UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ENGENHARIA & INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

Ruan Medina Carvalho

Investigação sobre a capacidade de predição de afinidade de ligação entre moléculas em sistemas hospedeiro-hóspede por meio de métodos de aprendizado de máquina.

Juiz de Fora
2021

Ruan Medina Carvalho

Investigação sobre a capacidade de predição de afinidade de ligação entre moléculas em sistemas hospedeiro-hóspede por meio de métodos de aprendizado de máquina.

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Faculdade de Engenharia & Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional. Área de concentração: Modelagem Computacional.

Orientador: Prof. D.Sc. Leonardo Goliatt da Fonseca

Coorientadora: Prof[a]. D.Sc. Priscila Vanessa Zabala Capriles Goliatt

Juiz de Fora

2021

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Carvalho, Ruan Medina.
Investigação sobre a capacidade de predição de afinidade de ligação entre moléculas em sistemas hospedeiro-hóspede por meio de métodos de aprendizado de máquina. / Ruan Medina Carvalho. – 2021.
89 f. : il.

Orientador: Leonardo Goliatt da Fonseca
Coorientadora: Priscila Vanessa Zabala Capriles Goliatt
Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Faculdade de Engenharia & Instituto de Ciências Exatas . Programa de Pós-Graduação em Modelagem Computacional, 2021.

1. Afinidade Molecular. 2. Aprendizado de Máquina. 3. Ciclodextrina. I. Fonseca, Leonardo Goliatt da, orient. II. Goliatt, Priscila Vanessa Zabala Capriles, coorient. III. Título.

Ruan Medina Carvalho

Investigação sobre a capacidade de predição de afinidade de ligação entre moléculas em sistemas hospedeiro-hóspede por meio de métodos de aprendizado de máquina

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional. Área de concentração: Modelagem Computacional

Aprovada em 14 de dezembro de 2021.

BANCA EXAMINADORA

Prof(a). Dr(a). Leonardo Goliatt da Fonseca - Orientador
Universidade Federal de Juiz de Fora

Prof(a). Dr(a). Priscila Vanessa Zabala Capriles Goliatt - Coorientadora
Universidade Federal de Juiz de Fora

Prof(a). Dr(a). Carlos Cristiano Hasenclever Borges
Universidade Federal de Juiz de Fora

Prof(a). Dr(a). Isabella Alvim Guedes
Laboratório Nacional de Computação Científica

Prof(a). Dr(a). Camila Martins Saporetti

Universidade do Estado de Minas Gerais

Juiz de Fora, 16/02/2022.

Documento assinado eletronicamente por **Leonardo Goliatt da Fonseca**, **Professor(a)**, em 09/03/2022, às 17:25, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Camila Martins Saporetti**, **Usuário Externo**, em 10/03/2022, às 10:25, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Isabella Alvim Guedes**, **Usuário Externo**, em 11/03/2022, às 23:09, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Carlos Cristiano Hasenclever Borges**, **Professor(a)**, em 12/03/2022, às 11:13, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Priscila Vanessa Zabala Capriles Goliatt**, **Professor(a)**, em 14/03/2022, às 13:34, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **0682774** e o código CRC **56C01F03**.

Dedico este trabalho a minha família e amigos, pelo apoio incondicional e intercessão nas necessidades.

AGRADECIMENTOS

Gratidão é o que me permite continuar caminhando, com a certeza de que com a devida companhia conseguimos ir mais longe. Obrigado a todos que diretamente ou indiretamente contribuíram para minha formação e para o desenvolvimento desse trabalho.

Agradeço a minha família pelo apoio em todo o meu processo de formação. Em especial agradeço a minha mãe, Maria do Rosário Prado Medina, por todo apoio, intercessão e proximidade em todos esses anos. Obrigado pela coragem de me deixar trilhar um caminho tão diferente da sua realidade. Tudo que eu atingir também é seu.

Agradeço a todo o corpo docente e técnico da Universidade Federal de Juiz de Fora que acredita e que luta pela importância de um ensino transformador: público e de qualidade. Agradeço os membros do Programa de Pós-Graduação em Modelagem Computacional pelos ensinamentos. Em especial, agradeço meus orientadores de longa data Leonardo Goliatt da Fonseca e Priscila Vanessa Zabala Capriles Goliat, pelos anos de paciência e companheirismo.

Agradeço a todos os meus amigos por serem exatamente como eles são. Em especial, agradeço aos membros do Grupo de Educação Tutorial da Engenharia Computacional pela família que fomos durante nosso período de formação! Levo todos os momentos para a vida.

Agradeço à Universidade Federal de Juiz de Fora pela infraestrutura que possibilitou minha formação e pelas politicas sociais que garantiram minha permanência por grande parte da minha formação. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Agradeço e testemunho pela importância do financiamento público da pesquisa no Brasil.

Peguei as malas e caminhei pela porta.
No portão da varanda me quebrei...
Frente ao escuro vi olhos que acendiam,
Se falassem, diriam ‘fica’!
Mas pra quem sente,
A fala é superficial.

RESUMO

A inserção de experimentações *in silico* no contexto científico nas últimas décadas permitiram a consolidação de áreas interdisciplinares como a bioinformática, biologia computacional, química computacional entre outras que buscam descrever, entender e prever eventos naturais por meio de equações matemáticas e métodos computacionais. Neste contexto, é comum que pesquisadores tenham interesse em prever medidas de interação entre moléculas, principalmente para viabilizar o estudo racional de fármacos. Realizar triagens de potenciais fármacos de forma computacional visa reduzir o tempo na descoberta de novas drogas, assim como reduzir o elevado número de testes em laboratório que encarece todo o processo. As triagens moleculares computacionais geralmente são realizadas por meio de processos chamados de *docking*, nos quais define-se graus de liberdade para representações moleculares no interior de uma *grid* de simulação. O objetivo do processo é evoluir uma otimização nesse espaço que visa encontrar a configuração geométrica de uma possível ligação entre as moléculas e calcular métricas relativas a esse estado de interação. Para isso, a literatura já apresenta diversas propostas para a formulação de funções objetivo para a busca, ora baseados em modelos matemáticos sob a ótica da física clássica, ora em modelos com base na teoria quântica. Mais recentemente, como alternativa, vêm sendo propostos modelos preditivos baseados em dados e ajustados por métodos computacionais de aprendizado de máquina. Alguns desses métodos vêm apresentando resultados superiores aos dos modelos físicos, além de possuírem tempo de predição inferiores, uma vez já treinados. Visto isso, as técnicas de aprendizado de máquina (ML, do inglês *Machine Learning*) estão se tornando parte integrante do desenho e descoberta racionais de fármacos e o estudo de uma série de moléculas. Nesse contexto, as Ciclodextrinas (CDs) são nano-gaiolas (*nanohorns*) usadas para melhorar a entrega de drogas insolúveis ou tóxicas para o organismo. Devido à semelhança química entre CDs e proteínas, abordagens ML podem beneficiar vastamente os estudos da área, identificando carreadores promisores para uma dada molécula de interesse. No presente trabalho, são avaliados o desempenho de três métodos de ML bem conhecidos na literatura - *Support Vector Regression* (ε-SVR), *Gaussian Process Regression* (GPR) e *eXtreme Gradient Boosting* (XGB) - para prever a afinidade de ligação da ciclodextrina e ligantes de interesse em um sistema hospedeiro-ligante (*host-guest*). Os hiperparâmetros dos métodos ML propostos foram ajutados em uma estratégia de busca randomizada (Random Search). Os resultados mostram a consistencia da metodologia utilizada por apresentar resultados médios de erro controlados. O melhor desempenho na predição foi obtido por um modelo GPR otmizado em busca randomizada, se ajustando bem aos dados ($R^2 = 0,803$) com baixos erros de predição ($RMSE = 1,811 kJ/mol$ e $MAE = 1,201 kJ/mol$).

Palavras-chave: Afinidade Molecular. Aprendizado de Máquina. Ciclodextrina.

ABSTRACT

The insertion of *in silico* experiments in the scientific context in recent decades has allowed the consolidation of interdisciplinary areas such as bioinformatics, computational biology, computational chemistry, among others, which seek to describe, understand and predict natural events through mathematical equations and computational methods. In this context, it is frequent that researchers are interested in predicting interaction measures between molecules, mainly to enable the rational study of drugs. Performing screenings of potential drugs computationally aims to reduce time to discover new drugs and reduce the high number of laboratory tests that make the whole process more expensive. Researchers usually perform computational molecular screenings through docking techniques, which define degrees of freedom for molecular representations within a simulation grid. The goal of the process is to evolve an optimization in this space that aims to find the geometric configuration of a possible bond between molecules and calculate metrics relating to this interaction state. To this end, the literature already presents several proposals for the formulation of objective functions for the search, sometimes based on mathematical models from the perspective of classical physics, sometimes based on models based on quantum theory. Recently, as an alternative, predictive models based on data and adjusted by computational machine learning methods have been proposed. Surprisingly, some of these methods have shown better results than the physical models, with lower prediction time once trained. Therefore, Machine Learning (ML) techniques are an integral part of rational drug design and discovery. Cyclodextrins (CDs) are nano-cages (nanohorns) used to enhance the delivery of insoluble or toxic drugs to the body. Due to the chemical similarity between CDs and proteins, ML approaches can vastly benefit studies in the field by identifying promising carriers for a given molecule of interest. In the present work, the performance of three well-known ML methods in the literature - Support Vector Regression (ε-SVR), Gaussian Process Regression (GPR), and eXtreme Gradient Boosting (XGB) - are evaluated to predict the binding affinity of cyclodextrin and ligands of interest in a host-ligand system (host-guest). We have tuned the hyperparameters of the proposed ML methods in a Random Search strategy. The results show the consistency of the methodology used by presenting controlled average error results. The best prediction performance was obtained by a GPR model optimized in random search, fitting the data well ($R^2 = 0.803$) with low prediction errors ($RMSE = 1.811 kJ/mol$ and $MAE = 1.201 kJ/mol$).

Keywords: Molecular Affinity. Machine Learning. Cyclodextrin.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CB7	Cucurbit[7]urila
CD	Ciclodextrinas
DB	Base de Dados (*Data Base*)
FC	Carga Formal do Ligante (*Ligand Formal Charge*)
GBSA	*Generalized Born Surface Area*
GPR	*Gaussian Process Regression*
HB	Ligação de Hidrogênio (Hydrogen Bond)
HG	Hospedeiro-Hóspede (*Host-Guest*)
MAPE	*Mean Absolute Percentage Error*
ML	Aprendizado de Máquina (*Machine Learning*)
MLR	*Multiple Linear Regression*
MM	*Molecular Mechanics*
PBSA	*Poisson-Boltzmann Surface Area*
PCA	Análise de Componentes Principais (*Principal Components Analysis*)
PPGMC	Programa de Pós-Graduação em Modelagem Computacional
RB	Ligações Rotáveis do Ligante (*Ligand Ratatable Bonds*)
RMSE	*Root-Mean-Square Deviation*
RS	Busca Randomizada (*Randomized Search*)
SAMPL	*Statistical Assessment of Modeling of Proteins and Ligands*
SVR	*Support Vector Regression*
UFJF	Universidade Federal de Juiz de Fora
XGB	*eXtreme Gradient Boosting*
ΔG	Afinidade de Ligação Molecular Medida em kJ/mol

SUMÁRIO

1 INTRODUÇÃO

A inserção de experimentações *in silico* no contexto científico nas últimas décadas permitiram a consolidação de áreas interdisciplinares como a bioinformática, biologia computacional, química computacional entre outras que buscam descrever, entender e prever eventos naturais por meio de equações matemáticas e métodos computacionais (2). Neste contexto, é comum que pesquisadores tenham interesse em prever medidas de interação entre moléculas, principalmente para viabilizar o estudo racional de fármacos (2). Realizar triagens de potenciais fármacos de forma computacional visa reduzir o tempo na descoberta de novas drogas, assim como reduzir o elevado número de testes em laboratório que encarecem todo o processo (3, 2).

As triagens moleculares computacionais são realizadas por meio de processos chamados de *docking*, nos quais define-se graus de liberdade para representações moleculares no interior de uma grade de simulação (*grid*) (4). O objetivo do processo é evoluir uma otimização nesse espaço que visa encontrar a configuração geométrica de uma possível ligação entre moléculas de interesse e calcular métricas relativas a esse estado de interação (4, 5). Para isso, a literatura já apresenta diversas propostas para a formulação de funções objetivo para a busca, ora baseadas em modelos matemáticos sob a ótica da física clássica, ora em modelos com base na teoria quântica (6, 7, 8, 5). Tais funções objetivo, apesar de serem as mais aplicadas na literatura, são altamente dependentes das escolhas dos termos matemáticos para a modelagem da dinâmica do campo de força que rege as interações na simulação, assim como dos parâmetros experimentais para a representação das ligações entre os átomos das moléculas de interesse. Quando a dependência desses valores são reduzidos, por exemplo na aplicação de métodos *ab initio* baseados em princípios quânticos, um alto custo computacional é demandado para a obtenção dos resultados (6, 7, 8, 5).

Mais recentemente, como alternativa, pesquisadores vêm propondo modelos preditivos baseados em dados, ajustados por métodos computacionais de aprendizado de máquinas (9, 10, 5). Alguns desses métodos integram técnicas de aprendizado de máquina às metodologias tradicionais, conseguindo assim apresentar resultados superiores aos dos modelos físicos tradicionais (7). Métodos de aprendizado de máquina podem ser bons aliados na tarefa desafiadora de calcular predições para afinidade de ligação molecular (10). Contudo, é importante ressaltar que o sucesso na aplicação de tais funções é relativo ao domínio de moléculas nas quais os testes estão sendo realizados, demandando certos níveis de homogeneidade com a natureza dos dados utilizados nos processos de treinamento dos modelos (9). Apesar dos métodos baseados em aprendizado de máquina ainda não substituírem os métodos físicos tradicionais, a aplicação dessas abordagem pode gerar novos *insights* no entendimento da natureza de certas interações moleculares, inclusive possibilitando o desenvolvimento de metodologias conjuntas que unem modelagem baseada em física e em dados para o entendimento de sistemas moleculares com complexidade

elevada (5).

Na aplicação de métodos *in silico* para a triagem de potenciais fármacos, geralmente seu conjunto de moléculas passa por um processo de *docking* ou dinâmica molecular para calcular posições geométricas de acoplamento na formação do complexo molecular. Com essas conformações encontradas são aplicadas funções de avaliação para o cálculo de métricas de interesse, como por exemplo a afinidade de ligação entre as moléculas do complexo e assim definir uma conformação preferencial (4, 5). Esse processo pode ser visualizado de forma esquemática na Figura 1. Todo esse processo tradicional é computacionalmente custoso e deve ser realizado para cada molécula ligante de interesse nas pesquisas. Com isso, propõe-se investigar a possibilidade de aproximação para os valores de funções de avaliação a partir de propriedades individuais de cada molécula, de forma menos custosa, sem a necessidade de executar grandes simulações da dinâmica física dos complexos. Esse processo pode ser aplicado como uma etapa de pré-processamento nas pesquisa com o objetivo de selecionar ligantes mais promissores e propor uma ordem para os testes mais computacionalmente custosos.

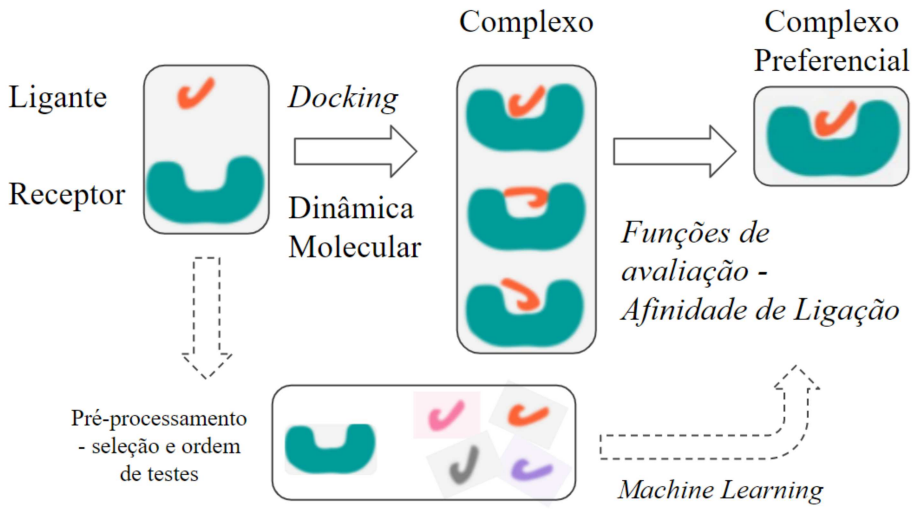


Figura 1 – Esquema para o processo de obtenção de métricas de interação molecular a partir de um conjunto de moléculas de interesse.

Fonte: Produzido pelo autor (2021).

Nesse contexto, este trabalho propõe uma investigação da utilização de diferentes naturezas de métodos computacionais para a predição de medidas de interação entre pares moleculares, conforme esquematizado na Figura 1. Para isso, definiu-se um escopo de trabalho que se preocupa igualmente com os dados a serem utilizados, com a representação molecular utilizada como entrada para os modelos de aprendizado, e a definição e ajuste dos métodos computacionais aplicados (8).

Como esse trabalho se trata do início de uma investigação, preferiu-se trabalhar com

sistemas moleculares simplificados de hospedeiro-hóspede (do inglês, *host-guest*), nos quais duas moléculas interagem entre si formando um complexo de inclusão (11). Nos sistemas simplificados tratados nesse trabalho, prevalecem características de hospedeiros simétricos e hóspedes com complexidade geométrica reduzida disponibilizados após um processo de curadoria na comunidade do *BindingDB* (11). Em um primeiro momento, por apresentar uma maior quantidade de artigos de referência disponíveis, escolheu-se trabalhar com sistemas que possuem como hospedeiros moléculas das classes α, β e γ de ciclodextrinas (12), além de moléculas ligantes conhecidas na literatura e listadas pela comunidade do *BindingDB* (*Acesse a lista de trabalhos aqui*). Apesar disso, ainda existe espaço na literatura para o entendimento de seus comportamentos em complexos moleculares (13). Além disso, os dados dessa classe de moléculas foram os mais recentes disponibilizados pela equipe do banco de dados *BindingDB* na categoria de sistemas hospedeiro-hóspede, o que justifica o empenho na utilização, expansão, difusão e a realização de experimentos com esse conjunto de dados.

A representação molecular escolhida para esse trabalho foi composta por descritores individuais calculados para cada molécula hospedeira ou hóspede, como informações topológicas e atributos físico-químicos (14). Buscou-se investigar se por meio de uma representação simplificada seria possível criar mecanismos computacionais que permitissem prever métricas de interação molecular entre duas moléculas analisando somente atributos que podem ser calculados por ferramentas de bioinformática e química computacional já difundidos na área (15, 14). Com mecanismos computacionais dessa natureza é possível, por exemplo, atuar em processos preliminares de triagem computacional clássica, direcionando os testes a serem executados, reduzindo cálculos e investigando de forma mais assertiva potenciais ligantes de interesse.

Por fim, a escolha dos métodos de aprendizado de máquina utilizados nesse trabalho seguiu de forma a proporcionar testes e comparações entre 3 métodos de diferentes classes de regressores para aprendizado supervisionado (16, 17, 18, 19, 8): *Gaussian Process Regressor*(GPR); *ε-Support Vector Regression*(ε-SVR); e *eXtreme Gradient Boosting* (XGB). Os algoritmos testados são treinados em um processo de validação cruzada e terão seus hiperparâmetros ajustados pela estratégia de Busca Randomizada (RS, do inglês *Randomized Search*) (20).

1.1 SISTEMAS HOSPEDEIRO-HÓSPEDE

Sistemas hospedeiro-hóspede compõem um ramo da química supramolecular que descreve a formação de complexos estruturais entre duas ou mais moléculas ou íons. No caso mais comum, uma molécula hospedeira (geralmente a com o maior número de átomos da dupla molecular) forma um composto químico com uma molécula ou íon hóspede (21). Os dois componentes do sistema são mantidos juntos por forças não covalentes, como por

meio de um conjunto de ligações de hidrogênio, dispostos em soluções aquosas (21).

Nos estudos da formação desses compostos, geralmente se observa um equilíbrio entre o estado não vinculado, no qual o hospedeiro e o hóspede estão separados um do outro, e o estado vinculado, no qual há um complexo estruturalmente definido, representado por

$$H + G \rightleftharpoons HG \tag{1.1}$$

onde H é o hospedeiro (*host*), G é o hóspede (*guest*), e HG é o complexo formado entre a interação hospedeiro-hóspede (22).

Em sistemas biológicos, os termos análogos de hospedeiro e hóspede são comumente referidos, respectivamente, como receptor (proteína ou enzima) e ligante (substrato ou inibidor) (22). Dessa forma, estudos de sistemas hospedeiro-hóspede auxiliam no entendimento da termodinâmica de sistemas mais complexos, pois da mesma forma que ocorrem com os sistemas proteicos, os complexos hospedeiro-hóspede ocorrem por proporcionarem ao sistema uma menor energia livre de Gibbs total (23).

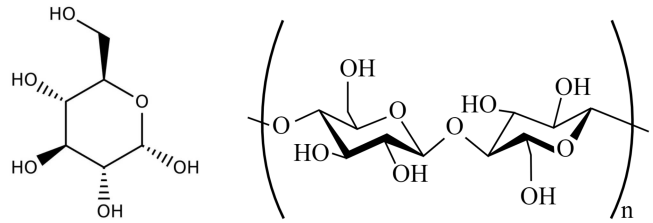
Para isso, sistemas simplificados de hospedeiro-hóspede (ou *Toy Systems*) vêm sendo utilizados para entender a base da natureza dessas interações não covalentes (24, 1). Nesses sistemas, prevalecem características de hospedeiros simétricos e com propriedades físicas bem conhecidas, assim como hóspedes com complexidade geométrica reduzida. Ao estudar esses sistemas, espera-se obter mais informações sobre o resultado combinatório de várias forças de menor magnitude e não covalentes que são usadas para gerar um efeito geral sobre a estrutura supramolecular estudada (5, 24).

Historicamente, tais estudos são incentivados na comunidade científica, como por exemplo da *Drug Design Data Resource Community*, financiada em parte pelos Institutos Nacionais de Ciências (NIH) dos Estados Unidos. Anualmente a comunidade oferece competições na área de desenvolvimento racional de fármacos, como a SAMPL (*Statistical Assessment of Modeling of Proteins and Ligands*) e o D3R *Grand Challenge* com devida publicação dos resultados obtidos nas competições (25), o que justifica os esforços de pesquisa na área e demostra a importância desses estudos no entendimento da formação de complexos moleculares.

1.2 CICLODEXTRINAS

As ciclodextrinas (CDs) são carboidratos (oligossacarídeos) compostos por unidades de glicose (α-D-glicopiranose) unidas em ligação covalente como mostrado na Figura 2.

O processo de síntese das CDs forma uma estrutura no formato de um tronco de cone com tamanhos diversos. Por exemplo, é comum observar a ciclização de seis (α-CD), sete(β-CD) ou oito (γ-CD) unidades de glicose, causando a diferenciação entre as classes consideradas nesse trabalho (13, 12, 27). A Figura 3 mostra as estruturas químicas dessa família de moléculas.



(a) α-D-glicopiranose (b) Ligações entre α-D-glicopiranose

Figura 2 – Representação 2D de moléculas de glicose (α-D-glicopiranose) e esquema da ligação covalente entre duplas dessas moléculas.

Fonte: Adaptado de (26).

A Figura 3 revela radicais de hidroxilas nas pontas das moléculas, o que torna as ciclodextrinas solúveis em água, devido sua facilidade na criação de ligações de hidrogênio com o solvente (28). Entender a solubilidade dessas moléculas é de suma importância para seu uso no desenvolvimento racional de fármacos (29, 30, 28). Isso corrobora com o fato de todos os experimentos considerados na atual pesquisa terem sido realizados em meio aquoso. Contudo, o interior da cavidade dessas moléculas possuem perfil hidrofóbico, devido ao alinhamento dos hidrogênios C(3)-H e C(5)-H e pelo oxigênio da união do éter C(1)-O-C(4). Tal característica pode facilitar a proximidade e interação do interior das moléculas de ciclodextrinas com os ligantes considerados em experimentos de bancada ou simulações computacionais de dinâmica molecular (30, 31, 28).

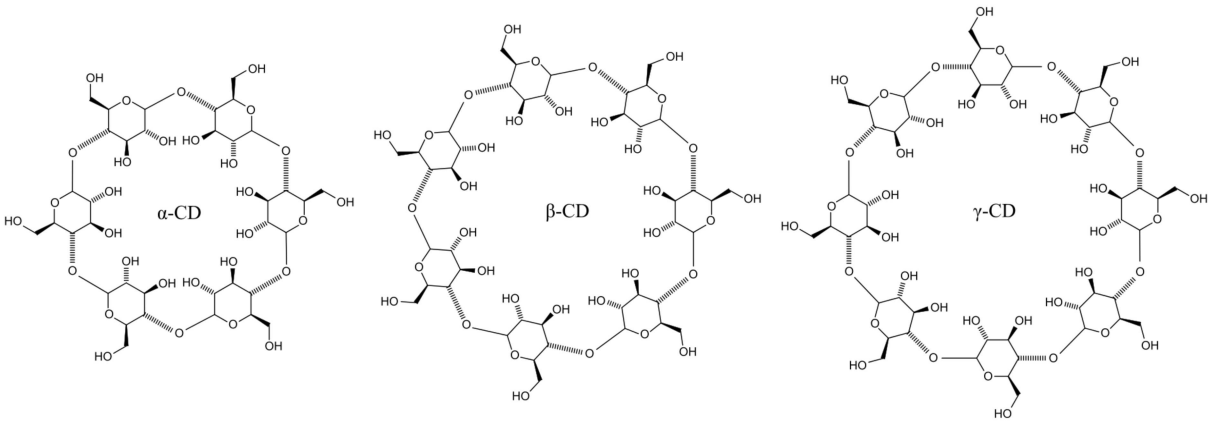


Figura 3 – Representação 2D da estrutura química das classes α, β e γ de ciclodextrinas.

Fonte: Adaptado de (32).

A aplicabilidade de moléculas de ciclodextrinas são diversas no campo de medicamentos, alimentos e agricultura. Sua propriedade solúvel permite a criação de complexos por meio da interação com diferentes moléculas hóspedes em seu interior apolar (33, 31, 28). A Figura 4 apresenta a disposição espacial de uma das moléculas tratadas nesse trabalho

mostrando sua característica estrutural de recipiente, o que possibilita sua aplicação como molécula carreadora. Na área farmacêutica, por exemplo, esse tipo de complexo de inclusão vem sendo utilizados, principalmente, como veículos que potencializam a farmacocinética de medicamentos no corpo ao incorporar medicamentos apolares, e por vezes citotóxicos, no interior do tronco do cone durante o processo de entrega do fármaco (ou *drug delivery*) (13, 33).

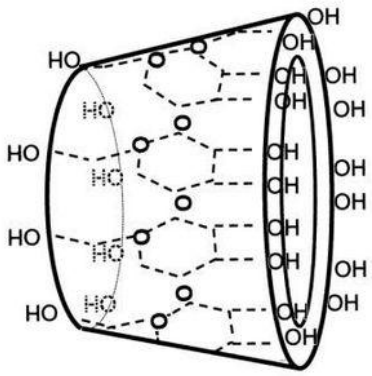


Figura 4 – Estrutura do arranjo espacial em toroide da γ-CD.

Fonte: Adaptado de (34).

A Figura 5 mostra esquematicamente como os complexos de inclusão ocorrem geometricamente entre a molécula hospedeira (ciclodextrinas) e a molécula hóspede (ou ligante). Tal característica coloca as ciclodextrinas como moléculas importantes no processo de desenvolvimento de tecnologias de liberação controlada de fármacos, assim como para o desenvolvimento de estruturas mais complexas de (*nanohorns*), nanotubos e nanoesferas (33, 28).

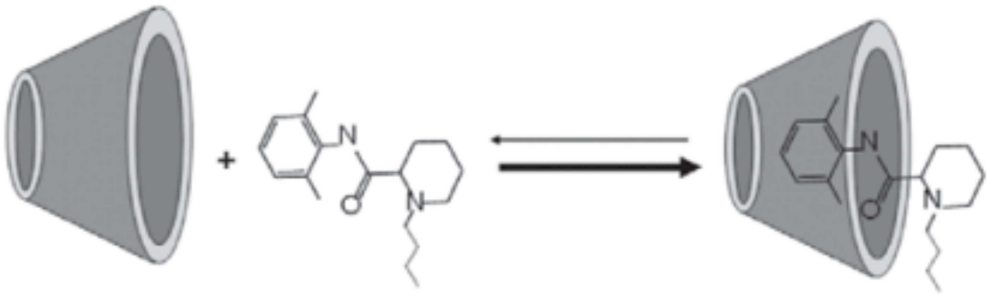


Figura 5 – Esquema de formação de complexo de inclusão em sistemas hospedeiro-hóspede com moléculas com estrutura de *nanohorns*, como é o exemplo das ciclodextrinas.

Fonte: Adaptado de (35).

1.3 OBJETIVOS

1.3.1 **Objetivo Geral**

Avaliar a performance de métodos de aprendizado de máquina para predição de afinidade de ligação molecular em sistemas hospedeiro-hóspede simplificados compostos por ciclodextrinas e ligantes disponíveis na literatura.

1.3.2 **Objetivos Específicos**

Como objetivos específicos necessários para atingir o objetivo geral, destacam-se:

- Apresentar um *pipeline* para o cálculo de descritores topológicos e físico-químicos para moléculas por meio de ferramentas computacionais já existentes a fim de definir uma representação molecular adequada.
- Compilar e disponibilizar uma base de dados para o problema hospedeiro-hóspede composta por descritores de moléculas de diferentes classes de ciclodextrinas, descritores de ligantes disponíveis na literatura, descritores para a solução de experimentação (pH e temperatura) e dados experimentais curados de afinidade de ligação molecular.
- Avaliar a qualidade das predições para as diferentes classes de dados considerados na base de dados compilada.
- Interpretar métricas de erro dos modelos computacionais no domínio de aplicação do problema.
- Comparar a performance dos métodos de aprendizado de máquina *Gaussian Process Regressor*(GPR), *Support Vector Regression*(ϵ-SVR) e *eXtreme Gradient Boosting* (XGB) para predição de afinidade de ligação molecular dos sistemas hospedeiro-hóspede considerados.

2 REVISÃO BIBLIOGRÁFICA

O presente capítulo destina-se a apresentar trabalhos disponíveis na literatura que contribuem para o entendimento do contexto da presente pesquisa, assim como referências que sustentam a justificativa de avanços em estudos sobre a predição de métricas de interação molecular de forma computacional.

Primeiramente apresenta-se um trabalho que exemplifica o desenvolvimento de métodos para a predição de métricas de interação molecular baseados apenas em modelagem física. Tais estudos são os mais clássicos na literatura, porém abrem mão da capacidade explicativa que métodos estatísticos podem apresentar em maiores conjuntos de dados. Apesar disso, para pequenos conjuntos de moléculas, quando não existe dados curados que possam servir em uma etapa anterior de treinamento de métodos computacionais, os métodos físicos ainda conseguem ser aplicados por meio da generalização de parâmetros estruturais e dinâmicos de campos de forças de moléculas simulares.

Dimas *et al.* (36) no trabalho intitulado "*Affinity calculations of cyclodextrin host–guest complexes*" aplica uma série de variações de metodologias de energia livre de ponto final, que buscam estimar diretamente as energias livres a partir de simulações sem restrições que amostram o espaço de fase de estados individuais. São exemplos de métodos aplicados o MM/PBSA (*Molecular Mechanics-Poisson-Boltzmann Surface Area*) e MM/GBSA (*Molecular Mechanics-Generalized Born Surface Area*) que buscam estimar a energia livre de uma biomolécula, combinando sua energia mecânica com a energia livre de solvatação e termos de entropia. Para cada complexo uma dinâmica molecular convencional (baseada em física newtoniana) de $1\mu s$ foi executada, seguida pela aplicação de métodos semiempíricos baseados em física quântica para os cálculos das métricas de entropia necessárias para a definição das energias de interação dos complexos simulados. A equipe de pesquisa selecionou um conjunto de complexos entre β-CD e 57 pequenas moléculas orgânicas que foram previamente estudadas com o método de análise de distribuição de energia de ligação em combinação com um modelo de solvente implícito (37). Segundo os autores, os complexos relativamente pequenos de hospedeiro-hóspede com ciclodextrinas, para os quais dados calorimétricos de alta qualidade estão geralmente disponíveis, estão se tornando modelos de referência para testar a precisão dos métodos de energia livre. Os níveis de erros obtidos foram $R^2 = 0,66$ e RMSE $= 9,330$ kJ/mol na melhor execução obtida pelo grupo.

Com relação aos trabalhos de predição de energia em sistemas hospedeiro-hóspedes simplificados, as ciclodextrinas não são as únicas moléculas hospedeiras utilizadas. A seguir, apresenta-se um exemplo de trabalho contendo o estudos de sistemas hospedeiro-hóspede focado em moléculas cucurbit[7]urila (CB7). Isso demostra que diferentes famílias de moléculas podem compor sistemas hospedeiro-hóspede, cada uma podendo demandar

conjuntos de descritores diferentes para a modelagem dos compostos, além de mostrar que estudos com diversas famílias são demandadas pela comunidade científica.

Xu *et al.* (38) no trabalho intitulado "*Computation of host–guest binding free energies with a new quantum mechanics based mining minima algorithm*" apresenta estudos sobre a interação de molécula hospedeira cucurbit[7]urila (CB7) com uma série de 15 ligantes. As estruturas das moléculas CB7 são apresentadas na Figura 6. Percebe-se que, diferente das ciclodextrinas, as moléculas CB7 se assemelham a anéis, ao invés de cones. As moléculas e os valores experimentais das energias de interação foram disponibilizadas pelo desafio SAMPL4 (39). Os ligantes considerados possuem diferentes valores de carga formal, variando positivamente de +1 a +3. O trabalho se preocupa em analisar os resultados para as instâncias contendo ligantes com as diferentes cargas formais separadamente, a fim de verificar a aplicação dos métodos testados para as diferentes classes de complexos estudados com relação à carga dos ligantes. Os melhores resultados estão concentrados nas instâncias com ligantes com carga formal +1. O trabalho aplica um método novo chamado QM-VM2 que combina conceitos de mecânica estatística e mecânica quântica para a modelagem dos potenciais energéticos a fim de calcular a energia de ligação não-covalente dos sistemas de hospedeiro-hóspede.

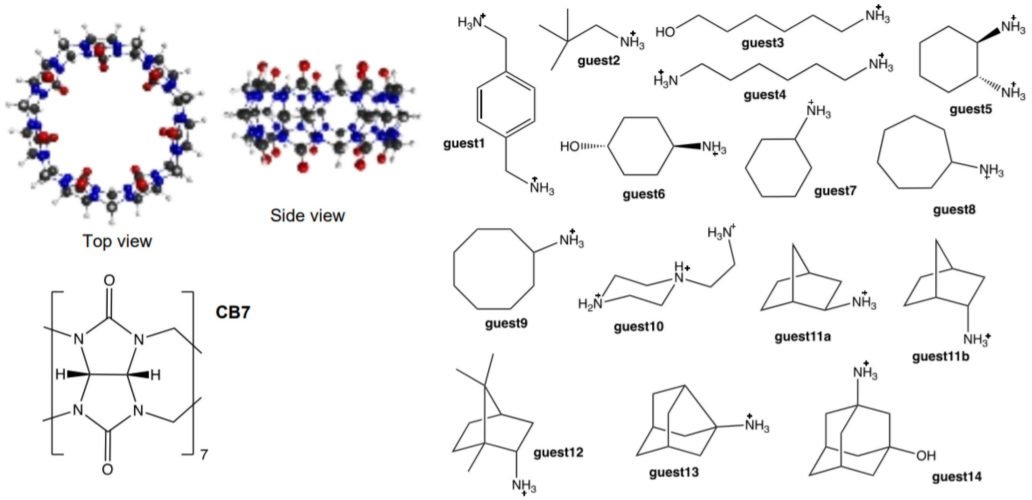


Figura 6 – Estrutura da molécula hospedeira cucurbit[7]urila (CB7) e das moléculas hóspedes que compõem os sistemas hospedeiro-hóspede do estudo de Xu *et al.*

Fonte: Disponível em (38).

A seguir, são apresentados uma sequencia de trabalhos focados no estudo de sistemas hospedeiro-hóspede compostos por moléculas hospedeiras da família das ciclodextrinas que aplicam metodologias estatísticas e computacionais como a base para realizar as predições de métricas de interação dos compostos. Estão listados alguns trabalhos entre 2011 e 2020 que estão bem alinhados com a metodologia seguida no desenvolvimento da pesquisa

apresentada nesse relatório.

Mano *et al.* (40) no trabalho intitulado "*Development of machine learning models of β-cyclodextrin and sulfobutylether-β-cyclodextrin complexation free energies*" apresenta um conjunto de 220 instâncias (142 novas instâncias apresentadas no trabalho somadas a 78 obtidas da literatura) com hospedeiros sulfo-butil-éter-β-CD e diversas moléculas orgânicas hóspedes aplicadas em dois métodos de regressão de aprendizagem da máquina para a predição de energias de interação molecular - *Cubist* e *Random Forest*. Modelos similares foram construídos para β-cyclodextrin usando um conjunto de dados de 233 compostos disponível na literatura. Os resultados demonstram que os métodos de regressão por meio de aprendizagem de máquina podem descrever com sucesso a formação de complexos moleculares entre as moléculas orgânicas e β-cyclodextrin ou sulfobutylether-β-cyclodextrin. As métricas de erro obtidos nos conjuntos de testes foram $RMSE = 1.9$ kJ/mol e $RMSE = 2.7$ kJ/mol, respectivamente.

Xu et al. (41) no trabalho intitulado "*Quantitative structure–property relationship study of β-cyclodextrin complexation free energies of organic compounds*" apresenta um estudo preditivo sobre afinidade de ligação em sistemas moleculares compostos por hospedeiros β-CD e um conjunto diverso de hóspedes orgânicos. O conjunto de dados disponível contém 218 compostos, dividido em um conjunto de treinamento de 160 compostos e um conjunto de teste de 58 compostos por meio da aplicação do algoritmo `DUPLEX` combinado com uma abordagem baseada em Análise de Componentes Principais (PCA) sobre uma base inicial de 865 descritores. O trabalho aplica Regressão Linear Múltipla obtendo métricas de erro $R^2 = 0,833$ e $MAPE = 1,911$ kJ/mol, além de uma arquitetura de Rede Neuronal Artificial obtendo métricas de erro $R^2 = 0,957$ e $MAPE = 0,925$. Ambos os métodos são aplicados em um espaço reduzido de 7 descritores obtidos após a redução de dimensionalidade. A aplicação não linear de Redes Neuronais apresentaram bons resultados, contudo o modelo é especializado apenas para aplicação de complexos compostos por moléculas β-CD. Ademais, o trabalho apresenta um estudo do domínio de aplicação dos modelos treinados por meio de análises baseados no Gráfico de Willians.

Solovev *et. al.* (42) no trabalho intitulado "*3D molecular fragment descriptors for structure–property modeling: predicting the free energies for the complexation between antipodal guests and β-cyclodextrins*" relata o uso de novos descritores de fragmentos 3D para modelar parâmetros e propriedades de moléculas estereoisoméricas ou confórmeros. O primeiro são moléculas que possuem fórmulas moleculares e arranjos de átomos idênticos, diferindo um do outro apenas na orientação espacial dos grupos na molécula, e o segundo são o mesmo composto mostrado com rotações diferentes sobre ligações simples. Tais moléculas não são diferenciadas pela maioria dos descritores moleculares, sendo necessário a adição descritores que levam em conta angulações das ligações atômicas. Um exemplo desse tipo de moléculas que podem apresentar confórmeros é o etano, como apresentado na Figura 7. No trabalho de Solovev *et. al*, foram aplicados uma implementação de

Decomposição em Valores Singulares para análise de descritores e uma Regressão Linear Múltipla como método de aprendizagem de máquina. Nas predições de energia livre da complexação de 76 hóspedes quirais com *beta*-ciclodextrina (β-CD) foram obtidas as métricas $R^2 = 0,918$ e $RMSE = 0,89$ kJ/mol e nas predições de 40 hóspedes quirais com 6-amino-6-deoxy-β-ciclodextrina (am-β-CD) foram obtidas as métricas $R^2 = 0,910$ e $RMSE = 0,89$ kJ/mol.

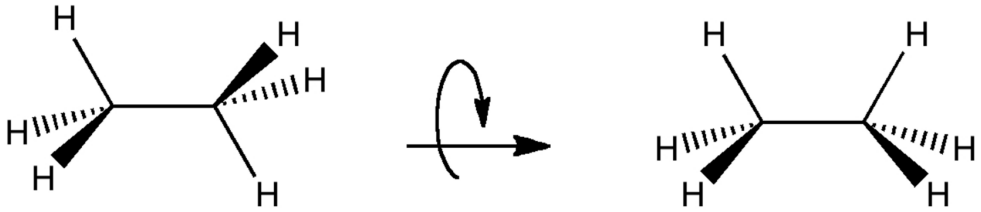


Figura 7 – Exemplo de confórmeros de etano.

Fonte: Disponível em (43).

Zhao *et al.* (1) no trabalho intitulado "*Predicting complexation performance between cyclodextrins and guest molecules by integrated machine learning and molecular modeling techniques*" aplicou métodos ML para prever a afinidade de ligação entre um vasto número de classes de ciclodextrina (8 classes) e diversas moléculas hóspedes (1320 ligantes). A Figura 8 mostra a separação da composição da base considerada no trabalho. Os testes preditivos são realizados utilizando metodologias de Floresta Randômica, Aprendizado Profundo e Métodos de *Boosting*. O conjunto de dados utilizados na pesquisa é o maior encontrado nas pesquisas realizadas durante o desenvolvimento desse trabalho (3000 instâncias hospedeiro-hóspede), contudo o conjunto não foi disponibilizado por questões de propriedade intelectual, segundo informado pelo próprio autor do trabalho. Isso impossibilita a reprodutibilidade dos resultados e a continuidade do desenvolvimento de modelos similares por outros grupos de pesquisa. O melhor resultado obtido para Zhao *et al.* foi baseado na abordagem *Light Gradient Boosting Machine* (LightGBM) com métricas de erro $R^2 = 0,86$, RMSE $= 1,83$ kJ/mol, e MAE $= 1,38$ kJ/mol.

Di *et al.* (44) no trabalho intitulado "*In silico prediction of binding capacity and interaction forces of organic compounds with α- and β-cyclodextrins*" apresenta um estudo com a aplicação de metodologias computacionais de consenso para prever as constantes de ligação e as forças de interação dos complexos de inclusão de ciclodextrinas das classes α-CD e β-CD. O trabalho utiliza uma base formada por 418 complexos de inclusão de ciclodextrina foram coletados de trabalhos da literatura com suas constantes de ligação experimental. A abordagem combina 6 naturezas de atributos para descrição molecular e 5 métodos de aprendizagem de máquina, construímos 60 modelos únicos, que

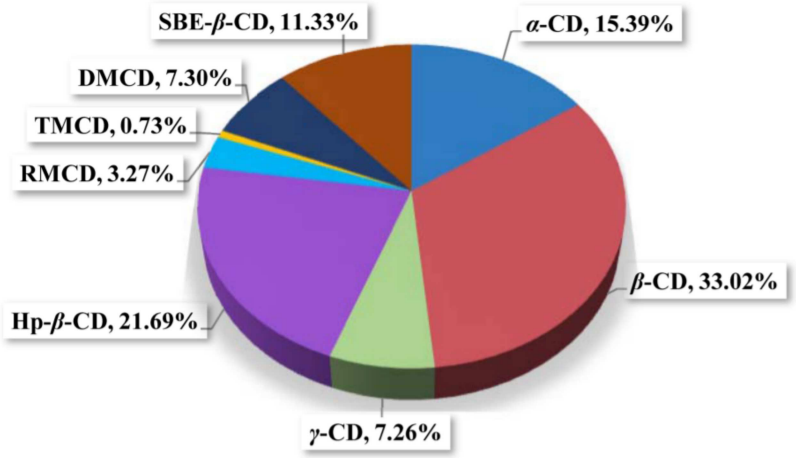


Figura 8 – Distribuições dos tipos moléculas hospedeiras do trabalho de Zhao *et al.*.

Fonte: Disponível em (1).

foram selecionados para desenvolver 104 modelos de consenso. De forma geral, modelos de consenso objetivam a combinação de resultados obtidos da predição de diferentes modelos de aprendizado de máquina para a criação de uma predição combinada média (45). Geralmente, tais abordagens estão associadas a um maior custo computacional. No trabalho de Di *et al.*, as métricas de ajuste entre os valores preditos e os valores experimentais do conjuntos de teste no melhor modelo de consenso para α-CD e no melhor modelo de consenso β-CD foram $R^2 = 0,84$ e $R^2 = 0,89$, respectivamente.

Os resultados de vários dos trabalhos acima podem ser utilizados para comparações com os resultados relativos ao desenvolvimento da presente pesquisa com o intuito de averiguar similaridade nos níveis de erros obtidos. Contudo, as métricas de erros não podem ser diretamente comparadas, uma vez que os conjuntos de dados e as divisões de conjuntos de trenamento e teste não são as mesmas utilizadas para o desenvolvimento das pesquisas. Além disso, muitos dos trabalhos definem modelos específicos para cada classe de molécula de ciclodextrina, não apresentando resultados que podem ser generalizados para a aplicação de ciclodextrinas compostas por diferentes números de glicoses. O trabalho atual se destaca ao propor a utilização de uma base de dados mista de ciclodextrinas que pode ser divulgada para garantir a continuidade da linha de pesquisa, definir o domínio de aplicação dos modelos propostos, além de aplicar métodos de aprendizado de máquina, pouco ou não utilizados em testes anteriores na literatura.

Por fim, são apresentados um conjunto de trabalhos de revisão que apresentam uma visão geral do contexto de aplicação da pesquisa. O intuito é justificar os esforços no entendimento de sistemas hospedeiro-hóspede, em especial de compostos contendo a família das ciclodextrinas como hospedeira, além de vislumbrar possibilidades futuras de continuidade dessa linha de pesquisa.

Zhao *et al.* (24) no trabalho intitulado "*Research Advances in Molecular Modeling in Cyclodextrins*" apresenta uma discussão justificada pela ampla aplicação de ciclodextrina como excipientes farmacêuticos. O trabalho se propõe em apresentar uma análise qualitativa e quantitativa dos resultados da modelagem molecular de CDs. O trabalho conclui que as pesquisas na áreas disponíveis na literatura possuem três características frequentes: (i) as moléculas de CD mais utilizadas: β-cyclodextrin, (2) ferramentas de modelagem mais comuns: cálculo de *docking* e dinâmica molecular, (3) campos de pesquisa em alta: propriedades estruturais, solubilidade, reconhecimento quiral e complexos de inclusão de estado sólido.

Gu *et al.* (13) no trabalho intitulado "Macrocycles as drug-enhancing excipients in pharmaceutical formulations" apresenta um estudo focado em um conjunto de moléculas macrocíclicas, entre elas: n-ciclodextrinas, cucurbit[n]urilas, calix[n]arenos e pilar[n]arenos, onde n representa o número de subunidades em cada homólogo. O trabalho discute por meio de uma revisão da literatura atual o potencial do uso dessas moléculas como excipientes de medicamentos em uma gama de aplicações farmacêuticas. Todos esses macrociclos podem formar complexos hospedeiro-hóspede com ativos farmacêuticos, nas quais a ligação entre as moléculas é estabilizada por efeitos hidrofóbicos dentro da cavidade de cada macrociclo e através de ligações de hidrogênio/ligações íon-dipolo/interações eletrostáticas nessas cavidades. A formação desses complexos de inclusão estão associados a maior solubilidade do medicamento, mudança do sabor na ingestão, liberação controlada e sustentada do medicamento, melhor estabilidade química e física do medicamento, entre outras. O trabalho discute a utilização dessas classes de moléculas, seja para agilizar a tomada de decisão em testes contendo moléculas de ciclodextrinas, já aprovadas como excipientes em medicamentos, ou para potencializar os estudos para uso futuro das demais famílias de moléculas citadas em ensaios clínicos em humanos.

Guedes *et al.* (5) no trabalho intitulado "New machine learning and physics-based scoring functions for drug discovery" apresenta um estudo mais focado no desenvolvimento de processos computacionais na descoberta de fármacos e triagem computacional sobre compostos proteína-ligante e proteína-proteína. As proteínas, por possuírem maior grau de liberdade se comparados com a família de ciclodextrinas, apresentam um desafio ainda maior para os pesquisadores que objetivam realizar uma previsão precisa da afinidade de ligação dos complexos moleculares. Para isso, Guedes *et al.* apresentam um conjunto de novas funções de pontuação empírica que consideram de forma explicita termos baseados em física combinados com técnicas de aprendizado por máquina. As funções de pontuação empírica são baseadas em um conjunto de descritores individuais das moléculas ou descritores da interação calibrados por regressão ou abordagens estatísticas usando um conjunto de dados de afinidade experimentais para complexos de proteínas-ligantes (46). As técnicas de aprendizado de máquina Regressão Linear Múltipla (MLR), Máquina de Vetor Suporte e Métodos de Floresta Randomizadas foram empregadas para obter funções de pontuação

geral e específicas do alvo envolvendo termos otimizados pelo campo de força MMFF94S, termos de solvatação e interações lipofílicas, e um termo modificado pelos autores que contabiliza a contribuição da entropia de torção do ligante na ligação. Mesmo com a utilização de termos físicos ajustados por métodos estatísticos, os autores discutem que a previsão precisa da afinidade proteína-ligante continua sendo um grande desafio e mostra que a literatura apresenta estudos demostrando que o desempenho das funções de pontuação é muito heterogêneo entre diferentes classes-alvo (47, 48). Os elevados graus de liberdade dos sistemas contendo proteínas justificam que os esforços no início de novas linhas de pesquisa busquem a aplicação em sistemas hospedeiro-hóspede mais tradicionais, como é proposto pela dissertação apresentada neste relatório. Contudo, os avanços nas linhas de estudos de sistemas hospedeiro-hóspede devem caminhar para a aplicação de moléculas cada vez mais complexas, acoplando dados propriamente curados, métodos estatísticos-computacionais e modelagem física da estrutura e dinâmica das moléculas do complexo de forma conjunta.

Desse forma, no capítulo a seguir, serão apresentadas as descrições dos materiais e métodos utilizados para o desenvolvimento da presente pesquisa, o que nos possibilita identificar onde o presente trabalho está inserido no contexto proposto.

3 MATERIAL E MÉTODOS

Conforme descrito por (8), o sucesso da previsão de propriedades químicas com o aprendizado de máquinas é dependente da interação de três componentes-chave: Representação Molecular, Algoritmo de Aprendizado de Máquina e dos Dados de entrada. A Figura 9 mostra esquematicamente como esses componentes se relacionam para formar mecanismos computacionais para o estudo de propriedades moleculares.

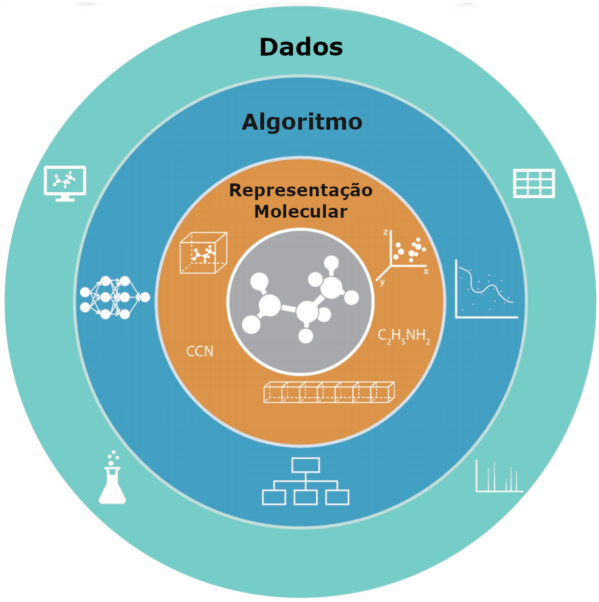


Figura 9 – Componentes-chave no estudo de previsão de propriedades químicas de moléculas por métodos de aprendizado de máquina.

Fonte: Adaptada de (8).

De forma geral, trabalhos no contexto de predições de propriedades químicas por métodos de aprendizado de máquina devem partir do entendimento da representação molecular considerada, ou seja, o que o seu modelo vai entender como uma molécula. Por exemplo, sua representação pode considerar apenas atributos descritivos fisico-químicos e/ou topológicos, reais ou categóricos, dados geométricos em duas dimensões ou três dimensões, ou mesmo um subconjunto dessas possibilidades.

Após isso, a classe de algoritmos considerados é de suma importância e deve ser preparada para receber o modelo de representação molecular considerado, como ilustrado na camada interna da Figura 9. Os métodos de aprendizado de máquina são os responsáveis pelo mapeamento das representações moleculares (*input*) na propriedade química de interesse (*output*). Nesse conjunto de métodos, incluem-se: regressões lineares, redes neuronais, redes de aprendizado profundo, métodos baseados em árvores de decisão, entre outros.

Por fim, a origem e a qualidade dos dados utilizados nas etapas de aprendizado

e validação dos métodos preditivos devem ser consideradas para garantir o sucesso do processo e definir o domínio de aplicação dos métodos treinados. Assim, a qualidade dos dados, sejam eles experimentais ou calculados, deve ser investigada, uma vez que a utilização de dados não tratados podem impedir a obtenção de bons resultados em todas as demais etapas do processo.

Dessa forma, o presente capítulo apresenta primeiramente na Seção 3.1 a descrição da origem dos dados experimentais, o cálculo de descritores físico-químicos, o processo de representação molecular considerado e os tratamentos efetuados nos dados que compõem a base dos experimentos desse trabalho. A seguir, na Seção 3.2, apresenta-se as características dos métodos de aprendizado de máquina usados no estudo, assim como as métricas utilizadas para a avaliação dos resultados.

3.1 DADOS E REPRESENTAÇÃO MOLECULAR

Nesta seção, explicita-se o escopo de trabalho que se preocupa com a origem dos dados a serem utilizados, a representação molecular para o modelo de aprendizado, e a definição e ajuste da base de dados já pré-tratada (8).

3.1.1 Base de Dados Experimentais: *BindingDB*

O `BindingDB` (https://www.bindingdb.org/) é um banco de dados público, acessível via *web*, de medidas de afinidade de ligação não covalente de moléculas em solução (49, 11). A plataforma oferece dados extraídos de diversas publicações validadas pela comunidade científica (49). Todos os dados selecionados passaram por processos de curadoria pela própria comunidade da plataforma.

O foco da plataforma é a compilação de dados referentes a sistemas biomoleculares para interação de proteínas (peptídeos de cadeia longa) com ligantes (moléculas pequenas) candidatos a novos medicamentos. Contudo, recentemente a plataforma também começou a disponibilizar dados sobre sistemas hospedeiro-hóspede (*host-guest*) para pesquisas em química supramolecular.

Cada instância do banco considera duas moléculas, um hospedeiro (moléculas de maior tamanho) e um ligante (moléculas de menor tamanho), as condições do ambiente do experimento realizado, e as métricas de interação molecular extraídas do experimento. Nesse contexto, o banco de dados foi utilizado para obter as informações estruturais das moléculas por meio dos *smiles* de cada uma delas, as condições de pH e Temperatura (°C) da solução do experimento e a afinidade de ligação por meio da medida de ΔG (KJ/mol). Foram considerados apenas os dados relativos aos experimentos com medidas de pH e Temperatura disponíveis e que tenham sido realizados em condições que satisfizessem os seguintes limites:

$$6.9 \leq \text{pH} \leq 7,4$$
$$14.5 \leq \text{Temp} \leq 30,1$$

Em um primeiro momento, escolheu-se trabalhar com instâncias do banco de dados nas quais os hospedeiros dos sistemas pertencessem às classes α (*alpha*), β (*beta*) e γ (*gamma*) das ciclodextrinas (12), por conta da maior quantidade de dados disponíveis.

3.1.2 Representação Molecular

No contexto de trabalho interdisciplinar da modelagem computacional e da química computacional, as moléculas podem ser modeladas estruturalmente em um espaço bidimensional ou tridimensional (coordenadas dos átomos, funções de densidade e/ou campos de força), ou pelo cálculo de descritores topológicos e/ou físico-químicos que caracterizam as moléculas no espaço de variáveis (8). Cada uma dessas informações apresentam diferentes níveis de informação e diferentes graus de dificuldade para sua obtenção.

A representação molecular no presente trabalho segue por meio da utilização de descritores para cada uma das moléculas, além do uso de descritores de ambiente para caracterizar a solução experimental da interação entre a molécula hospedeira e o seu ligante.

Dessa forma, o cálculo de métricas de interação entre duas moléculas poderia ser previsto utilizando descritores individuais facilmente calculáveis de cada molécula, auxiliando em processos preliminares ao da triagem computacional clássica. Além disso, predições dessa natureza podem ser utilizadas para definir ordens (cronograma) para testes experimentais posteriores, priorizando a investigação de moléculas que se mostrem mais promissoras.

Para isso, é necessário calcular os descritores individuais de cada uma das moléculas e garantir que esse processo consiga ser realizado para aplicações posteriores em moléculas pouco conhecidas. Assim, utilizou-se o programa `KNIME` para definir um *pipeline* para os cálculos dos descritores, apresentado na subseção 3.1.3.

3.1.3 Cálculo dos Descritores Moleculares: *KNIME*

O cálculo computacional dos descritores moleculares individuais para todas as moléculas consideradas nesses trabalho (moléculas hospedeiras e ligantes) foram realizadas por meio da ferramenta *KNIME*.

O `KNIME` (https://www.knime.com/) é um *software* gratuito e de código aberto que oferece ferramentas para tratar problemas da ciência de dados, desde as etapas de obtenção e tratamento de dados até as etapas de obtenção de conhecimento por ferramentas de aprendizado de máquinas. A interface intuitiva do programa permite que o usuário opere os dados por meio da criação de fluxogramas gráficos (ou *pipelines* de execução). O

conjunto de ferramentas oferecidas pelo *software* fez com que o KNIME se tornasse comum em aplicações de pesquisas da área farmacêutica e da química computacional, porém também vêm sendo sendo aplicado na área de inteligência de negócios e análise de dados financeiros.

A ferramenta foi utilizada no presente trabalho por possuir diversas opções para cálculos de descritores para dados moleculares. Além disso, sua interface intuitiva permite que profissionais da área de aplicação desse trabalho consigam utilizá-lo sem que conhecimento em linguagens de programação fossem requeridos. A Figura 10 apresenta o fluxograma de execução no programa KNIME, criado para obtenção dos descritores necessários para os testes realizados neste trabalho. As entradas são as estruturas químicas das três moléculas hospedeiras consideradas e o conjunto de estruturadas das moléculas hospede. Os cálculos dos descritores foram efetuados por meio do módulo RDKit Descriptor Calculation (https://hub.knime.com) presentes no KNIME. O restante do fluxograma integra os dados das moléculas em instâncias para o problema hospedeiro-hóspede.

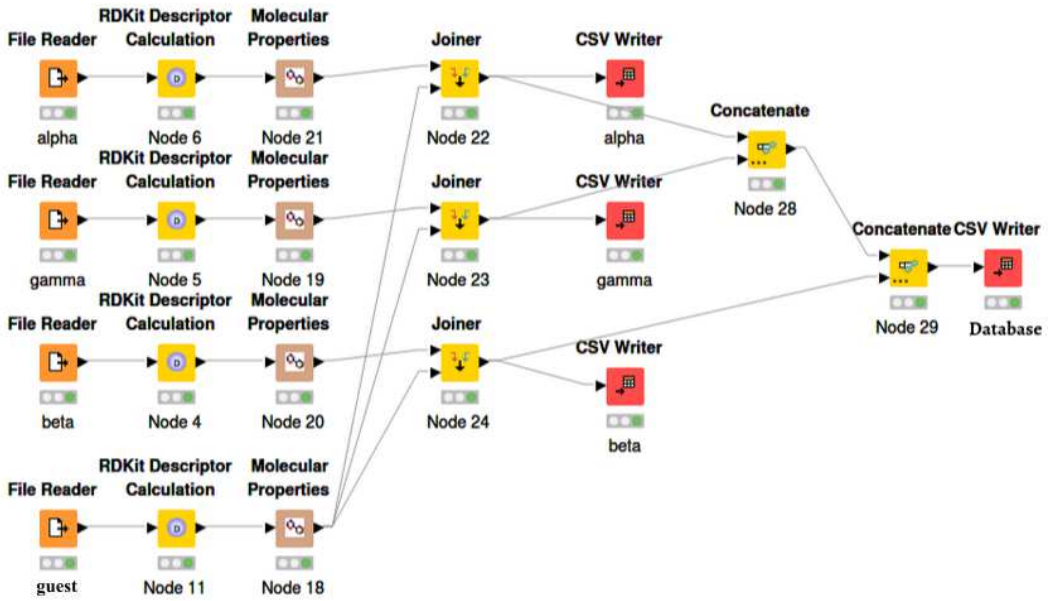


Figura 10 – Fluxograma utilizado para cálculo dos descritores moleculares no programa KNIME.

Fonte: Produzida pelo autor (2021).

Com esse ferramental, partindo da representação em *smiles* de cada molécula proporcionada pelo banco de dados BindingDB, calcula-se, utilizando o programa KNIME, as demais propriedades físico-químicas das instâncias da base (15).

3.1.4 Descritores Moleculares Considerados

A representação molecular proposta nesse trabalho passa pelo cálculo de descritores moleculares que consigam representar a natureza de cada molécula. Para isso, utilizou-se do módulo `RDKit Descriptor Calculation`. Para esse trabalho, considerou-se as seguintes propriedades fisico-químicas (descritores) para cada molécula no banco:

- **NumAtoms**: Número de átomos na molécula.
- **SlogP**: Se trata de uma predição para o valor de LogP molecular, que é o log do coeficiente de partição octanol/água (incluindo hidrogênios implícitos). O valor é calculado avaliando a contribuição de cada átomo da estrutura no modelo proposto por Wildman e Crippen (50). A predição é feita considerando a protonação molecular informada como entrada.
- **SMR** (*S. Molecular Refractivity*): Se trata de uma predição para a refratividade molecular (incluindo hidrogênios implícitos). O valor é calculado avaliando a contribuição de cada átomo da estrutura no modelo proposto por Wildman e Crippen (50). A predição é feita considerando a protonação molecular informada.
- **LabuteASA** (*Labute's Approximate Surface Area*): Se trata de uma predição para o valor da área de superfície molecular acessível ao solvente (à água). O valor é calculado considerando um raio de 1,4Å para a molécula de água e cada átomo é representado por uma geometria poliédrica, como proposto por Labute (51).
- **TPSA** (*Topological Polar Surface Area*): Se trata de uma predição para o valor da superfície polar, considerando os parâmetros moleculares propostos por Ertl *et. al.* (52).
- **AMW** (*Average Molecular Weight*): Se trata do valor média calculado para o peso molecular.
- **NumLipinskiHBA**: Se trata do valor da definição padrão de Lipinski para o número de receptores em ligações de hidrogênio (HBA).
- **NumLipinskiHBD**: Se trata do valor da definição padrão de Lipinski para o número de doadores em ligações de hidrogênio (HBD).
- **NumRotatableBonds**: Se trata do número de ligações entre átomos da molécula com grau de liberdade rotacional livre.
- **Formal Charge**: Carga total da molécula, ou seja, soma das cargas formais de cada átomo.

Os descritores explicitados acima foram calculados para cada uma das moléculas, ligantes e hospedeiros, exceto o valor de carga formal para hospedeiros. Isso se deu pois todos os hospedeiros considerados (ciclodextrinas) possuem carga formal nula. Dessa forma, o atributo não traria informação alguma para os modelos de predição. Como exemplo, a Tabela 1 apresenta os valores obtidos com os cálculos dos descritores para as três moléculas hospedeiras consideradas.

Tabela 1 – Informação dos atributos calculados para cada uma das 3 moléculas de ciclodextrina (CD) consideradas nesse trabalho.

	α-CD – BDBM4	**β-CD – BDBM11**	**γ-CD – BDBM36126**
SlogP	-13,06	-15,23	-17,41
SMR[a]	195,80	228,43	261,07
ASA[b]	372,13	433,92	495,72
TPSA[c]	474,90	554,05	633,20
AMW[d]	972,84	1134,99	1297,13
HBA[e]	30,0	35,0	40,0
HBD[f]	18	21	24
RB[d]	6	7	8
Átomos	126	147	168

[a]SMR: Molecular Refractivity. [b]ASA: Approximate Surface Area. [c]TPSA: Topological Polar Surface Area. [d]AMW: Average Molecular Weight. [e]HBA: HB Acceptor. [f]HBD: HB Donor. [g]RB: number of Rotatable Bonds.

Tendo em vista que cada instâncias do banco é composta por duas moléculas, um ligante e um hospedeiro, os atributos de cada instância de nossas bases podem ser divididos em três categorias: `lig_atr` são os atributos relativos ao ligante, `host_atr` são os atributos relativos aos hospedeiros (ciclodextrinas), e `env_atr` são os atributos relativos ao meio no qual o experimento foi realizado, categoria composta pelos atributos pH e Temperatura (°C). As instâncias ainda possuem o valor resposta de interesse, a afinidade de ligação entre o ligante e o hospedeiro, descrita por ΔG (KJ/mol).

Por fim, vale ressaltar que os atributos utilizados nos testes desse trabalho estão em sintonia com o que vem sendo utilizado em pesquisas similares na literatura. A Figura 11 apresenta uma comparação entre os atributos utilizados na pesquisa do laboratório de Defang Ouyang (1) e os atributos considerados no trabalho atual. A nomenclatura dos atributos segue a mesma utilizada no trabalho de Defang Ouyang para fácil comparação. Além disso, o trabalho citado utilizada técnicas baseadas em árvores de decisão que permitem uma ordenação da importância de cada atributo utilizado na tomada de decisão do regressor *LightGBM*. Vale ressaltar que alguns atributos utilizados na pesquisa de Defang Ouyang não possuem equivalência no trabalho atual. Isso é justificado pois pretendia-se manter uma representação simplificada a princípio e de fácil generalização para moléculas pouco conhecidas. Além disso, Defang Ouyang possuía uma base de dados com maior variedade de moléculas hospedeiras e hóspedes que justificam a utilização de um número superior de atributos descritivos.

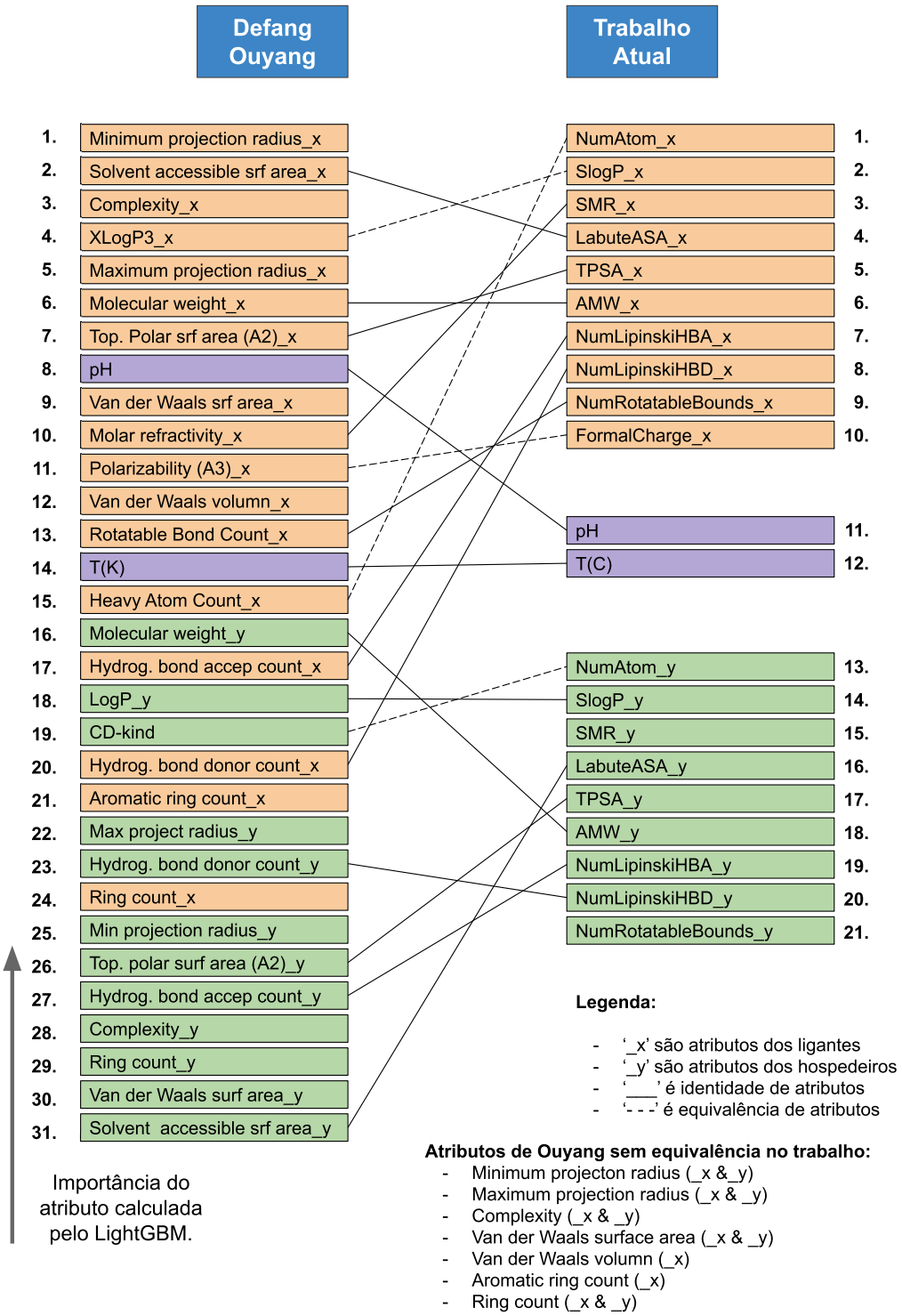


Figura 11 – Comparação entre os atributos utilizados no presente trabalho com os atributos utilizados por Defang Ouyang *et. al.* (1). Os atributos em salmão são relativos às moléculas hóspede, os atributos em verde são relativos às moléculas hospedeiras, e os atributos em violeta são relativos ao meio no qual os experimentos foram realizados.

Fonte: Produzida pelo autor (2021).

3.1.5 **Processo de Tratamento das Instâncias**

A base de dados obtida a partir do *BindingDB* continha 577 observações. O primeiro passo aplicado foi a remoção das observações com algum valor faltante em relação aos atributos do meio (pH e Temperatura) e da variável objetivo (ΔG). Essas instâncias não serviam para os testes, uma vez que não apresentavam os dados experimentais para a interação molecular do sistema. Dessa forma, a base passou a conter 485 observações.

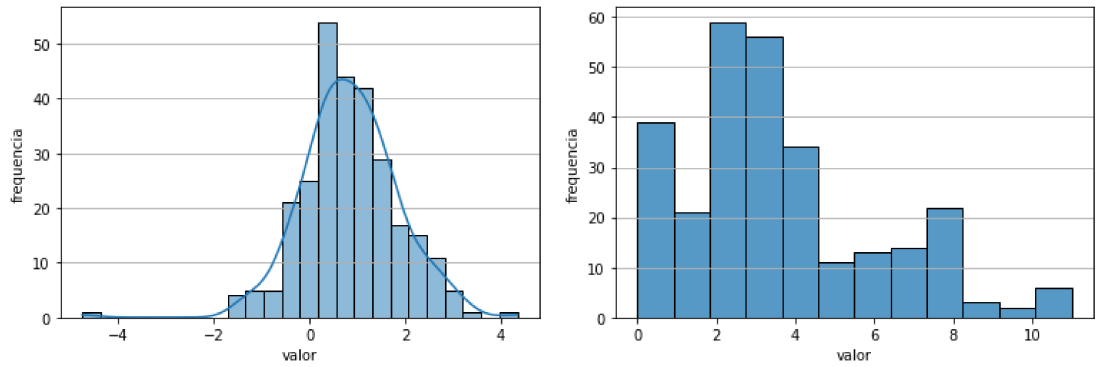
Partindo dessa premissa, aplicou-se os filtros em `pH` e `Temp` para adequar aos limites informados na Subseção 3.1.1. A partir deste momento, a base apresentou 369 instâncias. Por último, removeu-se as observações duplicadas na base de dados. Duplicatas acontecem na base, pois é possível que mais de um experimento tenha sido realizado sobre as mesmas moléculas dentro do ambiente considerado por diferentes laboratórios. As medições podem possuir diferenciações na definição da entalpia e entropia de interação. Como essas medidas não são objetivo nesse trabalho, apenas o valor total do potencial de interação, essas instâncias foram removidas para não causar problemas no treinamento das máquinas. Sendo assim, após a remoção das duplicatas, a base de dados utilizada para o desenvolvimento deste trabalho apresentava a composição mostrada na Tabela 2.

Tabela 2 – Composição da base de dados compilada.

Propriedade	Contagem
Quantidade de observações	280
Quantidade de colunas	25
Atributos dos ligantes	10
Atributos dos hospedeiros	9
Atributos do ambiente	2
Atributos identificadores	3
Variável resposta	1
Número de ligantes	145
Número de hospedeiros	3

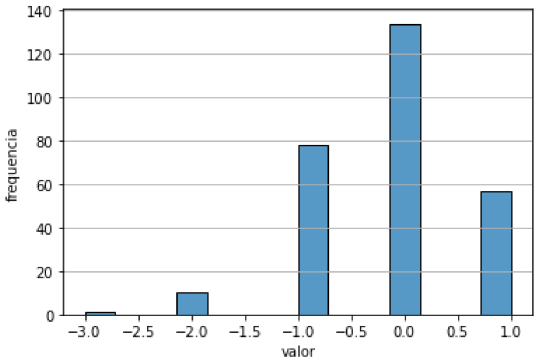
A Figura 12 mostra algumas das distribuições para os descritores de moléculas hóspedes, além disso, apresenta-se a distribuição geral para os valores da energia de complexação molecular considerados na base (ΔG). Informações adicionais sobre as demais variáveis são apresentadas na Tabelas 9 e 10, as quais apresentam estatísticas das distribuições do conjunto de dados.

Por fim, a Figura 13 apresenta o formato da base construída após todos os tratamentos efetuados e um resumo de suas dimensões e unidade de medida. Repare que o número de instâncias para cada tipo de CD hospedeira varia. Verificar tal característica será importante nas discussões apresentadas na seção a seguir (Seção 3.2), pois será necessário aplicar métodos capazes de lidar com essa natureza desbalanceada da base de dados.

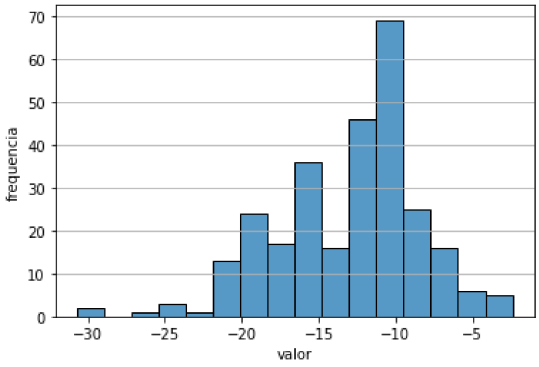


(a) SlogP

(b) RB



(c) FC



(d) ΔG

Figura 12 – Apresentação da distribuição dos valores de descritores para a molécula hóspede (a), (b), e (c), e ΔG em (d)

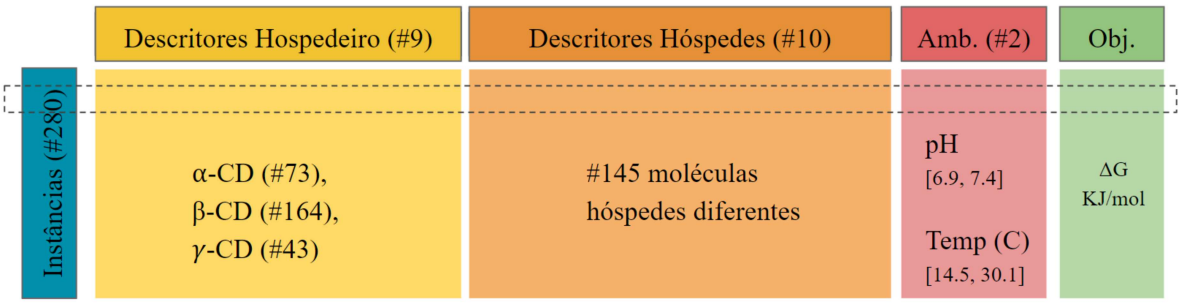


Figura 13 – Formato e dimensões da base de dados considerada no presente trabalho após os tratamentos realizados.

Fonte: Produzido pelo autor (2021).

3.2 ABORDAGEM POR APRENDIZADO DE MÁQUINA

Nesta seção, explicita-se como os dados preparados na seção anterior serão aplicados para o objetivo desse trabalho. Aqui são descritas as formulações dos métodos de aprendizado de máquina utilizados, as estratégias de otimização de hiperparâmetros e as métricas utilizadas na avaliação dos resultados.

Como visto anteriormente, o banco de dados considerado é comporto por 280 observações únicas, 25 descritores moleculares (9 para os hospedeiros, 10 para os hóspedes, 2 para o ambiente, 3 identificadores e 1 variável objetivo). Os descritores identificadores são de caráter informativos e não são utilizados nos treinamentos dos modelos de regressão, assim, trabalha-se com uma base descritiva com 21 variáveis explicativas e 1 variável explicada.

O conjunto de dados foi aleatoriamente dividido em um conjunto único de treinamento (224) e teste (56) usando o método *Stratified K-fold* (53) para manter a mesma proporção de instâncias de cada classe de CD em ambos os conjuntos (54), juntamente com uma análise de divergência de *Kullback-Leibler* (55) entre os conjuntos. O processo de geração de conjunto de treinamento e teste foram iterados para definir conjuntos que possuíssem distribuições mais homogêneas. Detalhes dessa seleção são apresentados na Subseção 3.2.1. A Figura 14 mostra esquematicamente as etapas descritas para a separação da base e avaliação dos métodos de aprendizado de máquina considerados. Os dados e código são disponibilizados pelos autores mediante solicitação.

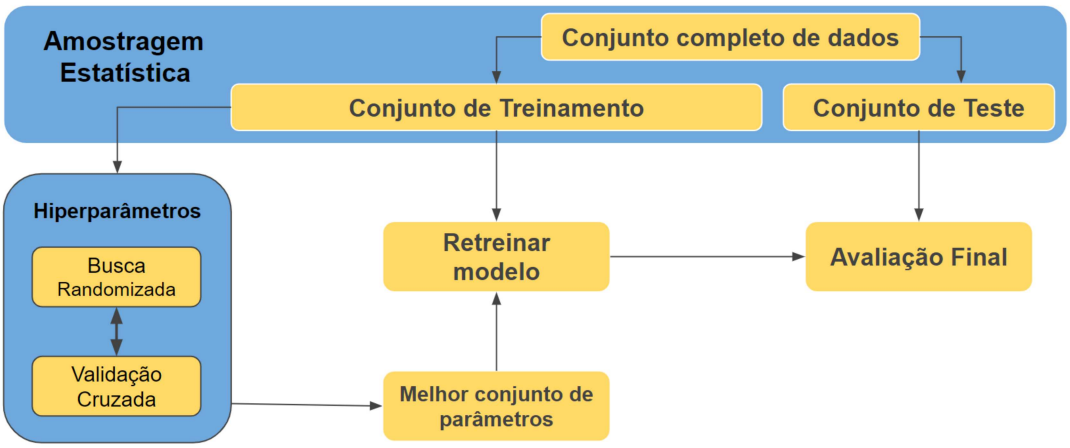


Figura 14 – Esquema das etapas para amostragem da base de dados, treinamento e avaliação dos modelos de aprendizado de máquina considerados no presente trabalho.

Fonte: Produzido pelo autor (2021).

No presente trabalho, compara-se o desempenho de três métodos de aprendizado de máquina para prever a afinidade de ligação de diferentes ligantes com as ciclodextrinas: (i) *ε-Support Vector Regression*, (ii) *Gaussian Process Regressor*, e (iii) *eXtreme Gradient Boosting*. A seguir, são apresentadas as tecnologias aplicadas no decorrer do desenvolvimento da pesquisa.

3.2.1 Seleção dos Conjuntos de Treinamento e Teste

No campo da matemática estatística, a métrica de Divergência de Kullback-Leibler (D_{kl}), também documentada como Entropia Relativa, é uma medida da diferença entre

duas distribuições de probabilidade (56, 57). A métrica é dada pela Equação 3.1, onde x e y são duas distribuições de entrada.

$$D_{kl}(x,y) = \begin{cases} x\log(x/y) - x + y & x > 0, y > 0 \\ y & x = 0, y \geq 0 \\ \infty & \text{caso contrário} \end{cases} \tag{3.1}$$

A fim de definir conjunto de treinamento e teste com propriedades semelhantes para garantir as análises de aplicabilidade dos métodos em um domínio de aplicação bem definido, executou-se uma estratégia com múltiplas amostragens na qual distintos conjuntos de treinamento e teste na proporção 80/20 eram selecionados a cada iteração, respeitando restrições de mesmas proporções de instâncias compostas por hospedeiros de mesma classe, como esquematizado na Figura 15 (discutida a seguir).

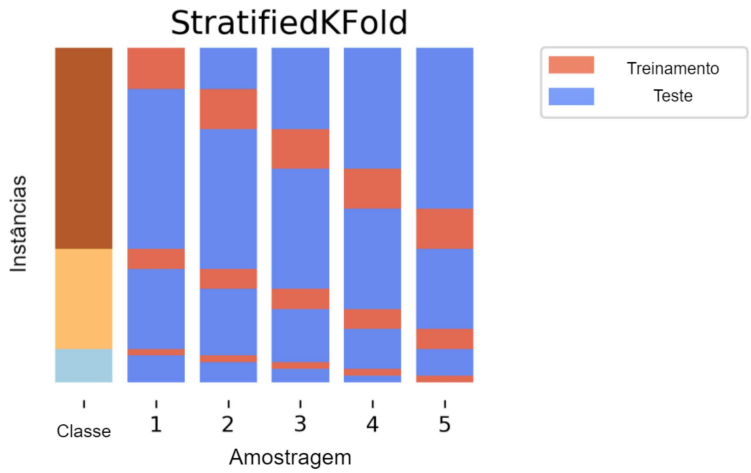


Figura 15 – Esquema de seleção de conjuntos de treinamento e teste utilizando o método *Stratified K-fold*.

Fonte: Traduzido de (58).

Em cada iteração, calculou-se a entropia para as distribuições de energia de iteração (ΔG) das instâncias dos conjuntos de treinamento e teste definidos. Nas iterações do processo de amostragem, buscou-se minimizar a entropia relativa entre as distribuições, ou seja, buscou-se os valores mais homogêneos de D_{kl} para as distribuições de treinamento e teste no espaço de busca. O método utilizado para o cálculo em questão está disponível na biblioteca `Scipy` (59).

Os conjuntos de treinamento e teste de mesma proporção de instâncias de moléculas hospedeiras foram definidos em cada iteração da busca por meio da aplicação do método *Stratified K-fold* (53), o qual garante sorteio respeitando as proporções de uma variável categórica informada como parâmetro. A Figura 15 mostra esquematicamente como a

seleção é feita. No nosso exemplo, a classe da instância é dada pelo tipo de ciclodextrina hospedeira do complexo em questão.

3.2.2 *ε*-*Support Vector Regression* (*ε*-SVR)

O *ε*-Support Vector Regression (*ε*-SVR) é uma adaptação da formulação clássica de Máquinas de Vetor Suporte (60) que vem sendo aplicada em diversos campos de pesquisa e se mostrado uma ferramenta promissora no desenvolvimento racional de fármacos (61). Estudos prévios realizados no decorrer do desenvolvimento dessa pesquisa indicam bons resultados com este método, embora haja espaço para melhorias (17, 16).

O *ε*-SVR é um modelo de regressão linear, como descrito pela Equação 3.2,

$$f(\mathbf{x}) = \sum_{j=1}^{N} w_j K\left(\mathbf{x}, \mathbf{x}_j\right) + b \tag{3.2}$$

onde $K(\cdot, \cdot)$ é uma função *kernel* geralmente associada a uma transformação não linear, $\mathbf{w} = [w_1, \cdots, w_N]$, é o vetor de pesos, b é o bias e N é o número de amostras. Nesse trabalho, aplicou-se função *kernel* de base radial (`RBF`) para realizar a transformação não linear dos dados, descrita por Eq. (3.3)

$$K\left(\mathbf{x}, \mathbf{x}_i\right) = \sum_{i=1}^{N} \exp\left(-\gamma \left\|\mathbf{x} - \mathbf{x}_i\right\|^2\right) \tag{3.3}$$

, onde γ é o fator de escala (*length scale*).

No método *ε*-SVR o vetor ótimo de pesos $\mathbf{w}$ e b são computados por meio da minimização de Eq. (3.4) (62):

$$J = \sum_{i=1}^{N} w_i^2 + \frac{C}{N} \sum_{i=1}^{N} L_\varepsilon\left(y_i - f\left(\mathbf{x}_i\right)\right) \tag{3.4}$$

onde

$$L_\varepsilon(y - f(\mathbf{x})) = \begin{cases} 0 \text{ if } |y - f(\mathbf{x})| \leq \varepsilon \\ |y - f(\mathbf{x})| \text{ caso contrário} \end{cases}$$

e y_i são dos dados de saída associados a $\mathbf{x}_i$. L_ε é a função de perda *ε*-insensível (63), C é um parâmetro de regularização e ε é um parâmetro de erro máximo de erro admitido sem penalização que caracteriza o método *ε*-SVR. A característica do método de possuir um parâmetro de erro máximo é de grande valia, pois isso permite a definição de uma faixa de erro aceitável na etapa de ajuste aos dados de treinamento.

3.2.3 *eXtreme Gradient Boosting* (XGB)

O *eXtreme Gradient Boosting* (XGB) (64) é um método *ensemble* que combina vários estimadores de baixa complexidade e alto erro, chamados de aprendizes fracos, para gerar um aprendiz robusto. No caso do XGB, os estimadores fracos são árvores de decisão

regularizadas (65). A proposta geral da maior parte dos métodos do tipo *boosting* é treinar preditores sequencialmente, de modo que cada iteração tente corrigir erros da etapa de treinamento anterior.

Dessa forma, a predição do XGB para uma dada instância i é formada por

$$\hat{y}_i = F_M(x_i) = \sum_{m=1}^{M} h_m(x_i) \tag{3.5}$$

onde `M` é o número de estimadores, e h_m são os estimadores fracos construídos com base nos parâmetros de profundidade máxima da árvore (`max_depth`) e um fator de perda mínimo para a partição de uma nova folha nas árvores (Γ_{tree}). Por se tratar de um método de *boosting* integrador, ele é construído de uma forma gulosa, conforme apresentado na Equação 3.6,

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \tag{3.6}$$

onde η é uma taxa de aprendizado para controlar situações de *overffiting* e proporcionar uma aprendizado gradual na evolução das iterações do método, e toda nova árvore adicionada ao conjunto de estimadores $h_m(x)$, dado por Eq. (3.7), é ajustada de forma a minimizar um somatório de perdas L_m.

$$h_m = \arg\min_h L_m = \arg\min_h \sum_{i=1}^{N} l(y_i, F_{m-1}(x_i) + h(x_i)) + \Omega(F_{m-1}) \tag{3.7}$$

Em Eq. (3.7), considera-se

$$l(y_i, F_m(x)) = \left[y_i - F_m\left(x_i\right)\right]^2 \tag{3.8}$$

onde N é o número de amostras e

$$\Omega(F_m) = \alpha_{reg} T + \frac{1}{2}\lambda_{reg}||\mathbf{w}||^2 \tag{3.9}$$

é um termo de regularização, T é o número de folhas nas árvores de decisão (estimadores fracos), $\mathbf{w}$ são os pesos das folhas, e α_{reg} e λ_{reg} são a L_1 e L_2 termos de regularização dos pesos, respectivamente.

Recentemente, pesquisadores vem obtendo resultados promissores com a aplicação do método XGB para prever a interação molecular em sistemas Hospedeiro-Hóspede (1, 54). O que justifica a inclusão da metodologia nesse estudo comparativo.

3.2.4 ***Gaussian Process Regressor*** **(GPR)**

Processos Gaussianos proporcionam uma forma prática e probabilística em uma abordagem Bayesiana que pode ser aplicável no contexto de regressão em aprendizado em máquina permitindo a definição de funções admissíveis com base linear ou não linear diversas (66). Nas últimas décadas, processos gaussianos têm recebido cada vez mais uma maior atenção na comunidade de pesquisa em aprendizagem em máquina (67, 68, 69).

Ao contrário de muitos algoritmos populares de aprendizado de máquina supervisionado que aprendem valores exatos para cada parâmetro em uma função, uma abordagem Bayesiana infere uma distribuição de probabilidade sobre todo um conjunto de funções admissíveis (66). Para o caso linear, dada uma função linear $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}$, podemos escrever $y = f(\mathbf{x}) + \epsilon$, onde $\mathbf{x}$ é o vetor de entradas, $\mathbf{w}$ é um vetor de parâmetros, y é o valor objetivo observado e ϵ é um erro que geralmente é assumido com comportamento normal com média zero (70). De forma geral, a abordagem Bayesiana define uma distribuição prévia, $p(w)$, sobre o parâmetro w e ajusta probabilidades baseadas em evidências retiradas dos dados de treinamento usando a Regra de Bayes (69):

$$p(w|y, X) = \frac{p(y|X, w)p(w)}{p(y|X)} \tag{3.10}$$

onde a distribuição atualizada $p(w|y, X)$, chamada de distribuição posteriori, incorpora informações tanto da distribuição *a priori* como do conjunto de dados de treinamento.

Além disso, a complexidade das funções admissíveis não se limita em aspectos lineares, podendo ser definidos por meio de funções *kernel* não lineares para a formação de matrizes de covariância (66). Por esse motivo, métodos de regressão baseados em processos gaussianos são modelos não paramétricos. De fato, ao realizar regressões utilizando modelos paramétricos, a complexidade ou flexibilidade dos modelos é limitada pelo número de parâmetros da forma funcional que a metodologia aplicada admite, como uma forma linear ou quadrática. Nos modelos não paramétricos o número de parâmetros cresce com o tamanho do conjunto de dados observado no treinamento. Como os dados de treinamento servem para definir a parte *a priori* da função, sua complexidade é relativa aos dados de entrada e com comportamento regido por uma função *kernel* matricial (68).

Para obter previsões em novos dados distintos do conjunto de treinamento, x^*, a distribuição preditiva (prevista) pode ser calculada ponderando todas as previsões possíveis por sua distribuição posteriori calculada pela Equação 3.11 (69).

$$p\left(f^* \mid x^*, y, X\right) = \int_w p\left(f^* \mid x^*, w\right) p(w \mid y, X) dw \tag{3.11}$$

Um Processo Gaussiano assume que $p(f(x_1, ..., f(x_N)$ é conjuntamente gaussiano com certa média e covariância $\mathbf{K}$ dada por $K_{i,j} = K\left(\mathbf{x}_i, \mathbf{x}_j\right)$, onde K(.) é uma função *kernel* positivamente definida (discutida a seguir). A ideia-chave é que se x_i e x_j são considerados semelhantes pelo *kernel*, esperamos que a saída da função nesses pontos também seja semelhante (71). Assumindo isso e resolvendo a Equação 3.11 para a distribuição posteriori preditiva, obtemos uma distribuição da qual podemos obter uma previsão pontual estimada pela média e uma quantificação da incerteza estimada pela variância relativa (71).

Visto isso, o método *Gaussian Process Regressor* (GPR) (72) é a implementação computacional de um processo gaussiano não paramétrico (não limitado por uma forma

funcional) que, ao invés de calcular a distribuição de probabilidade de parâmetros de uma função específica, calcula a distribuição de probabilidade sobre todas as funções admissíveis que se ajustam aos dados. Assim, semelhante ao mostrado acima, especifica-se uma função a priori (no espaço da funções consideradas), calcula-se uma função posteriori usando os dados de treinamento, e calcula-se a distribuição posteriori preditiva em dados não vistos pelos modelos (69).

Como descrito para Xiaoling Ou *et al.* (73), para abordagens de regressão, um modelo de processo gaussiano pode ser definido como:

$$\hat{y} = f(X) + \epsilon \tag{3.12}$$

onde $f(\cdot)$ é uma função não linear que associa o vetor de entrada X a uma saída escalar, $\hat{y}$ é o valor objetivo observado e ϵ é um ruído Gaussiano com média zero (i.e. $\epsilon \sim G(O, \sigma_v^2)$). A priori, assume-se $f(.)$ como um processo Gaussiano, $f(.) \sim GP(\mu, \mathbf{K}(X, X'))$, onde μ é usualmente definido como zero e $\mathbf{K}(X, X')$ é a matriz de covariância (74).

A covariância a priori é especificada através da aplicação de um objeto *kernel* nos dados. Nesse trabalho, aplica-se tanto o *kernel* `RBF`, como descrito na Equação 3.3, quanto o *kernel* `Matern`, dado pela Equação 3.13,

$$K_{i,j} = K\left(\mathbf{x}, \mathbf{x}_i\right) = \frac{1}{\Gamma_f(\nu) 2^{\nu-1}} \left(\frac{\sqrt{2\nu}}{\sigma} \left\| \mathbf{x} - \mathbf{x}_i \right\|^2 \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{\sigma} \left\| \mathbf{x} - \mathbf{x}_i \right\|^2 \right) \tag{3.13}$$

onde σ é um fator de escala (*length scale*), ν controla a suavização dos resultados da função, $\Gamma_f(.)$ é a Função Gama, e K_ν é uma adaptação da Função de Bessel (75).

Assim, conforme Wang *et. al.* (74), a distribuição predita (ou a média da distribuição para uma predição pontual) para a saída $\hat{y}$, dada sua entrada x, é calculada por:

$$\hat{y}(\mathbf{x}) = \mathbf{k}^{\mathrm{T}}(\mathbf{x})(\mathbf{K} + \mathbf{I})^{-1}\mathbf{y} \tag{3.14}$$

onde $\mathbf{y} = [y_1, y_2, \ldots, y_N]^{\mathrm{T}}$ é o vetor com os dados de treinamento do modelo, $\mathbf{k} = [K\left(\mathbf{x}, \mathbf{x}_1\right), K\left(\mathbf{x}, \mathbf{x}_2\right), \ldots, K\left(\mathbf{x}, \mathbf{x}_N\right)]^{\mathrm{T}}$ é formada por uma função *kernel* positivamente definida aplicada sobre os dados de treinamento e os dados de entrada e $\mathbf{K}$ é a função matricial de covariância calculada sobre os dados observados no treinamento.

Modelos GPR dentem a apresentar bons resultados preditivos mesmo em conjuntos de dados de tamanho reduzido, contanto que os novos dados, não observados na etapa de treinamento, mantenham certo grau de similaridade com os dados observados e se concentrem no interior dos intervalos considerados no treinamento. Como o conjunto de dados desse trabalho possui tamanho moderado e é mantém-se um controle do domínio dos conjuntos de treinamento e teste, justifica-se a aplicação do método GPR nessa comparação.

3.2.5 **Estratégia *Randomized Search* (RS)**

Além dos parâmetros internos ajustados na etapa de treinamento, os métodos de aprendizado de máquina são geralmente sensíveis às definições de hiperparâmetros (76). Neste trabalho, aplica-se uma abordagem híbrida acoplando os métodos aprendizado de máquina com o ajuste de hiperparâmetros através de uma estratégia de Busca Randomizada (*Randomized Search* - RS). O RS realiza uma escolha aleatória de valores, onde cada configuração é amostrada a partir de uma distribuição sobre possíveis valores de parâmetros (20). O processo permite que um conjunto de valores seja escolhido independentemente do número de parâmetros em questão. Além disso, adicionar parâmetros que não influenciam o desempenho não diminui a eficiência, o que é uma boa característica para análises iniciais.

A estratégia de Busca Randomizada, realizadas sobre um *Random Layout*, é geralmente confundida com estratégias de Busca em Grade, realizadas sobre um *Grid Layout* (20). A Figura 16 mostra esquematicamente a diferença entre essas estratégias de busca. Uma busca em grade cria pontos igualmente espaçados nos intervalos definidos para a busca. Buscas randomizadas partem da premissa da definição de uma distribuição de probabilidade contínua para o comportamento da busca de cada um dos hiperparâmetros. Diferentes distribuições vão direcionar amostragens em determinadas regiões da busca. Buscas em grade são sensíveis a variáveis pouco importantes, testando todas as possibilidades para essas variáveis de forma exaustiva. Buscas randomizadas tendem a amostrar regiões definidas a princípio como positivas, podendo reduzir o processo de amostragem, atingir resultados tão bons quando buscas em grade, além de não se limitar a valores de otimização definidos pela discretização do espaço de busca, podendo retornar qualquer valor real na distribuição contínua de probabilidade (20).

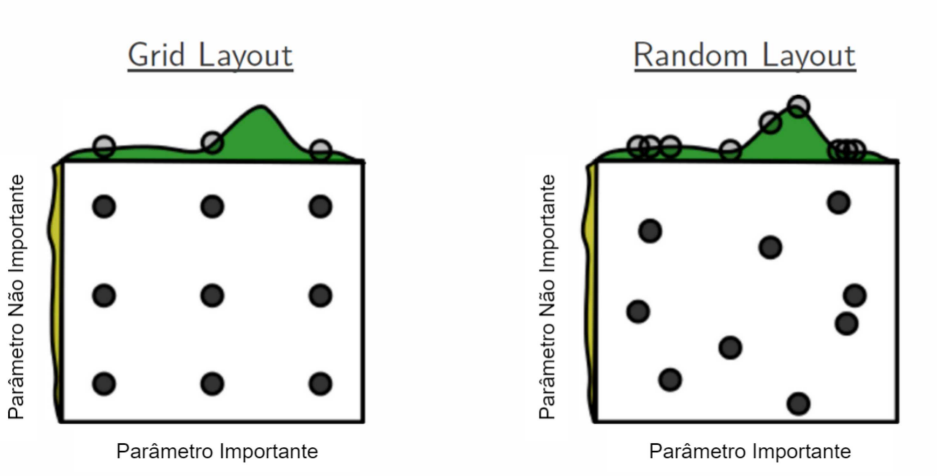


Figura 16 – Comparação entre estratégias de Busca em Grade e de Busca Randomizada.

Fonte: Traduzido de (20).

No presente trabalho, cada execução do RS considera uma amostragem de 1000

máquinas de aprendizado e executa treinamentos aplicando uma estratégia de validação cruzada 3-*fold* em cada uma delas. A estratégia busca encontrar o conjunto de hiperparâmetros relacionado no menor erro RMSE no conjunto de treinamento. Todos os parâmetros otimizados seguem distribuições uniformes como descrito a seguir:

- Para ε-SVR, `C` $\in [0, 10^4]$, $\varepsilon \in [0, 10]$, e $\Gamma \in [0, 10]$. Cada máquina é ajustada internamente com 10.000 iterações (no máximo), utilizando a função *kernel* `RBF`.

- Para XGB, `M` $\in [1, 300]$, `max_depth` $\in [1, 10]$, $\eta \in [10^{-5}, 1]$, $\Gamma_{tree} \in [10^{-5}, 5]$, $\alpha_{reg} \in [10^{-5}, 2]$, $\lambda_{reg} \in [10^{-5}, 2]$.

- Para GPR, `n_restarts` $\in [0, 10]$ e a função kernel é selecionada entre `RBF` ($\sigma \in [0.1, 30]$) ou `Matern` ($\sigma \in [0.1, 30]$ e $\nu \in [0.1, 3]$). Todas as execuções consideram $\alpha_{reg} = 50^{-5}$.

3.2.6 Validação Cruzada

Validação cruzada é uma técnica de amostragem estatística para avaliar a habilidade de generalização de um modelo a partir de um subconjunto de dados (77).

Neste trabalho, utiliza-se a técnica *K-Fold* (78). Entre as técnicas de validação cruzada, a técnica *K-Fold* é uma das mais utilizadas e consiste em dividir todas as amostras de entrada em K partes de tamanhos aproximadamente iguais. Parte desse conjunto é utilizada para treinar o modelo, enquanto uma parte é separada no início do processo e utilizada para avaliar a qualidade do que foi definido. Um conjunto de dados contendo N amostras é dividido em K subconjuntos ($K \geq 1$), onde $K-1$ são utilizados no treinamento e a seleção restante é usada para avaliação. Esse processo é repetido K vezes, utilizando um conjunto diferente para avaliação a cada iteração.

Aplicar técnicas de validação cruzada acopladas a métodos de otimização de hiperparâmetros ajudam a garantir bons resultados dos modelos treinados frente a novos conjuntos de dados. Como o processo de validação cruzada é aplicado no interior da Busca Randomizada, a estratégia é utilizada para separar o conjunto de treinamento de entrada em subconjuntos de treinamento e validação para definir os hiperparâmetros do modelo em questão. A Figura 17 esquematiza o funcionamento do processo de validação cruzada por meio da estratégia *K-Fold*.

3.2.7 Métricas para Avaliação dos Modelos

Foram utilizadas três (3) métricas para o cálculo dos erros de predição nos conjuntos de treinamento e de teste (79). Abaixo a descrição de cada uma delas:

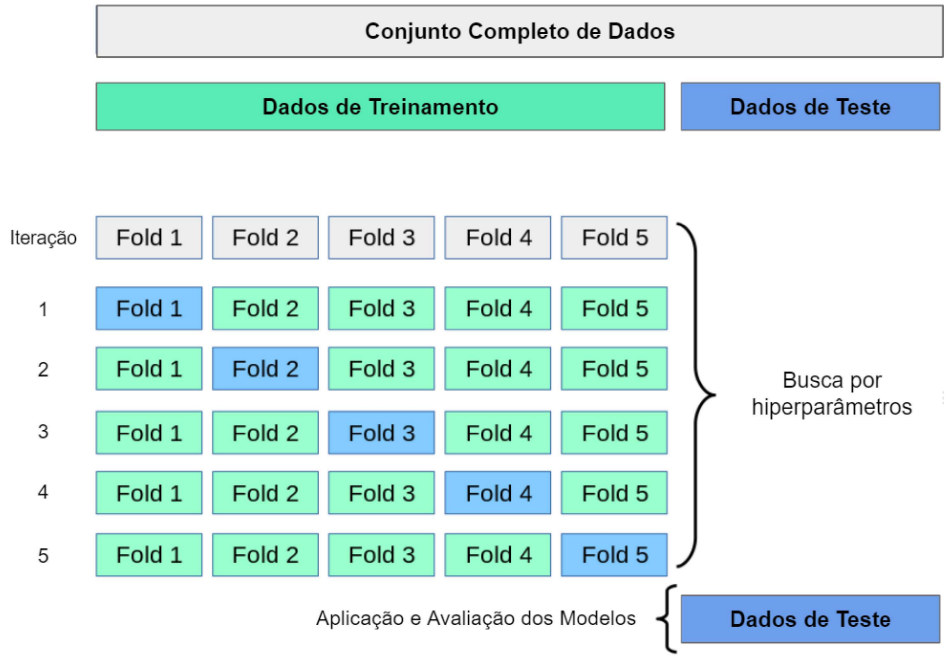


Figura 17 – Divisão de conjunto de treinamento e teste. No conjunto de treinamento, os dados são dividido em K = 5 subconjuntos no processo de validação cruzada utilizando a técnica *K-Fold*.

Fonte: Adaptado de (58).

- **MAPE** (*Mean Absolute Percentage Error*): Se trata de uma função de perda para as regressões calculada por meio do erro percentual absoluto médio. A métrica é definida pela Equação 3.15,

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100\% \tag{3.15}$$

onde y é o valor real da medida objetivo e $\hat{y}$ é o valor da predição obtida, calculados para cada uma das n instância i.

- $\mathbf{R^2}$: Se trata de do coeficiente de determinação entre os valores reais e preditos. Essa métrica representa a proporção de variância que foi explicada pelas variáveis independentes do modelo, portanto, indicando a qualidade do ajuste dos valores ao modelo. O melhor valor para essa métrica é 1. Valores negativos podem ser observados para modelos de qualidade ruim. A métrica pode ser calculada pela Equação 3.16,

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{3.16}$$

onde y é o valor real, $\hat{y}$ é o valor predito e $\bar{y}$ é a média entre os valores reais dada por $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ considerando as n instâncias i.

- **RMSE** (*Root-Mean-Square Deviation*): Se trata de um estimador para a raiz da diferença quadrática média entre os valores preditos e os valores reais. Nos exemplos tratados, essa medida representa a raiz quadrada da variância entre os valores reais e preditos, também descrito como erro padrão. Essa métrica se mostra muito útil, pois apresenta valores de erro médio nas mesmas unidades dos valores preditos. Essa característica permite realizar comparações e analisar limiares de erros aceitáveis no domínio de aplicação do problema. A métrica pode ser calculada pela Equação 3.17,

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.17}$$

  onde y é o valor real, $\hat{y}$ é o valor predito considerando as n instâncias i.

Além disso, analisou-se os domínios de aplicação do melhor modelo de regressão obtido através de um gráfico de William (*Williams plot*).

- **Gráfico de William:** Consiste em um método amplamente aplicado para avaliar domínios de aplicação que fornece valores de leverage dispostos contra os erros de previsão (80).

  O valor de leverage para uma instância i (h_i) mede a distância entre uma nova instância i e o centroide das instâncias utilizadas no treinamento. Podemos calculá-lo pela equação

$$h_i = x_i^T.(X^T X)^{-1} x_i, \tag{3.18}$$

  onde X é a matriz descritora do conjunto de treinamento (80). É comumente considerado um leverage de alerta o valor dado por

$$h^* = 3\frac{(p+1)}{n}, \tag{3.19}$$

  onde p é o número de descritores moleculares e n é o número de amostras de treinamento. Se uma nova instância de hospedeiro-ligante possui um leverage maior que o limite h^*, pode-se considerar o valor obtido pela predição como não confiável. Dizemos assim que a molécula está fora do espaço definido pelas instâncias no treinamento, ou seja, estão fora do domínio de aplicação do modelo.

4 RESULTADOS E DISCUSSÃO

O presente capítulo apresenta os resultados obtidos no decorrer da pesquisa, assim como a discussão da interpretabilidade e aplicabilidade dos modelos desenvolvidos. Primeiramente na Seção 4.1 descreve-se a base de dados obtida pelo processo descrito no capítulo anterior, proporcionando os materiais necessários para que pesquisas correlatas e trabalhos futuros sejam desenvolvidos a partir dos resultados atuais. Em seguida, a Seção 4.2 apresenta os resultados obtidos do treinamento e avaliação dos modelos regressivos considerados neste trabalho, assim como a comparação entre os valores médios atingidos por cada estratégia regressiva. A Seção 4.3 apresenta análises adicionais sobre o método regressivo que atingiu os melhores resultados, assim como a discussão sobre a qualidade e interpretabilidade das métricas de erro consideradas. Por fim, a Seção 4.4 apresenta a produção científica resultante do desenvolvimento deste projeto de dissertação no mestrado acadêmico em Modelagem Computacional.

4.1 DESCRIÇÃO E EXPLORAÇÃO DA BASE DE DADOS

Essa seção se destina a apresentar os resultados relativos ao tratamento da base de dados, a padronização da nomenclatura, Tabela 3,utilizada nas discussões posteriores, análises exploratórias sobre a distribuição dos dados e o resultado do processo de separação da base inicial em conjuntos de treinamento e teste.

A Tabela 3 mostra informações descritivas sobre a natureza das variáveis de entrada (X_i) e saída (y) consideradas no trabalho. Os resultados e dados divulgados a partir dessa pesquisa apresenta essa padronização. As variáveis de entrada são divididas em três grupos de acordo com o valor de i: (i) $1 \leq i \leq 2$ são variáveis de experimentais de ambiente de interação; (ii) $3 \leq i \leq 12$ são variáveis da molécula hóspede (ligante); e (iii) $13 \leq i \leq 21$ são variáveis da molécula hospedeira.

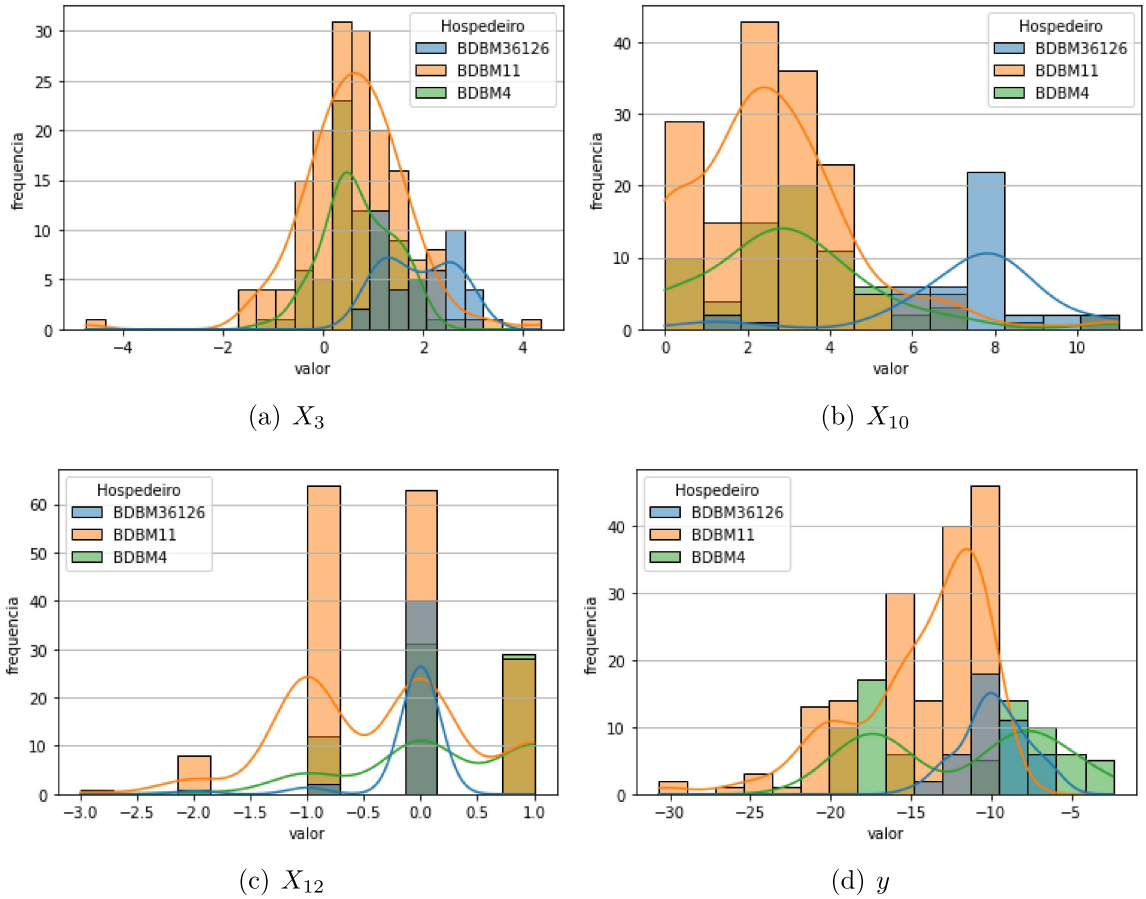
A Figura 18 mostra a distribuição de alguns descritores das instâncias da base agrupados pelo tipo de molécula hospedeira (identificadores na Tabela 1). É perceptível que há uma distinção entre os valores dos descritores das moléculas que se ligam a diferentes classes de ciclodextrinas. Por exemplo, a base de dados apresenta a tendência de moléculas α-CD e β-CD de se interagirem com ligantes com menores valores de X_3 e X_{10}, se comparados com as moléculas γ-CD. Com relação a variável objetivo y, hospedeiras γ-CD apresentam uma distribuição aproximadamente normal, β-CD apresentam um comportamento normal deslocado com alongamento de calda da distribuição para valores mais negativos, enquanto α-CD apresentaram um comportamento bimodal.

Além disso, como o número de instâncias com cada classe de hospedeiro é desbalanceado, justifica-se a aplicação de métodos de separação de conjuntos apropriados que mantenham as mesmas proporções e distribuições aproximadas dos conjuntos. Essas ca-

Tabela 3 – Variáveis de entrada e variáveis objetivo da base de dados com suas respectivas descrições.

	Variável	Descrição
X_1	**pH**[1]	pH da solução
X_2	**Temp**[1]	temperatura do ambiente de solução
X_3	**SlogP**[2]	SlogP (*log* do coeficiente de partição octanol/água) da molécula hóspede
X_4	**SMR**[2]	Refratividade da molécula hóspede
X_5	**LabuteASA**[2]	Aproximação para a área da superfície acessível ao solvente da molécula hóspede
X_6	**TPSA**[2]	Aproximação para a área da superfície polar da molécula hóspede
X_7	**AMW**[2]	Peso médio molecular da molécula hóspede
X_8	**NumLipinskiHBA**[2]	Número de receptores em ligações de hidrogênio da molécula hóspede
X_9	**NumLipinskiHBD**[2]	Número de doadores em ligações de hidrogênio da molécula hóspede
X_{10}	**NumRotatableBonds**[2]	Número de ligações rotáveis na molécula hóspede
X_{11}	**NumAtoms**[2]	Número de átomos da molécula hóspede
X_{12}	**Formal Charge**[2]	Carga formal da molécula hóspede
X_{13}	**SlogP**[3]	SlogP (*log* do coeficiente de partição octanol/água) da molécula hospedeira
X_{14}	**SMR**[3]	Refratividade da molécula hospedeira
X_{15}	**LabuteASA**[3]	Aproximação para a área da superfície acessível ao solvente da molécula hospedeira
X_{16}	**TPSA**[3]	Aproximação para a área da superfície polar da molécula hospedeira
X_{17}	**AMW**[3]	Peso médio molecular da molécula hospedeira
X_{18}	**NumLipinskiHBA**[3]	Número de receptores em ligações de hidrogênio da molécula hospedeira
X_{19}	**NumLipinskiHBD**[3]	Número de doadores em ligações de hidrogênio da molécula hospedeira
X_{20}	**NumRotatableBonds**[3]	Número de ligações rotáveis na molécula hospedeira
X_{21}	**NumAtoms**[3]	Número de átomos da molécula hospedeira
y	ΔG[4]	Energia de ligação livre (afinidade de ligação)

[1] Variáveis experimentais de ambiente. [2] Descritores das moléculas hóspedes (ligantes). [3] Descritores das moléculas hospedeiras (ciclodextrinas). [4] Variável objetivo.



(a) X_3 (b) X_{10}

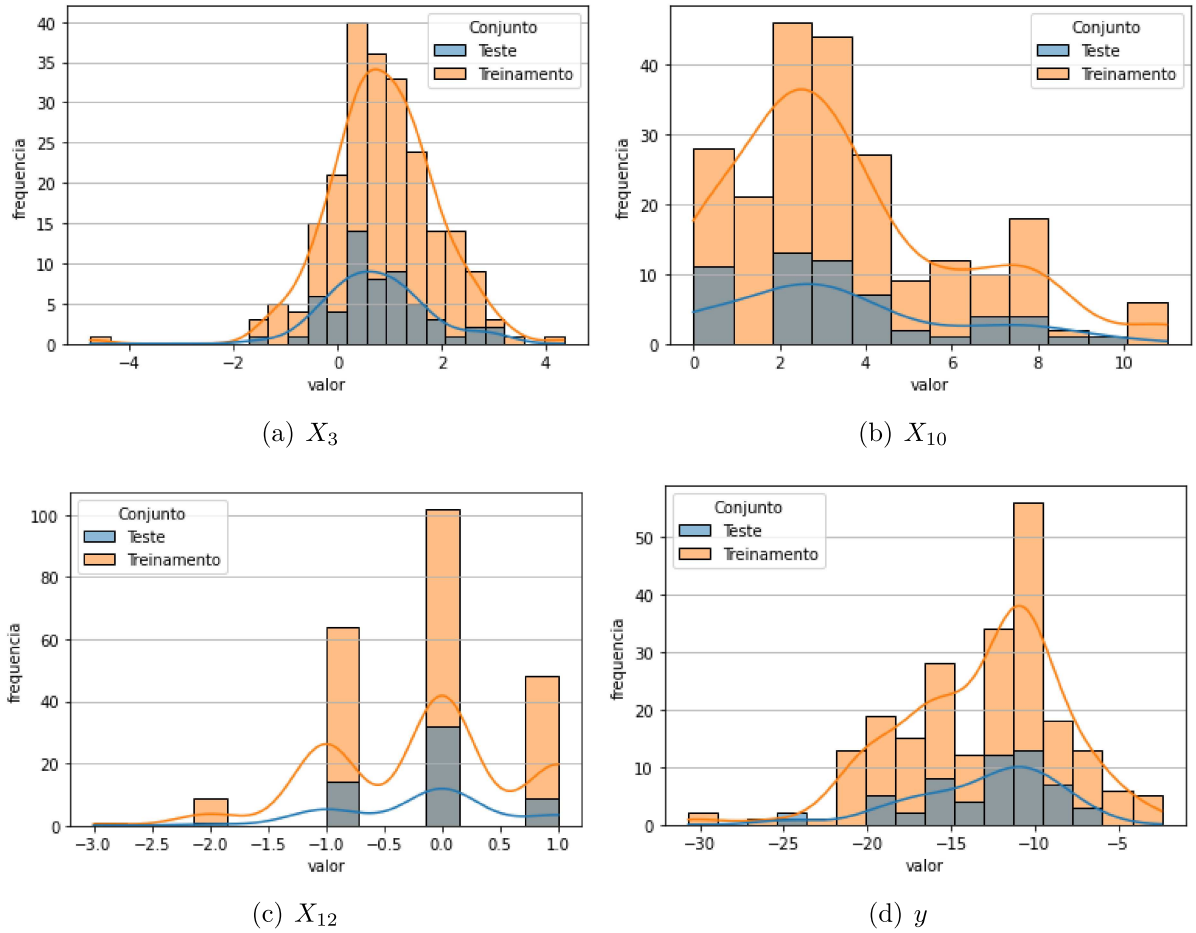
(c) X_{12} (d) y

Figura 18 – Apresentação da distribuição dos valores de descritores para a molécula hóspede (a), (b), e (c), e ΔG em (d) agrupadas por tipo de molécula hospedeira no sistema.

racterísticas foram levadas em consideração para a definição dos conjuntos de treinamento e teste. A Figura 19 apresenta as distribuições resultantes da aplicação dos processos acoplados de *Stratified K-fold* e *Kullback-Leibler* para o controle das proporções de instancias com moléculas hóspedes e distribuições de descritores nos conjuntos de treinamento e teste. Pode-se observar a similaridade das distribuições dos conjuntos, o que dificilmente seria obtida por um processo puramente randomizado de seleção de instancias. Esse mecanismo nos permite avaliar com mais assertividade a aplicação dos métodos treinados nos domínios de aplicação definidos.

Adicionalmente, as Tabelas 9 e 10, dispostas no Anexo A, mostram estatísticas descritivas de tendência central para cada uma das variáveis das intrâncias dos conjuntos de treinamento e teste, respectivamente. Para os descritores relativos aos hospedeiros, as métricas são calculadas sobre um conjunto de apenas três valores, uma vez que apenas três tipos de ciclodextrinas são consideradas no processo. O mesmo acontece com as métricas de ambiente, uma vez que elas foram controladas em um pequeno intervalo definido na Seção 3.1.1.

Com relação aos descritores dos ligantes nas Tabelas 9 e 10, pode-se observar uma



(a) X_3 (b) X_{10}

(c) X_{12} (d) y

Figura 19 – Apresentação das distribuições dos conjuntos de treinamento e teste para os valores de descritores para a molécula hóspede (a), (b), e (c), e ΔG em (d).

variabilidade dos valores considerados. De forma complementar, o Anexo B, apresenta uma forma alternativa de se analisar a variabilidade das moléculas hóspedes dispostas na base. É proposto um moledo de redução de dimensionalidade da base descritores dos hóspedes e agrupamento dos dados por modelo não supervisionado para a avaliação da dispersão dos descritores das moléculas hóspedes em um espeço de tamanho reduzido. O objetivo é conhecer melhor o espaço de instâncias utilizadas na pesquisa.

Essa linha de exploração por meio da transformação da dimensão dos atributos é introduzido nesse contexto, mas deverá ser melhor explorado em trabalhos futuros, contudo, entende-se que haja perspectiva nessa linha de avaliação e manipulação da base de dados. Essa linha de investigação corrobora com o que é apresentado na Figura 20, que mostra os valores encontrados de correlação entre os atributos descritores das moléculas ligantes e a variável objetivo considerada no presente trabalho. Os dados mostram que alguns dos atributos descritores das moléculas ligantes possuem elevado grau de correlação entre si, ao mesmo tempo que apresentam baixos níveis de correlação com a variável objetivo do sistema considerado.

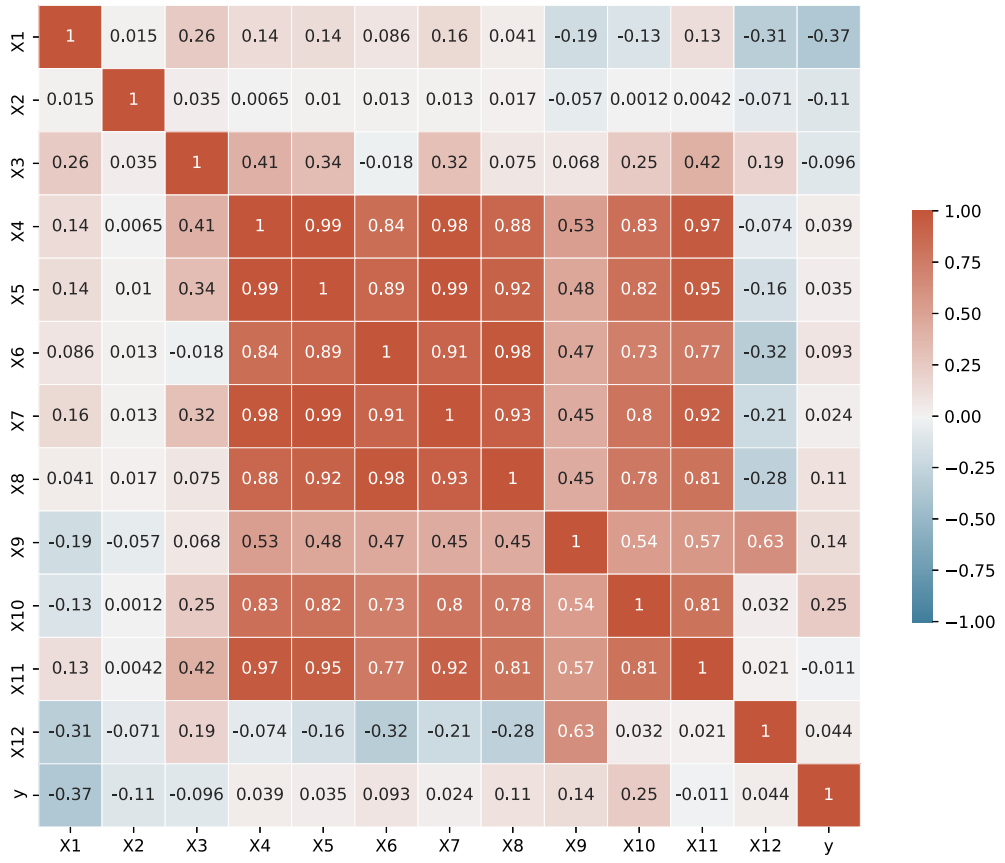


Figura 20 – Mapa de calor das correlações encontradas entre os atributos descritores das moléculas ligantes e a variável objetivo considerada no presente trabalho.

Fonte: Produzido pelo autor (2021).

4.2 COMPARAÇÃO ENTRE OS RESULTADOS DOS MÉTODOS DE APRENDIZADO DE MÁQUINA

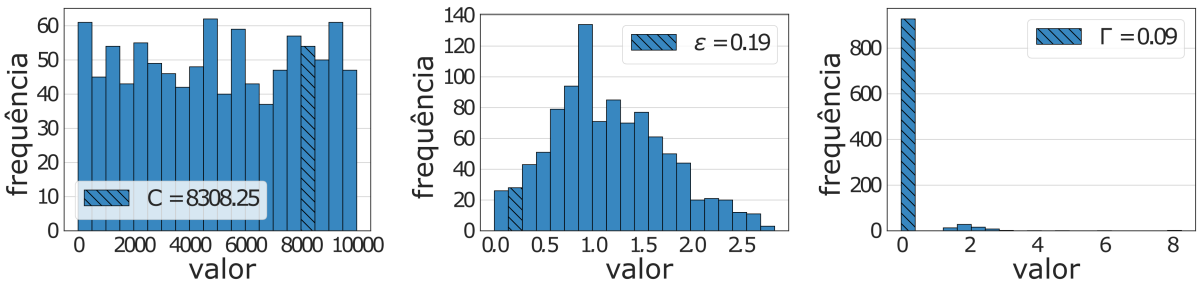
Essa seção se destina a apresentar os resultados obtidos com a aplicação dos métodos de aprendizado de máquina no problema tratado, desde o processo de ajuste de hiperparâmetros do modelos e ajuste dos parâmetros internos na etapa de treinamento, até os resultados médios obtidos com os conjuntos de testes definidos e a comparação dos resultados para cada estratégia.

Primeiramente, apresenta-se os resultados relacionados à busca por hiperparâmetros através da estratégia de Busca Randomizada (RS). Cada execução de RS tem como

objetivo encontrar uma máquina ajustada (um modelo) com um RMSE de treinamento associado. Nas análises aqui performadas, considera-se o modelo melhor otimizado a máquina associada ao menor RMSE no conjunto de treinamento. Abaixo apresenta-se a distribuição dos parâmetros otimizados selecionados nas mil (1000) execuções de RS para cada uma das estratégias ML consideradas. Os parâmetros SVR são mostrados na Figura 21, os parâmetros GPR aparecem na Figura 22, enquanto a Figura 23 exibe os parâmetros XGB. A barra hachurada indica a faixa de valores que envolve o modelo melhor otimizado (menor RMSE de treinamento) entre as execuções RS.

Note que para SVR, Figura 21, mesmo que o espaço de busca de ε tenha sido definido em uma faixa maior, os melhores modelos ficaram sempre concentrados no intervalo $\varepsilon \leq 3$. Um resultado semelhante é observado para Γ, uma vez que essencialmente converge para valores mais próximos de zero. O parâmetro `C` apresentou uma apresentação aproximadamente uniforme, o que mostra que a obtenção de bons resultados pode ocorre independente do valor escolhido para essa variável, uma vez que os demais parâmetros estejam ajustados para a dada situação.

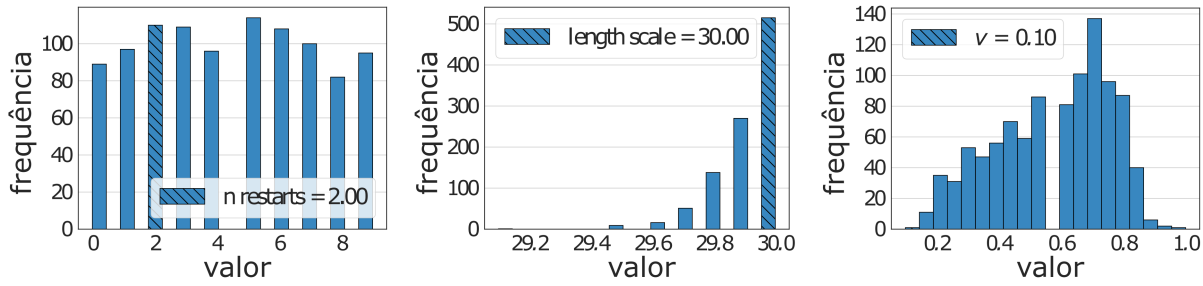
Para o GPR, nenhum dos melhores resultados utilizou a seleção de *kernel* `RBF`. Na Figura 22, ainda para o para GPR, nota-se que a medida de `length scale` aumenta no kernel `Matern` enquanto ν diminui nos melhores resultados. O parâmetro `n restarts` apresentou uma distribuição aproximadamente uniforme, o que demonstra pouca interferência desse parâmetro na obtenção dos de melhores resultados.



(a) `C` (b) ε (c) Γ

Figura 21 – Melhores hiperparâmetros para SVR obtidos nas 1000 execuções do processo de busca RS. A barra hachurada indica o intervalo de valores que contempla o valor do melhor resultado obtido.

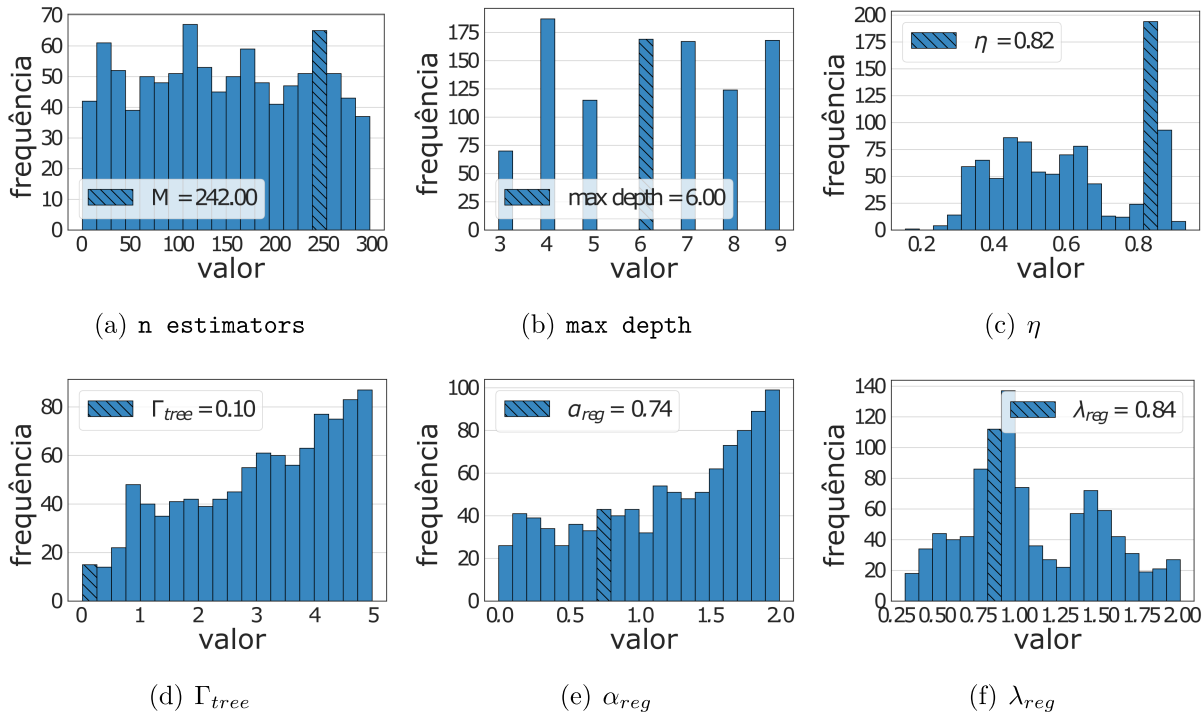
Analisando as distribuições obtidas nas execuções do XGB, observa-se comportamentos parecidos com os anteriores, contudo esse método possui um maior número de hiperparâmetros para serem ajustados. As variáveis `n estimators` e `max depth` apresentam comportamento aproximadamente uniforme. As variáveis η e λ_{reg} apresentam os melhores resultados nas proximidades dos valores mais frequentes entre as melhores execuções. Já as as variáveis Γ_{tree} e α_{reg} parecem apresentar uma tendencia a valores maiores entre as melhores máquinas, contudo, na melhor máquina os valores escolhidos não se



(a) `n restarts` (b) `length scale` (c) ν

Figura 22 – Melhores hiperparâmetros para GPR obtidos nas 1000 execuções do processo de busca RS. A barra hachurada indica o intervalo de valores que contempla o valor do melhor resultado obtido.

encontravam na extremidade de maior magnitude das distribuições, inclusive apresentando Γ_{tree} na extremidade inferior das distribuições.



(a) `n estimators` (b) `max depth` (c) η

(d) Γ_{tree} (e) α_{reg} (f) λ_{reg}

Figura 23 – Melhores hiperparâmetros para XGB obtidos nas 1000 execuções do processo de busca RS. A barra hachurada indica o intervalo de valores que contempla o valor do melhor resultado obtido.

A Tabela 4 mostra os valores médios para as métricas de avaliação obtidos das melhores máquinas de cada execução do RS sobre os conjuntos de treinamento e testes para cada abordagem ML. Os resultados demonstram a consistência dos métodos, já que os desvios entre as execuções não apresentam grande dispersões de qualidade de resultados. O GPR alcançou melhores resultados em comparação com os outros métodos.

Tabela 4 – Valores médios para as métricas de erro (média ± desvio padrão) sobre as 1000 execuções do processo de busca RS.

	Método	R^2 score	RMSE (kJ/mol)	MAE (kJ/mol)
Treinamento	**SVR**	0,881 ± 0,040	1,621 ± 0,256	1,188 ± 0,399
	XGB	0,894 ± 0,027	1,537 ± 0,189	1,002 ± 0,170
	GPR	**0,943 ± 0,000**	**1,132 ± 0,004**	**0,421 ± 0,005**
Teste	**SVR**	0,727 ± 0,041	2,127 ± 0,154	1,692 ± 0,220
	XGB	0,719 ± 0,078	2,142 ± 0,285	1,567 ± 0,156
	GPR	**0,755 ± 0,020**	**2,015 ± 0,082**	**1,270 ± 0,028**

Tabela 5 – Melhor valor obtido por cada método entre as 1000 execuções do processo de busca RS. RMSE e MAE são medidos em kJ/mol.

Método	Treinamento			Teste		
	R^2	RMSE	MAE	R^2	RMSE	MAE
SVR	0,922	1,331	0,505	0,776	1,932	1,351
XGB	0,940	1,163	0,558	0,688	2,277	1,468
GPR	**0,953**	**1,034**	**0,396**	**0,803**	**1,811**	**1,201**

Tabela 6 – Conjunto dos valores dos parâmetros otimizados para cada método de aprendizado de máquina obtidos nas 1000 execuções de busca por RS.

Método	Hiperparâmetros
SVR	`C`: 8308,25; ε:0,19; Γ:0,09
XGB	`M`: 242; `max_depth`: 6; η: 0,82; Γ_{tree}: 0,10; α_{reg}: 0,74; λ_{reg}: 0,84
GPR	`n_restarts`: 2; `Matern`(σ: 30; ν: 0,10)

A Tabela 5 mostra as métricas de avaliação para os modelos melhor otimizados de cada abordagem ML alcançado pela estratégia RS. A Tabela 6 apresenta o conjunto de hiperparâmetros otimizados associados a cada modelo melhor otimizado. Novamente, notamos melhores resultados de GPR em cada métrica. Constatando isso, as análises subsequentes serão focadas no estudo dessa abordagem.

4.3 AVALIAÇÃO, INTERPRETAÇÃO E DISCUSSÃO SOBRE O MELHOR RESULTADO

Essa seção se destina a apresentar os resultados discretizados obtidos com a melhor máquina obtida na Seção 4.2: abordagem GPR. Aqui apresenta-se um estudo sobre os erros do modelo, umas discussão sobre a interpretabilidade dos resultados e do domínio de aplicação da máquina obtida.

A Figura 24 mostra a comparação entre os valores de afinidade de ligação experimental e prevista obtida pelo melhor modelo GPR, discretizados pelas classes Carga Formal do Ligante (*Ligand Formal Charge* - FC), número de Ligações Rotáveis do Ligante

(*Ligand Ratatable Bonds* - RB), e o tipo de ciclodextrina hospedeira na instância (CD).

A Figura 24(a)-(c) mostra a alta adaptabilidade do modelo GPR aos dados de treinamento ($R^2 = 0,953$) junto com as demais métricas de avaliação. As classes escolhidas para a discretização dos resultados indicam atributos que frequentemente enganam as previsões em modelos tradicionais. A definição de cada classe é descrita na Tabela 7.

A Figura 24(d)-(f) mostra a capacidade de previsão do modelo no conjunto de teste. A capacidade de generalização do modelo GPR é evidente de acordo com os valores das métricas de desempenho apresentadas na Tabela 4. Em comparação com o desempenho do modelo obtido no conjunto de treinamento, o aumento dos erros de previsão para o conjunto de testes era esperado. Entretanto, observa-se que não houve *overfitting*, uma vez que a diminuição do desempenho no conjunto de teste não é significativo em comparação com o de treinamento, demonstrando o bom desempenho da estratégia em um conjunto de amostras não vistas na etapa de aprendizado.

A Figura 24 explícita a ocorrência de pelo menos uma previsão *outlier* no conjunto de teste (amostra com maior erro). A instância relacionada a esta previsão é a interação entre o hospedeiro α-CD e o hóspede *1-butylamine* – `CCCC[NH3+]`. O $\Delta G_{real} = -17,124$ e o $\Delta G_{pred} = -8,095$. Neste caso particular, os modelos SVR e XGB previram $\Delta G_{pred} \sim -13,5$, que ainda caracteriza um erro elevado, mas mais próximo do valor real do que o GPR.

A Tabela 7 mostra a média geral da métrica RMSE (sobre as 1000 execuções RS) para cada separação de instância: Carga Formal do Ligante (FC), número de Ligações Rotáveis do Ligante (RB), e o tipo de hospedeiro na instância (CD). De fato, há oscilações nas medidas previstas entre as classes. Embora a diferença no RMSE não ultrapasse $\sim 0,8$ kJ/mol, ela deve ser levada em conta antes de aplicar o método ML sobre novos conjuntos de dados.

Para evoluir a discussão sobre os resultados apresentados anteriormente, é necessário verificar se os níveis de erro são compatíveis com o domínio de aplicação de interações moleculares. Neste contexto, as interações de potencial eletrostático, van der Waals, pontes salinas e ligações de hidrogênio (HB) são formas predominantes de interação. Entre elas, as ligações de hidrogênio e as interações eletrostáticas são determinantes nos complexos receptor-ligante (81).

Comparativamente, ligações HB são mais fracas que as interações eletrostáticas, mas extremamente mais frequentes. A literatura destaca que as conexões HB têm um potencial de interação sempre inferior a 10 kJ/mol; na grande maioria dos casos, elas são inferiores a 3 kJ/mol (81), (82). Em complexos proteicos, por exemplo, as contribuições de afinidade por conta da formação HB são de 5 $\pm$ 2,5 kJ/mol (81). Assim, o nível de erro nas previsões pode ser relacionado com o número de HB inseridos na faixa não prevista, ou seja, o número de ligações de hidrogênio que podem ser justificadas pelos valor de erro

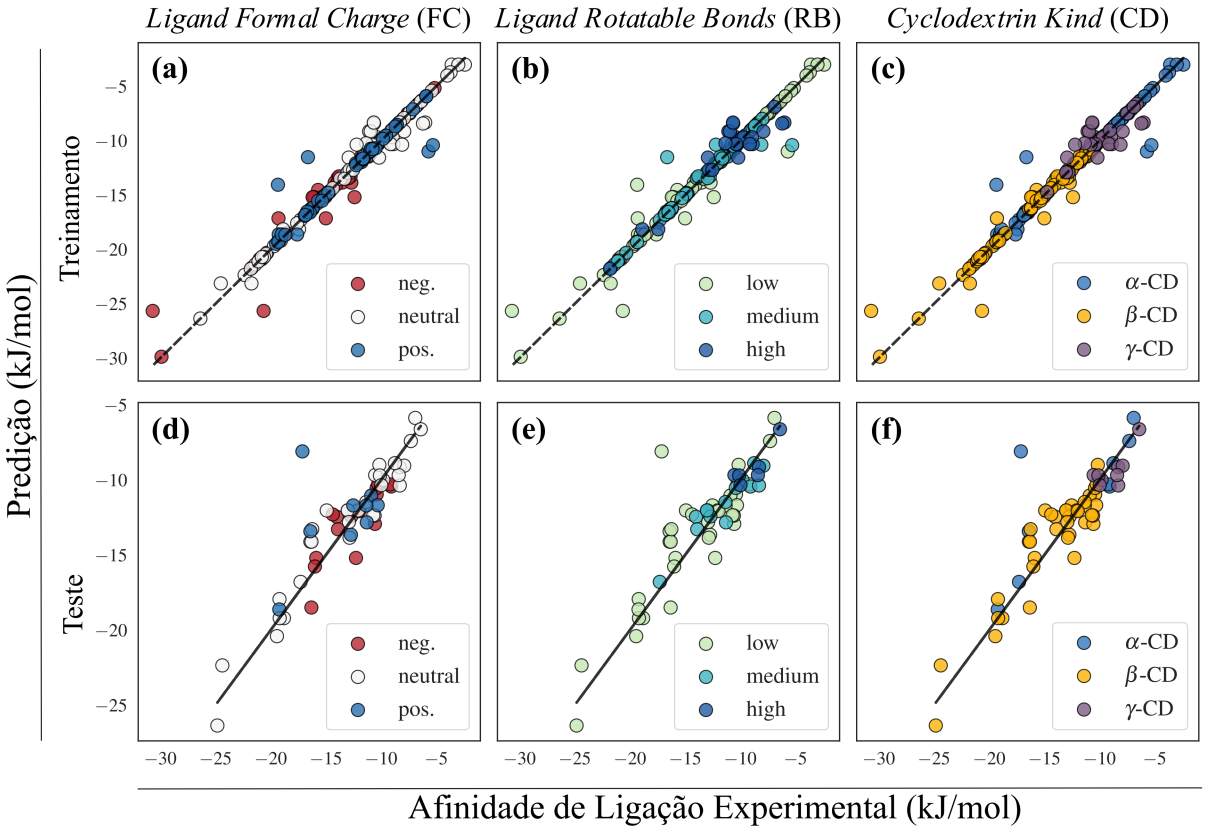


Figura 24 – Predições resultantes do melhor modelo GPR obtido. As subfiguras (a)-(c) apresentam o ajuste do modelo aos dados de treinamento ($R^2 = 0{,}953$, RMSE = 1,034 e MAE = 0,396). As subfiguras (d)-(f) apresentam a capacidade preditiva do modelo no conjunto de teste ($R^2 = 0{,}803$, RMSE = 1,811 e MAE = 1,201). Cada classe de instâncias consideradas (FC, RB, e CD) são descritas na Tabela 7.

Tabela 7 – Valores de RMSE em kJ/mol (mean ± std) para cada subconjunto de dados com relação a Carga Formal do Hóspede (FC), Número de Ligações Rotáveis do Hóspede (RB), e Classe de Hospedeiro (CD). Considera-se um RB baixo o intervalo $[0, 3]$, RB médio o intervalo $[4, 7]$, e um RB alto o intervalo $[8, 11]$. Para esses cálculos, desconsiderou-se as instâncias que geraram predições *outliers*.

Subconjunto		**Métrica RMSE**		
FC	**Classes**	**Neg.**	**Neutro**	**Pos.**
	Treinamento	1,017 ± 0,001	0,708 ± 0,001	1,821 ± 0,011
	Teste	1,564 ± 0,026	1,623 ± 0,114	1,196 ± 0,025
RB	**Classes**	**Baixo**	**Médio**	**Alto**
	Treinamento	1,122 ± 0,005	1,132 ± 0,003	1,180 ± 0,002
	Teste	1,760 ± 0,096	1,135 ± 0,024	0,999 ± 0,025
CD	**Classes**	**α-CD**	**β-CD**	**γ-CD**
	Treinamento	1,585 ± 0,009	0,801 ± 0,000	1,154 ± 0,001
	Teste	1,495 ± 0,115	1,646 ± 0,078	1,136 ± 0,015

cometido pelos modelos regressivos. Para este trabalho, um erro de até uma ligação de hidrogênio não predita é considerado aceitável.

Entre as métricas de avaliação calculadas, as métricas RMSE e MAE indicam um erro médio mantendo a mesma medida que nossa variável objetivo (kJ/mol). Analisando o melhor GPR com relação a RMSE e MAE nas etapas de treinamento (RMSE = 1,034 e MAE = 0,396) e teste (RMSE = 1,811 e MAE = 1,201), observa-se que os valores permanecem seguramente abaixo do limiar de 5 kJ/mol. O mesmo é válido para a análise média apresentada na Tabela 4. Nessas análises, verifica-se um erro médio de apenas uma ligação de hidrogênio média entre o valor experimental de afinidade (ΔG) e o previsto pelo modelo.

Outra forma de verificar o domínio de aplicação do melhor modelo GPR é através do gráfico Williams, apresentado na Figura 25. A maioria dos complexos nos conjuntos de treinamento e testes estão dispostos no interior da área confiável, indicando que a interação entre esses compostos são provavelmente são melhores previstos pelo modelo, e portanto o modelo funcionaria bem para prever instâncias com características descritivas (descritores moleculares) similares.

Contudo, existem quatro compostos da etapa de treinamento com valores de alavancagem superiores ao limite de h^*. Eles diferem das outras instâncias em valores zerados para as variáveis X_4, X_6 ou X_{10} relativas às moléculas hóspedes, como descrito na Tabela 3. Estas instâncias podem influenciar o desempenho do modelo no conjunto de treinamento, mas não são necessariamente *outliers* a serem excluídos do conjunto de dados de treinamento, uma vez que seus valores residuais padrão são relativamente baixos e dentro do limite estabelecido. Em geral, os resíduos padronizados são menores ou no limite de nossa região segura dada por:

$$\varepsilon_{std} = \pm\frac{5KJ/mol}{\sigma_{resid}} \sim \pm 4,07 \tag{4.1}$$

onde σ_{resid} é o desvio padrão dos resíduos, exceto por uma instância do conjunto de teste (instância #10). Esta instância já foi identificada na Figura 24(d)-(f). Uma vez que seus descritores não ultrapassem o valor h^*, esse pode ser considerado como o pior cenário possível em predição verificado.

Os dados rotulados na Figura 25 são instâncias dispostas no limite ou fora da região de segurança do Gráfico de William. A Tabela 8 relaciona o identificador BindingDB da instância, do hospedeiro e do hóspede com os respectivos valores calculados de h e com os resíduos padronizados obtidos nas predições do melhor modelo GPR.

Finalmente, compara-se os resultados obtidos com alguns trabalhos apresentados na literatura. Note que esta comparação não é sobre os resultados com base nos mesmos conjuntos de dados. A seguir, procura-se apenas demonstrar que, para a base de dados atual, os níveis de erro são compatíveis com o que é frequentemente apresentado na

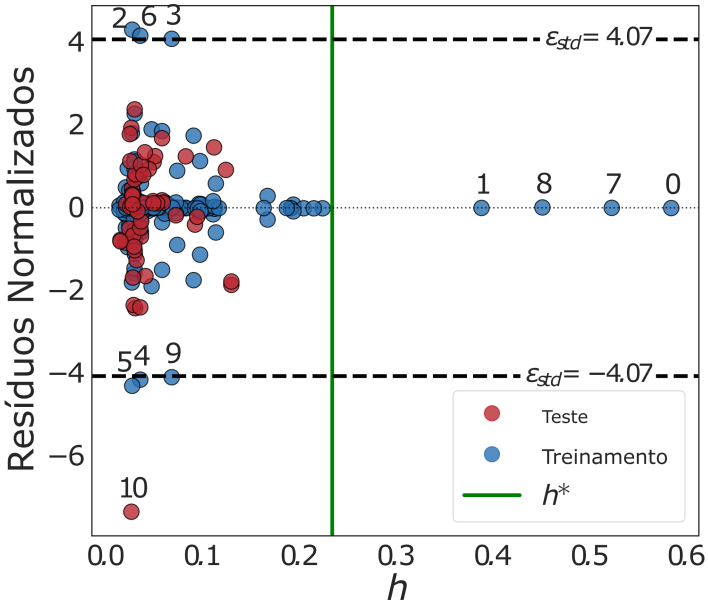


Figura 25 – Gráfico de Williams para o melhor modelo GPR obtido conforme descrito na Subseção 3.2.7.

Tabela 8 – Instâncias relacionadas a predições *outliers* ou de borda de acordo com os resultados obtidos pelo gráfico de Williams para o melhor modelo GPR obtido.

#	Instância	Hospedeiro	Hóspede	h	Resíduo Normalizado
0	**6_22**	BDBM11	BDBM5	0,586	-0,008
1	**3532_56**	BDBM11	BDBM36131	0,390	-0,005
2	**3529_16**	BDBM4	BDBM36019	0,029	4,303
3	**3535_4**	BDBM11	BDBM50029034	0,070	4,083
4	**3529_7**	BDBM11	BDBM36022	0,037	-4,157
5	**3529_14**	BDBM4	BDBM36018	0,029	-4,305
6	**3529_5**	BDBM4	BDBM36012	0,037	4,158
7	**3532_55**	BDBM11	BDBM36130	0,524	-0,003
8	**6_2**	BDBM4	BDBM5	0,453	0,006
9	**3535_5**	BDBM11	BDBM36146	0,070	-4,097
10	**3540_35**	BDBM4	BDBM36198	0,028	-7,349

literatura.

Dimas *et al.* (36) selecionou um conjunto de complexos entre β-CD e 57 pequenas moléculas orgânicas que foram previamente estudadas com o método de análise de distribuição de energia de ligação em combinação com um modelo de solvente implícito. Mesmo sendo um estudo focado apenas nos β-CD e aplicando um método baseado em física, os níveis de erros foram $R^2 = 0,66$ e RMSE $= 9,330$ kJ/mol, valores que são piores do que os obtidos pelas métricas médias da Tabela 4 e da Figura 24(d)-(f).

Solovev et al. (42) relatam o uso de novos descritores de fragmentos 3D para modelar parâmetros e propriedades de moléculas estereoisoméricas. Foi aplicado uma implementação de Decomposição em Valores Singulares para análise de descritores e

Regressão Linear Múltipla como método de aprendizagem de máquina. Nas predições de energia livre da complexação de 76 hóspedes quirais com *beta*-cyclodextrin (β-CD) foram obtidas as métricas $R^2 = 0,918$ e $RMSE = 0,89$ kJ/mol e nas predições de 40 hóspedes quirais com 6-amino-6-deoxy-β-cyclodextrin (am-β-CD) foram obtidas as métricas $R^2 = 0,910$ e $RMSE = 0,89$ kJ/mol. Ambos valores obtidos são extremamente promissores, contudo todos os testes são executados com modelos especializados para a classe de ciclodextrina em questão e produzidos sobre um conjunto de dados de tamanho bastante reduzido.

Xu et al. (41) apresenta um estudo preditivo sobre hospedeiros β-CD e um conjunto diverso de hóspedes orgânicos. O conjunto de dados disponível continha 218 compostos, dividido em um conjunto de treinamento de 160 compostos e um conjunto de teste de 58 compostos por meio da aplicação do algoritmo `DUPLEX`. O trabalho aplica Regressão Linear Múltipla obtendo métricas de erro $R^2 = 0,833$ e $MAPE = 1,911$ kJ/mol, além de uma arquitetura de Rede Neuronal Artificial obtendo métricas de erro $R^2 = 0,957$ e $MAPE = 0,925$. A aplicação não linear de Redes Neuronais apresentaram bons resultados, contudo o modelo é especializado apenas para aplicação de complexos compostos por moléculas β-CD.

Zhao *et al.* (1), por outro lado, aplicou métodos ML para prever a afinidade de ligação entre um vasto número de classes de ciclodextrina e diversas moléculas hóspedes. Os testes preditivos são realizados utilizando metodologias de Floresta Randômica, Aprendizado Profundo e Métodos de *Boosting*. O conjunto de dados utilizados na pesquisa é o maior encontrado nas pesquisas realizadas durante esse trabalho, contudo o conjunto não foi disponibilizado por questões de propriedade intelectual, segundo o próprio autor do trabalho. O melhor resultado obtido para Zhao foi baseado na abordagem *Light Gradient Boosting Machine* (LightGBM), semelhante à abordagem XGB aplicada nesse trabalho, com métricas de erro $R^2 = 0,86$, RMSE $= 1,83$ kJ/mol, e MAE $= 1,38$ kJ/mol. Comparando com os resultados obtidos nessa pesquisa para o modelo XGB apresentado na Tabela 5, percebe-se que as execuções atingiram resultados competitivos, considerando o menor tamanho de banco de dados da pesquisa atual. Entretanto, considerando os resultados para as execuções do XGB, é possível identificar a situação atual como uma situação de *overfitting*, já que a etapa de treinamento obteve métricas erro consideravelmente melhores do que as obtidas na etapa de teste, veja a Tabela 5. Isso acontece uma vez que o espaço do hiperparâmetro do XGB tem muitos graus de liberdade. Assim, geralmente sua aplicação exige grandes tamanhos de conjunto de dados para atingir seus níveis ótimos de generabilidade. Análises futuras podem abordar o espaço de hiperparâmetros de busca XGB, reduzindo possíveis valores que impulsionem a solução para um ótimo local com ocorrência de *overfitting*. Podemos indicar o hiperparâmetro taxa de aprendizagem (η) (Figura 23(c)) como um possível originador de *overfitting*. Nossa melhor máquina indica um alto valor neste parâmetro ($\eta = 0,82$), indicando a possibilidade de uma solução com viés

e uma alta tendência de acumulação de informações específicas dos dados de treinamento. Nesses casos, o método não obtém soluções generalizadas para serem aplicados em novos dados não vistos na etapa de teste, como desejamos. Através de nossos resultados na Figura 23(c), talvez seja o caso de realizar buscas futuras por melhores soluções para o XGB quando $\eta \leq 0,6$, ou com um valor de corte ainda menor.

Ao comparar os resultado de (1) *et al.* com melhor modelo GPR apresentado na Tabela 5 e na Figura 24(d)-(f), vemos que foi possível atingir melhores valores em questão de MAE e RMSE. Além disso, o GPR é um método de menor complexidade computacional em comparação com o XGB. Na hipótese de consideração da instância *outilier* apresentada anteriormente, as métricas são ainda melhores em valores de erro médio. Entretanto, os valores para a métrica R^2 não puderam ser ajustados com os mesmos valores apresentados nos resultados de Zhao *et al.* (1).

A critério de resumo das informações apresentadas na discussão dos últimos parágrafos, a Tabela 11, disposta no Anexo C desse trabalho, apresenta uma alternativa direta para a análise dos resultados discutidos.

4.4 PRODUÇÃO CIENTÍFICA

Durante o desenvolvimento da atual pesquisa e da formação do discente no mestrado acadêmico em modelagem computacional, pode-se listar algumas produções científicas ligadas diretamente ou indiretamente ao desenvolvimento da linha de pesquisa atual.

Produções diretamente relacionadas à pesquisa apresentada nesse relatório:

- Publicação na revista internacional *Journal of Inclusion Phenomena and Macrocyclic Chemistry* do trabalho intitulado "*Gaussian processes regression for cyclodextrin host-guest binding prediction*" (16).
- Apresentação de trabalho no congresso internacional *Research and Applications in Artificial Intelligence* (RAAI, 2020) e publicação do trabalho intitulado "*Prediction of Cyclodextrin Host-Guest Binding Through a Hybrid Support Vector Method*" (17).
- Trabalho final da disciplina Introdução à Ciência de Dados (PPGMC-219044) intitulado "Breve investigação sobre a capacidade de predição de afinidade de ligação entre moléculas em sistemas *Host-Guest* por meio de métodos de aprendizado de máquinas"no trimestre regular 2019-3.

Produções indiretamente relacionadas à pesquisa por meio da aplicação de métodos de aprendizado de máquina e manipulação de dados:

- Publicação do trabalho intitulado " *Hybrid Unsupervised Extreme Learning Machine Applied to Facies Identification*" (83).

- Apresentação no "*XLI Ibero-Latin-American Congress on Computational Methods in Engineering (CILAMCE)*" e publicação do trabalho intitulado "A comparative study of Singular Value Decomposition (SVD) and Discrete Cosine Transform (DCT) techniques for image compression applications" (84).
- Publicação do trabalho intitulado "Extensions and improvements of the Extreme Learning Machine (ELM) applied to face recognition" (85).
- Publicação em revista nacional do trabalho intitulado "Quality-Model Clustering Tool: a module for clustering protein models based on quality attributes" (86).
- Publicação do trabalho intitulado "ODS Mapeados: uma ferramenta computacional voltada à gestão acadêmica e científica para o apoio e alinhamento da Agenda 2030" nos anais do III SUSTENTARE e VI WIPIS (87).

5 CONCLUSÃO

Neste trabalho apresentou-se uma abordagem que combina métodos de aprendizagem de máquinas (ε-SVR, XGB e GPR), juntamente com uma estratégia de ajuste de hiperparâmetros por Busca Randomizada (RS), com o objetivo de prever a interação em sistemas hospedeiro-hóspede de ciclodestrinas e avaliar a qualidade das diferentes abordagens nessa tarefa.

Para isso, foi necessário um trabalho preliminar para compilar e expandir uma base de dados confiável que contivesse diferentes classes de moléculas de ciclodestrinas como hospedeiras e um conjunto bem conhecido de ligantes como hóspede. Os dados experimentais curados de afinidade de ligação molecular e condições de experimentos (pH e temperatura) foram obtidos por meio do banco de dados público *BidingDB* e os demais descritores moleculares foram calculados por meio de um *pipeline* construído no software *KNIME*. Todos os dados compilados foram disponibilizados, juntamente com dados referentes as separações de conjuntos de treinamento e teste efetuados no trabalho por meio da aplicação de métodos de amostragem estatística, o que proporciona a reprodutividade dos resultados e avanços futuros da linha de pesquisa.

Entre os métodos comparados o GPR alcançou os melhores resultados médios, se considerarmos as 1000 execuções realizadas para cada uma das estratégias, além de ter apresentado o melhor resultado em uma execução individual, garantindo o modelo treinado com melhor ajuste aos dados de entrada. A abordagem proposta foi suficiente para definir um modelo de GPR que reproduz predições com boa correlação com os dados experimentais ($R^2 = 0,803$) e baixos erros associados (RMSE $= 1,811$ e MAE $= 1,201$) para o domínio da aplicação do método. Os resultados são compatíveis com a literatura apresentada, ainda que utilizando um método menos complexo computacionalmente (1, 36, 42, 41).

Análises complementares sobre os resultados do melhor modelo GPR obtido mostraram a qualidade de predição para diferentes classes de instância da base de dados, considerando métricas de avaliação médias para agrupamento de instancias com mesmos valores de cargas formais dos ligantes, número de ligações rotáveis dos ligantes e tipo de molécula hospedeira. Na discussão dos resultados, as instâncias com piores predições e com composição de atributos descritivos mais distintos dos demais na base de dados foram identificados por meio do Gráfico de Willians. Juntas, essas análises apresentam o domínio de aplicação do modelo treinado, permitindo sua aplicação em novos dados de forma mais assertiva.

5.1 TRABALHOS FUTUROS

Diante das evoluções contantes em métodos de aprendizado de máquina e a possibilidade de utilização de conjuntos de moléculas mais variados e complexos, a presente

linha de pesquisa possibilita continuidade em diversas frentes:

- Realizar expansão sobre a diversidade de dados de moléculas hóspedes no conjunto de dados, a fim de amostrar diferentes naturezas de interações moleculares.
- Realizar expansão das análises sobre a utilização de diferentes moléculas hospedeiras, sendo elas da família das ciclodextrina ou de outras famílias de moléculas que podem compor sistemas hospedeiro-hóspede, como por exemplo cucurbit[n]urilas, calix[n]arenos e pilar[n]arenos.
- Realizar comparação dos resultados atuais com resultados de predição de interação molecular por meio de modelos específicos para cada classe de n-ciclodextrina considerada.
- Expandir a base de dados para permitir testes adicionais com métodos com alto grau de liberdade, como é o exemplo do XGB que pode estar enfrentando uma situação de *overfitting* nos testes aqui apresentados.
- Estudo sobre a expansão e/ou redução do espaço de descritores utilizados para descrever as variáveis de ambiente dos experimentos, as moléculas hóspedes e as moléculas hospedeiras.
- Aplicação e investigação de diferentes metodologias de métodos de aprendizado de máquina e comparação com os resultados atuais.
- Aplicação de abordagens evolutivas no processo de otimização de hiperparâmetros dos modelos, principalmente os que possuem espaços de busca com maior número de parâmetros.

REFERÊNCIAS

1 ZHAO, Q. et al. Predicting complexation performance between cyclodextrins and guest molecules by integrated machine learning and molecular modeling techniques. *Acta Pharmaceutica Sinica B*, Elsevier, v. 9, n. 6, p. 1241–1252, 2019.

2 KATSILA, T. et al. Computational approaches in target identification and drug discovery. *Computational and structural biotechnology journal*, Elsevier, v. 14, p. 177–184, 2016.

3 KUMAR, N. et al. Applying computational modeling to drug discovery and development. *Drug discovery today*, Elsevier, v. 11, n. 17-18, p. 806–811, 2006.

4 FORLI, S. et al. Computational protein–ligand docking and virtual drug screening with the autodock suite. *Nature protocols*, Nature Publishing Group, v. 11, n. 5, p. 905, 2016.

5 GUEDES, I. A. et al. New machine learning and physics-based scoring functions for drug discovery. *Scientific reports*, Nature Publishing Group, v. 11, n. 1, p. 1–19, 2021.

6 WANG, C.; ZHANG, Y. Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *Journal of computational chemistry*, Wiley Online Library, v. 38, n. 3, p. 169–177, 2017.

7 LU, J. et al. Incorporating explicit water molecules and ligand conformation stability in machine-learning scoring functions. *Journal of Chemical Information and Modeling*, ACS Publications, v. 59, n. 11, p. 4540–4549, 2019.

8 HAGHIGHATLARI, M. et al. Learning to make chemical predictions: the interplay of feature representation, data, and machine learning algorithms. *arXiv preprint arXiv:2003.00157*, 2020.

9 LI, H. et al. Machine-learning scoring functions for structure-based virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, Wiley Online Library, v. 11, n. 1, p. e1478, 2021.

10 SHEN, C. et al. Accuracy or novelty: what can we gain from target-specific machine-learning-based scoring functions in virtual screening? *Briefings in Bioinformatics*, 2021.

11 MOBLEY, D. L.; GILSON, M. K. Predicting binding free energies: frontiers and benchmarks. *Annual review of biophysics*, Annual Reviews, v. 46, p. 531–558, 2017.

12 HU, Q.-D.; TANG, G.-P.; CHU, P. K. Cyclodextrin-based host–guest supramolecular nanoparticles for delivery: from design to applications. *Accounts of chemical research*, ACS Publications, v. 47, n. 7, p. 2017–2025, 2014.

13 GU, A.; WHEATE, N. J. Macrocycles as drug-enhancing excipients in pharmaceutical formulations. *Journal of Inclusion Phenomena and Macrocyclic Chemistry*, Springer, p. 1–15, 2021.

14 BEISKEN, S. et al. Knime-cdk: Workflow-driven cheminformatics. *BMC bioinformatics*, Springer, v. 14, n. 1, p. 1–4, 2013.

15 BERTHOLD, M. R. et al. Knime-the konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter*, ACM New York, NY, USA, v. 11, n. 1, p. 26–31, 2009.

16 CARVALHO, R. M. et al. Gaussian processes regression for cyclodextrin host-guest binding prediction. *Journal of Inclusion Phenomena and Macrocyclic Chemistry*, Springer, v. 101, n. 1, p. 149–159, 2021.

17 CARVALHO, R. M. et al. Prediction of cyclodextrin host-guest binding through a hybrid support vector method. In: *Proceedings of Research and Applications in Artificial Intelligence*. [S.l.]: Springer, 2021. p. 309–317.

18 CARVALHO, T. P. et al. A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, Elsevier, v. 137, p. 106024, 2019.

19 DEY, A. Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies*, v. 7, n. 3, p. 1174–1179, 2016.

20 BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, JMLR. org, v. 13, n. 1, p. 281–305, 2012.

21 TEYSSANDIER, J.; FEYTER, S. D.; MALI, K. S. Host–guest chemistry in two-dimensional supramolecular networks. *Chemical Communications*, Royal Society of Chemistry, v. 52, n. 77, p. 11465–11487, 2016.

22 ANSLYN, E. V.; DOUGHERTY, D. A. *Modern physical organic chemistry*. [S.l.]: University science books, 2006.

23 PIÑEIRO, Á. et al. On the characterization of host- guest complexes: surface tension, calorimetry, and molecular dynamics of cyclodextrins with a non-ionic surfactant. *The Journal of Physical Chemistry B*, ACS Publications, v. 111, n. 17, p. 4383–4392, 2007.

24 ZHAO, Q. et al. Research advances in molecular modeling in cyclodextrins. *Current pharmaceutical design*, Bentham Science Publishers, v. 23, n. 3, p. 522–531, 2017.

25 MOBLEY, D. L. et al. The sampl7 host-guest challenge virtual workshop. *Zenodo*, 2020.

26 WILLGERT, M. et al. Cellulose nanofibril reinforced composite electrolytes for lithium ion battery applications. *Journal of Materials Chemistry A*, Royal Society of Chemistry, v. 2, n. 33, p. 13556–13564, 2014.

27 KATRITZKY, A. R. et al. Quantitative structure- property relationship modeling of β-cyclodextrin complexation free energies. *Journal of chemical information and computer sciences*, ACS Publications, v. 44, n. 2, p. 529–541, 2004.

28 COLEMAN, A. W. et al. Aggregation of cyclodextrins: An explanation of the abnormal solubility of β-cyclodextrin. *Journal of inclusion phenomena and molecular recognition in chemistry*, Springer, v. 13, n. 2, p. 139–143, 1992.

29 SID, D. et al. Solubility enhancement of mefenamic acid by inclusion complex with β-cyclodextrin: in silico modelling, formulation, characterisation, and in vitro studies. *Journal of enzyme inhibition and medicinal chemistry*, Taylor & Francis, v. 36, n. 1, p. 605–617, 2021.

30 SAOKHAM, P. et al. Solubility of cyclodextrins and drug/cyclodextrin complexes. *Molecules*, Multidisciplinary Digital Publishing Institute, v. 23, n. 5, p. 1161, 2018.

31 SZENTE, L.; SZEJTLI, J. Highly soluble cyclodextrin derivatives: chemistry, properties, and trends in development. *Advanced drug delivery reviews*, Elsevier, v. 36, n. 1, p. 17–28, 1999.

32 WIMMER, T. Cyclodextrins. *Ullmann's Encyclopedia of Industrial Chemistry*, Wiley Online Library, 2000.

33 CONCEICAO, J. et al. Cyclodextrins as drug carriers in pharmaceutical technology: the state of the art. *Current pharmaceutical design*, Bentham Science Publishers, v. 24, n. 13, p. 1405–1433, 2018.

34 ZAREH, M. M. β-cyclodextrin as an ionophore for membrane electrode. *Cyclodextrin: A Versatile Ingredient*, BoD–Books on Demand, p. 291, 2018.

35 MORAES, C. M. et al. Preparação e caracterização físico-química de complexos de inclusão entre anestésicos locais e hidroxipropil-beta-ciclodextrina. *Química Nova*, SciELO Brasil, v. 30, p. 777–784, 2007.

36 SUÁREZ, D.; DÍAZ, N. Affinity calculations of cyclodextrin host–guest complexes: assessment of strengths and weaknesses of end-point free energy methods. *Journal of chemical information and modeling*, ACS Publications, v. 59, n. 1, p. 421–440, 2018.

37 WICKSTROM, L. et al. Large scale affinity calculations of cyclodextrin host–guest complexes: understanding the role of reorganization in the molecular recognition process. *Journal of chemical theory and computation*, ACS Publications, v. 9, n. 7, p. 3136–3150, 2013.

38 XU, P. et al. Computation of host–guest binding free energies with a new quantum mechanics based mining minima algorithm. *The Journal of Chemical Physics*, AIP Publishing LLC, v. 154, n. 10, p. 104122, 2021.

39 MUDDANA, H. S. et al. The sampl4 host–guest blind prediction challenge: an overview. *Journal of computer-aided molecular design*, Springer, v. 28, n. 4, p. 305–317, 2014.

40 MERZLIKINE, A. et al. Development of machine learning models of β-cyclodextrin and sulfobutylether-β-cyclodextrin complexation free energies. *International journal of pharmaceutics*, Elsevier, v. 418, n. 2, p. 207–216, 2011.

41 XU, Q. et al. Quantitative structure–property relationship study of β-cyclodextrin complexation free energies of organic compounds. *Chemometrics and Intelligent Laboratory Systems*, Elsevier, v. 146, p. 313–321, 2015.

42 SOLOVEV, A.; SOLOV'EV, V. 3d molecular fragment descriptors for structure–property modeling: predicting the free energies for the complexation between antipodal guests and β-cyclodextrins. *Journal of Inclusion Phenomena and Macrocyclic Chemistry*, Springer, v. 89, n. 1, p. 167–175, 2017.

43 BRUNCK, T.; WEINHOLD, F. Quantum-mechanical studies on the origin of barriers to internal rotation about single bonds. *Journal of the American Chemical Society*, ACS Publications, v. 101, n. 7, p. 1700–1709, 1979.

44 DI, P. et al. In silico prediction of binding capacity and interaction forces of organic compounds with α-and β-cyclodextrins. *Journal of Molecular Liquids*, Elsevier, v. 302, p. 112585, 2020.

45 BOUREL, M.; CRISCI, C.; MARTÍNEZ, A. Consensus methods based on machine learning techniques for marine phytoplankton presence–absence prediction. *Ecological Informatics*, Elsevier, v. 42, p. 46–54, 2017.

46 GUEDES, I. A.; PEREIRA, F. S.; DARDENNE, L. E. Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. *Frontiers in pharmacology*, Frontiers, v. 9, p. 1089, 2018.

47 SU, M. et al. Tapping on the black box: how is the scoring power of a machine-learning scoring function dependent on the training set? *Journal of chemical information and modeling*, ACS Publications, v. 60, n. 3, p. 1122–1136, 2020.

48 POLITI, R. et al. Docking and scoring with target-specific pose classifier succeeds in native-like pose identification but not binding affinity prediction in the csar 2014 benchmark exercise. *Journal of chemical information and modeling*, ACS Publications, v. 56, n. 6, p. 1032–1041, 2016.

49 GILSON, M. K. et al. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research*, Oxford University Press, v. 44, n. D1, p. D1045–D1053, 2016.

50 WILDMAN, S. A.; CRIPPEN, G. M. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences*, ACS Publications, v. 39, n. 5, p. 868–873, 1999.

51 LABUTE, P. A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling*, Elsevier, v. 18, n. 4-5, p. 464–477, 2000.

52 ERTL, P.; ROHDE, B.; SELZER, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of medicinal chemistry*, ACS Publications, v. 43, n. 20, p. 3714–3717, 2000.

53 RODRÍGUEZ, J. D.; LOZANO, J. A. Repeated stratified k-fold cross-validation on supervised classification with naive bayes classifier: An empirical analysis. 2007.

54 GAO, H.; YE, Z.; AL et. Predicting drug/phospholipid complexation by the lightgbm method. *Chemical Physics Letters*, Elsevier, p. 137354, 2020.

55 SASAKI, H.; NOH, Y.-K.; SUGIYAMA, M. Direct density-derivative estimation and its application in kl-divergence approximation. In: PMLR. *Artificial Intelligence and Statistics*. [S.l.], 2015. p. 809–818.

56 KULLBACK, S.; LEIBLER, R. A. On information and sufficiency. *The annals of mathematical statistics*, JSTOR, v. 22, n. 1, p. 79–86, 1951.

57 KULLBACK, S. *Information theory and statistics*. [S.l.]: Courier Corporation, 1997.

58 LEARN, S. Cross-validation: evaluating estimator performance. *línea]. Available: https://scikit-learn. org/stable/modules/cross_validation. html# crossvalidation [Último acceso: 26 Mayo 2020]*, 2017.

59 GRANT, M.; BOYD, S.; YE, Y. *CVX: Matlab software for disciplined convex programming*. 2009.

60 CHANG, C.-C.; LIN, C.-J. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, Acm New York, NY, USA, v. 2, n. 3, p. 1–27, 2011.

61 KARTHIKEYAN, M.; VYAS, R. Machine learning methods in chemoinformatics for drug discovery. In: *Practical chemoinformatics*. [S.l.]: Springer, 2014. p. 133–194.

62 KARGAR, K.; AL. et. Estimating longitudinal dispersion coefficient in natural streams using empirical models and machine learning algorithms. *Engineering Applications of Computational Fluid Mechanics*, Taylor & Francis, v. 14, n. 1, p. 311–322, 2020.

63 GUNN, S. R. et al. Support vector machines for classification and regression. *ISIS technical report*, v. 14, n. 1, p. 5–16, 1998.

64 CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2016. (KDD '16), p. 785–794. ISBN 978-1-4503-4232-2. Disponível em: <http://doi.acm.org/10.1145/2939672.2939785>.

65 CHEN, T.; HE, T. Higgs boson discovery with boosted trees. In: *NIPS 2014 workshop on high-energy physics and machine learning*. [S.l.: s.n.], 2015. p. 69–80.

66 GöRTLER, J.; KEHLBECK, R.; DEUSSEN, O. A visual exploration of gaussian processes. *Distill*, 2019. Https://distill.pub/2019/visual-exploration-gaussian-processes.

67 GOLIATT, L. et al. Estimation of natural streams longitudinal dispersion coefficient using hybrid evolutionary machine learning model. *Engineering Applications of Computational Fluid Mechanics*, Taylor & Francis, v. 15, n. 1, p. 1298–1320, 2021.

68 WANG, J. An intuitive tutorial to gaussian processes regression. *arXiv preprint arXiv:2009.10862*, 2020.

69 RASMUSSEN, C. E. Gaussian processes in machine learning. In: SPRINGER. *Summer school on machine learning*. [S.l.], 2003. p. 63–71.

70 WILLIAMS, C. K.; RASMUSSEN, C. E. *Gaussian processes for machine learning*. [S.l.]: MIT press Cambridge, MA, 2006. v. 2.

71 MURPHY, K. P. *Machine learning: a probabilistic perspective*. [S.l.]: MIT press, 2012.

72 PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, JMLR. org, v. 12, p. 2825–2830, 2011.

73 OU, X.; MORRIS, J.; MARTIN, E. Gaussian process regression for batch process modelling. *IFAC Proceedings Volumes*, Elsevier, v. 37, n. 9, p. 817–822, 2004.

74 WANG, B.; CHEN, T. Gaussian process regression with multiple response variables. *Chemometrics and Intelligent Laboratory Systems*, Elsevier, v. 142, p. 159–165, 2015.

75 KUMAR, R. The generalized modified bessel function and its connection with voigt line profile and humbert functions. *Ad. in Applied Mathematics*, Elsevier, v. 114, p. 101986, 2020.

76 SCHMIDT, M. et al. On the performance of differential evolution for hyperparameter tuning. In: IEEE. *2019 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2019. p. 1–8.

77 KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA. *Ijcai*. [S.l.], 1995. v. 14, n. 2, p. 1137–1145.

78 BICKEL, P. et al. *Springer series in statistics*. [S.l.]: Springer, 2009.

79 PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

80 GOLBRAIKH, A. et al. Predictive qsar modeling: methods and applications in drug discovery and chemical risk assessment. In: *Handbook of computational chemistry*. [S.l.]: Springer Netherlands, 2012. p. 1309–1342.

81 ZERBE, O.; JURT, S. *Applied NMR spectroscopy for chemists and life scientists*. [S.l.]: John Wiley & Sons, 2013.

82 BLUNDELL, C. D.; NOWAK, T.; WATSON, M. J. Measurement, interpretation and use of free ligand solution conformations in drug discovery. In: *Progress in medicinal chemistry*. [S.l.]: Elsevier, 2016. v. 55, p. 45–147.

83 SAPORETTI, C. M. et al. Hybrid unsupervised extreme learning machine applied to facies identification. In: *Proceedings of Research and Applications in Artificial Intelligence*. [S.l.]: Springer, 2021. p. 319–326.

84 CARVALHO, R. M. et al. A comparative study of singular value decomposition (svd) and discrete cosine transform (dct) techniques for image compression aplications. *Proceedings of the XLI Ibero-Latin-American Congress on Computational Methods in Engineering*, Associação Brasileira de Métodos Computacionais em Engenharia (ABMEC), v. 2, p. 8209, 2020.

85 ROSA, I. G. et al. Extensions and improvements of the extreme learning machine (elm) applied to face recognition. *Proceedings of the XLI Ibero-Latin-American Congress on Computational Methods in Engineering*, Associação Brasileira de Métodos Computacionais em Engenharia (ABMEC), v. 2, 2020.

86 CARVALHO, R. M.; ROSSI, A. D.; GOLIATT, P. V. Z. C. Quality-model clustering tool: A module for clustering protein models based on quality attributes. *Revista Mundi Engenharia, Tecnologia e Gestão (ISSN: 2525-4782)*, v. 5, n. 2, 2020.

87 SILVA, A. C. E. P. et al. Ds mapeados: uma ferramenta computacional voltada à gestão acadêmica e científica para o apoio e alinhamento da agenda 2030. *Anais do Terceiro Sustentare e Sexto Wipis*, Even3, 2021. 10.29327/III_SUSTENTARE_VI_WIPIS.430788.

88 JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, The Royal Society Publishing, v. 374, n. 2065, p. 20150202, 2016.

89 ARTHUR, D.; VASSILVITSKII, S. *k-means++: The advantages of careful seeding*. [S.l.], 2006.

90 ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, Elsevier, v. 20, p. 53–65, 1987.

APÊNDICE A – CONJUNTOS DE TREINAMENTO E TESTE

A.1 ESTATÍSTICA DESCRITIVA DO CONJUNTO DE TREINAMENTO

Tabela 9 – Estatísticas descritivas sobre os dados de treinamento.

Variável	**Média**	**Desvio**	**Mínimo**	**25%**	**50%**	**75%**	**Máximo**
X_1	6.929	0.089	6.900	6.900	6.900	6.900	7.200
X_2	24.980	0.445	20.700	25.000	25.000	25.000	30.000
X_3	0.864	1.035	-4.756	0.298	0.825	1.557	4.342
X_4	52.243	26.540	0.000	35.376	44.481	57.051	129.975
X_5	83.718	43.094	23.783	52.684	71.665	92.797	209.562
X_6	58.354	39.443	0.000	27.640	43.280	81.248	171.600
X_7	199.471	105.300	58.085	125.920	165.168	231.761	514.705
X_8	3.326	2.500	0.000	1.000	2.000	5.000	10.000
X_9	1.781	1.591	0.000	0.000	1.000	3.000	5.000
X_{10}	3.473	2.693	0.000	2.000	3.000	5.000	11.000
X_{11}	28.393	12.766	3.000	21.000	25.000	30.000	79.000
X_{12}	-0.165	0.823	-3.000	-1.000	0.000	0.000	1.000
X_{13}	-14.939	1.413	-17.406	-15.231	-15.231	-13.055	-13.055
X_{14}	224.063	21.185	195.800	195.800	228.434	228.434	261.067
X_{15}	425.648	40.116	372.130	372.130	433.924	433.924	495.717
X_{16}	543.450	51.383	474.900	474.900	554.050	554.050	633.200
X_{17}	1113.272	105.260	972.846	972.846	1134.987	1134.987	1297.128
X_{18}	34.330	3.246	30.000	30.000	35.000	35.000	40.000
X_{19}	20.598	1.948	18.000	18.000	21.000	21.000	24.000
X_{20}	6.866	0.649	6.000	6.000	7.000	7.000	8.000
X_{21}	144.188	13.633	126.000	126.000	147.000	147.000	168.000
y	-13.084	4.762	-30.690	-16.320	-12.050	-10.278	-2.400

A.2 ESTATÍSTICA DESCRITIVA DO CONJUNTO DE TESTE

Tabela 10 – Estatísticas descritivas sobre os dados de teste.

Variável	**Média**	**Desvio**	**Mínimo**	**25%**	**50%**	**75%**	**Máximo**
X_1	6.911	0.056	6.900	6.900	6.900	6.900	7.200
X_2	24.981	0.050	24.850	25.000	25.000	25.000	25.000
X_3	0.776	0.924	-1.436	0.258	0.771	1.236	3.090
X_4	51.394	24.636	17.355	33.752	45.431	54.937	108.101
X_5	82.264	40.071	26.263	51.472	71.847	92.717	172.976
X_6	57.919	34.578	17.070	28.735	40.460	78.460	135.960
X_7	193.604	96.244	60.096	125.935	166.244	222.472	413.430
X_8	3.375	2.340	1.000	1.750	2.000	5.000	9.000
X_9	1.857	1.420	0.000	1.000	1.000	3.000	5.000
X_{10}	3.304	2.628	0.000	2.000	3.000	4.000	10.000
X_{11}	28.018	10.914	12.000	21.000	25.500	30.250	55.000
X_{12}	-0.125	0.689	-2.000	-1.000	0.000	0.000	1.000
X_{13}	-15.231	1.245	-17.406	-15.231	-15.231	-15.231	-13.055
X_{14}	228.434	18.669	195.800	228.434	228.434	228.434	261.067
X_{15}	433.924	35.351	372.130	433.924	433.924	433.924	495.717
X_{16}	554.050	45.280	474.900	554.050	554.050	554.050	633.200
X_{17}	1134.987	92.757	972.846	1134.987	1134.987	1134.987	1297.128
X_{18}	35.000	2.860	30.000	35.000	35.000	35.000	40.000
X_{19}	21.000	1.716	18.000	21.000	21.000	21.000	24.000
X_{20}	7.000	0.572	6.000	7.000	7.000	7.000	8.000
X_{21}	147.000	12.014	126.000	147.000	147.000	147.000	168.000
y	-12.951	4.115	-24.830	-16.052	-12.170	-10.232	-6.400

APÊNDICE B – VARIABILIDADE DAS MOLÉCULAS LIGANTES

B.1 AGRUPAMENTO E ANÁLISE DAS MOLÉCULAS LIGANTES

Para verificar a diversidade de dados para as moléculas hóspedes (ligantes), foi proposto uma análise alternativa por meio de métodos de agrupamento dos dados. Nesse cenário, considerou-se somente os atributos relativos aos hóspedes de cada instância (X_3 a X_{12} da Tabela 3), desconsiderando duplicatas, uma vez que um mesmo ligante pode ter sido testado com diferentes hospedeiros na base. A análise realizada partiu de uma redução de dimensionalidade dos atributos para um hiperespaço de dimensão 3 (`new_dim = 3`) utilizando a estratégia *Principal Components Analysis* (PCA) (88). Para o agrupamento dos dados foi utilizado o algoritmo *K-Means* (89). O *K-Means* demanda o número de grupos (k) *a priori*, assim, executou-se uma busca exaustiva entre 2 e 10 grupos que otimizassem o valor da métrica *silhueta* (90). A Figura 26 apresenta os resultados obtidos.

O algoritmo identificou dois grandes grupos de ligantes que apresentam certa variabilidade. Os resultados do PCA garantem que a maior parte da variância dos dados conseguiram ser representadas em 3 dimensões. Isso poderia nos direcionar para estudos futuros na redução da dimensionalidade dos atributos descritores dos ligantes antes mesmo do treinamento das máquinas de aprendizado. O baixo valor de *silhueta* encontrado com a

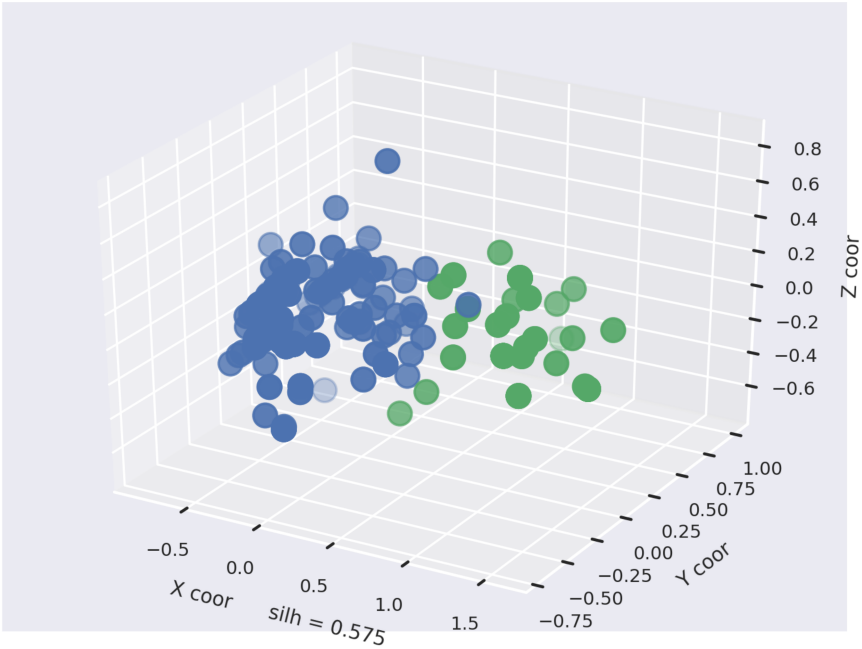


Figura 26 – Distribuição e agrupamento das moléculas de ligantes consideradas. *K-Means*: k = 2, Autovalor do PCA = (0.69 0.20 0.06).

busca exaustiva no espaço delimitado ($silh = 0.575$) garante que os dados não conseguiram ser agrupados de forma satisfatória, o que pode ser considerado uma característica positiva na aplicação atual, uma vez que se busca dados diversos para um treinamento geral das máquinas. A distribuição garante certa diversidade nos dados, o que permite a realização das investigações iniciais sobre a capacidade de predição com o atual conjunto de dados.

APÊNDICE C – RESUMO DE INFORMAÇÕES DA SESSÃO DE RESULTADOS E DISCUSSÃO

C.1 RESUMO DE RESULTADOS OBTIDOS DA LITERATURA

Referência	**Modelo**	**Obs.**	**#Dados**	**Tipo de CD**	R^2	**RMSE**	**MAPE**
	ε-SVR			α-CD (73)	0.776	1.932	1.351
	XGB		280	β-CD (164)	0.688	2.277	1.468
	GPR			γ-CD (43)	0.803	1.811	1.201
Dimas et al.	Molecular Dynamics		57	β-Cyclodextrin	0.66	9.330	-
Solovev et al.	Multiple Linear Regression			β-Cyclodextrin	0.918	0.89	-
				am-β-Cyclodextrin	0.910	0.89	-
Xu et al.	Multiple Linear Regression		218	β-Cyclodextrin	0.833	-	1.911
	Artificial Neural Netwotk				0.957	-	0.925
Zhao et al.	Random Forest		3000	Hp-β-CD; 2-hydroxypropyl-β-CD; RMCD; randomly methylated β-CD; TMCD; (2,3,6-tri-O-methyl)-β-CD; DMCD; (2,6-di-O-methyl)-β-CD; SBE-β-CD; sulfobutylether-β-CD	0.81	2.11	1.54
	Deep Learning		3000		0.62	3.36	2.56
	LightGBM		3000		0.86	1.83	1.38
			2500		$\sim$0.82	-	$\sim$1.5
			2000		$\sim$0.78	-	$\sim$1.6
			1500		$\sim$0.73	-	$\sim$1.9
			1000		$\sim$0.71	-	$\sim$2.0
			500		$\sim$0.58	-	$\sim$2.3

Tabela 11 – Tabela resumo com os resultados de predições de energia de interação de uma seleção de trabalhos da literatura com domínio de aplicação similar ao trabalho em questão.