Series C (Applied Statistics), 31 (1982) 300-303.
[9] S. Wold, M. Sjöström, L. Eriksson, Chemometr. Intell. Lab., 58 (2001) 109-130.
[10] S. Wold, L. Eriksson, S. Clementi, Chemometric Methods in Molecular Design, Wiley-VCH Verlag GmbH, Weinheim, 2008.
[11] P. Gramatica, QSAR Comb. Sci., 26 (2007) 694-701.
[12] R. Kiralj, M.M.C. Ferreira, J. Braz. Chem. Soc., 20 (2009) 770-787.
[13] A. Tropsha, Mol. Informatics, 29 (2010) 476-488.
[14] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, P. Gramatica, Environ. Health Persp., 111 (2003) 1361-1375.
[15] K. Baumann, N. Stiefl, J. Comput.-Aid. Mol. Des., 18 (2004) 549-562.
[16] R. Clark, P. Fox, J. Comp-Aid. Mol. Des., 18 (2004) 563-576.
[17] C. Rücker, G. Rücker, M. Meringer, J. Chem. Inf. Mod., 47 (2007) 2345-2357.
[18] L. Eriksson, J. Jaworska, A.P. Worth, M.T. Cronin, R.M. McDowell, P. Gramatica, Environ. Health Persp., 111 (2003) 1361-1375.
[19] R. Kiralj, M.M.C. Ferreira, J. Chemometr., 24 (2010) 681–693.
[20] R.D. Cramer, D.E. Patterson, J.D. Bunce, J. Am. Chem. Soc., 110 (1988) 5959-5967.
[21] G. Cruciani, Molecular Interaction Fields, Wiley-VCH Verlag GmbH & Co. KGaA, Perugia IT, 2006.
[22] H. Kubinyi, Drug Discov. Today, 2 (1997) 457-467.
[23] M. Arakawa, K. Hasegawa, K. Funatsu, Curr. Comput.-Aid. Drug, 3 (2007) 254-262.
[24] M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi, S. Clementi, Quant. Struct.-Act. Rel., 12 (1993) 9-20.
[25] M. Pastor, G. Cruciani, S. Clementi, J. Med. Chem., 40 (1997) 1455-1464.
[26] R. Grohmann, T. Schindler, J. Comput. Chem., 29 (2008) 847-860.
[27] H. Alonso, A.A. Bliznyuk, J.E. Gready, Med. Res. Rev., 26 (2006) 531-568.
[28] Steuber, Holger, Zentgraf, Matthias, Gerlach, Christof, Sotriffer, A. Christoph, Heine, Andreas, Klebe, Gerhard, J. Mol. Biol. 363 (2006), 174-187.
[29] A. Podjarny, A.P. Dejaegere, B. Kieffer, Biophysical Approaches Determining Ligand Binding to Biomolecular Targets, Detection, Measurement and Modelling RSC Publishig, Strasbourg, 2011.
[30] A.J. Hopfinger, S. Wang, J.S. Tokarski, B. Jin, M. Albuquerque, P.J. Madhav, C. Duraiswami, J. Am. Chem. Soc., 119 (1997) 10509-10524.
[31] A.J. Hopfinger, S. Wang, J.S. Tokarski, B. Jin, M. Albuquerque, P.J. Madhav, C. Duraiswami, J. Am. Chem. Soc., 119 (1997) 10509-10524.
[32] A. Vedani, M. Dobler, J. Med. Chem., 45 (2002) 2139-2149.

[33] B.J. Alder, T.E. Wainwright, J. Chem. Phys., 27 (1957) 1208-1209.
[34] A. Rahman, Phys. Rev., 136 (1964) A405.
[35] F.H. Stillinger, A. Rahman, J. Chem. Phys., 60 (1974) 1545-1557.
[36] J.A. McCammon, B.R. Gelin, M. Karplus, Nature, 267 (1977) 585-590.
[37] R. Galeazzi, Curr. Comput-Aid. Drug, 5 (2009) 225-240.
[38] S. David Van Der, L. Erik, H. Berk, G. Gerrit, E.M. Alan, J.C.B. Herman, J. Comput. Chem., 26 (2005) 1701-1718.
[39] M. Christen, P.H. Hunenberger, D. Bakowies, R. Baron, R. Burgi, D.P. Geerke, T.N. Heinz, M.A. Kastenholz, V. Krautler, C. Oostenbrink, C. Peter, D. Trzesniak, W.F. van Gunsteren, J. Comput. Chem., 26 (2005) 1719-1751.
[40] E.J. Sorin, V.S. Pande, Biophys. J., 88 (2005) 2472-2493.
[41] A.J. DePaul, E.J. Thompson, S.S. Patel, K. Haldeman, E.J. Sorin, Nucleic Acids Res., 38 4856-4867.
[42] J. Wang, W. Wang, P.A. Kollman, D.A. Case, J. Mol. Graph. Model., 25 (2006) 247-260.
[43] J.P.A. Martins, E.G. Barbosa, K.F.M. Pasqualoto, M.M.C. Ferreira, J. Chem. Info. Model., 49 (2009) 1428-1436.
[44] C.M. Breneman, K.B. Wiberg, J. Comput. Chem., 11 (1990) 361-373.
[45] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, J. Chemometr., 23 (2009) 32-48.
[46] World Health Organization, Chagas disease (American trypanosomiasis), em: http://www.who.int/topics/chagas_disease/en/ (2010).
[47] C. Chagas, Memórias do Instituto Oswaldo Cruz, 8 (1916) 5-36.
[48] C. Chagas, Memórias do Instituto Oswaldo Cruz, 8 (1916) 37-60.
[49] J.R. Coura, J. Borges-Pereira, Acta Tropica, 115 (2010) 5-13.
[50] Fundação Oswaldo Cruz, Doença de Chagas, em: *www.fiocruz.br* (2011).
[51] V.G. Duschak, A. Couto, Recent Pat. Anti-Infect., 2 (2007) 19-51. RECENT PAT ANTI-CANC
[52] M.N.C. Soeiro, S.L. de Castro, Expert Opin. Ther. Tar., 13 (2009) 105-121.
[53] C.C. Wang, Annu. Rev. Pharmacol. Toxicol., 35 (1995) 93-127.
[54] T.A. Shapiro, Acta Trop., 54 (1993) 251-260.
[55] G.E. Liñares, E.L. Ravaschino, J.B. Rodriguez. Curr. Med. Chem., 13 (2006) 335-360.
[56] M.H. el Kouni, Pharmacol. Therapeut., 99 (2003) 283-309.
[57] M.O.F. Khan, Drug Tar. Insights, 2007 (2007).
[58] A.H. Fairlamb, A. Cerami, Annu. Rev. Microbiol., 46 (1992) 695-729.
[59] C.H. Faerman, S.N. Savvides, C. Strickland, M.A. Breidenbach, J.A. Ponasik, B. Ganem, D. Ripoll, R. L. Krauth-Siegel, P. A. Karplus, Bioorgan. Med. Chem., 4 (1996) 1247-1253.
[60] C.S. Bond, Y. Zhang, M. Berriman, M.L. Cunningham, A.H. Fairlamb, W.N. Hunter, Struct. Fold. Des., 7 (1999) 81-89.
[61] D.S. Berkholz, H.R. Faber, S.N. Savvides, P.A. Karplus, J. Mol. Biol., 382 (2008) 371-384.
[62] B.L. Christina, S. Ilme, K. Wolfgang, F.P. Emil, R.L. Krauth-Siegel, Proteins- Structure Function and Genetics, 18 (1994) 161-173.
[63] Z. Yihong, S.B. Charles, B. Susan, L.C. Mark, H.F. Alan, N.H. William, Protein Sci., 5 (1996) 52-61.
[64] T.J. Benson, J.H. McKie, J. Garforth, A. Borges, A.H. Fairlamb, K.T. Douglas, Biochem. J., 286 (1992) 9-11.

[65] C. Chan, H. Yin, J. Garforth, J.H. McKie, R. Jaouhari, P. Speers, K.T. Douglas, P.J. Rock, V. Yardley, S.L. Croft, A.H. Fairlamb, J. Med. Chem., 41 (1998) 148-156.
[66] E.M. Jacoby, I. Schlichting, C.B. Lantwin, W. Kabsch, R.L. Krauth-Siegel, Proteins-Structure Function and Bioinformatics, 24 (1996) 73-80.
[67] K. Chibale, H. Haupt, H. Kendrick, V. Yardley, A. Saravanamuthu, A.H. Fairlamb, S.L. Croft, Bioorgan. Med. Chem. Lett., 11 (2001) 2655-2657.
[68] R. Fernandez-Gomez, M. Moutiez, M. Aumercier, G. Bethegnies, M. Luyckx, A. Ouaissi, A. Tartar, C. Sergheraert, Int. J. Antimicrob. Ag., 6 (1995) 111-118.
[69] M.O.F. Khan, S.E. Austin, C. Chan, H. Yin, D. Marks, S.N. Vaghjiani, H. Kendrick, V. Yardley, S.L. Croft, K.T. Douglas, J. Med. Chem., 43 (2000) 3148-3156.
[70] H. Luo, Y.-K. Cheng, QSAR Comb. Sci., 24 (2005) 968-975.
[71] S.D. Peterson, W. Schaal, A. Karãn, J. Chem. Inf. Model., 46 (2005) 355-364.
[72] A.H. Asikainen, J. Ruuskanen, K.A. Tuppurainen, Environ. Sci. Technol., 38 (2004) 6724-6729.
[73] J. Xu, S. Huang, H. Luo, G. Li, J. Bao, S. Cai, Y. Wang, Int. J. Mol. Sci., 11 (2010) 880-895.
[74] M. Liu, L. He, X. Hu, P. Liu, H.-B. Luo, Bioorgan. Med. Chem. Lett., 20 7004-7010.
[75] C. Sköld, A. Karlén, J. Mol. Graph. Model., 26 (2007) 145-153.
[76] O. Oltulu, M.M. Yasar, E. Eroglu, Eur. J. Med. Chem., 44 (2009) 3439-3444.
[77] B.E. Mattioni, P.C. Jurs, J. Mol. Graph. Model., 21 (2003) 391-419.
[78] D.J. Maddalena, G.A.R. Johnston, J. Med. Chem., 38 (1995) 715-724.
[79] S.A. DePriest, D. Mayer, C.B. Naylor, G.R. Marshall, J. Am. Chem. Soc., 115 (1993) 5372-5384.
[80] A. Golbraikh, P. Bernard, J.R. Chrétien, Eur. J. Med. Chem., 35 (2000) 123-136.
[81] J.L. Melville, K.R.J. Lovelock, C. Wilson, B. Allbutt, E.K. Burke, B. Lygo, J.D. Hirst, J. Chem Inf. Model., 45 (2005) 971-981.
[82] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, Taylor & Francis, Boca Raton, FL, 2006
[83] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, Arch. Biochem. Biophys., 185 (1978) 584-591.
[84] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, J. Comput. Chem., 25 (2004) 1605-1612.
[85] S. Parveen, M.O.F. Khan, S.E. Austin, S.L. Croft, V. Yardley, P. Rock, K.T. Douglas, J. Med. Chem., 48 (2005) 8087-8097.
[86] D. Horvath, J. Med. Chem., 40 (1997) 2412-2423.
[87] A. Saravanamuthu, T.J. Vickers, C.S. Bond, M.R. Peterson, W.N. Hunter, A.H. Fairlamb, J. Biol. Chem., 279 (2004) 29493-29500.
[88] C.B. Delphine, M.R. David, H.J. Jan, Proteins-Structure Function and Bioinformatics, 73 (2008) 765-783.
[89] Y. Zhao, N.E. Schultz, D.G. Truhlar, J. Chem. Theory Comput., 2 (2006) 364-382.
[90] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, J.A.M. Jr., T. Vreven, K.N. Kudin, J.C. Burant, J.M. Millam, S.S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G.A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J.E. Knox, H.P. Hratchian, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C.

Pomelli, J.W. Ochterski, P.Y. Ayala, K. Morokuma, G.A. Voth, P. Salvador, J.J. Dannenberg, V.G. Zakrzewski, S. Dapprich, A.D. Daniels, M.C. Strain, O. Farkas, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J.V. Ortiz, Q. Cui, A.G. Baboul, S. Clifford, J. Cioslowski, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R.L. Martin, D.J.F. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, C. Gonzalez, J.A. Pople, Gaussian 03, Revision E.01, in, Gaussian, Inc., Wallingford CT, 2004.
[91] K.B.W. Curt M. Breneman, J. Comput. Chem., 11 (1990) 361-373.
[92] A.W. Schuettelkopf, D.M.F.v. Aalten, Acta Crystallogr. D60, (2004) 1355-1363.
[93] M.M. Garrett, S.G. David, S.H. Robert, H. Ruth, E.H. William, K.B. Richard, J.O. Arthur, J. Comput. Chem., 19 (1998) 1639-1662.
[94] M.D. Wodrich, C. Corminboeuf, P.R. Schreiner, A.A. Fokin, P.R. Schleyer, Org. Lett., 9 (2007) 1851-1854.
[95] A. Saravanamuthu, T.J. Vickers, C.S. Bond, M.R. Peterson, W.N. Hunter, A.H. Fairlamb, J. Biol. Chem., 279 (2004) 29493-29500.
[96] A. Hernán, A.B. Andrey, E.G. Jill, Med. Res. Rev., 26 (2006) 531-568.
[97] D.C. Liu, J. Nocedal, Math. Program., 45 (1989) 503-528.
[98] J.J. Sutherland, D.F. Weaver, J. Comput.-Aid. Mol. Des., 18 (2004) 309-331.
[99] B. Lygo, B. Allbutt, S.R. James, Tetrahedron Letters, 44 (2003) 5629-5632.
[100] A. Jakalian, B.L. Bush, D.B. Jack, C.I. Bayly, J. Comput. Chem., 21 (2000) 132-146.
[101] R. Natesh, S.L.U. Schwager, E.D. Sturrock, K.R. Acharya, Nature, 421 (2003) 551-554.
[102] J. San, A. Amor, C.H.O. Seung Joo, B. Kor. Chem. Soc., 26, (2005) 952-958.
[103] A. Gangjee, E. Elzein, S.F. Queener, J.J. McGuire, J. Med. Chem., 41 (1998) 1409-1416.
[104] F. Iribarne, M. Paulino, S. Aguilera, O. Tapia, J. Mol. Graph. Model., 28 (2009) 371-381.

Anexos

Erros Relativos

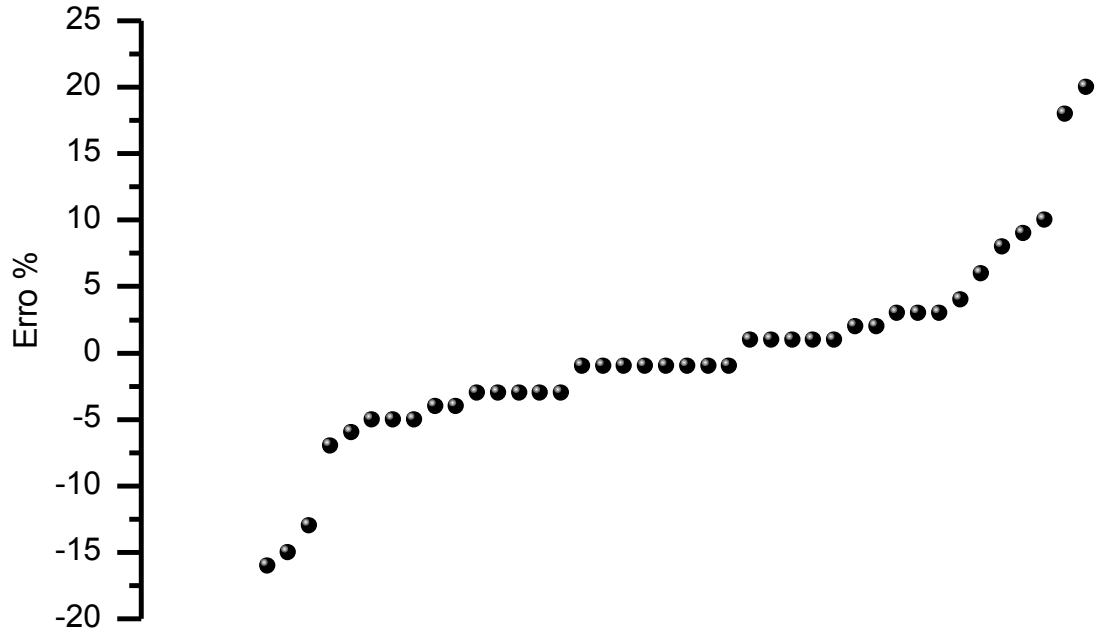


Figura A. Erros relativos para a previsão e validação para o modelo final do conjunto de dados (1) do em ordem crescente.

Figura B. Erros relativos para a previsão e validação para o modelo final do conjunto de dados (2) do em ordem crescente.

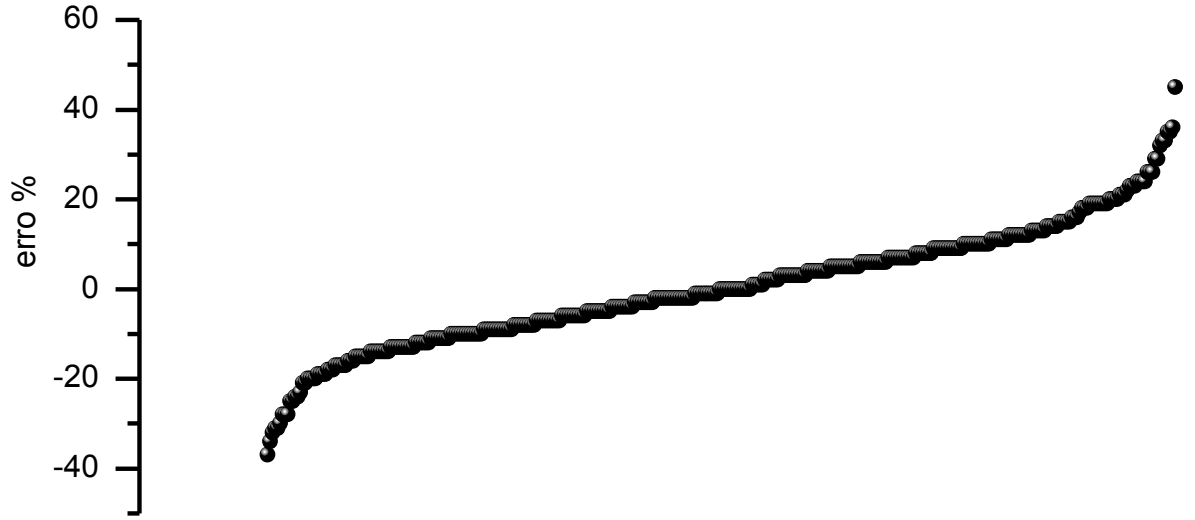


Figura C. Erros relativos para a previsão e validação para o modelo final do conjunto de dados (3) do em ordem crescente.

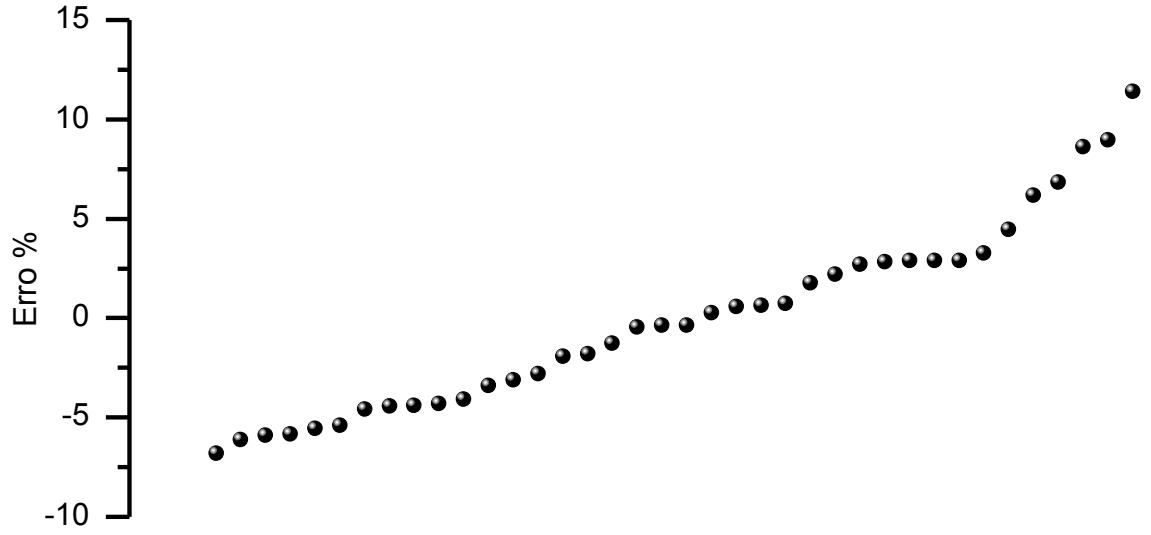


Figura D. Erros relativos para a previsão e validação para o modelo final LQTA-QSAR-DR do em ordem crescente.

Funções do Matlab®

Tratamento dos descritores de LJ e corte pela variância:

```
function [OUT,vari] = LJ_trunc_var(X,var_val)
%
% USAGE [LJ,vari] = LJ_trunc_var(VDW,var_val)
%
%%
[m,n]=size(X);
X2 = X;
for f=1:m
    for j=1:n
        if (X2(f,j) >= 30)
            X2(f,j) = 30 ;
        else
            X2(f,j) = X2(f,j);
        end
    end
end
vari=var(X2);
%%
OUT = X;
for f=1:m
    for j=1:n
        if (OUT(f,j) >= 30)
            OUT(f,j) = 30 + log10(1+OUT(f,j)-30);
        else
            OUT(f,j) = OUT(f,j);
        end
    end
end
OUT=OUT(:,vari>var_val);
end
```

Determinação de $|r|_{corte}$:

```
function val = corr_cut_val(y)
[m,~] = size(y);
C = rand(m,50000);
C_A = corr(C,y);
val= norminv(1-0.01, 0, std(C_A)); % val= norminv(1-0.01, 0, std(C_A))
end
```

Filtro de correlação e CDDA:

```
function [X_OK,names_OK,param] = cdda(X,y,names,cut,cdda_cut)
%
% usage =  [X_cut,names_cut] = cdda(X,y,names,cut,cdda_cut)
%
% X is the independente variable matrix, y is dependente variable vector,
% names is title of matrix and cut is the desired cut-off on correltation.
% cdda_cut is the desired cut-off on scatterness.
%
%% Correlation cut-off
X2=autoscale(X);
y2=autoscale(y);
C = abs(corr(X2,y2));
X(:,C < cut)=[];
names(:,C < cut)=[];
%% Distribution cut-off
jansize = 4;
clear m n , [m,n]=size(X);
if n > 5000 , display('You might face problems with memory')
else fprintf('Number of variable for the next step: %d\n' , n);
end
clear y2 , y2=normalization01(y);
Xn = X;
for i=1:n
    Xn(:,i)=normalization01(X(:,i));
end
X2=[y2 Xn];
clear F
F = ones(m,2^jansize,n+1);
jan = 1/(2^jansize);
    for g=1:(2^jansize)
        for f=1:m
         for j=1:n+1
             if (X2(f,j) >= ((g-1) * jan) && (X2(f,j) < (g * jan) ||
(X2(f,j) == (g * jan)) && (X2(f,j) == 1)));
            F(f,g,j) = 1;
                else
            F(f,g,j) = 0;
             end
         end
        end
    end
% Error comparasions
Best = (sum(F(:,:,1)));
ScoreForw = ones(m,2^jansize);
ScoreBack = ones(m,2^jansize);
ScoreOK = ones(m,2^jansize);
for i=1:n
    ScoreForw(i,:) = abs((sum(F(:,:,i+1))-Best));
```

```
    ScoreBack(i,:) = abs((fliplr(sum(F(:,:,i+1))))-Best);
    if sum(ScoreForw(i,:)) < sum(ScoreBack(i,:));
    ScoreOK(i,:)= ScoreForw(i,:);
    else
    ScoreOK(i,:)= ScoreBack(i,:);
    end
end
clear param
param = sum(ScoreOK,2);
param = param./2;
param = 1-(param./(m-2));
%% Ending Step
X(:,param <= cdda_cut)=[];
names(:,param <= cdda_cut)=[];
X_OK =  X;
names_OK = names;
display('Done')
end
```

Remoção de descritores inter-correlacionados mantendo o melhor correlacionado com y:

```
function [X,X_n,Groups,Final] = inter_corr_rem(X_in,X_n_in,y,parm)
%
% Usage [X,X_n] = inter_corr_rem(X_in,X_n_in,y,param);
% Where X and X_n are the output non intercorrelated descriptor matrix and
% descriptors headers;
% X_in,X_n_in and y are Input descriptor matrix, descriptors headers and
% dependete variable.
% param is the cut-off for intercorrelation
%
X = X_in;
X_n = X_n_in;
%% Intercorrelation matrix (intercorr)
n=size(X,2);
Cors=zeros(n,n);
Groups=NaN(n,n);
for i=1:n
    Cors(:,i) = abs(corr(X,X(:,i)));
    clear A; A = find(Cors(:,i)> parm );
    clear t ; t = size(A,1); A = [A' NaN(1,n-t)];
    Groups(i,:)=A;
end
%% Non intercorrelated set (ninterset)
Final = ones(1,1) ;
for i=1:n
    clear T ; T = Groups(i,:); T = T(isfinite(T));
    clear X2 ; X2 = X(:,T);
    clear CorG ; CorG = corr(X2,y);
```

```
    clear B ; B = find(CorG == max(CorG));
    Elem = T(:,B);
    Final = [Final Elem];
end
Final = Final';
Final(1,:)=[];
Final = unique(Final);
X=X_in(:,Final);
X_n = X_n_in(:,Final);
%% Iterative ending
while sum(isfinite(Groups(:,2))) ~= 0 ;
    [X,X_n,Groups,Final] = inter_corr_rem(X,X_n,y,parm);
end
end
```

Previsão conjunto externo para modelos LQTA-QSAR

```
function [y_pred,Q2ext,SEP] =
predLQTA(ExtDesMat,ExtDesMat_names,Model_names,RegrVec,IndepTerm,yext,ytrain
)
% Usage
% [y_pred,Q2ext,SEP] =
predLQTA(ExtDesMat,ExtDesMat_names,Model_names,RegrVec,IndepTerm,yext,ytrain
)
%
%% creat ext matrix
[m,~]=size(ExtDesMat);
extX2 = ExtDesMat;
moddes = find(ismember(ExtDesMat_names,Model_names)==1);
Mod_Desc_n = ExtDesMat_names(:,moddes);
Mod_Desc = extX2(:,moddes);
% sorting
[~,T2] = sort(Mod_Desc_n) ;
Mod_Desc_2 = Mod_Desc(:,T2); % sortcellchar por sort
[~,T3] = sort(Model_names) ;
VR_2 = RegrVec(T3,:); % sortcellchar por sort
% regression
y_pred = ((VR_2'*Mod_Desc_2')+IndepTerm)';
plot(y_pred,yext,'.')
hold ; xlabel('y Predicted') ; ylabel('y Experimental')
Q2ext = 1-(sum((yext-y_pred).^2)/(sum((yext-mean(ytrain)).^2)));
display(Q2ext);
SEP = sqrt((sum((yext-y_pred).^2))/(m-1));
display(SEP);
end
```

Shell Scripts de Linux

Alinhamento

```
#!/bin/bash
if [ -d $PWD/pconfs ]
then
  rm -r $PWD/pconfs
  echo " Conformation directory clear OK"
fi
if [ -e CEP_atoms.ndx ]
then
  echo " Aligment atoms set OK"
else
  echo " Missing CEP_atoms.ndx "
  exit 1
fi
mkdir $PWD/pconfs
echo 1 > l ; echo 1 >> l
trjconv -b 50 -f md300.trr -s md300.tpr -fit rot+trans -sep -o pconfs/.pdb <
l
rm l
a=`basename $PWD`
if [ "$a" = "ref" ]
then
  cd pconfs
  max_frame=`ls *pdb | sort -rn | sed -n '1p' | cut -d. -f1`
  cd ..
  for ((  i = 0 ;  i <= $max_frame ;  i++  ))
  do
    g_confrms -f1 pconfs/0.pdb -n1 CEP_atoms.ndx -f2 pconfs/${i}.pdb -n2
CEP_atoms.ndx -o pconfs/${i}_alg.pdb -one
  done
  for ((  i = 0 ;  i <= $max_frame ;  i++  ))
  do
    editconf -f pconfs/${i}_alg.pdb -o pconfs/${i}_alg.gro
  done
else
   if [ -d "../ref" ]
   then
    echo " ref/ directory OK"
   else
    echo "Error: Missing ../ref directory "
    exit 1
   fi
   if [ -d "../ref/pconfs" ]
   then
    echo " Conformation ref/ directory OK"
   else
    echo "Error: Missing ../ref/confs directory "
```

```
    exit 1
   fi
   cd pconfs
   max_frame=`ls *pdb | sort -rn | sed -n '1p' | cut -d. -f1`
   cd ..
   for ((  i = 0 ;  i <= $max_frame ;  i++  ))
   do
    g_confrms -f1 ../ref/pconfs/0_alg.pdb -n1 ../ref/CEP_atoms.ndx -f2
pconfs/${i}.pdb -n2 CEP_atoms.ndx -o pconfs/${i}_alg.pdb -one
   done
   for ((  i = 0 ;  i <= $max_frame ;  i++  ))
   do
    editconf -f pconfs/${i}_alg.pdb -o pconfs/${i}_alg.gro
   done
fi
cat pconfs/*alg.pdb > ${a}_ali.pdb
cat pconfs/*alg.gro > ${a}_ali.gro
echo " "
echo "Outputs: "
echo " ${a}_ali.pdb for checking "
echo " ${a}_ali.gro to be used as input for LQTAgrid "
echo " "
if [ "$a" = "ref" ]
then
echo "Done"
else
  rm -r $PWD/pconfs
  echo " Conformation directory clear OK"
fi
```

Dinâmica Molecular

```
#!/bin/bash

# PREP

if [ -e recep_t_2.gro ]
then
  rm *tpr
  echo " recep_t_2.gro OK "
  editconf -f recep_t_2.gro -o recep_OK.gro
else
  echo " Sem o arquivo recep_t_2.gro  "
  exit 1
fi

# ST

if [ -e recep_OK.gro ]
then
```

```
  echo " recep_OK.gro OK "
  grompp -f st.mdp -c recep_OK.gro -p recep.top -o st.tpr
  mdrun -s st.tpr -o st.trr -c gs.gro -g st.log -e st.edr
else
  echo " Sem o arquivo recep_OK.gro  "
  exit 1
fi

# GS

if [ -e gs.gro ]
then
  echo " gs.gro OK "
  grompp -f gs.mdp -c gs.gro -p recep.top -o gs.tpr
  mdrun -s gs.tpr -o gs.trr -c md300.gro -g gs.log -e gs.edr
else
  echo " Sem o arquivo gs.gro  "
  exit 1
fi

# MD300

if [ -e md300.gro ]
then
  echo " gs.gro OK "
  grompp -f md300.mdp -c md300.gro -p recep.top -o md300.tpr
  mdrun -s md300.tpr -o md300.trr -c pmd.gro -g md300.log -e md300.edr
else
  echo " Sem o arquivo md300.gro  "
  exit 1
fi
```

Cargas GAFF

```
#!/bin/bash
export AMBERHOME=/home/lqta/amber11
for ((  i = $1 ;  i <= $1 ;  i++  ))
do
cd mol$i/
babel -ihin mol${i}.hin -osd mol${i}.sd
vconf mol${i}.sd
csplit -k -f conf. mol${i}.hin_confs.sdf '/.hin/' '{*}'
rm conf.00
#
# antechamber
#
      for i in `ls conf*`
      do
      /home/lqta/amber11/bin/antechamber -i $i -fi mdl -o ${i}.mol2 -fo mol2
-c bcc -at sybyl -nc 0
```

```
     done
#
# linhas
#
lines=`sed -n "3 p" conf.01.mol2 | awk '{ print $1 }'`
lines=`expr $lines + 8`
#
     for i in `ls *mol2`
     do
     sed -n '9,'$lines' p' $i | awk '{ print $9 }' > ${i}.txt
     paste -d, *mol2.txt > columns.txt
     done
#
rm *mol2.txt
perl -pi.bak -e 's/,/\t/g' columns.txt
awk 'NF {s=0;for(i=1;i<=NF;i++) s+=$i;printf("%.4f\n", s/NF)}' columns.txt >
mean.txt
sed -n '1,8 p' conf.01.mol2 > pronto.mol2
sed -n '9,'$lines' p' conf.01.mol2 > file1
paste -d, mean.txt file1 | perl -pi.bak -e 's/,//g' | awk 'BEGIN { OFS =
"\t" } ; { print $2,$3,$4,$5,$6,$7,$8,$9,$1 }' >> pronto.mol2
tilend=`wc -l conf.01.mol2 | awk '{ print $1 }'`
lines=`expr $lines + 1`
sed -n ''$lines','$tilend' p' conf.01.mol2 >> pronto.mol2
#
# limpa pasta
#
mkdir conf
mv conf* conf/
#
# finaliza
#
cd ..
done
```

Tutorial 1

Using the LQTAgridAFF program.

What is LQTA-QSAR?

LQTA-QSAR[2] is a free package for time dependent QSAR. LQTA-QSAR incorporates GROMACS[3] molecular dynamics simulation package to obtain Conformational Ensemble Profiles (CEP) for a desired series of compound molecular models. For such compounds is known a certain biological activity. The main program LQTAgrid (Java™), makes use of GROMACS fixed-column coordinate file format (.gro) and topology file which defines the parameters for the molecular models (.tpr, .top) to build Molecular Interaction Field (MIF) descriptors.

The first version of LQTAgrid dealt only with GROMOS atom type. With the new LQTAgridAFF you can set any type of atom from you favorite force field.

MIF descriptors are selected to build PLS models using OPS variable selection algorithm.[4] The final models are presented as spheres in 3D space (**Figure 1**). QSAR model interpretation can be carried out as long as molecular modification designs to create new active molecules.

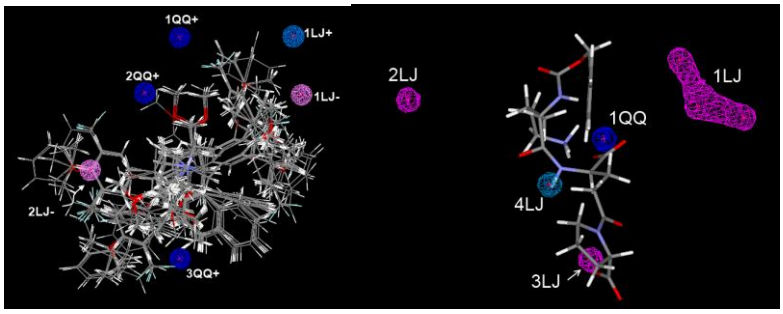


Figure 1. Examples of LQTA-QSAR models. Descriptors are shown as surfaces in 3D space.

LQTA-QSAR package is free and the source code can be acquired upon request. Go to lqta.iqm.unicamp.br and download the files required for this tutorial as well as LQTAgridAFF.

You can find a good tutorial on how to use LQTAgridAFF program in:

http://lqta.iqm.unicamp.br/

http://www.cenapad.unicamp.br/servicos/treinamentos/apostilas/apostila_QuimComp.pdf

Tutorial 2

LQTAgrid descriptors pretreatment, variable filtering and selection and PLS model visualization.

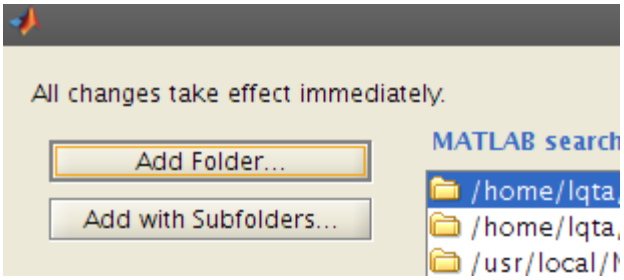
For this tutorial you will need:

MATLAB and basic usage.
http://www.mathworks.com/academia/student_center/tutorials/launchpad.html
Textpad. http://www.textpad.com/
The TUTORIAL.mat file
Microsoft windows. Or another OS where MATLAB is installed.
M-files contained in tutorial2files.rar.
QSARmodeling. Java applet for PLS models building
USCF chimera http://www.cgl.ucsf.edu/chimera/

For this tutorial you'll be using LQTAgrid descriptor matrix for a data set of 40 fase transfer catalysts. This data set was retrieved from the work of Melville *et al.*[1] The activities, expressed in % selectivity.

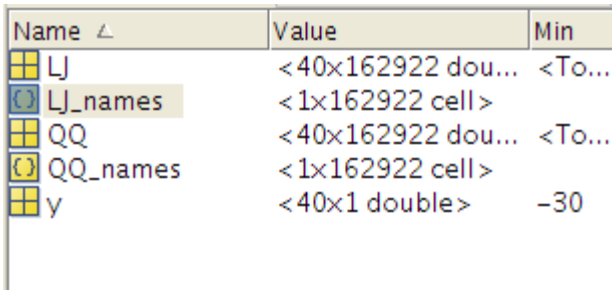
Before you start certify that the tutorial directory in in matlab path

> File > Set Path



1. Load the TUTORIAL.mat in matlab.

> load TUTORIAL.mat



2. Perform the Variance cut-off using the filter_var function, which will treat the LJ descriptors and remove those distant from the molecular surface.

> [LJ_out,QQ_out,LJ_names_out,QQ_names_out] = filter_var(LJ,QQ,LJ_names,QQ_names,0.01);
The out put will be the matrices LJ_out and QQ_out and their respective names. The variance cut was 0.01 kcal mol^{-1}.

```
>> whos
  Name              Size                    Bytes  Class     Attributes

  LJ              40x162922              52135040  double
  LJ_names         1x162922              24516846  cell
  LJ_names_out     1x34759                5169828  cell
  LJ_out          40x34759               11122880  double
  QQ              40x162922              52135040  double
  QQ_names         1x162922              24516846  cell
  QQ_names_out     1x34759                5169828  cell
  QQ_out          40x34759               11122880  double
  y               40x1                        320  double
```

3. Split the dataset into training set and test set by defining a external vector.

external_n = [6,9,13,12,19,21,25,28,34,36];

The sample above will be treated as the external data set.

4. Define the elements of the training and external datasets

> y_t=y ; y_t(external_n)=[];
> y_ext= y(external_n);
> LJ_t = LJ_out; LJ_t(external_n,:)=[];
> QQ_t = QQ_out; QQ_t(external_n,:)=[];
> EXT= [QQ LJ];
> EXT = EXT(external_n,:);

5. Now perform determine the $|r|_{cut}$ using the r_cut function.

> r_cut(y_t)

0.4312

Now you know that the correlation cut must be higher than 0.43

6. Now use CDDA to eliminate simultaneously the poorly correlated and badly distributed descriptors.

> [LJ_cdd,LJ_names_cdd] = cdda(LJ_t,y_t,LJ_names_out,0.43,0.5);
> [QQ_cdd,QQ_names_cdd] = cdda(QQ_t,y_t,QQ_names_out,0.43,0.5);

7. Save the QQ_cdd and LJ_cdd matrices and the dependent variable

> MIF = [QQ_cdd LJ_cdd];
> save "MIF.dat" MIF -ASCII
> save "y_t.dat" y_t -ASCII

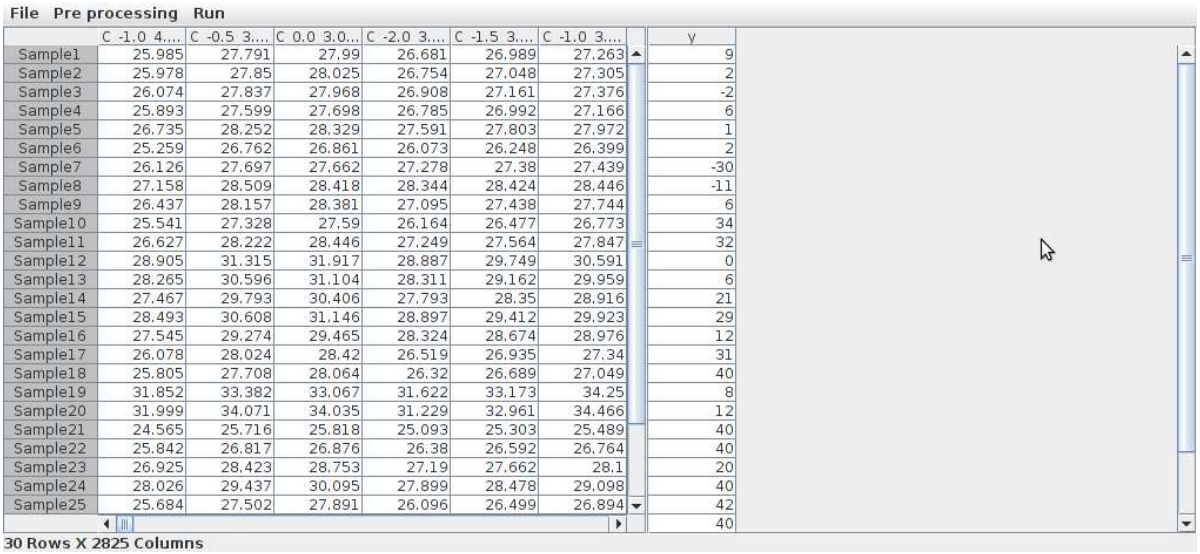
8. Edit manually the MIF.dat to insert the columns names QQ_names_cdd and LJ_names_cdd. To to that just copy and past the entire lines from the Variable Editor. Make sure to paste the QQ and LJ names in the same line. You can use textpad on windows or gedit on linux.

9. Replace the ' symbol from the MIF.dat file and save.

10. Open the MIF.dat file and the y_t.dat in QSARmodeling

> ***Warning: OPS is implemented with the QSAR Modeling software, developed and provided as a courtesy of the Theoretical and Applied Chemometrics Laboratory's research group (article in preparation).***

You should have the window below.

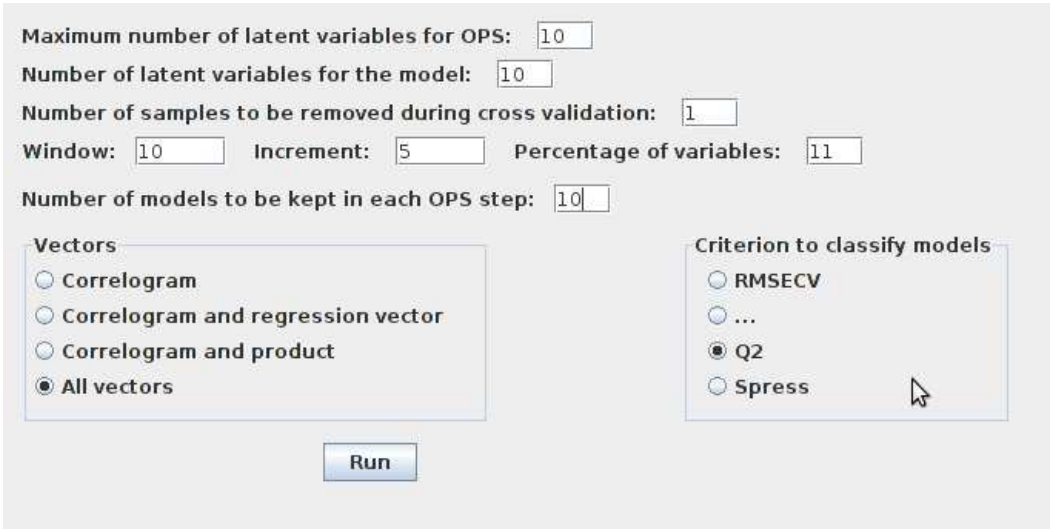


2 – ASCII file for biological activity (dependent variable, **y**). Do not use header or sample names. Example:

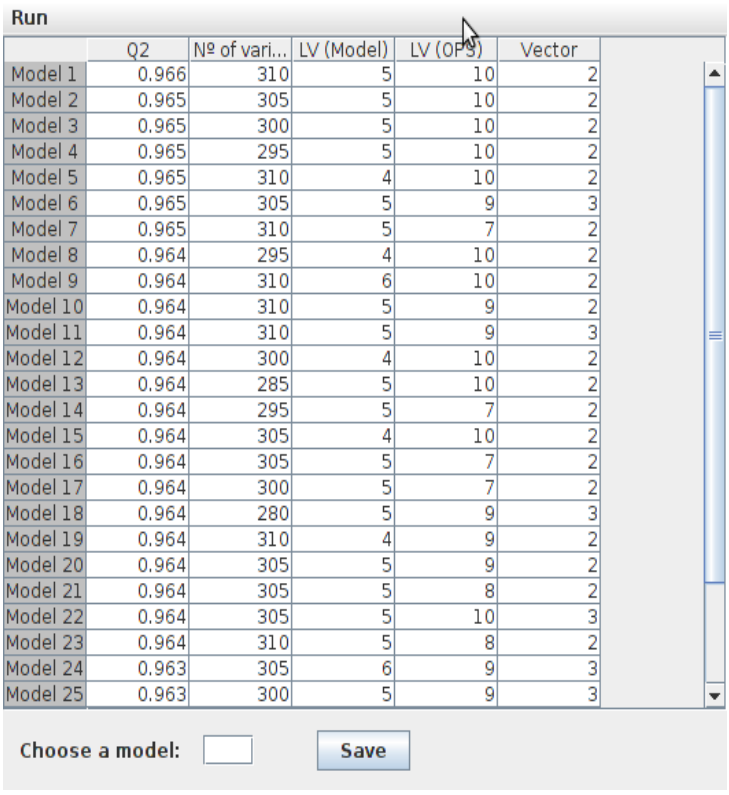
7.92
5.6
6.41
5.6
9.6
8.46
.
.
.

Save LJ_OK matrix and ytrain to produce these ASCII files.

11. Perform a course advance OPS search by filling out the gaps as follows. To accesses the window below go to Run > Advanced OPS



The calculation might take a while to finish.
When asked about how many models to see type 30.

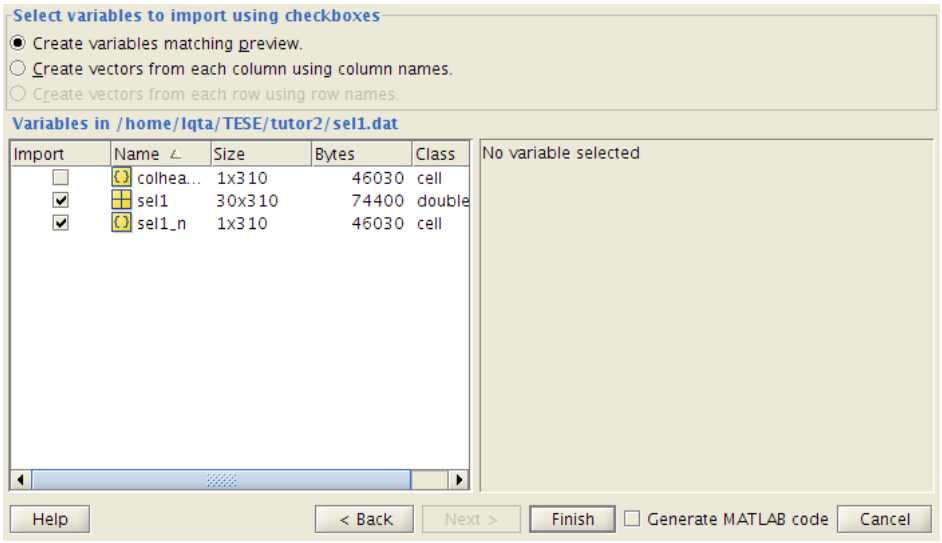


	Q2	Nº of vari...	LV (Model)	LV (OPS)	Vector
Model 1	0.966	310	5	10	2
Model 2	0.965	305	5	10	2
Model 3	0.965	300	5	10	2
Model 4	0.965	295	5	10	2
Model 5	0.965	310	4	10	2
Model 6	0.965	305	5	9	3
Model 7	0.965	310	5	7	2
Model 8	0.964	295	4	10	2
Model 9	0.964	310	6	10	2
Model 10	0.964	310	5	9	2
Model 11	0.964	310	5	9	3
Model 12	0.964	300	4	10	2
Model 13	0.964	285	5	10	2
Model 14	0.964	295	5	7	2
Model 15	0.964	305	4	10	2
Model 16	0.964	305	5	7	2
Model 17	0.964	300	5	7	2
Model 18	0.964	280	5	9	3
Model 19	0.964	310	4	9	2
Model 20	0.964	305	5	9	2
Model 21	0.964	305	5	8	2
Model 22	0.964	305	5	10	3
Model 23	0.964	310	5	8	2
Model 24	0.963	305	6	9	3
Model 25	0.963	300	5	9	3

The first model is good enough type 1 in the “Choose a model” field and save as sel1.dat

12. Edit the sel1.dat file in a text editor and replace the tabs for spaces. To do so just select a tab space anf find and replace by a space.

Now, load the file in matlab againd using > File > Import data



13. Now its time to remove the highly inter-correlated descriptors using the inter_corr_rem function.

> [sel2,sel2_n] = inter_corr_rem(sel1,sel1_n,y_t,0.9);

14. Save the sel2 matrix as sel2.dat and insert the column names (sel2_n) as mentioned above and load it on QSARmodeling with the y_t.dat