JOYCE KAROLINE DARÉ

INFLUENCE OF SPATIAL INFORMATION CODING INTO 2D AND 3D DESCRIPTORS FOR QSAR MODELLING PURPOSES

LAVRAS-MG
2023

JOYCE KAROLINE DARÉ

INFLUENCE OF SPATIAL INFORMATION CODING INTO 2D AND 3D DESCRIPTORS FOR QSAR MODELLING PURPOSES

Tese 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 Doutor.

Prof. Dr. Matheus Puggina de Freitas
Orientador

LAVRAS-MG
2023

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.
Influence of spatial information coding into 2D and 3D descriptors for QSAR modelling purposes / Joyce Karoline Daré. - 2023.
86 p. : il.

Orientador(a): Matheus Puggina de Freitas.

Tese (doutorado) - Universidade Federal de Lavras, 2023.
Bibliografia.

1. MIA-QSAR. 2. QSAR-3D. 3. Representações moleculares. I. de Freitas, Matheus Puggina. II. Título.

JOYCE KAROLINE DARÉ

INFLUÊNCIA DA CODIFICAÇÃO DE INFORMAÇÕES ESPACIAIS EM DESCRITORES 2D E 3D PARA FINS DE MODELAGEM QSAR

INFLUENCE OF SPATIAL INFORMATION CODING INTO 2D AND 3D DESCRIPTORS FOR QSAR MODELLING PURPOSES

Tese 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 Doutor.

APROVADA em 16 de fevereiro de 2023.
Dra. Daiana Teixeira Mancini, UFLA
Dra. Elaine Fontes Ferreira da Cunha, UFLA
Dr. João Eustaquio Antunes, UFJF
Dr. Teodorico Ramalho de Castro, UFLA

Prof. Dr. Matheus Puggina de Freitas
Orientador

LAVRAS-MG
2023

À minha família pelo amor incondicional e sacrifícios feitos para que eu chegasse até aqui.

Ao meu esposo Wilson por todo cuidado, ajuda, compreensão e incentivo nos momentos mais difíceis desta jornada.

À minha doce filha Cecília, que mesmo tão pequena já tem me ensinado tanto.

A Deus e à Nossa Senhora por serem meu sustento, fonte de paz e equilíbrio.

Dedico

AGRADECIMENTOS

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

Ao programa de pós-graduação em Agroquímica pela formação acadêmica proporcionada.

Ao professor Matheus pelo incentivo, direcionamento e disposição para ajudar.

À banca pela disponibilidade e contribuições para este trabalho.

A todos colegas do laboratório de Química Computacional pela ajuda e ensinamentos.

À minha família pelo apoio incondicional e a torcida.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

MUITO OBRIGADA!

“É justo que muito custe o que muito vale.” (Santa Teresa D’Ávila)

RESUMO

A análise multivariada de imagens aplicada em QSAR (MIA-QSAR) é uma técnica baseada no tratamento de imagens bidimensionais resultantes das projeções de estruturas moleculares perfeitamente congruentes (alinhadas) e com geometrias não-otimizadas. Destaca-se por ser uma metodologia que balanceia simplicidade e eficácia na geração de modelos de predição de propriedades biológicas e físico-químicas. Por se tratar de uma técnica 2D, a MIA-QSAR não codifica, de forma eficiente, informações espaciais em seus descritores moleculares. Nesse sentido, e tendo em mente o papel fundamental das informações 3D para modelagem e descrição de propriedades biológicas/físico-químicas de moléculas, o presente trabalho tem como objetivo principal investigar diferentes estratégias para codificar e modelar esse tipo de informação em descritores MIA-QSAR, bem como avaliar o papel da conformação em uma abordagem QSAR originalmente tridimensional. Três diferentes fontes de descritores foram propostas para codificação de informação 3D nos descritores MIA-QSAR: (I) imagens de projeções 2D de compostos com geometrias previamente otimizadas; (II) imagens de "fatias moleculares" obtidas após o escaneamento, ao longo de um dos eixos cartesianos, de moléculas com geometrias otimizadas; e (III) imagens da face frontal, direita e superior de estruturas químicas, com geometrias otimizadas, dispostas dentro de uma caixa teórica. Para a modelagem dos dados, duas ferramentas robustas de regressão multivariada foram empregadas: para os descritores oriundos das projeções 2D fez-se uso do método de máquina de vetores de suporte para regressão (SVR); para as duas outras estratégias o método dos mínimos quadrados parciais multilinear (N-PLS) foi empregado. As três rotinas foram empregadas em três diferentes grupos de compostos: uma série de moléculas com atividade contra o vírus da hepatite C (anti-HCV), outra com ação contra o coronavírus causador da síndrome respiratória aguda severa (SARS-CoV), e um grupo com atividade anti-HIV (vírus da imunodeficiência humana). Como resultado, parâmetros de boa qualidade, tanto de validação interna quanto externa, foram obtidos nas três estratégias e resultados estatísticos de correlação foram, no mínimo, similares aos reportados em outros estudos envolvendo os mesmos conjuntos de dados. No entanto, o risco de correlação casual não pôde ser excluído, como demonstrado por testes de randomização do bloco Y. Dessa forma, a metodologia MIA-QSAR tradicional, que faz uso de imagens de subestruturas farmacofóricas perfeitamente congruentes e com geometrias não-otimizadas, mostrou-se mais eficiente que as estratégias que codificaram informação tridimensional na modelagem. Para avaliar o papel da conformação em uma técnica QSAR originalmente 3D, foram comparados modelos construídos com variáveis que codificam aspectos tridimensionais completamente descritos, obtidos de estruturas químicas previamente ancoradas em seu alvo biológico, com descritores em que esse tipo de informação é suprimido (estruturas planas) ou apenas parcialmente descrito (estruturas químicas com geometrias computacionalmente otimizadas). Como resultado, os parâmetros de validação indicaram que a robustez dos modelos QSAR parece estar mais associada ao alinhamento das estruturas do que ao nível de detalhamento dos aspectos tridimensionais codificados pelos descritores moleculares.

Palavras-chave: MIA-QSAR. Informação tridimensional. Projeções 2D. Fatias moleculares. Faces moleculares. QSAR-3D.

ABSTRACT

The multivariate image analysis applied to QSAR (MIA-QSAR) is a technique based on the treatment of bidimensional images resulting from the projections of perfectly congruent, non-optimized geometries. It stands out for being a methodology that balances simplicity and efficiency in the generation of prediction models of biological/physicochemical properties. Because MIA-QSAR is a 2D technique, it does not efficiently encode spatial information in its molecular descriptors. In this sense, and keeping in mind the key role of 3D information for modeling and describing biological/physicochemical properties of molecules, the present work aims to investigate different strategies to encode and model this type of information in MIA-QSAR descriptors, as well as to evaluate the role of conformation in an originally tridimensional QSAR approach. Three different sources of descriptors have been proposed to codify 3D information into the MIA-QSAR descriptors: (I) images of 2D projections of compounds with previously optimized geometries; (II) images of "molecular slices" obtained after scanning molecules, with optimized geometries, along one of the cartesian axes; and, (III) images from the front, right and top faces of chemical structures, with optimized geometries, placed inside a theoretical box. For data modeling, two robust multivariate regression tools were used: for 2D projection descriptors, the support vector machine applied to regression (SVR) method was employed; for the other two strategies, the multilinear partial least squares (N-PLS) method was chosen. The three routines were applied to three different groups of compounds, a series of molecules with activity against the hepatitis C virus (anti-HCV), another with action against the coronavirus that causes severe acute respiratory syndrome (SARS-CoV), and a group with anti-HIV activity (human immunodeficiency virus). As a result, high quality parameters for both internal and external validation were achieved in all three strategies, and the statistical results of correlation were at least similar to those earlier reported for these series of compounds. Nevertheless, the risk of chance correlation could not be excluded as demonstrated by y-randomization tests. Accordingly, traditional MIA-QSAR method that uses perfectly congruent, non-optimized geometries of pharmacophoric substructures as images is still more efficient than the attempts to incorporate 3D information in the modelling. To evaluate the role of conformational information in an originally 3D-QSAR technique, one compared models built with variables codifying tridimensional aspects fully described, obtained from chemical structures previously docked in their biological target, with descriptors in which this type of information is either suppressed (flat structures) or only partially described (chemical structures with computationally optimized geometries). As a result, the validation parameters indicated that the robustness of the QSAR models seems to be more related to the alignment aspects of the structures than to the level of detail of tridimensional aspects encoded by the molecular descriptors.

Key-words: MIA-QSAR. Tridimensional information. 2D projections. Molecular slices. Molecular faces. 3D-QSAR.

LISTA DE SIGLAS

AFMoC – Adaptação de campos para comparação molecular

AMBER – Construção assistida de modelos com refinamento de energia

AM1 – Austin modelo 1

CHELPG – Cargas de potenciais eletrostáticos utilizando um método baseado em *Grid*

CoMMA – Análise comparativa do momento molecular

CoMFA – Análise comparativa de campo molecular

CoMSIA – Análise comparativa de índices de similaridade

CV – Validação cruzada

DYLOMMS – Sistema de modelagem molecular orientada por rede dinâmica

FFD – Planejamento fatorial fracionário

HCV – Vírus da hepatite C

HIFA – Análise do campo intermolecular HINT

HIV – Vírus da imunodeficiência humana

LOOCV – Validação cruzada do tipo *leave-one-out*

LV – Variáveis latentes

MIA – Análise multivariada de imagens

MIA-QSAR – Análise multivariada de imagens aplicada a QSAR

MIF – Campo de interação molecular

M^{pro} – Protease principal

N-PLS – Método dos mínimos quadrados parciais multinear

PDB – Banco de dados de proteína

PCA – Análise dos componentes principais

PLS – Método dos mínimos quadrados parciais

QPLS – Mínimos quadrados parciais quadráticos

QSAR – Relação quantitativa entre estrutura e atividade

QSPR – Relação quantitativa entre estrutura e propriedade

RGB – Sistema baseado nas cores vermelha, verde e azul.

RMSE – Raiz do erro quadrático médio

SARS-CoV – Síndrome respiratória aguda severa causada por coronavírus

SDEC – Erro de desvio padrão no cálculo

SDEP – Erro de desvio padrão na predição

SRD – Definição da região inteligente

SVM – Máquina de vetores de suporte

SVR – Máquina de vetores de suporte aplicada à regressão

2D – Bidimensional

3D – Tridimensional

LISTA DE SÍMBOLOS

C – Parâmetro referente à função custo

ε – Eletronegatividade de Pauling

Ɛ – Parâmetro referente à função de perda insensitiva

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

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

q^2 – Coeficiente de regressão resultante da validação cruzada

r_{atm} – Raio atômico

r^2 – Coeficiente de regressão

r^2_m – Coeficiente de regressão modificado

r^2_{test} – Coeficiente de regressão externo

${}^cr^2_p$ – Parâmetro para análise do risco de correlação casual

$r^2_{y\text{-}rand}$ – Coeficiente de regressão resultante do teste de randomização do bloco Y

¯ – Média

SUMÁRIO

PRIMEIRA PARTE

1 INTRODUÇÃO

Desde sua criação, no início da década de 1960, a técnica de modelagem molecular QSAR (do inglês *Quantitative Structure-Activity Relationship*) evoluiu substancialmente. Passou de simples modelos de regressão lineares a modelos robustos que incluem diferentes tipos de descritores moleculares, novas ferramentas matemáticas para modelagem, e representações moleculares cada vez mais próximas de sistemas reais. As diferentes rotinas que surgiram com o intuito de gerar modelos matemáticos eficientes para predição de propriedades biológicas/físico-químicas e interpretação estrutural são normalmente classificadas de 1D–7D, de acordo com o tipo de informação codificada pelos descritores moleculares. Cada abordagem possui vantagens e desvantagens que resultam da complexidade e da acurácia da técnica.

A metodologia MIA-QSAR (do inglês *Multivariate Image Analysis applied to QSAR*) destaca-se por balancear simplicidade de execução e eficácia dos modelos de predição. É uma técnica que se baseia no tratamento de imagens 2D para a obtenção de descritores moleculares. Por se tratar de uma abordagem bidimensional, originalmente, ela inclui informações topológicas e topoquímicas de moléculas, isto é, informações acerca da forma molecular, presença de centros estereogênicos, etc., mas não inclui, de forma eficiente, informações de cunho espacial, tal como conformação molecular. Sendo assim, uma crítica significativa a respeito da técnica MIA-QSAR e às abordagens 2D em geral se baseia na falta de representação conformacional necessária para explicar a bioatividade de moléculas.

Dessa forma, sabendo o papel fundamental do aspecto conformacional na atividade biológica/físico-química de moléculas, recentemente, esforços se voltaram para a inclusão desse tipo de informação nos descritores MIA-QSAR. Dois estudos foram conduzidos para a obtenção de descritores 3D: no primeiro caso foi proposto a utilização de imagens 2D obtidas de projeções das geometrias moleculares computacionalmente otimizadas como fonte de descritores espaciais (DARÉ; RAMALHO; FREITAS, 2018); no segundo estudo foi sugerida a utilização de imagens 2D obtidas das projeções de moléculas computacionalmente ancoradas em seu alvo biológico, *i.e.*, possíveis conformações bioativas (DARÉ; SILVA; RAMALHO; FREITAS, 2020). Os descritores moleculares obtidos em cada caso foram correlacionados à atividade biológica através do método de regressão dos mínimos quadrados parciais (PLS, do inglês *Partial Least Squares*). Como resultado, em ambas propostas os modelos de predição gerados mostraram-se inferiores àqueles construídos seguindo a rotina tradicional MIA-QSAR,

na qual são geradas imagens a partir de moléculas perfeitamente congruentes e com geometrias não-otimizadas.

Como justificativa para a realização deste trabalho, considerando que ambas as tentativas para a codificação e modelagem de informação 3D em descritores MIA-QSAR falharam e tendo em mente a importância da informação espacial para modelagem da propriedade biológica/físico-química de moléculas, verifica-se a necessidade de explorar novas rotas para lidar com esse tipo de informação na técnica MIA-QSAR em prol do seu avanço. Sendo assim, o presente trabalho visa investigar diferentes estratégias para codificar e modelar informação tridimensional em descritores MIA-QSAR, bem como avaliar os efeitos dessa inclusão no desempenho dos modelos construídos a partir dessas estratégias. Uma vez avaliado o papel da informação conformacional em uma abordagem 2D, estende-se o mesmo raciocínio para o QSAR-3D, ou seja, busca-se também compreender o papel da conformação na construção de modelos QSAR-3D, abordagem na qual, por padrão, a informação espacial já é eficientemente codificada. Portanto, o presente trabalho não só traz contribuições para a técnica MIA-QSAR, mas também amplia a compreensão acerca da influência da informação conformacional na modelagem QSAR tridimensional.

Três diferentes rotinas são apresentadas para a codificação e modelagem de informação 3D em descritores MIA-QSAR, as quais incluem três diferentes fontes de descritores moleculares: imagens de projeções 2D, fatias (*slices*) e faces de moléculas com geometrias otimizadas e inseridas dentro de uma caixa teórica; e dois métodos de regressão distintos: PLS trilinear e máquina de vetores de suporte para regressão (SVR, do inglês *Support Vector Machine applied to Regression*). Para validação e comparação das diferentes abordagens, três conjuntos de moléculas foram propostos para estudo: séries de compostos com atividade anti-HCV (do inglês *Hepatitis C Virus*), anti-SARS-CoV (do inglês *Severe Acute Respiratory Syndrome Coronavirus*), e anti-HIV (do inglês *Human Immunodeficiency Virus*).

Por fim, para avaliar o papel da conformação em uma abordagem QSAR tridimensional, são comparados modelos QSAR-3D construídos com variáveis que codificam aspectos tridimensionais completamente descritos, obtidos a partir de estruturas químicas previamente ancoradas em seu alvo biológico, com variáveis em que esse tipo de informação é suprimido (estruturas planas) ou apenas parcialmente descrito (estruturas químicas com geometrias computacionalmente otimizadas). A série de compostos com atividade anti-SARS-CoV foi escolhida para tal análise.

2 OBJETIVOS

2.1 Objetivo geral

Em suma, o presente trabalho tem como objetivo investigar diferentes estratégias para codificar e modelar informação tridimensional em descritores MIA-QSAR, bem como avaliar os efeitos dessa inclusão no desempenho dos modelos construídos a partir dessas estratégias, ou seja, busca-se avaliar o papel da informação conformacional em uma abordagem 2D. De forma semelhante, também busca-se compreender o papel da conformação na construção de modelos QSAR-3D, abordagem na qual, por padrão, a informação espacial já se encontra eficientemente codificada.

2.2 Objetivos específicos

Para a codificação e avaliação da informação conformacional nos descritores MIA-QSAR, os seguintes objetivos específicos foram traçados:

a. Selecionar conjuntos de compostos químicos que tenham sido previamente submetidos à modelagem MIA-QSAR tradicional para fins de comparação;
b. Desenhar os compostos, otimizar suas geometrias em um nível de teoria apropriado e alinhar as estruturas otimizadas;
c. Selecionar um *software* que permita o escaneamento e captura de imagens de diferentes posições das moléculas sem a perda das seguintes características: uniformidade da cor ao longo de todo o átomo/ligação (cores sólidas) e tamanho dos átomos com base nos respectivos raios de van der Waals.
d. Gerar imagens de projeções bidimensionais das geometrias moleculares otimizadas e alinhadas;
e. Fazer o escaneamento das estruturas e capturar imagens de cada "fatia" (*slice*) molecular;
f. Gerar imagens das faces frontal, topo e lateral direita das moléculas, posicionadas dentro de uma caixa teórica;
g. Extrair os descritores das imagens mencionadas, selecioná-los e organizá-los para que sejam posteriormente submetidos a métodos de regressão para geração de modelos;
h. Construir modelos QSAR utilizando ferramentas de regressão adequadas e validá-los *via* métricas comumente utilizadas em análises MIA-QSAR tradicionais.

Para avaliação do papel da informação conformacional em uma técnica QSAR-3D, traçou-se os seguintes objetivos:

a. Escolher um conjunto de dados que contenha um alvo biológico bem definido;
b. Encontrar um *software* de QSAR-3D livre e gratuito e se familiarizar com a ferramenta em questão;
c. Desenhar as estruturas planas de forma que as subestruturas coincidentes estejam perfeitamente congruentes, alinhá-las e submetê-las à rotina de QSAR-3D para obtenção de um modelo de predição para essas estruturas em que a informação conformacional é suprimida;
d. Desenhar as estruturas tridimensionais, otimizar suas geometrias em um nível de teoria apropriado, alinhá-las e construir o modelo de QSAR-3D para essas estruturas em que a informação conformacional é parcialmente descrita;
e. Ancorar as moléculas previamente otimizadas em seu alvo biológico de forma a obter as possíveis estruturas bioativas; alinhar as estruturas resultantes e construir o modelo de QSAR-3D para essas estruturas em que a informação conformacional é melhor descrita;
f. Comparar os três modelos QSAR obtidos.

3 REFERENCIAL TEÓRICO

3.1 Modelagem QSAR: fundamentos e princípios

QSAR (do inglês *Quantitative Structure-Activity Relationship*) é uma abordagem computacional utilizada para modelagem de dados químicos, que tem como objetivo gerar modelos matemáticos capazes de correlacionar os chamados descritores moleculares – propriedades computadas a partir de estruturas moleculares – e a bioatividade de moléculas. Tais modelos são gerados com o intuito de empregá-los posteriormente para a predição da atividade biológica de novos compostos ou para projetá-los com as propriedades desejadas *via* modificações estruturais indicadas pela análise QSAR (MURATOV et al., 2020). Dessa forma, o QSAR tornou-se uma ferramenta indispensável ao desenvolvimento e otimização de fármacos, e sua aplicabilidade se estendeu também para análises envolvendo propriedades físico-químicas, o que introduziu o conceito de QSPR (do inglês *Quantitative Structure-Property Relationship*) (YOUSEFINEJAD; HEMMATEENEJAD, 2015).

A essência de tal metodologia está em assumir que compostos novos ou não-testados para uma determinada função que possuam características estruturais semelhantes às observadas nas moléculas utilizadas para construção do modelo QSAR/QSPR, as quais já possuem propriedade biológica/físico-química determinada experimentalmente, também apresentarão propriedades biológicas/físico-químicas semelhantes (NANTASENAMAT et al., 2009).

Muitas rotinas QSAR/QSPR surgiram desde sua criação no início da década de 1960. Evoluíram de simples modelos lineares, publicados por Hansch e Fujita (1964), a modelos robustos que incluem diferentes tipos de descritores moleculares, novas ferramentas matemáticas e representações moleculares tridimensionais, o que foi permitido graças ao uso de métodos baseados em campos moleculares como CoMFA (CRAMER; PATTERSON; BUNCE, 1988) e CoMSIA (KLEBE; ABRAHAM; MIETZNER, 1994), por exemplo. Normalmente, as diferentes abordagens QSAR/QSPR são classificadas de 1D a 7D, de acordo com o tipo de informação codificada por seus descritores moleculares.

A abordagem 1D consiste na forma mais simplificada da técnica, onde o foco está voltado às propriedades macroscópicas dos compostos, os quais são representados, normalmente, por fórmulas moleculares, ou seja, representações do tipo linear 1D. Exemplos de dados 1D são escalares, tais como massa molecular, número de átomos, dentre outros. Por outro lado, metodologias 2D fazem uso de representações planas das moléculas e seus

descritores codificam informações topológicas na forma de matrizes 2D (KIRALJ; FERREIRA, 2003). As abordagens 3D são as mais exploradas e focam em todas as propriedades dos átomos que constituem a representação espacial de uma molécula (características estéricas, eletrostáticas, lipofílicas etc.) (DAMALE et al., 2014). O conjunto de técnicas 4D, além de ter como objetivo a modelagem de propriedades 3D de compostos químicos, também visa a construção e análise do perfil conformacional dos ligantes; para tal, envolve uma série de passos, dentre os quais destacam-se a geração e utilização de múltiplas conformações, alinhamento e a consideração de múltiplos grupos de subestruturas (ROY; KAR; DAS, 2015).

A utilização de múltiplas topologias do ligante para o estudo da conformação, protonação e orientação é considerada uma adição de dimensão ao QSAR-4D e, portanto, é referida como QSAR-5D; essa nova dimensão é construída considerando um receptor criado a partir de uma série de simulações e hipóteses de ajuste induzido. Já as rotinas classificadas como 6D são assim referidas por incluírem a função de solvatação ao QSAR-5D, ou seja, as características do meio passam a ser consideradas, o que permite o estudo de interações não-covalentes entre ligante e receptor. Por fim, a mais recente das abordagens, 7D, inclui o receptor real ou dados de um modelo do receptor (ROY; KAR; DAS, 2015).

Dentre as diversas abordagens apresentadas anteriormente, Fujita e Winkler (2016) observaram duas linhas distintas de pesquisa que se estabeleceram ao longo dos anos: uma que permanece ligada às origens do QSAR, na qual o modelo pode ser considerado relativamente simples, linear e interpretável, denominada pelos autores como QSAR clássico ou “puro”; e outra que tem como foco principal a modelagem de conjuntos grandes de dados (*big data*) e/ou com ampla diversidade química, utilizando uma série de métodos de regressão/classificação robustos, com o intuito de gerar predições confiáveis das propriedades de novos compostos. Nesse segundo caso, a interpretação do modelo é normalmente inviável. Apesar da tensão existente entre as duas áreas, os autores deixam claro que ambas são importantes e têm seu papel na pesquisa (FUJITA; WINKLER, 2016).

A classe de QSAR voltada à predição, de certa forma, complementa as técnicas clássicas, uma vez que é apropriada para casos em que o cálculo de descritores físico-químicos não é prático, tais como situações envolvendo conjuntos de dados formados por estruturas não-congêneres e análises com conjuntos muito grandes de dados (FUJITA; WINKLER, 2016). Na Seção 3.2, discute-se uma técnica chamada *Deep Snap*, a qual exemplifica essa classe de QSAR, que se baseia em métodos robustos de aprendizagem de máquina para fins de predição. Nesse caso, em específico, a técnica envolve o uso de imagens multivariadas como fonte de descritores.

3.2 O uso de imagens multivariadas para modelagem QSAR

A análise de imagens digitais se iniciou ainda na década de 1960 no campo da ciência da computação para fins de sensoriamento remoto (extração de informações presentes em imagens de satélite) e medicina diagnóstica (radiologia, microscopia, etc.) (PRATS-MONTALBÁN; de JUAN; FERRER, 2011). Desde então, esse tipo de análise tem feito importantes contribuições em diversas áreas do conhecimento, tais como engenharia, agricultura, biologia, química, dentre outras. Na medicina em especial, a análise de imagens digitais avançou ainda mais com o advento de poderosas ferramentas de aprendizagem de máquina, tais como a técnica de aprendizagem profunda, máquina de vetores de suporte, redes neurais, etc., e hoje é frequentemente utilizada em sistemas de diagnósticos empregados para a detecção precoce de doenças complexas, tais como ressonância magnética por imagem (MRI), tomografia por emissão de pósitrons (PET), tomografia computadorizada (CT), dentre outras (RAZA; SINGH, 2021).

A análise de imagens digitais faz parte de uma área mais ampla chamada de processamento de imagens. Em termos gerais, o processamento de imagens tem como objetivo tratar a informação em uma imagem para melhorar sua qualidade visual ou para extrair informações úteis, o que é baseado em diferentes propriedades, tais como cor, forma, área, textura, etc. (CHANDA; MAJUMDER, 2011).

Em 1989, Esbensen e Geladi (1989) introduziram o conceito de análise multivariada de imagens (MIA), uma técnica para lidar com imagens que possuam mais de uma medição por píxel. A partir disso, a análise de imagem no campo da química tomou grandes proporções, uma vez que imagens podem codificar informações químicas e muitos métodos químicos podem gerar dados de imagem (microscopia, tomografia de raio-X, etc.) (ESBENSEN; GELADI, 1989). O fato de a análise de imagens ser uma técnica não destrutiva também contribuiu para difusão da mesma.

O processo de extração de características (descritores) de imagens, para fins de análise, retorna valores numéricos e/ou gráficos relacionados a essas características, que são utilizados normalmente para a classificação, detecção de defeitos, para predição de parâmetros de qualidade de produtos, etc. (PRATS-MONTALBÁN; de JUAN; FERRER, 2011).

Muitas ferramentas quimiométricas foram adaptadas para o tratamento de imagens, desde as mais simples até as chamadas imagens hiperespectrais (imagem em que um espectro completo é gerado para cada píxel). A introdução da quimiometria no campo de análise de imagens permitiu o surgimento de novas áreas, tais como: análise exploratória de imagem,

monitoramento estatístico multivariado de processo, e regressão multivariada de imagem (MIR, do inglês *Multivariate Image Regression*) (PRATS-MONTALBÁN; de JUAN; FERRER, 2011).

Trabalhos envolvendo imagens multivariadas de moléculas com aplicações em QSAR já foram publicados na literatura. Freitas, Brown e Martins (2005) estabeleceram a técnica denominada MIA-QSAR (do inglês *Multivariate Image Analysis applied to QSAR*), a qual tem como objetivo a construção de modelos de regressão com base em descritores extraídos de imagens bidimensionais de moléculas. Desde sua criação, a técnica MIA-QSAR passou por diversas melhorias e hoje conta, inclusive, com ferramentas de interpretação acopladas a geração dos modelos de predição (BARIGYE et al., 2016); na Seção 3.3, tal técnica é descrita em detalhes. Uma outra técnica envolvendo o uso de imagens, denominada *Deep Snap*, foi sugerida por Uesawa (2018), onde a estrutura molecular tridimensional é rotacionada 360° ao longo de cada um dos eixos cartesianos e *snapshots* são capturados de diferentes visões. No entanto, diferentemente da técnica MIA-QSAR, a *Deep Snap* é pensada em associação com a técnica de aprendizagem profunda (do inglês *Deep Learning*) para fins apenas de classificação; além disso, segundo o autor, tal metodologia dispensa a etapa de cálculo de descritores estruturais, pois a técnica de *Deep Learning* é capaz de extrair as características da imagem sem intervenção humana, propriedade conhecida como aprendizagem de representação de recursos (UESAWA, 2018). Em outras palavras, as entradas para o algoritmo de aprendizagem de máquina profundo são as imagens na íntegra, o que dispensa qualquer tipo de conversão requerida por outras técnicas.

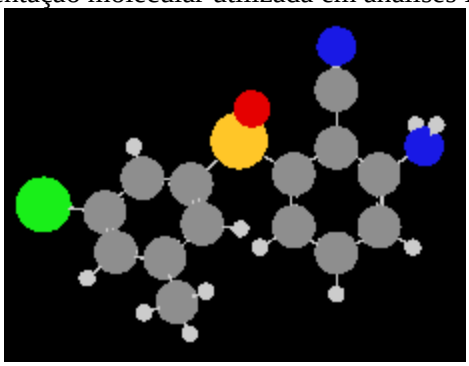
Em suma, a incorporação de imagens nas rotinas de análises químicas como fonte de informações tem se mostrado de grande valia e, associada às técnicas robustas de quimiometria e ao avanço dos recursos computacionais, tem permitido a exploração de novas fontes de descritores moleculares para construção de modelos QSAR eficientes, tal como apresentado pelas técnicas MIA-QSAR e *Deep Snap*.

3.3 Análise multivariada de imagens aplicada em QSAR (MIA-QSAR)

O método MIA-QSAR (do inglês *Multivariate Image Analysis applied to QSAR*) consiste em uma abordagem baseada no tratamento de imagens bidimensionais resultantes das projeções de estruturas moleculares perfeitamente congruentes (alinhadas) e com geometrias não-otimizadas. Em sua versão mais recente, os átomos são representados por círculos de

tamanho proporcional ao raio de van der Waals dos diferentes elementos e por cores sólidas que seguem o padrão RGB (FIGURA 3.1) (BARIGYE; FREITAS, 2016).

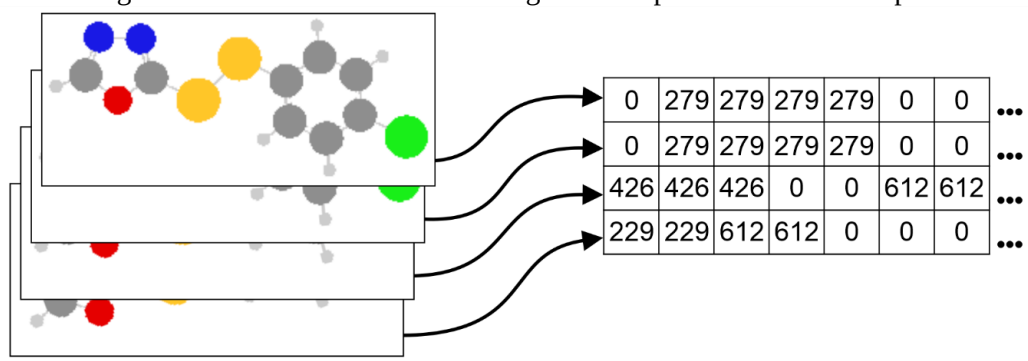
Figura 3.1 – Representação molecular utilizada em análises MIA-QSAR.



Fonte: Do autor (2023)

As informações estruturais químicas são extraídas em termos da propriedade ‘cor’, a qual consiste em um parâmetro resultante da intensidade dos canais RGB. Tal processo se dá através do desdobramento das imagens em uma matriz de píxeis, onde cada linha corresponde a uma molécula e cada coluna representa um píxel referente a uma posição na imagem, de forma que os píxeis se tornam os próprios descritores moleculares (FIGURA 3.2).

Figura 3.2 – Desdobramento das imagens bitmap em uma matriz de píxeis.



Fonte: Do autor (2023)

Os valores dos píxeis variam de 0 (cor preta) a 765 (cor branca), sendo que átomos de um mesmo elemento químico apresentam a mesma cor em toda a sua extensão, característica que precisa ser garantida pelo *software* utilizado para gerar as imagens. Tal exigência se faz necessária porque a etapa seguinte à geração da matriz de píxeis consiste na exclusão das colunas com variância nula (referente ao centro congênere e aos espaços vazios) e substituição dos valores referentes às cores dos píxeis por valores proporcionais a diferentes propriedades atômicas, tais como: eletronegatividade de Pauling, raio atômico, hidrofobicidade, etc.

(FREITAS; BARIGYE; FREITAS, 2015) (DARÉ; SILVA; FREITAS, 2017). Assim conclui-se a etapa de extração e seleção de descritores moleculares.

Vale ressaltar que, por se tratar de uma abordagem bidimensional, a MIA-QSAR, originalmente, codifica informações topológicas e topoquímicas das estruturas moleculares, mas não inclui, de maneira eficiente, características espaciais. (DARÉ; RAMALHO; FREITAS, 2018) (DARÉ; SILVA; RAMALHO; FREITAS, 2020).

Para modelagem dos dados obtidos, diferentes métodos multivariados de regressão têm sido empregados ao longo dos anos, dentre os quais destacam-se: o método dos mínimos quadrados parciais (PLS) (WOLD; SJÖSTRÖM; ERIKSSON, 2001) (NUNES; FREITAS, 2013), PLS multilinear (*n*-PLS) (BRO, 1996) (GOODARZI; FREITAS, 2009), e máquina de vetores de suporte (SVM) (VAPNIK; CORTES, 1995) (GOODARZI; FREITAS, 2011) (CORMANICH; GOODARZI; FREITAS, 2009). A Seção 3.3.1 se dedica ao detalhamento desses métodos.

Uma fonte de referência utilizada constantemente para validar o desempenho de modelos MIA-QSAR é o QSAR-3D, uma vez que, assim como explicado anteriormente, essa é a abordagem mais explorada. Com isso em mente, e considerando os objetivos do presente trabalho, a Seção 3.4 dedica-se ao detalhamento de tal abordagem.

3.3.1 O uso de métodos de aprendizagem de máquina em MIA-QSAR

Dentre os métodos de regressão citados anteriormente, o PLS (método dos mínimos quadrados parciais, do inglês *Partial Least Squares*) é um dos mais empregados em QSAR, o que também se aplica ao MIA-QSAR (BARIGYE et al., 2016) (NUNES; FREITAS, 2013) (GUIMARÃES et al., 2016). De maneira sucinta, trata-se de um método baseado na extração das chamadas variáveis latentes (LVs), a partir da matriz de covariância, para explicar o comportamento observado na variável resposta. Em tese, as LVs reúnem o máximo possível da variação do conjunto X (variáveis independentes) relacionada ao comportamento observado na variável Y (propriedade de interesse). Em outras palavras, o método dos mínimos quadrados parciais mede a covariância entre dois ou mais blocos de descritores (X e Y, por exemplo) e cria um novo conjunto de variáveis que é otimizado de forma a conter máxima covariância em relação à parte não explicada da variável dependente (PIROUZ, 2006).

Sua ampla aplicação em QSAR está relacionada, principalmente, ao fato desse método linear ser capaz de lidar com conjuntos de dados onde o número de descritores excede grandemente o número de observações experimentais, situação comum em análises envolvendo

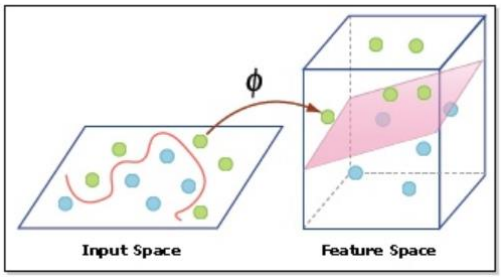
estudos de estrutura-atividade (CRAMER, 1993). Além disso, o PLS consegue lidar com conjuntos onde se têm o problema de multicolinearidade, bem como com dados contendo ruídos e valores faltando (PIROUZ, 2006).

Devido ao êxito em lidar com dados de primeira ordem, ou seja, situações em que para cada amostra/objeto diversas variáveis são determinadas, Bro (1996) propôs a extensão do método PLS de forma a contemplar também dados de ordens superiores, através do uso do algoritmo nomeado, por ele, N-PLS. Outros métodos que lidam com dados tridimensionais podem ser encontrados na literatura, tais como o modelo Tucker3 (T3) (TUCKER, 1966) e o modelo Candecomp/Parafac (CP) (HARSHMAN, 1970). No entanto, apenas o N-PLS maximiza a covariância com qualquer vetor multidimensional Y, por isso, é o mais adequado para fins de predição (HERVÁS et al., 2018).

Nesse sentido, Goodarzi e Freitas (2009) aplicaram tal método acoplado à rotina MIA-QSAR, com o intuito de comparar o desempenho desse método com o do PLS regular (bilinear) na modelagem de uma série de derivados de azol [compostos heterocíclicos aromáticos contendo um átomo de nitrogênio e um outro heteroátomo dispostos na posição-1,2 de um anel de cinco membros (NETTO; FREM; MAURO, 2008)]. Como resultado, os autores apontaram um melhor desempenho do método bilinear ao modelar os descritores MIA-QSAR, o que pode estar relacionado à perda de ajuste do N-PLS em relação ao método regular, apontada por Kiers (1991), devido às restrições mais severas impostas pelo método tridimensional, o que também é mencionado por Bro (1996).

O PLS, de uma forma geral, possui certas limitações que devem ser levadas em consideração. Talvez a mais evidente seja a incapacidade de modelar relações não-lineares entre os dados. Nesse sentido, métodos mais robustos, tal como SVM (do inglês *Support Vector Machine*), têm ganhado espaço em análises QSAR. A SVM consiste em um conjunto de métodos de aprendizagem supervisionada utilizado para fins de classificação, detecção de *outliers*, e regressão. Trata-se de uma metodologia cujo objetivo geral consiste em encontrar um hiperplano, em um espaço N-dimensional, que classifique distintamente os dados amostrais. Para isso, os vetores de entrada (amostras) são mapeados (*via* funções não lineares, ex.: função de base radial) para um espaço de alta dimensão, de forma a permitir a separação linear dos dados (ZHANG et al., 2016). A Figura 3.3 representa graficamente o funcionamento da técnica em questão.

Figura 3.3 – Representação do princípio de funcionamento da técnica SVM.



Fonte: Sethi (2020)

Em casos de classificação, o objetivo do método SVM é maximizar a margem de separação entre duas classes de forma a minimizar os erros de predição. Por outro lado, para casos de regressão, o objetivo é gerar um hiperplano que esteja o mais próximo da maior quantidade de dados possível (TRAFALIS; INCE, 2000). LS-SVM (do inglês *Least-Squares Support Vector Machine*) foi um método aplicado com sucesso em estudos envolvendo a construção de modelos de regressão MIA-QSAR (GOODARZI; FREITAS, 2011) (CORMANICH; GOODARZI; FREITAS, 2009) e, inclusive, apresentou desempenho superior ao PLS e N-PLS.

3.4 Construção de modelos QSAR-3D

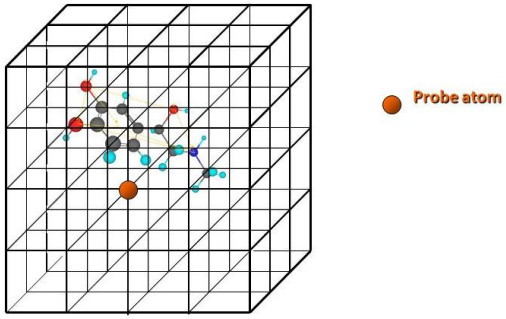
A terminologia QSAR-3D engloba todas as rotinas QSAR que correlacionam propriedades macroscópicas de interesse com descritores computados a partir de representações espaciais (tridimensionais) de moléculas (VERMA; KHEDKAR; COUTINHO, 2010). Devido ao surgimento de um grande número de técnicas, alguns critérios de classificação foram criados para agrupá-las em subcategorias. O primeiro critério consiste no tipo de informação utilizado para gerar o modelo QSAR, nesse sentido, tem-se o QSAR-3D baseado em ligante (CoMFA, CoMSIA, COMPASS, etc.) e baseado no receptor (COMBINE, AFMoC, HIFA etc.); um segundo critério se baseia na questão de alinhamento, tem-se o QSAR-3D baseado em alinhamento (CoMFA, CoMSIA, HIFA, etc.) e independente de alinhamento (COMPASS, CoMMA, etc.); o último critério se dá com base na técnica quimiométrica empregada para correlacionar propriedades estruturais e atividades, desse critério deriva-se as abordagens denominadas QSAR-3D linear (CoMFA, CoMSIA, COMMA, etc.) e não-linear (COMPASS e QPLS) (VERMA; KHEDKAR; COUTINHO, 2010).

Dentre as diferentes vertentes de técnicas 3D apresentadas, aquelas baseadas em campos de interação molecular (MIFs) tridimensionais, tais como CoMFA (do inglês *Comparative*

Molecular Field Analysis) (CRAMER; PATTERSON; BUNCE, 1988) e CoMSIA (do inglês *Comparative Similarity Indices Analysis*) (KLEBE; ABRAHAM; MIETZNER, 1994), estão entre as mais utilizadas. A técnica CoMFA surgiu como uma modificação do sistema de modelagem molecular orientado à malha, DYLOMMS (do inglês *DYnamic Lattice- Oriented Molecular Modelling System*), o qual envolvia a utilização de PCA (do inglês *Principal Component Analysis*) para a extração de vetores a partir de campos de interação molecular e posterior correlação desses com atividades biológicas. Cramer *et al.* (1988) modificaram a técnica DYLOMMS combinando duas outras técnicas disponíveis, GRID e PLS, e assim dá origem ao CoMFA, uma poderosa ferramenta de QSAR-3D que se tornou um protótipo para as demais técnicas 3D (VERMA; KHEDKAR; COUTINHO, 2010).

De forma geral, uma rotina QSAR-3D se inicia com a seleção ou *design* de representações de estruturas tridimensionais de moléculas ligantes, as quais, em seguida, passam por um processo de refinamento baseado em geometria e valores de energia. Os métodos mais utilizados para otimização de energia são os baseados na física clássica (mecânica molecular) e os de estrutura eletrônica (semiempíricos e mecânica quântica). O passo seguinte consiste na geração de um conjunto de confôrmeros para cada molécula. Dentre os múltiplos confôrmeros, representantes específicos, portando bioatividade, são procurados e selecionados para o estudo, o que é feito utilizando-se os chamados métodos baseados em conhecimento, através das medições de afinidade de ligação entre os confôrmeros e o receptor. Posteriormente, os compostos bioativos selecionados passam por uma etapa de alinhamento automática ou manual. Em seguida, as estruturas alinhadas são inseridas no interior de uma estrutura teórica em treliça e campos estéricos e eletrostáticos são computados através do posicionamento de diferentes grupos de sonda em todas as interseções da estrutura em treliça (FIGURA 3.4). Os valores de campo ou energia de interação calculados representam propriedades físico-químicas e biológicas de uma molécula. Após a obtenção desses descritores moleculares, métodos estatísticos [ex.: PLS, MLR (*Multiple Linear Regression*), *Cluster Analysis*, etc.] são utilizados para seleção e modelagem dos dados de forma a gerar um modelo de predição eficiente e que correlacione estrutura e atividade (DAMALE et al., 2014) (AKAMATSU, 2002).

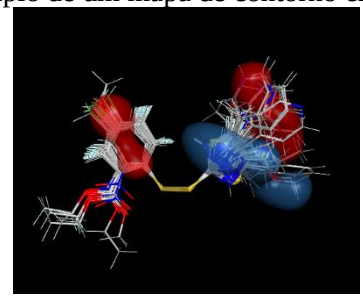
Figura 3.4 – Representação de uma estrutura teórica em treliça utilizada para cálculo dos campos de interação molecular.



Fonte: Reynolds (2016)

No caso específico do CoMFA, a técnica de PLS é utilizada para identificação e extração de informação química relacionada à atividade biológica a partir dos valores de campo calculados. A equação de correlação resultante é composta por variáveis que individualmente consistem na combinação linear dos descritores da estrutura em treliça originais. Para interpretação visual de tal equação, a técnica de CoMFA apresenta os resultados do PLS na forma de mapas de contorno das variáveis de campo correspondentes em cada interseção da estrutura em treliça; regiões favoráveis e desfavoráveis que estão consideravelmente associadas à atividade biológica são representadas nesses gráficos de acordo com um esquema de cores padrão; a Figura 3.5 exemplifica um mapa de contorno eletrostático; as regiões em vermelho indicam regiões onde grupos retiradores de elétrons favorecem o aumento da propriedade biológica modelada e as regiões em azul, onde grupos doadores de elétrons exercem tal função (VERMA; KHEDKAR; COUTINHO, 2010).

Figura 3.4 – Exemplo de um mapa de contorno eletrostático.



Fonte: Do autor (2023)

Existe uma série de ferramentas disponíveis para a implementação de modelos QSAR-3D baseados em campos de interação molecular, dentre as quais destaca-se o programa desenvolvido por Tosco e Balle (2011) chamado Open3DQSAR, por se tratar de uma

ferramenta gratuita e de código aberto, cujo objetivo consiste na investigação de moléculas através da análise quimiométrica de alto rendimento de MIFs.

Para a construção de um modelo de QSAR-3D utilizando a ferramenta Open3D-QSAR, inicialmente, o usuário deve importar o arquivo contendo todas as estruturas químicas em um dos formatos aceitos pelo programa. Em seguida, faz-se necessária a obtenção dos MIFs, os quais podem ser calculados internamente. Um campo estérico é computado com base nos parâmetros de van der Waals do campo de força AMBER FF99 e um campo eletrostático é obtido com base em um modelo de carga. O campo estérico é computado de acordo com o potencial de Lennard-Jones entre os *n* átomos da molécula e uma sonda de carbono sp^3. Enquanto isso, o campo eletrostático é computado somando as interações de Coulomb entre uma sonda carregada positivamente e os *n* átomos da molécula (TOSCO; BALLE, 2011).

Depois de reunidos os MIFs, o conjunto de dados pode ser dividido em treinamento e teste, de forma que a validação externa possa ser realizada posteriormente. Uma série de operações de pré-tratamento dos dados estão disponíveis para que o usuário remova possíveis variáveis não-informativas. Também estão disponíveis operações para transformações dos dados, caso necessário. Uma vez realizadas as operações de pré-tratamento e transformações dos dados, um modelo PLS pode ser construído com o algoritmo NIPALS, após a escolha do número de variáveis latentes por parte do usuário. A capacidade preditiva do modelo pode ser atestada por validação cruzada e validação externa. A robustez da correlação pode ser também avaliada por *progressive scrambling*. Por fim, o programa também reúne opções para agrupamento e seleção de variáveis, que são procedimentos eficientes para a construção de modelos robustos e preditivos (TOSCO; BALLE, 2011).

A técnica de QSAR é amplamente utilizada em conjunto com uma outra técnica de modelagem molecular chamada de ancoramento molecular. Tal técnica permite a predição dos diversos modos de ligação e afinidades envolvidos nos eventos de reconhecimento molecular (DU et al., 2016). Sendo assim, uma molécula proposta em um estudo de QSAR para o desempenho de uma determina atividade biológica pode ser avaliada no interior de seu respectivo alvo biológico, caso esse esteja disponível. Uma breve descrição da técnica é dada na seção 3.5.

3.5 Ancoramento molecular

A técnica de ancoramento molecular tem como objetivo principal a identificação e quantificação do modo de interação entre uma molécula ligante e seu respectivo receptor, de

modo que a atividade do receptor seja inibida ou potencializada (MORGON; COUTINHO, 2007). Tal processo se dá através da busca de poses (direção e orientação) do ligante no sítio ativo do receptor, de forma que cada pose receba uma pontuação que indica a probabilidade de ligação com o receptor (CAVASOTTO; AUCAR; ADLER, 2018).

A busca pelas poses, bem como pelas conformações que o ligante poderá assumir é executada pelo chamado algoritmo de busca. A probabilidade de ligação é quantificada pela função de pontuação, também chamada de função *score*. Tal função 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). Existem diferentes tipos de funções *score* e algoritmos de busca. Dentre as funções de pontuação, aquelas baseadas em campo de força estão entre as mais comuns. Quanto ao algoritmo de busca, o algoritmo genético é frequentemente empregado.

No presente trabalho, a ferramenta Glide (FRIESNER, et al., 2004) foi empregada para obtenção das possíveis conformações bioativas do grupo de compostos anti-SARS-CoV (WANG, et al., 2017). A função de pontuação em questão é baseada no campo de força OPLS (do inglês *Optimized Potentials for Liquid Simulations*). Já para a busca das poses, a ferramenta utiliza uma série de filtros hierárquicos para procurar as possíveis localizações do ligantes na região do sítio ativo do receptor (FRIESNER, et al., 2004).

4 CONSIDERAÇÕES GERAIS

Em suma, importantes avanços para a codificação de informação conformacional em descritores oriundos da metodologia MIA-QSAR foram feitos; no entanto, a maior eficiência da abordagem tradicional (2D) foi reforçada frente às modificações aqui propostas. Além disso, a análise do papel da conformação em uma abordagem 3D, realizada neste trabalho, demonstra a maior dependência da técnica em relação ao alinhamento do que aos tipos de informação conformacional considerados.

REFERÊNCIAS

AKAMATSU, M. Current state and perspectives of 3D-QSAR. **Current Topics in Medicinal Chemistry**, v. 2, p. 1381-1394, 2002.

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

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, p. 64–77, 2016.

BRO, R. Multiway calibration multilinear PLS. **Journal of Chemometrics**, v. 10, p. 47–61, 1996

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

CHANDA, B.; MAJUMDER, D. D. **Digital image processing and analysis**. 2.ed. New Delhi: PHI Learning Pyt. Ltd**.**, 2011.

CORMANICH, R. A.; GOODARZI, M.; FREITAS, M. P. Improvement of multivariate image analysis applied to quantitative structure–activity relationship (QSAR) analysis by using wavelet-principal component analysis ranking variable selection and least-squares support vector machine regression: QSAR study of checkpoint kinase WEE1 inhibitors. **Chemical Biology & Drug Design**, v. 73, p. 244–252, 2009.

CRAMER III, R. D.; PATTERSON, D. E.; BUNCE, J. D. Comparative molecular-field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. **Journal of the American Chemical Society**, v. 110, p. 5959−5967, 1988.

CRAMER III, R. D. Partial least squares (PLS): its strengths and limitations. **Perspectives in Drug Discovery and Design**, v. 1, p. 269–278, 1993

DAMALE, M. G.; HARKE, S. N.; KHAN, F. A. K.; SHINDE, D. B.; SANGSHETTI, J. N. Recent Advances in Multidimensional QSAR (4D-6D): A Critical Review. **Mini-Reviews in Medicinal Chemistry**, v. 14, p. 35–55, 2014.

DARÉ, J. K.; RAMALHO, T. C.; FREITAS, M. P. 3D perspective into MIA-QSAR: A case for anti-HCV agentes. **Chemical Biology & Drug Design**, v. 93, p. 1096–1104, 2018.

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

DARÉ, J. K.; SILVA, D. R.; RAMALHO, T. C.; FREITAS, M. P. Conformational fingerprints in the modelling performance of MIA-QSAR: a case for SARS-CoV protease inhibitors. **Molecular Simulation**, v. 46, p. 1055–1061, 2020.

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, p. 67–86, 1989.

FREITAS, M. R.; BARIGYE, S. J.; FREITAS, M. P. Coloured Chemical image-based models for the prediction of soil sorption of herbicides. **Royal Society of Chemistry Advances**, v. 5, p. 7547–7553, 2015.

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

FRIESNER, R. A.; BANKS, J. L.; MURPHY, R. B; HALGREN, T. A.; KLICIC, J. J.; MAINZ, D. T.; REPASKY, M. P.; KNOLL, E. H.; SHELLEY, M.; PERRY, J. K.; SHAW, D. E.; FRANCIS, P.; SHENKIN, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. **Journal of Medicinal Chemistry**, v. 47, p. 1739–1749, 2004.

FUJITA, T.; WINKLER, D. A. Understanding the Roles of the “Two QSARs”. **Journal of Chemical Information and Modeling**, v. 56, p. 269–274, 2016

GOODARZI, M.; FREITAS, M. P. On the use of PLS and N-PLS in MIA-QSAR: Azole antifungals. **Chemometrics and Intelligent Laboratory Systems**, v. 96, p. 59–62, 2009.

GOODARZI, M.; FREITAS, M. P. MIA-QSAR Coupled to different regression methods for the modeling of antimalarial activities of 2-aziridinyl and 2,3-*bis*-(aziridinyl)-1,4-naphtoquinonyl sulfate and acylate derivatives. **Medicinal Chemistry**, v. 7, p. 645–654, 2011.

GUIMARÃES, M. C.; DUARTE, M. H.; SILLA, J. M.; FREITAS, M. P. Is conformation a fundamental descriptor in QSAR? A case for halogenated anesthetics. **Beilstein Journal of Organic Chemistry**, v. 12, p. 760–768, 2016.

HANSCH, C.; FUJITA, T. ρ-σ-π Analysis. A method for correlation of biological activity and chemical structure. **Journal of the American Chemical Society**, v. 86, p. 1616−1624, 1964.

HARSHMAN, R.A. Foundations of the parafac procedure: models and conditions for an "explanatory" multimodal factor analysis, **UCLA Working Papers in Phonetics**, v. 16, p. 1–84, 1970.

HERVÁS, D.; PRATS-MONTALBÁN, J. M.; LAHOZ, A.; FERRER, A. Sparse N-way partial least squares with R package sNPLS. **Chemometrics and Intelligent Laboratory Systems**, v. 179, p. 54–63, 2018.

KIERS, H. A. L. Hierarchical relations among three-way methods. **Psychometrika**, v. 56, p. 449–470, 1991.

KIRALJ, R.; FERREIRA, M. M. C. A priori molecular descriptors in QSAR: a case of HIV-1 protease inhibitors I. The chemometrics approach. **Journal of Molecular Graphics and Modelling**, v. 21, p. 435–448, 2003.

KLEBE, G.; ABRAHAM, U; MIETZNER, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. **Journal of Medicinal Chemistry**, v. 37, p. 4130−4146, 1994.

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.

MURATOV, E. N.; BAJORATH, J.; SHERIDAN, R. P.; TETKO, I. V.; FILIMONOV, D.; POROIKOV, V; OPREA, T. I.; BASKIN, I. I.; VARNEK, A; ROITBERG, A.; ISAYEV, O.; CURTALOLO, S.; FOURCHES, D.; COHEN, Y.; ASPURU-GUZIK, A.; WINKLER, D. A.; AGRAFIOTIS, D.; CHERKASOV, A.; TROPSHA, A. QSAR without borders. **Chemical Society Reviews**, v. 49, p. 3525–3564, 2020.

NANTASENAMAT, C.; ISARANKURA-NA-AYUDHYA, C.; NAENNA, T.; PRACHAYASITTIKUL, V. A practical overview of quantitative structure-activity relationship. **Experimental and Clinical Sciences Journal**, v. 8, p. 74–88, 2009.

NETTO; A. V. G.; FREM, R. C. G.; MAURO, A. E. A química supramolecular de complexos pirazólicos. **Química Nova**, v. 31, p. 1678–7064, 2008.

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.

PIROUZ, D. An overview of partial least squares. **Social Science Research Network Electronic Journal**, 2006.

PRATS-MONTALBÁN, J. M.; de JUAN, A.; FERRER, A. Multivariate image analysis: A review with applications. **Chemometrics and Intelligent Laboratory Systems**, v. 107, p. 1–23, 2011.

RAZA, K.; SINGH, N. K. A tour of unsupervised deep learning for medical image analysis. **Current Medical Imaging**, v. 17, p. 1059–1077, 2021.

REYNOLDS, M. **Quantitative structure activity relationships QSAR and 3D-QSAR.** Disponível em: https://slideplayer.com/slide/6283595. Acesso em: 14 mar. 2023.

ROY, K.; KAR, S.; DAS, R. N. Newer QSAR techniques. In: ROY, K.; KAR, S.; DAS, R. N. **Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment**. 1. ed., Academic Press, 2015. Cap. 9, p. 319–356.

SETHI, Alakh. **Support Vector Regression Tutorial for Machine Learning.** Disponível em: https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning. Acesso em: 28 out. 2020.

TRAFALIS, T. B.; INCE, H. Support vector machine for regression and applications to financial forecasting. **Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks**, v. 6, p. 348–353, 2000.

TOSCO, P.; BALLE, T. An open-source software aimed at high-throughput chemometric analysis of molecular interaction fields. **Journal of Molecular Modeling**, v. 17, p. 201 – 208, 2011.

TUCKER, L. R. Some mathematical notes on three-mode factor analysis, **Psychometrika**, v. 31, p. 279–311, 1966.

UESAWA, Y. Quantitative structure-activity relationship analysis using deep learning based on a novel molecular input technique. **Bioorganic & Medicinal Chemistry Letters**, v. 28, p. 3400–3403, 2018.

VAPNIK, V.; CORTES, C. Support vector networks. **Machine Learning**, v. 20, p. 273 – 297, 1995.

VERMA, J.; KHEDKAR, V. M.; COUTINHO, E. C. 3D-QSAR in drug design – A Review. **Current Topics in Medicinal Chemistry**, v. 10, p. 95 – 115, 2010.

WANG, L.; BAO, B.-B.; SONG, G.-Q.; CHEN, C.; ZHANG, X-M; LU, W. WANG, Z.; CAI, Y.; LI, S.; FU, S.; SONG, F.-H; YANG, H.; WANG, J.-G. 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.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v. 58, p. 109 – 130, 2001.

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

ZHANG, F.; DEB, C.; LEE, S. E.; YANG, J.; SHAH, K. W. Time series forecasting for building energy consumption using weighted support vector regression with differential evolution optimization technique. **Energy and Buildings**, v. 126, p. 94–103, 2016.

SEGUNDA PARTE – ARTIGOS

ARTIGO 1 – DIFFERENT APPROACHES TO ENCODE AND MODEL 3D INFORMATION IN A MIA-QSAR PERSPECTIVE

* Artigo apresentado na íntegra segundo as normas do periódico científico no qual foi publicado (*Chemometrics and Intelligent Laboratory Systems)*.

Different approaches to encode and model 3D information in a MIA-QSAR perspective

Joyce K. Daré, Matheus P. Freitas*

Departamento de Química, Universidade Federal de Lavras, 37200–900, Lavras–MG, Brazil

* Corresponding author.
E-mail address: matheus@ufla.br (Matheus Puggina de Freitas)

Abstract

Tridimensional information is a fundamental aspect for modelling and explaining biological/physicochemical properties. In this sense, the goal of this study was to explore different approaches for encoding this type of information into MIA-QSAR (Multivariate Image Analysis applied to Quantitative Structure-Activity Relationships) descriptors and to effectively model these new features. Originally, MIA-QSAR is a technique based on the treatment of 2D images of molecules. The approaches explored in this work were: (I) the use of 2D image projections of computationally optimized molecular geometries as a source of 3D information for a powerful machine learning method (support vector machine applied to regression); (II) the use of slice images obtained from the optimized molecules placed inside a theoretical box as a source of 3D descriptors for a multi-way regression method (trilinear PLS); and (III) the use of images viewed from different faces of the previous box as an alternative source of 3D MIA-QSAR descriptors. These strategies were applied in three different data sets comprising anti-HCV, anti-SARS-CoV, and anti-HIV compounds. Satisfactory parameters for both internal and external validation were achieved in all three models, and the statistical results of correlation were at least similar to those earlier reported for these series of compounds. Nevertheless, the risk of chance correlation could not be excluded as demonstrated by *y*-randomization tests. Whereas the traditional MIA-QSAR method, that uses perfectly congruent, non-optimized geometries of pharmacophoric substructures as images, is more efficient than 3D MIA-QSAR, the latter uses tridimensional digital objects as descriptors for the first time in QSAR for regression purposes.

Keywords: 3D information, MIA-QSAR, 2D image projections, Molecular slice images,

Molecule-in-a-box face images

1. Introduction

Quantitative structure-activity relationship (QSAR) modelling is a widely spread computational technique that has effectively contributed to the development process of new drugs and agrochemicals. Since its roots in the 1960s, QSAR has evolved substantially, from simple linear regression models, published by Hansch and Fujita [1,2], to very robust models that include different types of molecular descriptors, new mathematical methods and tridimensional (3D) molecular representation *via* the use of molecular field-based methods (*e.g.*, CoMFA [3] and CoMSIA [4]) [5].

With the rapid emergence of a substantial number of different QSAR routines, a classification system has been established based on the type of information encoded by the different molecular descriptors; the different approaches are classified from 1D to 6D, and the essence of these methodologies is well reported in the literature [6-8]. More recently, an alternative classification system has emerged based on the main goal of the QSAR routines: classical QSAR, which focuses in a mechanistic interpretation of the model, and a second class that relies on machine learning methods to generate models for forecasting purposes [5]. Both classes have their importance and roles; the classical techniques are appropriate for cases where the data set is small and comprises chemically similar structures, from which interpretable descriptors can be generated; on another hand, the machine learning based methods, for forecasting purposes, are suitable for large data sets with chemically varied structures, which generate more arcane descriptors [5]. Therefore, the purpose of the analysis and the nature of the molecular descriptors dictate the most appropriate QSAR approach to be employed.

Multivariate image analysis applied to QSAR (MIA-QSAR) modelling is a technique based on the treatment of bidimensional molecular representations for building robust predictive models [9]. It is classified as a typical 2D QSAR technique, *i.e.*, the molecular descriptors encode, mainly, relevant topological and topochemical information [10]. In a MIA-QSAR analysis, the different atoms are represented by circles with size proportional to their van der Waals radii and color defined based on the RGB system. The essence of the chemical structure is encoded in image elements called pixels (MIA-QSAR molecular descriptors); a pixel can be defined as the smallest image unit and, in the case of a RGB image (8 bits), its value varies from 0 to 765. The MIA-QSAR routine can be found in detail elsewhere [9,11-13].

Since its foundation in 2005, MIA-QSAR has evolved significantly and, currently, it includes an interpretation tool that helps to understand how structural features affect different response properties [13]. More recently, efforts have been made in order to encode spatial molecular

information into MIA-QSAR descriptors, leading the technique to a 3D approach [14,15], once it is well-known that tridimensional features play an important role in explaining and modelling biological/physicochemical properties. In the first study, it was proposed the use of 2D image projections of computationally optimized molecular geometries of a series of anti-HCV compounds, as a source of tridimensional information; in the second study, it was suggested the use of image projections of bioactive conformations of a series of anti-SARS-CoV molecules, obtained through molecular docking technique. However, due to the dissimilar coordinates of the previously proposed MIA-QSAR descriptors, it has been observed that simple regression methods, such as PLS (Partial least Squares), are not efficient in explaining the existent correlation between these new features and the biological responses observed [14,15].

Accordingly, the main goal of this work was to investigate new approaches for encoding 3D information into MIA-QSAR descriptors and to effectively model these novel features. Furthermore, the strategy based on 2D image projections of computationally optimized molecular geometries, proposed in our previous work [14], was also used herein, but now with a more powerful machine learning regression method incorporated to its routine, support vector machine applied to regression (SVR), which is capable of modelling non-linear relationships .

In order to generate novel robust 3D descriptors, we have proposed in this study the use of images obtained after the slicing of compounds with computationally optimized geometries. This approach was based on the same principle employed in tomographic medical imaging techniques, where medical instruments pick slices at regular intervals and generalize three-dimensional (3D) volumetric images from a series of slices collected in a certain direction [16]. An alternative source of 3D MIA-QSAR descriptors was also suggested, the use of images generated from the front, right, and top faces of the molecules, when those are assumed inside a theoretical box.

Three different data sets were selected from the literature for validation and comparison purposes. A series of compounds with anti-HCV (Hepatitis C virus) activity [17], a set of anti-SARS-CoV (Severe Acute Respiratory Syndrome Coronavirus) molecules [18], and a group of anti-HIV (Human Immunodeficiency Virus) compounds [19].

This paper follows in Section 2 with the presentation of the proposed 3D MIA-QSAR descriptor sources, as well as the multivariate regression methods applied in each case. Section 3 is devoted to present a basic foundation of the machine learning methods employed herein. The results and a proper discussion are presented in Section 4. A final Section (5) summarizes the main conclusions of this study.

2. Materials and methods

The three data sets (anti-HCV, anti-SARS-CoV, and anti-HIV compounds) are presented in Tables S.1, S.2, and S.3 in the Supplementary Material. For each group, the following routine was performed: first, the compounds were designed with the aid of the GaussView 5.0 software [20]. Then, the molecular geometries were optimized at the ωB97X-D/6-31G(d,p) [21,22] level of the density functional theory (DFT), in the Gaussian 09 software [23]. The structurally optimized compounds were aligned in the Discovery Studio Visualizer software [24]. Fig.1 shows the resulting superposition of the molecules for each data set, as well as the type of molecular representation images normally employed in a MIA-QSAR analysis. These final aligned structures were employed in the three techniques described as following.

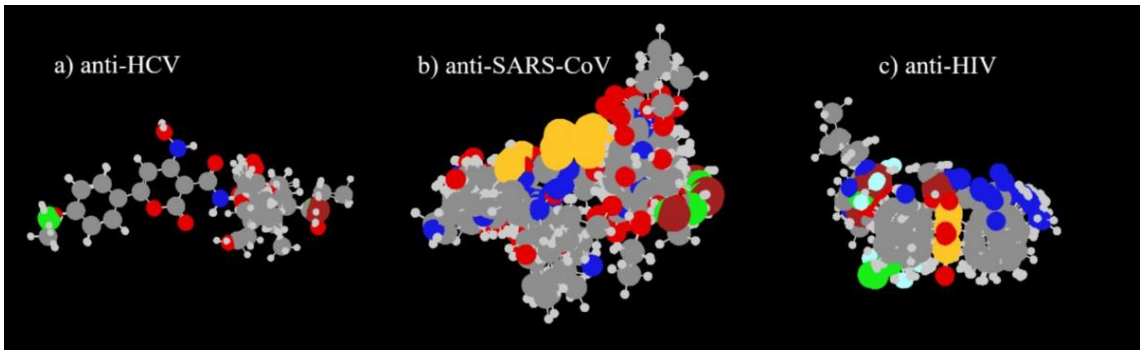


Fig. 1. Superposition of the images obtained from the computationally optimized molecular geometries.

2.1 2D image projections

2D image projections were generated after loading the aligned structures into the GaussView 5.0 software [20]. The image files were unfolded into a pixel matrix of dimensions $n \times (x \times y)$, where n is the number of molecules, and $x \times y$ is a result of the image measurements (x = width, y = height). From this matrix, coherent descriptors were selected after removing the zero variance columns, which refers to the empty spaces and coincident substructures. Lastly, the final 3D MIA-QSAR descriptors were obtained after replacing the RGB color values (0-765) by combinations of chemically meaningful values of atomic radii (r_{atm}), provided by Mathematica's Element Data function from Wolfram Research, Inc. [25], and Pauling´s electronegativity (ε). Table 1 displays the atomic radii/electronegativity replacement scheme employed herein.

Table 1
Atomic radii and electronegativity values used for chemical elements.

Element	GaussView color code	r_{atm} (×10)	ε (×100)	$r_{atm}/ε$ (×100)
C	426	670	255	263
H	612	530	220	241
O	229	480	344	139
N	279	560	304	184
F	688	420	398	105
Cl	289	790	316	250
Br	231	940	296	318
S	493	870	258	337

A support vector machine regression model was, then, built for each of the three final descriptor matrices (with pixel values proportional to I – r_{atm}, II – ε, and III – $r_{atm}/ε$), as described in Section 2.1.1.

2.1.1 Support vector machine regression models for 2D image projections of computationally optimized molecular geometries

All the molecular modelling process was performed using the R 3.5.3 software [26]. Initially, the data sets containing the selected descriptors were divided into training and test sets: the anti-HCV test set chosen was the same proposed by Li and collaborators [27] – samples **1**, **6**, **17**, **19**, **24**, **26**, **29**, **34**, **37**; for the anti-SARS-CoV group, molecules **8**, **23** and **40** were removed from the data set before the splitting step, once they were considered outliers by the authors [18], then the remaining data were divided into training and test sets *via* Kennard-Stone algorithm [~25% (9 samples) of the molecules were included in the test set]; lastly, the anti-HIV data set was also divided using the Kennard-Stone algorithm, 25% (16 samples) of the molecules were selected for the test set.

A model for each training set was built with a kernel function of the type radial, using the "Caret" package [28] for hyperparameter tuning. The optimum parameters were obtained *via* 10-fold cross-validation repeated ten times and are shown in the Supplementary Material (Table S.4). The validation metrics used to evaluate the quality of the models were: the regression coefficient of determination (r^2), cross-validated r^2 (q^2), the external r^2 (r^2_{test}), the modified r^2 (r^2_m) [29], and the RMSE (Root Mean Square Error) associated with each accuracy metric. The risk of chance correlation was analyzed using the ${}^cr^2_p$ parameter [${}^cr^2_p = r\,(r^2 - r^2_{y\text{-}rand})^{1/2}$, where $r^2_{y\text{-}rand}$ corresponds to the mean of the r^2 value obtained after randomizing the *y* block ten times] [30]. The validation parameters for each group of chemical compounds are presented in Section

4. The resulting plot of predicted × measured, for the best model in each case, is included in the Supplementary Material (Figures S.1, S.2, and S.3) along with a plot of the chemical space (PC1 × PC2 – Figures S.4, S.5, and S.6). Due to the complexity of the SVR model and the amount of data generated in this study, the applicability domain through William's plot analysis could not be performed, once memory limitation warnings were reported by the R software [26]. However, once the main goal of this type of analysis is to verify whether or not the samples belong to the same chemical space, independently of the descriptors, a William's plot for each chemical data set (anti-HCV, anti-SARS-CoV, and anti-HIV) was generated using the traditional MIA-QSAR technique (2D MIA-QSAR) and included in the Supplementary Material as well (Figure S.7). The plots for the anti-HCV and anti-SARS-CoV compounds are the same presented in our previous works [14][15] and, therefore, a proper discussion can be found in those studies. For the anti-HIV data set, the analysis of studentized residuals × sample leverages indicate the absence of outliers.

Two other kernel functions were tested in order to compare the influence of the mapping function on the accuracy metrics, linear and polynomial. The results from both functions are presented in Tables S.5, S.6 and S.7 in the Supplementary Material.

2.2 Molecular slicing and molecule-in-a-box approaches

In order to generate the slice images, the following routine was carried out: using a freely available version of the Pymol 1.8.2 software [31], the previous structurally optimized compounds were sliced, maintaining the depth coordinate fixed. In this case, the thickness of the slices was fixed at 1.0Å, and an equal number of slices – with the same dimensions (110×100px) – was obtained for each molecule inside the data set, based on the maximum amount required for that specific group. For the anti-HCV compounds, 13 slices were generated, anti-SARS-CoV data set required 11 slices, and the anti-HIV, 8 slices. Fig. 2 demonstrates the slicing process of a molecule.

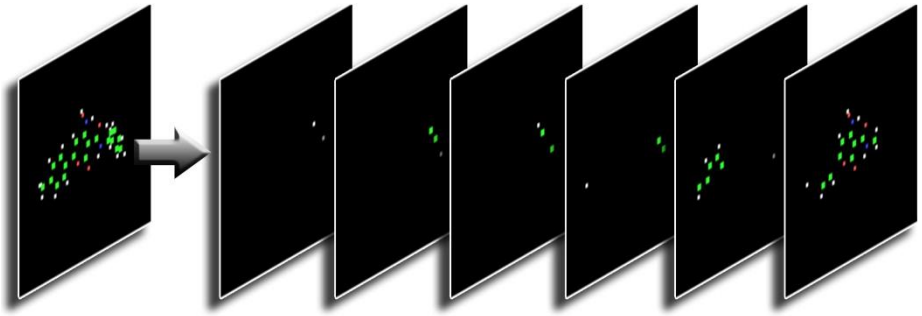


Fig. 2. Generation of slice images from a computationally optimized chemical structure.

After generating the slices, the image files were converted into matrices of pixel using the Python 2.7.13 software [32]. Seeking the reduction of the number of variables without losing the equal dimension aspect of the slices, the empty borders, common to all slices, were excluded. This step, as well as the following procedure, were performed using the R 3.5.3 software [26].

The reduced matrix of each slice was unfolded into a row vector and, then, the vectors of the slices from the same molecule were combined into a single row (side by side). The X-block was obtained after merging all rows into a new matrix, in such a way that each row corresponded to a single molecule and the number of columns was a result of the number of slices multiplied by the dimensional size of the image files (width × height), which was fixed.

Next, the pixel values were replaced according to the color pattern of Table 1. Thus, three data sets for each molecule group were obtained, one with descriptor values proportional to atomic radius (r_{atm}), another with variables proportional to the Pauling´s electronegativity (ε), and a final matrix with values proportional to the ratio r_{atm}/ε. The trilinear PLS algorithm was chosen to model the data sets generated in this approach and the modelling routine is described in Section 2.2.1. The results are shown in the Supplementary Material (Tables S.8, S.9, and S.10) and discussed in Section 4.

In a similar way, the image files for the molecule-in-a-box approach were generated. Using the Pymol 1.8.2 software [31], images from the front, top, and right faces of the structurally optimized molecules were obtained assuming the compounds inside a theoretical box, *i.e.*, the top right vertex was taken as reference. The dimensions of the facial images were equal for all molecules inside the same group (110×100px for the anti-HCV and anti-HIV compounds; 220×190px for the anti-SARS-CoV molecules). Fig. 4 represents the process for generating molecule-in-a-box face images of a molecule.

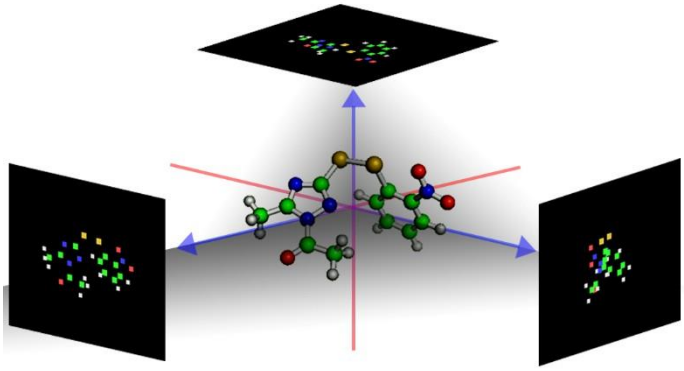


Fig. 3. Graphical representation of the generation of molecule-in-a-box face images.

After generating the three facial images of the molecules, a routine similar to that followed for the molecular slice images technique was employed. First, the image files were converted into matrices of pixel using the Python 2.7.13 software [32]. Then, the empty borders, common to all images, were excluded. The reduced matrix of each face was unfolded into a row vector and, then, the vectors of the faces from the same molecule were combined into a single row (side by side). The X-block was obtained after merging all rows into a new matrix, in such a way that each row corresponded to a single molecule and the number of columns was a result of 3 × the dimensional size of the image files (width × height). Next, the pixel values were replaced according to the color pattern established in Table 1. Thus, three data sets for each molecule group were obtained, one with descriptor values proportional to atomic radius (r_{atm}), other with variables proportional to the Pauling´s electronegativity (ε), and a final matrix with values proportional to the ratio r_{atm}/ε.

As in the slice image technique, the trilinear PLS algorithm was chosen to model the data sets generated in this approach. The exactly same routine, described in Section 2.2.1, was employed; the results are shown in the Supplementary Material (Tables S.11, S.12, and S.13) and discussed in Section 4 along with the slicing technique outcomes.

2.2.1 Trilinear PLS models for molecular slicing and molecule-in-a-box approaches

Initially, the data sets were divided into training and test sets. The splitting of the data followed the same distribution used in the SVR models. For the anti-HCV, the test set included the same samples used in the previous section. Similarly, molecules **8**, **23**, and **40** were removed from the anti-SARS-CoV data set and the remaining samples were divided *via* Kennard-Stone algorithm (9 samples were included in the test set). For the anti-HIV group, the splitting was also performed *via* Kennard-Stone algorithm (16 samples were included in the test set).

A trilinear PLS (tri-PLS) model requires a 3D-structured data set as an input. Therefore, a dimensional rearrangement was the first step after splitting the data. The row vectors corresponding to each slice/box face were reorganized along the *z*-coordinate, *i.e.*, the slices/box faces were arranged sequentially one after the other, and a final 3D-structure with dimensions of [*n* (number of molecules) × *p* (total number of pixels in a slice/box face) × *sf* (number of slices/box faces)] was obtained in each case. Fig. 4 is a graphical representation of the 3D-structured data used for building the tri-PLS models.

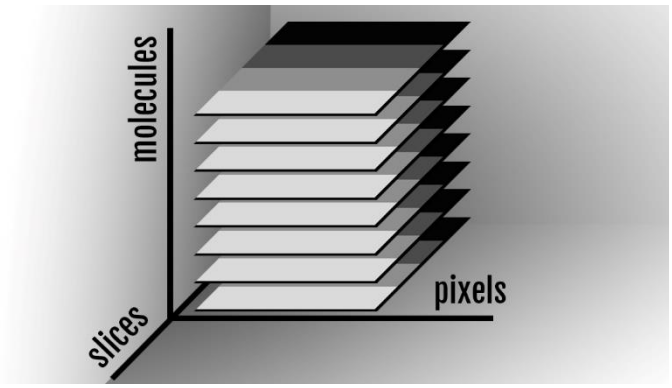


Fig. 4. Graphical representation of the 3D-structured data used in the trilinear-PLS analysis.

A model for each training set was built using the "Tripls" package [33]. The optimum number of latent variables was chosen *via* the analysis of the decay of the root mean square error (RMSE) in the leave-one-out cross-validation (LOOCV). The validation metrics used to evaluate the quality of the models were the same used for evaluating the SVR models (r^2, q^2, r^2_{test}, r^2_m, $^cr^2_p$, $r^2_{y\text{-}rand}$, and the RMSE associated with each parameter). The results are presented in the Supplementary Material (Tables S8 – S13) and discussed in Section 4. The plots of predicted × measured, for the best models in each group of compounds, are also included in the Supplementary Material (Figures S.8 – S.13) along with plots of the chemical spaces (PC1 × PC2 – Figures S.14 – S.19).

Figure 5 is a workflow that summarizes the approaches for encoding tridimensional information into MIA-QSAR descriptors and the methods for modelling the information described herein.

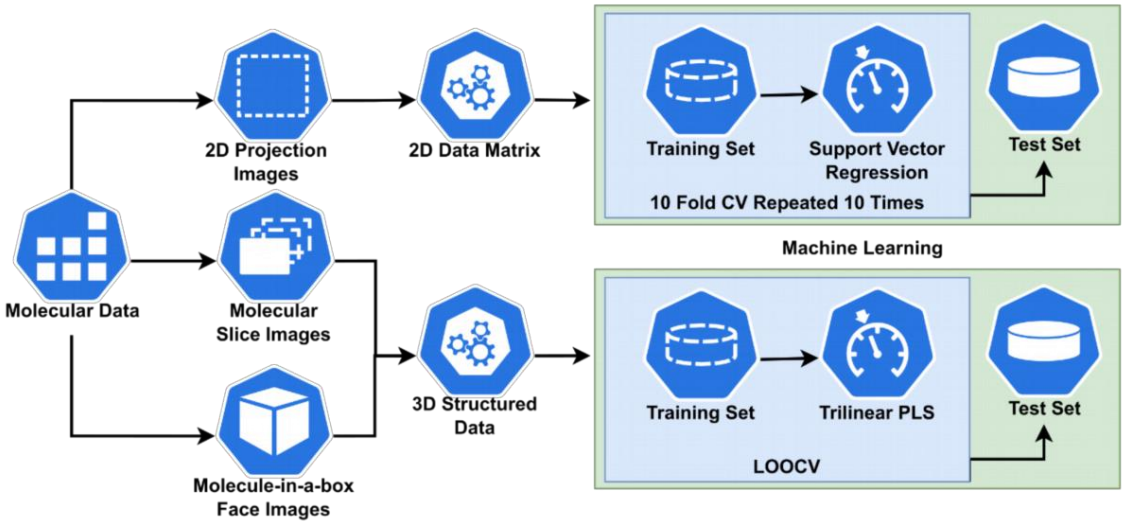


Fig. 5. Workflow of the strategies chosen for encoding and modelling tridimensional information in a MIA-QSAR perspective.

3. Theory

3.1 Support Vector Machine applied to regression (SVR): a brief description

Support vector machine (SVM) is a supervised machine learning method based on the structural risk minimization (SRM) principle and the statistical learning theory proposed by Vapnik and Chervonenkis [34,35]. This class of methods seeks the minimization of the VC (Vapnik-Chervonenkis) dimensions instead of the absolute value of an error or a squared error. This modification has proven to guarantee a better global generalization capability of the regression/classification model than the use of empirical risk minimization (ERM) parameters [36].

The essence of support vector machine for regression (SVR) is very similar to that in classification, *i.e.*, the aim is to find an optimum separating hyperplane (in a higher dimension space) that fits the data and minimizes errors. Technically, in SVR, the goal is to choose a hyperplane with small norm while simultaneously minimizing the sum of the distances from the data points to the hyperplane [37].

Two important concepts are constantly employed in SVM classification and regression problems: kernel and optimizer algorithm. The former consists of a function that maps a lower dimensional data into a higher dimensional space, where a linear separation of the data is possible. The second component is applied to solve the optimization problem. A linear SVM estimating function can be expressed as in Eq. (1).

$$f(x) = w\psi(x) + b. \quad (1)$$

In which, $\psi(x)$ is the kernel function previously described, 'w' is a weight vector and 'b' is a threshold value obtained after minimizing a regularized risk function. From the risk function, two parameters stand out: the epsilon (Ɛ – insensitive loss function), which expresses the Vapnik's linear loss function zone to measure empirical error, and the cost ('C' – cost function), which specifies the trade-off between the empirical risk and the model smoothness, in other words, it controls the empirical risk degree of the SVM model [38].

3.2 Tri-PLS: a brief description

Partial least squares (PLS) regression is a widely used multivariate technique designed specifically to deal with small datasets, missing values and multicollinearity, *i.e.*, cases in which ordinary least squares presented an undesirable performance (unstable models) [39]. These features have made PLS a very suitable modelling tool for chemical data, which normally contain few samples with highly collinear variables. Therefore, it is not surprising that PLS was incorporated by chemists in their data analysis routines, although it was originally designed to deal with econometric problems [39].

In 1996, Rasmus Bro presented a new multi-way regression method called N-way partial least squares (N-PLS), which works as an extension of the ordinary regression method PLS to multi-way cases [40]. In a multi-way method, the unfolding of the data set is not required; this aspect brings a series of advantages such as the maintenance of multi-way information, which is normally lost during the unfolding procedure.

There are many multi-way methods, but N-PLS presents the advantage of speed, a crucial feature when dealing with several variables, and a stabilized solution, which optimizes the interpretability and accuracy in predictions [40].

The algorithm proposed in N-PLS is a natural extension of the bilinear PLS algorithm. In a tri-PLS model, for example, the goal is to perform the decomposition of a cube $\underline{\mathbf{X}}$ into a set of triads, which are equivalent to bilinear factors, in a way that the covariance between $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$ is maximized [41]. According to the algorithm proposed, a triad consists of one score vector *t* and two weight vectors, w^J *and* w^K, the former is a second order vector and the latter a third order vector [40]. Eq. (2) is a general representation of the model of *X*.

$$x_{ijk} = t_i w_j^J w_k^K \quad (2)$$

4. Results and Discussion

4.1 SVR modelling

The optimal models for pixel proportional to r_{atm}, ε, and r_{atm}/ε, built with the best hyperparameters obtained in the tuning process with radial kernel, for each data set, are presented in Tables 2, 3 and 4.

Table 2
Internal and external validation parameters for anti-HCV compounds using SVR algorithm with radial kernel.

Parameters	$\mathbf{r_{atm}}$	**ε**	$\mathbf{r_{atm}/\varepsilon}$
RMSEC	0.0875	0.0754	0.0844
r^2	0.8939	0.9270	0.9054
$RMSE_{y\text{-}rand}$	0.1623	0.1639	0.1416
$r^2_{y\text{-}rand}$	0.8477	0.8440	0.8883
$^cr^2_p$ (y-rand)	0.2032	0.2773	0.1241
RMSECV	0.2685	0.2474	0.2729
q^2	0.6576	0.6207	0.6226
RMSEP	0.1479	0.1745	0.1494
r^2_{test}	0.5979	0.6234	0.6040
r^2_m (test)	0	0.1490	0

Results for 2D image projections of computationally optimized molecular geometries using PLS algorithm [14]: r^2: 0.7188, q^2: 0.4467, $r^2_{y\text{-}rand}$: 0.4480, $^cr^2_p$: 0.4112, r^2_{test}: 0.7395, and r^2_m: 0.5137.

Table 3
Internal and external validation parameters for anti-SARS-CoV compounds using SVR algorithm with radial kernel.

Parameters	$\mathbf{r_{atm}}$	**ε**	$\mathbf{r_{atm}/\varepsilon}$
RMSEC	0.1232	0.1975	0.0641
r^2	0.9930	0.9813	0.9982
$RMSE_{y\text{-}rand}$	0.3021	0.5248	0.1138
$r^2_{y\text{-}rand}$	0.9698	0.9208	0.9964
$^cr^2_p$ (y-rand)	0.1518	0.2435	0.0429
RMSECV	1.2465	1.2478	1.2305
q^2	0.7107	0.7273	0.6850
RMSEP	0.5351	0.4359	0.6356
r^2_{test}	0.7422	0.7979	0.6952
r^2_m (test)	0.5161	0.5384	0.5213

Results for 2D image projections of computationally optimized molecular geometries using PLS algorithm [15]: r^2: 0.9050, q^2: 0.3920, $r^2_{y\text{-}rand}$: 0.8710, $^cr^2_p$: 0.1760, r^2_{test}: 0.8230, and r^2_m: 0.8120.

Table 4
Internal and external validation parameters for anti-HIV compounds using SVR algorithm with radial kernel.

Parameters	$\mathbf{r_{atm}}$	**ε**	$\mathbf{r_{atm}/\varepsilon}$
RMSEC	0.1813	0.2542	0.1361
r^2	0.9460	0.9039	0.9748
$RMSE_{y\text{-}rand}$	0.5187	0.5464	0.4086
$r^2_{y\text{-}rand}$	0.7292	0.7509	0.8671
$^cr^2_p$ (y-rand)	0.4529	0.3719	0.3241
RMSECV	0.5883	0.7041	0.5748
q^2	0.6281	0.5555	0.6941
RMSEP	0.4289	0.3050	0.4182
r^2_{test}	0.6674	0.8238	0.7457
r^2_m (test)	0.2661	0.5060	0.3860

The SVR modelling outcomes for the anti-HCV compounds (Table 2) revealed similar modelling performance for the three type of descriptor values (proportional to r_{atm}, ε, and r_{atm}/ε). In all these cases, there were issues with both internal and external validation. The results for the *y*-block randomization test do not exclude the risk of chance correlation ($^cr^2_p < 0.5$), and when they are analyzed in parallel with the r^2_m values, this idea is reinsured and an overfitting problem is implied ($r^2_m < 0.5$). On another hand, the results from the 10-fold cross-validation (repeated 10×) indicate a moderate stability of the model for predicting the pEC_{50} property. When comparing these results with those of our previous work [14] (the validation parameters can be found in the footnote of Table 2), it is possible to observe some improvement in the cross-validation parameters, but a worse predictive performance for the test set. For both algorithms (PLS and SVR) the risk of chance correlation could not be excluded.

The SVR validation parameters for the anti-SARS-CoV compounds (Table 3) were also similar for the three data values, and in none of them the risk of casual correlation was dismissed, but, unlike the anti-HCV group, the external validation metrics do not suggest an overfitting of the data. Furthermore, 10-fold cross-validation (repeated 10×) indicates a satisfactory stability of the SVR prediction model. Comparing these findings with the results obtained in our previous work [15] (the validation parameters can be found in the footnote of Table 3), it is also possible to observe a meaningful improvement in the cross-validation parameters.

At this point, it is important mentioning that it is not surprising that the cross-validation parameters obtained *via* LOOCV, in the regular PLS approach, for both anti-HCV and anti-SARS-CoV, were worse than those obtained through repeated *k*-fold CV in the SVR models, because, although the test error in LOOCV is approximately an unbiased estimate of the true

prediction error, it has a high variance, since the training sets generated during the analysis only differ in one case [42]. Lastly, the anti-HIV SVR outcomes (Table 4) suggest a slightly higher dependence between pIC_{50} and electronegativity than between *y* and r_{atm}, once the model built for the data set with descriptor values proportional to ε produced better external validation metrics. The 10-fold cross-validation (repeated 10×) results were also acceptable.

Regarding the use of the linear and polynomial kernels, in an overall analysis, Tables S.5, S.6 and S.7 indicate a slightly better performance of those functions in the external validation step (the r^2_m values were superior) than the radial kernel for the anti-HCV group, but considerably worse results for the *y*-block randomization tests (the $^cr^2_p$ were nearly zero). On another hand, for the anti-SARS-CoV data base, the radial kernel was superior in the external validation and in the *y*-block randomization tests. Finally, for the anti-HIV data set, the same behavior observed for the anti-HCV was observed, *i.e.*, the linear and polynomial kernels were slightly better in the external validation step than the radial, but were considerably worse in the *y*-block randomization tests. Therefore, from a general point of view, the choice of the kernel function presented some influence on the external outcomes and on the chance correlation analysis, but it was not a strong effect. In summary, the SVR models obtained presented good quality parameters for the three data sets investigated; however, it is difficult to attribute these results to the multivariate regression method chosen (SVR), to the source of descriptors (simple 2D image projections), or a combination of both. Therefore, the proposal for new sources of features, as suggested herein (slice and box face images of molecules), is a natural and adequate thought, as well as the use of a different regression method (trilinear PLS).

4.2 Tri-PLS modelling

The slice and box face images strategies were accomplished in association with the trilinear PLS method with the goal of capturing 3D information, that is normally lost in a 2D projection routine (as in the traditional MIA-QSAR analysis or the 2D image projection method previously proposed), and maintaining the maximum possible amount of that information, avoiding unnecessary unfolding steps of regression methods that are not multi-way.

The three models (pixel values proportional to r_{atm}, ε, and r_{atm}/ε), for each data set, obtained through the slicing technique are shown in Tables S.8, S.9 and S.10, in the Supplementary Material, as well as the results obtained *via* the box face images (Tables S.11-S.13).

The resulting models for the anti-HCV compounds obtained through molecular slice images (Table S.8) reveal a higher dependence between pEC_{50} and electronegativity than with atomic radii. Nonetheless, in all models, the leave-one-out cross-validation metrics were not satisfactory, as well as the ${}^{c}r^2_p$ and the r^2_m parameters. These findings cast doubt on the quality of both internal and external validation results because they do not exclude the risk of casual correlation or support the performance observed in the external predictions. On another hand, the results for the anti-SARS-CoV molecules (Table S.9) show a meaningful improvement in the calibration parameters, as well as in the external validation when compared to the previous group. However, the *y*-randomization test and the leave-one-out cross-validation parameters still indicate a poor performance of the models. The results obtained for the anti-HIV compounds (Table S.10) demonstrate that the best model was the one built with the data set proportional to r_{atm}/ε, which indicates that the pIC_{50} is correlated to both types of chemical information. However, even the best model did not exclude the risk of chance correlation and the external validation parameters were also unsatisfactory.

Accordingly, an alternative source of descriptors was also explored, the box face images. The modelling performance of this approach for the anti-HCV compounds (Table S.11) is better than that observed in the slice images technique. Table S.11 demonstrates a similar performance among the three models built for pixel values proportional to r_{atm}, ε, and r_{atm}/ε. Satisfactory parameters of calibration and external validation were observed for the three data sets; however, the cross-validation and *y*-randomization test parameters were insufficient. On another hand, for the anti-SARS-CoV and anti-HIV compounds, a similar performance to that observed in the slice images technique is noticed. Table S.12 reveals poor leave-one-out cross-validation parameters in addition to unsatisfactory *y*-randomization test outcomes for the anti-SARS-CoV molecules. The results for the anti-HIV (Table S.13) indicate some improvement in the external validation, but the r^2_m values were still below the cut-off value of 0.5; furthermore, the risk of chance correlation could not be excluded.

At this point, it is interesting to compare the trilinear PLS algorithm modelling performance with that of the SVR, despite the difference in the dimensionality of the data bases. In general terms, for the anti-HCV compounds, it is possible to observe that the SVR generated better cross-validation parameters than the trilinear PLS. As previously mentioned, the LOOCV is a less stable method for estimating performance (in this case, RMSE) than repeated *k*-fold CV [42], therefore, this finding is not surprising. On another hand, the external validation parameters generated by the trilinear PLS algorithm were better than those obtained *via* SVR;

the *y*-block randomization test parameters were similar in both methods ($^{c}r^2_p < 0.5$). For the anti-SARS-CoV data set, the repeated *k*-fold CV generated notably higher q^2 values than LOOCV in the trilinear algorithm; other parameters presented similar behavior. Lastly, for the anti-HIV data, the SVR algorithm associated with the 2D image projections technique generated considerably better cross-validation and external validation outcomes than those observed for the trilinear algorithm; further parameters were alike.

Considering the extensive analysis performed in this work and the results obtained for each of the three approaches explored herein, the concept of overfitting seems to stand out, once the parameters $^{c}r^2_p$ and r^2_m were below the cut-off values in most of the cases. The reasons for this issue may be complex to analyze, but possible sources are: (I) noisy learning on the training set and (II) hypothesis complexity (too many inputs). According to Ying (2019), small data sets may contribute to increase noisy learning on the training set, as well as less representative data or too many noises [43]. In the case of the MIA-QSAR routines investigated herein, one may imply some possible sources of noise: the data sets chosen for the analysis were small; for the slicing technique, there were images completely empty, because it was necessary to maintain the same number of slices for all molecules and they have different molecular volumes; the computational optimization of molecular geometries also contributed for including more variability (noise) into the models, because even in cases where the molecules have identical substituents, at the same positions, there is not repeatability of variables due to the conformational freedom, *i.e.*, the congruence of the pharmacophoric substructures and substituents is lost during the process of geometry optimization. All these issues, in addition to the large number of descriptors used in the present approaches, may have resulted in overfitting. Possible solutions for such issue are proposed by Ying (2019): the use of the so-called "early stopping" strategy; the exclusion of noise from the training set; data expansion; and the "regularization strategy", which is based on variable selection [43].

Comparing the results obtained through the three techniques (2D image projections, molecular slice images, and molecule-in-a-box face images) explored in this work with those obtained *via* the traditional MIA-QSAR technique (Table S.14), one can see that the performance of classical 2D descriptors in describing the relationship between structural features, captured from "flat" molecular images, and biological/physicochemical properties is far more efficient than the performance of 3D descriptors obtained from images of computationally optimized geometries *via* the methods presented herein. This conclusion reinsures the efficiency of MIA-QSAR as a simple technique based on 2D image projections of congruent molecules obtained from molecular drawing programs without geometry

optimization or rotation. This does not exclude the importance of conformational features in explaining biological activity. Therefore, "flat" chemical structures may be used in a 3D molecular field method to attest whether molecules with superimposable common substructures are indeed more encoding than conformationally unconstrained molecules, or if digital image treatment of objects in three dimensions may not be a practical process for QSAR purposes. A critical evaluation addressing this issue is currently in progress.

So far, we have performed comparisons involving exclusively descriptors from image sources, but it is also important to analyze the performance of the present strategies against 3D QSAR approaches that use well-established molecular descriptors. The works used to retrieve the anti-HCV and anti-SARS-CoV data sets employed CoMSIA [27] and CoMFA [18], respectively, to model these compounds. For the anti-HCV group, CoMSIA modelling resulted in r^2: 0.904, q^2: 0.569, $^cr^2_p$: 0.609, r^2_{pred}: 0.653, and r^2_m: 0.680. Facing these results, the slice and molecule-in-a-box face strategies presented cross-validation parameters inferior to those obtained with CoMSIA; on another hand, the 2D projection approach generated better q^2 results (Table 2). Regarding the external validation, the molecule-in-a-box face strategy presented a similar performance to CoMSIA, with a considerably higher value of r^2_{pred} (0.7940) and a slightly smaller r^2_m (0.5592); the other strategies presented r^2_m values inferior to 0.5. Finally, with respect to $^cr^2_p$, as previously discussed, none of the proposed approaches eliminated the risk of chance correlation, *i.e.*, the $^cr^2_p$ results were inferior to the cut-off value of 0.5. For the anti-SARS-CoV data set, the results with CoMFA reported by the authors were r^2: 0.9160, standard error: 0.088, and q^2: 0.681, once external validation was not performed. Accordingly, only a partial comparison is possible. The 2D projection model (Table 3) generated superior validation parameters when compared to CoMFA; on another hand, the slice and molecule-in-a-box face models presented worse cross-validation results.

5. Conclusions

This work reports creative ways to encode tridimensional information from digital images of chemical structures, and also the use of modelling techniques based on nonlinear and three-way data regression. In the first attempt of including spatial information into the MIA-QSAR descriptors, a powerful machine learning regression method (SVR) was applied to three data sets obtained from 2D image projections of computationally optimized molecular geometries. As a result, some of the models presented good quality parameters for both internal and external validation, but the *y*-block randomization test results could not exclude the risk of casual

correlation. In a second attempt, a different source of 3D information (slice images) was explored, as well as a different multivariate regression model (trilinear PLS). The resulting models presented a moderate predictive performance, once the risk of chance correlation could not be excluded either. The last attempt was the use of molecule-in-a-box face images as an alternative source of spatial descriptors. As a result, some improvement on the modelling performance was observed for one of the compound data sets (anti-HCV) when compared to the previous approach, and a similar performance was identified for the others. In summary, important progress has been made towards the effective encoding and modelling of 3D information in MIA-QSAR. However, the conventional MIA-QSAR analysis based on the alignment of congruent pharmacophoric substructures has proven to be more efficient than the approaches explored herein.

Acknowledgements

The authors are thankful to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, funding code 001), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grant number 301371/2017-2), and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) for the financial support of this research.

References

[1] C. Hansch, T. Fujita, ρ-σ-π Analysis. A method for correlation of biological activity and chemical structure, J. Am. Chem. Soc. 86 (1964) 1616−1624. https://doi.org/10.1021/ja01062a035.

[2] C. Hansch, M. Streich, F. Geiger, R.M. Muir, P.P. Maloney, T. Fujita, Correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients, J. Am. Chem. Soc. 85 (1963) 2817−2824. https://doi.org/10.1021/ja00901a033.

[3] R.D. Cramer, D.E. Patterson, J.D. Bunce, Comparative molecular-field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins, J. Am. Chem. Soc. 110 (1988) 5959−5967. https://doi.org/10.1021/ja00226a005.

[4] G. Klebe, U. Abraham, Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. J. Comput.-Aided Mol. Des. 13 (1999) 1−10. https://doi.org/10.1023/A:1008047919606.

[5] T. Fujita, D. Winkler, Understanding the roles of the “two QSARs”, J. Chem. Inf. Model., 56 (2016) 269−274. https://doi.org/10.1021/acs.jcim.5b00229.

[6] M. Damale, S. Harke, F. Khan, D. Shinde, J. Sangshetti, Recent advances in multidimendional QSAR (4D-6D): A critical review, Mini Rev. Med. Chem. 14 (2014) 35−55. https://doi.org/10.2174/13895575113136660104.

[7] J. Verma, V. Khedkar, E. Coutinho, 3D-QSAR in drug design - a review, Curr. Top. Med. Chem. 10 (2010) 95−115. https://doi.org/10.2174/156802610790232260.

[8] D. Fourches, 4D- quantitative structure–activity relationship modeling: making a comeback, Expert Opin. Drug. Discov. 14 (2019) 1227−1235. https://doi.org/10.1080/17460441.2019.1664467.

[9] S.J. Barigye, M.P. Freitas, Ten years of the MIA-QSAR strategy: historical development and applications, Int. J. Quant. Struct.-Prop. Relatsh., 1 (2016) 64−77. https://doi.org/10.4018/IJQSAR.2016010103.

[10] J.K. Daré, S.J. Barigye, M.P. Freitas, Multi-objective modeling of herbicidal activity from an environmentally friendly perspective, Int. J. Quant. Struct.-Prop. Relatsh., 2 (2017) 16−26. https://doi.org/10.4018/IJQSAR.2017070102.

[11] M.P. Freitas, S.D. Brown, J.A. Martins, MIA-QSAR: A simple 2D image-based approach for quantitative structure-activity relationship analysis, J. Mol. Struct. 738 (2005) 149−154. https://doi.org/10.1016/j.molstruc.2004.11.065.

[12] C.A. Nunes, M.P. Freitas, Introducing new dimensions in MIA-QSAR: A case for chemokine receptor inhibitors, Eur. J. Med. Chem. 62 (2013) 297−300. https://doi.org/10.1016/j.ejmech.2013.01.005.

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

[14] J.K. Daré, T.C. Ramalho, M.P. Freitas, 3D perspective into MIA-QSAR: A case for anti-HCV agents, Chem. Biol. Drug. Des. 93 (2019) 1096−1104. https://doi.org/10.1111/cbdd.13440.

[15] J.K. Daré, D. Silva, T.C. Ramalho, M.P. Freitas, Conformational fingerprints in the modeling performance of MIA-QSAR: A case for SARS-CoV protease inhibitors, Mol. Simul. 46 (2020) 1055−1061. https://doi.org/10.1080/08927022.2020.1800691.

[16] Z. Li, Y. Wang, J. Yu, Reconstruction of thin-slice medical images using generative adversarial network, in: Q. Wang, Y. Shi, H. I. Suk, K. Suzuki (Eds.), Lecture notes in computer science, Springer, Cham, 2017, pp. 325−333.

[17] A.K. Konreddy, M. Toyama, W. Ito, C. Bal, M. Baba, A. Sharon, Synthesis and anti-HCV activity of 4-hydroxyamino α-pyranone carboxamide analogues, ACS Med. Chem. Lett. 5 (2013) 259–263. https://doi.org/10.1021/ml400432f.

[18] L. Wang, B.-B. Bao, G.-Q. Song, C. Chen, X.-M. Zhang, W. Lu, Z. Wang, Y. Cai, S. Li, S. Fu, F.-H. Song, H. Yang, J.-G. Wang, 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 (2017) 450–461. https://doi.org/10.1016/j.ejmech.2017.05.045.

[19] M. Goodarzi, M.P. Freitas, Augmented three-mode MIA-QSAR modeling for a series of Anti-HIV-1 compounds, QSAR Com. Sci. 27 (2008) 1092-1097. https://doi.org/10.1002/qsar.200810030.

[20] R.D. Dennington, T.A. Keith, J.M. Millam, GaussView 5.0., Gaussian, Inc., Wallingford CT, 2008.

[21] J.-D. Chai, M. Head-Gordon, Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections, Phys. Chem. Chem. Phys. 10 (2008) 6615–6620. https://doi.org/10.1039/B810189B.

[22] M.M. Francl, W.J. Pietro, W.J. Hehre, J.S. Binkley, M.S. Gordon, D.J. DeFrees, J.A. Pople, Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements, J. Chem. Phys. 77 (1982) 3654–3665. https://doi.org/10.1063/1.444267.

[23] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G.A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H.P. Hratchian, A.F. Izmaylov, J. Bloino, G. Zheng, J.L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J.A. Montgomery, Jr., J.E. Peralta, F. Ogliaro, M. Bearpark, J.J. Heyd, E. Brothers, K.N. Kudin, V.N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J.M. Millam, M. Klene, J.E. Knox, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, R.L. Martin, K. Morokuma, V.G. Zakrzewski, G.A. Voth, P. Salvador, J.J. Dannenberg, S. Dapprich, A.D. Daniels, O. Farkas, J.B. Foresman, J.V. Ortiz, J. Cioslowski, D.J. Fox, Gaussian 09, Revision D.01, Gaussian, Inc., Wallingford CT, 2013.

[24] BIOVIA, Dassault Systèmes, Discovery Studio Visualizer, R2. San Diego: Dassault Systèmes, 2017.

[25] WolframAlpha widgets, Atomic Radius, https://www.wolframalpha.com/widgets/view.jsp?id=58304cd7c0d99cf3f50b8a6f019a86ae, 2016 (accessed 27 March 2020).

[26] R 3.5.3 Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

[27] W. Li, F. Xiao, M. Zhou, X. Jiang, J. Liu, H. Si, M. Xie, X. Ma, Y. Duan, H. Zhai, 3D-QSAR study and design of 4-hydroxyamino α-pyranone carboxamide analogues as potential anti-HCV agents, Chem. Phys. Lett. 661 (2016) 36–41. https://doi.org/10.1016/j.cplett.2016.08.042.

[28] M. Kuhn, J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan,

Caret: Classification and Regression Training. R package version 6.0–71, 2016. https://CRAN.R-project.org/package=caret.

[29] K. Roy, P. Chakraborty, I. Mitra, P. K. Ojha, S. Kar, R. N. Das, 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, J. Comput. Chem. 34 (2013) 1071‑1082. https://doi.org/10.1002/jcc.23231.

[30] I. Mitra, A. Saha, K. Roy, Exploring quantitative structure‑activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants, Mol. Simul. 36 (2010) 1067‑1079. https://doi.org/10.1080/08927022.2010.503326.

[31] The PyMOL Molecular Graphics System, Version 1.8.2.0 Schrödinger, LLC.

[32] G. Van Rossum, F.L. Drake Jr. Python reference manual, Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[33] S. Serneels, Trilinear partial least squares regression. R package version 1.0, 2014. https://rdrr.io/github/SvenSerneels/tripls_r/man/tripls-package.html.

[34] V. N. Vapnik, An overview of statistical learning theory, IEEE Trans. Neural Netw. 10 (1999a) 988-999. https://doi.org/10.1109/72.788640.

[35] V. N. Vapnik, A. Chervonenkis, Theory of Pattern Recognition (in Russian), Nauka, Moscow, 1974.

[36] W. Niu, Z. Feng, B. Feng, Y. Min, C. Cheng, J. Zhou, Comparison of multiple linear regression, artificial neural network, extreme learning machine, and support vector machine in deriving operation rule of hydropower reservoir, Water 11 (2019) 88–104. https://doi.org/10.3390/w11010088.

[37] T. Trafalis, H. Ince, Support vector machine for regression and applications to financial forecasting, IEEE Int. Jt. Conf. Neural Netw. 6 (2000) 348–353. https://doi.org/10.1109/IJCNN.2000.859420.

[38] F. Zhang, C. Deb, S. Lee, J. Yang, K. Shah, Time series forecasting for building energy consumption using weightes support vector regression with differential evolution optimization technique, Energy and Build. 126 (2016) 94–103. http://dx.doi.org/10.1016/j.enbuild.2016.05.028.

[39] D. Pirouz, An overview of partial least squares, SSRN 1631359 (2006). http://dx.doi.org/10.2139/ssrn.1631359.

[40] R. Bro, Multiway calibration. Multilinear PLS, J. Chemom. 10 (1996) 47–61. https://doi.org/10.1002/(SICI)1099-128X(199601)10:1<47::AID-CEM400>3.0.CO;2-C

[41] D. Hérvaz, J. Prats-Montalbán, A. Lahoz, A. Ferrer, Sparse N-way partial least squares with R package sNPLS, Chemom. Intell. Lab. Syst., 179 (2018) 54–63. https://doi.org/10.1016/j.chemolab.2018.06.005.

[42] D. Berrar, Cross-Validation, Encycl. Bioinform. Comput. Biol. 1 (2019) 542−545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X.

[43] X. Ying, An overview of overfitting and its solutions, J. Phys.: Conf. Ser., 1168 (2019). https://doi.org/10.1088/1742-6596/1168/2/022022.

[44] C. N. Borges, S. J. Barigye, M. P. Freitas, Towards molecular design using 2D-molecular contour maps obtained from PLS regression coefficients, Mol. Phys., 115 (2017) 3044–3050. https://doi.org/10.1080/00268976.2017.1347294.

Supplementary Material

for

Different approaches to encode and model 3D information in a MIA-QSAR perspective

Joyce K. Daré, Matheus P. Freitas*

Address: Department of Chemistry, Federal University of Lavras, 37200–900, Lavras–MG, Brazil

Email: Matheus Puggina de Freitas – matheus@ufla.br

* Corresponding author

Table of contents

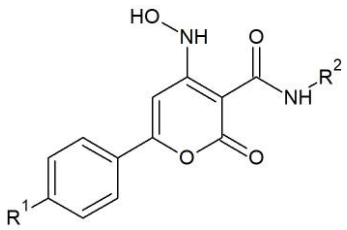
Page S16. Fig. S.17. Chemical space representation for the anti-HCV data set obtained from the molecule-in-a-box technique.

Page S16. Fig. S.18. Chemical space representation for the anti-SARS-CoV data set obtained from the molecule-in-a-box technique.

Page S16. Fig. S.19. Chemical space representation for the anti-HIV data set obtained from the molecule-in-a-box technique.

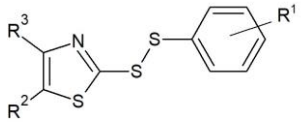
Page S17. Table S.14 Results of the traditional MIA-QSAR analysis from previous works [14, 15, 44].

Table S.1
Series of anti-HCV compounds used for MIA-QSAR modeling and the corresponding measured pEC_{50}.

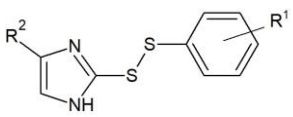


Molecule	R1	R2	pEC_{50}	Molecule	R1	R2	pEC_{50}
1	H	Ph	6.456	**36**	CH_3	CH_2COOMe	5.399
2	F	Ph	5.975	**37**	OCH_3	CH_2COOMe	5.529
3	Cl	Ph	6.252	**38**	H	CH_2Ph	6.538
4	CH_3	Ph	5.839	**39**	F	CH_2Ph	5.928
5	OCH_3	Ph	6.252	**40**	Cl	CH_2Ph	6.167
6	H	4-F-Ph	6.495	**41**	CH_3	CH_2Ph	6.041
7	H	4-Cl-Ph	6.398	**42**	OCH_3	CH_2Ph	6.244
8	H	4-Br-Ph	6.523				
9	H	4-OH-Ph	6.000				
10	H	4-OMe-Ph	6.770				
11	H	4-COOMe-Ph	6.481				
12	H	2-Me-Ph	6.268				
13	H	3-Me-Ph	6.569				
14	H	4-Me-Ph	6.745				
15	F	4-Me-Ph	5.710				
16	Cl	4-Me-Ph	5.845				
17	CH_3	4-Me-Ph	6.337				
18	OCH_3	4-Me-Ph	5.796				
19	H	Me	6.222				
20	F	Me	5.757				
21	Cl	Me	6.108				
22	CH_3	Me	5.848				
23	OCH_3	Me	6.155				
24	H	Et	6.000				
25	H	Pr	6.347				
26	H	*i*-Pr	6.347				
27	H	C_2H_4OH	5.830				
28	F	C_2H_4OH	5.526				
29	Cl	C_2H_4OH	5.398				
30	CH_3	C_2H_4OH	5.588				
31	OCH_3	C_2H_4OH	5.851				
32	H	C_3H_6OH	6.347				
33	H	CH_2COOMe	5.735				
34	F	CH_2COOMe	5.714				
35	Cl	CH_2COOMe	5.514				

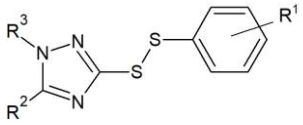
Table S.2
Series of anti-SARS-CoV compounds used for MIA-QSAR modeling and the corresponding measured IC_{50} (μM).



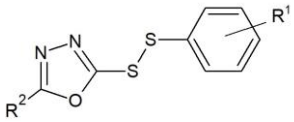
Comp.: 1-2, 5, 12- 13, 18-25



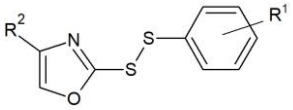
Comp.: 3, 14-17



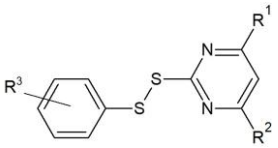
Comp.: 4, 6-11



Comp.: 26-29, 31



Comp.: 30

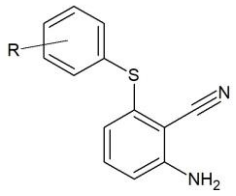


Comp.: 32-40

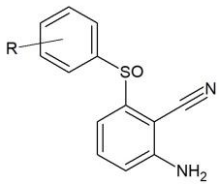
Molecule	R^1	R^2	R^3	IC_{50}	Molecule	R^1	R^2	R^3	IC_{50}
1	4-Cl	H	H	1.871	**32**	Me	Me	2-NO_2	0.921
2	4-Me	NHCOMe	H	2.803	**33**	Me	Me	4-Cl	1.437
3	4-Cl	$COOC_2H_5$	-	3.675	**34**	Me	Me	4-Br	1.121
4	2-NO_2	Me	$COCH_3$	3.130	**35**	Me	Me	H	1.991
5	H	NHCOMe	H	1.506	**36**	Me	Me	4-Me	1.495
6	4-Me	Ph	$COCH_3$	4.344	**37**	H	H	2-NO_2	0.883
7	4-OMe	Ph	$COCH_3$	4.100	**38**	H	H	4-Cl	0.684
8*	2-NO_2	3-pyridyl	$COCH_3$	1.762	**39**	H	H	4-Br	0.697
9	2-$COOC_2H_5$	3-pyridyl	$COCH_3$	5.654	**40***	H	H	4-Me	1.522
10	2-$COOC_2H_5$	4-pyridyl	$COCH_3$	4.511					
11	4-OCH_3	3-pyridyl	$COCH_3$	5.794					
12	4-Cl	NHCOMe	H	2.626					
13	4-Br	NHCOMe	H	1.651					
14	2-NO_2	$COOCH_3$	-	2.075					
15	2-$COOC_2H_5$	$COOCH_3$	-	5.954					
16	$COOCH_3$	$COOCH_3$	-	3.957					
17	4-Cl	$COOCH_3$	-	4.126					
18	4-F	NHCOMe	H	2.565					
19	2-NO_2	NHCOMe	H	1.947					
20	2-NO_2	H	H	2.029					
21	4-Me	H	H	1.250					
22	4-F	H	H	2.211					
23*	4-Br	H	H	3.321					
24	2-NO_2	H	Me	2.555					
25	2-$COOC_2H_5$	H	Me	2.452					
26	2-$COOCH_3$	Me	-	1.679					
27	2-$COOC_2H_5$	Me	-	1.557					
28	2-NO_2	Me	-	1.713					
29	2-$COOCH_3$	H	-	1.118					
30	2-$COOCH_3$	Me	-	1.264					
31	4-Cl	H	-	0.516					

*outliers.

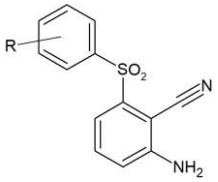
Table S.3
Series of anti-HIV compounds used for MIA-QSAR modeling and the corresponding measured pIC_{50}.



Comp.: **1-19**



Comp.: **20-34**



Comp.: **35-64**

Molecule	R	pIC_{50}	Molecule	R	pIC_{50}
1	H	1.836	**33**	3-Cl, 5-Me	3.495
2	2-OMe	2.367	**34**	3-OMe, 5-CF_3	2.684
3	3-OMe	2.222	**35**	H	2.699
4	2-Me	1.796	**36**	2-OMe	3.222
5	3-Me	2.215	**37**	3-OMe	3.046
6	4-Me	0.939	**38**	4-OMe	1.602
7	2-Cl	2.387	**39**	2-Me	2.638
8	3-Cl	2.131	**40**	3-Me	3.398
9	2-Br	1.523	**41**	4-Me	2.022
10	3-Br	2.292	**42**	2-Cl	2.387
11	3-F	2.009	**43**	3-Cl	3.229
12	3-CN	2.762	**44**	4-Cl	2.523
13	4-CN	1.359	**45**	2-Br	2.301
14	3-CF_3	1.893	**46**	3-Br	3.268
15	3-NH_2	1.502	**47**	4-Br	1.699
16	3,5-Me_2	3.367	**48**	2-F	2.523
17	3-Cl, 5-Me	2.754	**49**	3-F	2.523
18	3-OMe, 5-Me	2.699	**50**	2-CN	2.268
19	3-OMe, 5-CF_3	2.292	**51**	3-CN	2.620
20	2-OMe	2.319	**52**	4-CN	1.097
21	3-OMe	1.796	**53**	3-CF_3	2.456
22	2-Me	1.032	**54**	2,5-Cl_2	3.523
23	3-Me	1.534	**55**	3,5-Cl_2	4.155
24	4-Me	1.310	**56**	3,5-Me_2	5.000
25	2-Br	1.407	**57**	3-Br, 5-Me	4.699
26	3-Br	4.097	**58**	3-Cl, 5-Me	4.523
27	4-Br	1.694	**59**	3-OMe, 5-Me	4.301
28	2-CN	2.409	**60**	3-OMe, 5-CF_3	4.046
29	3-CN	1.848	**61**	3-OH, 5-Me	3.367
30	3-CF_3	1.398	**62**	3-OCH_2CH_3, 5-Me	4.222
31	3,5-Me_2	3.469	**63**	3-$O(CH_2)_2CH_3$, 5-Me	4.222
32	2,5-Cl_2	2.007	**64**	3-$O(CH_2)_3CH_3$, 5-Me	3.222

Table S.4
Hyperparameters employed in the SVR model with radial kernel.

Chemical group	Data set	Ɛ	C
Anti-HCV	r_{atm}	1.43×10^{-9}	2
	ε	7.45×10^{-9}	2
	r_{atm}/ε	8.27×10^{-9}	4
Anti-SARS	r_{atm}	3.77×10^{-10}	8
	ε	3.87×10^{-9}	4
	r_{atm}/ε	2.78×10^{-9}	16
Anti-HIV	r_{atm}	9.25×10^{-10}	2
	ε	3.29×10^{-9}	4
	r_{atm}/ε	1.07×10^{-8}	2

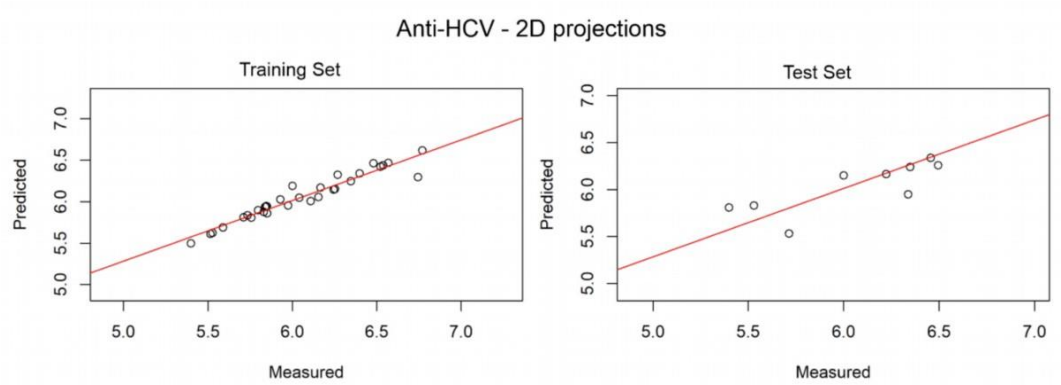


Fig. S.1. Predicted pEC_{50} × Measured pEC_{50} for the anti-HCV group with pixel values proportional to ε and SVR as regression method with radial kernel.

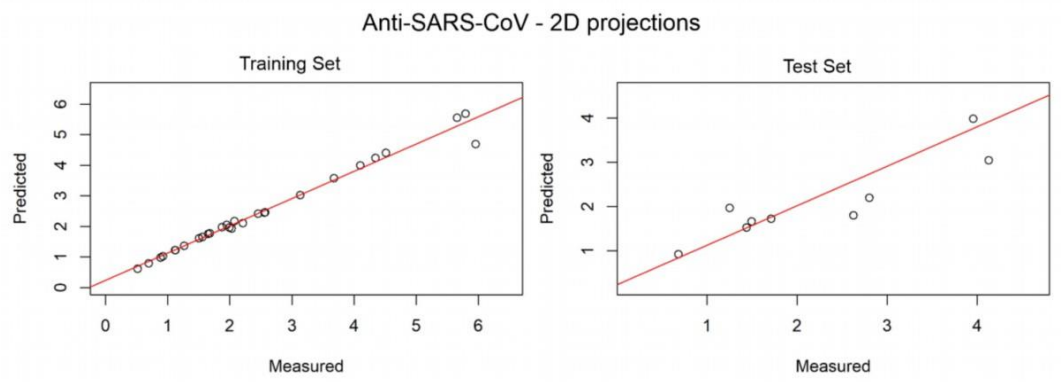


Fig. S.2. Predicted IC_{50} × Measured IC_{50} for the anti-SARS-CoV group with pixel values proportional to ε and SVR as regression method with radial kernel.

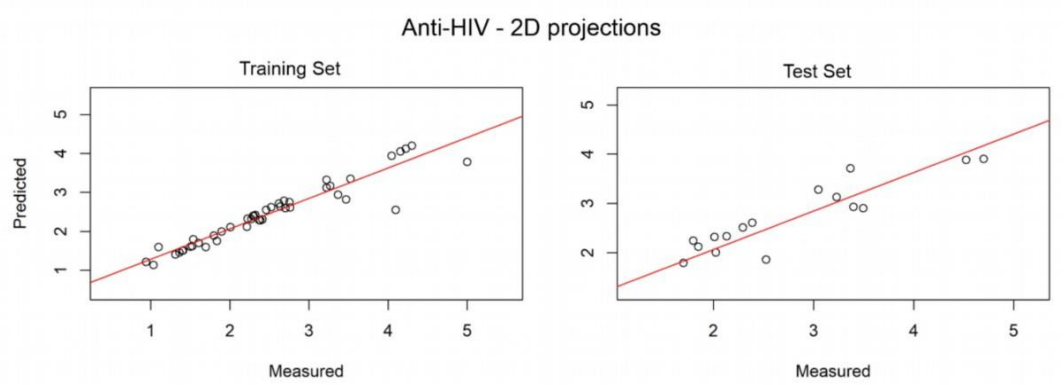


Fig. S.3. Predicted pIC_{50} × Measured pIC_{50} for the anti-HIV group with pixel values proportional to ε and SVR as regression method with radial kernel.

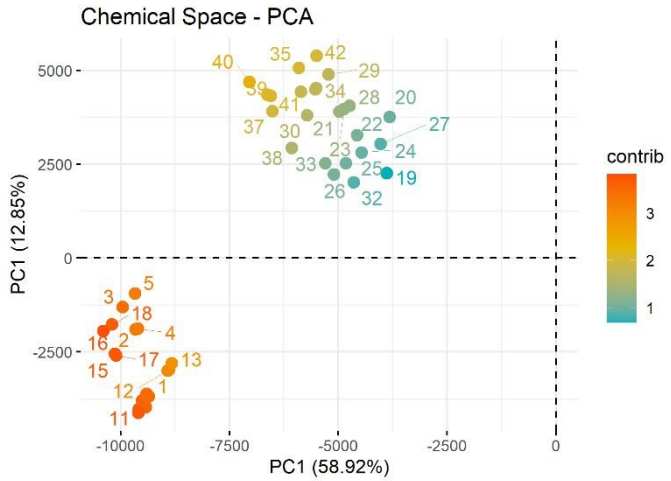


Fig. S.4. Chemical space representation for the anti-HCV data set obtained from 2D projections.

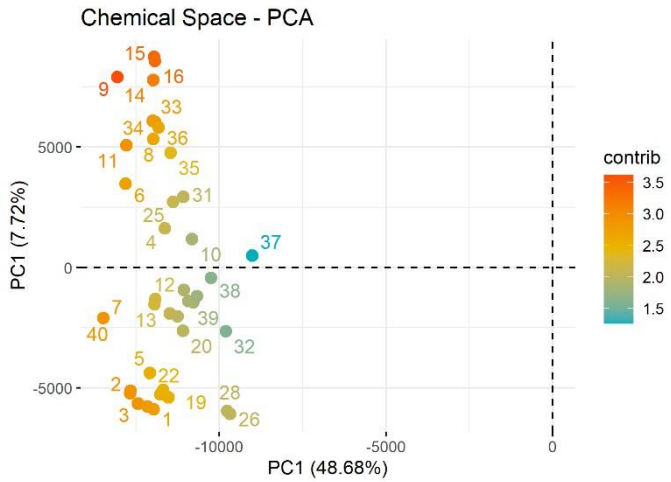


Fig. S.5. Chemical space representation for the anti-SARS-CoV data set obtained from 2D projections.

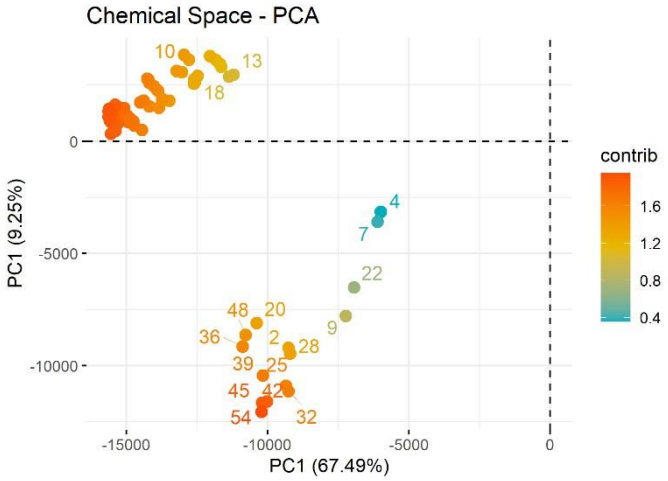


Fig. S.6. Chemical space representation for the anti-HIV data set obtained from 2D projections.

Table S.5
Internal and external validation parameters for anti-HCV compounds using SVR algorithm with linear and polynomial kernels.

Data set	Kernel function	Parameters	r_{atm}	ε	$r_{atm}/ε$
Anti-HCV	**Linear**	C	1	1	1
		ε	0.1000	0.1000	0.1000
		RMSEC	0.0699	0.0771	0.0635
		r^2	0.9513	0.9460	0.9573
		$RMSE_{y\text{-}rand}$	0.0938	0.0940	0.0962
		$r^2_{y\text{-}rand}$	0.9479	0.9496	0.9489
		$^cr^2_p$ (y-rand)	0.0564	0	0.0900
		RMSECV	0.2791	0.2852	0.2814
		q^2	0.5987	0.6558	0.6185
		RMSEP	0.2096	0.2480	0.1981
		r^2_{test}	0.6804	0.6312	0.6866
		r^2_m (test)	0.4273	0.4460	0.3952
	Polynomial	C	0.25	0.25	0.25
		Degree	1	1	2
		Scale	0.0010	0.0100	0.001
		RMSEC	0.0698	0.0771	0.066
		r^2	0.9514	0.9460	0.9536
		$RMSE_{y\text{-}rand}$	0.0957	0.0964	0.0942
		$r^2_{y\text{-}rand}$	0.9499	0.9493	0.9510
		$^cr^2_p$ (y-rand)	0.0377	0	0.0500
		RMSECV	0.2736	0.2737	0.2957
		q^2	0.5800	0.6665	0.6117
		RMSEP	0.2098	0.2480	0.1611
		r^2_{test}	0.6797	0.6312	0.6473
		r^2_m (test)	0.4264	0.4459	0.1188

Table S.6
Internal and external validation parameters for anti-SARS-CoV compounds using SVR algorithm with linear and polynomial kernels.

Data set	Kernel function	Parameters	r_{atm}	ε	$r_{atm}/ε$
Anti-SARS-CoV	**Linear**	C	1	1	1
		ε	0.1000	0.1000	0.1000
		RMSEC	0.0498	0.0510	0.0507
		r^2	0.9479	0.9447	0.9460
		$RMSE_{y\text{-}rand}$	0.0908	0.0899	0.0910
		$r^2_{y\text{-}rand}$	0.9348	0.9384	0.9367
		$^cr^2_p$ (y-rand)	0.1114	0.0768	0.0940
		RMSECV	0.2104	0.2166	0.2113
		q^2	0.7689	0.7217	0.7003
		RMSEP	0.0916	0.0908	0.0907
		r^2_{test}	0.7406	0.7568	0.7322
		r^2_m (test)	0.2616	0.2860	0.2269
	Polynomial	C	0.25	0.25	0.25
		Degree	2	1	1
		Scale	0.0100	0.0100	0.0100
		RMSEC	0.0500	0.0510	0.0507
		r^2	0.9466	0.9447	0.9460
		$RMSE_{y\text{-}rand}$	0.0912	0.0905	0.0917
		$r^2_{y\text{-}rand}$	0.9374	0.9349	0.9359
		$^cr^2_p$ (y-rand)	0.0932	0.0962	0.0976
		RMSECV	0.2251	0.2131	0.2170
		q^2	0.6955	0.6908	0.7241
		RMSEP	0.0905	0.0908	0.0907
		r^2_{test}	0.7234	0.7568	0.7322
		r^2_m (test)	0.1994	0.2860	0.2269

Table S.7
Internal and external validation parameters for anti-HIV compounds using SVR algorithm with linear and polynomial kernels.

Data set	Kernel function	Parameters	r_{atm}	ε	$r_{atm}/ε$
Anti-HIV	Linear	C	1	1	1
		ε	0.1000	0.1000	0.1000
		RMSEC	0.0840	0.0833	0.0807
		r^2	0.9904	0.9923	0.9919
		$RMSE_{y\text{-}rand}$	0.0980	0.0971	0.0977
		$r^2_{y\text{-}rand}$	0.9929	0.9940	0.9934
		$^cr^2_p$ (y-rand)	0	0	0
		RMSECV	0.5995	0.7290	0.5885
		q^2	0.6078	0.5342	0.6254
		RMSEP	0.5001	0.3499	0.4607
		r^2_{test}	0.6464	0.8536	0.7107
		r^2_m (test)	0.3627	0.6857	0.3876
	Polynomial	C	0.25	0.25	0.25
		Degree	0.0100	0.0100	0.0010
		Scale	2	2	2
		RMSEC	0.0848	0.0824	0.0809
		r^2	0.9902	0.9925	0.9918
		$RMSE_{y\text{-}rand}$	0.0983	0.0983	0.0972
		$r^2_{y\text{-}rand}$	0.9932	0.9944	0.9941
		$^cr^2_p$ (y-rand)	0	0	0
		RMSECV	0.5866	0.7575	0.5777
		q^2	0.6498	0.4987	0.6397
		RMSEP	0.4866	0.8517	0.4435
		r^2_{test}	0.6640	0.3718	0.7257
		r^2_m (test)	0.3836	0.7401	0.3947

Table S.8
Internal and external validation parameters for the trilinear PLS model obtained from molecular slice images of the anti-HCV compounds.

Data set	Parameters	r_{atm}	ε	$r_{atm}/ε$
Anti-HCV	LV's	2	2	2
	RMSEC	0.1836	0.1804	0.2501
	r^2	0.5333	0.5984	0.5382
	$RMSE_{y\text{-}rand}$	0.2890	0.2702	0.2782
	$r^2_{y\text{-}rand}$	0.3428	0.4207	0.3889
	$^cr^2_p$ (y-rand)	0.3187	0.3260	0.2841
	RMSECV	0.2661	0.2377	0.2500
	q^2	0.2350	0.2977	0.2277
	RMSEP	0.1376	0.1278	0.2281
	r^2_{test}	0.7539	0.8298	0.7351
	r^2_m (test)	0.1999	0.3913	0.1706

Table S.9
Internal and external validation parameters for the trilinear PLS model obtained from molecular slice images of the anti-SARS-CoV compounds.

Data set	Parameters	r_{atm}	ε	$r_{atm}/ε$
Anti-SARS-CoV	LV's	3	3	3
	RMSEC	0.4545	0.4119	0.4267
	r^2	0.9120	0.9298	0.9234
	$RMSE_{y\text{-}rand}$	0.5077	0.4552	0.4982
	$r^2_{y\text{-}rand}$	0.8875	0.9111	0.8946
	$^{c}r^2_p$ (y-rand)	0.1494	0.1315	0.1631
	RMSECV	1.3268	1.3381	1.3142
	q^2	0.0847	0.0885	0.0924
	RMSEP	0.6354	0.6168	0.6254
	r^2_{test}	0.7071	0.6918	0.7135
	r^2_m (test)	0.4845	0.5161	0.4900

Table S.10
Internal and external validation parameters for the trilinear PLS model obtained from molecular slice images of the anti-HIV compounds.

Data set	Parameters	r_{atm}	ε	$r_{atm}/ε$
Anti-HIV	LV's	4	4	4
	RMSEC	0.3103	0.2997	0.3090
	r^2	0.8912	0.8891	0.9052
	$RMSE_{y\text{-}rand}$	0.5356	0.4826	0.5568
	$r^2_{y\text{-}rand}$	0.6974	0.7326	0.7073
	$^{c}r^2_p$ (y-rand)	0.4156	0.3730	0.4232
	RMSECV	0.6506	0.5627	0.5467
	q^2	0.3763	0.4288	0.5694
	RMSEP	0.6401	0.5807	0.5191
	r^2_{test}	0.2958	0.3321	0.1135
	r^2_m (test)	0.0779	0.0766	0

Table S.11
Internal and external validation parameters for the trilinear PLS model obtained from box face images of the anti-HCV compounds.

Data set	Parameters	r_{atm}	ε	$r_{atm}/ε$
Anti-HCV	LV's	3	3	3
	RMSEC	0.1407	0.1413	0.1372
	r^2	0.8223	0.8202	0.8331
	$RMSE_{y\text{-}rand}$	0.2182	0.2158	0.2191
	$r^2_{y\text{-}rand}$	0.6204	0.6300	0.6194
	$^{c}r^2_p$ (y-rand)	0.4074	0.3949	0.4219
	RMSECV	0.2412	0.2350	0.2453
	q^2	0.3410	0.3423	0.3111
	RMSEP	0.1685	0.2244	0.1785
	r^2_{test}	0.7940	0.6877	0.7598
	r^2_m (test)	0.5592	0.5010	0.4984

Table S.12
Internal and external validation parameters for the trilinear PLS model obtained from box face images of the anti-SARS-CoV compounds.

Data set	Parameters	r_{atm}	ε	$r_{atm}/ε$
Anti-SARS-CoV	LV's	2	3	2
	RMSEC	0.5056	0.2502	0.4965
	r^2	0.8894	0.9756	0.8939
	$RMSE_{y\text{-}rand}$	0.5959	0.3343	0.5885
	$r^2_{y\text{-}rand}$	0.8499	0.9525	0.8542
	${}^{c}r^2_p$ (y-rand)	0.1874	0.1502	0.1884
	RMSECV	1.1080	1.1533	1.1023
	q^2	0.1950	0.1612	0.1982
	RMSEP	0.7375	0.6136	0.7417
	r^2_{test}	0.6191	0.6692	0.6062
	r^2_m (test)	0.4721	0.5718	0.4522

Table S.13
Internal and external validation parameters for the trilinear PLS model obtained from box face images of the anti-HIV compounds.

Data set	Parameters	r_{atm}	ε	$r_{atm}/ε$
Anti-HIV	LV's	4	4	4
	RMSEC	0.2902	0.3054	0.2716
	r^2	0.8977	0.8909	0.9198
	$RMSE_{y\text{-}rand}$	0.4764	0.4651	0.5080
	$r^2_{y\text{-}rand}$	0.7403	0.7610	0.7294
	${}^{c}r^2_p$ (y-rand)	0.3759	0.3402	0.4184
	RMSECV	0.5594	0.5080	0.4986
	q^2	0.4828	0.5444	0.5826
	RMSEP	0.4804	0.4170	0.5956
	r^2_{test}	0.5905	0.5644	0.3499
	r^2_m (test)	0.2990	0.3445	0.1189

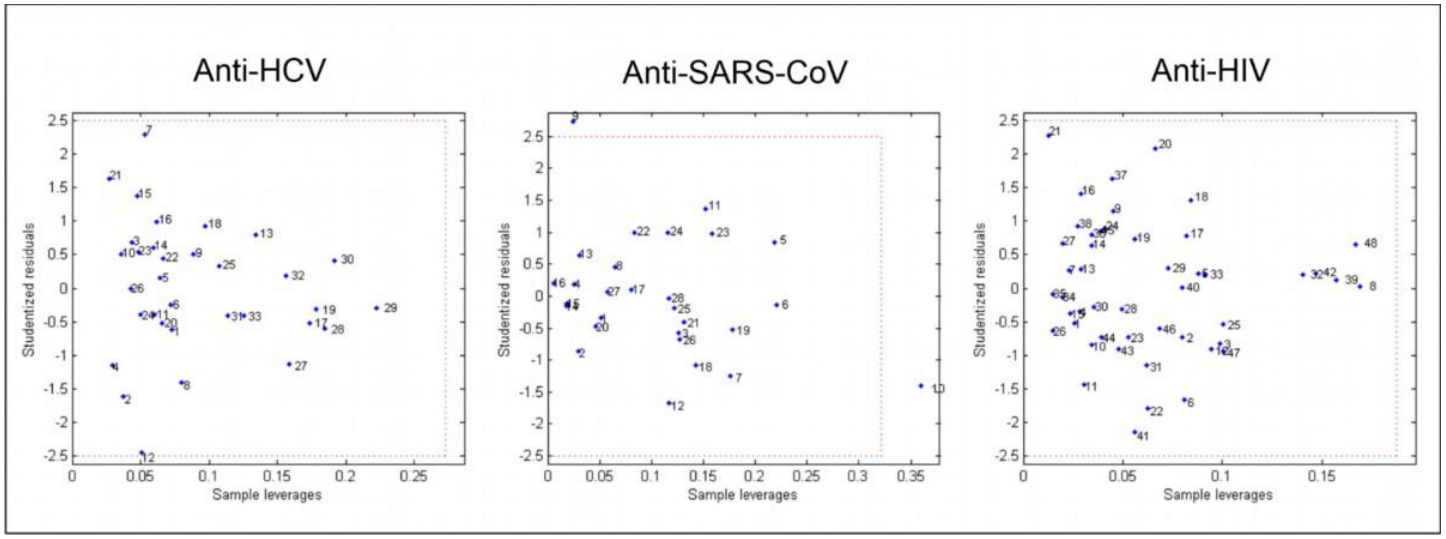


Fig. S.7. Applicability domain analysis by William's plot.

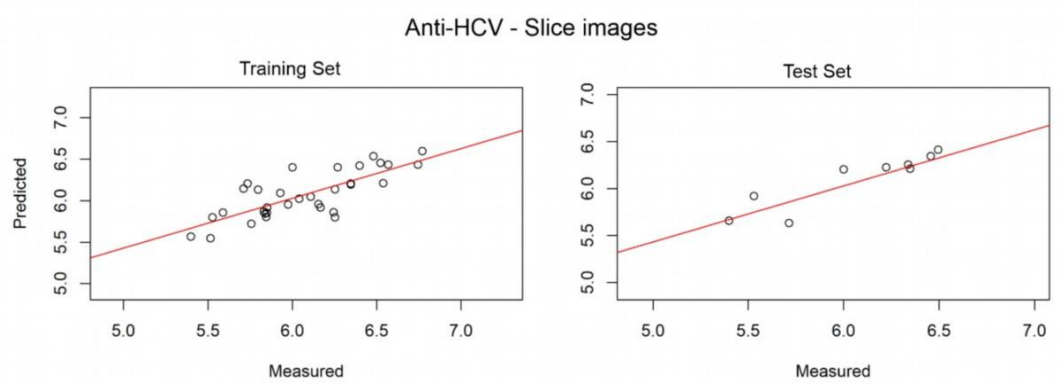


Fig. S.8. Predicted pEC_{50} × Measured pEC_{50} for the anti-HCV group with pixel values proportional to ε and trilinear PLS as regression method for the slice approach.

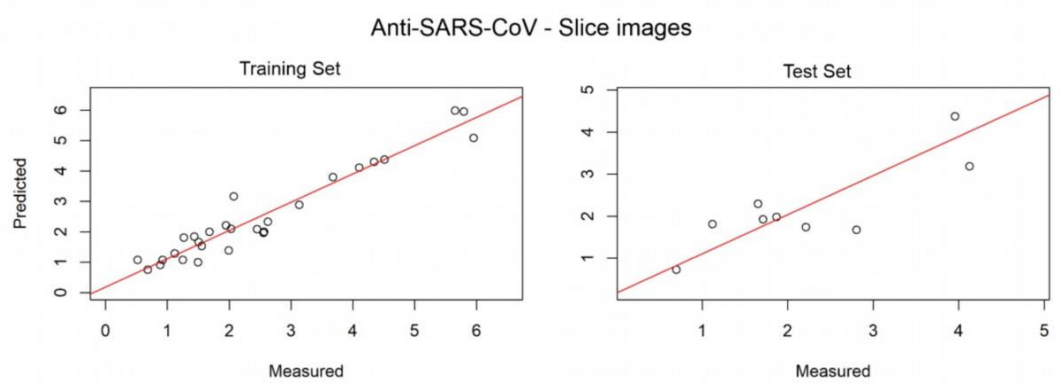


Fig. S.9. Predicted IC_{50} × Measured IC_{50} for the anti-SARS-CoV group with pixel values proportional to ε and trilinear PLS as regression method for the slice approach.

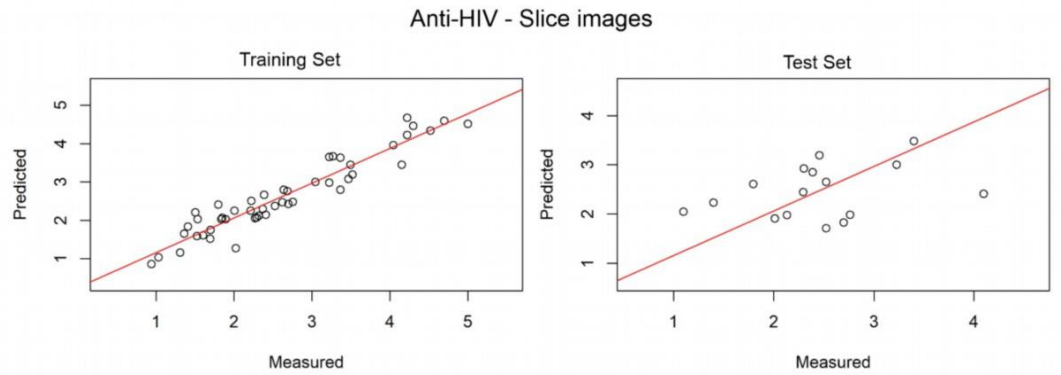


Fig. S.10. Predicted pIC_{50} × Measured pIC_{50} for the anti-HIV group with pixel values proportional to r_{atm}/ε and trilinear PLS as regression method for the slice approach.

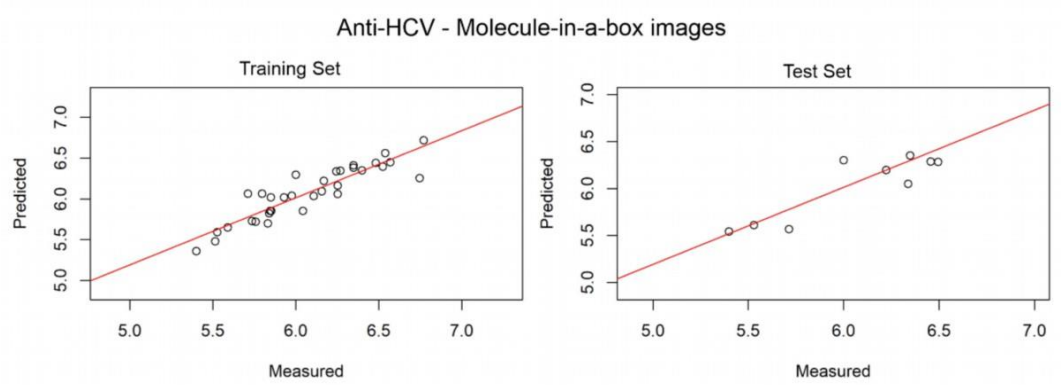


Fig. S.11. Predicted pEC_{50} × Measured pEC_{50} for the anti-HCV group with pixel values proportional to r_{atm} and trilinear PLS as regression method for the molecule-in-a-box approach.

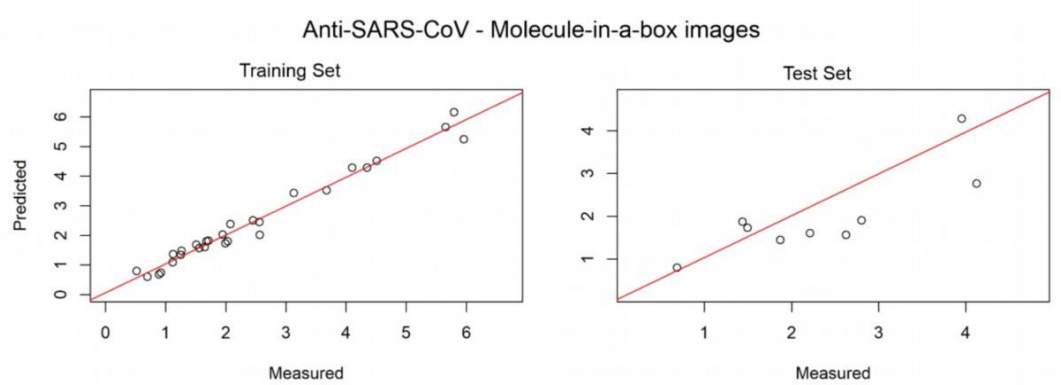


Fig. S.12. Predicted IC_{50} × Measured IC_{50} for the anti-SARS-CoV group with pixel values proportional to ε and trilinear PLS as regression method for the molecule-in-a-box approach.

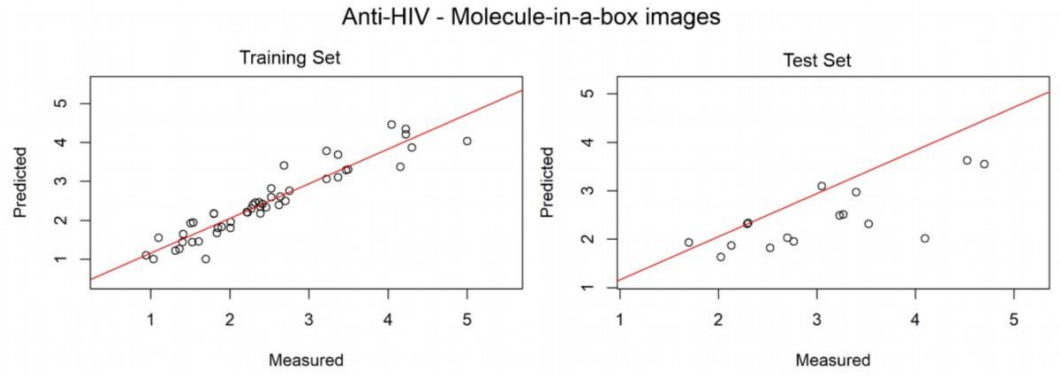


Fig. S.13. Predicted pIC_{50} × Measured pIC_{50} for the anti-HIV group with pixel values proportional to ε and trilinear PLS as regression method for the molecule-in-a-box approach.

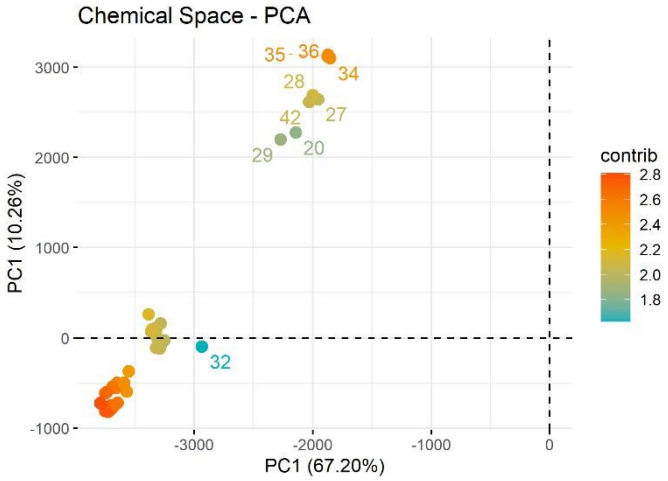


Fig. S.14. Chemical space representation for the anti-HCV data set obtained from the slice technique.

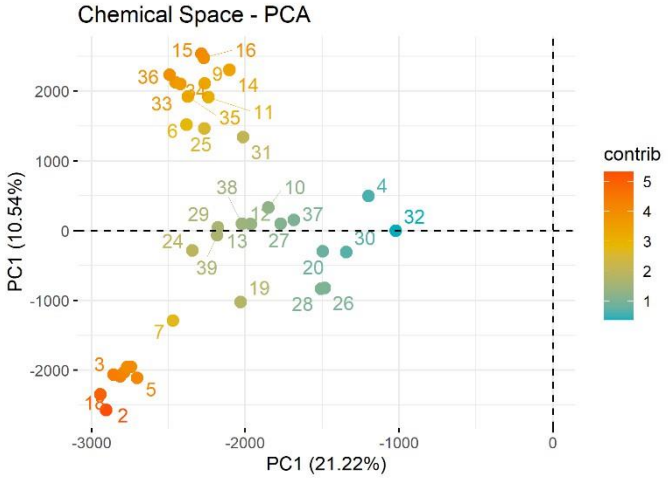


Fig. S.15. Chemical space representation for the anti-SARS-CoV data set obtained from the slice technique.

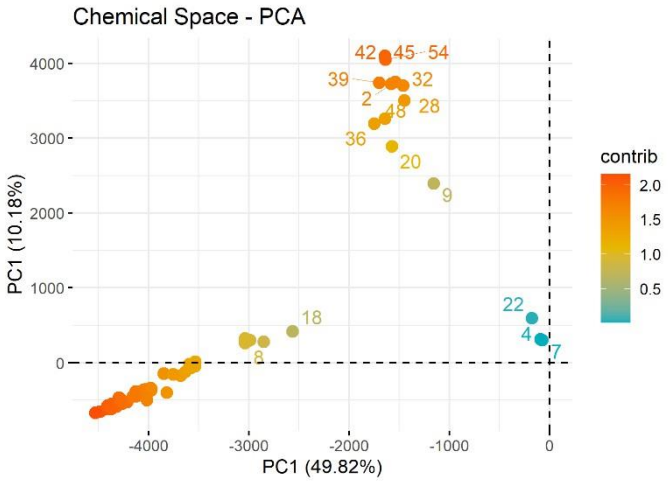


Fig. S.16. Chemical space representation for the anti-HIV data set obtained from the slice technique.

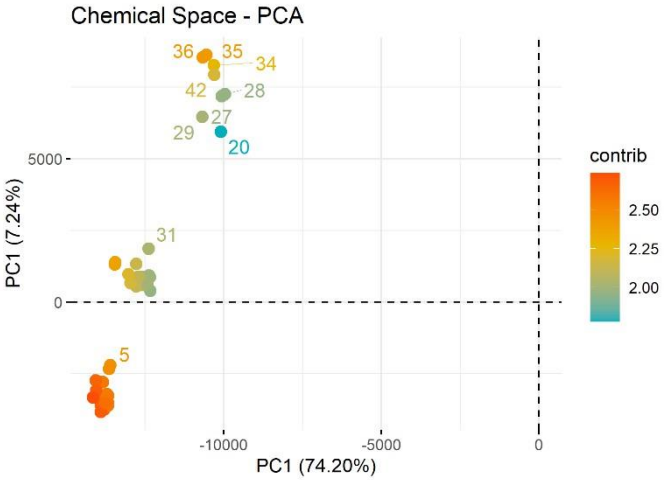


Fig. S.17. Chemical space representation for the anti-HCV data set obtained from the molecule-in-a-box technique.

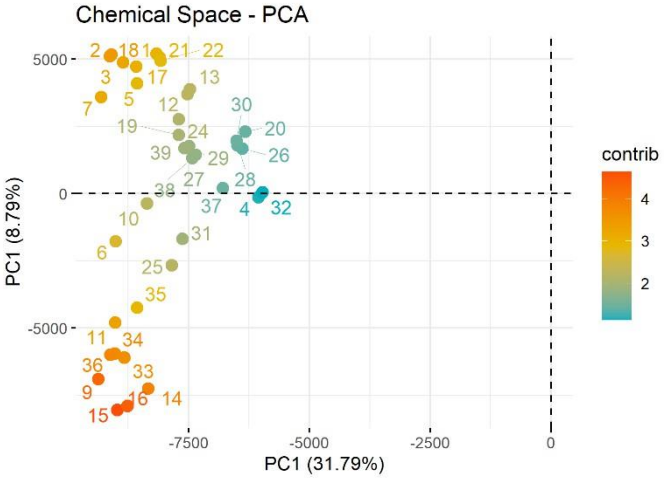


Fig. S.18. Chemical space representation for the anti-SARS-CoV data set obtained from the molecule-in-a-box technique.

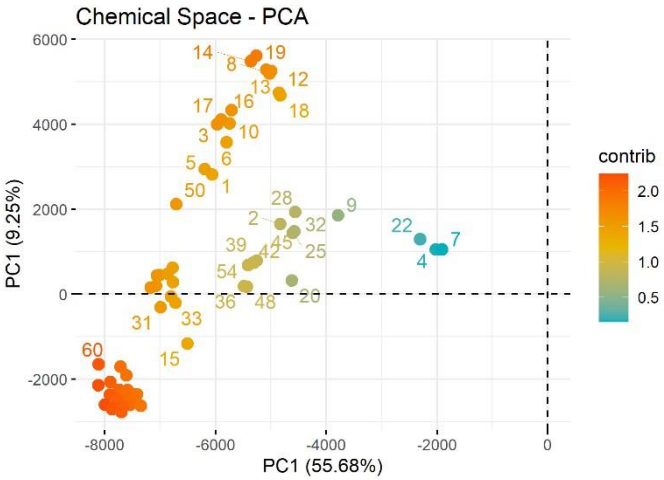


Fig. S.19. Chemical space representation for the anti-HIV data set obtained from the molecule-in-a-box technique.

Table S.14
Results of the traditional MIA-QSAR analysis from previous works [14, 15, 44].

Parameters	Anti-HCV	Anti-SARS-CoV	Anti-HIV
LV's	3	3	4
RMSEC	0.156	0.465	0.45
r^2	0.809	0.897	0.78
$RMSE_{y\text{-}rand}$	0.273	1.036	0.81
$r^2_{y\text{-}rand}$	0.411	0.488	0.28
${}^{c}r^2_{p}$ (y-rand)	0.567	0.606	0.55
RMSECV	0.225	0.794	0.62
q^2	0.604	0.701	0.60
RMSEP	0.269	0.715	0.55
r^2_{test}	0.635	0.781	0.80
r^2_{m} (test)	0.557	0.653	0.75

ARTIGO 2 – IS CONFORMATION RELEVANT FOR QSAR PURPOSES? 2D CHEMICAL REPRESENTATION IN A 3D-QSAR PERSPECTIVE

* Artigo apresentado na íntegra segundo as normas do periódico científico no qual foi publicado (*Journal of Computational Chemistry*).

Is conformation relevant for QSAR purposes? 2D chemical representation in a 3D-QSAR perspective

Joyce Karoline Daré[1] and Matheus P. Freitas[2]

Correspondence to: Matheus P. Freitas (E-mail: *matheus@ufla.br*)

[1,2] *Departamento de Química, Instituto de Ciências Naturais, Universidade Federal de Lavras, Lavras, MG, Brazil, 37200-900.*

ABSTRACT

Conformation has a key role in the mechanism of interaction between small molecules and biological receptors. However, encoding this type of information in molecular descriptors for the construction of robust QSAR models is not an easy task and, so far, the dependence of these models on such feature has not been thoroughly investigated. In the present study, the authors explore the effects of conformational information on a 3D-QSAR technique by comparing models built with descriptors that encode fully described tridimensional aspects (structures docked inside a biological target), with descriptors in which this information is suppressed (flat structures) or not fully described (structures with quantum-chemically optimized geometries). As a result, the validation parameters indicate that the robustness of the models seems to be more related to the alignment aspect of the structures than to how well their tridimensional features are described.

Introduction

Quantitative Structure-Activity Relationships (QSAR) are computational approaches used to model biological properties with both the goals of understanding the chemical features that explain these data and predicting the target response for new drug or agrochemical candidates, without the need for tests with living organisms.[1] The predictor variables used to correlate the chemical structures with the respective response variable are named molecular descriptors, which vary in dimension from simple chemical information (*e.g.* atom counting and connectivity) to the complex spatial representation of conformational ensembles.[2] One of the most successful QSAR techniques is that based on three-dimensional molecular force fields, such as CoMFA[3] and CoMSIA.[4] According to these methods, the molecules of a data set are placed inside a three-dimensional lattice and the fields (*e.g.* steric and electrostatic) are sampled at the intersections of this grid box. However, it has been debated whether a single conformation or even the method for selection of relevant conformations would provide reliable three-dimensional information;[5] therefore, a multi-instance learning approach has been proposed.[6,7] In turn, our question goes beyond this: which relevant conformational information do the 3D methods provide that 2D approaches should inform?

Multivariate Image Analysis applied to Quantitative Structure-Activity Relationships (MIA-QSAR) is an essentially 2D-QSAR technique, in which the MIA molecular descriptors correspond to pixels of chemical structure images plotted in a blackboard.[8] Until recently, these pixels were employed in QSAR modeling as simple combinations of values from the three corresponding RGB channels, then ranging from zero (black) to 765 (white). However, a deeper analysis has been proposed, in which the combined RGB values are replaced by atomic properties such as the van der Waals radius and the Pauling´s electronegativity.[9] Despite the chemical information decoded by this procedure, a major criticism lying on MIA-QSAR, and on 2D descriptors as a whole, is the lack of conformational representation necessary to explain the spatial requirement for ligand-enzyme fit-based bioactivity. Accordingly, efforts to evolve the MIA-QSAR technique by adding another dimension have been performed, but the images analyzed from the slicing of a 3D-array with the molecule inside, as well as from images obtained in three perspectives of the molecule placed in a box, did not improve the prediction capability of the models.[10] Therefore, we inquire whether conformational information is indeed an essential condition to obtain predictive and interpretable QSAR models. Accordingly, the goal of this study is to perform a thorough investigation of the effects of conformational information on a 3D-QSAR modeling technique, a commonly conformation-dependent

approach, by comparing models built with descriptors that encode fully described tridimensional aspects, with descriptors in which this information is suppressed or not fully described.

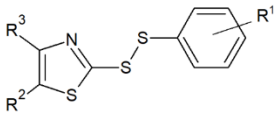
In this sense, 3D-QSAR models were built using molecules with anti-SARS-CoV activities[11] represented according to three different approaches: 1) with no pre-treatment and coincident substructures perfectly congruent and planar, such as in a 2D representation; 2) with molecular geometries quantum-chemically optimized to obtain the most stable conformations in the gas phase; 3) the molecules were previously docked into a biological target (PDB code: 2AMD) to generate the most likely bioactive conformations. In this way, the authors seek to obtain full evidences of the relevance of conformational description for QSAR purposes by comparing the results obtained from these three approaches.

Methods

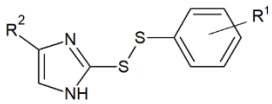
Design of the datasets

The chemical data set is formed by a series of unsymmetrical aromatic disulfide compounds, selected from the literature,[11] and is shown in Table 1 along with their respective biological activity expressed in terms of half maximal inhibitory concentration (IC_{50}, in μM).

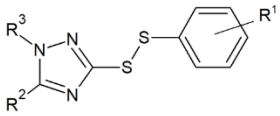
Table 1. Series of anti-SARS-CoV compounds used for MIA-QSAR modeling and the corresponding measured IC_{50} (μM).



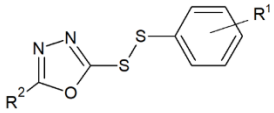
Comp.: **1-2, 5, 12-13, 18-25**



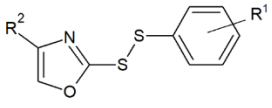
Comp.: **3, 14-17**



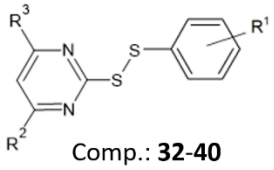
Comp.: **4, 6-11**



Comp.: **26-29, 31**



Comp.: **30**



Comp.: **32-40**

Molecule	R^1	R^2	R^3	IC_{50}	Molecule	R^1	R^2	R^3	IC_{50}
1	4-Cl	H	H	1.871	**21**[a]	4-Me	H	H	1.250
2	4-Me	NHCOMe	H	2.803	**22**	4-F	H	H	2.211
3	4-Cl	$COOC_2H_5$	-	3.675	**23**[b]	4-Br	H	H	3.321
4	2-NO_2	Me	$COCH_3$	3.130	**24**	2-NO_2	H	Me	2.555
5[a]	H	NHCOMe	H	1.506	**25**	2-$COOC_2H_5$	H	Me	2.452
6	4-Me	Ph	$COCH_3$	4.344	**26**	2-$COOCH_3$	Me	-	1.679

7[a]	4-OMe	Ph	$COCH_3$	4.100	**27**	2-$COOC_2H_5$	Me	-	1.557
8[b]	2-NO_2	3-pyridyl	$COCH_3$	1.762	**28**	2-NO_2	Me	-	1.713
9[b]	2-$COOC_2H_5$	3-pyridyl	$COCH_3$	5.654	**29**	2-$COOCH_3$	H	-	1.118
10	2-$COOC_2H_5$	4-pyridyl	$COCH_3$	4.511	**30**	2-$COOCH_3$	Me	-	1.264
11	4-OCH_3	3-pyridyl	$COCH_3$	5.794	**31**	4-Cl	H	-	0.516
12[a]	4-Cl	NHCOMe	H	2.626	**32**	2-NO_2	Me	Me	0.921
13	4-Br	NHCOMe	H	1.651	**33**[a]	4-Cl	Me	Me	1.437
14	2-NO_2	$COOCH_3$	-	2.075	**34**	4-Br	Me	Me	1.121
15	2-$COOC_2H_5$	$COOCH_3$	-	5.954	**35**	H	Me	Me	1.991
16	$COOCH_3$	$COOCH_3$	-	3.957	**36**[a]	4-Me	Me	Me	1.495
17	4-Cl	$COOCH_3$	-	4.126	**37**	2-NO_2	H	H	0.883
18[a]	4-F	NHCOMe	H	2.565	**38**	4-Cl	H	H	0.684
19	2-NO_2	NHCOMe	H	1.947	**39**[a]	4-Br	H	H	0.697
20	2-NO_2	H	H	2.029	**40**[b]	4-Me	H	H	1.522

[a] Test set. [b] Outliers.

Initially, the molecules were designed with the aid of the GaussView software;[12] they were submitted to three different routines to generate the desired representations, listed in the previous section. The first molecule group, used to construct model 1, was designed in such a way that the coincident substructures were perfectly congruent and planar; no pre-treatments were applied. For the second model, an initial conformational search was performed at the semi-empirical AM1 level of theory[13] using the Spartan'16 software;[14] the global energy minimum conformation of each compound was then re-optimized in a higher level of theory, ωB97X-D/6-31G(d,p),[15,16] in the Gaussian 09 software,[17] resulting in the final poses employed for the construction of the second QSAR model. To obtain the most likely bioactive conformations, required for the construction of the last model, the previously optimized structures were first submitted to charge calculations using the CHELPG (CHarges from ELectrostatic Potentials using a Grid-based method) function,[18] at same level of theory of the optimization step. The prepared ligands were then docked into the SARS-CoV M^{pro} (main protease) enzyme (PDB code: 2AMD) with the aid of the Glide ligand-receptor docking tool[19,20] using the standard precision protocol. A grid box with dimensions of 10×10×10Å was established and 100 conformations were generated. The most likely bioactive conformations were identified analyzing the similarity between the generated poses and the compound considered the most effective inhibitor (compound **31**) by Wang and collaborators.[11] More details can be found elsewhere.[21]

Next, the three groups of molecules were aligned using the tether tool available in the Discovery Studio Visualizer software;[22] the congeneric center was used as the reference moiety. Figure 1 shows the resulting aligned structures.

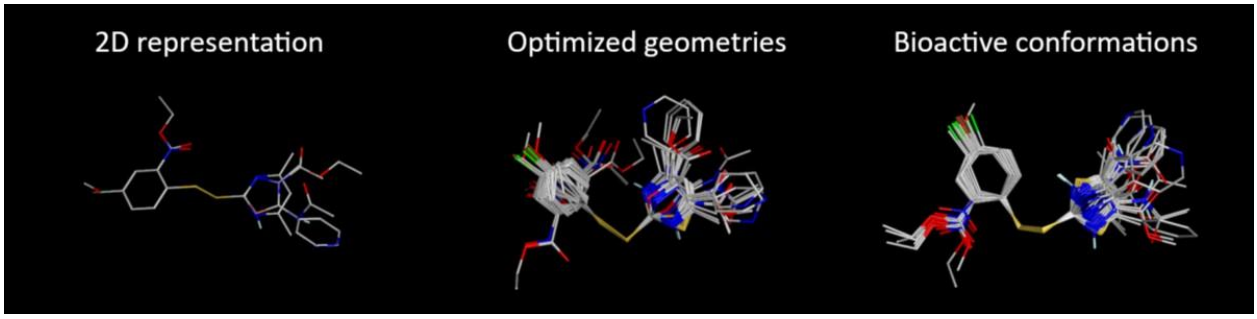


Figure 1. Resulting aligned structures from the three different designing procedures described herein.

QSAR analysis

QSAR analysis was performed using the Open3DQSAR software, a freely available tool for QSAR purposes.[23] Initially, the aligned subgroups were loaded into the software, then, for comparison purposes, samples **8**, **23** and **40** were removed from the data set, once they were considered outliers by Wang and collaborators.[11] After the removal, a grid box with a 1.5Å step size and a 4.0Å outgap was set. Next, steric and electrostatic molecular interaction fields (MIF) were computed using a carbon atom probe and a volume-less probe with +1 charge. The steric field is calculated based on AMBER FF99 van der Waals parameters[24] and the electrostatic field is based on a point charge model.[23] Their energy values were truncated at 30 kcal mol^{-1}.[25] Low energy values (< 0.05 kcal mol^{-1}) were set to zero.

After performing the pre-treatment operations above, the data sets were split into training and test set. Samples **5**, **7**, **9**, **12**, **18**, **21**, **33**, **36**, and **39** were selected for external validation, which were the same objects chosen by Wang and collaborators,[11] to maintain the same sample space for comparison purposes. After splitting the data, an operation for reducing the size of the PLS matrix was performed, the variables with low standard deviation (<0.1) were excluded. This procedure is normally performed after the splitting step to avoid the inclusion of test information in the selection of the training set; although, in this case, the order would not interfere, once the selection of the training set followed specific and independent criteria. A final operation was the removal of the *N*-level variables, that is, variables which assume only *N* values across the training set, most of which are distributed on a small number of objects. Their removal avoids overweighting the importance of particular substituent groups, maybe present in a single molecule, which might otherwise jeopardize the whole model. To give the same importance to all molecular interaction fields (MIFs) in the PLS matrix, regardless of the absolute magnitude of energy values, a collective block weighting coefficient (block unscaled weighting) was assigned to each MIF.

The molecular descriptors were regressed against the biological data using partial least squares (PLS) regression to obtain the 3D QSAR models. Cross-validation (CV) carried out by the leave-one-out (LOO) procedure was chosen to select the optimal number of components for PLS modeling and to check the statistical significance of the models. Also, variable clustering and selection was implemented to remove the less influential variables and facilitate the interpretation of the model. First, variable grouping based on their approximation in 3D space with Smart Region Definition (SRD) algorithm[26] was performed; this procedure reduces redundancy arising from the existence of multiple nearby descriptors, which basically encode the same kind of information.[23] Next, Fractional Factorial Design (FFD)[27,28] variable selection was carried out; it aims at selecting the groups identified at the previous SRD run which have the largest effect on predictability. It is worth mentioning that the last two steps were carried out using the optimal number of components defined in the LOOCV. After variable selection, the PLS models were re-computed, LOOCV was used again to determine the new optimal number of components and external validation was performed. The validation metrics used to evaluate the quality of the models were: the regression coefficient of determination (r^2), cross-validated r^2 (q^2), the external r^2 (r^2_{test}), and the SDEP (standard deviation error in prediction)/SDEC (standard deviation error in calculation) associated with each accuracy metric. The stability of the models was further assessed through a bootstrapping procedure, in which multiple test sets (10 subsets containing 9 samples each) were randomly selected, and the same calibration and validation steps previously described were performed for each cycle; the average of the statistical parameters were stored. Lastly, contour maps were generated for the models built with the Wang´s test set for visualization and interpretation purposes, which are shown in the next section along with the validation parameters obtained for the QSAR models.

Results and Discussion

From Figure 1, one can easily observe that the non-optimized molecules are the best-aligned structures; their variability is only associated with the difference on the nature of the substituent groups. On another hand, the two other data sets vary both in the chemical nature of the group and their spatial arrangements.

Regarding the QSAR analysis, the validation parameters for the final models, obtained after pre-treatment and variable selection, are shown in Table 2. To facilitate the comparison

among the statistical metrics, a scoring parameter (the average of the determination coefficients) was introduced.

Table 2. Internal and external validation parameters for anti-SARS-CoV compounds using PLS algorithm.

Validation Parameter	2D	Optimized	Docked
LV	2	3	2
r^2	0.827	0.914	0.772
SDEC	0.603	0.426	0.692
q^2	0.672	0.701	0.434
SDEP	0.829	0.649	1.092
r^2_{test}	0.899	0.814	0.473
SDEP	0.477	0.649	1.092
Score	0.799	0.809	0.560

CoMFA results from Wang et al.:[11] 6 LV´s, r^2 = 0.916, Standard Error = 0.088, and q^2= 0.681.

Comparing the values of the score parameter in Table 2, it can be seen that the simple optimization of the geometries does not seem to significantly improve the predictability of the QSAR model, once the scores of the models built with 2D and optimized structures were very similar. On another hand, approximating the chemical structures to their most likely bioactive conformations seems to jeopardize the quality of the model, indicating that no useful information has been encoded by this procedure.

The bootstrapping outcomes, shown in Table 3, assure the stability of the models and reassure the pattern observed before, i.e., the predictability of the models built with flat and optimized structures are similar, when their score parameters are compared, and both are superior to the model generated with docked structures.

Table 3. Average of the statistical results obtained for the 10 models built in the bootstrapping procedure.

Validation Parameter	2D	Optimized	Docked
$\overline{\mathbf{LV}}$	2	3.7 ± 0.675	2.1 ± 0.316
$\overline{\mathbf{r^2}}$	0.857 ± 0.026	0.935 ± 0.016	0.800 ± 0.0438
$\overline{\mathbf{SDEC}}$	0.544 ± 0.062	0.364 ± 0.044	0.642 ± 0.0694
$\overline{\mathbf{q^2}}$	0.736 ± 0.047	0.722 ± 0.043	0.483 ± 0.151
$\overline{\mathbf{SDEP}}$	0.739 ± 0.082	0.759 ± 0.058	1.028 ± 0.140
$\overline{\mathbf{r^2_{test}}}$	0.764 ± 0.094	0.680 ± 0.085	0.396 ± 0.162
$\overline{\mathbf{SDEP}}$	0.739 ± 0.216	0.864 ± 0.208	1.184 ± 0.269
$\overline{\mathbf{Score}}$	0.786 ± 0.056	0.779 ± 0.048	0.559 ± 0.119

Facing these results, the robustness of the models seems to be more related to the alignment aspect of the structures than to how well their tridimensional features are described. This finding is in agreement with our most recent study,[10] where the inclusion of 3D information

into MIA-QSAR descriptors implicated in worse models than those obtained with the conventional MIA-QSAR analysis, which is based on the alignment of congruent pharmacophoric substructures and disregards the spatial aspects of the molecules as in traditional 2D QSAR techniques.

Comparing the internal validation results obtained herein with those of Wang and collaborators[11] (see footnote of Table 2), one may notice that the models 1 and 2 present similar performances, although, a much smaller number of latent variables were required in our models, which is an advantage. On another hand, model 3 presented an inferior predictability performance.

Lastly, the electrostatic and steric contour maps are shown in Figure 2. These maps are used to rationalize the regions in 3D space around the molecules where changes in the steric and electrostatic fields are predicted to increase or decrease the observed property.[29] The electrostatic maps are represented by blue and red contours, which demonstrate the regions where an electron donating group and an electron-withdrawing group would be favorable, respectively.[30] On another hand, the steric field is represented by green and yellow contours, in which green contours indicate regions where bulky groups would be favorable, while the yellow contours represent regions where the opposite behavior is observed, that is, regions in which bulky groups decrease the observed property.[30]

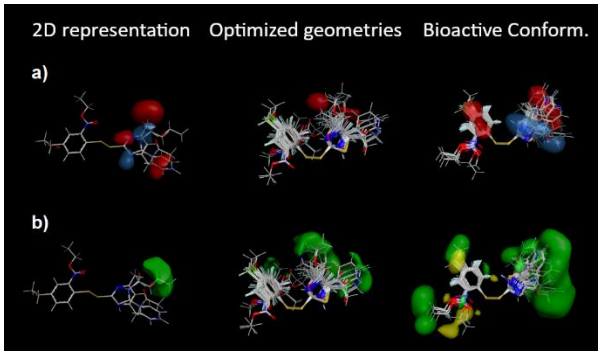


Figure 2. Electrostatic (a) and steric (b) contour maps.

Comparing the three electrostatic maps, one can notice a certain similarity in their distribution, all three attribute the strongest electrostatic effects on the activity to the moiety including the substituted ring; *i.e.*, the right side of the disulfide bond. However, a remarkable difference among them is that the map of the docked structures indicates a strong favorable contribution of electron-withdrawing groups on the substituted benzene ring, which is only slightly pointed out in the map of the planar structures and ignored in the map of optimized compounds. An important note is that the property of interest is expressed in terms of IC_{50}, thus, the positive effects observed in the maps implies in higher values of IC_{50}, which indicates a

smaller inhibitory activity; on another hand, the negative effects imply in a higher activity. Another consideration is that both maps from docked and planar structures indicate a significant favorable contribution (for increasing IC_{50}) of electron-donating groups around the right-side substituent, while this effect is only slightly pointed out in the map of optimized structures.

Regarding the steric contour maps, all three maps show clear evidences of regions where bulky groups would be favorable for increasing IC_{50}, but only the map of docked structures captured regions where bulky groups would decrease the values of the biological property.

In summary, the contour maps generated herein presented certain similarities in their distribution around the molecular regions, but also significant differences. The main differences in interpretation come from the map constructed with docked molecules, which is not surprising, considering the deviation among the statistical parameters obtained with these structures and the other two molecule groups. Moreover, the contributions indicated by the map of docked structures should be considered carefully, once the QSAR model generated with these structures did not present reliable statistical parameters.

Conclusions

A thorough analysis on the influence of conformational information in QSAR models was performed in this study with the goal of understanding the importance of this type of information in the generation of predictive and interpretable QSAR models. As a result, in terms of accuracy, the models built with descriptors encoding 2D molecular aspects and 3D information obtained from optimized structures were very similar, both presented robust validation parameters. On another hand, the model built with the most likely bioactive conformations showed a considerably worse predictive performance than the others. This finding indicates that the robustness of QSAR models seems to be more related to the alignment of the structures than to how well their tridimensional properties are described. In terms of interpretability, the contour maps generated from the three models showed significant discrepancies. Electrostatic maps obtained from the 2D structures and bioactive conformations showed more information than that generated with optimized geometries. On the other hand, both steric maps built with 3D molecular representations were more informative than that obtained with 2D structures. Therefore, if the purpose of a QSAR analysis is only to predict biological data for unknown samples, the 2D approximation seems to be the most suitable, as it avoids the complex steps of conformational screening and 3D alignment, and provides robust predictive models. On the other hand, if the objective is to explain the effects responsible for

the variation in biological data, then one should indeed consider 3D approaches, although there is no guarantee that the actual bioactive conformations are being analyzed.

Acknowledgments

Authors are thankful to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, funding code 001), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grant number 301371/2017-2), and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) for the financial support of this research.

Keywords: Conformational Information. 3D-QSAR. 2D representation. Optimized structures. Bioactive conformation.

References and Notes

1. Fujita, T. and Winkler, D. A. *J. Chem. Inf. Model.* **2016**, *56*, 269–274.
2. Todeschini, R.; Consonni, V.; Ballabio, D.; Grisoni, F. In Comprehensive Chemometrics; Brown, S. D.; Tauler, R.; Walczak, B., Eds.; Elsevier: Amsterdam, **2020**; Vol. 2, Chapter 4.25, pp 599-634.
3. Cramer, R. D., Patterson, D. E. and Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*,5959–5967.
4. Klebe, G., Abraham, U. and Mietzner, T. *J. Med. Chem.* **1994**, *37*, 4130–4146.
5. Gürsoy, O. and Smiesko, M. *J. Cheminform.* **2017**, *9*, 1–13.
6. Zankov, D.V., Matveieva, M., Nikonenko, A., Nugmanov, R., Varnek, A., Polishchuk, P. and Madzhidov, T. *ChemRxiv.* **2020**.
7. Nikonenko, A., Zankov, D., Baskin, I., Madzhidov, T. and Polishchuk, P. *Mol. Inf.* **2021**, *40*.
8. Barigye, S. J. and Freitas, M. P. *Int. J. Quant. Struct.-Prop. Relat.* **2016**, *1*, 64–77.
9. Freitas, M. R., Barigye, S. J. and Freitas, M. P. *RSC Adv.* **2015**, *5*, 7547–7553.
10. Daré, J. K. and Freitas, M. P. *Chemom. Intell. Lab. Sys.* **2021**, *212*, 104286 (2021).
11. Wang, L., Bao, B. - B., Song, G. - Q., Chen, C., Zhang, X. - M., Lu, W., Wang, Z., Cai, Y., Li, S., Fu, S., Song, F. - H., Yang, H. and Wang, J - G. *Eur. J. Med. Chem.* **2017**, *137*, 450–461.

12. Dennington, R. D., Keith, T. A. and Millam, J. M. GaussView 5.0, Gaussian, Inc., Wallingford, CT, USA, **2008**.

13. Dewar, M. J. S., Zoebisch, E. G., Healy, E. F. and Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

14. Spartan'16, Wavefunction, Inc.: Irvine, CA, USA, **2017**.

15. Chai, J. D. and Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*.

16. Krishnan, R., Binkley, J. S. and Seeger, R. P. *J. Chem. Phys.* **1980**, *72*, 650–654.

17. 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., Peralta, J. E., Ogliaro, F., Bearpark, M., Heyd, J. J., Brothers, E., Kudin, K. N., Staroverov, V. N., Keith, T., 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, O., Foresman, J. B., Ortiz, J. V., Cioslowski, J. and Fox, D.J. Gaussian 09, Revision D.01, Gaussian, Inc., Wallingford CT, USA, **2013**.

18. Breneman, C. M. and Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361–373.

19. Friesner, R. A., Banks, J. L., Murphy, R. B, Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H.; Shelley, M., Perry, J. K., Shaw, D. E, Francis, P., and Shenkin, P. S. *J Med Chem.* **2004**, *47*, 1739–1749.

20. Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L, Pollard, W. T., and Banks, J. L. *J Med Chem.* **2004**, *47*, 1750–1759.

21. Daré, J. K., Ramalho, T. C., and Freitas, M. P. *Mol. Simul.* **2020**, *46*,1055–1061.

22. Discovery Studio Visualizer 2017R2, Dassault Systèmes, BIOVIA: San Diego, CA, USA, **2017**.

23. Tosco, P. and Balle, T. *J. Mol. Model.* **2011**, *17*, 201–208.

24. Wang, J. M., Cieplak, P. and Kollman, P. A. *J. Comput. Chem.* **2000**, *21*,1049–1074.

25. Politi, A., Durdagi, S., Moutevelis-Minakakis, P., Kokotos, G., Papadopoulos, M. G. and Mavromoustakos, T. *Eur. J. Med. Chem.* **2009**, *44*, 3703-3711.

26. Pastor, M., Cruciani, G. and Clementi, S. *J. Med. Chem.* **1997**, *40*,1455–1464.

27. Baroni, M., Clementi, S., Cruciani, G., Costantino, G., Riganelli, D. *J. Chemometr*. **1992**, *6*, 347–356.

28. Baroni, M., Costantino, G., Cruciani, G., Riganelli, D., Valigi, R., Clementi, S. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9–20.

29. Pohjala, L., Alakurtti, S., Ahola, T., Yli-Kauhaluoma, J. and Tammela, P. *J. Nat. Prod.* **2009**, *72*, 1917–1926.

30. Ghasemi, J. B. and Shiri, F. *Med. Chem. Res.* **2012**, *21*, 2788–2806.

GRAPHICAL ABSTRACT

AUTHOR NAMES

Joyce Karoline Daré and Matheus P. Freitas

TITLE

Is conformation relevant for QSAR purposes? 2D chemical representation in a 3D-QSAR perspective

TEXT

The dependence of a 3D-QSAR technique on conformational information is investigated by comparing three QSAR models built with molecular descriptors containing different levels of tridimensionality description. The first model is built from 2D representations, where no conformational information is considered; a second model uses structures with optimized geometries; and, a final model employs the most likely bioactive conformations, obtained after docking the molecules into a biological target. The latter was not reliable.

GRAPHICAL ABSTRACT FIGURE