JOYCE KAROLINE DARÉ

INCLUSION OF CONFORMATIONAL INFORMATION IN MIA-QSPR

LAVRAS – MG
2019

JOYCE KAROLINE DARÉ

INCLUSION OF CONFORMATIONAL INFORMATION IN MIA-QSPR

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Agroquímica, área de concentração em Química Computacional aplicada na agricultura, para a obtenção do título de Mestre.

Prof. Dr. Matheus Puggina de Freitas

Orientador

Prof. Dr. Teodorico de Castro Ramalho

Coorientador

LAVRAS – MG
2019

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Daré, Joyce Karoline.
Inclusion of conformational information in MIA-QSPR / Joyce Karoline Daré. - 2019.
83 p. : il.

Orientador(a): Matheus Puggina de Freitas.
Coorientador(a): Teodorico de Castro Ramalho.
Dissertação (mestrado acadêmico) - Universidade Federal de Lavras, 2019.
Bibliografia.

1. MIA-QSPR. 2. Informação Conformacional. 3. Otimização estrutural. I. de Freitas, Matheus Puggina. II. Ramalho, Teodorico de Castro. III. Título.

JOYCE KAROLINE DARÉ

INCLUSÃO DE INFORMAÇÃO CONFORMACIONAL EM MIA-QSPR
INCLUSION OF CONFORMATIONAL INFORMATION IN MIA-QSPR

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Agroquímica, área de concentração em Química Computacional aplicada na agricultura, para a obtenção do título de Mestre.

APROVADA em 21 de fevereiro de 2019
Melissa Soares Caetano - UFOP
Elaine Fontes Ferreira da Cunha - UFLA

Prof. Dr. Matheus Puggina de Freitas

Orientador

Prof. Dr. Teodorico de Castro Ramalho

Coorientador

LAVRAS – MG
2019

Aos meus pais, por todo apoio e incentivo.
Ao meu noivo Wilson, por toda paciência e amor.
Aos meus amigos Bruna, Daniela e Francisco por todos os ótimos momentos que me proporcionaram e as tantas risadas.
Aos demais membros de minha família por se alegrarem nas minhas vitórias e me apoiarem nas derrotas.
A Deus pela vida, saúde e por sempre estar comigo.
Dedico.

AGRADECIMENTOS

À Universidade Federal de Lavras, especialmente ao Departamento de Química, pela oportunidade.

Aos professores Matheus e Teodorico pelo incentivo, direcionamento e disposição para ajudar.

À banca pela disponibilidade e contribuições para este trabalho.

A todos colegas do laboratório de Química Computacional por todos os ensinamentos.

À minha família e noivo pelo apoio incondicional e a torcida.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – pela concessão da bolsa de mestrado (Código de Financiamento 001).

À fundação de amparo à pesquisa do estado de Minas Gerais (FAPEMIG) e ao conselho nacional de desenvolvimento científico e tecnológico (CNPQ) pelo apoio financeiro.

MEU MUITO OBRIGADA!

RESUMO

MIA-QSPR (*Multivariate Image Analysis applied to Quantitative Structure-Property Relationships*) é uma técnica livre e eficiente de modelagem molecular desenvolvida com o intuito de simplificar análises quantitativas que buscam correlacionar características estruturais de compostos químicos com uma propriedade físico-química/biológica de interesse. Fundamentalmente, ela se baseia na ideia de que, para uma série de imagens de moléculas congêneres, as mudanças nos substituintes gerarão variações na orientação (posição) dos pixels, o que, por sua vez, explica as mudanças na variável resposta (propriedade modelada). Os descritores moleculares, nessa técnica, são gerados a partir de projeções bidimensionais (2D) das estruturas químicas de interesse. Por se tratar de uma ferramenta 2D, intuitivamente, o MIA-QSPR não incorpora informações espaciais (3D) acerca das estruturas químicas, o que, de certa forma, pode ser considerado uma limitação da técnica, visto que as propriedades espaciais são conhecidas por regerem grande parte do comportamento físico-químico/biológico de moléculas. No entanto, tratando-se de MIA-QSPR, não existem comprovações na literatura de que a inclusão desse tipo de informação realmente contribua para criação de modelos de predição mais efetivos. Em decorrência, a proposta do presente trabalho consistiu no desenvolvimento de uma estratégia para inserção de informações conformacionais e bioconformacionais nos descritores MIA-QSPR, bem como uma análise detalhada dos resultados da inclusão de tais propriedades. Para fins de comparação, o trabalho foi divido em três etapas: i) primeiro, elaborou-se um modelo MIA-QSPR tradicional com uma classe de compostos com propriedade independente de um receptor biológico (logKoc); ii) para uma segunda classe de compostos, construiu-se um modelo MIA-QSPR com as projeções 2D das moléculas com estruturas completamente otimizadas em um meio livre do receptor (*i.e.*, fez-se uma busca pelo arranjo espacial com menor energia para cada molécula); e, como etapa principal, iii) determinou-se as bioconformações mais prováveis de uma classe de ligantes, através da técnica de ancoramento molecular (*molecular docking*), e, com base nessas estruturas, construiu-se um modelo MIA-QSPR com as projeções bidimensionais das conformações identificadas. Vale ressaltar que, para a segunda parte, o grupo de compostos selecionado possui um alvo biológico; porém, até o momento, tal receptor não é conhecido e, consequentemente, um estudo bioconformacional não se fez necessário. Por outro lado, o terceiro grupo de moléculas possui um alvo biológico bem elucidado, o que permitiu um estudo envolvendo tal receptor. Como resultado, comprovou-se que a técnica MIA-QSPR é capaz de englobar informações de cunho conformacional; no entanto, a inclusão desse tipo de informação acarretou em um alinhamento imperfeito das subestruturas congêneres, o que é crucial para a eficiência dos modelos de predição da maioria das técnicas QSPR. Sendo assim, os modelos gerados a partir de estruturas otimizadas/ancoradas resultaram em modelos de calibração menos satisfatórios do que os convencionais e pode-se, portanto, concluir que a conectividade e as propriedades atômicas (tais como raio de van der Waals e eletronegatividade) são descritores moleculares mais eficientes em MIA-QSPR do que as características conformacionais atribuídas nesse trabalho. Tal estudo não só reforçou a eficiência da técnica MIA-QSPR tradicional como também levantou questionamentos acerca da complexidade de outras abordagens.

Palavras-chave: MIA-QSPR. Informação conformacional. Otimização estrutural. Ancoramento molecular.

ABSTRACT

MIA-QSPR (*Multivariate Image Analysis applied to Quantitative Structure-Property Relationships*) is an efficient technique for molecular modeling developed with the aim of simplifying quantitative analyses that seek correlating structural features of chemical compounds with physical-chemical/biological properties (QSPR studies). Essentially, it relies on the idea that, for a congeneric series of molecule images, the changing in substituent groups will generate variations in the pixel orientations (positions), which, in turn, can explain the visual changing observed in the response variable values (physical-chemical/biological property). The molecular descriptors employed in this technique are obtained from bidimentional (2D) projections of the chemical structures under analysis. Since MIA-QSPR has a 2D approach, intuitively, it does not include spatial information (3D), which can be considered a limitation of this technique, once it is well-known that spatial features rule a significative part of the physical-chemical/biological molecular behavior. However, regarding MIA-QSPR, there are no evidences in literature that the inclusion of this type of information would, indeed, improve the prediction ability of QSPR models. In this sense, this work aimed to develop a strategy for inserting conformational and bioconformational information into the MIA-QSPR descriptors, and a detailed analysis of the results of such inclusion. For means of comparison, this work was divided into three parts: i) initially, a traditional MIA-QSPR model was built for a class of compounds with a physical-chemical property independent of a biological receptor (logKoc); ii) a second MIA-QSPR model was built using bidimentional projections resulted from the fully structural optimization of a set of molecules in a receptor free environment; iii) finally, for a third set of compounds, the most likely bioconformations were determined through molecular docking technique and, based on the bidimentional projection of these structures, the last MIA-QSPR model was built. It is worth mentioning that the second molecule set has a biological target, but, until currently, this receptor is not completely known, then a bioconformational study was not required; on another hand, for the last set of compounds, the biological target is already fully elucidated, which supports the molecular docking step performed herein. As a result, it was verified that the MIA-QSPR technique is capable of encoding conformational information; however, including this type of information resulted in an imperfect alignment of the congeneric substructures, which is crucial for building efficient MIA-QSPR prediction models. Accordingly, the models obtained from the optimized/docked structures generated less satisfactory calibration models than the conventional ones, and, therefore, one can conclude that the atomic connectivity and properties (*e.g.* van der Waals radii and electronegativity) are more efficient molecular descriptors for MIA-QSPR technique than spatial parameters considered herein.

Key-words: MIA-QSPR. Conformational information. Structural optimization. Molecular docking.

LISTA DE ILUSTRAÇÕES

PRIMEIRA PARTE

Figura 1- (a) Superposição de imagens utilizadas no modelo MIA-QSPR original; (b) Superposição de imagens utilizadas no modelo *aug*-MIA-QSPR 10

Figura 2- MIA-Plots baseados nos parâmetros VIP e b para derivados de triazinas..... 11

Figura 3- Representação esquemática utilizada na metodologia MIA-QSPR do composto Clorpirifos: agente agroquímico organofosforado 20

Figura 4- Representação esquemática utilizada na metodologia MIA-QSPR do composto anti-HCV mais ativo..... 21

Figura 5- Representação esquemática utilizada na metodologia MIA-QSPR do composto anti-SARS-CoV mais ativo (Composto **31**) 22

SEGUNDA PARTE

Artigo 1

Figure 1- The superposed images for the 24 organophosphorus compounds 38

Figure 2- Plot of actual vs. predicted $\log K_{OC}$ values for the 24 organophosphorus insecticides 39

Figure 3- MIA-contour maps based on variable importance in projection (VIP) scores and PLS regression coefficients (b), used to analyze the descriptor contributions for the QSPR modeling of $\log K_{OC}$ values of the organophosphorus insecticides..... 41

Artigo 2

Figure 1- (a) Superposed images for the 42 anti-HCV compounds. The common nucleus is used for 2D alignment, while the variable substructures explain the variance in the pEC_{50} 59

Figure 2- Plot of the experimental × predicted pEC50 values using the MIA-QSAR model obtained from atomic colors proportional to ε 59

Figure 3- MIA-contour maps based on variable importance in projection (VIP) scores and PLS regression coefficients (b), used to analyze the descriptor contributions for the MIA-QSAR modeling of pEC_{50} values of the anti-HCV molecules 60

Figure 4- Proposed molecular structure for enhanced anti-HCV bioactivity 60

Figure S1- Summary of the steps for performing 2D and 3D MIA-QSAR 63

Figure S2- Applicability domain showed through William´s plot..... 64

Artigo 3

Figure 1- Superposition of the disulfide molecules images used in the (1a) traditional MIA-QSAR procedure, (1b) MIA-QSAR with optimized geometries, and (1c) MIA-QSAR with docked structures..... 81

Figure 2- Conformation selected for compound 31 used as reference for the other disulfide structures 81

Figure 3- Plot for predicted × experimental IC50 values using the MIA-QSAR model B 82

Figure 4- b-plot and VIP plot for model B 82

Figure 5- Applicability domain obtained from the William´s plot 83

Figure 6- Proposed disulfide compounds, where X = halogen atom. The first structure contains an o-nitrophenyl group at R1, a nitro group at R2, and a halogen at R3. The second compound would have a p-phenyl R1 substituent, a nitro group at R2 and a halogen at R3 83

LISTA DE TABELAS

SEGUNDA PARTE

Artigo 1

Table 1- Series of organophosphorus insecticides used in the MIA-QSPR modeling and the corresponding actual and predicted (in calibration, leave-one-out cross-validation and external validation) $\log K_{OC}$ data 37

Table 2- Statistical parameters used to attest the quality of the MIA-QSPR model 39

Artigo 2

Table 1- Validation parameters for MIA-QSAR models with pixel values proportional to r_{vdW}/ε, r_{vdW}, and ε, for the traditional analysis 58

Table 2- Validation parameters for MIA-QSAR models built using the optimized geometries..... 58

Table S1- Series of anti-HCV (Hepatitis C Virus) compounds used in MIA-QSAR modeling and the corresponding actual and predicted (in calibration, leave-one-out-validation, and external validation) pEC_{50} data using the descriptor values proportional to ε for the traditional analysis (model 3) 60

Table S2- Validation parameters for double cross validation technique 62

Artigo 3

Table 1- Data set of the disulfide compounds with their respective measured IC50 (μM), the calibration results for Model B, as well as their cross- and external validation results 78

Table 2- Statistical parameters obtained through traditional MIA-QSAR technique (Models A and B), MIA-QSAR applied to optimized molecular geometries (Models C and D), and the same technique employed to bioconformation-like images (Models E and F)..... 80

LISTA DE SIGLAS E ABREVIATURAS

aug-MIA-QSPR - *Augmented Multivariate Image Analysis applied to Quantitative Structure-Property Relationship*

b-plot - *PLS regression coefficients plot*

CMYK - *Cyan, Magenta, Yellow, and Black*

HCV - *Hepatitis C Virus*

HSB - *Hue, saturation e brightness*

HSV - *Hue, saturation e value*

IC_{50} - Concentração necessária de um inibidor para que a resposta seja reduzida pela metade

logKoc - *the logarithm of the soil/water normalized to organic carbon*

logKow - *the logarithm of the octanol/water partition coefficient*

LOOCV - *leave-one-out cross-validation*

LV - *Latent Variable*

MC - *Monte Carlo*

MIA - *Multivariate Image Analysis*

MIA-QSPR - *Multivariate Image Analysis applied to Quantitative Structure-Property Relationship*

MLR - *Multiple Linear Regression*

OF - Organofosforados

PCA - *Principal Component Analysis*

pEC_{50} - Concentração necessária de um composto para que a metade da resposta máxima seja obtida $\left(\frac{Vmáx}{2}\right)$

PLS - *Partial Least Squares*

QSAR - *Quantitative Structure-Activity Relationship*

QSPR - *Quantitative Structure-Property Relationship*

QSTR - *Quantitative Structure-Toxicity Relationship*

QSRR - *Quantitative Structure-Reactivity Relationship*

RGB - *Red, Green, and Blue*

RMN- Ressonância Magnética Nuclear

RMSE - *root mean square error*

SARS-CoV - *Severe acute respiratory syndrome Coronavirus*

VIP - *the variable importance in projection (VIP) scores*

YIQ - espaço de cores em que Y representa informação luma (brilho) e I e Q representam crominância (diferença colorimétrica entre uma dada cor na imagem de uma televisão e uma cor padrão de mesma luminância)

2D - Bidimensional

3D - Tridimensional

SUMÁRIO

PRIMEIRA PARTE.....--
1. INTRODUÇÃO..... 1
2.1. Correlação quantitativa entre estrutura e propriedade (QSPR) e a geração de descritores..... 4
2.2. Análise Multivariada de Imagens..... 6
2.3. Análise multivariada de imagens aplicada em QSPR (MIA-QSPR)..... 8
2.4. Métodos de calibração multivariada aplicados em QSAR/QSPR..... 12
2.5. Conformação em análise QSPR..... 14
2.6. Ancoramento molecular..... 16
2.7. Grupos de compostos empregados nas análises MIA-QSPR..... 19
2.7.1. Inseticidas Organofosforados..... 19
2.7.2. Compostos anti-HCV..... 20
2.7.3. Compostos anti-SARS-CoV..... 21
3.1. Modelagem MIA-QSPR Tradicional..... 23
3.2. Modelagem MIA-QSPR utilizando projeções 2D obtidas a partir de compostos com geometrias otimizadas..... 24
3.3. Modelagem MIA-QSPR utilizando projeções 2D obtidas a partir de compostos com geometrias otimizadas e bioativas..... 25
4. **CONSIDERAÇÕES GERAIS**..... 27
SEGUNDA PARTE..... 33
ARTIGO 1..... 34
ARTIGO 2..... 44
ARTIGO 3..... 65

PRIMEIRA PARTE

1. INTRODUÇÃO

MIA-QSPR (Análise Multivariada de Imagens aplicada à Correlação Quantitativa entre Propriedade e Estrutura, do inglês *Multivariate Image Analysis applied to Quantitative Structure-Property Relationship*) é uma metodologia bidimensional utilizada em análises quantitativas que buscam relacionar estrutura química com propriedades físico-químicas/biológicas de compostos (QSPR), na qual os descritores são gerados a partir de projeções bidimensionais (2D) das imagens de tais moléculas. Nesse contexto, descritores moleculares são definidos como o resultado de um procedimento lógico e matemático, que transforma uma informação química, codificada dentro de uma representação simbólica (como uma imagem *bitmap*) de uma molécula, em um número obtido experimentalmente.

A utilização de imagens para modelagem molecular (MIA) é antiga, no entanto, nunca havia se empregado tais imagens como fonte para extração de descritores na forma de *pixels* e construção de modelos de predição a partir deles. Nesse sentido, o MIA-QSPR é uma metodologia inovadora e eficiente que tem se reafirmado em diversos trabalhos, disponíveis na literatura. Os modelos gerados nesses trabalhos se mostram confiáveis e comparáveis aos gerados através de abordagens tridimensionais (3D). No entanto, sabe-se que, por se tratar de uma metodologia bidimensional, essa abordagem, usualmente, não codifica informações conformacionais de uma forma eficiente. Por outro lado, os descritores, nesse caso, buscam descrever características relacionadas puramente à conectividade dos átomos.

Na literatura, não existem trabalhos que comprovem que a inclusão ou exclusão de informações de cunho conformacional altere a qualidade dos modelos de predição MIA-QSPR gerados. Porém, sabe-se que diversas propriedades de moléculas são fortemente dependentes do arranjo espacial (estereoquímica) que elas apresentam. Um exemplo da importância da conformação de moléculas é a reatividade das mesmas em meio biológico. Um ligante só se encaixa em uma enzima se a conformação de ambos for apropriada; da mesma forma, uma enzima só desempenha seu papel biológico se estiver em sua conformação nativa/bioativa (bioconformação).

Diante desse impasse, o presente trabalho propõe (tem como objetivos) a averiguação da capacidade dos descritores MIA-QSPR em codificar informações espaciais

(conformacionais e bioconformacionais) moleculares e a investigação das mudanças acarretadas nos modelos de predição em função dessa inclusão.

Com isso em mente, um dos grandes desafios desse estudo foi determinar a melhor forma para incluir informações de cunho conformacional nos descritores MIA-QSPR; uma vez que a obtenção desses parâmetros (e sua natureza) é considerada uma das etapas determinantes em uma análise QSPR. A estratégia escolhida é descrita com detalhes na seção destinada à metodologia.

Para fins de comparação, esse projeto foi dividido em três etapas principais: (i) elaboração de um modelo MIA-QSPR tradicional; (ii) construção de um modelo MIA-QSPR com as projeções das moléculas estruturalmente otimizadas em um meio livre do receptor (*i.e.,* fez-se uma busca pelo arranjo espacial com menor energia); e, como etapa principal, (iii) determinação das bioconformações mais prováveis de uma classe de ligantes e construção do modelo MIA-QSPR com as projeções bidimensionais das conformações identificadas.

Por questão de coerência, para a primeira parte do trabalho, selecionou-se um grupo de moléculas com propriedade independente de um receptor biológico (logKoc); para a segunda parte, selecionou-se um grupo com propriedade biológica dependente de um alvo, porém, esse receptor não é conhecido e, logo, um estudo bioconformacional não se faz necessário; por último, o grupo de moléculas escolhido possui um alvo biológico conhecido e bem elucidado, o que permite um estudo acerca das conformações ativas (bioconformações) desses ligantes.

Nas abordagens que utilizam informações conformacionais para modelagem molecular, a busca conformacional exaustiva e as regras empíricas de alinhamento tridimensional são os maiores consumidores de recursos computacionais e complicadores dessas metodologias. Diante disso, o desenvolvimento do presente trabalho se justifica pelo fato de que, caso se comprove que os descritores moleculares contendo informações espaciais são menos eficientes que os puramente 2D, ou que não contribuem consideravelmente à técnica, ambas etapas (busca conformacional e alinhamento molecular) poderiam, em princípio, ser dispensadas. Essa redução traria melhorias significativas à etapa de modelagem molecular, pois estaria se reduzindo tempo e, consequentemente, economizando recursos. Por outro lado, se a inserção de informações espaciais se mostrar de fato relevante, a técnica MIA-QSPR se tornará ainda mais robusta para geração de bons modelos de predição; ou seja, ambas conclusões acerca da inclusão

de informações espaciais nos descritores MIA-QSPR trazem contribuições relevantes à área de modelagem molecular.

Por fim, caso a inserção de informação estereoquímica, da maneira como é proposta no presente trabalho, se mostre relevante para a MIA-QSPR, uma técnica 3D, *free*, robusta e capaz de fornecer interpretação química estará disponível para diversos fins.

2. REFERENCIAL TEÓRICO

2.1. Correlação quantitativa entre estrutura e propriedade (QSPR) e a geração de descritores

O QSAR (Correlação quantitativa entre estrutura química e atividade biológica, do inglês *Quantitative Structure-Activity Relationship*) é entendido como uma metodologia quantitativa que tem como objetivo encontrar um modelo matemático que correlacione **características estruturais** químicas de compostos (codificado na forma de **descritores**) e **atividade biológica** dos mesmos (YOUSEFINEJAD; HEMMATEENEJAD, 2015). Um termo mais abrangente, muitas vezes empregado como sinônimo de QSAR é o QSPR (do inglês *Quantitative Structure-Property Relationship*); no entanto, tal terminologia pode incluir diferentes propriedades químicas e físicas de compostos, ou seja, não só a atividade biológica. Como exemplos de diferentes propriedades biológicas que podem ser correlacionadas à estrutura têm-se: toxicidade (QSTR do inglês *Quantitative Structure-Toxicity Relationship*), reatividade (QSRR do inglês *Quantitative Structure-Reactivity Relationship*), dentre outras (YOUSEFINEJAD; HEMMATEENEJAD, 2015).

O QSPR, nos dias atuais, já é considerado uma técnica introduzida e aceita no desenvolvimento de moléculas com diferentes propósitos. O QSAR, por exemplo, está inserido na indústria farmacêutica como uma ferramenta essencial e inseparável na **descoberta** e **otimização** de fármacos (YOUSEFINEJAD; HEMMATEENEJAD, 2015). A essência de tal metodologia está em assumir que compostos novos ou não-testados que sejam estruturalmente similares a substâncias já conhecidas terão chances significativas de possuir atividades/propriedades biológicas parecidas (NANTASENAMAT et al., 2009).

Historicamente, o QSAR, da maneira como é entendido atualmente (vide primeiro parágrafo), surgiu há cerca de 57 anos com o trabalho publicado por Corwin Hansch e colegas (1962), no qual os autores tentam correlacionar estrutura e atividade pesticida de uma série de moléculas (DEARDEN et al., 2009). Desde 1962, muitos trabalhos envolvendo análises QSAR/QSPR foram desenvolvidos, sendo que os mais recentes apresentam diversos métodos estatísticos sofisticados, tanto para calibração quanto para validação dos modelos construídos. Devido a essa vasta gama de ferramentas estatísticas disponíveis, algumas publicações buscam oferecer guias práticos do que são consideradas (ou não) boas práticas em QSPR (DEARDEN et al., 2009) (TROPSHA, 2010).

Independentemente da metodologia adotada, os guias disponíveis na literatura destacam que esse tipo de análise, que busca correlacionar características estruturais

(**descritores**) de compostos a propriedades biológicas/físico-químicas dos mesmos, apresenta, no mínimo, cinco etapas, as quais são mostradas a seguir (YOUSEFINEJAD; HEMMATEENEJAD, 2015):

1- Seleção ou *design* de um subgrupo de compostos químicos ou biológicos a ser analisado;
2- Geração de **descritores** potentes que sejam capazes de refletir a essência estrutural dos compostos;
3- Seleção dos descritores que serão incluídos no modelo de predição;
4- Construção do modelo de calibração **(vide seção 2.4);**
5- Avaliação da estabilidade e validade do modelo gerado.

Os **descritores** (**etapa 2)** podem ser de dois tipos: **empíricos**, ou seja, propriedades obtidas experimentalmente, ou **teóricos**, os quais são computados com base na estrutura molecular **(descritores moleculares teóricos)** (KATRITZKY et al., 2010) (ESTRADA, 2008).

Descritores moleculares teóricos são representações matemáticas de informações que estão codificadas na molécula e que são obtidas a partir de sua estrutura por um caminho essencialmente teórico (SAHOO et al., 2016). Ainda, de acordo com Consonni e Todeschini:

> O descritor molecular é o resultado final de um procedimento lógico e matemático que transforma informações químicas codificadas com uma representação simbólica de uma molécula em um número útil ou o resultado de algum experimento padronizado. (CONSONNI & TODESCHINI, 2000)

Existem diversas formas de se obter os descritores moleculares teóricos, bem como uma variedade de tipos de descritores. De acordo com o tipo de descritor empregado, a metodologia é classificada de **1D** a **6D**. A abordagem 1D contempla parâmetros relacionados a restrições eletrônicas, hidrofóbicas e estéricas. A 2D se concentra em fragmentos específicos da molécula, sendo que a maioria de seus descritores codificam informações constitucionais, topológicas, químico-quânticas, dentre outras propriedades específicas dos compostos (DARÉ; BARIGYE; FREITAS, 2017). Os descritores 3D contemplam informações espaciais unicamente do ligante, tais como volume da molécula e grupos farmacofóricos. As demais abordagens consistem em complementos ao QSPR-3D, de forma a tratar os sistemas químicos da maneira mais realística possível (DAMALE et al., 2014).

Nessas novas abordagens, busca-se melhorar a representação e a descrição dos compostos químicos. A abordagem 4D, por exemplo, gera uma gama de conformações e considera múltiplos grupos subestruturais. A aproximação em 5D, além de considerar propriedades estruturais do receptor, também inclui o fator flexibilidade do receptor em sua rota de análises. Finalmente, a mais recente abordagem (6D) foi desenvolvida de forma a incluir uma função de solvatação à análise QSPR (DAMALE et al., 2014).

Cada uma dessas abordagens apresenta suas vantagens e desvantagens particulares (DARÉ; BARIGYE; FREITAS, 2017). O QSPR-3D, por exemplo, é uma das metodologias mais empregadas para geração de descritores atualmente; no entanto, apesar de gerar modelos confiáveis, tais modelos são normalmente difíceis de manipular, visto que se faz necessário uma exaustiva busca conformacional e o alinhamento de estruturas de acordo com regras pré-estabelecidas (NUNES; FREITAS, 2013). Além disso, a maioria dos métodos de QSPR-3D utilizam varreduras conformacionais e regras de alinhamento estrutural que não consideram o alvo biológico em si, o que pode gerar interpretações químico-biológicas equivocadas (GUIMARÃES et al., 2016). Por sua vez, abordagens que se baseiam na utilização de descritores físico-químicos (abordagens 2D) para codificar mudanças estruturais, apesar de serem simples e de gerarem modelos de predição confiáveis, muitas vezes não oferecem modelos interpretáveis (NUNES; FREITAS, 2013).

Portanto, uma abordagem que consiga balancear simplicidade e eficácia na geração de descritores seria benéfico. Com isso em mente, a metodologia MIA-QSPR (Análise Multivariada de imagens aplicada em QSPR, do inglês *Multivariate Image Analysis applied to* QSPR) parece ser uma interessante abordagem para modelagem molecular. No entanto, antes de se focar no funcionamento técnico da metodologia MIA-QSPR, uma introdução acerca de análise multivariada de imagens se faz necessária e é dada na seção seguinte.

2.2. Análise Multivariada de Imagens

A análise de imagens com implicações químicas passou por um processo de desenvolvimento relativamente mais lento do que nas demais áreas do conhecimento, como medicina e biologia (GELADI et al., 1992). Os primeiros trabalhos envolvendo esse tipo de análise ocorreram na década de sessenta e eram voltados, principalmente, para a área de medicina (JUAN; FERRER, 2011). No entanto, à medida que métodos para

tratamento de imagens foram se tornando mais acessíveis à comunidade acadêmica, o uso de tais técnicas sofreu uma significativa expansão no campo da química, principalmente por aqueles que lidam com microscopia, RMN (ressonância magnética nuclear), entre outros (GELADI et al., 1992).

Uma imagem, tecnicamente, é descrita como uma distribuição bidimensional de intensidades de luz capturadas de algum meio físico. Espera-se que uma imagem tenha algum significado visual, ou seja, que ela seja capaz de representar algo. Além disso, tratando-se de uma imagem *bitmap* (formada por pixels e não vetores de valores matemáticos), um pixel não deve assumir um valor aleatório ou não-correlacionado de intensidade em relação aos seus vizinhos, pois, dessa forma, não se teria uma imagem. Em outras palavras, os pixels devem apresentar alguma autocorrelação espacial, para que, juntos, correspondam a uma imagem (GELADI et al., 1992).

Existem, basicamente, dois tipos de imagens, as ditas univariadas e as multivariadas. As primeiras são imagens em escala de cinza, compostas por duas dimensões geométricas (x e y) para indexação de pixels individuais, cada um caracterizado por um nível de intensidade (ESBENSEN; GELADI, 1989). Em outras palavras, em uma imagem univariada, cada pixel possui uma coordenada (x,y) própria, à qual é atribuída um valor de intensidade.

Por outro lado, imagens multivariadas funcionam como uma pilha de imagens univaridas, que podem ser vistas como uma matriz de pixels. Nessas imagens, o valor de um pixel não corresponde a um único valor de intensidade de luz, mas sim à contribuição de diferentes canais (bandas) com diferentes comprimentos de onda, como é o caso de uma imagem que segue o modelo RGB (*Red, Green, and Blue*) de cores. Nesse caso, cada pixel possui a contribuição de três bandas correspondentes às cores vermelho, verde e azul. Para se capturar uma imagem RGB por uma câmera fotográfica, por exemplo, são necessários três sensores, cada qual com leituras em intervalos de comprimento de onda correspondentes às cores RGB (ESBENSEN; GELADI, 1989).

Ao se tratar de uma análise multivariada de imagens (MIA, do inglês *Multivariate Image Analysis*), deve-se ter em mente que a correlação (ou autocorrelação), destacada anteriormente, entre as variáveis independentes é um ponto crítico para a análise (ESBENSEN; GELADI, 1989). Nesse caso, faz-se necessário o uso de métodos multivariados (vide **seção 2.4**) para caracterizar a covariância e estrutura de correlação entre as variáveis.

A MIA pode ser empregada para diversos fins, como por exemplo, **classificação, calibração, identificação de padrões, detecção de defeitos,** entre outros (JUAN; FERRER, 2011). Para isso, as imagens precisam ser processadas em um computador. Assim, faz-se necessário que as informações visuais desejadas (descritores) estejam na forma numérica, de forma que o computador consiga interpretá-las (GELADI et al., 1992). Existem diferentes descritores que podem ser utilizados para caracterizar uma informação visual (cor, textura, etc.), sendo que um dos mais utilizados é o parâmetro "cor" (SILVA, 2016). Isso se deve ao fato de ser uma característica robusta que independe do tamanho e da orientação da imagem (SILVA, 2016).

Os modelos de cor podem ser classificados em orientados ao *hardware* (RGB, CMYK, YIQ) e orientados ao usuário (HSI, HSV, etc.) (SILVA, 2016). Dentre esses, o RGB é o mais empregado para descrição de tal propriedade. O MIA-QSPR, descrito na **seção 2.3,** faz uso desse tipo de descritor como metodologia.

Os primeiros trabalhos com MIA consideravam cada pixel da imagem como uma amostra e cada canal de cor do sistema RGB como variável. Depois de decompor as imagens em uma matriz de *m* pixels × 3 variáveis (RGB), o procedimento clássico de redução do espaço das variáveis, análise dos componentes principais (PCA, do inglês *Principal Component Analysis*), era empregado como forma de se preparar os dados para a análise final de interesse. Uma das limitações dessa abordagem é que as informações quanto à posição relativa de um pixel em relação ao outro é perdida (JUAN; FERRER, 2011).

Dessa forma, quando o objetivo principal da análise de imagens não é segmentar objetos particulares da figura, mas sim utilizar toda a imagem para propósitos de classificação ou predição, faz-se mais apropriado resumir as propriedades relevantes da imagem em um grupo de características (descritores) que sejam capazes de caracterizar aquele tipo de imagem. Nesse sentido, cada imagem passa a ser uma amostra para a análise, como é o caso da técnica de MIA-QSPR apresentada a seguir (JUAN; FERRER, 2011).

2.3. Análise multivariada de imagens aplicada em QSPR (MIA-QSPR)

O MIA-QSPR é um método que se baseia no tratamento de imagens bidimensionais (2D) para modelagem molecular. Entende-se que tais imagens contêm informações

topoquímicas e topoestruturais (informações acerca da forma da molécula, efeitos estéricos, centros estereogênicos, etc.) relevantes e úteis para modelagem de propriedades químicas, físico-químicas e biológicas. Como uma imagem digital é composta integralmente por **pixels** (valores binários que constituem a menor unidade de uma imagem digital), seus respectivos valores se tornam as variáveis independentes (**descritores**) que serão utilizadas para a construção do modelo de predição. Em suma, as informações de caráter topoquímico e topoestrutural estão codificadas na forma de valores binários correspondente aos pixels (DARÉ; BARIGYE; FREITAS, 2017).

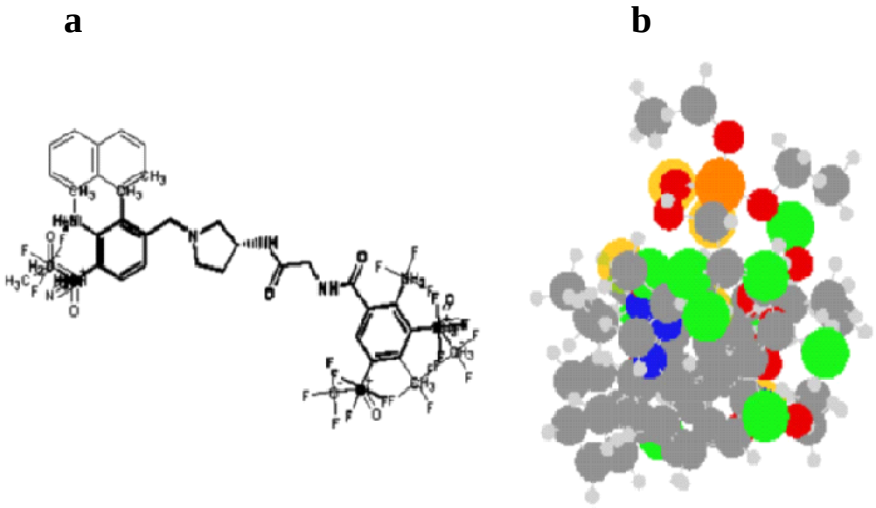
A metodologia MIA-QSPR tem como essência a ideia de que, para uma série de imagens de moléculas congêneres, as mudanças nos substituintes gerarão variações na orientação (posição) dos pixels, o que, por sua vez, explica as variações na variável resposta (propriedade biológica) (NUNES; FREITAS, 2013).

Na primeira versão do MIA-QSPR (FREITAS; BROWN; MARTINS; 2005), os átomos e ligações eram representados unicamente por contornos de cores preta e branca, ou seja, as moléculas consistiam em simples *wireframes*. Tal metodologia, apesar de ter alcançado resultados promissores em várias análises que podem ser encontradas na literatura, possuía algumas limitações, em função do alinhamento impreciso dos *wireframes*. A imprecisão no alinhamento gerava variações nas posições dos pixels, o que, por sua vez, influenciava no modelo de regressão, ou seja, a dificuldade na etapa de alinhamento se tornava uma fonte de erro sistemático para a análise (GOODARZI; FREITAS; FERREIRA; 2009). Além disso, a diferença de tamanho dos átomos não era considerada, o que fazia com que o método falhasse em diferenciar heteroátomos e/ou tipos de átomos (NUNES; FREITAS, 2013).

Dessa forma, uma nova versão denominada *aug*-MIA-QSPR (*augmented* MIA-QSPR) foi desenvolvida (NUNES; FREITAS, 2013). Essa é baseada nas cores oriundas do sistema RGB (*Red*, *Green* and *Blue*) e considera os diferentes raios atômicos na molécula. Os átomos são representados por círculos de tamanhos proporcionais aos respectivos raios de van der Waals e suas cores (valores de pixels) são proporcionais à eletronegatividade. No sistema RGB, o espectro de cores é considerado um produto da contribuição das três cores: vermelho, verde e azul. Cada cor no sistema RGB-8 *bits* é representada por 1 *byte* de memória (8 *bits*), o que permite 256 valores diferentes. Como um único pixel tem a contribuição das três cores, então, ele poderá assumir valores que vão de 0 (ausência de cor - preto) a 765 (contribuição das 3 cores - branco).

O uso de cores e volumes atômicos para diferenciar átomos e grupos químicos fez com que o *aug*-MIA-QSPR melhorasse sua habilidade de predição se comparado com a metodologia convencional (NUNES; FREITAS, 2013). Para um melhor entendimento, a Figura 1 mostra como eram feitas as representações dos compostos segundo o modelo MIA-QSPR original (a) e de acordo com o modelo *aug*-MIA-QSPR (b).

Figura 1- (a) Superposição de imagens utilizadas no modelo MIA-QSPR original; **(b)** Superposição de imagens utilizadas no modelo *aug*-MIA-QSPR.



Fonte: (a) Introducing New Dimensions in MIA-QSAR: A case for chemokine receptor inhibitors (NUNES; FREITAS, 2013);
(b) Revealing chemophoric sites in organophosphorus insecticides through the MIA-QSPR modeling of soil sorption data (DARÉ; SILVA; FREITAS, 2017).

A princípio, os valores (das cores) atribuídos aos descritores da técnica *aug*-MIA-QSPR eram aleatórios (variando de 0-765), ou seja, não continham nenhum significado químico; eles serviam apenas para a distinção das diferentes partes das moléculas. No entanto, em 2016, a técnica passou por novas modificações, e, a partir de então, esses valores sem significado passaram a ser substituídos por valores proporcionais à eletronegatividade de Pauling, ao raio atômico, e ainda pela razão desses dois parâmetros (BARIGYE; FREITAS, 2016).

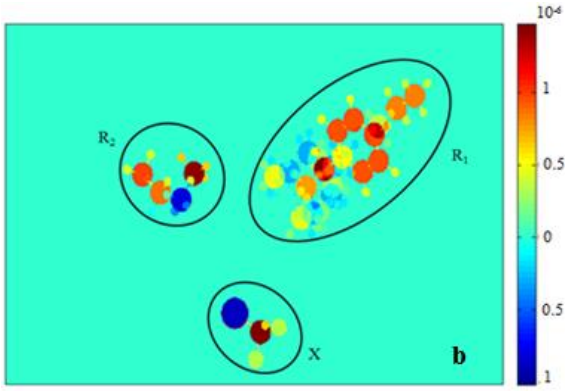
Essa modificação facilitou a interpretação química dos modelos de predição gerados, pois se passou a observar os seguintes aspectos: se o modelo envolvendo valores proporcionais à eletronegatividade for melhor que os demais, isso significa que tal parâmetro dos substituintes é mais relevante para o padrão da propriedade biológica/físico-química observado do que o volume atômico dos mesmos; o recíproco também é verdadeiro; por outro lado, caso o modelo com valores proporcionais à razão

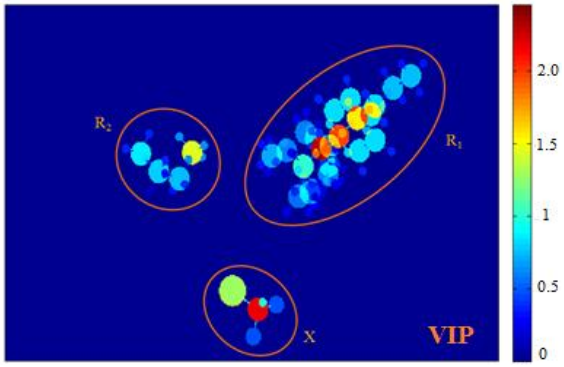
ε/r_{vdW} seja melhor que os demais, isso significa que tanto a eletronegatividade quanto o volume dos grupos substituintes influenciam de maneira relevante na propriedade estudada e, consequentemente, também são importantes para a modelagem da mesma.

Novos melhoramentos da técnica MIA-QSPR ocorreram ainda em 2016. Foi desenvolvida uma ferramenta gráfica (MIA-Plot) que auxilia na interpretação química dos modelos gerados, ou seja, ela contribui para o entendimento de como os diferentes substituintes e suas respectivas posições afetam a propriedade de interesse. A necessidade do desenvolvimento de tal ferramenta surgiu do fato de que a redução significativa da quantidade de descritores pelo método de **seleção de variáveis** era um verdadeiro desafio e a aplicação de tal metodologia resultava em perda de informações relevantes. Em termos práticos, essa nova ferramenta consiste na utilização de dois gráficos de cores, o primeiro é chamado de gráfico de coeficientes de regressão do PLS (**b**) e o segundo de gráfico de projeção da importância das variáveis para o grupo de compostos (**VIP**) (BARIGYE et al., 2016).

Um *score* VIP mostra a importância dos átomos nas moléculas para a determinação do modelo de projeção PLS, tanto para as varáveis preditoras quanto para a variável resposta, ou seja, ele permite a visualização da importância das variáveis para o modelo. Por outro lado, um coeficiente de regressão indica como uma variável afeta a resposta, *i.e.*, um valor alto de **b** (positivo ou negativo) indica que tal variável tem uma maior relevância (correlação) para a propriedade modelada. Assim sendo, uma análise conjunta dos gráficos VIP e **b** é recomendada. A Figura 2, mostrada a seguir, corresponde a um exemplo de utilização do MIA-Plot (BARIGYE et al., 2016).

Figura 2- MIA-Plots baseados nos parâmetros **VIP** e **b** para derivados de triazinas.





Fonte: Multi-Objective Modeling of Herbicidal Activity from an Environmentally Friendly Perspective (DARÉ, et al., 2017)

Existem diferentes formas de extrair os descritores de interesse de imagens, dependendo do tipo de dado com que se está lidando. No caso do MIA-QSPR, as informações relevantes são aquelas relacionadas com mudanças estruturais que levam a variações na propriedade biológica estudada. Assim sendo, os descritores de interesse são selecionados por meio de exclusão dos que não apresentam variações ao longo do conjunto de imagens (espaços em branco e núcleo comum). Depois de se obter o conjunto de descritores, deve-se utilizar um método adequado de análise multivariada para lidar com o conjunto de dados (**seção 2.4**).

2.4. Métodos de calibração multivariada aplicados em QSAR/QSPR

Uma técnica comumente utilizada em laboratório quando se deseja encontrar a concentração de um analito, através do parâmetro absorbância, é a construção da chamada curva analítica. Segundo essa metodologia, o analista escolhe um comprimento de onda adequado e define a amostra sem analito como sendo aquela com absorbância ‘0’ e, a partir disso, faz leituras consecutivas de amostras com concentrações crescentes e conhecidas do analito, até que se atinja um valor máximo de absorbância (‘1’). Então, uma curva de absorbância × [analito] é construída e a equação (modelo matemático, onde y = absorbância e x = concentração) que a descreve é determinada. Diz-se que a equação matemática é obtida por um método de calibração univariada, pois, considerando um único comprimento de onda, existe apenas uma variável correlacionada com a propriedade de interesse (absorbância). A partir disso, qualquer amostra de analito com absorbância

dentro da faixa limite, pode ter sua concentração determinada pela simples leitura da absorbância e aplicação do modelo matemático.

Da mesma forma, em uma análise QSPR, um dos passos principais consiste na geração do modelo de predição (equação) a partir de um grupo de moléculas do qual se sabe os valores da propriedade biológica (Y) e dos descritores (X). No entanto, tal análise se diferencia de uma calibração comum pela quantidade de variáveis independentes que descrevem o sistema. Enquanto em uma análise comum de absorbância, o analista lida apenas com uma variável ([analito] em um comprimento de onda específico), em uma análise de QSPR pode ser que exista até dezenas de milhares de variáveis, como é o caso do MIA-QSPR. Portanto, há a necessidade de métodos de uma classe diferente para se obter a equação que descreva a correlação entre descritores e propriedade biológica. Tal classe reúne os chamados **métodos de calibração multivariada.**

O método mais simples de calibração multivariada é a regressão linear múltipla (MLR, do inglês *Multiple Linear Regression*). O modelo gerado assume a seguinte forma geral:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n \qquad \textbf{(Eq. 1)}$$

em que b_0 - b_n são os coeficientes de regressão. Esses coeficientes podem ser estimados por regressão dos mínimos quadrados ou por métodos de regressão robustos. O modelo gerado (Eq. 1) pode ser utilizado tanto para se investigar como cada descritor influencia na variável resposta, quanto para avaliar o efeito conjunto dos descritores na variável resposta (LIU; LONG, 2009). Essa última característica (efeito conjunto), de certa forma, justifica a utilização de MLR em vez de múltiplas regressões univariadas.

A modelagem por MLR funciona bem desde que os descritores sejam totalmente não correlacionados, o que é uma condição difícil de ser satisfeita quando se lida com dados químicos, pois a quantidade de descritores costuma ser bem maior do que o número de amostras (WOLD; SJOSTROM, 2001). O problema de se ter muitos descritores é que o modelo gerado, normalmente, superajusta-se aos dados que o geraram (grupo *training*); a esse problema se dá o nome de *overfitting*.

Como método alternativo, a técnica dos mínimos quadrados parciais (PLS, do inglês *Partial Least Squares*) é um método linear que consegue, ao contrário do MLR, lidar com conjuntos de dados que tenham mais descritores do que amostras e que possuam alta colinearidade (correlação entre variáveis independentes) (WOLD; SJOSTROM, 2001). Essa vantagem faz com que o PLS seja um dos métodos de regressão linear mais utilizados na atualidade.

O funcionamento do PLS, de uma forma geral, se baseia na tentativa de extrair as chamadas variáveis latentes (LV) da matriz X, as quais, em teoria, trazem consigo a maior parte da variação (informações) do conjunto de dados responsável pelo perfil observado da propriedade biológica modelada, ou seja, esse método tenta extrair os descritores que realmente influenciam na variável resposta (TOBIAS, 2016).

A matriz que contém as variáveis latentes das amostras é chamada de *T* e a que contém as variáveis com informações acerca da resposta é chamada de *U*. As variáveis latentes *T* (também chamadas de *X-scores*) são usadas para predizer os *Y-scores* (*U*), e esses *Y-scores* são, então, utilizados para construir predições para as respostas. Vale ressaltar, ainda, que os *X* e *Y-scores* são escolhidos de forma que a correlação entre pares sucessivos de *scores* seja a maior possível. Em princípio, a técnica PLS procura direções no espaço de variáveis que estejam associadas com as maiores variações na variável resposta (TOBIAS, 2016).

Existem muitos outros métodos multivariados de calibração, tanto lineares, como é o caso do MLR e do PLS, quanto não-lineares como é caso do *Random Forest* e das chamadas redes neurais (VARMUZA; FILZMOSER, 2009). Normalmente, o desempenho desses diferentes tipos de algoritmo é avaliado por meio dos chamados parâmetros de validação. A etapa de validação, citada na seção 2.1, consiste no processo de avaliação da estabilidade de um modelo gerado, bem como sua acurácia e exatidão em prever uma variável de interesse (TROPSHA, 2010).

Dentre os diversos parâmetros utilizados para validação, os mais comuns são: r^2 (coeficiente linear de correlação), q^2 (validação interna cruzada), r^2_{teste} (validação externa), $r^2_{y\text{-}rand}$ (teste y), RMSE (*Root Mean Square Deviation*) e MSE (*Mean Square error*). Por se tratar de um procedimento quantitativo, existem valores específicos que caracterizam um modelo como bom ou ruim. Tais valores possuem significado estatístico e estão tabelados na literatura em trabalhos como de Tropsha (2010) e Dearden et al. (2009).

2.5. Conformação em análise QSPR

A abordagem QSPR tridimensional (QSPR-3D) é a mais explorada em estudos de correlação entre estrutura e propriedade biológica. Um dos motivos de tal notoriedade é o fato desse tipo de abordagem contemplar informações espaciais (estereoquímicas) das moléculas nos modelos, mais especificamente, informações de cunho conformacional. De

fato, os descritores da metodologia 3D são baseados em conformações com valores pequenos de energia, ou seja, conformações estáveis (LEWIS; WOOD, 2014). No entanto, não existem evidências de que a utilização de uma conformação no lugar de outra altere modelos QSPR, apesar de sua importância em outros quesitos, como em interações e reações biológicas envolvendo enzimas e ligantes (DE FREITAS; RAMALHO, 2013).

A estereoquímica de moléculas é responsável por governar diversas propriedades de compostos, como por exemplo, a seletividade e reatividade dos mesmos. Uma molécula que apresente um ou mais graus de liberdade para a rotação exibirá, necessariamente, isomerismo conformacional (DE FREITAS; RAMALHO, 2013).

Uma conformação, ou rotâmero, como é muitas vezes chamada, consiste em uma dada distribuição espacial (estado) da molécula gerada pela rotação de uma ou mais ligações simples, sendo que cada uma das conformações tem um determinado valor de energia. Sendo assim, existem conformações mais estáveis (com um menor valor de energia associado a ela) e conformações menos estáveis (com valores de energia mais altos). O exemplo mais simples que apresenta conformações com diferentes estabilidades é o etano (POPHRISTIC; GOODMAN, 2001) (SOLOMONS & FRYHLE, 1999).

A conformação desempenha um papel fundamental em processos biológicos (NUSSINOV; TSAI, 2014). A função de uma proteína, por exemplo, está intimamente ligada com sua estrutura e organização espacial (estereoquímica). Essa função é determinada com base no quanto uma macromolécula ocupa sua conformação ativa (complexo proteína-ligante), ou seja, ela é determinada com base no ligante (NUSSINOV; TSAI, 2014). A essa conformação nativa da proteína dá-se o nome de **bioconformação**. Com isso em mente, faz-se interessante que técnicas que envolvam a predição de atividade proteína-ligante, como o QSPR, considerem informações relacionadas à bioconformação da proteína e do ligante em questão. Porém, vale ressaltar, novamente, que não existem evidências que comprovem uma relação entre inclusão desse tipo de informação e melhoria dos modelos QSPR.

Os métodos de QSAR 2D, como explicado na seção dedicada à técnica de MIA-QSPR, normalmente, são baseados puramente em características relacionadas à conexão dos átomos na estrutura, ou seja, informações conformacionais não são contempladas (LEWIS; WOOD, 2014). No entanto, isso não significa que esses descritores não sejam capazes de codificar informações tridimensionais de moléculas (ESTRADA; MOLINA; PERDOMO-LÓPEZ, 2001).

Por último, conformação é um conceito essencial para a técnica de ancoramento molecular discutida a seguir.

2.6. Ancoramento molecular

Muitas vezes empregado em conjunto com a técnica de QSPR, o **ancoramento molecular**, também chamado de *docking* molecular, consegue complementar o processo de planejamento de novos candidatos a fármacos/agroquímicos em aspectos bastante relevantes (SANT'ANNA, 2009). Essa ferramenta computacional, amplamente utilizada, relativamente rápida e econômica, é empregada para predição *in silico* dos modos de ligação e afinidades envolvidos nos eventos de reconhecimento molecular (DU et al., 2016). Essas predições são de extrema importância no raciocínio de novos candidatos a fármacos/agroquímicos, uma vez que a vida celular, de uma forma geral, está intimamente ligada a um gigantesco número de interações seletivas e específicas entre biomacromoléculas (CAVASOTTO; AUCAR; ADLER, 2018).

O chamado *docking* proteína-ligante é uma subclasse do termo mais geral muito importante, pois tal metodologia está envolvida no planejamento de novos candidatos a fármacos, ou seja, está envolvida na triagem virtual de compostos com possível atividade biológica (DU et al., 2016). O objetivo central do *docking* proteína-ligante é encontrar e quantificar o modo de interação de uma molécula ligante em um receptor, de modo que a atividade do receptor seja inibida/potencializada (MORGON; COUTINHO, 2007); isso é feito através da otimização da pose (estrutura) de uma molécula pequena (ligante) no sítio de ligação do receptor e, então, a melhor pose (ou as moléculas de menor energia) recebe uma pontuação, a qual indica a probabilidade de ligação com o receptor (CAVASOTTO; AUCAR; ADLER, 2018).

Sendo assim, tecnicamente, o *docking* proteína-ligante possui dois componentes principais: o algoritmo de busca de confôrmeros e a função de pontuação (função *score*). O algoritmo tem como tarefa a busca pelas diversas conformações que o ligante pode assumir, bem como suas respectivas posições (poses) no sítio de ligação da proteína em questão. A função *score*, por sua vez, deve estimar as afinidades de ligação (energias livres de ligação) para cada uma das poses geradas, ranqueá-las e eleger o modo de ligação mais favorável do ligante com a proteína (DU et al., 2016).

A maneira como a função *score* irá estimar as afinidades de ligação dependerá da classe a que ela pertence. Existem três classes principais de funções: funções baseadas em campo de força, funções empíricas e aquelas baseadas em conhecimento (KITCHEN et al., 2004).

Considerando a primeira classe, faz-se necessário, inicialmente, definir o que é um **campo de força**. Um campo de força consiste na representação física de um sistema em simulações *in silico*. De uma forma mais técnica, ele é um funcional (função dependente de outras funções) que inclui parâmetros de ajuste às suas respectivas funções (MORGON; COUTINHO, 2007). Esses parâmetros podem ser de fonte experimental ou teórica (cálculos de mecânica quântica) (DU et al., 2016). Exemplos comuns de campos de força são: MM3, MM4, AMBER, OPLS, CHARMM e MMFF94 (WANG et al., 2004).

A estimativa das afinidades de interação por campo de força, normalmente, requer um alto custo computacional e, por isso, de forma a minimizar a complexidade desses cálculos, geralmente, apenas forças de interações intermoleculares não-covalentes (contribuição da entalpia) são estimadas para o complexo proteína-ligante (DU et al., 2016). Uma forma de tornar os resultados mais acurados é levar em consideração o efeito do solvente de forma explícita ou por um modelo implícito. Vale ressaltar, por fim, que a maioria dos campos de força dividem a função energia potencial (utilizada para o cálculo da energia livre de ligação) em termos de contribuições entre átomos ligados e não-ligados (MORGON; COUTINHO, 2007).

Funções *score* do tipo **empírico** são uma alternativa mais veloz à abordagem de campo de força, o que se deve ao alto grau de simplificação que elas apresentam. Fundamentalmente, elas se baseiam na ideia de que as energias de ligações podem ser aproximadas pela soma de termos individuais não correlacionados. Os coeficientes desses vários termos são obtidos através de análises de regressão usando energias de ligação determinadas experimentalmente e, caso disponível, informações estruturais cristalográficas (Raio-X) (KITCHEN et al., 2004).

Sendo assim, o maior desafio da segunda abordagem é obter parâmetros de energia acurados para serem empregados na determinação da afinidade. Por fim, uma das grandes desvantagens de se trabalhar com este tipo de função *score* é a possibilidade de um superajuste (*overfitting*) decorrente da utilização de diversos parâmetros empíricos pré-otimizados (DU et al., 2016).

As funções baseadas em conhecimento se fundamentam na hipótese de que as interações interatômicas mais próximas entre ligantes, as quais ocorrem mais frequentemente do que aquelas com distribuição aleatória, são prováveis de serem energicamente favoráveis e de, portanto, contribuírem na afinidade de ligação. Nesse tipo de abordagem, grupos considerados estatisticamente próximos em um grupo de calibração (*training set*) são utilizados para se derivar potenciais estatísticos. Essa abordagem tem a vantagem de ser mais rápida do que a de campo de força e menos provável de se superajustar a um grupo de moléculas (MORGON; COUTINHO, 2007).

Cada uma das classes de função *score* tem suas vantagens e desvantagens. Sendo assim, com o intuito de melhorar a acurácia e aplicabilidade destas funções, recentemente, tem-se buscado combinar *scores* provenientes de múltiplas funções *scores*. A esta estratégia deu-se o nome de estratégia de *scores* consensual (DU et al., 2016).

Com relação ao algoritmo de busca de confôrmeros, sabe-se que considerar todas as conformações e poses da proteína e do ligante (tanto em suas formas livres quanto interagindo uma com a outra) torna a técnica de ancoramento molecular inviável (considerando os recursos computacionais da atualidade); assim, surge um impasse.

Com isso em mente, o maior desafio em uma análise de *docking* é como lidar, de forma eficiente, com a flexibilidade da proteína, pois existe um dilema a ser considerado: levar em consideração todos os graus de liberdade não é possível, mas negligenciá-los gera resultados ruins de predição das poses. Dessa forma, com o intuito de melhorar a técnica de ancoramento, os algoritmos de busca evoluíram daqueles que consideravam tanto o ligante quanto a proteína rígidos para os que passam a considerar ambos parcialmente flexíveis (DU et al., 2016).

Em suma, com relação aos valores preditos de energia livre de ligação (afinidade) pela função *score* e aos confôrmeros ranqueados pelo algoritmo, tais resultados podem ser assegurados desde que a técnica tenha sido bem executada e tenha-se levado em consideração todos os fatores envolvidos, como por exemplo, a protonação ou desprotonação de grupos da proteína ou do ligante. A predição computacional acurada e robusta das forças de interação entre proteína e ligante pode ser considerada a essência do *design* virtual de moléculas bioativas (YILMAZER; KORTH, 2016).

Por fim, vale ressaltar que o *docking* é uma ferramenta pertencente ao método de mecânica molecular, o qual foi desenvolvido com base em aproximações clássicas (equações de Newton). Na mecânica molecular, as moléculas são descritas como um

conjunto de “átomos conectados”, em vez de núcleos e elétrons. Tal abordagem se baseia nas primícias de que os parâmetros associados a conjuntos de átomos permanecem razoavelmente constantes entre estruturas diferentes, desde que o tipo e a hibridação dos átomos envolvidos sejam os mesmos (SANT’ANNA, 2009).

2.7. Grupos de compostos empregados nas análises MIA-QSPR

2.7.1. Inseticidas Organofosforados

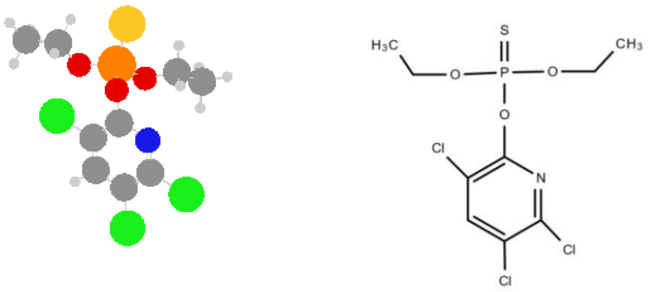
Pesticidas são compostos orgânicos tóxicos utilizados para proteção de culturas agrícolas e sementes, o que é feito através da destruição de ervas daninhas (herbicidas), bactérias (bactericidas), fungos (fungicidas), insetos (inseticidas), dentre outras pragas (MALVANO et al., 2017). De acordo com a Organização Mundial da Saúde, existem mais de 1000 pesticidas empregados atualmente durante o cultivo de alimentos para garantir que esses não sejam atacados por pragas (WORLD HEALTH ORGANIZATION, 2018). Dentre os diversos pesticidas, os organofosfatos (OF) e os carbamatos estão entre os mais utilizados em razão de sua alta atividade inseticidas e baixa persistência se comparados com outros, tais como os organoclorados e lindano (já banido em países que assinaram o tratado da Convenção de Estocolmo) (MALVANO et al., 2017) (WORLD HEALTH ORGANIZATION, 2018). No entanto, apesar de muito utilizados, tais produtos podem oferecer riscos de contaminação para recursos hídricos, grãos, alimentos e, consequentemente, ao ser humano (KAPOOR; RAJAGOPAL, 2011).

Os OF, em especial, têm sido alvo de diversos estudos, pois uma pesquisa recente revelou que mais de 3 milhões de pessoas são expostas a esse tipo de agente químico a cada ano, sendo causa de cerca de 300.000 mortes; vale ressaltar que, apesar da maior parte dessa contaminação ser por pesticidas, organofosfatos também estão presentes em medicações e itens domésticos, como por exemplo os *sprays* contra insetos (ROBB; BAKER, 2018). A exposição a esse composto químico pode se dar de forma direta (durante aplicação) ou indireta (através de alimentos contaminados), e sua toxicidade está relacionada, principalmente, ao efeito inibitório da acetilcolinesterase, uma enzima com papel fundamental na transmissão nervosa. A inibição dessa enzima gera fraqueza nos

músculos, falha respiratória, perda de consciência e, eventualmente, morte (MALVANO et al., 2017).

No caso dos pesticidas organofosforados, as moléculas bioativas possuem em sua estrutura um átomo de fósforo central, ligado a átomos de oxigênio, conforme exemplificado na Figura 3.

Figura 3 – Representação esquemática utilizada na metodologia MIA-QSPR do composto Clorpirifos: agente agroquímico organofosforado. O = vermelho, Cl = verde, N = azul, C = cinza, H = cinza claro, P = laranja, e S = amarelo.



Quanto à persistência no solo, ainda que os OF apresentem uma menor permanência se comparado a outras classes de agroquímicos, isso não deixa de ser uma limitação. Sendo assim, estudos que busquem detalhar a relação entre estrutura química e persistência no solo (logKoc por exemplo) são indispensáveis, de forma que possibilite o *design* de compostos menos agressivos ao meio ambiente.

2.7.2. Compostos anti-HCV

De acordo com a Organização Mundial da Saúde, em 2015, 71 milhões de pessoas viviam afetadas com um dos seis principais genótipos do vírus da hepatite C e 399.000 morreram em decorrência dessa infecção (BULTERYS; HAMID, 2018).

O vírus da hepatite C (HCV) faz parte da família *Flaviviridae* e age, predominantemente, através da infecção dos hepatócitos. A infecção por HCV é uma doença silenciosa que, em longo prazo, pode levar a sérios problemas de fígado, tais como fibrose e cirrose, seguido de carcinoma hepatocelular (KONREDDY et al., 2014). Quanto ao seu tratamento, os medicamentos atuais têm se mostrado de eficácia limitada e não existem vacinas disponíveis (BULTERYS; HAMID, 2018). Um dos fatores que contribuem para essa limitação é o fato de que o mecanismo de ação dessa classe de vírus ainda não foi completamente elucidado (KONREDDY et al., 2014). Diante disso, diversos

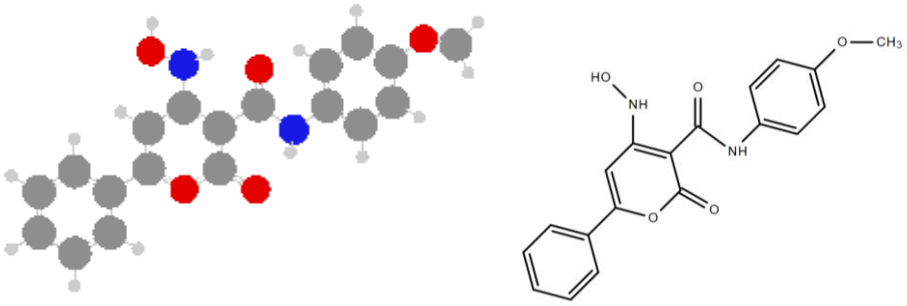
estudos têm sido realizados com o intuito de se elucida seu mecanismo de ação e identificar medicamentos antivirais de ação direta (sigla em inglês: DAAs).

Konreddy e colegas (2014) sintetizaram uma série de análogos da 4-hidroxiamino-α-piranona carboxamida e avaliaram sua eficácia como agentes anti-HCV. Foram testados mais de 40 compostos e uma quantidade significativa dessas moléculas apresentaram resultados promissores no combate ao vírus da hepatite C.

Diante disso, Li et al. (2016) optaram por investigar mais detalhadamente essa série de compostos. Para tal, foram realizados estudos de QSAR-3D a fim de construir um modelo capaz de predizer a atividade antiviral de compostos análogos à 4-hidroxiamino-α-piranona carboxamida que ainda não foram testados e, ainda, identificar os requisitos estruturais para a bioatividade dessas moléculas. Com base nos resultados encontrados, os autores estabeleceram guias para a síntese de novos compostos mais eficientes (LI et al., 2016). No entanto, vale ressaltar que estudos mais aprofundados precisam ser realizados, pois essa classe de compostos ainda não tem um alvo biológico completamente definido (KONREDDY et al., 2014) e, bem se sabe que, além de ser bioativo, o medicamento precisa atender a outros requisitos, tais como: ter uma boa farmacocinética, baixa toxicidade, dentre outros.

A Figura 4 mostra o composto que apresentou o maior valor de pEC_{50} (concentração necessária de um composto para que a metade da resposta máxima seja obtida $\left(\frac{Vmáx}{2}\right)$) nos estudos de Li et al. (2016).

Figura 4 - Representação esquemática utilizada na metodologia MIA-QSPR do composto anti-HCV mais ativo. O = vermelho, N = azul, C = cinza, e H = cinza claro.



2.7.3. Compostos anti-SARS-CoV

A síndrome respiratória aguda severa (SARS) é uma doença infecciosa, descoberta em 2002, que teve suas primeiras aparições no sul da China e afetou cerca de 8000

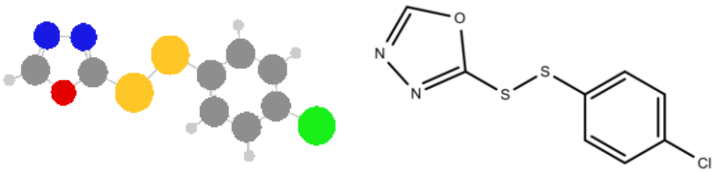
pessoas ao redor do mundo, das quais 800 chegaram a óbito (WORLD HEALTH ORGANIZATION, 2003).

Apesar de a epidemia estar controlada atualmente (desde 2005 não foram reportados mais casos), a reincidência de tal doença ainda é um risco eminente e novas manifestações podem ser ainda mais perigosas que a primeira (KUMAR; JUNG; LIANG; 2013). Em função desse contínuo receio, vários alvos biológicos, críticos para etapa de replicação do vírus da SARS (SARS coronavírus ou SARS-CoV), vêm sendo estudados nos últimos anos com o objetivo de se desenvolver novos inibidores. Dentre os alvos identificados, a M^{pro} (protease principal) chamou a atenção de diversos pesquisadores e vários grupos de compostos vêm sendo propostos como inibidores em potencial para tal enzima (WANG et al., 2017).

Wang e colaboradores (2017) identificaram uma série de 40 dissulfetos aromáticos assimétricos que, quando testados *in vitro* contra a M^{pro} do SARS-CoV, apresentaram bons valores de IC_{50} (concentração necessária de um inibidor para que a resposta seja reduzida pela metade). O modo de ligação dessas moléculas também foi estudado através da técnica de *molecular docking* e estudos de QSAR-3D também foram efetuados.

Um exemplar dessa série de compostos anti-SARS é exibido na Figura 5.

Figura 5 - Representação esquemática utilizada na metodologia MIA-QSPR do composto anti-SARS-CoV mais ativo (Composto 31). O = vermelho, Cl = verde, N = azul, C = cinza, H = cinza claro, e S = amarelo.



3. MATERIAL E MÉTODOS

3.1. Modelagem MIA-QSPR Tradicional

Para a primeira parte do trabalho, um grupo de compostos organofosforados, contendo 24 moléculas com propriedade inseticida, foi selecionado (Artigo 1 – Tabela 1). De forma a garantir uniformidade, os valores de logKoc (parâmetro de sorção no solo) foram obtidos de um mesmo banco de dados, o "ChemSpider" (WILLIAMS, 2016). Utilizando o programa *GaussView* (DENNINGTON RD, KEITH TA, 2008), as moléculas foram construídas, de forma que o núcleo comum permanecesse na mesma posição, em termos de pixels, em todas elas, para garantir que, assim, a superposição dos compostos fosse perfeita nos pontos em comum. As imagens tinham a dimensão de 300 × 354 pixels cada.

Em seguida, utilizando-se o programa *Chemoface* (NUNES et al., 2012), gerou-se um arranjo tridimensional de dimensão 24 × 300 × 354, resultante da sobreposição da série congênere. Esse arranjo foi desdobrado em uma matriz de dimensão 24 × 106200, na qual cada linha corresponde a uma molécula e cada coluna corresponde a um pixel de uma determinada posição da imagem; em outras palavras, a coluna 1, por exemplo, corresponde ao pixel de posição (0,0) de cada uma das imagens.

As colunas com variância '0' foram excluídas e, em seguida, os valores das cores correspondentes a cada átomo foram substituídos por valores proporcionais a r_{vdW}/ε, em que r_{vdW} corresponde ao raio de van der Waals e ε corresponde à eletronegatividade de Pauling. Em seguida, deu-se início à análise QSPR em si. Primeiro, a matriz desdobrada foi dividida nos grupos de calibração (75% das amostras) e teste (25% das amostras) empregando-se o algoritmo Kennard-Stone. O procedimento de calibração foi feito utilizando-se PLS, sendo que o número de variáveis latentes foi escolhido analisando-se o decaimento do RMSE (erro quadrático médio, do inglês *Root Mean Square Error*) na validação cruzada do tipo *Leave-One-Out*.

A qualidade do modelo de predição foi avaliada por intermédio de parâmetros provenientes da validação interna feita por *leave-one-out* (q^2 e RMSE correspondente) e externa (r^2_{teste} e RMSE correspondente). Além disso, a proximidade entre os valores experimentais de logKoc e os calculados para o grupo teste foi avaliada estatisticamente utilizando os parâmetros r^2_m $\{r^2_m = r^2 \times [1-(r^2 - r^2_0)^{1/2}]\}$, em que r^2 e r^2_0 são os coeficientes de determinação entre o logKoc observado e predito com e sem intercepto, respectivamente. O

risco de *chance correlation* foi analisado através do parâmetro $^{c}r^2_p$ [$^{c}r^2_p = r \times (r^2 - r^2_{y\text{-}random})^{1/2}$], em que $r^2_{y\text{-}random}$ corresponde ao valor do coeficiente de determinação médio obtido depois de randomizar o bloco y dez vezes. O domínio de aplicabilidade foi avaliado plotando-se o gráfico de *sample leverages* e *Student's residuals*. Por último, para fins de interpretação, os mapas de contorno MIA foram obtidos plotando-se os coeficientes de regressão PLS (**b**) e os *scores* do gráfico VIP (importância das variáveis na projeção, do inglês *Variable Importance in Projection*).

3.2. Modelagem MIA-QSPR utilizando projeções 2D obtidas a partir de compostos com geometrias otimizadas

Para a segunda parte do trabalho, um grupo de compostos, contendo 42 moléculas com atividade anti-HCV (vírus da hepatite C, do inglês *Hepatitis C Virus*), foi selecionado na literatura (LI et al., 2016). Inicialmente, modelos MIA-QSPR tradicionais foram construídos para esse conjunto de amostras, seguindo os mesmos passos citados no **item 3.1**, sendo que a variável resposta (atividade anti-HCV) utilizada no modelo foi expressa em termos de pEC_{50}. O conjunto teste escolhido, constituído por 9 compostos, foi o mesmo utilizado por Li e colaboradores (LI et al., 2016); isso foi feito de forma a garantir o mesmo *domain space*, para que, posteriormente, os resultados pudessem ser comparados.

Para gerar o modelo com projeções obtidas de moléculas com geometrias otimizadas, inicialmente, fez-se uma varredura conformacional utilizando-se a distribuição estocástica de Monte Carlo, para cada uma das 42 moléculas, no *software Spartan'16* (WAVEFUNCTION INC, 2017) (nível de teoria: método semi-empírico AM1 (DEWAR; ZOEBISCH; HEALY; STEWART, 1985)). A conformação de menor energia, em cada caso, foi selecionada e, em seguida, otimizada no nível de teoria ωB97XD/6-31G(d,p) (CHAI; HEAD-GORDON, 2008) (KRISHNAN; BINKLEY; SEEGER, 1980) utilizando-se o programa *Gaussian* (FRISCH et al., 2013).

O próximo passo consistiu no alinhamento (sobreposição) das estruturas otimizadas, o que foi feito empregando-se o *software Discovery Studio Visualizer* (BIOVIA, 2017). As moléculas sobrepostas foram, então, carregadas no programa *GaussView* (DENNINGTON; KEITH, 2008) e salvas com as mesmas configurações e formato daquelas utilizadas no processo de modelagem tradicional MIA-QSPR. Utilizando essa estratégia, informação

conformacional foi incluída nos descritores MIA-QSPR. Após a geração de imagens, os modelos QSPR foram gerados, seguindo a marcha tradicional.

Por fim, os modelos de predição construídos foram comparados e os resultados são discutidos na próxima sessão. Vale ressaltar, ainda, que os compostos inclusos nessa etapa são mostrados na Tabela S1 (Artigo 2).

3.3. Modelagem MIA-QSPR utilizando projeções 2D obtidas a partir de compostos com geometrias otimizadas e bioativas

Um conjunto de moléculas com 40 representantes que apresentam atividade anti-SARS-CoV (coronavírus da síndrome respiratória aguda severa, do inglês *Severe acute respiratory syndrome Coronavirus*) foi escolhido na literatura (WANG et al., 2017) para a realização da terceira etapa desse projeto. A escolha desse grupo de compostos levou em consideração a existência de um alvo biológico bem definido. A variável resposta é expressa em termos de IC_{50}, sendo que esses valores foram medidos diretamente na enzima alvo (M^{pro}, também conhecida como proteína principal).

Assim como nas etapas anteriores, modelos MIA-QSPR tradicionais foram construídos para fins de comparação, sendo que o conjunto teste selecionado era composto por 10 representantes. A Tabela 1 (Artigo 3) mostra os compostos empregados nessa etapa.

Modelos MIA-QSPR também foram construídos para as projeções bidimensionais dos compostos com geometrias unicamente otimizadas, assim como na etapa anterior. Esse procedimento foi realizado com o intuito de consolidar e comparar os resultados da estratégia escolhida para inclusão de informação conformacional nos descritores MIA-QSPR.

Como etapa crucial dessa parte, um tipo mais significativo de informação 3D foi incluído aos descritores MIA-QSPR: informações bioconformacionais, ou seja, informações obtidas de moléculas previamente ancoradas dentro do sítio ativo de seu respectivo receptor. Para tal, inicialmente, obteve-se a estrutura cristalizada SARS-CoV M^{pro} com 1.85 Å de resolução do PDB (banco de dados de proteínas, do inglês *Protein Data Bank* – código: 2AMD), a qual se encontrava complexada com seu inibidor (N9). A estrutura proteica passou pelo processo de preparação para ancoramento molecular no ambiente *Maestro* disponível na suíte de aplicativos *Schrödinger* (Schrödinger LLC, 2011). Os compostos a serem ancorados tiveram suas estruturas previamente otimizadas no nível de teoria ωB97XD/6-31G(d,p) (CHAI; HEAD-GORDON, 2008) (KRISHNAN; BINKLEY; SEEGER, 1980) e, em seguida, foram

submetidos a cálculo de carga (*CHelpG*, no mesmo nível de teoria da otimização estrutural). Por último, após adequar a *Grid Box* ao sítio de interesse, deu-se início ao ancoramento molecular em si, o que foi feito carregando-se as estruturas previamente preparadas (ligantes e receptor) no ambiente de trabalho *Maestro*. Cem conformações foram geradas para cada composto.

Com o intuito de se selecionar as conformações com provável bioatividade, tomou-se como referência o composto **31**, o qual foi identificado como o inibidor mais eficiente por Wang et al. (2017). Uma vez localizada a conformação bioativa do composto **31** (também estabelecida por Wang et al. (2017), procurou-se por conformações com boa sobreposição a essa, dentre aquelas geradas durante o processo de ancoramento molecular. Foi selecionada uma conformação por composto e essas foram salvas para análises posteriores.

As bioconformações obtidas através do processo de ancoramento molecular foram, então, sobrepostas utilizando-se o programa *Discovery Studio Visualizer* (BIOVIA, 2017). Em seguida, as moléculas sobrepostas foram carregadas no programa *GaussView* (DENNINGTON; KEITH, 2008) e salvas com as mesmas configurações e formato daquelas utilizadas no processo de modelagem tradicional MIA-QSPR. Utilizando essa estratégia, informação bioconformacional foi incluída aos descritores MIA-QSPR. Após a geração de imagens, os modelos QSPR foram gerados e avaliados estatisticamente seguindo a sequência tradicional.

4. CONSIDERAÇÕES GERAIS

O presente trabalho consistiu em uma análise detalhada acerca do potencial dos descritores MIA-QSPR em incluir informações conformacionais, além de avaliar os impactos nos modelos de calibração construídos a partir desses descritores modificados. De forma a garantir resultados embasados e concisos, optou-se por dividir tal análise em três partes distintas: i) construção de um modelo tradicional MIA-QSPR a partir de um conjunto de compostos com propriedade biológica independente de sítio ativo ($\log K_{OC}$); ii) construção de modelos MIA-QSPR (tradicional e 3D sem informações do receptor) para um conjunto de moléculas com sítio ativo desconhecido (dispensa a etapa de ancoramento molecular); iii) construção de modelos MIA-QSPR (tradicional, 3D sem informações do receptor, e 3D com informações do receptor) para moléculas com sítio ativo conhecido.

Em suma, concluiu-se que, apesar dos descritores MIA-QSPR serem capazes de codificar informações de cunho conformacional, tal inclusão trouxe variações na estrutura espacial das moléculas que dificultaram a identificação de padrões na qual a técnica se baseia, uma vez que um mesmo grupo funcional passou a ocupar diferentes posições nos diferentes compostos. Sendo assim, tal resultado sugere que a técnica MIA-QSPR, da maneira como é aplicada atualmente, é mais eficiente quando considera apenas representações planas, congêneres e com grupos funcionais iguais nas mesmas posições em todas moléculas.

Por fim, da análise feita, surgem questionamentos relevantes: caso a inclusão das informações conformacionais tenham tornado a relação descritores-variável não linear e, por isso, o método PLS tenha sido ineficaz na determinação de um bom modelo preditivo, outros métodos de análise multivariada, que considerem tal não-linearidade, poderiam resolver o problema? Embora os descritores MIA-QSPR tenham sido capazes de explicar informações espaciais, a técnica não deixou de ser 2D, pois continuou se trabalhando com imagens; nesse sentido, poderiam os princípios dessa técnica serem expandidos para uma abordagem puramente 3D? As respostas a esses questionamentos, as quais são perspectivas futuras para a continuidade desse trabalho, podem trazer contribuições importantes à área de modelagem molecular e, especialmente, às análises QSPR.

REFERÊNCIAS

BARIGYE, S. J.; FREITAS, M. P. Ten Years of the MIA-QSAR Strategy: Historical Development and Applications. **International Journal of Quantitative Structure-Property Relationships**, v. 1, n. 1, p. 64–77, 2016

BARIGYE, S. J.; DUARTE, M; NUNES, C.; FREITAS, M. RSC Advances MIA-plot : a graphical tool for viewing descriptor contributions in MIA-QSAR. **RSC Advances**, v. 6, p. 49604–49612, 2016.

BULTERYS, M.; HAMID, S. S. Treatment of people diagnosed with chronic hepatitis C virus infection. p. 515, 2018.

CHAI, J.-D.; HEAD-GORDON, M. Systematic optimization of long-range corrected hybrid density functionals. **Journal of Chemical Physics**, v. 128, n. 8, 2008.

CAVASOTTO, C.; AUCAR, M.; ADLER, N. Computationa chemistry in drug lead discovery and design. **International Journal of Quantum Chemistry**, v. 119, p. 1-19, 2018.

CONSONNI, V.; TODESCHINI, R. Methods and Principles of Medicinal Chemistry. In: Mannhold, R.; Kubibyi, H. e Timmerman, H. Handbook of molecular descriptors, Wileyn-VCH publishers, 2000.

DAMALE, M. et al. Recent Advances in Multidimensional QSAR (4D-6D): A Critical Review. **Mini-Reviews in Medicinal Chemistry**, v. 14, n. 1, p. 35–55, 2014.

DARÉ, J. K.; BARIGYE, S. J.; FREITAS, M. P. Multi-Objective Modeling of Herbicidal Activity from an Environmentally Friendly Perspective. **International Journal of Quantitative Structure-Property Relationships**, v. 2, n. 2, p. 16–26, 2017.

DARÉ, J. K.; SILVA, C. F.; FREITAS, M. P. Revealing chemophoric sites in organophosphorus insecticides through the MIA-QSPR modeling of soil sorption data. **Ecotoxicology and Environmental Safety**, v. 144, p. 560–563, 2017.

Dassault Systèmes BIOVIA -Discovery Studio Visualizer 2017R2. San DiegoDassault Systèmes BIOVIA, , 2017.

DE FREITAS, M. P.; RAMALHO, T. DE C. Empregando análise conformacional na modelagem molecular de agroquímicos: Observações sobre parâmetros QSAR do 2,4-D. **Ciencia e Agrotecnologia**, v. 37, n. 6, p. 485–494, 2013.

DEARDEN, J. C.; CRONIN, M. T. D.; KAISER, K. L. E. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). **SAR and QSAR in Environmental Research**, v. 20, n. 3–4, p. 241–266, 2009.

DENNINGTON RD, KEITH TA, M. J. **GaussView 5.0**Wallington, 2008.

DEWAR, M. J. S.; ZOEBISCH, E. G.; HEALY, E. F.; STEWART, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: A new general purpose quantum mechanical molecular model. **Journal of the American Chemical Society**, v. 107, p. 3902–3909, 1985.

DU, X. et al. Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods. **International Journal of Molecular Sciences**, v. 17, n. 2, p. 144, 2016.

ESBENSEN, K.; GELADI, P. Strategy of multivariate image analysis (MIA). **Chemometrics and Intelligent Laboratory Systems**, v. 7, n. 1–2, p. 67–86, 1989.

ESTRADA, E. Quantum-chemical foundations of the topological substructural molecular design. **Journal of Physical Chemistry A**, v. 112, n. 23, p. 5208–5217, 2008.

ESTRADA, E.; MOLINA, E.; PERDOMO-LÓPEZ, I. Can 3D Structural Parameters Be Predicted from 2D (Topological) Molecular Descriptors? **Journal of Chemical Information and Computer Sciences**, v. 41, n. 4, p. 1015–1021, 2001.

FREITAS, M.; BROWN, S.; MARTINS, J. MIA-QSAR: A simple 2D image-based approach for quantitative strucuture-activity relationship analysis. **Journal of Molecular Structure**, v. 738, p. 149-154, 2005.

FRISCH, M. J.; TRUCKS, G. W.; SCHLEGEL, H. B.; SCUSERIA, G. E.; ROBB, M. A.; CHEESEMAN, J. R.; SCALMANI, G.; BARONE, V.; MENNUCCI, B.; PETERSSON, G. A.; NAKATSUJI, H.; CARICATO, M.; LI, X.; HRATCHIAN, H. P.; IZMAYLOV, A. F.; BLOINO, J.; ZHENG, G.; SONNENBERG, J. L.; HADA, M.; EHARA, M.; TOYOTA, K.; FUKUDA, R.; HASEGAWA, J.; ISHIDA, M.; NAKAJIMA, T.; HONDA, Y.; KITAO, O.; NAKAI, H.; VREVEN, T.; MONTGOMERY, J. A., JR.; PERALTA, J. E.; OGLIARO, F.; BEARPARK, M.; HEYD, J. J.; BROTHERS, E.; KUDIN, K. N.; STAROVEROV, V. N.; KOBAYASHI, R.; NORMAND, J.; RAGHAVACHARI, K.; RENDELL, A.; BURANT, J. C.; IYENGAR, S. S.; TOMASI, J.; COSSI, M.; REGA, N.; MILLAM, J. M.; KLENE, M.; KNOX, J. E.; CROSS, J. B.; BAKKEN, V.; ADAMO, C.; JARAMILLO, J.; GOMPERTS, R.; STRATMANN, R. E.; YAZYEV, O.; AUSTIN, A. J.; CAMMI, R.; POMELLI, C.; OCHTERSKI, J. W.; MARTIN, R. L.; MOROKUMA, K.; ZAKRZEWSKI, V. G.; VOTH, G. A.; SALVADOR, P.; DANNENBERG, J. J.; DAPPRICH, S.; DANIELS, A. D.; FARKAS, Ö.; FORESMAN, J. B.; ORTIZ, J. V.; CIOSLOWSKI, J.; FOX, D. J. **Gaussian 09, Revision D. 01**,Wallingford CTGaussian Inc., 2009.

GELADI, P. et al. Image analysis in chemistry II. Multivariate image analysis. **Trends in Analytical Chemistry**, v. 11, n. 3, p. 121–130, 1992.

GOODARZI, M.; FREITAS, M.; FERREIRA, E. Influence of changes in 2-d chemical structure drawings and image formats on the prediction of biological properties using MIA-QSAR. **Molecular Informatics**, v.28, n. 4, p. 458-464, 2009.

GUIMARÃES, M. C. et al. Is the bioconformation of 5-deoxy-5-fluoro-d-xylulose affected by intramolecular hydrogen bonds? **RSC Advances**, v. 6, n. 113, p. 111681–111687, 2016.

JUAN, A. DE; FERRER, A. Multivariate image analysis: A review with applications. **Chemometrics and Intelligent Laboratory Systems**, v. 107, n. 1, p. 1–23, 2011.

KAPOOR, M.; RAJAGOPAL, R. Enzymatic bioremediation of organophosphorus insecticides by recombinant organophosphorous hydrolase. **International Biodeterioration and Biodegradation**, v. 65, n. 6, p. 896–901, 2011.

KATRITZKY, A. et al. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. **Chemical reviews**, v. 110, n. 10, p. 5714–5789, 2010.

KITCHEN, D.; DECRONEZ, H; FURR, J.; BAJORATH, J. Docking and scoring in virtual screening for drug discovery: methods and applications. **Nature Reviews Drug Discovery**, v. 3, p. 935-949, 2004.

KONREDDY, A. K. et al. Synthesis and anti-HCV activity of 4-hydroxyamino α-pyranone carboxamide analogues. **ACS Medicinal Chemistry Letters**, v. 5, n. 3, p. 259–263, 2014.

KRISHNAN R, BINKLEY JS, SEEGER R, P. J. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. **J. Chem. Phys.**, v. 72(1), p. 650, 1980.

KUMAR V, JUNG Y-S, LIANG P-H. Anti-SARS coronavirus agents: a patent review (2008 - present). **Expert Opinion on Therapeutic Patents**,v. 23(10), p. 1337-1348, 2013.

LEWIS, R. A.; WOOD, D. Modern 2D QSAR for drug discovery. **Wiley Interdisciplinary Reviews: Computational Molecular Science**, v. 4, n. 6, p. 505–522, 2014.

LI, W. et al. 3D-QSAR study and design of 4-hydroxyamino α-pyranone carboxamide analogues as potential anti-HCV agents. **Chemical Physics Letters**, v. 661, p. 36–41, 2016.

LIU, P.; LONG, W. Current mathematical methods used in QSAR/QSPR studies. **International Journal of Molecular Sciences**, v. 10, n. 5, p. 1978–1998, 2009.

MALVANO, F. et al. A New Label-Free Impedimetric Affinity Sensor Based on Cholinesterases for Detection of Organophosphorous and Carbamic Pesticides in Food Samples: Impedimetric Versus Amperometric Detection. **Food and Bioprocess Technology**, v. 10, n. 10, p. 1834–1843, 2017.

MORGON, N.; COUTINHO, K. Métodos de Docking Receptor-Ligante para o Desenho Racional de

Compostos Bioativos. In: **Métodos de Química Teórica e Modelagem Molecular**. Primeira ed. Campinas: Livraria da Física, 2007. p. 489–526.

NANTASENAMAT, C. et al. A practical overview of quantitative structure-activity relationship. **EXCLI Journal**, v. 8, p. 74–88, 2009.

NUNES, C. A. et al. Chemoface: A novel free user-friendly interface for chemometrics. **Journal of the Brazilian Chemical Society**, v. 23, n. 11, p. 2003–2010, 2012.

NUNES, C. A.; FREITAS, M. P. Introducing new dimensions in MIA-QSAR: A case for chemokine receptor inhibitors. **European Journal of Medicinal Chemistry**, v. 62, p. 297–300, 2013.

NUSSINOV, R.; TSAI, C. J. Unraveling structural mechanisms of allosteric drug action. **Trends in Pharmacological Sciences**, v. 35, n. 5, p. 256–264, 2014.

POPHRISTIC, V.; GOODMAN, L. Hyperconjugation not steric repulsion leads to the staggered structure of ethane. **Nature**, v. 411, n. 6837, p. 565–568, 2001.

ROBB, E.; BAKER, M. **Organophosphate Toxicity**. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK470430/>. Acesso em: 5 dez. 2018.

SAHOO, S. et al. A short review of the generation of molecular descriptors and their applications in quantitative structure property/activity relationships. **Current Computer-Aided Drug Design**, v. 12, n. 3, p. 181–205, 2016.

SANT'ANNA, C. M. R. Molecular modeling methods in the study and design of bioactive compounds: An introduction. **Revista Virtual de Química**, v. 1, n. 1, 2009.

Schrödinger Release 2011: Glide, Schrödinger, **LTC**, New York, NY, 2011.

SILVA, A. T. DA. Descritores de Imagem. p. 29, 2016.

Spartan'16 Software, *Wavefunction Inc.*, Irvine, 2017.

SOLOMONS, G.; FRYHLE. Química Orgânica. LTC, Rio de Janeiro, RJ, 1999.

TOBIAS, R. D. An Introduction to Partial Least Squares Regression. **Institute of Digital Research and Education**, 2016.

TROPSHA, A. Best practices for QSAR model development, validation, and exploitation. **Molecular Informatics**, v. 29, n. 6–7, p. 476–488, 2010.

VARMUZA, Kurt; FILZMOSER, Peter. **Introduction to multivariate statistical analysis in chemometrics**. CRC press, 2016.

WANG, Junmei et al. Development and testing of a general amber force field. **Journal of**

computational chemistry, v. 25, n. 9, p. 1157-1174, 2004.

WANG, L. et al. Discovery of unsymmetrical aromatic disulfides as novel inhibitors of SARS-CoV main protease: Chemical synthesis, biological evaluation, molecular docking and 3D-QSAR study. **European Journal of Medicinal Chemistry**, v. 137, p. 450–461, 2017.

WILLIAMS, A. J. **Chemspider**. Disponível em: <http://www.chemspider.com/>.

WOLD, S.; SJOSTROM, M. PLS-regression : a basic tool of chemometrics. p. 109–130, 2001.

WORLD HEALTH ORGANIZATION. **Pesticide residues in food**. Disponível em: <http://www.who.int/en/news-room/fact-sheets/detail/pesticide-residues-in-food>. Acesso em: 5 dez. 2017.

WORLD HEALTH ORGANIZATION. S**ummary of Probable SARS Cases with Onset of Illness from 1 November 2002 to 31 July 2003**. Disponível em <http://www.who.int/csr/sars/country/table2004_04_21/en/>. Acesso em: 04 Jan. 2019.

YILMAZER, N. D.; KORTH, M. Recent progress in treating protein–ligand interactions with quantum-mechanical methods. **International Journal of Molecular Sciences**, v. 17, n. 5, 2016.

YOUSEFINEJAD, S.; HEMMATEENEJAD, B. Chemometrics tools in QSAR/QSPR studies: A historical perspective. **Chemometrics and Intelligent Laboratory Systems**, v. 149, p. 177–204, 2015.

SEGUNDA PARTE

ARTIGO 1

Journal: Ecotoxicology and Environmental Safety 144 (2017) 560–563

Revealing chemophoric sites in organophosphorus insecticides through the MIA-QSPR modeling of logK_{OC}

Joyce K. Daré,[a] Cristina F. Silva,[b] Matheus P. Freitas[a,*]

[a] *Department of Chemistry, Federal University of Lavras, 37200-000, Lavras, MG, Brazil*

[b] *Department of Biology, Federal University of Lavras, 37200-000, Lavras, MG, Brazil*

* matheus@dqi.ufla.br

ABSTRACT

Soil sorption of insecticides employed in agriculture is an important parameter to probe the environmental fate of organic chemicals. Therefore, methods for the prediction of soil sorption of new agrochemical candidates, as well as for the rationalization of the molecular characteristics responsible for a given sorption profile, are extremely beneficial for the environment. A quantitative structure-property relationship method based on chemical structure images as molecular descriptors provided a reliable model for the soil sorption prediction of 24 widely used organophosphorus insecticides. By means of contour maps obtained from the partial least squares regression coefficients and the variable importance in projection scores, key molecular moieties were targeted for possible structural modification in order to obtain novel and more environmentally friendly insecticide candidates. The image-based descriptors applied encode molecular arrangement, atoms connectivity, groups size and polarity; consequently, the findings in this work cannot be achieved by simple relationship with hydrophobicity parameters, such as the octanol-water partition coefficient.

KEYWODS: QSPR; organophosphorus compounds; soil sorption; environment.

INTRODUCTION

Hydrophobicity, often described in terms of the easily calculable $\log K_{OW}$ parameter (the logarithm of the octanol/water partition coefficient), has long been recognized as a physicochemical property closely related to the soil sorption of organic chemicals.[1] However, the direct correlation between $\log K_{OC}$ (the logarithm of the soil/water normalized to organic carbon) and $\log K_{OW}$ can fail, due to the molecular structural complexity within families of compounds.[2] In addition, $\log K_{OW}$ as a single descriptor is not enough to provide detailed information on molecular features responsible for decreased or increased $\log K_{OC}$, which could be useful to drive the design/synthesis of novel chemical targets through structural modifications in key regions.[3]

Organophosphorus insecticides, as well as other pesticides, have the potential for environmental fate, since they move in soil, sediment, and groundwater. These agrochemicals are among the most well-studied pesticides with respect to the sorption on various soil components as a function of controlling parameters, such as pH.[4] The $\log K_{OW}$ values of a series of phosphate herbicides correlates, in a reasonable extent, with $\log K_{OC}$ (r^2 = 0.726).[2] However, as far as we are aware, the soil sorption of an extensive series of commonly used phosphate, phosphorothioate and phosphorodithioate insecticides has not been modeled yet using more informative descriptors than hydrophobicity. Such a modeling can be carried out using modern quantitative structure-property relationship (QSPR) methods.

Multivariate image analysis applied to QSPR (MIA-QSPR) appeared some time ago as a modeling technique of drugs.[5] Since then, the MIA descriptors used to correlate the molecular structures with the corresponding response variables (such as the soil sorption) have been improved to better encode them. Actually, descriptors in MIA-QSPR are pixels of images forming chemical structures; changes in their coordinates in a workspace reflect the molecular changes responsible for the variance in the response variables block (**y** block). While the former MIA-QSPR models used to represent chemical structures as black wireframes in a white background,[6] the modern approach benefits from augmented images, in which atoms in molecules are circles with sizes proportional to the respective van der Waals radii and numbered (colored) proportionally to a given atomic property (in the present study, the atoms were numbered according to the ratio van der Waals radius/Pauling´s electronegativity (r_{vdW}/ε).[7,8]

Despite more rationally descriptive, the so called augmented MIA-QSPR used to fail in providing clear explanation on which structural moiety most affected the studied property.[9]

Such a problem was solved by introducing MIA-contour maps, in which the descriptor contributions for the studied property are analyzed by plotting the PLS regression coefficients (**b**) and the variable importance in projection (VIP) scores.[10] Thus, the present study aimed at modeling the $\log K_{OC}$ values of a series of organophosphorus insecticides widely used in agriculture, as well as to find structural characteristics responsible for the observed property. Ultimately, our findings can be useful to drive the synthesis of agrochemicals with agricultural (minimum impact on efficacy) and environmental (minimum runoff contamination) benefits.

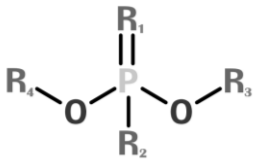
MATERIALS AND METHODS

The QSPR model was built for a series of 24 widely used organophosphorus insecticides (Table 1), whose $\log K_{OC}$ values were obtained from the ChemSpider database in order to guarantee uniformity and considering the proximity with the mean values available elsewhere.[11] The procedure to build MIA-QSPR models has been described in detail elsewhere,[5-10] thus only a brief description is given herein. The chemical structures of the 24 compounds were constructed using the GaussView program,[12] in such a way that the structural moiety in common along the series of compounds (centered in the phosphorus atom) was perfectly superposed to allow for the bidimensional alignment of the images (of 300 × 354 pixels each), which can be checked in Figure 1. The three-way array obtained from the superposition of the 24 images was unfolded to a data matrix of 24 × 106,200 dimension, which was subsequently split into training (75%) and test sets (25%), in order to proceed with calibration and external validation, respectively. Such a procedure was performed using partial least squares (PLS) regression and the optimum number of latent variables was chosen by analyzing the decay of the root mean square error (RMSE) in the leave-one-out cross-validation (LOOCV).

The model´s quality was analyzed by taking into account the determination coefficients in calibration (r^2), cross-validation (q^2) and external validation (r^2_{test}), as well as the corresponding RMSE´s. In addition, the proximity between the actual and predicted $\log K_{OC}$ values for the test set was statistically evaluated using the r^2_m parameter {$r^2_m = r^2 \times [1 - (r^2 - r_0^2)^{1/2}]$, where r^2 and r_0^2 are the determination coefficient values between the observed and predicted $\log K_{OC}$ data with and without the intercept}. The risk of chance correlation in calibration was analyzed using the ${}^c r^2_P$ parameter [${}^c r^2_P = r \times (r^2 - r^2_{y\text{-random}})^{1/2}$, where $r^2_{y\text{-random}}$ corresponds to the mean determination coefficient value obtained after randomizing the **y**

block ten times]. The reference values for these parameters to attest the reliability of the QSPR models are given in the literature.[13-15] The applicability domain of the molecules,[16,17] used to estimate the uncertainty in the prediction of a specific molecule based on how similar it is to the compounds employed to construct the model, was analyzed by plotting the sample leverages and Student´s residuals. The MIA-contour maps[10] were obtained by plotting the PLS regression coefficients (**b**) and the variable importance in projection (VIP) scores, using the Chemoface program.[18] These parameters capture the contributions of the MIA descriptors to the model. Since the pixels were numbered according to the ratio r_{vdW}/ε for each atom, it is expected that the atomic and substructure contributions are related to steric and electrostatic properties, which indeed relate to hydrophobicity.

Table 1. Series of organophosphorus insecticides used in the MIA-QSPR modeling and the corresponding actual and predicted (in calibration, leave-one-out cross-validation and external validation) $\log K_{OC}$ data.



ID	Name	R_1	R_2	R_3	R_4	Measured	Cal.	LOOCV	External val.
1	Dimethoate	S	$-CH_2CONHMe$	Me	Me	1.389	1.589	2.061	
2	Malathion	S	$-CH\ (COOC_2H_5)(CH_2COOC_2H_5)$	Me	Me	1.484	1.489	1.855	
3[a]	Dichlorvos	O	$-OCHCCl_2$	Me	Me	1.604			2.045
4	Phosmet	S	$-CH_2\ (NC_2O_2C_6H_4)$	Me	Me	1.632	1.711	1.943	
5	Crotoxyphos	O	$-C(Me)(CHCOOCHCH_3Ph)$	Me	Me	1.699	1.728	2.376	
6	Trichlorfon	O	$-CH(OH)(CCl_3)$	Me	Me	1.731	1.691	2.263	
7[a]	Azinphos-methyl	S	$-CH_2\ (N_3C_7OH_4)$	Me	Me	1.843			1.828
8	Mevinphos	O	$-C(CH_3)(CHCOOCH_3)$	Me	Me	2.366	2.263	2.124	
9	Monocrotophos	O	$-C(CH_3)(CHCONHCH_3)$	Me	Me	2.366	2.324	2.215	
10	Demeton	S	$-CH_2PS_2(OCH_3)_2$	C_2H_5	C_2H_5	2.478	2.564	2.816	
11[a]	Dicrotophos	O	$-C(CH_3)(CHCON(CH_3)_2)$	Me	Me	2.564			2.533
12	Phorate	S	$-SCH_2SC_2H_5$	C_2H_5	C_2H_5	2.647	2.529	2.487	
13	Parathion-methyl	S	$-O(p\text{-}NO_2\text{-}Ph)$	Me	Me	2.718	2.736	2.775	
14	Chorfenvinphos	O	$-OCH(Cl)$ (p-Cl-	C_2H_5	C_2H_5	2.772	2.852	2.633	

			Ph)						
15[a]	Phenthoate	S	-SCH(Ph)(COO C_2H_5)	Me	Me	2.868			1.794
16	Disulfolton	S	-S($C_2H_4SC_2H_5$)	C_2H_5	C_2H_5	2.913	2.730	2.547	
17	Fenitrothion	S	-O(*m*-CH_3-*p*-NO_2-Ph)	Me	Me	2.937	2.925	2.903	
18	Terbufos	S	-$SCH_2SC(CH_3)$	C_2H_5	C_2H_5	2.991	2.844	2.672	
19[a]	Diazinon	S	-O(3-$CH(CH_3)_2$-5-Me-C_4HN_2)	C_2H_5	C_2H_5	3.126			3.249
20	Parathion	S	-O-*p*-NO_2-Ph	C_2H_5	C_2H_5	3.250	3.276	3.328	
21	Ronnel	S	-O-2,4,5-Cl-Ph	C_2H_5	C_2H_5	3.303	3.324	3.192	
22	Fenthion	S	3-Me-4-SMe-Ph-	Me	Me	3.370	3.316	3.002	
23[a]	Chlorpyrifos	S	-O-3,4,6-Cl-C_5HN	C_2H_5	C_2H_5	3.834			3.835
24	Ethion	S	-SCH_2-$PS_2(OCH_3)_2$	C_2H_5	C_2H_5	4.119	4.275	3.190	

[a] Test set compounds.

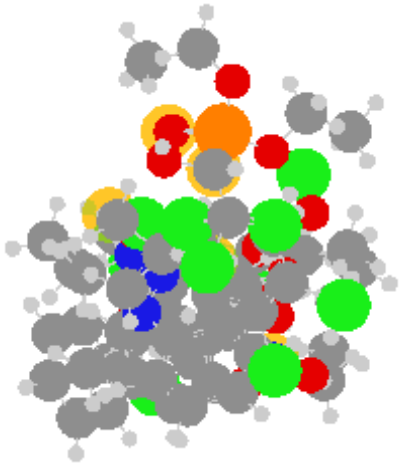


Figure 1. The superposed images for the 24 organophosphorus compounds. The common nucleus (the phosphorus atom) is used for 2D alignment, while the variable substructures explain the variance in the logK_{OC} values. P = orange (pixel number proportional to r_{vdW}/ε = 610), O = red (r_{vdW}/ε = 211), Cl = green (r_{vdW}/ε = 330), N = blue (r_{vdW}/ε = 307), S = yellow (r_{vdW}/ε = 508), C = gray (r_{vdW}/ε = 364), H = light gray (r_{vdW}/ε = 176).

RESULTS AND DISCUSSION

The predictive ability of the MIA-QSPR model can be evaluated on the basis of calibration and validation parameters (Table 2 and Figure 2), and considering that all compounds fell within the applicability domain, *i.e.* a theoretical region in chemical space encompassing both the model descriptors and modeled response. While the calibration step provided a good perspective for estimation and low risk of overfitting, the validation data

indicate that the predictions are reliable. These outcomes are promising, even considering the residual value of 1.074 for the test set compound Phentolate (**15**), which is not an outlier according to the applicability domain test.

Table 2. Statistical parameters used to attest the quality of the MIA-QSPR model.

Parameter	Value
LVs	4
RMSEC	0.098
r^2	0.982
$RMSE_{y\text{-}rand}$	0.349
$r^2_{y\text{-}rand}$	0.726
${}^{c}r^2_{P}$	0.501
RMSECV	0.403
q^2	0.775
RMSEP	0.477
r^2_{test}	0.656
r^2_{m}	0.564

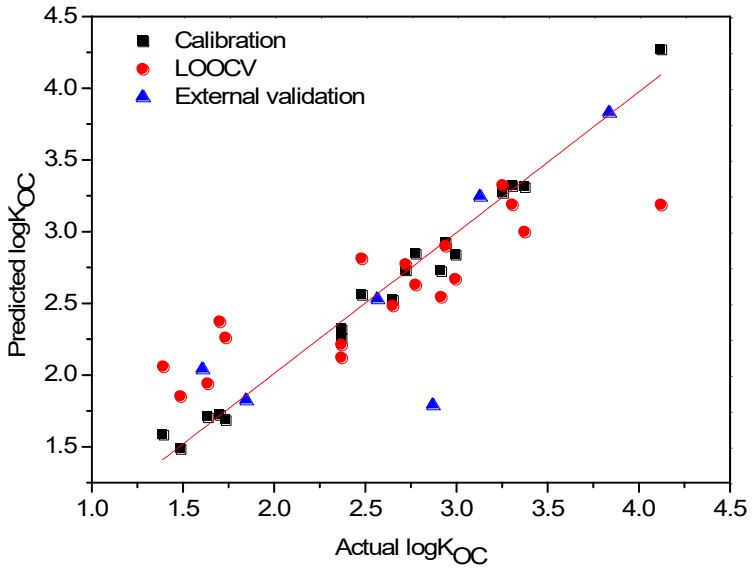


Figure 2. Plot of actual *vs.* predicted $logK_{OC}$ values for the 24 organophosphorus insecticides.

The modeling performance of the MIA descriptors (pixel values proportional to the ratio r_{vdW}/ε), indicates that $logK_{OC}$ for the organophosphorus insecticides indeed relates to hydrophobicity, since this parameter is controlled by the size and polarity of chemical substituents. However, a quick inspection of the correlation between $logK_{OC}$ and calculated $logK_{OW}$ values (obtained from the ACD/Labs Percepta platform) shows that hydrophobicity alone is not a sufficient parameter to describe completely the soil sorption of the

organophosphorus compounds, since the corresponding r^2 was only 0.44. Actually, hydrophobicity is dependent on the spatial arrangement of atoms/substituents and isomerism (even conformational),[19] while $\log K_{OW}$ estimation is based on group contributions, at most considering connectivity of fragments to characterize intramolecular hydrogen bonding and charge interactions. Thus, molecular images appear to be more suitable to describe accurately the variance in $\log K_{OC}$ as the chemical structure changes.

Since the regression method used to build the MIA-QSPR model was the partial least squares (PLS), the descriptors in the linear combination used to yield the modeled response are weighted by the regression coefficients **b**. When plotted, these PLS regression coefficients explain how descriptors affect the $\log K_{OC}$ values. In turn, the variable importance in projection (VIP) scores indicates the strength of this effect. Together, these MIA-contour maps (Figure 3) can provide deep insight on structural characteristics responsible for increasing or decreasing the $\log K_{OC}$ values.

From the VIP plot, the yellow to red regions are meaningful. The two terminal methyl groups of ethylthio groups in *e.g.* Ethion (**24**) have important influence to increase $\log K_{OC}$. This can also be observed in the **b** plot, as denoted by carbons with an orange color. Other regions marked with yellow to red colors in the VIP plot correspond to the sulfur, oxygen and phosphorus atoms of the phosphorodithioate moiety of Ethion. Since Ethion exhibits the largest $\log K_{OC}$ within this series of compounds, these regions are indicated as mostly affecting positively the $\log K_{OC}$ values. However, other groups also affect $\log K_{OC}$: green to electric blue regions in the VIP plot, also observed in the **b** plot as yellow or dark blue regions, should impact $\log K_{OC}$. Despite no significant difference between oxygen and sulfur bonded to phosphorus can be observed in both plots, the effect of a carbon bonded to P (such as in Trichlorfon - **6**), which is electric blue in the VIP plot and dark blue in the **b** plot, appears to be important to decrease $\log K_{OC}$. Chlorine atoms (at positions 2 and 4 in the phenyl ring of *e.g.* Chlorpyrifos - **23** and Ronnel - **21**) appear to be important to increase $\log K_{OC}$, given the electric blue (in the VIP plot) and yellow (in the **b** plot) colors in the MIA-contour maps. So, these regions should be taken into account when designing more environmentally friendly organophosphorus insecticides.

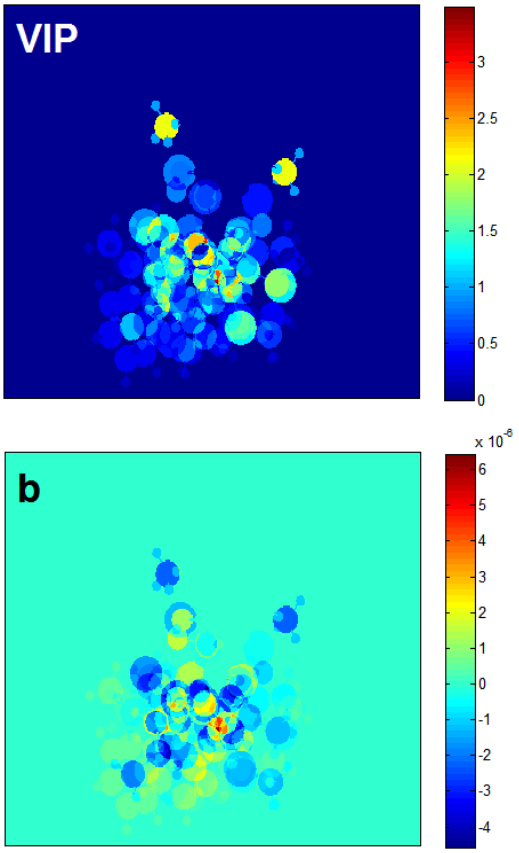


Figure 3. MIA-contour maps based on variable importance in projection (VIP) scores and PLS regression coefficients (**b**), used to analyze the descriptor contributions for the QSPR modeling of $\log K_{OC}$ values of the organophosphorus insecticides.

Pesticides with low $\log K_{OC}$ values are not always advantageous, either because their biological action mechanism requires permeation through lipophilic membranes of the living system or, from the environmental point of view, they have weaker affinity to the soil and, consequently, pesticides from the land's surface can move through the soil and end up in the groundwater. In this way, better than modeling pesticides with high soil sorption is to gain insight on new chemical entities capable of persisting in the soil during time enough for biodegradation. Most persistent organic pollutants lie on the organochlorine family of compounds, while organophosphorus pesticides can suffer rapid microbiological biodegradation.[20] Thus, efforts focused on modeling non-chlorinated organophosphorus insecticides containing some of the structural characteristics for high soil sorption discussed above, are in progress, as well as microbial assays to assess the biodegradation profile of the above compounds.

CONCLUSION

MIA descriptors encoding hydrophobicity and more were capable of modeling the logK_{OC} values of a series of organophosphorus insecticides with high degree of confidence. Consequently, this parameter can be predicted for congeneric compounds using the MIA-QSPR model built and, most importantly, key structural features were pointed out for possible chemical modifications to achieve more environmentally friendly insecticides. Such insights were driven by analysis of MIA-contour maps, which highlighted regions in the molecules most affecting, either increasing or decreasing, the soil sorption.

ACKNOWLEDGMENTS

Authors are thankful to FAPEMIG for the financial support of this research, as well as to CAPES for a studentship (to Joyce K. Daré) and to CNPq for the fellowships (to Cristina F. S. Batista and Matheus P. Freitas).

REFERENCES

1. Wen, Y.; Su, L. M.; Qin, W. C.; Fu, L.; He, J.; Zhao, Y. H. Linear and non-linear relationships between soil sorption and hydrophobicity: Model, validation and influencing factors. *Chemosphere* **2012**, *86*, 634-640.
2. Sabljic, A.; Güsten, H.; Verhaar, H.; Hermens, J. QSAR modelling of soil sorption. Improvements and systematics of log K_{OC} vs. log K_{OW} correlations. *Chemosphere* **1995**, *31*, 4489-4514.
3. Freitas, M. R.; Barigye, S. J.; Daré, J. K.; Freitas, M. P. Aug-MIA-SPR/PLS-DA classification of carbonyl herbicides according to levels of soil sorption. *Geoderma* **2016**, *268*, 1-6.
4. Uchimiya, M.; Wartelle, L. H.; Boddu, V. M. Sorption of triazine and organophosphorus pesticides on soil and biochar. *J. Agric. Food Chem.* **2012**, *60*, 2989-2997.
5. Barigye, S. J.; Freitas, M. P. Ten years of the MIA-QSAR strategy: Historical development and applications. *Int. J. Quantit. Struct. Prop. Relat.* **2016**, *1*, 64-77.
6. Freitas, M. P.; Brown, S. D.; Martins, J. A. MIA-QSAR: a simple 2D image-based approach for quantitative structure–activity relationship analysis. *J. Mol. Struct.* **2005**, *738*, 149-154.
7. Da Mota, E. G.; Duarte, M. H.; Barigye, S. J.; Ramalho, T. C.; Freitas, M. P. Exploring MIA-QSPR's for the modeling of biomagnification factors of aromatic organochlorine pollutants. *Ecotox. Environ. Saf.* **2017**, *135*, 130-136.

8. Freitas, M. R.; Barigye, S. J.; Freitas, M. P. Coloured chemical image-based models for the prediction of soil sorption of herbicides. *RSC Adv.* **2015**, *5*, 7547-7553.

9. Nunes, C. A.; Freitas, M. P. Introducing new dimensions in MIA-QSAR: A case for chemokine receptor inhibitors. *Eur. J. Med. Chem.* **2013**, *62*, 297-300.

10. Barigye, S. J.; Duarte, M. H.; Nunes, C. A.; Freitas, M. P. MIA-Plot: A graphical tool for viewing descriptor contributions in MIA-QSAR. *RSC Adv.* **2016**, *6*, 49604-49612.

11. Mackay, D.; Shiu, W.-Y.; Ma, K.-C. *Illustrated Handbook of Physical–Chemical Properties and Environmental Fate for Organic Chemicals*; Lewis Publishers: New York, 1997.

12. Dennington, R. D., II; Keith, T. A.; Millam, J. M. GaussView 5.0.8; Gaussian, Inc.: Wallingford, U.K., 2008.

13. Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inf.* **2010**, *29*, 476-488.

14. Roy, K.; Chakraborty, P.; Mitra, I.; Ojha, P. K.; Kar, S.; Das, R. N. Some case studies on application of "r^2_m" metrics for judging quality of quantitative structure-activity relationship predictions: Emphasis on scaling of response data. *J. Comput. Chem.* **2013**, *34*, 1071-1082.

15. Mitra, I.; Saha, A.; Roy, K. Exploring quantitative structure–activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants. *Mol. Simul.* **2010**, *36*, 1067-1079.

16. Roy, K.; Kar, S.; Ambure, P. On a simple approach for determining applicability domain of QSAR models. *Chemom. Intell. Lab. Sys.* **2015**, *145*, 22-29.

17. Roy, K.; Ambure, P.; Aher, R. B. How important is to detect systematic error in predictions and understand statistical applicability domain of QSAR models? *Chemom. Intell. Lab. Sys.* **2017**, *162*, 44-54.

18. Nunes, C. A.; Freitas, M. P.; Pinheiro, A. C. M.; Bastos, S. C. Chemoface: A novel free user-friendly interface for chemometrics. *J. Braz. Chem. Soc.* **2012**, *23*, 2003-2010.

19. O'Hagan, D.; Young, R. J. Accurate lipophilicity (log*P*) measurements inform on subtle stereoelectronic effects in fluorine chemistry. *Angew. Chem. Int. Ed.* **2016**, *55*, 3858-3860.

20. Deng, S.; Chen, Y.; Wang, D.; Shi, T.; Wu, X.; Ma, X.; Li, X.; Hua, R.; Tang, X.; Li, Q. X. Rapid biodegradation of organophosphorus pesticides by *Stenotrophomonas* sp. G1. *J. Hazard. Mat.* **2015**, *297*, 17-24.

ARTIGO 2

Journal: Chemical Biology and Drug Design (2019)

3D perspective into MIA-QSAR: A case for anti-HCV agents

Joyce K. Daré,[a] Teodorico C. Ramalho,[a] Matheus P. Freitas[a,*]

[a] *Department of Chemistry, Federal University of Lavras, 37200-000, Lavras, MG, Brazil*

* corresponding author: matheus@dqi.ufla.br

Running Head: Conformation in MIA-QSAR

ABSTRACT

Quantitative structure-activity relationship (QSAR) is a molecular modeling technique widely used in the discovery of novel drugs. Currently, there are many approaches for performing such analysis, which are commonly classified from 1D to 6D. 2D and 3D techniques are amongst the most exploited ones. Multivariate image analysis applied to QSAR (MIA-QSAR) is an example of 2D methodology that has presented a satisfactory performance in the generation of effective prediction models for biological/physicochemical properties. However, once this is a 2D method, conformational information is not explicitly considered, despite the well-known role of such type of information in explaining the biochemical behavior. Thus, the importance of conformation is undeniable, but the requirement of this information for QSAR analysis still needs to be studied. Therefore, this work aimed to provide a method for encoding 3D information into MIA-QSAR descriptors and analyze the consequences of this inclusion on this methodology. The strategy consisted in fully optimizing the molecular geometries of anti-HCV compounds and three-dimensionally align them before performing the MIA-QSAR procedure. As a result, it was possible to verify that this type of information does not improve the MIA-QSAR modeling performance; instead, the traditional procedure consisting of maximally congruent substructures generated a more reliable prediction model.

Keywords: alignment, conformation, geometry optimization, MIA-plots, MIA-QSAR.

INTRODUCTION

From a general view, computational chemistry has completely changed the classical process of screening and identifying novel drugs. Among the various modifications, the shortening in the duration of this usually long process, as a result of the replacement of a substantial part of an experimental analysis by computational simulations, can be highlighted, which results in saving money and time (Benfenati, 2012).

Accordingly, molecular modeling has gained the attention of researchers from all over the globe, which resulted in the development of a variety of computational techniques and software for aiding in the accomplishment of the given task. Among the various techniques, QSAR (quantitative structure–activity relationship) has been massively employed. QSAR can be understood as a quantitative methodology that aims to correlate the compounds' chemical structural features to a property of interest (bioactivity) through the generation of a mathematical model (regression equation) (Hansch, Leo, & Hoekman, 1995). In other words, this equation can predict the biological property under analysis for compounds with structural similarities that have not been studied yet (Dearden, 2016).

There are different approaches when performing a QSAR analysis. Commonly, these methods are classified into six categories (from 1D to 6D) according to their mode of obtaining molecular descriptors, as well as to the type of descriptors considered during the modeling step (Dearden, 2016). Since molecular descriptors are crucial for any QSAR analysis, it is coherent to define them. A molecular descriptor can be understood as a variable (either continuous or discrete) that describes/represents relevant structural properties of compounds and is employed as an independent variable in the construction of the QSAR model (Benfenati, 2012; Cherkasov et al., 2014).

Each approach mentioned above presents either advantages or drawbacks (Consonni & Todeschini, 2009). Among these methodologies, one of the most popular is 2D QSAR, which, in its essence, considers topological representations of molecules and explains the modeled property in terms of the presence and nature of chemical bonds (Consonni & Todeschini, 2009). The main advantages of this approach are (a) the parameters (descriptors) contain simple and efficient information of the molecular structure, (b) they are also invariant to molecule roto-translations, and, finally, (c) these parameters can be obtained avoiding geometry optimization (Consonni & Todeschini, 2009). On the other hand, the 2D models, although reliable, are not easily interpretable in a meaningful way, that is, it is not always

possible to find a biochemical/physicochemical explanation for the propriety under analysis based on the QSAR model (de Freitas & Ramalho, 2013).

3D QSAR methodologies, in turn, are as popular as 2D QSAR. The foundations of this approach lie on the idea that biological selectivity results from each target forming highly specific interactions, such as hydrogen bonds with a ligand and that the binding preferences emerge primarily from non-covalent field effects exerted in its spatial vicinity (Dare, Silva, & Freitas, 2017). In other words, 3D-QSAR is a broad term that covers all those QSAR techniques based on computed descriptors derived from spatial representation of the molecular structures (Verma, Khedkar, & Coutinho, 2010). Among the many techniques, CoMFA can be highlighted as one of the most important. It was the first 3D-QSAR methodology to be developed, and it still has great importance in QSAR studies (Cramer & Wendt, 2007). Indeed, CoMFA is still widely used, but, nowadays, it is mainly employed in association with CoMSIA or another enhanced 3D technique (Halder, Amin, Jha, & Gayen, 2017).

The advantages of 3D-QSAR are that models can be interpreted in terms of chemical regions in a 3D space most affecting the biological property. However, the model development is complex due to the exhaustive conformational search and structure alignment required in such analysis (Dare et al., 2017). Among the various methods of conformational searching, it is remarkable for the systematic search, random search, Monte Carlo, molecular dynamics, simulated annealing, distance geometry algorithm, and genetic and evolutionary algorithm (Verma et al., 2010). These methods, although efficient in their tasks, present a high level of complications, which contributes to the 3D-QSAR complexity (Verma et al., 2010). Furthermore, 3D methods usually employ conformational screening and structural alignment rules that do not consider the biological target itself, which can lead to incoherent conclusions (de Freitas & Ramalho, 2013).

Facing the advantages and drawbacks of 2D and 3D approaches, a method capable of balancing efficiency, interpretability, and simplicity would be very desirable. In this sense, a unique and efficient technique, named MIA-QSAR (multivariate image analysis applied to QSAR), was developed in 2005. This method is based on the treatment of bidimensional images to generate descriptors for molecular modeling. The descriptors, in this case, are pixels (binary values defined as the smallest unit in a digital image), once this type of data constitutes a digital image (Freitas, Brown, & Martins, 2005).

MIA-QSAR foundations lie on the idea that, for a congeneric series of compounds, changes in the substituents generate variations in pixel positions, which is useful in explaining the observed differences at the dependent variable pattern (studied property) (Guimaraes, Silla, da Cunha, Ramalho, & Freitas, 2016). In other words, the differences observed in the property, among a congeneric series of compounds, are attributed to the different chemical groups that these compounds have.

Basically, in the current version of MIA-QSAR, the atoms are represented by circles with sizes proportional to the respective van der Waals radii and their colors (pixel values) can be proportional to electronegativity or other atomic property. The color system employed is the RGB (24 *bits*); therefore, the pixels can assume values that varies from "0" (black color) to 765 (white color), since each pixel in this system is a result of the combination of three color channels (red, green, and blue) and each of these channels assumes values that vary from 0 to 255. These considerations have shown satisfactory outcomes, often comparable to those results obtained using 3D methodologies (Dare, Barigye, & Freitas, 2017; de Freitas & Ramalho, 2013; Freitas et al., 2005; Guimaraes et al., 2016). Moreover, MIA-QSAR also includes a graphical tool denominated MIA-plot that allows the user to obtain a detailed chemical interpretation of the regression models (Barigye, Duarte, Nunes, & Freitas, 2016; Borges & Barigye, 2017).

Traditionally, a bidimensional technique does not consider the spatial features of compounds, once it focuses, normally, on the presence and nature of chemical bonds. However, it is possible to include 3D information into the 2D descriptors as proved by Estrada and colleagues (2001), but the implications of this inclusion remain unclear (Estrada, Molina, & Perdomo-Lopez, 2001). Therefore, there is still no evidence that the inclusion of spatial information into bidimensional descriptors improves the results obtained in 2D-QSAR analysis (Guimaraes, Duarte, Silla, & Freitas, 2016).

Accordingly, this work aimed to provide a method to include stereochemical information into MIA-QSAR descriptors and to analyze the implications of such modification compared to the traditional analysis, which is based on non-optimized, flat-shape chemical representations. Specifically, this study targeted on (a) finding a method to align 3D chemical structures and generate 2D projections from these images to be used in the MIA-QSAR protocol and (b) to probe the benefits (or not) of using compound geometries optimized in an enzyme-free environment in MIA-QSAR modeling. The molecular data set selected for this study consists of 42 compounds with anti-HCV (hepatitis C virus) activity, synthesized and

characterized by Konreddy et al. (2014). Hepatitis C virus is a single-stranded RNA virus that belongs to the *Flaviviridae* family. It is known for infecting the hepatocytes, which in long term may lead to very serious liver diseases, such as fibrosis and cirrhosis followed by hepatocellular carcinoma (Konreddy et al., 2014).

MATERIALS AND METHODS

The data set comprised of 42 4-hydroxyamino α-pyranone carboxamide analogs, as well as the respective biological data described in terms of pEC_{50} (EC_{50} in mol/L), was obtained from the literature (Konreddy et al., 2014) and is shown in supporting Information Table S1. Initially, for purposes of comparison, a traditional MIA-QSAR model was built based on the following steps.

The RGB images were obtained using GaussView program (Dennington, Keith, & Millam, 2008). The options "Spotlight" were unselected in the preferences tab, and the option "Use van der Waals radii" was selected. These modifications generate the 2D molecule images used in MIA-QSAR models. Each image was generated maintaining the molecular congeneric center at the same position in all compounds. Then, these images were superposed as shown in Appendix A: Figure 1a.

Using Chemoface (Nunes, Freitas, Pinheiro, & Bastos, 2012), a free software for chemometrics, the three-way array, generated by the superimposition of the images, was unfolded into a data matrix with dimensions proportional to the area of the images. In this specific case, the compound images had a length of 300 pixels per 354 in width, which generated an unfolded matrix of 42 rows per 1 06 200 columns. The next step consisted in excluding the columns corresponding to the common moieties of the compounds, that is, the columns with null variance.

The pixel values were, then, replaced by numbers proportional to (a) the respective values of rvdW/ε (van der Waals radii/Pauling's electronegativity), (b) the van der Waals radii values, and (c) the respective electronegativity values. In all cases, white pixels (765) were replaced by "0", which corresponds to the black color. For the first model (values proportional to rvdW/ε), the employed pixel values were as follows: O = red (229) replaced by 211 ($\propto$ rvdW/ε); Cl = green (289) replaced by 330; N = blue (279) substituted for 307; F = ciano (688) replaced by 178; C = gray (426) substituted for 364; H = light gray (612) replaced by 176; and finally Br = dark red (231) was replaced by 373. For the second model (values

proportional to rvdW), O = 730, Cl = 990, N = 940, F = 710, C = 930, H = 370, and Br = 1140. Finally, for the model with pixel values proportional to electronegativity, O = 344, Cl = 316, N = 304, F = 400, C = 255, H = 220, and Br = 296. Bonds were valued as 0.1.

For each generated matrix, a PLS (partial least squares) model was built; in all situations, the data set was split into training (75%) and test (25%) sets, in order to proceed with calibration and external validation, respectively. The chosen test set was the same selected by Li et al. (2016), who also performed a QSAR analysis and is going to have their results used as a source for comparison herein; the same test set was chosen to guarantee the same domain space with the aim of comparing the results obtained herein with those obtained by them. The optimum number of latent variables was chosen by analyzing the decay of the root mean square error (RMSE) in the leave-one-out cross-validation (LOOCV). The validation parameters considered were the regression coefficients from the calibration (r^2), cross-validation (q^2), and external validation (r^2_{test}), as well as their respective RMSEs (root mean square errors).

In addition, the proximity between the actual and predicted pEC50 values for the test set was statistically evaluated using the r^2_m parameter {$r^2_m = r \times [1 - (r^2 - r_0^2)^{1/2}]$, where r^2 and r_0^2 are the determination coefficient values between the observed and predicted pEC_{50} data with and without the intercept} (Roy et al., 2013). The risk of chance correlation in calibration was analyzed using the $^c r^2_p$ parameter [$^c r^2_P = r \times (r^2 - r^2_{y\text{-random}})^{1/2}$, where $r^2_{y\text{-random}}$ corresponds to the mean determination coefficient value obtained after randomizing the y block ten times] (Mitra, Saha, & Roy, 2010). Finally, b and VIP plots (MIA contours maps based on PLS regression coefficients and variable importance in projection to account for descriptor weights for the model) (Barigye et al., 2016; Borges & Barigye, 2017) were generated for the third model (pixel values proportional to electronegativity), once this was asserted as the best model among all.

Regarding the inclusion of 3D information into the MIA-QSAR descriptors, as a starting point, a conformational screening was performed using a Monte Carlo stochastic distribution through the Spartan'16 programa for all 42 compounds (level of theory: semi-empirical AM1 method (Dewar, Zoebisch, Healy, & Stewart, 1985). The lowest energy conformation was selected for each case and subsequently optimized through the Gaussian 09 program (Frisch et al., 2013), using density functional theory (DFT) at the ωB97x-D/ 6-31G(d,p) level of theory (Chai & Head-Gordon, 2008; Krishnan, Binkley, & Seeger, 1980).

Next, the optimized structures were loaded on the Discovery Studio Visualizer program,b one at a time and the superposition of each molecule was performed using the Tether tool available on the manual alignment tab. This superposition generates the three-way array, and it was performed having as reference the congeneric center. Then, the superposed molecules were loaded in GaussView program (Dennington et al., 2008) and saved with the same settings and format of the molecules included in the traditional modeling step (Appendix A: Figure 1b). Employing this strategy, conformational information was taken into account. Next, the MIA-QSAR modeling was carried out following the same steps as the traditional technique, including the validation parameters. Finally, the models were compared, and the results are displayed and discussed in the next section.

A schematic representation summarizing the whole methodology is presented in Supporting Information Figure S1 with the aim of facilitating the comprehension of the traditional (2D) and 3D MIA-QSAR technique.

In addition to the analysis performed above, a double cross-validation was carried out to evaluate whether different training sets could result in better outputs, as proposed by Roy & Ambure (2016). The best model for both the traditional and 3D-MIA- QSAR is shown in Supporting Information. It is worth mentioning that the genetic algorithm was chosen as the variable selection method.

RESULTS AND DISCUSSION

Initially, it is important mentioning that the compound set used in this analysis does not have a well-defined biological target and neither an action mechanism completely figured out (Konreddy et al., 2014). Furthermore, the group responsible for the synthesis of these molecules performed the anti-HCV assays in cells harboring subgenomic HCV RNA replicons (genotype 1b) with a luciferase reporter (LucNeo#2) instead of direct measurements on a specific receptor (Konreddy et al., 2014). Therefore, conclusions relating the 3D information encoded during the QSPR analysis performed herein and the observed values of the property (pEC_{50}) must be thought carefully, once the chemical interactions are not clear in this case. However, a general structural analysis can be performed to try finding a reasonable explanation covering the observed pEC_{50} values.

The validation parameters for the three models with (a) pixel values replaced by values proportional to r_{vdW}/ε, (b) pixel values replaced by values proportional to van der Waals radii, and (c) pixel values replaced by values proportional to electronegativity are

organized in Appendix A: Table 1. Accordingly, all the models have been found to be statistically satisfactory, once their internal and external parameters are within the acceptable ranges established elsewhere (Dearden, Cronin, Kaiser, & Environ, 2009; Tropsha, 2010). Furthermore, their results are quite similar, in such a way that they could be considered equivalent. However, focusing on the optimal number of latent variables, the third model seems to be more parsimonious than the others, once it uses only three latent variables for explaining variability, which characterizes a more consistent and suitable model when compared with higher values of LVs as those from models 1 and 2. Therefore, the model in question was chosen for further analysis (construction and interpretation of MIA-plots). It is also worth mentioning that the result obtained with model 3 is comparable to those obtained through 3D-QSAR by Li et al. (2016) (see footnote in Supporting Information Table S1). The predicted and measured values of pEC_{50} are summarized in Appendix A: Figure 2, while the individual values of the predicted response variable are shown in Supporting Information Table S1.

Regarding the optimized molecules, that is, the second part of this study, the resulting superposed molecules are shown in Appendix A: Figure 1b. From this figure, one can observe that the superposition of the symbolic molecules presented more variability than those images obtained through the traditional analysis, as a result of the allowed conformational flexibility. The validation parameters for the three corresponding models are displayed in Appendix A: Table 2.

Although the calibration parameters are acceptable, the y-randomization results do not eliminate the chance correlation in any case. On the contrary, the ${}^{c}r^{2}_{p}$ parameter shows that all models are overfitted, despite the appropriate results for external validation in the last model. Furthermore, the internal validation was not satisfactory in all cases. Therefore, the inclusion of 3D information into MIA-QSAR technique, isolated from the active site, presented worse results than the traditional analysis, and, consequently, this type of information can be considered unnecessary for this 2D technique.

A successful MIA-QSAR model requires pattern recognition provided by a perfect alignment of pharmacophoric substructures and substituents placed at specific positions in a molecule, which is, in fact, a general QSAR foundation. However, the crowded image originated from the superposition of the optimized structures, as a result of conformational freedom, disfavors the above pattern requirement and impairs the calibration step. Despite the well-known role of conformation on biological effects, it follows that connectivity of atoms,

as well as atomic properties (e.g., van der Waals radii and electronegativity), is more encoding molecular descriptors in MIA-QSAR than conformational fingerprints.

Once the inclusion of conformational features presented itself as an unnecessary step in MIA-QSAR procedure, the MIA-plots were built based on the best model obtained through traditional analysis, that is, the third model in Appendix A: Table 1. VIP and b plots are shown in Appendix A: Figure 3. The applicability domain (William's plot) for the chosen model was also checked and is shown in Supporting Information Figure S2.

The PLS regression coefficients (b) characterize how descriptors affect the response variable values. In turn, the variable importance in projection (VIP) scores indicates the strength of this effect. Therefore, a joint analysis of b and VIP plots is necessary to avoid misinterpretation.

Initially, it is important observing that the pEC50 interval is small (the pEC_{50} values vary from 5.398 to 6.770), which complicates the task of finding a pattern of the different substituent influences.

Starting from a general point of view, one can observe that substituent 1 (R1) is more important to explain pEC_{50} than the second substituent (R2), once its contribution in the VIP plot is higher (varying from 1 to 3.25) than the second moiety (varying from 0.75 to 2). Focusing on the first portion (R1), the hydrogen atom (cyan color on the b plot) seems to have a strong negative contribution to the response variable. However, when comparing this observation with the actual values of pEC_{50} in Table S1, it is observed an exactly opposite pattern, that is, molecules with H in the R1 position tends to have higher values of pEC_{50} (>6.000). This happens when in the math equation (calibration model), the coefficient attributed to a specific descriptor has a negative sign. Therefore, hydrogen atoms increase the value of pEC_{50}, which is a desirable effect. There are two exceptions to this generalization, compounds **38** and **44**, which can be attributed to the R2 substituents. Compound **38** has the group C_2H_4OH in that position, which has its oxygen in cyan color and the H atom in green, then, this substituent seems to decrease the pEC_{50} value, confirming the previous observation. Compound **44** has a CH2COOMe group, in which the COO portion is cyan, that is, this group contributes negatively to the property. Therefore, the previous generalization regarding the positive contribution of H in R1 still stands. It is also worth mentioning that the compound with the highest pEC_{50} value (compound 21) contains a H atom in the R1 position.

On the other hand, fluorine (the most electronegative element) presents a dark blue color, which indicates a negative contribution toward pEC_{50} and, analyzing the actual values,

this hypothesis is confirmed; all molecules with a F atom bearing the R1 position exhibit a low value of pEC_{50}, independent of the substituent nature at R2.

The substituent OCH_3 presents an orange color, which indicates a positive contribution to the property. Observing the actual values of pEC_{50}, three of the six compounds with this type of substituent exhibit high values of pEC_{50}, while three others have lower results. This means that the influence of R2 is probably counterbalancing the contribution of OCH_3 in increasing the property, which is confirmed when analyzing these three compounds. Compound **29** has a *p*-methylphenyl substituent and, although the phenyl portion seems to have no relevant influence in the property, the *para* methyl group presents a light green color, indicating a decrease in the property value. Compound **42** has the same R2 group of compound **38**. Finally, compound **48** has the same R2 group of compound **44**. On another hand, analyzing the three other compounds with high values of pEC_{50}, it follows that a non-substituted phenyl does not affect the property, as well as a methyl group alone, that is, the effects observed in these cases come exclusively from the group OCH_3 in the R1 position.

Compounds containing chlorine atoms in the R1 position have, in general, small pEC_{50} values. However, the maintenance of the same R2 substituent and replacement of the Cl with F atom in the R1 position yield a decrease in the pEC_{50} values. This may be due to the smaller electronegativity of chlorine when compared to fluorine. There are only two exceptions to this pattern, compounds **40**—which has the smallest pEC_{50} value of all—and **46**. This atypical behavior might be related to some synergistic effect of both substituents or to a different effect not explained by the MIA-plot graphs.

It can be seen that the pEC_{50} values are highly dependent of the R2 substituents, which indicates that a methyl group is not significantly relevant to the model.

Regarding specifically the R2 substituents, the monosubstituted phenyl group presents a positive contribution to the property in most of the cases. But the same does not occur for phenyl group alone; in this occasion, the property seems to be more dependable on the R1 moiety than R2. On the other hand, alcohol and ester residues have a strong negative contribution to the property, which is confirmed by the light green color of all the oxygen atoms of these chemical functions in the b plot.

Therefore, based on all information collected from the MIA-plots, as discussed above, it was possible to propose a general structure (Appendix A: Figure 4) for a compound that may present a good value of bioactivity. This proposal took into account the observations that compounds containing H atom at R1 and a monosubstituted phenyl group at R2 showed the

highest pEC_{50} values. In this sense, we proposed a disubstituted phenyl ring, containing groups that were already modeled at those positions and presented good results.

Then, the best MIA-QSAR model was applied to predict the bioactivity of the proposed compound and a pEC_{50} value of 6.669 was subsequently obtained, which is close to the highest value (6.770) of the analyzed data set.

Regarding the double cross-validation technique, the results obtained are shown in Supporting Information Table S2. Even the best models were not superior than the previous results. This observation might be related to the fact that the PLS models were built using only a few number of descriptors (three to be more accurate), once the chosen technique requires the user to choose a variable selection method (stepwise or genetic algorithm). In this way, most of the structural molecular information is excluded once the descriptors are the pixels themselves. Based on these findings, and on the fact that this study aimed to compare the results obtained through MIA-QSAR technique to those obtained by Li et al. (2016), the original training set (Table S1) was kept.

CONCLUSIONS

The geometry optimization and the alignment step to generate 2D image projections from 3D chemical structures were successful, but the model obtained from flat-shape chemical structures provided a better QSAR model. A tentative explanation for this behavior is that image superposition of the flexible structures does not allow a perfectly aligned perspective of the congeneric moieties; it follows that connectivity of atoms, as well as atomic properties (e.g., van der Waals radii and electronegativity), is still more encoding molecular descriptors in MIA-QSAR than conformational fingerprints. In the specific case of anti-HCV agents, the effects of incorporating three-dimensional information into MIA-QSAR appear to be detrimental. However, this observation may also be related to the fact that such a stereochemical information was obtained from protein-free structures; thus, a deeper study undertaking the enzyme target is open. On another hand, the results obtained with the regular technique application were comparable to those obtained employing a 3D method, that is, regular MIA-QSAR generated a reliable predictive model capable of explaining the pattern observed in the studied property. Lastly, the MIA-plots were also very helpful in the chemical interpretation of the observed pattern for the pEC50 variable.

ACKNOWLEDGEMENTS

Authors are thankful to FAPEMIG for the financial support of this research (grant numbers: CEX-APQ-00383-15 and PPM-00344-17), as well as to CAPES for a studentship (to J.K.D.), and to CNPq for fellowships (to T.C.R. and M.P.F.).

CONFLICT OF INTEREST

The authors declare no conflict of interest

END NOTES

[a] Spartan'16 Software, *Wavefunction Inc.*, Irvine, 2017.

[b] Dassault Systemes BIOVIA -Discovery Studio Visualizer 2017R2 2017.

ORCID

Matheus P. Freitas https://orcid.org/0000-0002-7492-1801

REFERENCES

Barigye, S. J., Duarte, M. H., Nunes, C. A., & Freitas, M. P. (2016). MIA-plot: A graphical tool for viewing descriptor contributions in MIA-QSAR. RSC Advances, 6, 49604–49612. https://doi.org/10.1039/C6RA09593C

Benfenati, E. (**2012**). E-book: Theory, guidance and applications on QSAR and REACH. Milan, Italy: Istituto di Ricerche Farmacologiche Mario Negri.

Borges, F. M. C. N., & Barigye, S. J. (**2017**). Towards molecular design using 2D-molecular contour maps obtained from PLS regression coefficients. *Molecular Physics*, *115*, 3044–3050. https://doi.org/10.1080/00268976.2017.1347294

Chai, J.-D., & Head-Gordon, M. (**2008**). Systematic optimization of long-range corrected hybrid density functionals. *Journal of Chemical Physics*, *128*, 084106.

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., … Tropsha, A. (**2014**). QSAR modeling: Where have you been? Where are you going to? *Journal of Medicinal Chemistry*, *57*, 4977–5010. https://doi.org/10.1021/jm4004285

Consonni, V., & Todeschini, R. (**2009**). *Molecular descriptors for chemoinformatics*. Weinheim, Germany: Wiley-VCH, Verlag. Cramer, R., & Wendt, B. (**2007**). Pushing the boundaries of 3D-QSAR. *Journal of Computer-Aided Molecular Design*, *21*, 23–32. https://doi.org/10.1007/s10822-006-9100-0

Dare, J. K., Barigye, S. J., & Freitas, M. P. (**2017**). Multi-objective modeling of herbicidal activity from an environmentally friendly perspective. *International Journal of Quantitative Structure-Property Relationships*, *2*, 16–26.

Dare, J. K., Silva, C. F., & Freitas, M. P. (**2017**). Revealing chemophoric sites in organophosphorus insecticides through the MIA-QSPR modeling of soil sorption data. *Ecotoxicology and Environmental Safety*, *144*, 560–563. https://doi.org/10.1016/j.ecoenv.2017.06.072

Dearden, J., (**2016**). The history and development of quantitative structure-activity relationships (QSARs). *International Journal of Quantitative Structure-Property Relationships*, *1*, 1–44.

Dearden, J. C., Cronin, M. T. D., Kaiser, K. L. E., & Environ, S. A. R. Q. S. A. R. (**2009**). How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *SAR and QSAR in Environmental Research*, *20*, 241–266.

Dennington, R. D., Keith, T. A., & Millam, M. J. (**2008**). *GaussView 5.0*, Wallingford, CT.

Dewar, M. J. S., Zoebisch, E. G., Healy, E. F., & Stewart, J. J. P. (**1985**). Development and use of quantum mechanical molecular models. 76. AM1: A new general purpose quantum mechanical molecular model. *Journal of the American Chemical Society*, *107*, 3902–3909. https://doi.org/10.1021/ja00299a024

Estrada, E., Molina, E., & Perdomo-Lopez, I. (**2001**). Can 3D structural parameters be predicted from 2D (topological) molecular descriptors? *Journal of Chemical Information and Computer Sciences*, *41*, 1015–1021. https://doi.org/10.1021/ci000170v

Freitas, M. P., Brown, S. D., & Martins, J. A. (**2005**). MIA-QSAR: A simple 2D image-based approach for quantitative structure–activity relationship analysis. *Journal of Molecular Structure*, *738*, 149–154. https://doi.org/10.1016/j.molstruc.2004.11.065

de Freitas, M. P., & Ramalho, T. C. (**2013**). Employing conformational analysis in the molecular modeling of agrochemicals: Insights on QSAR parameters of 2,4-D. *Ciencia e Agrotecnologia*. *37*, 485–494. https://doi.org/10.1590/S1413-70542013000600001

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., … Fox, D. J. (**2013**). *Gaussian 09, Revision D01*. Wallingford, CT: Gaussian, Inc.

Guimaraes, M. C., Duarte, M. H., Silla, J. M., & Freitas, M. P. (**2016**). Is conformation a fundamental descriptor in QSAR? A case for halogenated anesthetics. *Beilstein Journal of Organic Chemistry*, *12*, 760–768. https://doi.org/10.3762/bjoc.12.76

Guimaraes, M. C., Silla, J. M., da Cunha, E. F. F., Ramalho, T. C., & Freitas, M. P. (**2016**). Is the bioconformation of 5-deoxy-5-fluoro-D-xylulose affected by intramolecular hydrogen bonds? *RSC Advances*, *6*, 111681–111687. https://doi.org/10.1039/C6RA23423B

Halder, A., Amin, S., Jha, T., & Gayen, S. (**2017**). Insight into the structural requirements of pyrimidine-based phosphodiesterase 10A (PDE10A) inhibitors by multiple validated 3D QSAR approaches. *SAR and QSAR in Environmental Research*, *28*(3), 253–273. https://doi.org/10.1080/1062936X.2017.1302991

Hansch, C., Leo, A., & Hoekman, D. H. (**1995**). *Exploring QSAR: Hydrophobic, electronic, and steric constants*. Washington, DC: ACS Publisher.

Konreddy, A. K., Toyama, M., Ito, W., Bal, C., Baba, M., Sharon, A., & Med, A. C. S. (**2014**). Synthesis and anti-HCV activity of 4-Hydroxyamino α-Pyranone Carboxamide analogues. *Chemistry Letters*, *5*, 259–263.

Krishnan, P. J. R, Binkley, J. S., & Seeger, R. (**1980**). Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *Journal of Chemical Physics*. *72*(1), 650–654. https://doi.org/10.1063/1.438955

Li, W., Xiao, F., Zhou, M., Jiang, X., Liu, J., Si, H., … Zhai, H. (**2016**). 3D-QSAR study and design of 4-hydroxyamino α-pyranone carboxamide analogues as potential anti-HCV agents. *Chemical Physics Letters*, *661*, 36–41. https://doi.org/10.1016/j.cplett.2016.08.042

Mitra, I., Saha, A., & Roy, K. (**2010**). Exploring quantitative structure–activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants. *Molecular Simulation*, *36*, 1067–1079. https://doi.org/10.1080/08927022.2010.503326

Nunes, C. A., Freitas, M. P., Pinheiro, A. C. M., & Bastos, S. C. (**2012**). Chemoface: A novel free user-friendly interface for chemometrics. *Journal of the Brazilian Chemical Society*, *23*, 2003–2010.

Roy, K., & Ambure, P. (**2016**). The "double cross-validation" software tool for MLR QSAR model development. *Chemometrics and Intelligent Laboratory Systems*, *159*, 108–126. https://doi.org/10.1016/j.chemolab.2016.10.009

Roy, K., Chakraborty, P., Mitra, I., Ojha, P. K., Kar, S., & Das, R. N. (**2013**). Some case studies on application of "r(m)2" metrics for judging quality of quantitative structure-activity relationship predictions: Emphasis on scaling of response data. *Journal of Computational Chemistry*, *34*, 1071–1082. https://doi.org/10.1002/jcc.23231

Tropsha, A. (**2010**). Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*, *29*, 476–488. https://doi.org/10.1002/minf.201000061

Verma, J., Khedkar, V., & Coutinho, E. (**2010**). 3D-QSAR in drug design—a review. *Current Topics in Medicinal Chemistry*, *10*, 95–115. https://doi.org/10.2174/156802610790232260

How to cite this article: Dare JK, Ramalho TC, Freitas MP. 3D perspective into MIA-QSAR:

A case for anti-HCV agents. *Chem Biol Drug Des*. 2018; 00:1–9. https://doi.org/10.1111/cbdd.13440

APPENDIX A

TABLE 1 Validation parameters for MIA-QSAR models with pixel values proportional to rvdW/ε, rvdW, and ε, for the traditional analysis

Parameters	Model 1 (LV=11) ($\propto r_{vdW}/\varepsilon$)	Model 2 (LV=8) ($\propto r_{vdW}$)	Model 3 (LV=3) ($\propto \varepsilon$)
$RMSE_{cal}$	0.1104	0.1178	0.1561
r^2_{cal}	0.9043	0.8910	0.8085
$RMSE_{y\text{-}rand}$	0.2497	0.2442	0.2727
$r^2_{y\text{-}rand}$	0.5062	0.5196	0.4110
$^{c}r^2_{p}$	0.6000	0.5753	0.5669
$RMSE_{cv}$	0.2424	0.2291	0.2250
r^2_{cv}	0.5733	0.6085	0.6035
$RMSE_{test}$	0.2926	0.2856	0.2686
r^2_{test}	0.6422	0.6007	0.6345
r^2_{m}	0.3147	0.4757	0.5571

TABLE 2 Validation parameters for MIA-QSAR models built using the optimized geometries

Parameters	Values (LV=4) ($\propto r_{vdW}/\varepsilon$)	Values (LV=3) ($\propto r_{vdW}$)	Values (LV=2) ($\propto \varepsilon$)
$RMSE_{cal}$	0.1394	0.1575	0.1892
r^2_{cal}	0.8473	0.8050	0.7188
$RMSE_{y\text{-}rand}$	0.1759	0.2113	0.2648
$r^2_{y\text{-}rand}$	0.7526	0.6386	0.4480
$^{c}r^2_{p}$	0.2832	0.3660	0.4112
$RMSE_{cv}$	0.2891	0.2744	0.2655
r^2_{cv}	0.3704	0.4211	0.4467
$RMSE_{test}$	0.2384	0.2503	0.2006
r^2_{test}	0.6422	0.5932	0.7395
r^2_{m}	0.3147	0.3339	0.5137

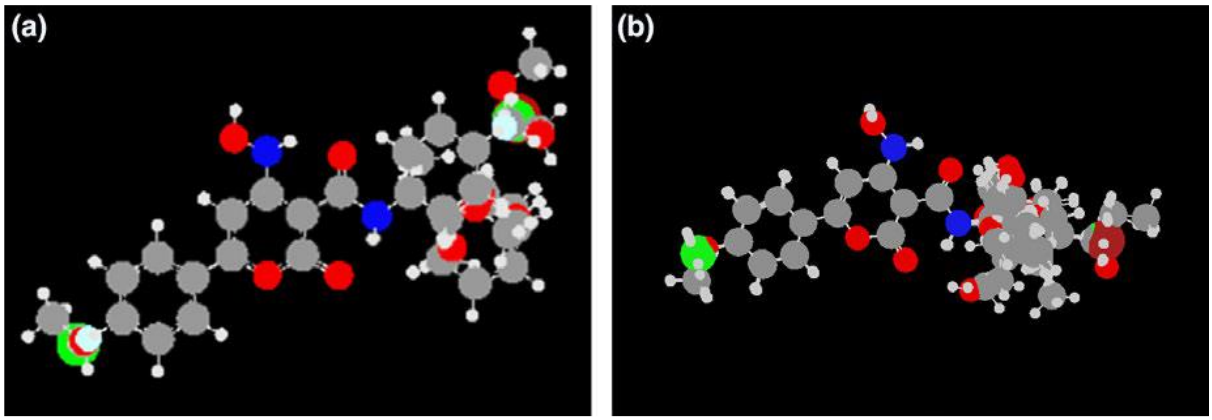


Figure 1 (a) Superposed images for the 42 anti-HCV compounds. The common nucleus is used for 2D alignment, while the variable substructures explain the variance in the pEC_{50}. O = red, Cl = green, N = blue, F = ciano, C = gray, H = light gray, and Br = dark red. (b) The superposed images of the 42 fully optimized anti-HCV compounds. O = red, Cl = green, N = blue, F = ciano, C = gray, H = light gray, and Br = dark red.

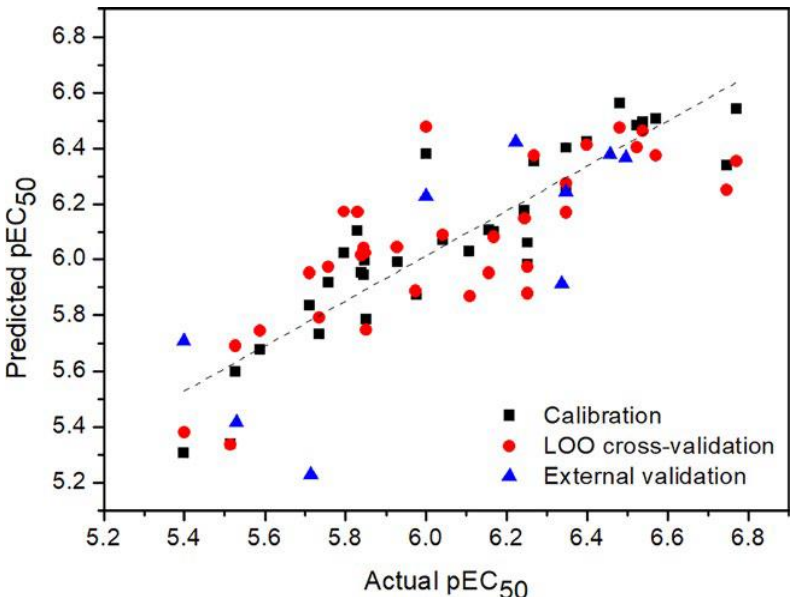


Figure 2 Plot of the experimental × predicted pEC_{50} values using the MIA-QSAR model obtained from atomic colors proportional to ε

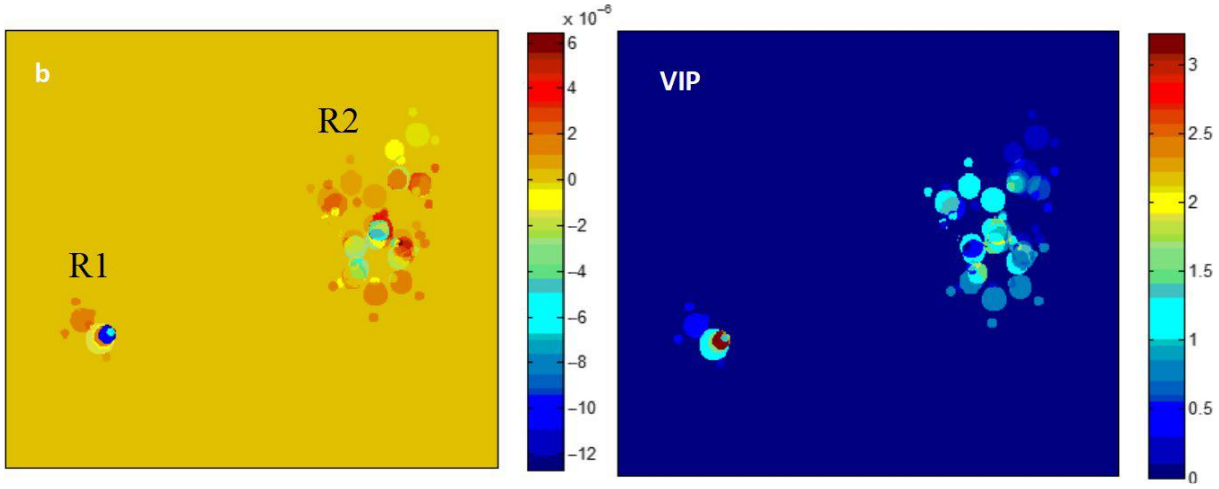


Figure 3 MIA-contour maps based on variable importance in projection (VIP) scores and PLS regression coefficients (b), used to analyze the descriptor contributions for the MIA-QSAR modeling of pEC_{50} values of the anti-HCV molecules.

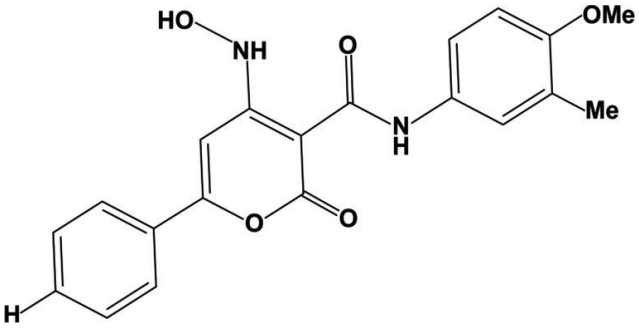
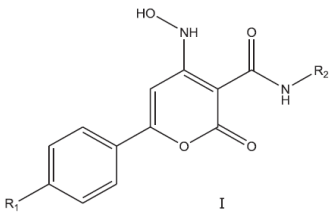


Figure 4. Proposed molecular structure for enhanced anti-HCV bioactivity.

SUPPORTING INFORMATION

Table S1. Series of anti-HCV (Hepatitis C Virus) compounds used in MIA-QSAR modeling and the corresponding actual and predicted (in calibration, leave-one-out-validation, and external validation) pEC_{50} data using the descriptor values proportional to ε for the traditional analysis (model 3).



I

Number	R1	R2	pEC_{50}	Calibration	LOOCV (3LV)	ExternalValidation
12[a]	H	Ph	6.456	-	-	6.377
13	F	Ph	5.975	5.873	5.887	-
14	Cl	Ph	6.252	5.983	5.878	-

15	CH_3	Ph	5.839	5.952	6.017	-
16	OCH_3	Ph	6.252	6.060	5.974	-
17[a]	H	4-F-Ph	6.495	-	-	6.366
18	H	4-Cl-Ph	6.398	6.424	6.411	-
19	H	4-Br-Ph	6.523	6.483	6.402	-
20	H	4-OH-Ph	6.000	6.378	6.476	-
21	H	4-OMe-Ph	6.770	6.540	6.354	-
22	H	4-COOMe-Ph	6.481	6.562	6.473	-
23	H	2-Me-Ph	6.268	6.354	6.373	-
24	H	3-Me-Ph	6.569	6.505	6.373	-
25	H	4-Me-Ph	6.745	6.339	6.249	-
26	F	4-Me-Ph	5.710	5.835	5.952	-
27	Cl	4-Me-Ph	5.845	5.945	6.041	-
28[a]	CH_3	4-Me-Ph	6.337	-	-	5.914
29	OCH_3	4-Me-Ph	5.796	6.023	6.173	-
30[a]	H	Me	6.222	-	-	6.423
31	F	Me	5.757	5.919	5.973	-
32	Cl	Me	6.108	6.029	5.868	-
33	CH_3	Me	5.848	5.998	6.024	-
34	OCH_3	Me	6.155	6.106	5.952	-
35[a]	H	Et	6.000	-	-	6.228
36	H	Pr	6.347	6.263	6.169	-
37[a]	H	*i*-Pr	6.347	-	-	6.243
38	H	C_2H_4OH	5.830	6.102	6.170	-
39	F	C_2H_4OH	5.526	5.598	5.690	-
40[a]	Cl	C_2H_4OH	5.398	-	-	5.708
41	CH_3	C_2H_4OH	5.588	5.677	5.744	-
42	OCH_3	C_2H_4OH	5.851	5.786	5.747	-
43	H	C_3H_6OH	6.347	6.401	6.273	-
44	H	CH_2COOMe	5.735	5.733	5.791	-
45[a]	F	CH_2COOMe	5.714	-	-	5.229
46	Cl	CH_2COOMe	5.514	5.338	5.336	-
47	CH_3	CH_2COOMe	5.399	5.308	5.380	-
48[a]	OCH_3	CH_2COOMe	5.529	-	-	5.416
49	H	CH_2Ph	6.538	6.495	6.463	-
50	F	CH_2Ph	5.928	5.991	6.045	-
51	Cl	CH_2Ph	6.167	6.100	6.080	-
52	CH_3	CH_2Ph	6.041	6.070	6.089	-
53	OCH_3	CH_2Ph	6.244	6.178	6.148	-

[a] – Test group

COMSIA results: [17] r^2: 0.904, q^2: 0.569, r^2_p: 0.609, r^2_{pred}: 0.653, and r^2_m: 0.680

Table S2. Validation parameters for double cross validation technique.

	Calibration				
	Traditional			**Optimized**	
	Model 2	**Model 4**	**Model 7**	**Model 5**	**Model 7**
R^2	0.7075	0.6840	0.7577	0.6582	0.728
Q^2_{LOO}	0.641	0.5735	0.6928	0.5867	0.6718
Avg Rm^2_{LOO}	0.5201	0.4437	0.5858	0.4581	0.5569
MAE (95%)	0.153	0.177	0.1339	0.1731	0.1598
	ExternalValidation				
	Traditional			**Optimized**	
Q^2F_1(95%)		0.6942		0.5305	
Q^2F_2(95%)		0.6937		0.5119	
Scaled Average r_m^2 (95%)		0.6392		0.4966	
Scaled Delta r_m^2 (95%)		0.0325		0.1581	
CCC (95%)		0.8364		0.7714	
MAE (95%)		0.1733		0.1918	
SD (95%)		0.1373		0.1411	

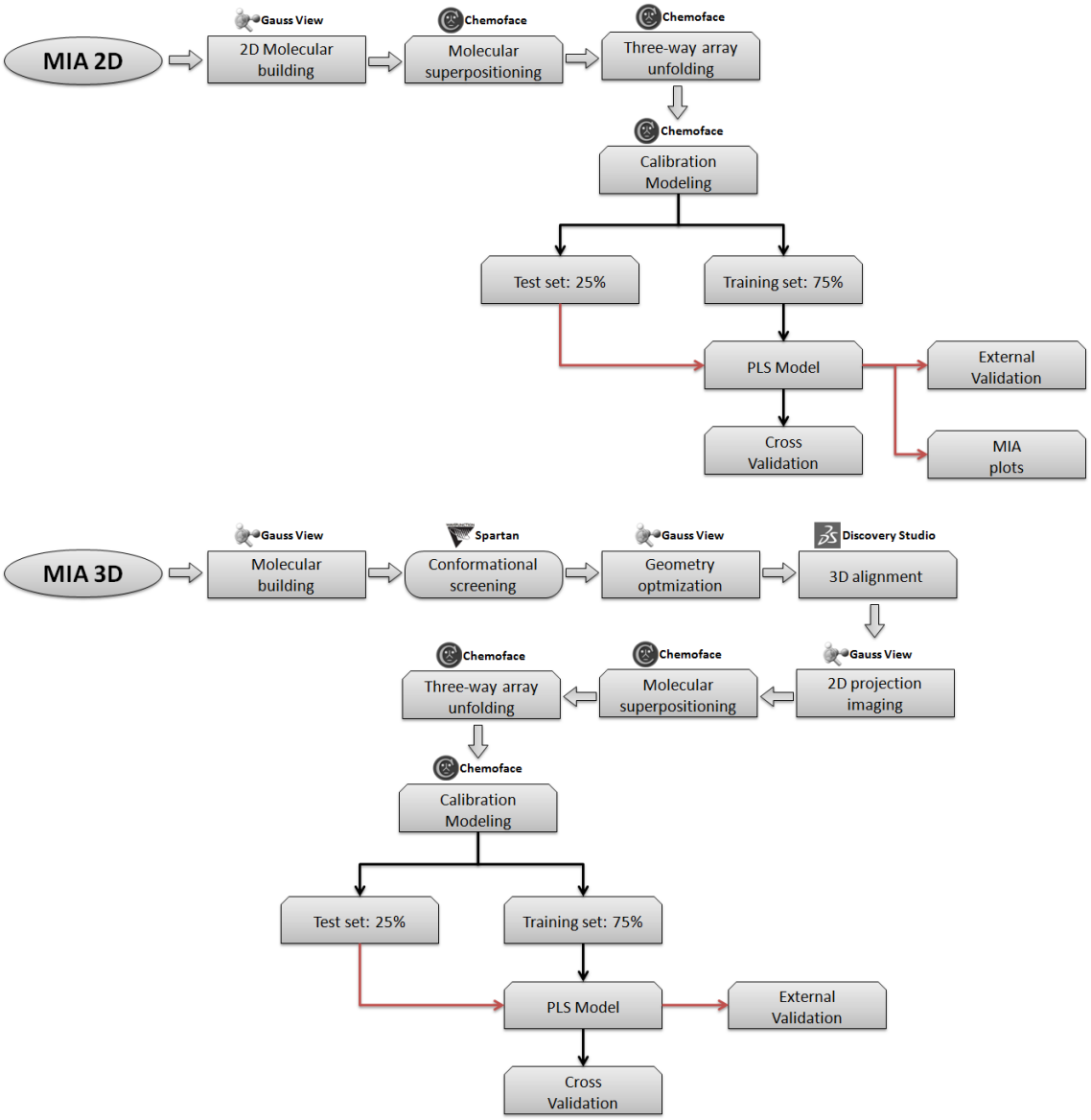


Figure S1. Summary of the steps for performing 2D and 3D MIA-QSAR.

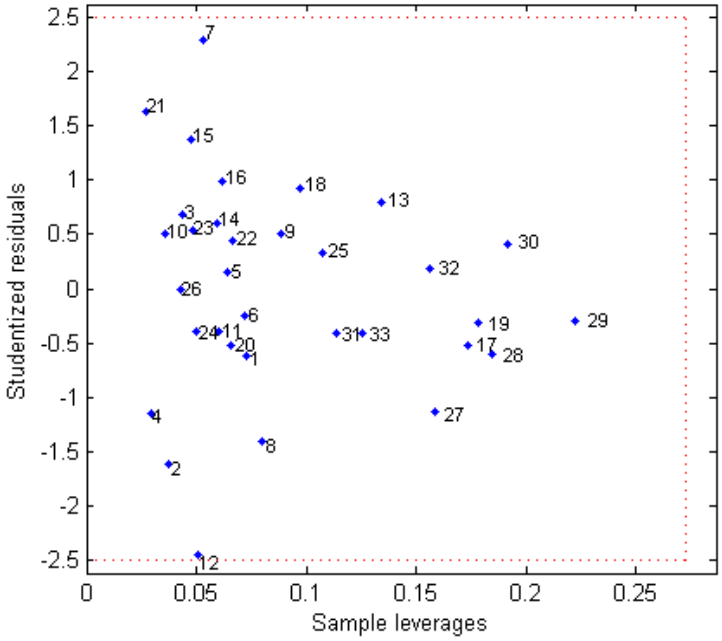


Figure S2. Applicability domain showed through William´s plot.

ARTIGO 3

Journal: *Submitted for publication*

Conformational fingerprints in the modeling performance of MIA-QSAR: A case for SARS-CoV protease inhibitors

Joyce K. Daré,[a] Daniela R. Silva,[a] Teodorico C. Ramalho,[a] Matheus P. Freitas[a,*]

[a] *Department of Chemistry, Federal University of Lavras, 37200-000, Lavras, MG, Brazil*

* corresponding author: matheus@dqi.ufla.br

Abstract

Multivariate image analysis applied to quantitative structure-activity relationships (MIA-QSAR) has been shown to be a high-performance 2D tool for drug design purposes. However, an interesting task that aims at evolving the MIA-QSAR strategy is to understand whether incorporation of conformational information in QSAR analysis would result in significant benefits in terms of predictability and interpretability of the model. Conformational information was included considering the optimized geometries and the docked structures of a series of disulfide compounds as SARS-CoV protease inhibitors. The traditional analysis proved itself as the most effective technique, which means that, despite the undeniable importance of conformation for biomolecular behavior, this type of information did not bring relevant contributions for MIA-QSAR modeling.

Keywords: Conformation, MIA-QSAR, geometry optimization, molecular docking, MIA-plots.

Introduction

Computers have become electronic devices of almost every day use for a great part of society, and their importance is undeniable. Among their diverse applications, a specific one has gained lots of attention from researchers, their use for molecular modeling. Many current techniques employ computational features as a mean for modeling molecules; molecular docking and QSAR (Quantitative Structure-Activity Relationship) are well-known examples of this type of techniques.

Quantitative Structure-Property Relationship (QSPR) is a broad subarea from molecular modeling dedicated to correlate compounds' chemical structural features to a property of interest (biological activity for example) through the generation of a predictive mathematical model (regression equation) [1]. In other words, a QSPR model can predict a specific property for compounds that have not been studied, based on structurally similar molecules that have already been investigated. This predictive ability has great contribution in identifying novel drugs, once it accelerates the screening part of this long process.

Although it may seem that QSPR is a recent molecular modeling tool due to the boom of works published from the 1970's, its roots goes back to 1868-1869, when Crum Brown and Fraser established that there was no doubt that, in fact, there was a relation between the physiological action of a substance and its chemical composition and constitution [2]. Nevertheless, the area of QSAR suffered many advances associated with the improvements in computer's hardware and software achieved in the last decades, once this equipment facilitates and accelerates the extraction of descriptors and generation of prediction models. Furthermore, associated with molecular docking, QSAR avoids the experimental tests of lots of unpromising compounds, which results in a large saving of costs. Due to these important contributions, QSAR has become the target of many studies.

There is a variety of methods to generate QSAR models and, basically, the remarkable difference among them is the type of descriptors employed in the analysis and the way these parameters are obtained from the data set. Commonly, the different approaches are classified from 1D to 6D; each of them has its own advantages and drawbacks [3]. MIA-QSAR (Multivariate Image Analysis applied to QSAR) is an interesting 2D approach that has been employed in many occasions in the last decade yielding satisfactory outcomes, comparable to 3D results [4-7].

Essentially, MIA-QSAR is a method based on the treatment of 2D colored molecular images for obtaining the required descriptors in a QSAR analysis. The molecules are built in a way that their atoms have sizes proportional to van der Waals radii and their colors are numerically described to be proportional to the respective electronegativities. The color system employed is the RGB (Red, Green, and Blue). Furthermore, in the current version of this technique, the user counts on the aid of what is called MIA-plots tool, which consists on colored graphics that help interpreting the generated model [8,9].

The main advantages of MIA-QSAR over other techniques are: (a) the descriptors do not require complex approaches and computations to be obtained and, therefore, the model

can be easily built; (b) the models can be chemically interpreted through the aid of the MIA-plots tool. Of course, there are some drawbacks as well, among them, it can be highlighted the fact that conformational information (tridimensional features) of the compounds are, usually, not deeply considered.

In this sense, this work aims to investigate whether incorporation of conformational information in MIA-QSAR analysis would result in significant benefits in terms of predictability and interpretability or not. For evaluating the influence of tridimensional molecular representations on the modeling performance of MIA-QSAR models, a series of SARS-CoV M^{pro} (severe acute respiratory syndrome coronavirus main protease) inhibitors, synthesized by Wang and coworkers, was chosen [10]. The authors also performed a 3D-QSAR modeling (CoMFA) analysis, which will be used herein as a reference model for means of comparison.

SARS is an infective disease that affects the respiratory system and is caused by a RNA type of virus known as coronavirus (CoV) [11]. This large family of viruses affects not just humans, but also a variety of other animals. Currently, six human coronaviruses (HCoVs) have been identified, and, although they have been known for decades, only in 2002 they gained clinical importance, which started due to an outbreak of SARS and MERS (Middle East respiratory syndrome – another disease caused by a coronavirus agent) [11]. Since these severe epidemics, which are now under control, many targets have been studied; among them, the $3CL^{pro}$, also known as M^{pro} (main protease), has been considered an interesting approach for inhibition [12]. Therefore, different inhibitors for SARS-CoV-M^{pro} have been proposed; Wang *et al.* [10] synthesized unsymmetrical aromatic disulfide compounds, which exhibited an encouraging biological potency.

It is worth mentioning that there is no doubt that molecular conformation plays a decisive role in affecting the bioactivity of a drug. However, there are not proofs that considering this type of parameter would generate more efficient/reliable QSAR models [13].

Materials & methods

The data set containing the anti-SARS molecules is shown in Table 1. For a clearer view of the goals of this project, this section was divided into three parts, as follows.

TABLE 1 (Appendix)

Traditional MIA-QSAR modeling

Initially, the anti-SARS molecules underwent a traditional MIA-QSAR modeling. The procedure to build MIA-QSAR models has been described in details elsewhere [4–9], thus,

only the main aspects are highlighted herein. At first, the compounds were drawn in GaussView program [14] maintaining the congeneric center at the same exactly position in all molecules. Then, they were saved in a bitmap (.bmp) format. Subsequently, the 2D images were loaded and superposed (Figure 1a), yielding a three-way array, which was further unfolded in a new matrix with dimensions 40 × 108340, using the Chemoface software for chemometrics (available at http://ufla.br/chemoface/) [15].

FIGURE 1 (Appendix)

The next step consisted in the exclusion of the invariant columns (which correspond to the blank spaces and the congeneric center), which generated a new matrix with dimensions 40 × 12215. The pixel values were, subsequently, substituted by numbers proportional to (i) the respective values of r/ε ratio (van de Waals radius/Pauling's electronegativity), (ii) the van der Waals radius values, and (iii) the respective electronegativity values.

The pixel colors of RGB 24 bits system are a result of the combination of its three raw color channels (red, green, and blue). Each channel can assume values that varies from 0 to 255; as consequence, the pixel value varies from 0 ('black') to 765 ('white'). Accordingly, the substitution of the pixel values by numbers proportional to chemical properties performed herein, followed the subsequent pattern; for those substituted by 'r/ε', the employed pixel values were: O = red (229) replaced by 211, Cl = green (289) replaced by 330, N = blue (279) replaced by 307, F = ciano (688) replaced by 178, C = gray (426) replaced by 364, H = light gray (612) replaced by 176, Br = dark red (231) replaced by 373. For those substituted by 'r': O = 730, Cl = 990, N = 940, F = 710, C = 930, H =370, Br = 1140. Finally, for those substituted by values proportional to electronegativity: O = 344, Cl = 316, N = 304, F = 400, C = 255, H =220, Br = 296.

Subsequently, to perform the QSAR model and for reasons of comparison with Wang's work, compounds **8**, **23**, and **40** were deleted, once they were considered outliers by the authors [10]. In Wang's work external validation was not performed, thus, it was necessary to build two models for each data set (proportional to r/ε, r and ε): one not considering a test set, in this way the model would have the same dimensional space; and, one considering a group for external validation, since it is well-known that this type of validation is extremely informative on the efficiency and reliability of the model. Therefore, six prediction models were built, in total, through traditional analysis. The validation parameters for the two best models, one without test set (A) and one with test set (B), are shown in Table 2.

The test sets were selected through Kennard-Stone algorithm and contained 9 elements (highlighted with the symbol '[a]' on Table 1), which corresponds to *ca.* 25% of the data set. For all models, partial least squares (PLS) regression was employed, and the optimum number of latent variables (LV) was chosen by analyzing the decay of the root mean square error (RMSE) in the leave-one-out cross-validation (LOOCV).

For models without test set, the following parameters were employed to analyze the quality of the model: regression coefficients in calibration (r^2) and cross-validation (q^2), and their respective RMSE's. The risk of chance correlation in calibration was analyzed using the $^{c}r^2_P$ parameter [$^{c}r^2_P = r \times (r^2 - r^2_{y\text{-}random})^{1/2}$, where $r^2_{y\text{-}random}$ corresponds to the mean determination coefficient value obtained after randomizing the y block a few times] [16].

For models containing test set, in addition to r^2 and q^2, r^2_{test} and its corresponding RMSE were also considered. The proximity between the actual and predicted IC_{50} values for the test set was statistically evaluated using r^2_m parameter [17]. The risk of chance correlation was studied using the same parameter mentioned above.

MIA-QSAR modeling using optimized molecular geometries

For the second part, the main goal was to introduce conformational information into the MIA-QSAR descriptors and analyze the consequences of such attempt in the QSAR model. However, it is worth mentioning that these 3D features were obtained from molecules isolated from their receptor, the main protease SARS-CoV M^{pro}. In light of this approach, the molecules were built, once more, in GaussView program, but now there was not a concern about maintaining the congeneric center at the same position for the compounds, once they would still undergo through the process of geometry optimization.

As a starting point for the geometry optimization process, a conformational screening was performed using the Spartan'16 program [18] for all 40 compounds at the semi-empirical AM1 (Austin Model 1) level of theory [19]. The lowest energy conformation was selected for each case and, then, they were fully optimized, using the Density Functional Theory (DFT) method at the ωB97X-D/6-31G(d,p) [20,21] level of theory, in the Gaussian 09 program [22].

Next, the optimized structures were loaded on Discovery Studio Visualizer [23], one at a time, and the superposition of each molecule was performed using the *Tether tool* available on the manual alignment tab. This superposition was performed having as reference the congeneric center (*i.e.*, the disulfide moiety bonded to a phenyl group on the right side) and the resulting three-way-array can be observed at Figure 1b. Then, the aligned molecules were loaded on GaussView program and saved as bitmap images. Employing this strategy, conformational information was included into the MIA-QSAR descriptors. Finally, the MIA-QSAR models were generated following the same steps of the traditional technique, *i.e.*, six more models were obtained; the validation parameters for the two best models are shown in Table 2 [model C (without a test) and model D (with a test)].

MIA-QSAR modeling using bioconformation-like images

For the last part of this work, a different and more significant type of tridimensional information was employed: bioconformations, *i.e.* conformational information obtained

from the molecules docked into the active site of their receptor (SARS-CoV M^{pro}). These bioconfomations were obtained employing the molecular docking technique.

In order to perform molecular docking, the crystal structure of SARS-CoV M^{pro} with 1.85 Å of resolution was obtained from the protein data bank (PDB code: 2AMD [11]). It is worth mentioning that this PDB structure was complexed with the inhibitor (N9). This raw structure was subsequently prepared for docking using the 'Protein Preparation Tool' available in Schrodinger software. During this preparation, some tests were performed to select the best features for further analysis. The tests were analyzed taking into account the results obtained by Wang and coworkers [10] and also those results obtained by Yang *et al.* [11]. Finally, the best preparation process was that in which all water molecules were removed and only the chain A was used during docking analysis, which is discussed in more details in the next section.

The ligands were prepared in a two-step way process. First, their geometries were optimized at Gaussian 09 program, using Density Functional Theory (DFT) method at the ωB97x-D/6-31G(d,p) level of theory. Secondly, a charge calculation (CHelpG at the same theoretical level of the optimization) was performed to prepare the ligands for the molecular docking process.

A final step, before docking, was establishing the Grid Box and the active site of the receptor. Then, molecular docking was initiated by loading the protein prepared structure as well as the ligand structures on Maestro workspace. No constrains were established and one hundred conformations were obtained for each compound. Compound **31** was used as reference, once it was the most efficient inhibitor determined by Wang *et al.* [10] and, also, the same (bio)conformation was searched among the results obtained herein. Once localized the sought conformation for compound **31** (see Figure 2), similar bioconformations for all the other molecules were identified.

FIGURE 2 (Appendix)

The 40 identified conformations were loaded into Discovery Studio Visualizer and superposed in a similar way of the optimized molecules, *i.e.* maintaining the congeneric center at the same position (Figure 1c). The aligned molecules were, then, loaded on GaussView program and saved as bitmap images. Employing this strategy, bioconformational information was incorporated into the MIA-QSAR descriptors, since the images correspond to 2D projections of 3D (bio)chemical structures. Finally, the 3D-MIA-QSAR models could be prepared, which was performed following the same steps as the earlier analysis. Therefore, six more models were obtained; the validation parameters for the two best models are shown in Table 2 [model E (without a test) and model F (with a test)] along with all the other best models (A-F).

Results and Discussion

QSAR Analysis

Regarding the traditional analysis, the resulting superposed molecules are shown in Figure 1a.

The quality parameters for the best model (A), *i.e.* the model without a test set, is shown on Table 2. On the same table, the quality parameters for all the other models (the best one in each case) are also displayed. In this way, a general comparison can be easily performed.

TABLE 2 (Appendix)

It can be observed that the best models were obtained from the traditional analysis. Model A, employing pixel descriptors proportional to r/ε, generated good results for internal validation and calibration, although the RMSEs associated with r^2 and q^2 are slightly high. The ${}^{c}r^2_P$ parameter eliminated the risk of chance correlation, which ensures that the model was not a result of randomness. Therefore, comparing to Wang's [9] results obtained from CoMFA technique (see Table 1 footer), model A was as good as those earlier obtained. Model B (r/ε) reinsures the quality and predictive ability of models obtained through traditional technique, once the external parameters were all good. This last observation shows that model B is not overfitted, *i.e*. it is a useful equation for predicting IC_{50} for compounds not included in the analysis, which is the main goal for a QSAR technique.

Models C (r/ε) and D (r) obtained from the optimized anti-SARS structures were not as good as those obtained from the traditional procedure. The values of r^2 and q^2 for model C were acceptable, but the risk of chance correlation could not be eliminated by the ${}^{c}r^2_P$. For model D, another problem was the q^2 value, since it was not within an acceptable range (≥ 0.5). An interesting observation in this case is that, for model D, the external parameters presented reliable results; this just confirms what Tropsha made very clear in his work entitled "Best practices for QSAR model development, validation, and exploitation", *i.e.*, both internal and external validations are required in order to consider a QSAR model as a reliable prediction equation [24]. The resulting superposed molecules employed on the generation of models C and D are shown in Figure 1b.

Lastly, those models obtained for molecules that passed through molecular docking procedure (E (ε) and F (ε)) did not present good quality parameters either. The risk of chance correlation could not be eliminated from neither models and even the external validation for model F presented a bad result for r^2_m. The resulting superposed molecules employed on the generation of models E and F are shown in Figure 1c.

Therefore, it can be concluded that, although conformational information is responsible for a great part of biomolecular behavior, this type of information does not contribute positively for MIA-QSAR methodology. This finding can be related with the loss of reference that happens with the optimized and docked molecular structures. In other words, in a traditional analysis the model can be built assuming that a chemical group is always in a

specific position and, based on that position, the effect (increasing/decreasing) on the response variable can be determined and, then, predicted later for a different structure containing the same group at the same position. On another hand, for the optimized and docked molecular structures, due to the position variability, the identification of a pattern for the effect of the different groups on the response variable is jeopardized and, consequently, the prediction is less precise. Of course, it is known that in real biological systems, a chemical group is not always at the same position, but for a MIA-QSAR analysis, apparently, this type of information does not contribute with the generated model.

Furthermore, once the focus of this discussion involves bitmap images, one can rationalize the whole problem in terms of pixels. In both analysis (traditional and 3D), the resulting matrix, obtained after the unfolding step, contains columns that correspond to every single pixel of all molecular images; in other words, a set of columns corresponds to a part of a molecule. In the traditional analysis, all the samples (lines) containing the same chemical group will, for sure, have the same values attributed to those columns corresponding to that specific portion; on another hand, in a 3D analysis, although two (or more) molecules may have the same chemical group in the same spatial position, they will not, necessarily, be at the same pixel position (columns), because their images include variation in angles and direction; in summary, samples containing the same moiety will not have the same values attributed to the columns that correspond to that part. This lack of pattern brings a considerable amount of variability to the descriptors, which interferes in the quality of the prediction models.

After determining the best models (A and B), a chemical interpretation can be performed employing MIA-plots. Model B was chosen for the interpretation process, once it is a more complete QSAR analysis compared to model A. The calculated results for calibration, internal and external validation are shown on Table 1, while the predicted $\times$ measured IC_{50} plot is shown in Figure 3. Figure 4, in turn, shows the MIA-plots obtained for model B. These plots are based on an analysis of the structural moieties most affecting (either enhancing or attenuating) the biological data in terms of PLS regression coefficients (b) and variable importance in projection (VIP) [8]. The applicability domain (William´s plot) for the chosen model was also checked and is shown in Figure 5.

FIGURE 3 (Appendix)

FIGURE 4 (Appendix)

FIGURE 5 (Appendix)

William's plot shows that samples 9 and 10 are considered outliers which corresponds to compounds **10** and **11**, once compound **8** was previously excluded. However, in order to obtain the same space domain for means of comparison with Wang's work, both samples were kept in the data set for further analysis; furthermore, both *outliers* are close to the established borders.

From the fact that the best model for the data set was the one with the descriptor values proportional to radii/Pauling's electronegativity, it can be observed that both steric and electrostatic effects are relevant for the inhibition activity of the disulfide compounds. In addition, since r/ε-based descriptors best explained the data variance in our models A and B, it follows that both atomic size and electronegativity play a more significant role for this QSAR modeling than these parameters alone.

From Figure 4, more specifically looking at the b-plot, it is possible to visualize specific moieties (herein called R1, R2, and R3) responsible for increasing/decreasing IC_{50}. Figure 1a can collaborate in a more detailed understanding.

Focusing on the VIP plot and its general aspects, it is possible to recognize a specific pattern on how strong the three groups affect the response variable. The R1 substituents comprise the most important variation in colors, which includes dark shades of red, yellow, and orange; accordingly, it can be considered the most significative moiety in explaining the observed response pattern. The R2 group also presents some significative variety in shades, but much less than R1, which means that it has less influence in explaining the IC_{50} than the former. Lastly, R3, in a general view, is the one with the smallest contribution for clarifying the IC_{50} behavior. In summary, R1>R2>R3 is the importance scale proposed.

Based on the previous generalization, it is reasonable to start the b-plot interpretation with the R1 substituents. Furthermore, compounds **31** and **15** can be taken as references for this analysis, once the former corresponds to the smallest value of IC_{50} (most active) and the latter to the biggest value (less active) of this series of molecules.

In order to analyze the R1 influence, it seems reasonable to select a subset of compounds containing the same R2 and R3 substituents in all cases. The chosen group comprises the following molecules: 1, 3, 12, 17, 31, 33, and 38; their structures are shown in Table 1.

Compound 1 has a five-membered ring with endocyclic Nitrogen and Sulfur in the positions 2 and 5; this ring presents a light green color in the b-plot, which indicates a slight increase in the IC_{50} value. Taking into account that the IC_{50} parameter varies from 0.516 μM to 5.954 μM, this interpretation agrees with the actual value measured for this compound – 1.871. Actually, one can see that, except for compound 31, all further molecules presenting a five-membered ring (3, 12, and 17) have bigger values of IC_{50} than those with a six-membered ring; furthermore, these compounds - 3, 12, and 17 - contain carboxyl/carbonyl groups (red color in the b-plot) attached to the five-membered ring, which also contributes, significantly, for the observed increasing in the IC_{50} parameter. Further analysis of the b-plot shows that a blue tone is attributed to the six-membered ring substituent; therefore, the previous conclusion agrees with the MIA-plot interpretation. Moreover, the unsubstituted six-membered ring seems to decrease the IC_{50} even more than the one with two methyl groups attached.

Regarding compound 31, not only it has a smaller value of IC_{50} than those with a six-membered ring, but also it is the most active compound of the series. Turning the attention to its groups on the regression coefficient plot, one can see that the moiety that seems to

strongly decrease the response variable is the Chlorine atom in the R3 position, which is an exception to the previous importance scale stablished. This fact might be related to a synergetic effect of the three substituents or its specific behavior inside the active site of the M^{pro} enzyme, which is not explained by MIA-QSPR plots. It is also important to understand that the MIA-plot tool shows the general behavior of a data set and, once compound 31 is an exception, it might not be well explained.

Following the same thought, in order to analyze the influence on IC_{50} due to the changing in the R2 substituent, a subset with invariable R1 and R3 groups was selected, which comprises compounds 14, 15, and 16. Focusing on the b-plot, it is possible to observe that the ester moieties present strong red/orange colors, which characterizes an increasing contribution for IC_{50}. On the other side, the nitro group presents a cyan color, which shows a decreasing contribution for IC_{50}.

Lastly, for analyzing R3 and following the same rationalization than previously, two subsets were selected: one with compounds 1, 33, and 38; and one with compounds 23, 34, and 39. It is easy to conclude from these subsets that a pattern cannot be identified for the R3 contribution in explaining the response variable. This observation agrees with the general assertation that the R3 group did not have a significative influence in explaining the observed IC_{50} pattern, which was made based on the VIP plot.

Lastly, from this analysis, it can be proposed that a disulfide compound containing a six-membered ring with 2 endocyclic Nitrogen atoms (positions 2 and 6) at the R1 position, a nitro group at R2 and a Chlorine at R3 (Figure 6) seems to be an interesting approach on designing a new SARS-CoV M^{pro} inhibitor. The IC_{50} of the proposed molecule was obtained applying the best MIA-QSPR model and it was equal to 0.440 µM, which is even better than compound 31 (most active compound of the series).

FIGURE 6 (Appendix)

Molecular Docking Analysis

The pose for compound **31** used as reference is shown in Figure 2 as well as some hydrogen bonds established with amino acid residues considered important for the inhibition of SARS-CoV M^{pro}.

The conformation shown above was chosen because of its similarity with that selected by Wang *et al.* [10], including the interactions established with Cys145 and Gly143 amino acid residues (also found by Wang *et al.* [10]). Cys 145, according to Wang *et al.* [10], plays a main role in the activity against SARS-CoV M^{pro}. Apparently, this residue binds to the disulfide compound covalently, which prevents the protein biological action. Therefore, it is important that the interaction with this residue has been identified herein. Furthermore, the superposition of compound **31**'s conformation with N9 was also evaluated and it presented a good alignment.

Finally, it is worth discussing the use of only protomer A instead of both units. According to Yang *et al.* [11], N1 (a similar inhibitor crystallized inside SARS-CoV M^{pro}) binds to protomers A and B of SARS-CoV M^{pro} in an identical and normal manner. Therefore, it was assumed a similar binding mode for the disulfide data set [10]. A test for compound **31** including both protomers was also performed and no meaningful differences were found between including both units and only the chain A. Then, looking for less time-consuming analysis, only protomer A was used.

Conclusions

MIA-QSAR, as currently performed, is a powerful tool for building prediction models for biological properties. In other words, conformation does not play a significant role in such technique, maybe due to an alignment rule in which chemical structures are superposed in a non-optimal way, as a result of different (bio)conformational behaviors. In this way, connectivity features encoding atomic properties in a well-defined space and comprising the congeneric center perfectly aligned, such as in the traditional MIA-QSAR approach, appear to be more instructive than 3D information. Thus, from the QSAR models built, the best one was that obtained in a traditional way and, from it, a meaningful chemical interpretation was possible. Lastly, the docking procedure helped to bring meaning to the entire analysis, once an acceptable binding mode for the disulfide compounds with their receptor could be proposed.

Although this work has shown that conformation does not play a major role in MIA-QSAR, which is a 2D methodology, a question may arise: could bidimensional molecular representation be employed in a 3D technique (*e.g.* CoMFA, CoMSIA, etc.) as an efficient source for QSAR descriptors? In other words, is conformation really a requirement in this type of analysis? Further studies using perfectly congruent flat-shape structures into a 3D-QSAR approach may aid answering this question. In addition, the lack of correlation using the 3D-MIA-QSAR attempt can be due to the appearance of a non-linearity profile, which can be further accounted for using proper techniques. Also, generation of MIA descriptors in a 3D space and subsequent regression of the 4-way array against the response variable block using multilinear regression methods, *e.g.* N-PLS, is a promising perspective in the field of MIA-QSAR.

Acknowledgments

Authors are thankful to FAPEMIG for the financial support of this research (grant numbers: CEX-APQ-00383-15 and PPM-00344-17), as well as to CAPES for a studentship (to J.K.D.), and to CNPq for a studentship (to D.R.S.) and fellowships (to T.C.R. and M.P.F.).

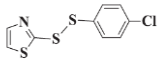
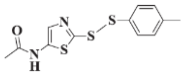
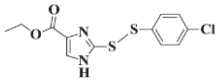
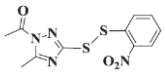
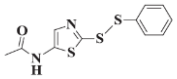
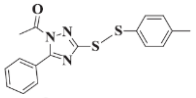
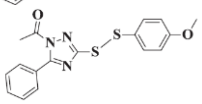
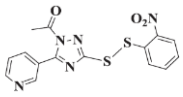
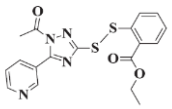
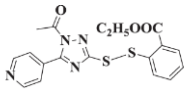
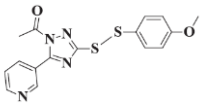
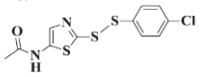
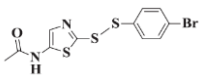
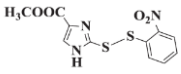
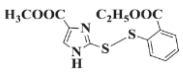
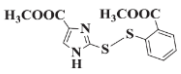
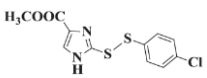
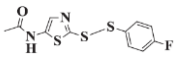
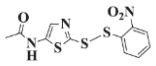
References

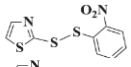
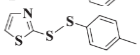
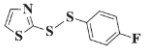
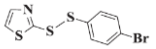
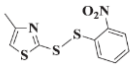
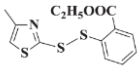
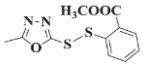
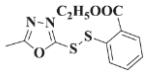
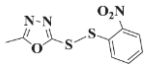
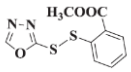
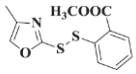
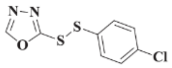
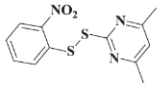
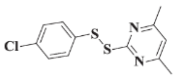
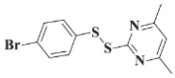
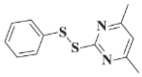
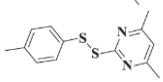
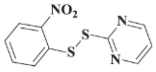
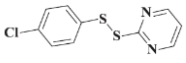
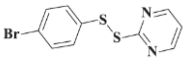
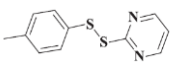
1. Hansch C, Leo A HD. Exploring QSAR. Hydrophobic, Electronic, and Steric Constants. *ACS Prof. Ref. B.* Washington (1995).
2. Dearden J. The History and Development of Quantitative Structure-Activity Relationships (QSARs). *Int. J. Quant. Struct. Relationships*. 1(1)(2018)
3. Consonni, V.; Todeschini R. Molecular Descriptors for Chemoinformatics. *Wiley-VCH*. Verlag (2009).
4. Daré JK, Barigye SJ, Freitas MP. Multi-Objective Modeling of Herbicidal Activity from an Environmentally Friendly Perspective. *Int. J. Quant. Struct. Relationships*. 2(2), 16–26 (2017).
5. Daré JK, Silva CF, Freitas MP. Revealing chemophoric sites in organophosphorus insecticides through the MIA-QSPR modeling of soil sorption data. *Ecotoxicol. Environ. Saf.* 144, 560–563 (2017).
6. Guimarães MC, Silla JM, da Cunha EFF, Ramalho TC, Freitas MP. Is the bioconformation of 5-deoxy-5-fluoro-d-xylulose affected by intramolecular hydrogen bonds? *RSC Adv.* 6(113), 111681–111687 (2016).
7. Freitas MP, Brown SD, Martins JA. MIA-QSAR: A simple 2D image-based approach for quantitative structure-activity relationship analysis. *J. Mol. Struct.* 738(1–3), 149–154 (2005).
8. Barigye SJ, Duarte MH, Nunes CA, Freitas MP. RSC Advances MIA-plot: a graphical tool for viewing descriptor contributions in MIA-QSAR. *RSC Adv.* 6, 49604–49612 (2016).
9. Borges CN, Barigye SJ FM. Towards molecular design using 2D-molecular contour maps obtained from PLS regression coefficients. *Mol. Phys.* 115(23), 3044–3050 (2017).
10. Wang L, Bao BB, Song GQ, *et al.* Discovery of unsymmetrical aromatic disulfides as novel inhibitors of SARS-CoV main protease: Chemical synthesis, biological evaluation, molecular docking and 3D-QSAR study. *Eur. J. Med. Chem.* 137, 450–461 (2017).
11. Yang H, Xie W, Xue X, *et al.* Design of Wide-Spectrum Inhibitors Targeting Coronavirus Main. *Plos Biol.* 3(11), 30428 (2005).
12. Yin Y, Wunderink RG. MERS, SARS and other coronaviruses as causes of pneumonia. *Respirology*. 23(2), 130–137 (2018).
13. Guimarães MC, Duarte MH, Silla JM, Freitas MP. Is conformation a fundamental descriptor in QSAR? A case for halogenated anesthetics. *Beilstein J. Org. Chem.* 12, 760–768 (2016).
14. Dennington RD, Keith TA, Millam MJ. *GaussView 5.0.* Wallingford (2008).
15. Nunes CA, Freitas MP. Introducing new dimensions in MIA-QSAR: A case for chemokine receptor inhibitors. *Eur. J. Med. Chem.* 62, 297–300 (2013).
16. Mitra I, Saha A, Roy K. Exploring quantitative structure-activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants. *Mol. Simul.* 36(13), 1067–1079 (2010).
17. Roy K, Chakraborty P, Mitra I, Ojha PK, Kar S, Das RN. Some case studies on application of “rm2” metrics for judging quality of quantitative structure-activity relationship predictions: Emphasis on scaling of response data. *J. Comput. Chem.* 34(12), 1071–1082 (2013).
18. Spartan’16 Software. *Wavefunction Inc.* Irvine (2017).

19. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* 107(13), 3902–3909 (1985).
20. Chai J-D, Head-Gordon M. Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* 128(8) (2008).
21. Krishnan R, Binkley JS, Seeger R PJ. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys.* 72(1), 650 (1980).
22. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Ro.. MA CJ, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M LX-, Hratchian HP, Iimaylov AF, Bloino J, Zheng G SD. Gaussian 09, Revision D. 01, Gaussian Inc., Wallington, CT. (2013).
23. Dassault Systèmes BIOVIA - Discovery Studio Visualizer 2017R2. (2017).
24. Tropsha A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inf.* 29(6–7), 476–488 (2010).

Appendix

Table 1. Data set of the disulfide compounds with their respective measured IC_{50} (μM), the calibration results for Model B, as well as their cross- and external validation results.

Id. Number	Compound	IC_{50}	Calibration	LOOCV (3 LV's)	External Validation
1		1.871	1.692	1.730	-
2		2.803	2.342	2.091	-
3		3.675	3.386	3.317	-
4		3.130	3.227	3.276	-
5[a]		1.506	-	-	2.517
6		4.344	4.747	5.353	-
7[a]		4.100	-	-	5.206
8[*]		1.762	-	-	-
9[a]		5.654	-	-	4.501
10		4.511	4.449	4.298	-
11		5.794	5.183	4.522	-
12[a]		2.626	-	-	2.034
13		1.651	1.888	2.306	-
14		2.075	3.531	4.012	-
15		5.954	5.348	3.804	-
16		3.957	4.636	4.793	-
17		4.126	3.280	3.010	-
18[a]		2.565	-	-	2.487
19		1.947	2.285	2.385	-

20		2.029	1.944	1.890	-
21[a]		1.250	-	-	2.001
22		2.211	2.145	2.107	-
23[*]		3.321	-	-	-
24		2.555	2.664	2.662	-
25		2.452	2.502	2.340	-
26		1.679	1.136	0.969	-
27		1.557	1.304	1.227	-
28		1.713	1.466	1.457	-
29		1.118	0.915	0.920	-
30		1.264	1.773	2.130	-
31		0.516	0.995	1.460	-
32		0.921	1.422	1.678	-
33[a]		1.437	-	-	1.170
34		1.121	1.025	1.170	-
35		1.991	1.654	1.429	-
36[a]		1.495	-	-	1.479
37		0.883	0.920	1.008	-
38		0.684	0.669	0.903	-
39[a]		0.697	-	-	0.523
40[*]		1.522	-	-	-

[*] - outliers [a] – test set

Table 2. Statistical parameters obtained through traditional MIA-QSAR technique (Models A and B), MIA-QSAR applied to optimized molecular geometries (Models C and D), and the same technique employed to bioconformation-like images (Models E and F).

	Traditional		Optimized		Docked	
Parameters	**Model A (r/ε)**	**Model B (r/ε)**	**Model C (r/ε)**	**Model D (r)**	**Model E (ε)**	**Model F (ε)**
LV	3	3	3	2	3	3
$RMSE_{cal}$	0.5266	0.4653	0.4452	0.4837	0.3999	0.3110
r^2_{cal}	0.8704	0.8969	0.9074	0.9049	0.9253	0.9523
$RMSE_{y\text{-}rand}$	1.1829	1.0362	0.7384	0.5568	0.6023	0.3882
$r^2_{y\text{-}rand}$	0.3446	0.4880	0.7416	0.8708	0.8299	0.9250
$^{c}r^2_{p}$	0.6765	0.6056	0.3879	0.1755	0.2970	0.1613
$RMSE_{cv}$	0.7805	0.7943	1.0223	1.2338	1.1187	1.2132
r^2_{cv}	0.7154	0.7009	0.5239	0.3924	0.4204	0.2839
$RMSE_{test}$	-	0.7146	-	0.5926	-	0.9458
r^2_{test}	-	0.7806	-	0.8232	-	0.6006
r^2_{m}	-	0.6526	-	0.8122	-	0.2682

[a] CoMFA results from the literature [10]: 6 LV´s, r^2 = 0.916, Standard Error = 0.088, and q^2 = 0.681.

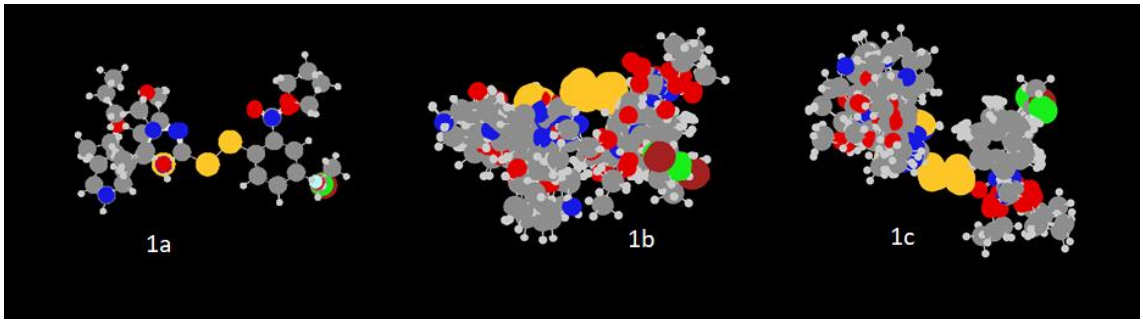


Figure 2. Superposition of the disulfide molecules images used in the (**1a**) traditional MIA-QSAR procedure, (**1b**) MIA-QSAR with optimized geometries, and (**1c**) MIA-QSAR with docked structures.

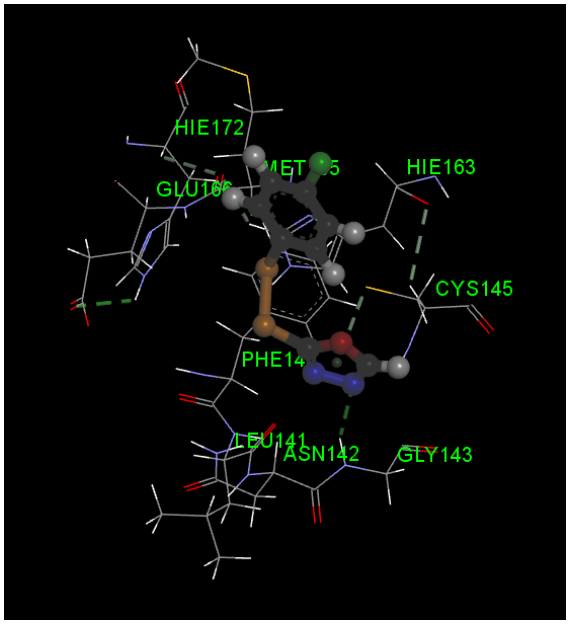


Figure 2. Conformation selected for compound **31** used as reference for the other disulfide structures.

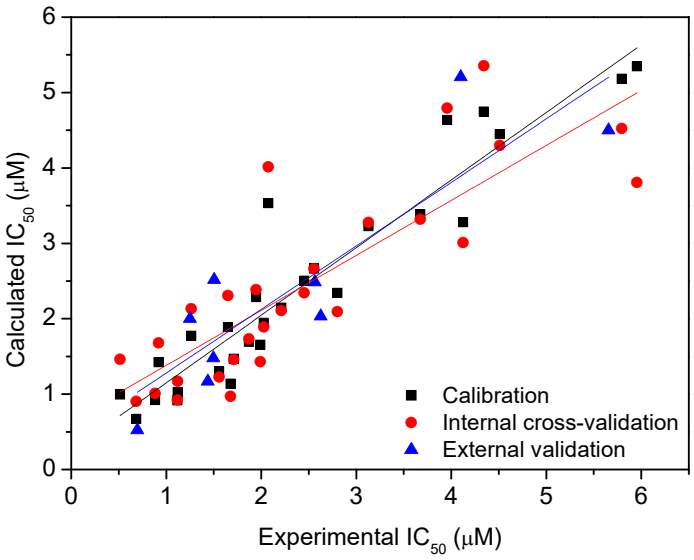


Figure 3. Plot for predicted × experimental IC_{50} values using the MIA-QSAR model **B**.

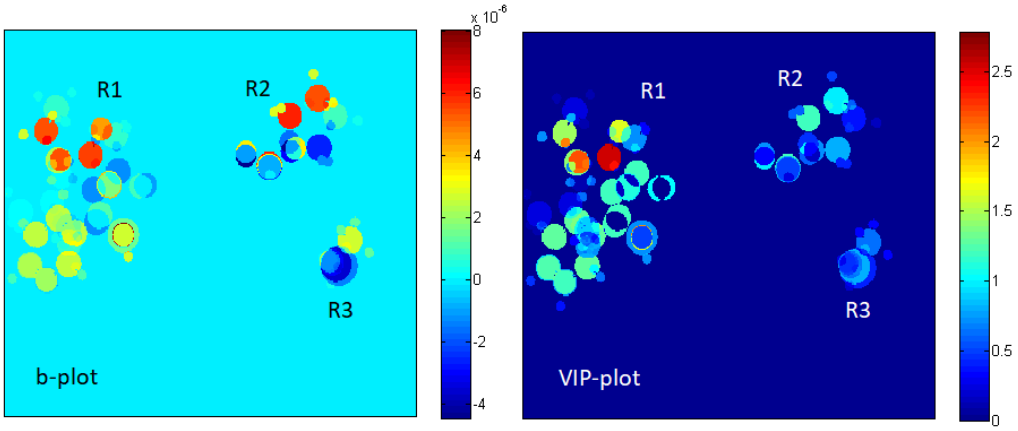


Figure 4. b-plot and VIP plot for model B.

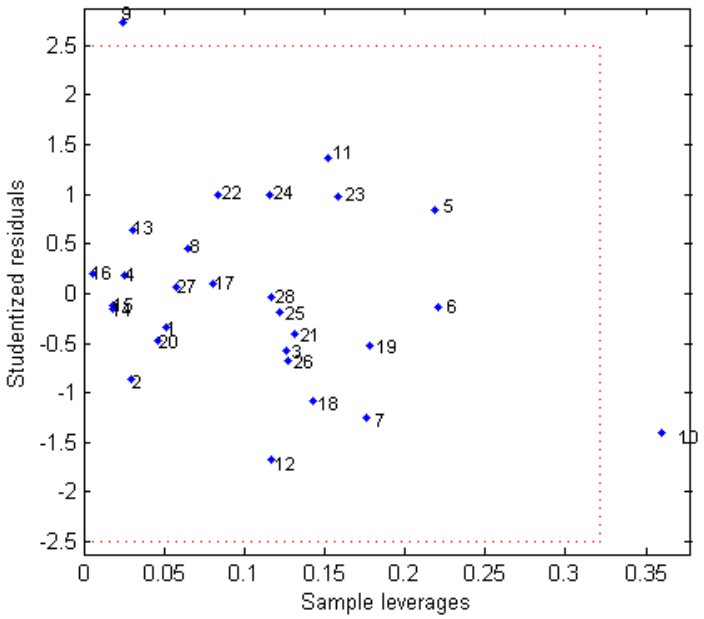


Figure 5. Applicability domain obtained from the William´s plot.

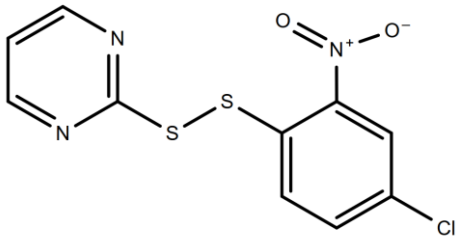


Figure 6. Proposed disulfide compound. The structure contains a six-membered ring with two endocyclic Nitrogen atoms at R1, a nitro group at R2, and a Chlorine at R3.