

Heliton Martins Reis Filho

**Química é Linguagem: Descritores 4D-QSAR via
Fecho Convexo para predição de atividade biológica**

Belo Horizonte

2025

Heliton Martins Reis Filho

**Química é Linguagem: Descritores 4D-QSAR via Fecho
Convexo para predição de atividade biológica**

Trabalho de conclusão de curso apresentado
ao Departamento de Química da Universi-
dade Federal de Minas Gerais como requisito
parcial para a obtenção de grau de Bacharel
em Química com ênfase em Química Com-
putacional.

Universidade Federal de Minas Gerais

Instituto de Ciências Exatas

Departamento de Química

Orientador: Prof. Dr. João Paulo Ataíde Martins

Belo Horizonte

2025

*“There are more things in heaven and earth, Horatio
Than are dreamt of in our philosophy.”¹*

— Hamlet ([Shakespeare, 1603](#))

¹ Tradução: “Há mais coisas no céu e na terra, Horácio, do que sonha a nossa pobre filosofia.”

Resumo

O desenvolvimento de novos fármacos é um processo custoso, longo e com baixas taxas de sucesso. Nesse contexto, metodologias de Relação Quantitativa Estrutura-Atividade (QSAR) consolidaram-se como ferramentas essenciais no Planejamento de Fármacos Assistido por Computador, permitindo a triagem virtual de compostos e a otimização do processo de descoberta. Este trabalho analisa e valida uma nova abordagem de geração de descritores 4D-QSAR denominada HullQSAR, que utiliza o conceito geométrico de fecho convexo para delimitar regiões de amostragem do espaço molecular fisicamente mais realistas. Essa metodologia foi validada em dois conjuntos de dados previamente modelados via LQTA-QSAR, os quais apresentam potencial de aplicação no tratamento do Câncer de Próstata e do vírus HIV-1. Os resultados demonstraram superioridade consistente do HullQSAR nos dois conjuntos, em comparação aos modelos obtidos na literatura. Assim, o HullQSAR representa um avanço significativo na modelagem 4D-QSAR, oferecendo uma representação molecular mais fiel à física das interações ligante-receptor, com potencial aplicação no planejamento racional de fármacos e em outras áreas da química computacional.

Palavras-chave: QSAR 4D, Descritores moleculares, Fecho convexo, Planejamento de fármacos, Modelagem preditiva.

Lista de ilustrações

Figura 1 – Evolução das bases de dados químicos e das modelagens QSAR ao longo dos anos.	15
Figura 2 – Fórmula estrutural da Água (H_2O)	16
Figura 3 – Ilustração de diferentes tipos de energia à luz da mecânica molecular	24
Figura 4 – Exemplo de um perfil de amostragem conformacional numa grade do LQTA-QSAR de 1 Å.	35
Figura 5 – Representação espacial dos melhores descritores LQTA-QSAR selecionados no conjunto CANCER para os compostos (a) menos e (b) mais ativos. Observe que a maioria dos descritores, representados pelas esferas coloridas, se localizam no interior da estrutura molecular. . .	37
Figura 6 – Exemplo de fecho convexo $\mathcal{H}(X)$ (linha vermelha) de um conjunto arbitrário $X \subset \mathbb{R}^2$ (pontos azuis)	39
Figura 7 – Estrutura base do conjunto de dados CANCER	42
Figura 8 – Estrutura base do conjunto de dados CARBOX	42
Figura 9 – Ilustração da diferença entre “precisão” e “exatidão”	57
Figura 10 – Vários conjuntos de pontos (x, y) , com o coeficiente de correlação entre x e y indicado para cada conjunto	59

Lista de tabelas

Tabela 2	– Exemplos de alguns compostos e suas representações em SMILES, fórmula molecular (FM) e nomenclatura IUPAC em português. . . .	21
Tabela 3	– Parâmetros hipotéticos que levariam o LQTA-QSAR a calcular energias fisicamente inverossímeis.	38
Tabela 4	– Hiper parâmetros utilizados no HullQSAR. Todos os valores numéricos estão em Å, exceto ℓ	43
Tabela 5	– Comparação entre o melhor modelo HullQSAR e o LQTA-QSAR nos conjuntos CANCER e CARBOX.	45
Tabela 6	– Resultados obtidos pelo HullQSAR para cada conjunto de dados, com o respectivo melhor modelo destacado em azul.	65

Lista de Abreviaturas e Siglas

- CADD** Planejamento de Fármacos Assistido por Computador. (Do inglês: *Computer-Aided Drug Design*). [p. 9, 12]
- CEP** Perfil de Amostragem Conformacional. (Do inglês: *Conformational Ensemble Profile*). [pp. 27, 34, 35, 38, 39, 40]
- CoMFA** Análise Comparativa de Campo Molecular. (Do inglês: *Comparative Molecular Field Analysis*). [pp. 34, 35, 57]
- DFT** Teoria do Funcional de Densidade. (Do inglês: *Density-Functional Theory*). [pp. 25, 26, 42]
- FDA** Administração Federal de Alimentos e Medicamentos dos Estados Unidos. (Do inglês: *Food and Drug Administration*). [p. 9]
- LLS** Quadrados mínimos lineares. (Do inglês: *Linear least squares*). [p. 32]
- LOO** *Leave-one-out*. [p. 54]
- MAD** Desvio Absoluto Médio. (Do inglês: *Mean Absolute Deviation*). [p. 52]
- MD** Dinâmica Molecular. (Do inglês: *Molecular Dynamics*). [pp. 26–28]
- ML** Aprendizado de Máquina. (Do inglês: *Machine Learning*). [p. 10]
- MM** Mecânica Molecular. [pp. 22–26]
- PLS** Quadrados mínimos parciais. (Do inglês: *Partial least squares*). [pp. 32, 36, 43, 47, 57]
- PRESS** Soma dos Quadrados dos Resíduos de Previsão. (Do inglês: *Prediction Residual Sum of Squares*). [pp. 32, 45, 47, 53, 54]
- QSAR** Relação Quantitativa Estrutura–Atividade. (Do inglês: *Quantitative Structure–Activity Relationship*). [pp. 9, 13, 16, 23, 27–30, 32, 34, 35, 40, 43, 45, 47, 50, 52, 54, 55, 58]
- RMS** Desvio Quadrático Médio. (Do inglês: *Root Mean Square Deviation*). [p. 52]
- RMSE** Erro Quadrático Médio. (Do inglês: *Root Mean Square Error*). [p. 45, 47]
- SMILES** Simplified Molecular Input Line Entry System. [pp. 19, 21, 23]

Lista de Símbolos

Símbolo	Descrição
$\Delta\alpha$	espessura angular considerada na discretização [p. 42]
Δr	incremento radial entre cascas sucessivas [p. 42]
IC_{50}	[p. 40]
ℓ	número total de camadas radiais incluídas no modelo [p. 42]
$\log P$	Coefficiente de Partição Octanol-Água [p. 13, 55]
pIC_{50}	Definido como $pIC_{50} = -\log(IC_{50})$ [p. 40]
Q^2	Poder de Predição do modelo. Também conhecido como Capacidade Preditiva ou <i>Goodness of Prediction</i> em inglês. [p. 32, 54]
r_0	raio inicial a partir do qual se inicia a amostragem radial (??) [p. 42]
R^2	Coefficiente de determinação ou Qualidade do Ajuste (do inglês <i>Goodness of Fit</i>). [p. 53, 54]

Sumário

	Sumário	8
1	INTRODUÇÃO	10
1.1	Objetivos	10
I	FUNDAMENTAÇÃO TEÓRICA	12
2	DESENVOLVIMENTO DE NOVOS FÁRMACOS	13
2.1	Planejamento de Fármacos Assistido por Computador (CADD)	13
2.2	Relação Quantitativa Estrutura-Atividade	14
3	MODELAGEM MOLECULAR	16
3.1	Todos os modelos estão errados!	17
3.2	Representação computacional de estruturas moleculares	20
3.3	Modelando à luz da física clássica: Mecânica Molecular	23
3.4	Modelando à luz da Mecânica Quântica	25
3.5	Dinâmica Molecular (MD)	27
3.6	Descritores Moleculares	29
4	MODELAGEM ESTATÍSTICA EM QSAR	31
4.1	Formalização Matemática da Relação Estrutura–Atividade	31
4.2	Regressão Linear	32
II	DESENVOLVIMENTO	34
5	MÉTODO PROPOSTO	35
5.1	LQTA-QSAR	35
5.2	Limitações do LQTA-QSAR	36
5.3	Fecho convexo	38
5.4	HullQSAR	38
6	METODOLOGIA EXPERIMENTAL	41
6.1	Dados experimentais	41
6.2	Modelagem Molecular	42
6.3	Simulações de Dinâmica Molecular	43
6.4	Cálculo dos descritores com o HullQSAR	43

6.5	Pré-processamento e Seleção de Variáveis	44
6.6	Construção e Validação dos Modelos QSAR	44
7	RESULTADOS E DISCUSSÃO	45
7.1	Análise Comparativa dos Modelos QSAR	45
7.2	Validação e Análise dos Critérios Estatísticos	46
III	CONSIDERAÇÕES FINAIS	47
8	CONCLUSÃO	48
	Referências	49
	Glossário	54
	APÊNDICES	55
	APÊNDICE A – MEDIDAS ESTATÍSTICAS	56
	APÊNDICE B – CLASSIFICAÇÃO DE DESCRITORES QSAR QUANTO À DIMENSIONALIDADE	61
B.1	QSAR “clássico”: 0D, 1D e 2D	61
B.2	QSAR 3D	62
B.3	QSAR 4D	63
B.4	Dimensionalidades mais altas: 5D, 6D e além	64
	APÊNDICE C – RESULTADOS DETALHADOS DO HULLQSAR	65
	APÊNDICE D – BLOCOS DE CÓDIGO COMPLETOS	66

1 Introdução

Em geral, de 3 a 6 anos são necessários para a descoberta de novos fármacos e o desenvolvimento pré-clínico; e os ensaios clínicos podem durar até 10 anos ou mais antes que o produto chegue ao mercado. Isso totaliza um investimento aproximado de 12 a 15 anos e cerca de US\$ 1,3 bilhão para trazer um novo fármaco ao mercado. Em média, entre 5000 e 10 000 compostos testados, cerca de 250 são selecionados para ensaios pré-clínicos. Desses, apenas 5 sobrevivem para entrar em ensaios clínicos, enquanto apenas um é aprovado pela FDA após uma rigorosa revisão do novo medicamento. (Das; Saha, 2017)

Diante desse panorama de alto custo, longo tempo de desenvolvimento e baixíssima taxa de sucesso, o Planejamento de Fármacos Assistido por Computador (CADD) consolidou-se como uma abordagem indispensável. Ao empregar a modelagem *in silico*, o CADD permite acelerar as etapas de descoberta, reduzir custos e aumentar as chances de sucesso ao priorizar os compostos mais promissores. Dentro deste campo, as metodologias de Relação Quantitativa Estrutura–Atividade (QSAR) são cruciais. Elas buscam estabelecer modelos matemáticos que correlacionam a estrutura química de uma molécula à sua atividade biológica, permitindo a triagem virtual de milhares de compostos e, assim, otimizando o processo de planejamento racional de novos fármacos.

1.1 Objetivos

O objetivo geral deste trabalho é demonstrar que a incorporação de restrições fisicamente realistas na geração de descritores 4D-QSAR resulta em modelos estatisticamente mais robustos e com maior capacidade preditiva. Para isso, uma minuciosa fundamentação teórica guia preliminarmente o leitor por uma reflexão aprofundada sobre a predição de propriedades biológicas (e químicas) e as questões filosóficas inerentes. Esse exercício técnico e epistemológico lançará as bases para analisarmos dois métodos de descrição molecular distintos, comparando-se seus resultados.

Dentre os objetivos específicos, destacam-se:

- Contextualizar a modelagem molecular preditiva nos âmbitos técnico e filosófico;
- Analisar algumas metodologias de geração de descritores para QSAR, como foco no LQTA-QSAR (Martins; Barbosa *et al.*, 2009) e no HullQSAR (Tenório, 2018);
- Selecionar conjuntos de dados reais que tenham sido previamente modelados via LQTA-QSAR;

- Modelar os conjuntos selecionados utilizando a metodologia HullQSAR;
- Validar estatisticamente os modelos obtidos;
- Comparar criticamente o desempenho dos modelos com as duas técnicas de descrição.

Com isso, espera-se contribuir para o avanço das ferramentas de modelagem molecular utilizadas no planejamento racional de fármacos, ampliando o repertório de descritores disponíveis para representar interações molécula-alvo de forma precisa, informativa e compatível com técnicas modernas de Aprendizado de Máquina.

Parte I

Fundamentação Teórica

2 Desenvolvimento de novos fármacos

Um fármaco¹ é, em essência, uma molécula de origem natural ou sintética que interage com processos biológicos e altera seu funcionamento normal. O objetivo central do planejamento de fármacos é descobrir o que pode ser considerado um “fármaco ideal” para prevenir, diagnosticar ou tratar uma doença. Este composto deve atender a um conjunto rigoroso de critérios: precisa ter ação específica no alvo biológico, ser seguro, não tóxico e apresentar o mínimo possível de efeitos colaterais. Além disso, deve ser estável, viável de sintetizar e possuir um equilíbrio delicado de solubilidade, sendo solúvel em água (para transporte na corrente sanguínea) e em lipídios (para atravessar as membranas celulares e atingir seu local de ação). (Doytchinova, 2022)

O planejamento de novos fármacos sempre representou um grande desafio científico, tecnológico e econômico. Historicamente, esse processo baseou-se em abordagens empíricas de tentativa e erro, nas quais compostos eram sintetizados e testados experimentalmente para se avaliar sua eficácia terapêutica. Essas abordagens, embora tenham levado à descoberta de medicamentos relevantes, frequentemente demandam décadas de pesquisa, mesmo com investimentos financeiros significativos e riscos ambientais e éticos consideráveis. (Niazi; Mariam, 2023)

2.1 Planejamento de Fármacos Assistido por Computador (CADD)

No final do século XX, com os avanços em biologia estrutural revelando as arquiteturas tridimensionais de biomoléculas (como o sequenciamento de proteínas por cristalografia de raios-X) e o crescimento exponencial do poder computacional, começava a se tornar viável a realização de simulações complexas de sistemas biológicos.

Essa convergência entre avanços computacionais, disponibilidade de dados estruturais e métodos estatísticos sofisticados culminou na consolidação e difusão do CADD, um campo interdisciplinar que visa otimizar o processo de descoberta de fármacos por meio da modelagem *in silico* de interações moleculares e da previsão de propriedades bioquímicas relevantes. O CADD tem como principal objetivo reduzir custos, acelerar etapas do desenvolvimento pré-clínico e aumentar as chances de sucesso de compostos candidatos ao orientar experimentalistas na priorização de estruturas promissoras (Niazi; Mariam, 2023).

Atualmente, diversas empresas farmacêuticas e instituições de pesquisa ao redor

¹ O termo fármaco deriva do grego *pharmakon*, que significa remédio ou veneno. O termo “droga” é um anglicismo derivado de “*drug*”, sendo amplamente utilizado na literatura científica como sinônimo de fármaco.

do mundo já contam com departamentos dedicados ao CADD, investindo continuamente no desenvolvimento e na otimização de metodologias e *softwares* voltados ao planejamento racional de fármacos (Lu *et al.*, 2018).

2.2 Relação Quantitativa Estrutura-Atividade

A ideia de que a estrutura de um composto químico determina suas propriedades e atividades biológicas remonta ao século XIX. Brown e Fraser (1868) foram provavelmente os primeiros a sugerir explicitamente a existência dessa relação, ao observarem a influência de características estruturais na ação fisiológica de diferentes substâncias. Posteriormente, Mills (1884) aprofundou esse entendimento ao demonstrar como grupos funcionais específicos afetavam propriedades físico-químicas mensuráveis, consolidando os primeiros indícios de uma correlação sistemática entre estrutura molecular e comportamento químico.

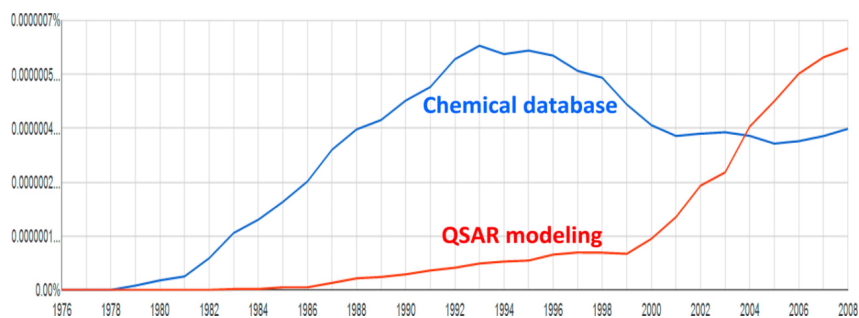
Esse arcabouço conceitual ganharia sofisticação nas décadas seguintes, culminando nos estudos de Hammett (1935). Investigando a influência de substituintes aromáticos sobre as constantes de equilíbrio e velocidades de reação de ácidos benzoicos, Hammett introduziu uma equação quantitativa baseada em parâmetros eletrônicos, hoje conhecidos como Constantes de Hammett (σ). Este trabalho é amplamente reconhecido como a primeira análise formal de uma relação quantitativa entre estrutura e atividade, estabelecendo um elo matemático entre propriedades estruturais e reatividade química.

Entretanto, a verdadeira consolidação da QSAR enquanto metodologia preditiva só ocorreria décadas mais tarde, com a contribuição seminal de Hansch e Fujita (1964). Empregando regressão linear múltipla, os autores integraram múltiplos parâmetros físico-químicos — incluindo as constantes de Hammett, o coeficiente de partição octanol-água ($\log P$) e parâmetros de lipofilicidade (π) — para modelar a atividade biológica de compostos orgânicos. Seus estudos abrangeram, entre outros exemplos, a avaliação de ácidos benzóicos em larvas de mosquitos, fenóis em bactérias gram-positivas e gram-negativas, e inseticidas organofosforados em moscas. Pela primeira vez, tornou-se possível prever valores de atividade para novos compostos dentro de limites estatísticos aceitáveis, a partir de propriedades estruturais previamente determinadas.

A partir dessa base, o campo da QSAR evoluiu de forma exponencial. Inicialmente restrito a pequenas séries de compostos congêneres e a modelos estatísticos simples, como a regressão linear, o QSAR ampliou seu escopo metodológico e computacional nas décadas seguintes. Hoje, abrange o processamento de grandes bases de dados químicas contendo milhares de estruturas moleculares diversas, empregando técnicas de aprendizado de máquina e algoritmos de inteligência artificial. A Figura 1 ilustra a correlação entre a expansão das bases de dados químicos disponíveis e o crescimento exponencial do número

de estudos QSAR publicados ao longo dos anos, evidenciando a crescente importância dessa abordagem na química medicinal e na ciência dos materiais. (Cherkasov *et al.*, 2014)

Figura 1 – Evolução das bases de dados químicos e das modelagens QSAR ao longo dos anos.



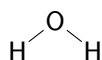
Fonte: Cherkasov *et al.* (2014).

3 Modelagem Molecular

A elucidação da estrutura molecular constituiu, em sucessivas etapas históricas, um dos mais paradigmáticos avanços científicos da Química e das Ciências da Vida. Cada salto conceitual — do plano bidimensional de fórmulas empíricas à descrição tridimensional e eletrônica das moléculas — abriu novos domínios de investigação e de aplicação tecnológica. Todas essas representações são, em essência, codificações simbólicas e parciais da realidade molecular.

Cada notação, conforme o contexto e objetivos com os quais é desenvolvida, privilegia determinados aspectos da molécula, por exemplo: sua composição atômica (como a fórmula molecular H_2O para a água), a conectividade entre os átomos (como a fórmula estrutural plana, $\text{H}-\text{O}-\text{H}$, para mesma molécula), a geometria espacial (veja [Figura 2](#)), suas propriedades eletrônicas *et cetera*.

Figura 2 – Fórmula estrutural da Água (H_2O)



Fonte: elaborado pelo autor.

Essas representações, conquanto eficazes para a comunicação e manipulação de estruturas químicas, não codificam, de maneira explícita, informações relativas à atividade biológica, à toxicidade, à farmacocinética ou a outras propriedades complexas de interesse humano. Essa limitação natural decorre do fato de que a atividade biológica resulta de múltiplas interações contextuais — com receptores, enzimas, transportadores, membranas e fluidos biológicos. Cada uma dessas interações depende de características específicas da estrutura molecular, como fatores estereoeletrônicos, conformacionais e dinâmicos ([Jensen, 2017](#)).

Focalizemos, por exemplo, a polaridade, uma propriedade que reflete a distribuição de cargas elétricas ao longo da estrutura. “A polaridade dos grupos funcionais tem profundo impacto nas propriedades eletrônicas dos fármacos e, em consequência, em suas propriedades físico-químicas, farmacocinéticas e farmacodinâmicas” ([Ferreira *et al.*, 2022](#)). Essa característica, por si só, já afeta diretamente a solubilidade, a capacidade de atravessar membranas biológicas e, sobretudo, a afinidade da molécula por sítios ativos de proteínas.

Um sítio ativo enzimático com resíduos polares (como serina, asparagina ou glutamina) tende a estabelecer interações mais estáveis com ligantes que também possuem regiões polares, por meio de ligações de hidrogênio ou interações dipolo-dipolo. Se, por outro lado, o ligante for predominantemente apolar, ele poderá ter dificuldade em se

posicionar corretamente no sítio ativo, levando a uma menor afinidade — e, portanto, menor atividade biológica.

Uma propriedade como a polaridade poderia ser descrita numericamente em termos de métricas como momento dipolar, superfície polar, dentre outras. Essas mesmas métricas são, portanto, descritores promissores cuja variação entre diferentes moléculas é potencialmente correlacionável quantitativamente com a atividade observada. Esse tipo de consideração estrutural é essencial em modelagem QSAR: uma molécula com uma distribuição de carga mais compatível com a do alvo terá maior probabilidade de formar interações estáveis e eficazes, o que se traduz em maior potência farmacológica.

Todavia, a polaridade é apenas um dentre os inúmeros fatores que governam essa suposta interação. Esses fatores não atuam de forma isolada, mas se combinam e se modulam mutuamente, compondo um panorama multifatorial que desafia abordagens puramente qualitativas. Em suma, as propriedades de uma substância não são diretamente legíveis a partir de sua estrutura química — ao menos não sem um processo de inferência mediado por modelos e conhecimento empírico (Martins, 2010).

3.1 Todos os modelos estão errados!

A representação computacional da realidade é uma tarefa de grande complexidade, que envolve primariamente decompor um universo contínuo e infinito em representações numéricas discretas e de precisão finita. Tratando-se de uma tarefa logicamente impossível, depreende-se que, na verdade, todo e qualquer modelo matemático ou computacional da realidade será, em última instância, uma representação parcial da realidade. “Como todos os modelos estão errados, o cientista não pode obter um modelo ‘correto’ através da elaboração excessiva” (Box, 1976).

Além disso, como teria dito Hamlet a Horácio, “há mais coisas no céu e na terra do que sonha nossa pobre filosofia”¹ (Shakespeare, 1603). Isso significa que, por melhor e mais qualificadas que sejam a habilidade e a intenção do cientista, sempre haverá simplificações excessivas e detalhes importantes desprezados.

Com efeito, a máxima “todos os modelos estão errados, mas alguns são úteis” (Box, 1979, tradução própria) causou tamanho impacto que se popularizou e tem intitulado diversos trabalhos na área ainda hoje (Warren; Ryan; Gan, 2025; Curchoe, 2020; Fisher; Rudin; Dominici, 2018; Wit; Heuvel; Romeijn, 2012).

Esse questionamento, próprio da filosofia, produz eco suficiente em suas áreas correlatas — dentre as quais figura a química, sobre a qual se centra o presente trabalho. Sob risco de ceder ao reducionismo, não será, portanto, possível prosseguir sem o enfrentar

¹ “*There are more things in heaven and earth, Horatio, than are dreamt of in our philosophy.*”

adequadamente.

O filósofo alemão Albert Camus, em seu “O Mito de Sísifo” (Camus, 1942), trata de analisar o inevitável *absurdo* existente no abismo entre a *realidade* propriamente dita e a sua *interpretação* feita por nós humanos. Ele reconhece que já naquele momento não se havia esperança no “verdadeiro conhecimento”:

Excetuando-se os racionalistas por profissão, hoje já não se tem esperança do verdadeiro conhecimento. [...] De quem e de que, de fato, posso dizer “conheço isso”? Este coração, em mim, posso experimentá-lo e julgo que ele existe. Este mundo, posso tocá-lo e julgo ainda que ele existe. Para aí toda a minha ciência, o resto é construção. Porque, se tento agarrar este eu de que me apodero, se tento defini-lo e sintetizá-lo, ele não é mais do que uma água que corre entre meus dedos. [...] Até este coração que é o meu continuará sendo sempre, para mim, indefinível. Entre a certeza que tenho da minha existência e o conteúdo que tento dar a essa segurança, o fosso jamais será preenchido. Serei para sempre um estranho diante de mim mesmo (Camus, 2018).

Para o autor, não é possível conhecer, com *certeza*, nem nossa própria individualidade — que dirá as particularidades do universo! Para ele, a lógica se aproxima da psicologia na medida em que existem verdades, mas não há verdade:

Em psicologia, como em lógica, há verdades mas não há verdade. O “conhece-te a ti mesmo” de Sócrates tem tanto valor quanto o “sê virtuoso” do nossos confessionários. Revelam uma nostalgia, ao mesmo tempo que uma ignorância. São jogos estéreis sobre grandes assuntos. São legítimos apenas na medida exata em que são aproximativos (Camus, 2018).

De fato, todos os modelos que somos ou seremos capazes de elaborar sofrem dessa mesma limitação última, fruto da inacessibilidade da realidade aos nossos métodos, ou fruto do *absurdo* que Camus defende.

Se desejarmos voltar nossas atenções à química, poderemos seguir analisando o referido ensaio de Camus (1942):

Contudo, toda a ciência dessa terra não me dará nada que me possa garantir que este mundo é para mim. Vocês o descrevem e me ensinam a classificá-lo. Vocês enumeram suas leis na minha sede de saber, concordo que elas sejam verdadeiras. Vocês desmontam seu mecanismo e minha esperança aumenta. Por último, vocês me ensinam que esse universo prestigioso e colorido se reduz ao átomo e que o próprio átomo se reduz ao elétron. Tudo isso é bom e espero que vocês continuem. Mas vocês me falam de um invisível sistema planetário em que os elétrons gravitam ao redor de um núcleo. Vocês me explicam esse mundo com uma imagem. Reconheço, então, que vocês enveredam pela poesia: nunca chegarei ao conhecimento. Tenho tempo para me indignar com isso? Vocês já mudaram de teoria. Assim, essa ciência que devia me ensinar tudo se limita à hipótese, essa lucidez se perde na metáfora, essa certeza se resolve como obra de arte. Para o que é que eu precisava de tantos esforço? (Camus, 2018)

A dura crítica que Camus faz à ciência positivista nesse trecho exige de nós tranquilidade e clareza para ser analisada. Sua reflexão nos remete à própria essência do que compreendemos como *representação científica*: uma construção abstrata, uma imagem formulada com símbolos e conceitos cuja validade se resume à sua utilidade e coerência interna, não à fidelidade absoluta à realidade última. Para Camus, cada avanço científico, ainda que eficaz na resolução de problemas práticos, é essencialmente uma aproximação, permeada por analogias e metáforas.

A imagem do átomo como um pequeno sistema planetário serve como um exemplo icônico dessa abordagem. Embora seja didática e inspiradora, Camus nos alerta que tal analogia é uma “poesia conveniente”, mais do que uma verdade fundamental. A história da química, com suas constantes revoluções teóricas, testemunha esse movimento incessante em direção a metáforas mais refinadas, modelos que buscam capturar a essência das observações empíricas sem jamais esgotá-las. A ciência positivista criticada pelo autor falha não por incompetência, mas por abominar a natureza intrinsecamente humana do empreendimento científico, perseguindo ilusões de “imparcialidade” absoluta.

Sob essa ótica, a Química deve ser compreendida não apenas como ciência experimental, mas como uma linguagem de representação. Fenômenos fundamentais — como o reconhecimento molecular entre um fármaco e seu receptor, a colisão entre reagentes ou a ação de um catalisador — ocorrem no domínio das propriedades moleculares. Visto que tais propriedades são consequências diretas da estrutura da matéria, é a linguagem química que torna esse comportamento inteligível. Ao codificarmos a estrutura em símbolos, traduzimos a realidade física para uma gramática manipulável, permitindo-nos prever interações que dependem da arquitetura invisível das moléculas.

Tal mediação simbólica não é uma idiosincrasia da Química, mas o *modus operandi* de toda construção científica robusta. Na Ciência da Computação, a complexidade do *hardware* e os fluxos binários são encapsulados por linguagens de programação; operamos sobre uma camada de abstração que traduz a lógica humana para a máquina. De modo análogo, a Biologia ordena a vertiginosa diversidade orgânica através da taxonomia, criando categorias que, embora sejam construções intelectuais, são indispensáveis para compreendermos as relações evolutivas. Em última instância, a ciência avança não por tocar a “coisa em si”, mas por aprimorar as interfaces que nos permitem interagir com ela.

É sob essa perspectiva crítica e filosófica que devemos interpretar a química computacional. Os modelos moleculares que exploraremos a seguir são representações simbólicas carregadas de simplificações práticas. Compreender suas incertezas inerentes é o primeiro passo para utilizá-las com rigor.

Doravante, nos debruçamos sobre esse interessante desafio e exploraremos as técnicas de representação molecular no espaço computacional.

3.2 Representação computacional de estruturas moleculares

A [Figura 2](#) representa, de forma visual, a molécula de água com seus três átomos, as duas ligações químicas entre eles e a distribuição geométrica desses átomos. Ela utiliza-se, portanto, do sentido da visão que muitos de nós, humanos, possuímos. Imagine-se, agora, ensinando química a um aluno com deficiência visual. Essa representação rapidamente perde sua relevância e se impõe a necessidade de representar a mesma estrutura de outras formas.

Isso reforça a noção de que não procuramos uma representação meramente “correta” da realidade², mas sobretudo uma representação “útil” aos nossos interesses!

Computacionalmente, nos encontramos na mesma situação. Embora a [Figura 2](#) seja apresentada como uma bonita e reconhecível notação bidimensional, ela foi, na verdade, gerada computacionalmente através do código $\text{\LaTeX}$ a seguir.

```
1 \chemfig{H-[:37.775]O-[:-37.775]H}
```

Este código, aparentemente em nada relacionado à molécula de água, instrui claramente ao compilador $\text{\LaTeX}$ as informações importantes para gerar a figura. Em particular, são dadas as instruções:

```
1 \chemfig {
2   H           % represente um átomo de H
3   -[:37.775]  % represente uma ligação simples a 37,775°
4   O           % represente um átomo de O
5   -[:-37.775] % represente uma ligação simples a -37,775°
6   H           % represente um átomo de H
7 }
```

Embora útil para a representação estrutural em um documento como este, outros formatos mais simples e poderosos surgiram. [Weininger \(1988\)](#) propôs a notação **SMILES**.

Embora o **SMILES** tenha se tornado e permaneça como um dos principais formatos de transporte digital de informações químicas, ele pode ser considerado “notadamente subespecificado”:

Mas a despeito de sua ampla adoção, o **SMILES** é notavelmente subespecificado. Por exemplo, o artigo original de Weininger, datado de 1988, deixou vários temas como exercício ao leitor. A palavra “estereoquímica” não foi mencionada nenhuma vez. A sintaxe dessa “linguagem” foi descrita apenas de forma anedótica. Algumas considerações semânticas — como

² Essa simples hipótese poderia causar, no leitor mais interessado pela Filosofia, uma avalanche de questões incontornáveis, disparadas pelo emprego de palavras como “correta” e “realidade”. Embora questões de interessante natureza, as reservaremos oportunamente a um estudo futuro.

Tabela 2 – Exemplos de alguns compostos e suas representações em SMILES, fórmula molecular (FM) e nomenclatura IUPAC em português.

SMILES	FM	IUPAC	SMILES	FM	IUPAC
CC	C ₂ H ₆	etano	C1CCCCC1	C ₆ H ₁₂	cicloexano
O=C=O	CO ₂	dióxido de carbono	c1ccccc1	C ₆ H ₆	benzeno
C#N	HCN	cianeto de hidrogênio	F/C=C/F	C ₂ H ₂ F ₂	<i>trans</i> -difluoroeteno
CCN(CC)CC	C ₆ H ₁₅ N	trietilamina	F/C=C\F	C ₂ H ₂ F ₂	<i>cis</i> -difluoroeteno
CC(=O)O	CH ₃ COOH	ácido acético	N[C@@H](C)C(=O)O	C ₃ H ₇ NO ₂	L-alanina
[OH3+]	H ₃ O ⁺	íon hidrônio	N[C@H](C)C(=O)O	C ₃ H ₇ NO ₂	D-alanina

Fonte: Adaptado de Daylight CIS (2011).

a contagem de hidrogênios e a aromaticidade — foram abordadas apenas de maneira parcial. (Apodaca, R. L., 2021, tradução própria)

Desde então, distintos esforços têm sido direcionados à criação de superconjuntos do SMILES (Apodaca, R., 2022) ou mesmo linguagens totalmente novas (Murray-Rust; Rzepa, 1999).³

3.2.1 Trade-off: dilemas computacionais intrínsecos

Em Ciência da Computação, o termo em inglês *trade-off*, que pode ser traduzido livremente como “troca” ou “compromisso”, descreve um dilema onipresente: ao melhorar um aspecto de um sistema, quase sempre somos obrigados a renunciar a outro. Trata-se de um equilíbrio inevitável, no qual ganhos em uma dimensão significam perdas em outra.

Um exemplo bem ilustrativo vem da área de proteção de dados. Ao coletar informações pessoais, como a localização de usuários em um projeto de ciência cidadã, existe uma tensão clara entre proteger a *privacidade* das pessoas e garantir a *utilidade* dos dados para análises futuras. Quanto mais protegemos a privacidade — por exemplo, adicionando ruído estatístico aos dados —, menos informativos e precisos eles se tornam. Por outro lado, ao buscar dados mais precisos e detalhados, inevitavelmente reduzimos a privacidade dos participantes (Franzen; Müller-Birn; Wegwarth, 2024). Essa situação representa um *trade-off* clássico: ganhos em privacidade significam perdas em utilidade, e vice-versa.⁴

Outro exemplo frequente em aprendizado de máquina é o *trade-off* entre *custo computacional* e *desempenho preditivo*. Melhorar o desempenho de um modelo — seja

³ Ao leitor que se interessar pelos detalhes envolvidas no processo de elaboração de uma linguagem de marcação para estruturas químicas, sugere-se o trabalho de Murray-Rust e Rzepa (1999) como um bom ponto de partida.

⁴ Nesse exemplo em particular, a “solução” ética proposta por Franzen, Müller-Birn e Wegwarth (2024) consiste em comunicar claramente esse *trade-off* aos usuários, permitindo que eles façam escolhas informadas sobre até que ponto estão dispostos a abrir mão de sua privacidade em prol de benefícios coletivos.

aumentando sua precisão, seja melhorando sua capacidade de generalização — quase sempre exige o uso de algoritmos mais complexos, redes neurais maiores ou conjuntos de dados mais extensos, o que, por sua vez, demanda mais tempo de processamento, mais memória e mais energia computacional (Mulyo *et al.*, 2024).

Na modelagem molecular, enfrentamos um dilema muito parecido. Quanto mais detalhada e próxima da realidade for a nossa representação computacional de uma molécula, maior será o volume de dados gerados, o custo de armazenamento, o tempo de processamento e a complexidade matemática dos modelos necessários para lidar com essa representação. Por outro lado, representações mais simples e computacionalmente econômicas — como as que utilizam descritores 0D ou 2D — podem ser mais rápidas de calcular, mas deixam inevitavelmente de capturar aspectos importantes da interação molecular.

Em outras palavras, quando escolhemos um tipo de descritor ou uma estratégia de modelagem, estamos sempre navegando entre dois polos: de um lado, o desejo por maior realismo e qualidade descritiva; do outro, a necessidade de manter o custo computacional dentro de limites aceitáveis. Não existe uma solução única e universalmente melhor. O que há, na prática, é uma escolha consciente de qual aspecto sacrificar e qual privilegiar, conforme os objetivos específicos do estudo. Como veremos nas próximas seções, a escolha da dimensionalidade dos descritores é, em si, uma manifestação direta desse *trade-off*.

3.2.2 Fluxo de trabalho na modelagem computacional de pequenas moléculas

Para modelar os sistemas moleculares de interesse com efetividade, precisamos primeiro representá-los computacionalmente, o que pode ser feito de diversas formas.

A mais simples delas é utilizar um *software* de modelagem molecular, como o ChemSketch (ACD, 2024), e salvar a estrutura em um formato adequado, como o *.mol (Dalby *et al.* (1992) ou o mais moderno *.mol2 (Tripos, 1994).

Cada arquivo *.mol2 armazena, em um único documento estruturado, informações detalhadas sobre a geometria molecular tridimensional (*e.g.*, as coordenadas cartesianas (x, y, z) de cada átomo), a conectividade entre os átomos, os tipos de ligação, os tipos atômicos e, opcionalmente, as cargas parciais atribuídas. Não seria exagero compará-lo a uma “fotografia” da molécula que representa — uma imagem não apenas da disposição espacial dos átomos, mas também de suas propriedades eletrônicas.

Por exemplo, podemos utilizar o Open Babel (O’Boyle *et al.*, 2011) para converter a representação SMILES (Weininger, 1988) da Aspirina, CC(=O)OC1=CC=CC=C1C(=O)O, em um arquivo `aspirina.mol2`. Para isso, utilizamos o comando:

```
obabel -:"CC(=O)OC1=CC=CC=C1C(=O)O" -O aspirina.mol2 --gen2d
```

O conteúdo resumido do arquivo `aspirina.mol2` pode ser visto no [Bloco de Código 3.1](#). A seção `@<TRIPOS>MOLECULE`, fornece informações gerais sobre a molécula representada. O bloco `@<TRIPOS>ATOM`, por sua vez, lista os átomos da molécula, especificando suas coordenadas tridimensionais, o tipo atômico segundo a Tripos Force Field ([Tripos, 1994](#)) e, quando disponíveis, as cargas parciais. Já o bloco `@<TRIPOS>BOND` descreve as ligações químicas, incluindo os átomos envolvidos e o tipo de ligação.

Bloco de Código 3.1 – Partes do arquivo `aspirina.mol2`, gerado pelo Open Babel. O arquivo completo pode ser visto no [Bloco de Código D.1](#).

```

1 @<TRIPOS>MOLECULE
2 *****
3 13 13 0 0 0
4 SMALL
5 GASTEIGER
6
7 @<TRIPOS>ATOM
8      1 C      -2.5981  -1.5000  0.0000 C.3      1 UNL1      0.1216
9      2 C      -1.7321  -2.0000  0.0000 C.2      1 UNL1      0.3220
10     3 O      -1.7321  -3.0000  0.0000 O.2      1 UNL1     -0.2497
11     4 O      -0.8660  -1.5000  0.0000 O.3      1 UNL1     -0.4239
12     5 C      -0.8660  -0.5000  0.0000 C.ar     1 UNL1      0.1547
21 @<TRIPOS>BOND
22      1      1      2      1
23      2      2      3      2
24      3      2      4      1
25      4      4      5      1
26      5      5      6      ar

```

Fonte: elaborado pelo autor.

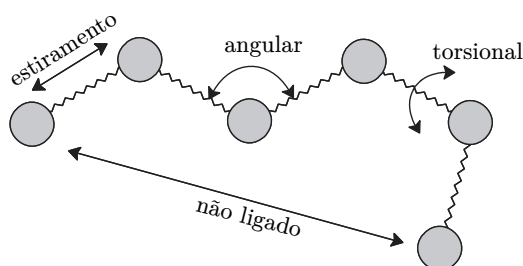
Note que, no [Bloco de Código 3.1](#), os átomos apenas dispõem de coordenadas x e y , mas que todas as coordenadas z são iguais a 0. Isso se deve ao fato de que geramos a estrutura 2D a partir da flag `--gen2d`, mas não há informação tridimensional sobre a molécula. Contudo, elas podem ser facilmente calculadas no processo de *otimização de geometria*. Essa otimização pode ser feita utilizando mecânica molecular ou mecânica quântica.

3.3 Modelando à luz da física clássica: Mecânica Molecular

A Mecânica Molecular (MM) representa uma das abordagens mais consagradas e acessíveis para a modelagem de estruturas moleculares em química computacional. Seu princípio fundamental é a representação simplificada dos átomos e de suas interações: os núcleos atômicos são tratados como esferas puntiformes dotadas de massa, enquanto as ligações químicas são modeladas como molas, obedecendo aos preceitos da física clássica. Os elétrons, por sua vez, não são tratados explicitamente; seus efeitos são incorporados de forma indireta, por meio de parametrizações empíricas.

O comportamento energético de uma molécula é descrito por uma função chamada campo de força (*force field*), que expressa a energia potencial total do sistema, E_{total} ,

Figura 3 – Ilustração de diferentes tipos de energia à luz da mecânica molecular



Fonte: Adaptado de Jensen (2017, p. 22).

em função das posições dos núcleos atômicos. Essa energia total é calculada como a soma de diferentes contribuições físicas, que incluem termos de estiramento de ligações, deformações angulares, torções, interações de van der Waals e interações eletrostáticas, conforme esquematizado na Equação 3.1 (Roy; Kar; Das, 2015, p. 165). Esses diferentes termos são ilustrados na Figura 3.

$$E_{\text{total}} = \sum E_{\text{estiramento}} + \sum E_{\text{angular}} + \sum E_{\text{torsão}} + \sum E_{\text{van der Waals}} + \sum E_{\text{elettrônico}} + \dots \quad (3.1)$$

Cada um desses termos apresenta uma forma matemática específica: interações de ligação são geralmente descritas por funções parabólicas (como no modelo harmônico de estiramento), enquanto as interações não ligantes costumam ser representadas por potenciais de Lennard-Jones para forças de dispersão e repulsão, além de um termo de Coulomb para as interações eletrostáticas.

O conjunto completo desses termos forma a chamada superfície de energia potencial da molécula. Navegar por essa superfície significa calcular a energia total do sistema em diferentes configurações geométricas, até encontrar o ponto de energia mais baixa — correspondente ao estado de equilíbrio estrutural da molécula.

O processo de encontrar essa configuração de mínima energia é denominado *otimização de geometria*. Na prática, trata-se de um procedimento iterativo, conduzido por algoritmos numéricos de minimização, que ajustam gradualmente as coordenadas dos átomos até que o sistema atinja uma configuração em que as forças resultantes sobre cada núcleo se aproximem de zero.

No contexto de estudos QSAR, a Mecânica Molecular desempenha um papel estratégico e recorrente. Suas principais aplicações incluem:

- Geração de coordenadas tridimensionais iniciais a partir de representações planas (2D), como aquelas derivadas de notações SMILES;

- Otimização preliminar da geometria molecular (por exemplo, como pré-tratamento para cálculos mais robustos);
- Cálculo de descritores conformacionais e energéticos, como ângulos diedros, volumes moleculares e energia de minimização.

Sua maior vantagem reside no baixo custo computacional: modelos baseados em MM permitem a análise de grandes bibliotecas químicas em tempo reduzido. Contudo, esse benefício vem acompanhado de limitações intrínsecas. Por tratar-se de uma abordagem baseada em física clássica e parametrizações empíricas, a MM não é capaz de descrever fenômenos eletrônicos detalhados, como redistribuições de carga, efeitos de ressonância ou quebras de ligação química. Para situações onde tais efeitos são relevantes, métodos de maior rigor teórico, como a Mecânica Quântica, tornam-se imprescindíveis.

3.4 Modelando à luz da Mecânica Quântica

Embora a MM seja computacionalmente eficiente, ela não passa de uma aproximação “fisicamente suficiente”: descreve núcleos e ligações com grande economia de recursos, mas ignora de forma explícita a distribuição eletrônica que, em muitos casos, domina o comportamento químico.

Sempre que buscamos uma caracterização mais fiel — por exemplo, capturar redistribuição de carga, quebra de ligação ou estados excitados — precisamos admitir, de antemão, que pagaremos um preço em custo de processamento e memória, conforme o *trade-off* custo-desempenho discutido na subseção 3.2.1. Assumido esse ônus, o passo natural é recorrer a métodos de Mecânica Quântica, que tratam os elétrons de forma explícita e fornecem uma representação muito mais rica e detalhada do sistema.

Enquanto a Mecânica Molecular encara os átomos como “esferas-mola” regidas por leis clássicas, a Mecânica Quântica parte de um pressuposto distinto: núcleos e elétrons são descritos por funções-de-onda que obedecem à equação de Schrödinger independente do tempo, (3.2), em que $\hat{H}$ é o Hamiltoniano do sistema, $|\Psi\rangle$ é a função-de-onda total (em função dos vetores de coordenadas eletrônicas e nucleares e, possivelmente, do tempo) e E é o autovalor de energia associado.

$$\hat{H}|\Psi\rangle = E|\Psi\rangle \quad (3.2)$$

Como a solução exata de (3.2) é inviável para moléculas reais, recorre-se a duas aproximações basilares:

1. **Born–Oppenheimer**: considera que a velocidade de translação dos núcleos atômicos é desprezível em relação à velocidade dos elétrons; a partir disso, a parte

eletrônica é resolvida considerando que os núcleos são estáticos e independentes do tempo;

2. **Base finita:** expande-se $|\Psi\rangle$ em um conjunto limitado de funções orbitais, convertendo o problema contínuo em matrizes tratáveis numericamente.

3.4.1 Métodos *ab initio*

Os métodos *ab initio*⁵ buscam resolver a equação de Schrödinger diretamente a partir dos princípios da física quântica, sem parametrizações empíricas extraídas experimentalmente. Diferentemente dos campos de força da **MM**, que ajustam seus parâmetros a dados experimentais pré-existentes, os métodos *ab initio* calculam interações eletrônicas e energias moleculares diretamente da estrutura eletrônica fundamental, oferecendo uma descrição significativamente mais precisa e detalhada.

Esses métodos, entretanto, demandam recursos computacionais elevados. Enquanto a **MM** simplifica o sistema molecular ao tratar núcleos e ligações como objetos clássicos com parâmetros fixos (tais como constantes de mola e cargas parciais pré-determinadas), os métodos *ab initio* realizam cálculos detalhados das interações elétron-elétron e elétron-núcleo em cada passo, capturando fenômenos complexos como polarização eletrônica e estados excitados.

Por isso, os métodos *ab initio* são preferíveis quando a precisão química é essencial, mas são frequentemente impraticáveis para sistemas grandes ou para simulações de longa duração, onde a eficiência computacional torna-se essencial. Já os campos de força são particularmente adequados para simular grandes biomoléculas ou estudar dinâmicas moleculares prolongadas, situações nas quais a precisão absoluta é menos crítica do que o custo computacional.

3.4.2 Métodos DFT

Os métodos fundamentados na Teoria do Funcional de Densidade (**DFT**) fornecem uma alternativa eficiente ao rigor dos métodos *ab initio*, comumente adotados quando é necessário equilibrar precisão e custo computacional. Em vez de abordar diretamente a função-de-onda eletrônica total, a **DFT** emprega a densidade eletrônica como variável principal, simplificando consideravelmente o problema e tornando-o computacionalmente viável mesmo para sistemas relativamente grandes.

Tal simplificação, contudo, depende criticamente da escolha adequada do funcional de troca-correlação — uma aproximação matemática que descreve a energia eletrônica em termos da densidade eletrônica. Funcionais comuns incluem B3LYP, PBE e M06-2X,

⁵ Literalmente “desde o princípio”, em latim.

cada qual com vantagens específicas dependendo do sistema estudado. Ainda assim, apesar dessa dependência empírica implícita, métodos **DFT** são amplamente utilizados para determinar propriedades estruturais e energéticas em química orgânica, bioquímica e materiais.

Em resumo, a **DFT** oferece um compromisso equilibrado entre eficiência computacional e fidelidade física, sendo frequentemente o método escolhido em situações nas quais uma precisão intermediária é suficiente para capturar os efeitos essenciais da distribuição eletrônica sem exigir o custo proibitivo dos métodos *ab initio*.

3.4.3 Métodos Semi-empíricos

Situando-se em um patamar intermediário de precisão e custo computacional, os métodos semi-empíricos constituem uma abordagem híbrida: partem da estrutura formal da Mecânica Quântica, porém simplificam substancialmente o Hamiltoniano, introduzindo parâmetros empíricos ajustados experimentalmente ou derivados de cálculos teóricos mais precisos.

Métodos clássicos como AM1, PM6 e MNDO exemplificam essa categoria. Esses métodos são notavelmente mais rápidos do que os *ab initio* e **DFT**, permitindo análises rápidas e qualitativas de grandes conjuntos de moléculas, especialmente em triagens iniciais de compostos bioativos. Entretanto, sua precisão é limitada justamente por dependerem de ajustes empíricos que podem não ser universais nem transferíveis para todos os tipos de sistemas químicos.

Esses métodos representam uma opção valiosa quando se deseja realizar análises preliminares ou simular grandes sistemas moleculares em escala razoável de tempo, sabendo-se que os resultados fornecerão mais tendências qualitativas do que rigor quantitativo detalhado.

Dessa forma, o pesquisador precisa sempre equilibrar cuidadosamente os objetivos específicos do estudo e as limitações práticas impostas pelos recursos disponíveis. A escolha do nível teórico deve refletir uma decisão consciente sobre quais aspectos do sistema são essenciais e quais concessões podem ser feitas sem comprometer significativamente a relevância científica dos resultados obtidos.

3.5 Dinâmica Molecular (MD)

A Dinâmica Molecular (**MD**) constitui uma das abordagens mais importantes e difundidas para investigar o comportamento dinâmico e conformacional de sistemas moleculares ao longo do tempo. Em essência, trata-se de simular numericamente a evolução temporal das posições e velocidades dos átomos de um sistema molecular

segundo as leis clássicas da mecânica, geralmente fundamentadas no formalismo da MM.

Nesta abordagem, a trajetória dos átomos é obtida integrando-se numericamente as equações de movimento de Newton. Isso permite capturar explicitamente a flexibilidade estrutural, as flutuações térmicas e as transições conformacionais, resultando em uma descrição detalhada das propriedades dinâmicas do sistema estudado. A precisão da simulação depende diretamente da qualidade do campo de força utilizado, que define as interações entre os átomos e, portanto, influencia os resultados da simulação.

3.5.1 Espaço conformacional

Uma das principais aplicações da MD é explorar o espaço conformacional acessível a moléculas flexíveis. Cada molécula bioativa possui múltiplas conformações possíveis devido ao livre movimento das ligações rotacionáveis. Cada arranjo espacial particular, denominado *confômero*, pode apresentar propriedades físico-químicas e biológicas distintas. Assim, a geração e a amostragem abrangente desses *confômeros* são essenciais para entender a atividade biológica de compostos químicos.

O “espaço conformacional” de uma molécula refere-se, de forma abstrata, ao conjunto de todas as conformações fisicamente possíveis que uma determinada molécula pode assumir. Embora teoricamente infinito, ele pode ser *amostrado* a partir dos quadros da simulação. Por exemplo, para uma dinâmica de 500 ps, podemos descartar os primeiros 50 ps e amostrar uma determinada quantidade de quadros entre 50 ps a 500 ps. Cada amostra (ou “pose”) corresponderá a um ponto diferente no espaço conformacional daquela molécula.

Durante uma simulação de MD, é possível obter um grande conjunto de conformações por meio da trajetória temporal resultante. A partir desta coleção de poses moleculares, é construído o chamado Perfil de Amostragem Conformacional (CEP), que representa uma média estatística das conformações acessíveis, ponderada pela distribuição de Boltzmann das energias associadas a cada pose.

A obtenção de um CEP envolve a superposição e o alinhamento sistemático de todas as poses moleculares amostradas ao longo da trajetória simulada. Este perfil pode ser interpretado como uma representação média da distribuição espacial (e temporal!) dos átomos, refletindo a probabilidade de ocupação de cada região do espaço durante a simulação de MD (Shi *et al.*, 2015).

Mais do que um mero retrato estrutural, o CEP incorpora, de forma implícita, a contribuição temporal de cada conformação, traduzindo não apenas a geometria tridimensional da molécula, mas também a frequência relativa com que diferentes regiões do espaço são exploradas ao longo do tempo. Por essa razão, o CEP constitui a base fundamental para o cálculo de descritores 4D, funcionando como elo conceitual entre a

natureza intrinsecamente dinâmica das moléculas e as representações estáticas requeridas pelos modelos QSAR.

Esse rigor na amostragem conformacional é especialmente importante em aplicações como o estudo das propriedades farmacocinéticas e farmacodinâmicas, onde pequenas variações estruturais podem resultar em efeitos biológicos significativamente distintos (Hawkins, 2017).

3.5.2 Aplicações e ferramentas comuns

A MD desempenha papel central no estudo da flexibilidade estrutural tanto de pequenos ligantes quanto de grandes biomoléculas como proteínas e ácidos nucleicos. Em QSAR, ela é especialmente valiosa para gerar descritores que dependem explicitamente da flexibilidade e da ocupação espacial dos compostos químicos, melhorando significativamente o poder preditivo dos modelos desenvolvidos.

Entre as ferramentas livres mais utilizadas na comunidade científica para realizar simulações de MD, destacam-se GROMACS e OpenMM (Eastman *et al.*, 2023), sendo que este último é distribuído como uma biblioteca para a linguagem Python. Essas plataformas oferecem funcionalidades robustas para a preparação dos sistemas moleculares, definição dos parâmetros de simulação, execução dos cálculos e análise detalhada dos resultados obtidos. Além disso, permitem empregar diversos campos de força reconhecidos, garantindo a reprodutibilidade e a confiabilidade dos estudos realizados.

3.6 Descritores Moleculares

No exemplo apresentado na abertura do Capítulo 3, a polaridade foi destacada como uma dentre inúmeras propriedades capazes de descrever, ainda que parcialmente, as interações entre um fármaco e seu receptor. O valor numérico associado a essa propriedade constitui um exemplo de *descriptor molecular*: um parâmetro quantitativo derivado da estrutura química, que captura informações latentes e potencialmente relevantes para a propriedade de interesse. Em essência, os descritores moleculares operam como um elo quantitativo entre a estrutura de uma molécula e sua atividade observada.

Segundo a definição clássica de Todeschini e Consonni (2000):

O descritor molecular é o resultado final de um procedimento lógico e matemático que transforma a informação química codificada em uma representação simbólica de uma molécula em um número útil ou no resultado de algum experimento padronizado (Todeschini; Consonni, 2000, tradução própria).

De forma operacional, os descritores podem ser vistos como vetores numéricos que representam características estruturais, físico-químicas, eletrônicas ou topológicas das moléculas. São, portanto, a única fonte de informação disponível para os modelos matemáticos durante o treinamento. Toda a “realidade molecular” acessível ao algoritmo encontra-se, inevitavelmente, condensada nesses vetores. Não surpreende, assim, que a escolha e a qualidade dos descritores figurem entre os fatores mais determinantes para o desempenho de um modelo [QSAR](#).

Cada descritor, por sua vez, representa uma abstração numérica de um aspecto específico da molécula: seja um parâmetro eletrônico, uma propriedade topológica ou um perfil tridimensional de interação. A literatura é praticamente unânime ao afirmar que o desempenho preditivo de modelos [QSAR](#) depende mais da relevância, diversidade e qualidade dos descritores empregados do que propriamente da complexidade algorítmica da técnica de modelagem utilizada ([Grisoni et al., 2018](#)).

Com um conjunto de descritores adequadamente selecionado e representativo, o modelo pode estabelecer correlações robustas entre estrutura química e atividade biológica, permitindo a previsão de bioatividades para novas moléculas estruturalmente semelhantes às do conjunto de treinamento. Essa limitação à similaridade estrutural define o chamado *domínio de aplicabilidade* do modelo, conceito crítico que deve sempre ser considerado na interpretação das previsões. A extrapolação de um modelo [QSAR](#) para compostos fora desse domínio compromete seriamente a validade de suas estimativas.

A construção desse espaço de descritores, contudo, está longe de ser trivial. Requer-se um método capaz de capturar, com fidelidade, os elementos estruturais responsáveis pela bioatividade observada. Diversas metodologias foram propostas ao longo das últimas décadas para gerar descritores moleculares, cada uma com seus pressupostos teóricos, vantagens e limitações.

Por fim, a seleção criteriosa de variáveis — visando eliminar descritores redundantes, irrelevantes ou colinearmente problemáticos — tornou-se uma etapa indispensável para maximizar a robustez estatística e a interpretabilidade dos modelos [QSAR](#) ([Danishuddin; Khan, 2016](#)).

No [Apêndice B](#), discutimos as principais categorias de descritores utilizadas na modelagem [QSAR](#), classificando-os conforme sua dimensionalidade e natureza da informação representada.

4 Modelagem Estatística em QSAR

No contexto da modelagem preditiva, como aquela conduzida em estudos de QSAR, os modelos de *regressão* ganham especial relevância. Tais modelos estatísticos têm por objetivo “aprender” uma função matemática que relacione um vetor de variáveis independentes $\mathbf{x}_i$ (como os descritores moleculares, discutidos na [seção 3.6](#)) com uma variável dependente y_i , correspondente a uma propriedade de interesse da i -ésima molécula (como a atividade biológica ou uma medida físico-química). Em última instância, busca-se estabelecer uma relação quantitativa (matemática) entre a estrutura molecular (descrita por $\mathbf{x}_i$) e a atividade y_i apresentada por essa mesma molécula — eis o “coração” da Relação Quantitativa Estrutura–Atividade.

Esta seção se dedica a formalizar matematicamente o problema da QSAR e apresentar brevemente alguns métodos de regressão linear frequentemente utilizados para resolvê-lo. O [Apêndice A](#) aborda os conceitos fundamentais em Estatística para a análise e validação de um modelo QSAR, sem os quais seria impossível tratar do assunto sobre o qual nos debruçamos agora.

4.1 Formalização Matemática da Relação Estrutura–Atividade

Em primeiro lugar, formularemos matematicamente o problema em questão, para então passar aos métodos estatísticos e computacionais para abordá-lo.

Formalmente, para um sistema de M moléculas no qual cada uma é codificada por N descritores, denota-se por $\mathbf{X} \in \mathbb{R}^{M \times N}$ a matriz de descritores. Cada linha $\mathbf{x}_i \in \mathbb{R}^N$ de $\mathbf{X}$ representa o vetor de descritores correspondente à i -ésima molécula ($i = 1, \dots, M$). Analogamente, o vetor $\mathbf{y} \in \mathbb{R}^M$ contém os valores experimentais da propriedade de interesse para as M moléculas, de modo que, se nos interessarmos pela atividade biológica, y_i será o valor da atividade observada experimentalmente para a i -ésima molécula.

Busca-se, portanto, construir uma função de regressão $f: \mathbb{R}^N \rightarrow \mathbb{R}$ que, dada uma entrada $\mathbf{x}_i \in \mathbb{R}^N$ — isto é, o vetor de descritores associado à molécula i —, forneça uma predição $f(\mathbf{x}_i) \equiv \hat{y}_i$ que aproxime, com o menor erro possível, o valor experimental da propriedade de interesse, ou seja, tal que $\hat{\mathbf{y}} \approx \mathbf{y}$.

Essa formulação pode ser alcançada facilmente ao escrever essas hipóteses em notação matricial. Nesses termos, podemos supor que exista uma relação linear entre a matriz de descritores $\mathbf{X} \in \mathbb{R}^{M \times N}$ e o vetor de predições $\hat{\mathbf{y}} \in \mathbb{R}^M$, mediada por um vetor de coeficientes $\boldsymbol{\beta} \in \mathbb{R}^N$.¹ Essa suposição nos permite escrever o modelo na forma (4.1):

¹ A hipótese de relação não-linear é igualmente possível e importante, sobretudo quando se utilizam

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} \approx \mathbf{y} \quad (4.1)$$

Com isso, o problema em tela se reduz à resolução de um sistema linear em que a incógnita a ser determinada é o vetor de parâmetros $\boldsymbol{\beta}$. Esse sistema também pode ser representado por (4.2):

$$\hat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta} = 0 \quad (4.2)$$

4.2 Regressão Linear

Seguindo a definição do modelo (Equação 4.2), sabemos que os erros experimentais existentes nas entradas tornam o sistema linear inconsistente, o que implica na inexistência de soluções exatas. Já que nenhuma solução exata é possível, buscamos um vetor $\hat{\boldsymbol{\beta}}$ que seja o mais próximo possível de uma solução, isto é, que minimize $\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|$. Enunciado de outro modo, queremos $\hat{\boldsymbol{\beta}}$ tal que

$$\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\| \approx 0. \quad (4.3)$$

Podemos reescrever (4.1) na forma de um sistema normal (4.4) (Anton; Rorres, 2013, p. 381).

$$\begin{aligned} \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} &= \mathbf{X}^T \mathbf{y} \\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \end{aligned} \quad (4.4)$$

Em (4.4), o termo $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ é conhecido como a pseudoinversa de Moore-Penrose e é denotada por $\mathbf{X}^+$.

$$\mathbf{X}^+ \equiv (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (4.5)$$

Com (4.5), podemos reescrever (4.4) de forma mais enxuta:

$$\hat{\boldsymbol{\beta}} = \mathbf{X}^+ \mathbf{y} \quad (4.6)$$

Por fim, a predição final $\hat{\mathbf{y}}$ é calculada por (4.8), em que (4.7) é a *matriz de projeção* no espaço $\mathcal{V}$ gerado pelas colunas de $\mathbf{X}$.

$$\mathbf{P} \equiv \mathbf{X}\mathbf{X}^+ \quad (4.7)$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{y} \quad (4.8)$$

No entanto, essa formulação exige cuidado quando os descritores estão altamente correlacionados entre si (multicolinearidade) ou quando o número N de descritores utilizados no modelo se aproxima do número M de moléculas, pois nesses casos a matriz $\mathbf{X}^T\mathbf{X}$ pode ser mal condicionada ou mesmo singular.

4.2.1 Quadrados Mínimos Parciais (PLS)

Quando o número de descritores N se aproxima ou mesmo supera o número de observações M , ou ainda quando há forte colinearidade entre os descritores, o modelo de Quadrados mínimos lineares (LLS) torna-se instável ou inaplicável. Nesse cenário, uma alternativa amplamente empregada em estudos QSAR é o método dos Quadrados mínimos parciais (PLS).

O PLS busca construir uma sequência de componentes latentes (ou variáveis projetadas), denominadas *scores*, que são combinações lineares dos descritores originais, com a particularidade de que cada componente maximiza sua covariância com a variável de resposta $\mathbf{y}$. Em outras palavras, o método projeta os dados em uma subestrutura de menor dimensionalidade que é otimizada não apenas com base na estrutura de $\mathbf{X}$, mas também levando em conta a relação com $\mathbf{y}$.

A matriz de descritores é, assim, decomposta como

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T, \quad (4.9)$$

enquanto a variável de resposta é aproximada como

$$\hat{\mathbf{y}} = \mathbf{T}\mathbf{q}, \quad (4.10)$$

onde $\mathbf{T}$ é a matriz de *scores* (componentes latentes), $\mathbf{P}$ contém os *loadings* de $\mathbf{X}$, e $\mathbf{q}$ representa os pesos associados a cada componente para modelar $\mathbf{y}$.

A escolha do número ótimo de componentes K é geralmente realizada por validação cruzada, analisando-se um gráfico de Qualidade de predição ou Soma dos Quadrados dos Resíduos de Previsão (PRESS) em função dos diferentes valores de K . Ao utilizar apenas os primeiros $K \ll N$ componentes, o modelo resultante é mais robusto, reduzindo a variância dos coeficientes estimados e mitigando problemas de sobre-ajuste. Dessa forma, o PLS combina a redução de dimensionalidade com a preservação da relevância preditiva, sendo especialmente eficaz em contextos de alta dimensionalidade e colinearidade, como frequentemente ocorre em conjuntos de dados moleculares.

Parte II

Desenvolvimento

5 Método proposto

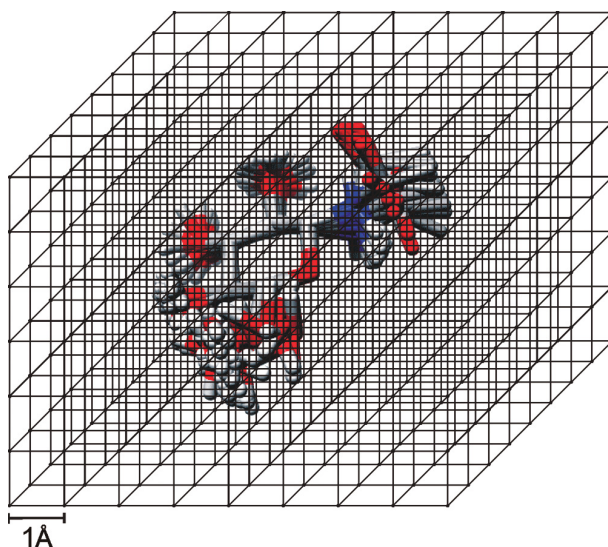
A [seção 3.6](#) apresentou as principais técnicas de descrição molecular consagradas na literatura de [QSAR](#). Dentre elas, destaca-se a abordagem LQTA-QSAR, que combina os princípios da Análise Comparativa de Campo Molecular ([CoMFA](#)) com o formalismo 4D, incorporando descritores energéticos derivados do espaço conformacional da molécula.

Neste capítulo, revisitamos o método LQTA-QSAR, detalhando sua formulação, para então aprofundar a análise crítica de suas limitações conceituais e metodológicas. A partir desse diagnóstico, introduzimos o HullQSAR, uma nova abordagem desenvolvida neste trabalho com o intuito de mitigar tais deficiências.

5.1 LQTA-QSAR

Proposto por [Martins, Barbosa *et al.* \(2009\)](#), o método LQTA-QSAR visa superar as limitações de abordagens 4D-QSAR anteriores, que não contemplavam diretamente as interações eletrostáticas entre a molécula e o ambiente. Para isso, o CEP da molécula é inserido em uma malha tridimensional, e em cada ponto dessa malha, uma sonda (como o íon NH_3^+) calcula as interações de Coulomb ([Equação 5.1](#)) e de Lennard-Jones ([Equação 5.2](#)) com todos os átomos do CEP. A [Figura 4](#) ilustra esse processo.

Figura 4 – Exemplo de um perfil de amostragem conformacional numa grade do LQTA-QSAR de 1 Å.



Fonte: [Martins, Barbosa *et al.* \(2009\)](#).

Sejam r_{ij} a distância entre o ponto z_j da malha e o átomo i do CEP; ε , a permissividade elétrica do meio; n , o número de conformações no CEP; e $\{K_{ii}^{12}, K_{ii}^6, K_{jj}^{12}, K_{jj}^6\}$ os parâmetros do campo de força. As interações são dadas por:

$$C_j = \frac{1}{n} \sum_i \frac{q_i q_j}{4\pi\epsilon r_{ij}} \quad (5.1)$$

$$LJ_j = \sum_i \left(\frac{K_{ij}^{12}}{r_{ij}^{12}} - \frac{K_{ij}^6}{r_{ij}^6} \right) \quad (5.2)$$

$$K_{ij}^{12} = \sqrt{\frac{K_{ii}^{12} \cdot K_{jj}^{12}}{n}} \quad (5.3)$$

$$K_{ij}^6 = \sqrt{\frac{K_{ii}^6 \cdot K_{jj}^6}{n}} \quad (5.4)$$

Para cada CEP, gera-se um vetor de descritores que representa as energias calculadas nos pontos da malha, formando uma matriz de descritores que alimenta o modelo QSAR.

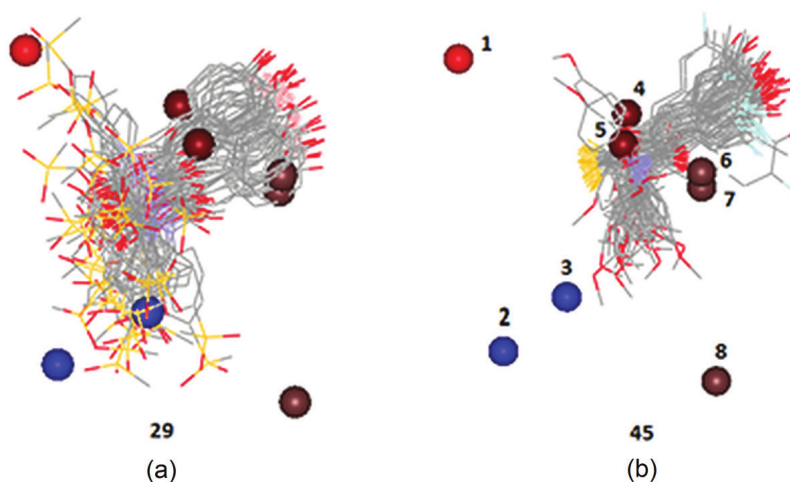
5.2 Limitações do LQTA-QSAR

Uma das limitações mais evidentes do LQTA-QSAR — e de outros métodos baseados em amostragem espacial fixa, como o CoMFA — decorre do fato de que a sonda pode ser posicionada a distâncias fisicamente irreais em relação aos átomos do CEP, chegando, por vezes, a ser colocada virtualmente dentro da esfera de volume ocupado por um átomo. Nessa condição-limite (quando $r_{ij} \rightarrow 0$), os termos de Coulomb e Lennard-Jones que compõem os descritores de interação tendem ao infinito, produzindo valores energeticamente absurdos e destituídos de significado químico. A Figura 5 ilustra perfeitamente este problema ao registrá-lo, quase em flagrante, no conjunto CANCER (ver subseção 6.1.1), objeto de estudo central deste trabalho.

O impacto dessa falha estrutural do método vai além de um mero artefato numérico: a geração de descritores com magnitude tão desproporcional compromete diretamente a robustez estatística dos modelos preditivos, degrada a qualidade da inferência e fragiliza a interpretabilidade físico-química dos resultados.

Para termos a dimensão do problema, retomemos à malha cartesiana da Figura 4. Suponha que, após a discretização espacial empregada pelo LQTA-QSAR, um dos nós amostrais coincida com um ponto localizado a apenas 0,1 Å de um átomo de oxigênio cuja carga parcial seja $q_a = -0,5e$. Considerando uma sonda NH_3^+ , com os parâmetros sintetizados na Tabela 3, obtemos, para essa posição específica, os valores de energia apresentados na Equação 5.5.

Figura 5 – Representação espacial dos melhores descritores LQTA-QSAR selecionados no conjunto **CANCER** para os compostos (a) menos e (b) mais ativos. Observe que a maioria dos descritores, representados pelas esferas coloridas, se localizam no interior da estrutura molecular.



Fonte: Adaptado de [Martins e Melo \(2019\)](#).

$$\begin{aligned}
 E^{\text{Cl}} &= \frac{1}{N_A} k_e \frac{q_a q_s}{r_{ij}} \\
 &\approx -6,95 \times 10^3 \text{ kJ mol}^{-1} ; \\
 E^{\text{LJ}} &= \frac{1}{N_A} 4\epsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^6 \right] \\
 &\approx 1,87 \times 10^{18} \text{ kJ mol}^{-1} !
 \end{aligned} \tag{5.5}$$

Para contextualizar o grau de desvio físico envolvido, vale lembrar que as energias típicas de ligações covalentes oscilam entre 150 kJ mol^{-1} a 500 kJ mol^{-1} , podendo atingir até cerca de 1000 kJ mol^{-1} no caso de ligações particularmente fortes, como a tripla ligação $\text{N}\equiv\text{N}$ (N_2). No entanto, a energia de Lennard-Jones calculada para este exemplo específico ($1,87 \times 10^{18} \text{ kJ mol}^{-1}$) excede essas referências por mais de 15 ordens de grandeza, o que a transforma em um *outlier* de proporção astronômica, sem qualquer correspondência no domínio das interações moleculares reais.

Além do problema conceitual evidente, tal discrepância gera um grave efeito colateral estatístico: a presença de um único descritor com esta magnitude seria suficiente para introduzir um grau extremo de **heterocedasticidade** nos dados, violando premissas básicas de métodos estatísticos multivariados como o **PLS** e comprometendo a qualidade da modelagem.

Diante dessa limitação estrutural, [Tenório \(2018\)](#) propuseram o **HullQSAR**, que introduz o conceito de “**fecho convexo**” como um critério geométrico adicional para a seleção dos pontos de amostragem. A [seção 5.3](#) a seguir apresenta esse conceito

Tabela 3 – Parâmetros hipotéticos que levariam o LQTA-QSAR a calcular energias fisicamente inverossímeis.

Parâmetro	Valor (hipotético)
Constante de Coulomb	$k_e = 8,9875 \times 10^9 \text{ N m}^2 \text{ C}^{-2}$
Distância sonda-átomo	$r_{ij} = 0,1 \text{ \AA} = 1,0 \times 10^{-11} \text{ m}$
Carga elementar	$e = 1,602\,176\,634 \times 10^{-19} \text{ C}$
Carga do átomo (O)	$q_a = -0,5 e = -8,01 \times 10^{-20} \text{ C}$
Carga da sonda (NH_3^+)	$q_s = 1 e = 1,602 \times 10^{-19} \text{ C}$
Profundidade do poço de Lennard-Jones (O)	$\varepsilon = 0,21 \text{ kcal mol}^{-1}$
Diâmetro efetivo de Lennard-Jones (O)	$\sigma = 3,0 \text{ \AA}$

Fonte: elaborado pelo autor.

matemático, estabelecendo o alicerce necessário para a subsequente descrição detalhada do HullQSAR na [seção 5.4](#).

5.3 Fecho convexo

Em geometria computacional, o fecho convexo $\mathcal{H}(X)$ de um conjunto de pontos $X \subset \mathbb{R}^d$ é definido como o menor conjunto convexo que o contém. Formalmente:

$$\mathcal{H}(X) = \left\{ \sum_{i=1}^k \lambda_i x_i \mid x_i \in X, \lambda_i \geq 0, \sum_{i=1}^k \lambda_i = 1 \right\} \quad (5.6)$$

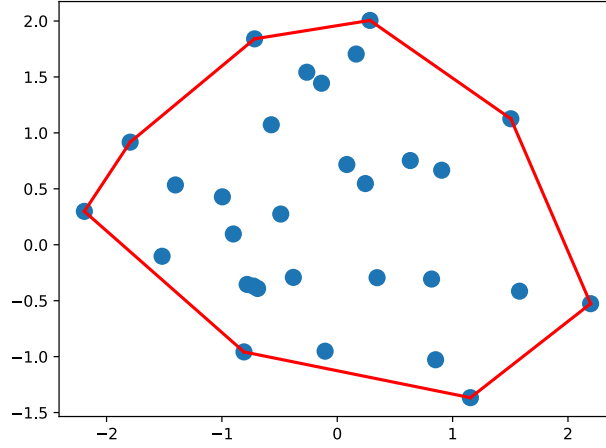
O cálculo computacional do fecho convexo pode ser realizado por diferentes algoritmos, sendo um dos mais eficientes em dimensões moderadas o algoritmo de [Chan \(1996\)](#), que computa o fecho convexo em tempo $\mathcal{O}(|X| \log h)$, onde $|X|$ é o número de pontos no conjunto X e h é o número de vértices do fecho convexo resultante. A [Figura 6](#) ilustra esse conceito em duas dimensões, com os pontos de um conjunto $X \subset \mathbb{R}^2$ (em azul) e seu fecho convexo $\mathcal{H}(X)$ (delineado pela linha vermelha).

5.4 HullQSAR

No contexto da 4D-QSAR, cada átomo i em uma dada conformação t do CEP é representado como um ponto tridimensional $x_i^{(t)} \in X$, com $X \subset \mathbb{R}^3$. Define-se o fecho convexo $\mathcal{H}(X)$ como o menor volume convexo que contém todos esses pontos. Essa construção geométrica atua como uma barreira natural, impedindo que a amostragem de descritores inclua pontos localizados dentro da distribuição atômica, que não representam áreas de interação realista com um possível receptor ([Tenório; Cerf; Martins, 2017](#)).

Para realizar uma amostragem adequada, expandimos a superfície externa do fecho convexo radialmente por uma distância inicial r_0 , gerando uma nova superfície $\mathcal{H}'(X)$:

Figura 6 – Exemplo de fecho convexo $\mathcal{H}(X)$ (linha vermelha) de um conjunto arbitrário $X \subset \mathbb{R}^2$ (pontos azuis)



Fonte: Tenório (2018).

$$\mathcal{H}'(X) = \{x + r_0 \cdot \hat{n}_x \mid x \in \partial\mathcal{H}(X), \hat{n}_x \text{ é a normal externa em } x\} \quad (5.7)$$

onde $\hat{n}_x$ denota o vetor normal unitário externo à superfície $\partial\mathcal{H}(X)$ no ponto x . Computacionalmente, nos vértices e arestas do poliedro convexo, onde a normal não é única, utiliza-se a média ponderada das normais das faces adjacentes.¹

A partir dessa superfície inicial, são definidas camadas adicionais em intervalos radiais constantes Δr , formando uma sequência de superfícies sucessivas:

$$\mathcal{H}^{(k)}(X) = \{x + (r_0 + k\Delta r) \cdot \hat{n}_x \mid x \in \partial\mathcal{H}(X)\}, \quad k = 0, 1, \dots, \ell - 1 \quad (5.8)$$

O número total de camadas ℓ é determinado pela distância máxima $r_{\max}$ definida pelo usuário:

$$\ell = 1 + \left\lfloor \frac{r_{\max} - r_0}{\Delta r} \right\rfloor \quad (5.9)$$

Em cada camada, são posicionados pontos de amostragem definidos por coordenadas esféricas (θ, ϕ) , espaçadas por um incremento angular constante $\Delta\alpha$. O número total de pontos Z gerados pelo método é calculado como:

$$Z = \ell \left[\left\lfloor \frac{360}{\Delta\alpha} \right\rfloor \cdot \left\lfloor \frac{180}{\Delta\alpha} \right\rfloor - 2 \left(\left\lfloor \frac{180}{\Delta\alpha} \right\rfloor - 1 \right) \right] \quad (5.10)$$

Em cada um desses pontos z_j , posiciona-se uma sonda (geralmente o íon NH_3^+), com carga q_j , para calcular as interações com cada átomo i do CEP. Utilizando-se das

¹ Para detalhes sobre o cálculo de normais em poliedros convexos, ver Tenório (2018).

cargas parciais atômicas q_i , previamente determinadas, calculam-se os descritores de Coulomb e Lennard-Jones:

$$E_j^{\text{Cl}} = \sum_{i=1}^N \frac{q_i q_j}{4\pi\epsilon_0 |z_j - x_i|} \quad (5.11)$$

$$E_j^{\text{LJ}} = \sum_{i=1}^N \left(\frac{K_{ij}^{12}}{|z_j - x_i|^{12}} - \frac{K_{ij}^6}{|z_j - x_i|^6} \right) \quad (5.12)$$

Cada ponto gera, portanto, dois descritores, resultando em um total de $2Z$ descritores por CEP. Essa abordagem assegura que os descritores representem condições mais realistas de interação molecular, eliminando contribuições não-físicas provenientes de regiões internas da densidade atômica (Tenório, 2018).

6 Metodologia Experimental

A fundamentação teórica apresentada nos capítulos anteriores evidenciou que a escolha da estratégia de amostragem espacial e da região do espaço químico onde os descritores são calculados influencia diretamente a qualidade da representação molecular. Como discutido na [seção 5.2](#), metodologias como o LQTA-QSAR permitem que a sonda de cálculo seja posicionada arbitrariamente próxima — ou mesmo no interior — da distribuição eletrônica da molécula, gerando valores de energia fisicamente irreais e comprometendo a robustez estatística dos modelos.

Conforme argumentado no [Capítulo 3](#), todos os modelos são aproximações da realidade, e sua utilidade depende fundamentalmente de quão bem os fenômenos que se deseja capturar resistem a essas simplificações. No contexto de modelagem QSAR, isso significa que a forma como representamos o espaço de interação molecular deve estar alinhada com a física subjacente ao reconhecimento ligante-receptor.

Nesse sentido, a hipótese central deste trabalho é que a utilização de uma restrição física mais próxima da realidade — delimitando a região amostrada ao exterior do volume molecular via fecho convexo — permite uma representação que captura aspectos mais relevantes da molécula e, conseqüentemente, fornece modelos estatisticamente superiores. Este capítulo descreve, portanto, a metodologia experimental desenhada para validar a hipótese central deste trabalho.

6.1 Dados experimentais

Para avaliar o impacto prático do método HullQSAR ([seção 5.4](#)), foram utilizados dois conjuntos de dados (CANCER e CARBOX) cujas modelagens via LQTA-QSAR já são descritas na literatura.

6.1.1 Conjunto de dados: CANCER

O conjunto CANCER consiste em 49 derivados benzilidenos de oxazolidinediona e tiazolidinediona, potenciais inibidores da enzima 17 β -hidroxiesteróide desidrogenase tipo 3 (17 β -HSD3), com aplicabilidade no tratamento do câncer de próstata. Os dados experimentais foram obtidos da literatura ([Harada *et al.*, 2012](#); [Martins; Melo, 2019](#)), onde as atividades biológicas foram reportadas como valores de Concentração Inibitória Média, posteriormente convertidos para a escala logarítmica negativa (pIC₅₀). A estrutura base dos compostos está representada na [Figura 7](#).

Tabela 4 – Hiper parâmetros utilizados no HullQSAR. Todos os valores numéricos estão em Å, exceto ℓ .

Hiperparâmetro	Valor utilizado
Sonda	NH_3^+
r_0	2,0 Å
Δr	0,5 Å
$\Delta\alpha$	4°, 6°, 8° e 10°
ℓ	10

Fonte: elaborado pelo autor.

o método DFT (B3LYP/6-311G(d,p)) com o software Gaussian 09, e as cargas atômicas parciais foram calculadas pelo método CHELPG.

Essas estruturas otimizadas foram então submetidas ao software UCSF Chimera para atribuição de cargas AM1-BCC e tipos atômicos compatíveis com o campo de força AMBER ff03, necessários para as simulações de dinâmica molecular.

6.3 Simulações de Dinâmica Molecular

As simulações de dinâmica molecular foram realizadas utilizando o pacote GRO-MACS (Abraham *et al.*, 2015), seguindo o protocolo descrito por Martins e Melo (2019) e Martins, Oliveira e Queiroz (2018). Cada molécula foi simulada individualmente em meio aquoso explícito (modelo SPC/E), com condições periódicas de contorno e neutralização por contraíons quando necessário.

O protocolo de simulação captou as etapas sequenciais da minimização energética por 500 ps, com registro das trajetórias a cada 1 ps. As conformações obtidas entre 50 ps e 500 ps foram sobrepostas e alinhadas para compor o CEP de cada molécula.

6.4 Cálculo dos descritores com o HullQSAR

Os descritores foram calculados utilizando a metodologia HullQSAR (seção 5.4).

A sonda utilizada foi o íon NH_3^+ , e os hiperparâmetros empregados estão apresentados na Tabela 4. Foram testados quatro valores diferentes para o parâmetro angular $\Delta\alpha$ (4°, 6°, 8° e 10°), resultando em quatro matrizes de descritores para cada conjunto de dados.

6.5 Pré-processamento e Seleção de Variáveis

As matrizes geradas pelo HullQSAR possuem alta dimensionalidade (da ordem de 10^5 descritores). Para reduzir essa dimensionalidade, foram aplicadas três filtrações iniciais:

1. **Variância:** todas as colunas com variância $\text{var}(\mathbf{x}_i) < 0,1$, por serem consideradas quase constantes, foram removidas;
2. **Correlação:** em seguida, foram removidas todas as colunas $\mathbf{x}_i$ cuja correlação com o vetor $\mathbf{y}$ fosse tal que $\text{cor}(\mathbf{x}_i, \mathbf{y}) < 0,3$;
3. **Autocorrelação:** por fim, qualquer par de colunas $\mathbf{x}_i$ e $\mathbf{x}_j$ com intercorrelação tal que $\text{cor}(\mathbf{x}_i, \mathbf{x}_j) > 0,9$ foi considerado redundante e simplificado, eliminando-se a segunda coluna ($\mathbf{x}_j$, neste caso).

Por fim, com o número de colunas drasticamente reduzido e um conjunto de descritores com potencial relevância pré-selecionados, o Algoritmo Genético (Fortin *et al.*, 2012; Leardi; Boggia; Terrile, 1992) foi utilizado para a seleção das variáveis mais relevantes, com base no critério de validação cruzada (Q^2).

6.6 Construção e Validação dos Modelos QSAR

A construção e validação dos modelos QSAR foram realizadas com o software QSARMODELINGPY (Martins; Reis Filho, 2020). Para cada conjunto de dados, as moléculas foram divididas em conjuntos de treino e teste na proporção de 80% e 20%, respectivamente, utilizando o algoritmo de Kennard-Stone (Kennard; Stone, 1969). Essa abordagem garante que o conjunto de teste seja representativo do espaço químico do conjunto de treino.

Os modelos foram construídos utilizando o método de regressão por Mínimos Quadrados Parciais (PLS). A robustez e a capacidade preditiva dos modelos foram avaliadas por meio de validação interna e externa. Essas duas etapas de validação estão disponíveis nativamente no QSARMODELINGPY, e portanto não foram necessárias implementações adicionais para este trabalho.

7 Resultados e Discussão

Neste capítulo, são apresentados e discutidos os resultados obtidos com a metodologia proposta, HullQSAR. O desempenho dos modelos gerados é rigorosamente avaliado e comparado com a abordagem de referência, LQTA-QSAR, utilizando os conjuntos de dados **CANCER** e **CARBOX**, conforme descrito na metodologia.

7.1 Análise Comparativa dos Modelos QSAR

Para avaliar a eficácia do método HullQSAR, foram construídos modelos para os dois conjuntos de dados, variando-se o hiperparâmetro de resolução angular, $\Delta\alpha$. Os resultados detalhados para cada configuração podem ser consultados no [Apêndice C](#).

A [Tabela 5](#) apresenta uma comparação direta entre o melhor modelo obtido com o HullQSAR para cada conjunto de dados e os respectivos modelos de referência da literatura, construídos com o método LQTA-QSAR.

A inspeção da tabela revela uma superioridade clara e sistemática da abordagem HullQSAR. Para ambos os conjuntos, o método proposto não apenas atingiu, mas superou o desempenho do método de referência em todas as métricas estatísticas fundamentais (R^2 , Q^2 e R_{Ext}^2).

7.1.1 Desempenho no Conjunto CANCER

Para o conjunto **CANCER**, o modelo de referência LQTA-QSAR já apresentava um desempenho robusto. Contudo, o melhor modelo HullQSAR ($\Delta\alpha = 4^\circ$) demonstrou um ganho de performance notável. O coeficiente de determinação (R^2) aumentou, enquanto a capacidade preditiva, avaliada por Q^2 e R_{Ext}^2 , apresentou melhorias ainda mais expressivas, com um incremento de 12 e 9 pontos percentuais, respectivamente.

Este resultado sugere que os descritores gerados pelo HullQSAR, que integram a informação de interação eletrostática com a de volume molecular (via fecho convexo),

Tabela 5 – Comparação entre o melhor modelo HullQSAR e o LQTA-QSAR nos conjuntos **CANCER** e **CARBOX**.

	CANCER		CARBOX	
	HullQSAR	LQTA-QSAR	HullQSAR	LQTA-QSAR
R^2	0,78	0,77	0,95	0,77
Q^2	0,73	0,61	0,88	0,70
R_{Ext}^2	0,83	0,74	0,96	0,63

Fonte: elaborado pelo autor.

capturam nuances estereoelétrônicas que não são adequadamente representadas pelo método LQTA-QSAR tradicional. O ganho de poder preditivo foi alcançado com apenas duas variáveis latentes, indicando que a melhoria não se deve a um sobreajuste, mas sim à maior qualidade informacional dos descritores.

7.1.2 Desempenho no Conjunto CARBOX

No conjunto CARBOX, a vantagem do HullQSAR foi ainda mais pronunciada. O modelo ótimo ($\Delta\alpha = 8^\circ$) alcançou um desempenho excepcional, com um R^2 de 0,95 e um R^2_{Ext} de 0,96. Em comparação, o modelo LQTA-QSAR obteve valores de 0,77 e 0,63 para as mesmas métricas. Trata-se de um salto de 33 pontos percentuais na capacidade de previsão do modelo em um conjunto de teste externo.

A expressiva redução nos erros de calibração e validação (PRESS e RMSE, ver [Apêndice C](#)) reforça a hipótese central deste trabalho: a combinação de descritores de ocupação e de interação, filtrados pela superfície do fecho convexo, gera uma representação molecular mais rica e informativa, resultando em modelos QSAR significativamente mais robustos e preditivos.

7.2 Validação e Análise dos Critérios Estatísticos

Todos os modelos HullQSAR gerados atenderam com folga aos rigorosos critérios de validação recomendados na literatura por [Roy, Kar e Das \(2015\)](#). Os valores de Q^2 (acima de 0,5) e R^2_{Ext} (acima de 0,6) confirmam a robustez e a capacidade de generalização dos modelos.

Parte III

Considerações Finais

8 Conclusão

À luz da fundamentação teórica e da discussão epistemológica desenvolvidas ao longo deste trabalho, pode-se afirmar que o objetivo central foi atingido: demonstrou-se que a incorporação de restrições fisicamente mais realistas na geração de descritores 4D-QSAR, por meio do conceito geométrico de fecho convexo, resulta em modelos estatisticamente mais robustos e com maior capacidade preditiva.

Para isso, este estudo analisou comparativamente duas abordagens de descrição molecular, o LQTA-QSAR e o HullQSAR, aplicadas a dois conjuntos de dados reais com relevância farmacológica (CANCER e CARBOX), previamente modelados na literatura. Em ambos os casos, partiu-se de um protocolo consistente de modelagem molecular e dinâmica, construção do CEP, cálculo de descritores e regressão por PLS, seguido de rigorosa validação interna e externa. Os resultados obtidos indicaram desempenho superior do HullQSAR em relação ao LQTA-QSAR em diferentes métricas estatísticas (R^2 , Q^2 , RMSE, PRESS), tanto nos conjuntos de treinamento quanto nos conjuntos de teste, sugerindo que a delimitação da região de amostragem ao exterior do volume molecular, via fecho convexo, constitui uma estratégia efetiva de filtragem de contribuições não físicas.

Do ponto de vista conceitual, esses resultados reforçam a hipótese de que uma representação molecular mais alinhada com o cenário de reconhecimento ligante-receptor — isto é, focada nas regiões de espaço onde a interação é plausível — produz descritores mais informativos e, por conseguinte, modelos QSAR mais úteis. Como nos lembra Box (1979), não se trata de buscar um modelo “correto”, mas de aproximar-se de modelos que errem de modo controlado e, sobretudo, relevante para o fenômeno de interesse. Ao restringir a amostragem à vizinhança externa do fecho convexo, o HullQSAR torna explícito esse compromisso: aceita-se a inevitável incompletude da representação, mas rejeitam-se, tanto quanto possível, contribuições que contradigam a física elementar do sistema.

Em síntese, este trabalho buscou articular uma visão na qual se reconhece a indivisibilidade entre as várias áreas do conhecimento. A esperança é que, ao reconhecer explicitamente os limites de nossos modelos e ao refinar suas aproximações, possamos construir ferramentas mais transparentes e eficazes para o planejamento racional de fármacos, ampliando o repertório de descritores disponíveis e contribuindo, ainda que modestamente, para o diálogo contínuo, ativo e altivo entre nós que pensamos o mundo.

Referências

ABRAHAM, Mark James *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. **SoftwareX**, Elsevier BV, v. 1-2, p. 19–25, set. 2015. ISSN 2352-7110. DOI: [10.1016/j.softx.2015.06.001](https://doi.org/10.1016/j.softx.2015.06.001).

ADVANCED CHEMISTRY DEVELOPMENT, INC. **ACD/ChemSketch**. [S. l.: s. n.], 2024. Software de desenho químico. Disponível gratuitamente para uso acadêmico. Disponível em: <https://www.acdlabs.com/resources/free-chemistry-software-apps/chemsketch-freeware/>. Acesso em: 11 maio 2025.

ANTON, Howard; RORRES, Chris. **Elementary Linear Algebra**. 11. ed. Hoboken: Wiley, 2013. ISBN 9781118434413.

APODACA, Richard. Balsa: A Compact Line Notation Based on SMILES. **Theoretical and Computational Chemistry**, American Chemical Society (ACS), 2022. Working paper. DOI: [10.26434/chemrxiv-2022-011tp](https://doi.org/10.26434/chemrxiv-2022-011tp).

APODACA, Richard L. **Beyond SMILES**. 22 set. 2021. Disponível em: <https://depth-first.com/articles/2021/09/22/beyond-smiles>. Acesso em: 30 maio 2025.

BOX, George E. P. Robustness in the Strategy of Scientific Model Building. *In: ROBUSTNESS in Statistics*. [S. l.]: Elsevier, 1979. p. 201–236. DOI: [10.1016/b978-0-12-438150-6.50018-2](https://doi.org/10.1016/b978-0-12-438150-6.50018-2).

BOX, George E. P. Science and Statistics. **Journal of the American Statistical Association**, Informa UK Limited, v. 71, n. 356, p. 791–799, 1976. DOI: [10.1080/01621459.1976.10480949](https://doi.org/10.1080/01621459.1976.10480949).

BROWN, A C; FRASER, T R. On the connection between chemical constitution and physiological action: With special reference to the physiological action of the salts of the ammonium bases derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia. **J. Anat. Physiol.**, v. 2, n. 2, p. 224–242, 1868. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/17230757/>. Acesso em: 3 maio 2025.

CAMUS, Albert. **Le Mythe de Sisyphe**. Paris: Gallimard, 1942.

CAMUS, Albert. **O mito de Sísifo**. Brazil: Record, 2018. ISBN 978-85-0111164-7.

CHAN, T. M. Optimal output-sensitive convex hull algorithms in two and three dimensions. **Discrete & Computational Geometry**, Springer Science e Business Media, v. 16, n. 4, p. 361–368, abr. 1996. DOI: [10.1007/bf02712873](https://doi.org/10.1007/bf02712873).

CHERKASOV, Artem *et al.* QSAR modeling: where have you been? Where are you going to? **Journal of Medicinal Chemistry**, v. 57, n. 12, p. 4977–5010, 2014. DOI: [10.1021/jm4004285](https://doi.org/10.1021/jm4004285).

CRAMER, Richard D.; PATTERSON, David E.; BUNCE, Jeffrey D. Comparative molecular field analysis (CoMFA): 1. Effect of shape on binding of steroids to carrier proteins. **Journal of the American Chemical Society**, v. 110, n. 18, p. 5959–5967, 1988. DOI: [10.1021/ja00226a005](https://doi.org/10.1021/ja00226a005).

CURCHOE, Carol Lynn. All Models Are Wrong, but Some Are Useful. **Journal of Assisted Reproduction and Genetics**, Springer Science e Business Media LLC, v. 37, n. 10, p. 2389–2391, 2020. DOI: [10.1007/s10815-020-01895-3](https://doi.org/10.1007/s10815-020-01895-3).

DALBY, Arthur *et al.* Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. **Journal of Chemical Information and Computer Sciences**, American Chemical Society, v. 32, n. 3, p. 244–255, maio 1992. DOI: [10.1021/ci00007a012](https://doi.org/10.1021/ci00007a012).

DAMALE, Manoj *et al.* Recent Advances in Multidimensional QSAR (4D-6D): A Critical Review. **Mini-Reviews in Medicinal Chemistry**, Bentham Science Publishers Ltd., v. 14, n. 1, p. 35–55, jan. 2014. DOI: [10.2174/13895575113136660104](https://doi.org/10.2174/13895575113136660104).

DANISHUDDIN; KHAN, Aamir U. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. **Drug Discovery Today**, v. 21, n. 8, p. 1291–1302, 2016. DOI: [10.1016/j.drudis.2016.05.016](https://doi.org/10.1016/j.drudis.2016.05.016).

DAS, Pratik Swarup; SAHA, Puja. A review on Computer Aided Drug Design in Drug Discovery. **World Journal of Pharmacy and Pharmaceutical Sciences**, v. 6, n. 7, p. 279–291, 2017. DOI: [10.20959/wjpps20177-9450](https://doi.org/10.20959/wjpps20177-9450). Disponível em: https://www.researchgate.net/publication/322163551_A_REVIEW_ON_COMPUTER_AIDED_DRUG_DESIGN_IN_DRUG_DISCOVERY. Acesso em: 3 maio 2025.

DAYLIGHT CHEMICAL INFORMATION SYSTEMS, INC. **Daylight Theory Manual**. 2011. Disponível em: <https://www.daylight.com/dayhtml/doc/theory/index.html>. Acesso em: 29 maio 2025.

DOYTCHINOVA, Irini. Drug Design: Past, Present, Future. **Molecules**, v. 27, n. 5, p. 1496, fev. 2022. DOI: [10.3390/molecules27051496](https://doi.org/10.3390/molecules27051496). Acesso em: 20 maio 2025.

EASTMAN, Peter *et al.* OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. **The Journal of Physical Chemistry B**, American Chemical Society (ACS), v. 128, n. 1, p. 109–116, dez. 2023. DOI: [10.1021/acs.jpcb.3c06662](https://doi.org/10.1021/acs.jpcb.3c06662).

FERREIRA, Elizabeth Igne *et al.* **Fundamentos de química farmacêutica medicinal**. 1. ed. Santana de Parnaíba: Manole, 2022. ISBN 9786555762600.

FISHER, Aaron; RUDIN, Cynthia; DOMINICI, Francesca. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. **ArXiv e-prints**, 2018. DOI: [10.48550/arXiv.1801.01489](https://doi.org/10.48550/arXiv.1801.01489). Acesso em: 31 maio 2025.

FORTIN, Félix-Antoine *et al.* DEAP: Evolutionary Algorithms Made Easy. **Journal of Machine Learning Research**, v. 13, n. 70, p. 2171–2175, 2012. Disponível em: <https://www.jmlr.org/papers/volume13/fortin12a/fortin12a.pdf>. Acesso em: 5 mar. 2025.

FRANZEN, Daniel; MÜLLER-BIRN, Claudia; WEGWARTH, Odette. Communicating the Privacy-Utility Trade-off: Supporting Informed Data Donation with Privacy Decision Interfaces for Differential Privacy. **Proceedings of the ACM on Human-Computer Interaction**, Association for Computing Machinery (ACM), v. 8, CSCW1, p. 1–56, abr. 2024. ISSN 2573-0142. DOI: [10.1145/3637309](https://doi.org/10.1145/3637309).

GRISONI, Francesca *et al.* Molecular Descriptors for Structure-Activity Applications: A Hands-On Approach. In: **METHODS in Molecular Biology**. [S. l.: s. n.], 2018. v. 1800. p. 3–53. DOI: [10.1007/978-1-4939-7899-1_1](https://doi.org/10.1007/978-1-4939-7899-1_1).

- HAMMETT, Louis P. Some Relations between Reaction Rates and Equilibrium Constants. **Chemical Reviews**, v. 17, n. 1, p. 125–136, 1935. DOI: [10.1021/cr60056a010](https://doi.org/10.1021/cr60056a010).
- HANSCH, Corwin; FUJITA, Toshio. p- σ - π Analysis: A Method for the Correlation of Biological Activity and Chemical Structure. **Journal of the American Chemical Society**, v. 86, n. 8, p. 1616–1626, 1964. DOI: [10.1021/ja01062a035](https://doi.org/10.1021/ja01062a035).
- HARADA, Koichiro *et al.* Discovery of potent and orally bioavailable 17 β -hydroxysteroid dehydrogenase type 3 inhibitors. **Bioorg. Med. Chem.**, Elsevier, v. 20, n. 10, p. 3242–3254, maio 2012. ISSN 1464-3391. DOI: [10.1016/j.bmc.2012.03.052](https://doi.org/10.1016/j.bmc.2012.03.052).
- HAWKINS, Paul C. D. Conformation Generation: The State of the Art. **Journal of Chemical Information and Modeling**, American Chemical Society, v. 57, n. 8, p. 1747–1756, ago. 2017. DOI: [10.1021/acs.jcim.7b00221](https://doi.org/10.1021/acs.jcim.7b00221). Acesso em: 30 maio 2025.
- HOPFINGER, A. J. *et al.* Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. **Journal of the American Chemical Society**, v. 119, n. 43, p. 10509–10524, 1997. DOI: [10.1021/ja9718937](https://doi.org/10.1021/ja9718937).
- JENSEN, Frank. **Introduction to Computational Chemistry**. Hoboken: Wiley, jul. 2017. ISBN 978-1118825990.
- KENNARD, R. W.; STONE, L. A. Computer Aided Design of Experiments. **Technometrics**, Taylor & Francis, v. 11, n. 1, p. 137–148, 1969. ISSN 00401706. DOI: [10.2307/1266770](https://doi.org/10.2307/1266770).
- LANDRUM, Greg *et al.* RDKit: Open-source cheminformatics. Zenodo, 2025. DOI: [10.5281/ZENODO.591637](https://doi.org/10.5281/ZENODO.591637). Disponível em: <https://www.rdkit.org>. Acesso em: 16 jun. 2025.
- LEARDI, Riccardo; BOGGIA, Rita; TERRILE, Monica. Genetic algorithms as a strategy for feature selection. **Journal of Chemometrics**, v. 6, n. 5, p. 267–281, 1992. DOI: [10.1002/cem.1180060506](https://doi.org/10.1002/cem.1180060506).
- LU, Wenchao *et al.* Computer-Aided Drug Design in Epigenetics. **Frontiers in Chemistry**, Frontiers Media SA, v. 6, mar. 2018. ISSN 2296-2646. DOI: [10.3389/fchem.2018.00057](https://doi.org/10.3389/fchem.2018.00057).
- MARTINS, João Paulo A.; BARBOSA, Euzébio G. *et al.* LQTA-QSAR: A New 4D-QSAR Methodology. **Journal of Chemical Information and Modeling**, v. 49, n. 6, p. 1428–1436, 2009. DOI: [10.1021/ci900014f](https://doi.org/10.1021/ci900014f).
- MARTINS, João Paulo Ataíde. Química computacional aplicada a QSAR. Campinas, mar. 2010.
- MARTINS, João Paulo Ataíde; MELO, Eduardo de. 4D-QSAR Study of 17 β -Hydroxysteroid Dehydrogenase Type 3 Inhibitors. **Journal of the Brazilian Chemical Society**, Sociedade Brasileira de Química (SBQ), 2019. DOI: [10.21577/0103-5053.20190052](https://doi.org/10.21577/0103-5053.20190052).
- MARTINS, João Paulo Ataíde; OLIVEIRA, Marco Antônio Rougeth de; QUEIROZ, Mário Sérgio Oliveira de. Web-4D-QSAR: A web-based application to generate 4D-QSAR descriptors. **Journal of Computational Chemistry**, Wiley, v. 39, n. 15, p. 917–924, fev. 2018. DOI: [10.1002/jcc.25166](https://doi.org/10.1002/jcc.25166).
- MARTINS, Joao Paulo Ataíde; REIS FILHO, Héilton Martins. QSARModelingPy: a Python package to build and validate QSAR models. In: 43ª REUNIÃO ANUAL DA SBQ. **Anais da 43ª reunião anual virtual da SBQ**. Campinas: Galoá, 2020. SBQ. Disponível em: <https://proceedings.science/rasbq->

2020/trabalhos/qsarmodelingpy-a-python-package-to-build-and-validate-qsar-models. Acesso em: 4 dez. 2021.

MAURI, Andrea *et al.* DRAGON software: An easy approach to molecular descriptor calculations. **MATCH Communications in Mathematical and in Computer Chemistry**, v. 56, p. 237–248, jan. 2006. Disponível em: https://match.pmf.kg.ac.rs/electronic_versions/Match56/n2/match56n2_237-248.pdf. Acesso em: 16 jun. 2025.

MILLS, Edmund J. XXIII. On melting-point and boiling-point as related to chemical composition. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, Taylor & Francis, v. 17, n. 105, p. 173–187, 1884. DOI: [10.1080/14786448408627502](https://doi.org/10.1080/14786448408627502).

MULYO, Budi Mukhamad *et al.* Document Classification Based on Computational Models: Insights from Computer Science Case Studies. *In: 2024 IEEE 10th Information Technology International Seminar (ITIS)*. [S. l.]: IEEE, nov. 2024. p. 1–5. DOI: [10.1109/itis64716.2024.10845503](https://doi.org/10.1109/itis64716.2024.10845503).

MURRAY-RUST, Peter; RZEPA, Henry S. Chemical Markup, XML, and the Worldwide Web: 1. Basic Principles. **Journal of Chemical Information and Computer Sciences**, American Chemical Society (ACS), v. 39, n. 6, p. 928–942, set. 1999. ISSN 0095-2338. DOI: [10.1021/ci990052b](https://doi.org/10.1021/ci990052b).

NIAZI, Sarfaraz K.; MARIAM, Zamara. Computer-Aided Drug Design and Drug Discovery: A Prospective Analysis. **Pharmaceuticals (Basel)**, v. 17, n. 1, p. 22, 2023. DOI: [10.3390/ph17010022](https://doi.org/10.3390/ph17010022).

O'BOYLE, Noel M. *et al.* Open Babel: An open chemical toolbox. **J. Cheminf.**, BioMed Central, v. 3, n. 1, p. 1–14, dez. 2011. ISSN 1758-2946. DOI: [10.1186/1758-2946-3-33](https://doi.org/10.1186/1758-2946-3-33).

PETROCCHI, Alessia *et al.* From dihydroxypyrimidine carboxylic acids to carboxamide HIV-1 integrase inhibitors: SAR around the amide moiety. **Bioorganic & Medicinal Chemistry Letters**, Elsevier BV, v. 17, n. 2, p. 350–353, jan. 2007. ISSN 0960-894X. DOI: [10.1016/j.bmcl.2006.10.054](https://doi.org/10.1016/j.bmcl.2006.10.054).

ROY, Kunal; KAR, Supratik; DAS, Rudra Narayan. **Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment**. London: Academic Press, 3 mar. 2015. p. 484. ISBN 9780128016336.

SHAKESPEARE, William. **Hamlet, Prince of Denmark**. Londres: N. Ling, 1603. Ato I, Cena V, verso 166.

SHI, Xuesong *et al.* Quantifying Nucleic Acid Ensembles with X-ray Scattering Interferometry. Chapter Three. *In: WOODSON, Sarah A.; ALLAIN, Frédéric H.T. (ed.). Structures of Large RNA Molecules and Their Complexes*. [S. l.]: Academic Press, 2015. v. 558. (Methods in Enzymology). p. 75–97. DOI: [10.1016/bs.mie.2015.02.001](https://doi.org/10.1016/bs.mie.2015.02.001).

TENÓRIO, João Vitor; CERF, Loïc; MARTINS, João Paulo Ataíde. A New Approach for Sampling Descriptors in 4D-QSAR Methodology Using Computational Geometry. *In: XIX SIMPÓSIO BRASILEIRA DE QUÍMICA TEÓRICA*, 2017. Atos do XIX Simpósio Brasileiro de Química Teórica. Águas de Lindóia: [s. n.], 2017. Disponível em: https://homepages.dcc.ufmg.br/~lcerf/publications/articles/A_New_Approach_for_Sampling_Descriptors_in_4D-QSAR_Methodology_Using_Computational_Geometry.pdf. Acesso em: 4 dez. 2021.

TENÓRIO, João Vitor Soares. **Uma Nova Amostragem de Descritores para Predição de Atividade Biológica**. 2018. Diss. (Mestrado) – Universidade Federal de Minas Gerais, Belo Horizonte.

TODESCHINI, Roberto; CONSONNI, Viviana. **Handbook of Molecular Descriptors**. [S. l.]: Wiley, set. 2000. ISBN 9783527613106. DOI: [10.1002/9783527613106](https://doi.org/10.1002/9783527613106).

TRIPOS INC. **SYBYL Mol2 File Format**. [S. l.: s. n.], 1994. Formato de arquivo molecular desenvolvido para o software SYBYL. Disponível em: <https://zhanggroup.org/DockRMSD/mol2.pdf>. Acesso em: 11 maio 2025.

VEDANI, Angelo; DOBLER, Max. 5D-QSAR: The Key for Simulating Induced Fit? **Journal of Medicinal Chemistry**, American Chemical Society (ACS), v. 45, n. 11, p. 2139–2149, abr. 2002. DOI: [10.1021/jm011005p](https://doi.org/10.1021/jm011005p).

WARREN, Geoffrey J.; RYAN, Laura; GAN, Baoqing. All Models Are Wrong, Some Might Be OK! Elsevier BV, 2025. DOI: [10.2139/ssrn.5165699](https://doi.org/10.2139/ssrn.5165699).

WEININGER, David. SMILES, a chemical language and information system: 1. Introduction to methodology and encoding rules. **Journal of Chemical Information and Computer Sciences**, v. 28, n. 1, p. 31–36, 1988. DOI: [10.1021/ci00057a005](https://doi.org/10.1021/ci00057a005).

WIKIMEDIA COMMONS. **Correlation examples**. 9 maio 2011. Disponível em: https://en.wikipedia.org/wiki/File:Correlation_examples2.svg. Acesso em: 22 maio 2025.

WIT, Ernst; HEUVEL, Edwin van den; ROMELIJN, Jan-Willem. “All models are wrong...” an introduction to model uncertainty. **Statistica Neerlandica**, Wiley, v. 66, n. 3, p. 217–236, 2012. DOI: [10.1111/j.1467-9574.2012.00530.x](https://doi.org/10.1111/j.1467-9574.2012.00530.x).

YAP, Chun Wei. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. **Journal of Computational Chemistry**, v. 32, n. 7, p. 1466–1474, 2011. DOI: [10.1002/jcc.21707](https://doi.org/10.1002/jcc.21707).

Glossário

Este glossário reúne os principais termos técnicos utilizados ao longo deste trabalho, com o objetivo de facilitar a compreensão e garantir a clareza na leitura.

fecho convexo

(ou *convex hull* em inglês) de um conjunto de pontos X se refere, em geometria computacional, ao menor conjunto convexo que contém X . Sua definição é dada pela [Equação 5.6](#). pp. [36–38](#), [44](#), [45](#), [47](#)

heterocedasticidade

característica oposta à homocedasticidade. p. [36](#)

homocedasticidade

refere-se à igualdade da variância dos erros em um modelo de regressão, ou seja, a dispersão dos pontos de dados em torno da linha de regressão é constante para todos os valores da variável independente. Em outras palavras, os erros do modelo são independentes dos valores da variável explicativa.

in silico

é uma expressão latina usada no âmbito da simulação computacional e áreas correlatas para indicar algo ocorrido “em ou através de uma simulação computacional”. Foi inspirada nas expressões *in vivo* e *in vitro*, amplamente usadas nas ciências biológicas para designar, respectivamente, experimentos feitos com animais reais e aqueles realizados em laboratório. p. [9](#), [12](#)

outlier

Valor ou observação que se desvia significativamente do padrão geral dos dados. Em contextos estatísticos, um *outlier* é uma instância cuja magnitude ou comportamento é atípico em relação à distribuição da amostra, podendo influenciar de maneira desproporcional ajustes de modelos e medidas como média e variância. A detecção e o tratamento de *outliers* são etapas fundamentais na análise de dados e na validação de modelos preditivos.. p. [36](#), [52](#)

sonda

ou *probe atom*. pp. [34](#), [35](#), [37](#), [38](#), [40](#), [42](#), [57](#)

Apêndices

APÊNDICE A – Medidas Estatísticas

A avaliação de modelos **QSAR** requer, invariavelmente, o uso criterioso de parâmetros estatísticos que caracterizem o comportamento dos dados experimentais e preditos. Dentre esses parâmetros, a média e a variância desempenham papéis centrais tanto na descrição quanto na quantificação da qualidade dos ajustes e previsões.

Considere um conjunto de dados unidimensionais $x = \{x_1, x_2, \dots, x_N\}$ com N elementos. A “média aritmética” dos valores, representada por $\bar{x}$ ou $\langle x \rangle$, é definida pela [Equação A.1](#):

$$\bar{x} \equiv \langle x \rangle \equiv \frac{1}{N} \sum_{i=1}^N x_i \quad (\text{A.1})$$

Esta grandeza corresponde ao valor médio do conjunto, ou seja, ao seu centro de gravidade estatístico. Embora seja uma medida de tendência central amplamente utilizada, a média, por si só, não é suficiente para caracterizar a distribuição dos dados — duas distribuições com a mesma média podem apresentar dispersões significativamente distintas.

Para quantificar essa dispersão em torno da média, introduz-se a “variância amostral”, denotada por σ^2 e dada pela [Equação A.2](#):

$$\sigma^2 \equiv \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 \quad (\text{A.2})$$

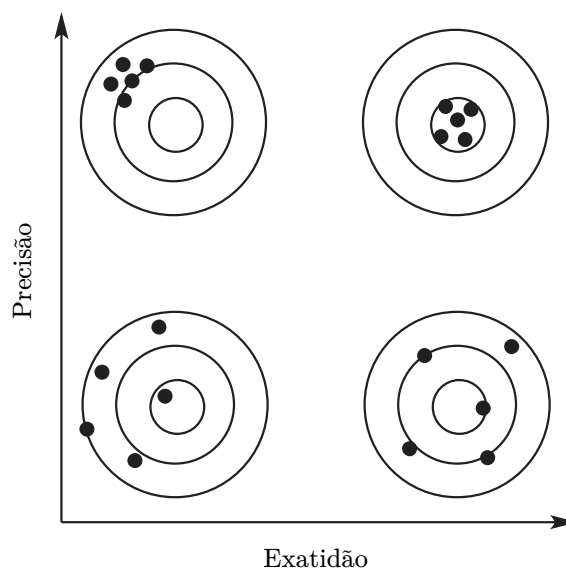
A variância mede a expectativa do quadrado do desvio em relação à média, penalizando mais fortemente valores afastados do centro.¹ A raiz quadrada da variância, conhecida como desvio padrão (σ), possui a mesma unidade dos dados originais e permite uma interpretação mais direta da dispersão:

$$\sigma = \sqrt{\sigma^2}$$

Essas medidas — média e desvio padrão — serão empregadas de forma recorrente ao longo deste trabalho, tanto na avaliação de propriedades físico-químicas quanto na análise de resíduos e na validação interna e externa dos modelos **QSAR**.

¹ O fator $N - 1$ no denominador corresponde à correção de Bessel, necessária quando a variância é estimada a partir de uma amostra finita da população.

Figura 9 – Ilustração da diferença entre “precisão” e “exatidão”



Fonte: Adaptado de [Jensen \(2017\)](#).

A.0.1 Precisão e Exatidão

Uma distinção conceitual frequentemente negligenciada, mas de suma importância para a interpretação adequada do processo de modelagem, é a diferença entre a *precisão* e a *exatidão* dos dados.

A precisão (*precision* em inglês) diz respeito à baixa variabilidade das medições: um conjunto de dados é considerado preciso quando os valores observados se concentram em torno de um mesmo ponto, apresentando pequena dispersão. Por outro lado, a exatidão (*accuracy* em inglês) refere-se à proximidade entre a média das medições e o valor verdadeiro ou esperado.

Essa diferença conceitual é ilustrada de maneira visual na [Figura 9](#). Imagine que cada conjunto de círculos concêntricos representa um alvo, e os pontos marcados correspondem a dardos lançados em sua direção. No quadrante inferior esquerdo, observa-se uma distribuição dispersa (alta variância, baixa precisão) cuja média também está afastada do centro (baixa exatidão). No extremo oposto, no quadrante superior direito, os lançamentos são agrupados (baixa variância, alta precisão) e concentram-se em torno do centro (alta exatidão).

Os demais quadrantes ilustram situações intermediárias: no canto superior esquerdo, há alta precisão com baixa exatidão — os lançamentos são consistentes entre si, mas sistematicamente desviados do centro. Já no quadrante inferior direito, observa-se o oposto: os valores são dispersos (baixa precisão), mas a média coincide com o centro (alta exatidão), revelando um acerto em média, apesar da variabilidade individual.

A.0.2 Medidas de Erro em Modelos Preditivos

Nesse cenário, é essencial dispor de ferramentas que quantifiquem, de forma objetiva, o desempenho do modelo ajustado. A qualidade dessa aproximação pode ser avaliada por meio de métricas que comparam os valores *preditos* $\hat{y}_i$ com os *observados* y_i , resultando em medidas formais de erro.

Dentre as medidas mais utilizadas destacam-se o Desvio Quadrático Médio (**RMS**) e o Desvio Absoluto Médio (**MAD**), que quantificam, de formas distintas, a discrepância entre os valores preditos $\hat{y}_i$ e os valores reais y_i para cada instância i do conjunto de dados.

A métrica **RMS**, definida na [Equação A.3](#), penaliza mais severamente os erros grandes, pois eleva ao quadrado cada desvio antes da média ser computada. Por esse motivo, é particularmente sensível à presença de *outliers*. Em contrapartida, o **MAD** ([Equação A.4](#)) calcula a média dos módulos dos desvios, oferecendo uma visão mais robusta da magnitude média dos erros, sem amplificar de forma desproporcional os desvios extremos.

$$\text{RMS} \equiv \sqrt{\frac{1}{M} \sum_{i=1}^M (y_i - \hat{y}_i)^2} \quad (\text{A.3})$$

$$\text{MAD} \equiv \frac{1}{M} \sum_{i=1}^M |y_i - \hat{y}_i| \quad (\text{A.4})$$

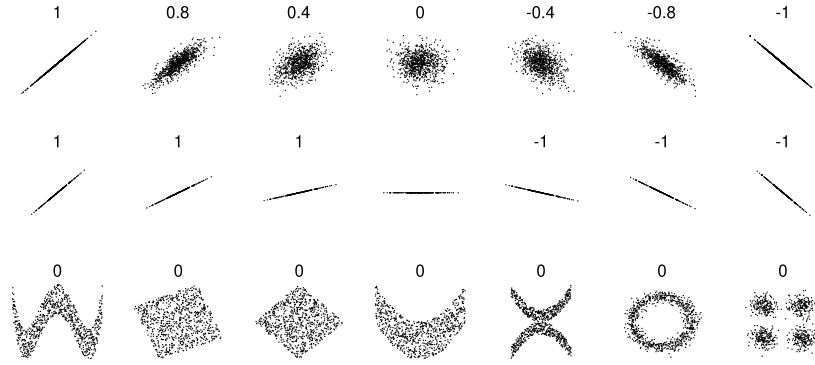
Ambas as métricas estão expressas nas mesmas unidades da variável dependente y e desempenham papel crucial na avaliação quantitativa da qualidade preditiva dos modelos. O uso simultâneo de **RMS** e **MAD** permite uma análise mais completa do comportamento dos resíduos e das propriedades estatísticas do ajuste.

A.0.3 Correlação entre dois conjuntos de dados

Ao avaliar modelos de regressão, como aqueles empregados em estudos **QSAR**, não basta quantificar apenas o erro médio entre os valores preditos e experimentais. É igualmente importante examinar a *associação estatística* entre os conjuntos de dados — em especial, o grau de dependência linear entre eles. Essa associação é expressa por meio de medidas de *correlação*, que descrevem em que medida as variações em uma variável se acompanham, de forma sistemática, por variações em outra.

A mais amplamente utilizada dessas medidas é o *coeficiente de correlação de Pearson*, definido pela [Equação A.5](#), onde y_i são os valores experimentais ou esperados, $\hat{y}_i \equiv f(\mathbf{x}_i)$ são os valores preditos, $\bar{y}$ são suas médias, e M é o número total de observações.

Figura 10 – Vários conjuntos de pontos (x, y) , com o coeficiente de correlação entre x e y indicado para cada conjunto



Fonte: [Wikimedia Commons](#) (2011).

$$R \equiv \frac{\sum_{i=1}^M (\hat{y}_i - \bar{y})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^M (\hat{y}_i - \bar{y})^2} \sqrt{\sum_{i=1}^M (y_i - \bar{y})^2}} \quad (\text{A.5})$$

O valor de R está contido no conjunto $[-1, 1]$ e sua interpretação é tal que: $R > 0 \Rightarrow$ associação linear positiva; $R < 0 \Rightarrow$ associação linear negativa; e $R = 0 \Rightarrow$ ausência de correlação linear. A [Figura 10](#) ilustra diferentes conjuntos de pontos e suas respectivas correlações R . Note que a correlação reflete a intensidade e a direção de uma relação linear (linha superior na figura), mas não a inclinação dessa relação (linha do meio), nem diversos aspectos de relações não lineares (linha inferior). A figura central, por sua vez, apresenta variância nula em y e, portanto, R não está definido.

A.0.4 Medidas de Qualidade

De forma similar ao coeficiente de correlação ([Equação A.5](#)), outras grandezas relevantes na medição da qualidade do ajuste produzido por um dado modelo são a PRESS, a variância de y (σ_y^2) e o coeficiente de determinação R^2 .

$$\text{PRESS} \equiv \sum_{i=1}^M (y_i - \hat{y}_i)^2 \quad (\text{A.6})$$

$$\sigma_y^2 \equiv \frac{1}{N} \sum_{i=1}^M (y_i - \bar{y})^2 \quad (\text{A.7})$$

$$R^2 \equiv 1 - \frac{\text{PRESS}}{N\sigma_y^2} = 1 - \frac{\sum_{i=1}^M (y_i - \hat{y}_i)^2}{\sum_{i=1}^M (y_i - \bar{y})^2} \quad (\text{A.8})$$

O *coeficiente de determinação*, $R^2 \in [0, 1]$, representa a proporção da variância de y que pode ser explicada pela variância de $\hat{y}$, sob uma relação linear. Assim, R^2 pode ser interpretado como uma medida da *qualidade do ajuste* de um modelo de regressão.² Por exemplo, $R^2 = 0,85$ indica que 85% da variância dos valores de y é explicada pela variância em $\hat{y}$. Em outras palavras, o modelo preserva 85% das informações do conjunto de dados original.

No raro extremo, um $R^2 = 1$ indicaria que a totalidade da variância de y pode ser explicada por $\hat{y}$, o que, em outros termos, significa $y = \hat{y}$: um ajuste perfeito dos dados.³

A.0.5 Validação Cruzada

É importante salientar que o R^2 quantifica o grau de correlação entre os valores preditos e os valores experimentais. Um valor elevado de R^2 indica que o modelo *se ajusta bem* aos dados observados y . No entanto, mais do que reproduzir os dados disponíveis, desejamos avaliar a *capacidade preditiva* do modelo — isto é, sua habilidade de generalizar para novas observações a partir das variáveis independentes.

Para isso, recorre-se a técnicas de *validação cruzada*, as quais visam estimar o desempenho preditivo do modelo com base apenas em subconjuntos dos dados. A forma mais elementar dessa abordagem consiste em ajustar o modelo a um subconjunto com $M - 1$ moléculas (por exemplo, $x_1, \dots, x_{m-1}$), e utilizar esse modelo para prever o valor da observação deixada de fora, $\hat{y}_m$. Repetindo esse procedimento para cada uma das M observações, obtém-se o método conhecido como *leave-one-out* (LOO).

A partir dessas predições externas, pode-se calcular a PRESS, cujo valor serve como indicativo da capacidade preditiva do modelo. Com base nela, define-se também o *coeficiente de determinação preditivo*, denotado por Q^2 e definido em analogia direta ao R^2 apresentado na [Equação A.8](#). Em termos gerais, valores de $Q^2 > 0,5$ são comumente considerados aceitáveis para modelos QSAR, indicando desempenho preditivo satisfatório (Jensen, 2017).

² Na literatura em língua inglesa, o termo *goodness of fit* é amplamente utilizado para descrever a qualidade do ajuste.

³ Note que este é um caso muito raro, especialmente em análises de conjuntos de dados reais. Embora desejável, a obtenção de $R^2 = 1$ deve evocar, no analista, suspeitas em relação à implementação do modelo e das métricas de validação.

APÊNDICE B – Classificação de descritores QSAR quanto à dimensionalidade

Uma das formas mais tradicionais de classificar os descritores moleculares é pela sua dimensionalidade, ou seja, pelo nível de complexidade estrutural que cada descritor incorpora sobre a molécula. Essa classificação vai desde informações mais simples, como o número total de átomos (0D), até representações que incluem o espaço tridimensional e até a dinâmica conformacional da molécula (4D ou mais).

Cada aumento na dimensionalidade traz consigo a possibilidade de capturar aspectos mais sutis e relevantes das interações moleculares, mas também aumenta o custo computacional e a complexidade de interpretação.

B.1 QSAR “clássico”: 0D, 1D e 2D

As primeiras abordagens de QSAR, historicamente chamadas de “clássicas”, basearam-se em descritores simples, hoje conhecidos como descritores 0D, 1D e 2D.

Descritores 0D: Informação escalar e global.

Os descritores 0D representam características moleculares que não dependem da forma ou da conectividade dos átomos na molécula. São propriedades globais, como massa molecular, ponto de fusão, número total de átomos, coeficiente de partição octanol/água ($\log P$)...

Esses descritores são fáceis de calcular, rápidos e oferecem uma visão geral da molécula. No entanto, são pouco informativos quando o objetivo é prever propriedades que dependem da forma ou da distribuição eletrônica.

Descritores 1D: Contagem de grupos químicos.

Os descritores 1D introduzem um pouco mais de detalhe, considerando a presença de grupos funcionais ou certos tipos de átomos. Exemplos típicos incluem:

- Número de doadores ou aceptores de ligação de hidrogênio;
- Contagem de heteroátomos (oxigênio, nitrogênio, enxofre, etc.);

- Número de ligações rotacionáveis;
- Frequência de determinados grupos funcionais...

Esses descritores já começam a fornecer uma impressão química mais específica da molécula, refletindo algumas de suas propriedades de reatividade ou solubilidade.

Descritores 2D: A forma como os átomos estão conectados.

Ao avançarmos para a segunda dimensão, passamos a considerar a topologia molecular, ou seja, como os átomos estão conectados entre si. Aqui entram descritores derivados diretamente do grafo molecular, como:

- Índices de ramificação e de caminho (ex.: Wiener, Balaban, Zagreb, Hosoya);
- Índices de conectividade (ex.: Randić, Kier-Hall);
- Número de ciclos, tamanho dos anéis e grau de ramificação;
- Propriedades extraídas de matrizes de adjacência...

Esses descritores capturam informações sobre o “formato lógico” da molécula, mesmo sem considerar sua geometria real no espaço tridimensional. Por exemplo, dois isômeros de cadeia podem ter diferentes valores de índice de ramificação, mesmo que tenham a mesma fórmula molecular.

De forma geral, os descritores 0D, 1D e 2D são extremamente úteis quando o objetivo é trabalhar com grandes bancos de moléculas, devido ao baixo custo computacional. No entanto, eles não conseguem diferenciar isômeros espaciais (estereoisômeros) nem prever com precisão propriedades que dependem da geometria molecular.

Algumas ferramentas livres e amplamente utilizadas (Yap, 2011; Landrum *et al.*, 2025; Mauri *et al.*, 2006) permitem a geração automática de milhares desses descritores em questão de segundos.

B.2 QSAR 3D

Quando a posição relativa dos átomos no espaço passa a influenciar de forma significativa a bioatividade, torna-se necessário utilizar descritores tridimensionais, inaugurando o campo da 3D-QSAR.

Ao contrário dos descritores topológicos, os descritores 3D incorporam a geometria real da molécula: as coordenadas cartesianas dos átomos após otimização estrutural. Isso permite capturar aspectos como:

- Forma tridimensional (volume, superfície);
- Distribuições de carga;
- Potenciais estéricos (relacionados a repulsão espacial entre átomos);
- Potenciais eletrostáticos (relacionados à distribuição de cargas)...

O método mais representativo dessa abordagem é a CoMFA, proposta por [Cramer, Patterson e Bunce \(1988\)](#). Nesse método, cada molécula é posicionada dentro de uma grade tridimensional e, em cada ponto dessa grade, calcula-se como uma sonda virtual (geralmente um íon $C^+ sp^3$) interage com a molécula via forças de van der Waals e interações eletrostáticas.

O resultado é um grande vetor numérico, no qual cada valor corresponde a uma interação potencial em um ponto específico do espaço. Esses vetores alimentam modelos estatísticos como o PLS, permitindo a correlação entre o campo molecular e a atividade biológica.

A grande vantagem da 3D-QSAR é que ela reflete com maior realismo o processo de reconhecimento molecular. Sua principal limitação é que ela exige que todas as moléculas estejam bem alinhadas em um referencial comum e que se escolha, previamente, uma única conformação representativa de cada molécula. Isso pode introduzir erros se a molécula apresentar muita flexibilidade.

B.3 QSAR 4D

A 4D-QSAR surge para resolver justamente essa limitação da 3D-QSAR. Em vez de trabalhar com uma única conformação fixa, o 4D-QSAR considera um conjunto de várias conformações possíveis que a molécula pode assumir ao longo do tempo. Essa coleção de estruturas é geralmente obtida por técnicas como a simulação de Dinâmica Molecular ([seção 3.5](#)).

A sobreposição e alinhamento de todas essas conformações produz o CEP ([subseção 3.5.1](#)), que pode ser interpretado como um “perfil médio” do espaço conformacional da molécula ao longo do tempo. O descritor final reflete, portanto, uma média de Boltzmann sobre o espaço conformacional acessível, incorporando efeitos de flexibilidade que a 3D-QSAR tradicional não contempla.

A grande vantagem da 4D-QSAR é que ela incorpora, de forma explícita, a flexibilidade molecular e o comportamento dinâmico das moléculas, aspectos cruciais para muitas interações biológicas. Além disso, ao suavizar os efeitos de alinhamento rígido, o método reduz o risco de erro associado à escolha de uma única conformação.

Essa metodologia, proposta inicialmente por Hopfinger *et al.* (1997), representa um passo importante rumo a modelos mais realistas de reconhecimento molecular, sendo particularmente útil em sistemas onde múltiplos modos de ligação coexistem ou onde a adaptação conformacional é determinante para a atividade.

B.4 Dimensionalidades mais altas: 5D, 6D e além

Com a consolidação da 4D-QSAR, persistiram duas fontes de incerteza que limitavam a acurácia preditiva dos modelos: (i) a incapacidade de representar, de forma explícita, a acomodação conformacional do sítio de ligação (*induced-fit*) e (ii) a negligência dos efeitos de solvatação diferencial que acompanham o processo de reconhecimento molecular. As extensões 5D e 6D-QSAR surgem precisamente para suprir essas lacunas.

Na quinta dimensão (5D-QSAR), introduz-se um *ensemble* de hipóteses de *induced-fit*: para cada ligante, múltiplas variantes topológicas do receptor virtual são geradas e evoluídas geneticamente durante a otimização do modelo. Essa estratégia, segundo Vedani e Dobler (2002), reduz o viés associado à escolha da conformação bioativa e melhora a robustez estatística do conjunto de modelos.

Já no 6D-QSAR, a sexta dimensão incorpora cenários múltiplos de solvatação (explícita ou implícita), permitindo que diferentes modelos de retirada de solvente (desolvatação do ligante) e de reorganização da água no sítio sejam simultaneamente ponderados pelo algoritmo evolutivo. Damale *et al.* (2014) relatam que a inclusão dessa dimensão elevou o R^2 médio de 0,872 (5D) para 0,903 (6D), destacando o contributo decisivo do termo de solvatação na afinidade calculada. Em termos práticos, o 6D-QSAR é particularmente útil quando efeitos de hidratação, ou a competição entre estados protônicos em meio aquoso, sejam determinantes para a bioatividade.

Propostas recentes aventam dimensões adicionais para acomodar, por exemplo, múltiplos estados protônicos (5D'), heterogeneidade conformacional do próprio receptor (6D') ou até variações de pH e força iônica (7D). Ainda que promissoras, tais extensões carecem de validação sistemática e demandam custo computacional crescente; portanto, sua adoção deve ser criteriosa, ponderando-se a relação benefício-complexidade.

Em síntese, as abordagens 5D e 6D representam um salto qualitativo na modelagem QSAR ao integrar explicitamente flexibilidade do receptor e solvatação, respectivamente, elevando o rigor físico-químico sem sacrificar a viabilidade computacional — conquistas que se refletem em ganhos empiricamente mensuráveis de capacidade preditiva (Vedani; Dobler, 2002; Damale *et al.*, 2014).

APÊNDICE C – Resultados Detalhados do HullQSAR

Este apêndice contém os resultados estatísticos detalhados para todos os modelos HullQSAR construídos, variando-se o parâmetro $\Delta\alpha$.

Tabela 6 – Resultados obtidos pelo HullQSAR para cada conjunto de dados, com o respectivo melhor modelo destacado em azul.

$\Delta\alpha$	CANCER				CARBOX			
	4°	6°	8°	10°	4°	6°	8°	10°
R^2	0,784	0,711	0,692	0,754	0,861	0,882	0,947	0,916
PRESS	5,263	7,049	7,520	5,991	5,521	4,703	2,110	3,340
RMSEcal	0,328	0,379	0,392	0,350	0,409	0,378	0,253	0,318
rcal	0,885	0,843	0,832	0,868	0,928	0,939	0,973	0,957
avgRmcal	0,697	0,600	0,575	0,657	0,802	0,831	0,923	0,879
deltaRmcal	0,174	0,222	0,233	0,194	0,117	0,101	0,047	0,073
Q^2	0,725	0,723	0,663	0,687	0,784	0,802	0,883	0,861
PRESScv	6,695	6,741	8,222	7,630	8,597	7,879	4,667	5,506
RMSEcv	0,370	0,371	0,410	0,395	0,510	0,489	0,376	0,408
rcv	0,853	0,853	0,814	0,831	0,886	0,896	0,941	0,928
avgRmcv	0,627	0,626	0,538	0,579	0,700	0,724	0,837	0,805
deltaRmcv	0,131	0,163	0,249	0,146	0,121	0,125	0,021	0,077
R^2_{Ext}	0,829	0,759	0,709	0,814	0,870	0,881	0,957	0,917
F	83,523	115,529	105,355	26,392	92,936	71,994	96,280	76,259
nLV	2	1	1	5	2	3	5	4

Fonte: elaborado pelo autor.

APÊNDICE D – Blocos de código completos

Bloco de Código D.1 – Arquivo aspirina.mol2, gerado pelo Open Babel. Ver [seção 3.1](#).

```

1 @<TRIPOS>MOLECULE
2 *****
3 13 13 0 0 0
4 SMALL
5 GASTEIGER
6
7 @<TRIPOS>ATOM
8      1 C      -2.5981   -1.5000   0.0000 C.3      1 UNL1      0.1216
9      2 C      -1.7321   -2.0000   0.0000 C.2      1 UNL1      0.3220
10     3 O      -1.7321   -3.0000   0.0000 O.2      1 UNL1     -0.2497
11     4 O      -0.8660   -1.5000   0.0000 O.3      1 UNL1     -0.4239
12     5 C      -0.8660   -0.5000   0.0000 C.ar     1 UNL1      0.1547
13     6 C      -1.7321   -0.0000   0.0000 C.ar     1 UNL1      0.0378
14     7 C      -1.7321    1.0000   0.0000 C.ar     1 UNL1      0.0031
15     8 C      -0.8660    1.5000   0.0000 C.ar     1 UNL1      0.0004
16     9 C      -0.0000    1.0000   0.0000 C.ar     1 UNL1      0.0097
17    10 C       0.0000   -0.0000   0.0000 C.ar     1 UNL1      0.1143
18    11 C       0.8660   -0.5000   0.0000 C.2      1 UNL1      0.3907
19    12 O       0.8660   -1.5000   0.0000 O.co2     1 UNL1     -0.2404
20    13 O       1.7321    0.0000   0.0000 O.co2     1 UNL1     -0.2404
21 @<TRIPOS>BOND
22      1      1      2      1
23      2      2      3      2
24      3      2      4      1
25      4      4      5      1
26      5      5      6      ar
27      6      6      7      ar
28      7      7      8      ar
29      8      8      9      ar
30      9      9     10      ar
31     10     5     10      ar
32     11     10     11      1
33     12     11     12      ar
34     13     11     13      ar

```

Fonte: elaborado pelo autor.