PUCRS

ESCOLA DE CIÊNCIAS DA SAÚDE E DA VIDA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR
MESTRADO EM BIOLOGIA CELULAR E MOLECULAR

ROSANA DA SILVA WASZAK

DESENVOLVIMENTO E APLICAÇÃO DE FUNÇÕES ESCORE OTIMIZADAS PARA PREVISÃO DE AFINIDADE ENTRE PROTEÍNAS E LIGANTES

Porto Alegre
2020

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA DE CIÊNCIAS DA SAÚDE E DA VIDA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR
MESTRADO EM BIOLOGIA CELULAR E MOLECULAR

ROSANA DA SILVA WASZAK

DESENVOLVIMENTO E APLICAÇÃO DE FUNÇÕES ESCORE OTIMIZADAS PARA PREVISÃO DE AFINIDADE ENTRE PROTEÍNAS E LIGANTES

Porto Alegre
2020

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA DE CIÊNCIAS DA SAÚDE E DA VIDA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR
MESTRADO EM BIOLOGIA CELULAR E MOLECULAR

ROSANA DA SILVA WASZAK

DESENVOLVIMENTO E APLICAÇÃO DE FUNÇÕES ESCORE OTIMIZADAS PARA PREVISÃO DE AFINIDADE ENTRE PROTEÍNAS E LIGANTES

Dissertação de Mestrado apresentada como requisito para a obtenção do grau de Mestre pelo Programa de Pós-Graduação em Biologia Celular e Molecular da Escola de Ciências da Saúde e da Vida da Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Dr. Walter Filgueira de Azevedo Jr.

Porto Alegre
2020

Ficha Catalográfica

W323d Waszak, Rosana da Silva

Desenvolvimento e Aplicação de Funções Escore Otimizadas para Previsão de Afinidade Entre Proteínas e Ligantes / Rosana da Silva Waszak . – 2020.

76.

Dissertação (Mestrado) – Programa de Pós-Graduação em Biologia Celular e Molecular, PUCRS.

Orientador: Prof. Dr. Walter Filgueira de Azevedo Junior.

1. Funções Escore. 2. Interações Proteína-Ligante. 3. Aprendizado de Máquina. 4. Bioinformática. I. Junior, Walter Filgueira de Azevedo. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS com os dados fornecidos pelo(a) autor(a).

Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

ROSANA DA SILVA WASZAK

DESENVOLVIMENTO E APLICAÇÃO DE FUNÇÕES ESCORE OTIMIZADAS PARA PREVISÃO DE AFINIDADE ENTRE PROTEÍNAS E LIGANTES

Dissertação de Mestrado apresentada como requisito para a obtenção do grau de Mestre pelo Programa de Pós-Graduação em Biologia Celular e Molecular da Escola de Ciências da Saúde e da Vida da Pontifícia Universidade Católica do Rio Grande do Sul.

Área de concentração: Biologia Molecular da Interação Droga/Alvo

Aprovada em:_____de_____de_____.

BANCA EXAMINADORA

_____
Prof. Dr. Maurício Reis Bogo

_____
Prof. Dr. Rodrigo Ligabue Braun

_____
Prof. Dr. Cristiano Valim Bizarro

Porto Alegre
2020

AGRADECIMENTOS

Agradeço à minha família e, em especial, aos meus pais Graça e Rui, pelo apoio e amor durante toda a minha vida. Sou grata por todos os sacrifícios que fizeram por mim e por todo o incentivo nos momentos bons e ruins. Pelas lições de vida, persistência, coragem e tudo o que sempre me ensinaram e continuam ensinando. Por todas as oportunidades que vocês têm me proporcionado e pela união de nossa família, que, com certeza, se deve à dedicação diária de vocês.

Ao meu orientador, Prof. Dr. Walter Filgueira de Azevedo Jr., agradeço por me acolher no laboratório, por todo o conhecimento que me proporcionou ao longo de quatro anos de convivência e por toda a compreensão que teve comigo nos momentos de dificuldade e de minhas inúmeras falhas ao longo do desenvolvimento desta pesquisa. Muito obrigada, professor.

Agradeço a todos os professores que fizeram parte da minha trajetória até aqui, em especial aos que me ajudaram na decisão de seguir na pós-graduação: Dra. Janaina Gomes, Dra. Márcia Häfele Islabão Franco, Dr. Fabio Yoshimitsu Okuyama, Dra. Karen Selbach Borges, Dr. Rodrigo Prestes Machado, Dra. Monica Ryff Moreira Vianna, Dr. Eduardo Eizirik e Dr. Cristiano Valim Bizarro. Obrigada por sua dedicação incansável ao ensino e pesquisa, pelo incentivo e conhecimentos ao longo de todo o tempo de nossa convivência.

A todos os amigos e colegas que a vida me presenteou e que sempre estiveram ao meu lado e me apoiaram apesar de minhas faltas. Agradeço por compartilharem a jornada da vida e me alegrarem com sua presença, seu apoio, suas experiências, alegrias e tristezas. Tudo isto me faz crescer enormemente.

À CAPES pela bolsa concedida, à PUCRS e ao PPGBCM pela oportunidade de realização deste trabalho.

E por último, mas não menos importante, agradeço à força motriz do Universo, que permite que cada átomo esteja precisamente no lugar onde deve estar e que faz tudo agir como deve agir.

"*Let us strive for the impossible. The great achievements throughout history have been the conquest of what seemed the impossible.*"

Charles Chaplin

RESUMO

Cinases são as proteínas mais intensamente estudadas no desenvolvimento e desenho de fármacos. Dentre as cinases, as serino/treonino cinases não específicas representam um sistema proteico interessante para propósitos de modelagem, devido à grande disponibilidade de dados experimentais estruturais e funcionais. As serino/treonino cinases não específicas compreendem uma importante classe de proteínas alvo usadas para o desenvolvimento de fármacos para tratamento de câncer. Neste estudo, foi descrita a criação de modelos de Aprendizado de Máquina para predição de afinidade entre proteínas e ligantes para esta classe enzimática. Foram utilizados para tal termos de energia calculados por Funções Escore clássicas disponíveis em programas como *AutoDock4* e *AutoDock Vina*. Estes termos foram empregados para a criação de novas Funções Escore específicas para um conjunto de dados composto por aproximadamente 100 complexos proteína-ligante, para os quais dados experimentais como a estrutura cristalográfica e a constante de inibição estavam disponíveis. Foi aplicado também um método híbrido que utiliza a simulação de um sistema de massa-mola para determinar a afinidade de ligação, usando o programa Taba. Todas as Funções Escore geradas tiveram sua performance preditiva analisada. Os resultados mostram claramente que os modelos de aprendizado de máquina apresentam capacidade preditiva superior quando comparados com as Funções Escore clássicas. Além disso, os modelos de Aprendizado de Máquina gerados foram capazes de identificar características estruturais, responsáveis pela afinidade de ligação junto a serino/treonino cinases não específicas.

Palavras-chave: Funções Escore. Interações Proteína-Ligante. Aprendizado de Máquina. Bioinformática.

ABSTRACT

Kinases are the most intensively studied protein in drug design and development. Among kinases, non-specific serine/threonine protein kinase represents an interesting protein system for modeling purposes due to the availability of structural and functional experimental data. Non-specific serine/threonine protein kinase comprises an important class of protein targets used to develop drugs to treat cancer. In this study, we describe the creation of machine learning models to predict protein-ligand binding affinity for this enzymatic class. We make use of energy terms available in classical scoring functions such as Autodock4 and AutoDock Vina. We use these terms to build a novel scoring function targeted to a dataset comprised of nearly 100 protein-ligand complexes for which experimental crystallographic structure and inhibition constant data are available. We also applied a hybrid mass-spring method to determine binding affinity using the program Taba. We carried out predictive performance analysis of all scoring functions. Our study clearly shows that machine learning models present superior predictive performance when compared with classical scoring functions. Also, our machine learning models can capture structural features responsible for the binding affinity against non-specific serine/threonine protein kinases.

Keywords: Scoring Functions. Protein-ligand Interactions. Machine Learning. Bioinformatics.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

AM – Aprendizado de Máquina

AMS – Aprendizado de Máquina Supervisionado

DM – Dinâmica Molecular

MDM – *Molegro Data Modeller*

MLS – *Multiple Linear Regression*

MVD – *Molegro Virtual Docking*

NN – *Neural Networks*

PDB – *Protein Data Bank*

PLS – *Partial Least Squares*

RMSE – *Root Mean Square Error*

RSS – *Residual Sum of Squares*

SD – *Standard Deviation*

SVM – *Support Vector Machine*

LISTA DE SÍMBOLOS

EC_{50}	Concentração Efetiva Máxima a 50%
IC_{50}	Concentração Inibitória Máxima a 50%
K_i	Constante de Inibição
K_d	Constante de Dissociação
ΔG	Energia Livre de Gibbs
(ρ)	Coeficiente de Correlação de Spearman
(R^2)	Coeficiente de Correlação de Pearson

SUMÁRIO

1. INTRODUÇÃO

Atualmente, diferentes técnicas e abordagens modernas têm sido utilizadas no processo de desenvolvimento de fármacos, como parte da busca contínua por moléculas mais eficazes e por soluções menos custosas, mais rápidas e mais precisas. Uma destas abordagens, chamada de abordagem fisiológica, tem como base o mecanismo de ação farmacológico pretendido, e, para isto, é necessário eleger um alvo terapêutico, cujo processo fisiopatológico já tenha sido identificado. Este alvo, geralmente uma enzima ou receptor, pode ter sua estrutura tridimensional conhecida ou não. Caso a tenha, há possibilidade de desenho de inibidores/ativadores enzimáticos ou antagonistas/agonistas de receptores, através de processos de complementariedade molecular planejada (Barreiro; Fraga, 2015). A afinidade de um receptor por uma molécula alvo pode ser prevista e medida através de simulações *in silico,* o que agiliza e otimiza a busca por fármacos, na medida em que os custos e o tempo dedicado para testes *in vitro* e *in vivo* acabam sendo consideravelmente diminuídos (Sliwoski et al., 2014).

Diversos programas de código aberto que visam a auxiliar no desenho de novos fármacos estão disponíveis para utilização pela comunidade científica. Aqueles que realizam a docagem molecular, ou seja, o processo de busca pelas melhores posições para que um ligante possa se encaixar ao sítio de ligação de uma proteína, utilizam as Funções Escore para encontrar a menor energia de ligação entre as possíveis conformações do complexo proteína-ligante (Huang; Grinter; Zou, 2010). A predição de afinidade entre os elementos deste complexo, por sua vez, pode ser vista como uma etapa posterior ao processo de docagem molecular e prevê a energia da interação não covalente entre as duas moléculas, através do cálculo das interações intra e intermoleculares, por exemplo, que é realizado por algumas das Funções Escore clássicas (Bitencourt-Ferreira; De Azevedo Jr, 2018).

Cada *software* que implementa estas funcionalidades utiliza equações específicas com diferentes termos para o cálculo da energia de ligação. Como exemplo é possível citar o *Molegro Virtual Docking* (MVD) (Thomsem; Christensen, 2006), o *AutoDock Vina* (Trott; Olson, 2010) e o *AutoDock4* (Morris et al., 2009). Devido às particularidades de cada equação, muitas vezes o uso destes programas

acaba sendo restrito a um sistema biológico específico, ou seja, tende a funcionar melhor para um conjunto de moléculas do que para outros. Isto demonstra a necessidade de implementação de novos métodos e ferramentas que possam cumprir o papel de ajustar-se à família de proteínas de interesse, levando em conta suas características singulares, mas que, ao mesmo tempo, sejam flexíveis para utilização com os mais diversos sistemas biológicos.

1.1. PROTEÍNAS

As proteínas são estruturas complexas e sofisticadas, compostas por cadeias polipeptídicas formadas entre um grupo amina e um grupo carboxila e cuja estrutura tem sido refinada ao longo de bilhões de anos de evolução. A forma de uma proteína é especificada por sua sequência de aminoácidos e sua função é determinada pela conformação que esta estrutura adota de acordo com o ambiente em que se encontra (Alberts, et al*.,* 2017).

As propriedades biológicas das proteínas dependem de sua interação física com outras moléculas. Cada proteína pode ligar-se a uma ou mais moléculas através de uma região em sua superfície chamada de sítio de ligação. Pequenas moléculas interagem com o sítio de ligação de forma específica, e no caso das enzimas, que são um grupo de proteínas responsáveis por catalisar reações químicas, essa ligação a pequenas moléculas as permite modificar seus substratos e dar início ou sequência a diferentes processos celulares (Alberts, et al., 2017).

Dentre as enzimas, as proteínas cinase, que catalisam a transferência de um fosfato do ATP para os resíduos de seus substratos, desempenham um papel essencial no controle de quase todas as funções celulares, com especial referência à transdução de sinais (Battistutta, et al. 2005). Possuem ainda grande importância em atividades celulares tais como apoptose, proliferação, neurotransmissão e oncogênese, e sua desregulação ou superexpressão, está associada a doenças como asma, câncer, e doenças do sistema nervoso central, entre outras (Silva*,* et al., 2009*).*

Esta é uma das maiores famílias de enzimas, com mais de 500 tipos sendo codificados pelo genoma humano (Battistutta, et al. 2005). Desse modo, as cinases têm sido amplamente estudadas, como potenciais alvos para o desenvolvimento de fármacos, incluindo antitumorais e antirretrovirais (Tutone, 2017; Mielecki, 2016).

Levando em consideração sua relevância para este fim, podem-se destacar as seguintes enzimas deste grupo: serino/treonino cinases não específicas, cinases dependentes de ciclina, caseína cinase 1 e cinase dependente de cálcio/calmodulina.

1.2. FUNÇÕES ESCORE

No âmbito das interações entre proteína e ligante, as Funções Escore podem ser definidas como métodos matemáticos que modelam uma situação física e retornam um valor diretamente relacionado à energia de ligação, chamada de Energia Livre de Gibbs ou ΔG (Jain, 2006).

Ainda segundo Jain (2006), uma das dificuldades em se estimar o ΔG de uma ligação é a de que as muitas interações físicas do complexo proteína-ligante possuem múltiplos efeitos entálpicos e entrópicos, ainda que indiretos. Por exemplo, a mais simples interação entre duas superfícies hidrofóbicas implica em efeitos entálpicos e entrópicos no ligante, na proteína e nas moléculas do solvente. Desta forma, o resultado da energia pode ser definido como uma combinação de efeitos que envolvem a formação e destruição das interações entre estes componentes do complexo, na forma de interações de Van der Waals, de ligações de Hidrogênio e interações Carga-Carga, entre outras.

De acordo com Heck et al. (2017), as Funções Escore clássicas podem ser divididas de forma abrangente em três famílias: a primeira, cujas funções são baseadas em campos de força, a segunda formada por Funções Escore empíricas e a terceira que compreende as funções baseadas em conhecimento. Como as equações utilizadas para o cálculo da energia de ligação variam de acordo com a natureza da situação que está sendo modelada, cada uma das famílias citadas engloba funções que possuem termos de energia diferentes. Se analisadas do ponto de vista computacional, estas particularidades exigem mais ou menos processamento para realização dos cálculos.

As Funções Escore baseadas em campos de força utilizam as interações intra e intermoleculares do complexo e levam em conta a geometria ideal para comprimentos de ligação, ângulo de ligação e ângulo diedro. As Funções Escore empíricas variam de uma abordagem para outra, mas a ideia básica é somar as interações intermoleculares multiplicadas por coeficientes de peso para a predição de afinidade de ligação. Por fim, as Funções Escore baseadas em conhecimento, em sua maior

parte, derivam um potencial intermolecular através da análise experimental de determinadas estruturas complexas (Heck et al., 2017).

Computacionalmente, as Funções Escore clássicas têm sido utilizadas na predição de afinidade entre receptores e seus ligantes há mais de duas décadas (Morris et al., 1998) e, mais recentemente, com o uso de algoritmos de Aprendizado de Máquina (AM), há a possibilidade de um refinamento em sua utilização através da atribuição de novos pesos a seus termos, sendo esta abordagem classificada como "técnica de ajuste fino". Há ainda uma segunda classificação ou nova família de Funções Escore, desenvolvidas por meio de AM não-paramétrica (Heck et al., 2017), onde são calculados termos de energia a partir de uma equação e, para cada termo, são dados pesos que vão determinar a criação de novas funções matemáticas.

1.2.1. *AutoDock4*

O *AutoDock4* é uma ferramenta para realização de docagem molecular desenvolvida por Morris et al. (2009), que utiliza como parte de seu processo de docagem uma Função Escore baseada em um campo de força semi-empírico de energia livre para o cálculo dos termos de energia. Sua equação tem por pretensão abranger a maior parte dos possíveis efeitos citados por Jain (2006), que possam vir a impactar o valor da energia obtido. Ela está demonstrada abaixo:

$$V = \gamma_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \quad + \quad \gamma_{HB} \sum_{i,j} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \quad + \quad \gamma_{elec} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} \quad +$$

$$\gamma_{Sol} \sum_{i,j} \left(S_i V_j + S_j V_i \right) e^{\left(-\frac{r_{ij}^2}{2\sigma^2} \right)} \quad + \quad \gamma_{tor} N_{tor}$$

Equação 1. Equação utilizada pelo programa *AutoDock4*. A equação inclui os termos de energia para interações de Van der Waals, Ligações de Hidrogênio, interações Carga-Carga, Solvatação e Torção. O peso relativo a cada termo é representado por γ (Bitencourt-Ferreira et al., 2020).

O primeiro termo diz respeito às interações de Van der Waals, que estima o potencial de Lennard-Jones, e o segundo às ligações de hidrogênio, aplicando uma adaptação do potencial de Lennard-Jones com uma equação que utiliza expoentes 12 e 10. O terceiro termo corresponde às interações eletrostáticas carga-carga, o quarto termo representa a solvatação e último o termo diz respeito à torção, ou seja, o número

de ligações rotáveis na molécula de interesse (Bitencourt-Ferreira et al., 2020). O cálculo destes termos é efetuado para interações relevantes (dentro de uma área específica) no complexo proteína-ligante e leva em consideração as coordenadas atômicas e a distância euclidiana entre os pares de átomos da proteína e do ligante, representados na equação pelas variáveis *i* e *j*, respectivamente (Huey et al., 2007).

1.2.2. *AutoDock Vina*

AutoDock Vina é um programa de código aberto, desenvolvido por Trott e Olson (2010) com o propósito de possibilitar a realização de docagem molecular e triagem virtual com maior velocidade. Para o cálculo dos termos energéticos o *software* utiliza uma Função Escore, que soma as interações intra e intermoleculares, conforme representada pela seguinte equação:

$$c = c_{\text{inter}} + c_{\text{intra}} = \sum_{i<j} h_{t_i t_j}(d_{ij})$$

Equação 2. Equação utilizada pelo programa *AutoDock Vina* (Bitencourt-Ferreira et al., 2020).

Na equação acima, as variáveis *i* e *j* dizem respeito aos átomos do complexo proteína-ligante, representados por t_i e t_j, respectivamente. O termo $h_{t_i t_j}$ corresponde à soma ponderada das interações estéricas, idênticas para todos os pares de átomos, interações hidrofóbicas entre átomos hidrofóbicos e ligações de hidrogênio, onde aplicáveis (Trott; Olson, 2010). Os pesos utilizados na soma ponderada são mostrados abaixo:

$$\text{gauss}_1(d) = e^{-(d/0.5\,\text{Å})^2}$$

$$\text{gauss}_2(d) = e^{-((d-3\,\text{Å})/2\,\text{Å})^2}$$

$$\text{repulsion}(d) = \begin{cases} d^2, & \text{if } d < 0 \\ 0, & \text{if } d \geq 0 \end{cases}$$

Equação 3. Pesos utilizados no programa *AutoDock Vina* (Trott; Olson, 2010).

Na equação acima, o termo d_{ij} representa a superfície de distância, que é dado pela diferença entre a distância interatômica (r_{ij}) e o raio de Van der Waals (R_t) para os átomos do tipo *t*.

$$d_{ij} = r_{ij} - R_{t_i} - R_{t_j}$$

Equação 4. Superfície de distância utilizada na equação do programa *AutoDock Vina* (Trott; Olson, 2010).

As interações hidrofóbicas são representadas no programa *AutoDock Vina* por funções lineares por pares e, quando ambos os átomos são hidrofóbicos, a primeira equação apresentada a seguir *(hydrophobic)* é adicionada à soma das interações estéricas. Quando o par de átomos é formado por um átomo de hidrogênio doador e um átomo receptor, a segunda equação abaixo (*Hbond*) é utilizada (Bitencourt-Ferreira et al., 2020).

$$\text{hydrophobi c}(d) = \begin{cases} 1, & \text{if } d < 0.5\text{Å} \\ \text{linearly interpolat ed} & \text{if } 0.5\text{Å} < d < 1.5\text{Å} \\ 0, & \text{if } d > 1.5\text{Å} \end{cases}$$

$$\text{Hbond}(d) = \begin{cases} 1, & \text{if } d < -0.7\text{Å} \\ \text{similarly linearly interpolated} & \text{if } -0.7\text{Å} < d < 0 \\ 0, & \text{if } d > 0 \end{cases}$$

Equação 5. Equações utilizadas no programa *AutoDock Vina* para equilibrar a soma das interações estéricas (Bitencourt-Ferreira et al., 2020).

1.2.3. Taba

O programa Taba (Da Silva, et al., 2020) foi desenvolvido com o intuito de gerar novas Funções Escore para predição de afinidade com base em métodos de Aprendizado de Máquina Supervisionado (AMS). Assim como os programas *AutoDock4 e AutoDock Vina*, termos de energia são calculados para serem utilizados posteriormente, no entanto, a equação utilizada é diferente dos demais programas,

sendo apresentada com detalhes mais abaixo. Para tal, o *software* interpreta as interações proteína-ligante como um sistema de massa-mola. Para esta abordagem, são consideradas as distâncias intermoleculares médias para fazer uma estimativa de energia potencial, considerando a energia potencial mínima para um par de átomos específicos.

Abaixo está representada a equação geral utilizada pelo programa para predição de afinidade:

$$PBA = \alpha_0 + \sum_i \sum_j \alpha_{i,j} (d_{i,j} - d_{0,i,j})^2$$

Equação 6. Equação utilizada pelo programa Taba para o cálculo dos termos de energia (Da Silva, et al., 2020).

Na equação acima, α_0 representa a constante de regressão e $\alpha_{i,j}$ corresponde ao peso relativo de cada variável explanatória. Estes dois termos são determinados pela aplicação de métodos de AM. A soma dupla é feita sobre todos os átomos da proteína (i) e do ligante (j) dentro de uma determinada área de corte. O termo $d_{0,i,j}$ representa a distância média para um determinado par de átomos i e j, calculado para todas as estruturas do conjunto de dados. Os termos α_0 e $\alpha_{i,j}$ também levam em conta todo o *dataset* e o termo $d_{i,j}$ considera a distância para um determinado par de átomos de uma estrutura específica, não a média de todas (Da Silva, et al., 2020).

O programa Taba implementa também as técnicas de AM capazes de gerar os novos modelos computacionais utilizando os cálculos obtidos anteriormente.

1.3. APRENDIZADO DE MÁQUINA

Os sistemas desenvolvidos com técnicas de AM são programados para que possam aprender com a experiência passada, empregando para tal, um princípio denominado indução, no qual se obtêm conclusões genéricas a partir de um conjunto particular de exemplos (Faceli, 2015). Com isso, os programas podem aprender e melhorar com a experiência e com o tempo, refinando um modelo que pode ser usado para prever novas entradas e questões baseadas no aprendizado anterior.

Muitas são as aplicações dos métodos de AM na resolução de problemas biológicos e relacionados à área da saúde, como por exemplo, o uso de Redes

Neurais para modelagem de dados complexos ou a predição de regiões de ligação de RNA através da abordagem de Máquina de Vetores de Suporte (SVM, do inglês *Support Vector Machine*) (Heck et al., 2017).

No contexto específico da predição de afinidade entre proteína e ligante destaca-se a utilização de técnicas de AMS, que caracteriza-se pelo uso de dados que sejam rotulados e conhecidos para treinamento de um sistema, onde para cada exemplo nos dados de treino há um objeto de entrada e um objeto de saída (Bell, 2015).

Diversos estudos que fazem uso das técnicas de AMS para predição de afinidade de ligação utilizam as estruturas tridimensionais de proteínas complexadas com seus ligantes ativos, que podem ser obtidas através do site PDB (*Protein Data Bank*) (Berman et al., 2000). Para tal, é necessário preparar previamente os dados antes de realizar o treinamento das ferramentas computacionais, a fim de evitar ruídos e o comprometimento da qualidade dos resultados obtidos na geração dos modelos matemáticos.

A partir da determinação e do preparo das estruturas para uso nos programas, o conjunto de dados escolhido é dividido em dois subconjuntos, um de treino, para aprendizagem do sistema, e outro de teste, para validação dos resultados. A divisão normalmente resulta em uma proporção de 70% para treino e 30% para teste, por apresentar maior confiabilidade dependendo do tamanho do conjunto de dados. Esta abordagem tem sido utilizada na calibragem das Funções Escore empíricas (De Azevedo; Dias, 2008).

Os métodos de AMS têm sido empregados tanto como forma de “ajuste fino” para Funções Escore já existentes quanto no desenvolvimento de novas Funções Escore a partir de pesos atribuídos a termos de energia calculados (Heck et al., 2017). A atribuição de peso em algumas destas técnicas é feita através de Regressão Linear, como ocorre, por exemplo, nos programas Taba e *Molegro Data Modeller* (MDM) (CLC BIO, 2013).

O programa MDM consiste em um ambiente integrado para análise e mineração de dados, permitindo a criação de modelos de classificação e regressão. Para a análise de regressão estão disponíveis quatro métodos: *Multiple Linear Regression* (MLR)*, Partial Least Squares* (PLS)*, Support Vector Machines* (SVM) e *Neural Networks* (NN). O algoritmo utilizado para o treinamento das redes neurais é chamado de *back-propagation* (CLC BIO, 2013).

Os métodos de regressão empregados nos programas Taba e MDM seguem a equação geral da Regressão Linear, conforme mostrado na Equação 7.

$$RSS = \sum_{i=1}^{M} (y_i - PA_i)^2 + \lambda_1 \sum_{j=1}^{N} |\omega_j| + \lambda_2 \sum_{j=1}^{N} |\omega_j|^2$$

Equação 7. Equação geral da Regressão Linear (Bitencourt-Ferreira et al., 2020).

Algumas de suas variações dão origem a diferentes métodos, com ajustes para cada caso. Sua explicação detalhada segue abaixo:

- Linear Regression: A Regressão Linear se adapta a um modelo linear com coeficientes w = (w_1, ..., w_p) para minimizar a soma residual de quadrados (RSS, do inglês *Residual Sum of Squares*) entre as respostas observadas no conjunto de dados e as respostas preditas pela aproximação linear (Legendre, 1805). Para a aplicação do método comum de Regressão Linear, os termos λ_1 e λ_2 são igualados a zero. A validação cruzada é frequentemente usada para a escolha dos parâmetros de regularização.
- Lasso: O método Lasso (*Least Absolute Shrinkage and Selection Operator*) é um modelo linear que estima coeficientes esparsos. Na equação geral da Regressão Linear, o método Lasso possui $\lambda_1 > 0$ e $\lambda_2 = 0$. É útil em alguns contextos devido à sua tendência de preferir soluções com um número menor de parâmetros, reduzindo efetivamente o número de variáveis das quais a solução dada é dependente. Por esta razão, o Lasso e suas variantes são fundamentais para o campo de detecção comprimida. Sob certas condições, ele pode recuperar o conjunto exato de pesos não–zero (Tibshirani, 1996).
- LassoCV: O LassoCV implementa a regressão do Lasso com uma iteração apropriada juntamente com um caminho de regularização. O melhor modelo é selecionado por validação cruzada (Tibshirani, 1996).
- Ridge: A regressão feita pelo método Ridge trata de alguns dos problemas dos mínimos quadrados comuns, impondo uma penalidade ao tamanho dos coeficientes. No método Ridge, temos $\lambda_1 = 0$ e $\lambda_2 > 0$, aplicados à equação da Regressão Linear (Tikhonov, 1963).

- RidgeCV: Este método implementa a regressão Ridge com uma validação cruzada pré-construída do parâmetro alfa (Tikhonov, 1963).
- ElasticNet: Trata-se de um modelo de regressão linear treinado com L1 e L2 (termos que representam penalidades) como regularizadores. Esta combinação permite aprender um modelo esparso onde alguns dos pesos não são zero como no Lasso, enquanto ainda mantém as propriedades de regularização do Ridge. A ideia por trás do método ElasticNet é a combinação dos métodos de regressão Lasso e Ridge, com $\lambda_1 > 0$ e $\lambda_2 > 0$ aplicados na equação geral de Regressão Linear. É útil quando existem várias características que estão correlacionadas entre si. É provável que o Lasso escolha um desses ao acaso, enquanto que o ElasticNet provavelmente escolha os dois (Zou; Hastie, 2005).
- ElasticNetCV: O ElasticNetCV implementa a regressão do ElasticNet com uma iteração apropriada juntamente com um caminho de regularização. O melhor modelo é selecionado por validação cruzada (Zou; Hastie, 2005).

2. JUSTIFICATIVA

O desenvolvimento de métodos computacionais para avaliar as interações proteína-ligante tem um impacto significativo nos estágios iniciais do descobrimento de fármacos. A avaliação de interações intermoleculares usando as coordenadas atômicas de complexos de proteína-ligante é de fundamental importância para entender as bases estruturais para a especificidade de inibidores em relação a um alvo (Labute, 2018).

Nas últimas décadas, o número de estruturas tridimensionais de macromoléculas (proteínas ou ácidos nucleicos) resolvidas por cristalografia e por difração de raios X tem aumentado consideravelmente, enriquecendo os bancos de dados que armazenam estas informações, como o PDB e o *PDBbind* (Wang et al., 2004), e potencializando as oportunidades de pesquisa com o uso destes dados. Da mesma forma, a capacidade computacional tem crescido em ritmo acelerado, o que confere um avanço significativo na manipulação e processamento destes dados, permitindo que sejam construídos modelos biológicos mais precisos e complexos de forma progressiva. A disponibilização de outras informações associadas à estrutura das macromoléculas, como a informação experimental de afinidade com seus possíveis ligantes (Constante de Inibição - K_i, Constante de Dissociação - K_d, Concentração Inibitória Máxima a 50% - IC_{50}, Concentração Efetiva Máxima a 50% - EC_{50} ou Energia Livre de Gibbs - ΔG), permite que sejam empregadas abordagens computacionais robustas, como por exemplo, as técnicas de Dinâmica Molecular (DM) e de AM na predição desta informação experimental através de modelos matemáticos. Simulações por DM podem gerar metodologias confiáveis para a predição de afinidade, no entanto, embora seja uma abordagem clássica para este fim, tais cálculos ainda custam muito, em termos de processamento computacional e de tempo (Bitencourt-Ferreira; De Azevedo Jr., 2018).

Dentre os métodos de AM empregados na previsão de afinidade de ligação destaca-se o desenvolvimento de Funções Escore por meio da técnica de AM não-paramétrica, onde interações intermoleculares relevantes entre ligante e proteína são extraídas, de modo a ser gerada uma nova Função Escore otimizada com base nestas interações, que tenta explicar matematicamente o poder de interação entre o receptor e sua molécula inibidora ou ativadora.

Com base nos paradigmas estabelecidos de ”Espaço de Proteínas” (Bohacek et al., 1996) e “Espaço Químico” (Dobson, 2004), sendo o primeiro composto por proteínas presentes na natureza e o segundo por pequenas moléculas que poderiam ligar-se àquelas proteínas, é possível imaginar um “Espaço de Funções Escore”, que seria um “espaço matemático” composto por infinitos modelos computacionais utilizados para esta predição de afinidade de ligação, onde para cada elemento no “Espaço de Proteínas”, haveria uma Função Escore que seria capaz de predizer sua afinidade com as moléculas de um subconjunto do “Espaço Químico”, servindo assim, de conexão entre eles (Heck et al*.,* 2017). Esta abstração pode ser melhor entendida através da Figura 1.

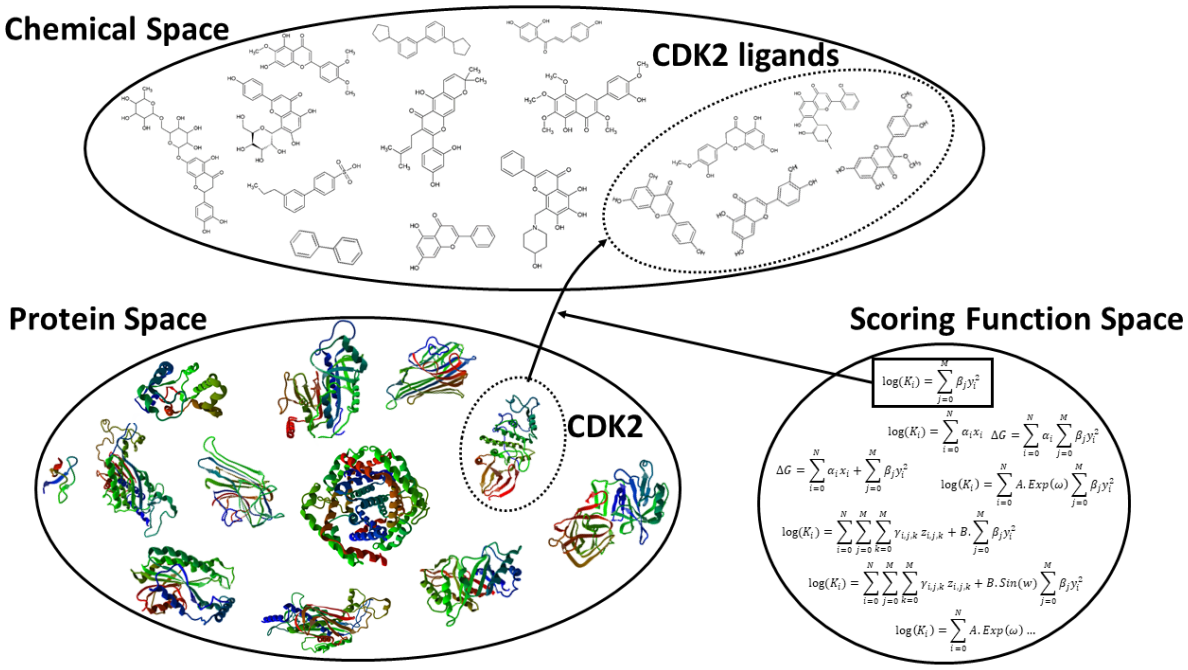


Figura 1. Relação entre Espaço de Proteínas e Espaço Químico, mediado pelo Espaço de Funções Escore. Para cada elemento do Espaço de Proteínas, existe uma Função Escore que o liga a elementos do Espaço Químico (Bitencourt-Ferreira *et at*., 2020).

Técnicas de AMS possuem um grande potencial para explorar o “Espaço de Funções Escore” e encontrar um modelo adequado para previsão de afinidade, já tendo sido utilizadas com sucesso em estudos prévios (Bitencourt-Ferreira; De Azevedo Jr, 2018; De Ávila et al., 2017).

A proposta deste trabalho é desenvolver Funções Escore otimizadas para a predição de inibição de serino/treonino cinases não específicas (EC 2.7.11.1), dada

sua importância como alvo para o desenvolvimento de fármacos, através do uso de duas abordagens distintas, chamadas de “Abordagem A” e “Abordagem B”, explicadas na seção 4. Estas podem ser vistas como formas de explorar o “Espaço de Funções Escore” a fim de encontrar um modelo computacional adequado para prever a inibição de proteínas desta classe enzimática.

3. OBJETIVO GERAL

Explorar o “Espaço de Funções Escore” a fim de encontrar um modelo computacional adequado que possa explicar matematicamente a relação entre as proteínas da classe enzimática EC 2.7.11.1 (serino/treonino cinases não específicas) e seus possíveis ligantes, para prever sua afinidade de ligação.

3.1. OBJETIVOS ESPECÍFICOS

1. Realizar o cálculo da predição de afinidade de ligação entre proteínas da classe enzimática EC 2.7.11.1 e seus ligantes através de Funções Escore clássicas utilizadas nos programas *AutoDock4* e *AutoDock Vina* (corresponde à “Abordagem A”).
2. Utilizar os termos calculados previamente para treinamento de métodos de AMS implementados no programa MDM para geração de novas Funções Escore otimizadas para esta família de proteínas (corresponde à “Abordagem A”).
3. Utilizar o programa Taba para cálculo dos termos de energia e treinamento de métodos de AMS implementados na própria ferramenta (corresponde à “Abordagem B”).
4. Aplicar sobre um subconjunto de teste o melhor modelo gerado na etapa de treino para validar o modelo (correspondente a ambas as abordagens).
5. Comparar os resultados gerados pelas Funções Escore otimizadas e os resultados gerados por Funções Escore clássicas, de modo a aferir o melhor modelo para predição de afinidade de ligação (correspondente a ambas as abordagens).

4. MÉTODOS

A partir dos objetivos apresentados, a metodologia utilizada para o desenvolvimento desta pesquisa será detalhada nas subseções que seguem.

4.1. SELEÇÃO E PREPARAÇÃO DOS DADOS

Com o crescente volume de informações disponíveis nas bases de dados a cada dia, faz-se necessário o emprego de técnicas de limpeza e organização destes dados antes que eles possam ser utilizados pelos algoritmos de AM. Em alguns casos, a presença de informações duplicadas ou falta de atributos, por exemplo, pode causar distorção nos resultados, e, por isso, técnicas de pré-processamento e filtragem são empregadas (Faceli, 2015).

O site PDB reúne dados experimentais de outras três bases de dados: *BindingDB* (THE BINDING DATABASE, 2019), *MOAD* (Binding MOAD, 2019) e *PDBbind* (Wang et al., 2004). Desta forma, as informações de afinidade disponíveis são robustas, entretanto, algumas estruturas podem apresentar ligantes repetidos dentro de um conjunto de interesse.

Para a seleção e montagem do *dataset* utilizado neste trabalho, foi realizada uma busca no site do PDB por estruturas cristalográficas de serino/treonino cinases não específicas (classe enzimática EC 2.7.11.1), resolvidas pelo método de difração de raios X e para as quais a informação sobre a constante de inibição (K_i) estivesse disponível. As estruturas foram então submetidas a uma etapa de preparação em cada abordagem, “A” e “B”, conforme descrito a seguir.

4.2. GERAÇÃO DE NOVAS FUNÇÕES ESCORE – “ABORDAGEM A”

Esta abordagem, utilizada para a geração de novos modelos de Funções Escore a partir de funções clássicas e técnicas de AMS, teve como base a combinação dos programas *AutoDock4, AutoDock Vina,* SAnDReS (Xavier, et al. 2016) e MDM.

Primeiramente, a preparação das estruturas consistiu na filtragem dos arquivos com uso do programa SAnDReS, um *software* que integra diferentes técnicas para realizar docagem molecular e calcular a afinidade de ligação. Após o *download* das estruturas, as moléculas de água foram retiradas do complexo, os ligantes que apareciam mais de uma vez foram removidos e o conjunto inicial foi separado em dois,

um subconjunto de treino composto por 70% das estruturas e um de teste, abrangendo os 30% restantes. Esta divisão foi feita utilizando-se uma semente aleatória.

Além disso, as coordenadas atômicas do ligante e da proteína foram separadas em arquivos distintos em formato *pdbqt,* pois este tipo de arquivo é necessário para uso no programa *AutoDock4*.

A próxima etapa consistiu no cálculo dos termos de energia das interações inter e intramoleculares das estruturas por meio dos programas *AutoDock4* e *AutoDock Vina,* de forma independente. As cargas atômicas das proteínas foram mantidas com os valores padrão usados nos dois programas.

Os dados de treino de cada programa foram utilizados separadamente como variáveis explanatórias para os métodos de AMS no programa MDM, sendo empregadas quatro técnicas para análise de Regressão Linear: MLS (*Multiple Linear Regression*), PLS (*Partial Least Squares*), SVM (*Support Vector Machines*) e NN (*Neural Networks*). Para este último método, foram usadas duas camadas ocultas (*hidden layers*) e o número de neurônios em cada camada variou de 1 a 10.

Após o treinamento dos métodos de AM e geração dos modelos para cada um dos programas, o modelo com a melhor correlação de cada um foi aplicado no conjunto de teste para a validação dos resultados.

4.3. GERAÇÃO DE NOVAS FUNÇÕES ESCORE – “ABORDAGEM B”

A segunda opção para geração de modelos computacionais utiliza os métodos desenvolvidos no programa Taba, de modo que, todas as etapas de *download,* filtragem, separação do *dataset*, cálculo dos termos de energia e geração dos modelos por técnicas de AMS descritos na “Abordagem A”, foram feitas utilizando-se unicamente o Taba nesta segunda abordagem.

Na etapa do processo em que é realizado o cálculo dos termos de energia, o programa Taba utiliza uma equação específica, apresentada na seção 1.2.3. As opções de Regressão Linear e suas variações, implementadas no *software,* foram explicadas na seção 1.3 e todos os sete diferentes métodos foram usados neste estudo para a busca pelo melhor modelo matemático.

O programa Taba escolhe os termos a serem usados como variáveis explanatórias pela maior correlação entre eles e a informação experimental, bastando apenas que se indique a quantidade de variáveis desejadas para o cálculo. É preciso

também indicar a distância de corte desejada, dentro da qual serão executados os cálculos de distância para os pares de átomos específicos. Nesta pesquisa, foram aplicadas as seguintes distâncias de corte: 3.5, 4.5, 6.0, 9.0 e 12.0 Ångstrons.

Em ambas as abordagens (“A” e “B”) aplicadas neste trabalho, a equação polinomial para avaliação do log(K_i) dos complexos de proteína-ligante do *dataset* utiliza 5 variáveis explanatórias e tem a seguinte forma geral:

$$y = \lambda_0 + \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4 + \lambda_5 x_5$$

Equação 8. Equação polinomial geral utilizada na criação das Funções Escore otimizadas.

A variável λ_0 representa a constante de regressão e os demais λ são os pesos relativos a cada variável explanatória. A variável resultante da soma é *y*, ou seja, o logaritmo da Constante de Inibição (log(K_i)).

Com relação à escolha de quais variáveis seriam utilizadas para compor as novas equações, na primeira abordagem, foram selecionados aleatoriamente 5 dos termos energéticos disponibilizados por cada programa, tendo em vista algumas combinações realizadas entre os termos que obtiveram maior correlação em relação ao dado experimental. Na aplicação da “Abordagem B”, alternou-se o número de variáveis explanatórias de 2 a 5.

4.4. ANÁLISE ESTATÍSTICA

Em geral, para a validação dos resultados obtidos por meio de métodos computacionais, é necessário que seja feita uma comparação dos dados resultantes da simulação com as informações obtidas experimentalmente.

A etapa final para ambas as abordagens consiste em uma análise estatística que tem por objetivo validar a capacidade de previsão de afinidade dos modelos gerados, por meio da comparação dos valores calculados pelos novos modelos com os valores de informação experimental (K_i, por exemplo) retirados das estruturas tridimensionais. Para tal, foram realizadas análises baseadas no cálculo do erro quadrático médio da raiz (RMSE), do desvio padrão (SD) e dos coeficientes de correlação de Spearman (ρ) e Pearson (R^2) (Zar, 1972), além dos valores de p-value para Spearman e p-value para Pearson, onde o valor da energia calculado pelos

novos modelos gerados é comparado com o informado como dado experimental junto às estruturas utilizadas (Heck et al., 2017). Quanto mais próximos estejam os dois valores, melhores os resultados.

Pode-se afirmar, assim, que o modelo matemático gerado é o melhor - ou seja, é o que idealmente pode explicar de forma matemática a afinidade de ligação entre proteína e ligante - quando possui um valor de correlação mais próximo de 1 (em uma escala de -1 a 1), para os coeficientes de Spearman e de Pearson.

A Figura 2 apresenta um fluxograma contendo as etapas descritas até aqui para a geração de modelos computacionais a partir de um *dataset* montado e separado em dois subconjuntos, com uso associado de Funções Escore clássicas e técnicas de AMS.

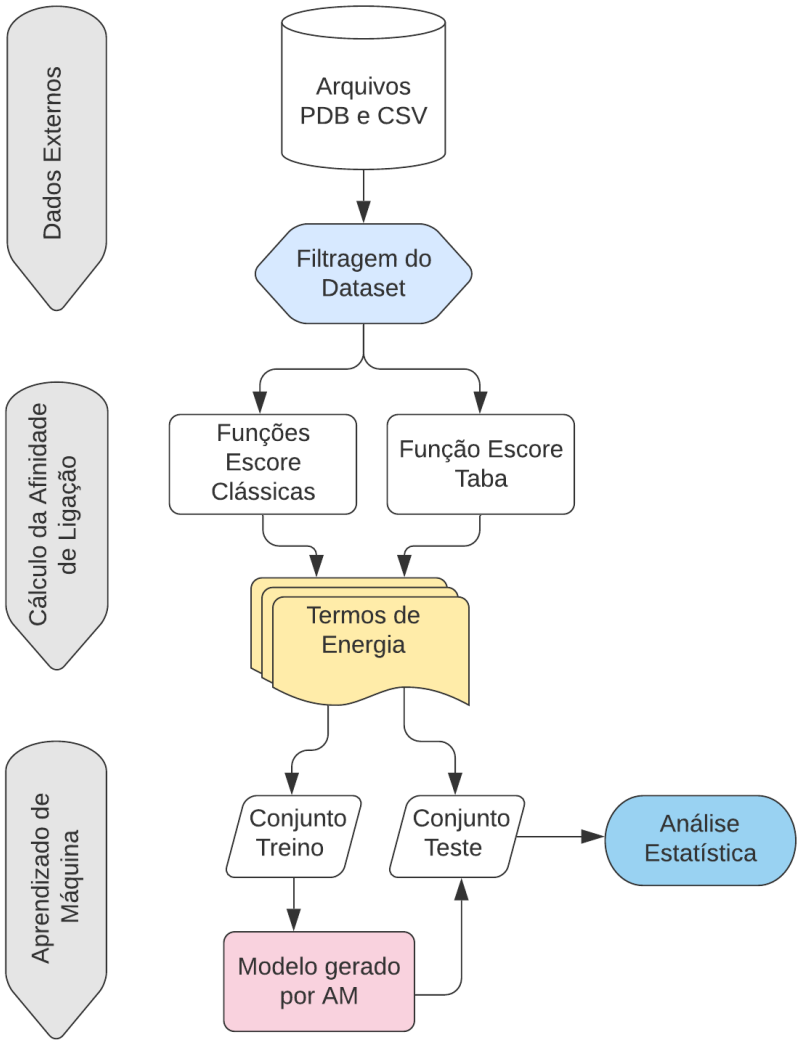


Figura 2. Etapas realizadas nas abordagens "A" e "B" para a geração das Funções Escore otimizadas. Fonte: A autora (2020).

Após a geração do melhor modelo matemático pelos métodos de AMS para um determinado conjunto de proteínas, tendo passado pelas etapas de treino e teste, é possível utilizá-lo na previsão de afinidade de outras proteínas e ligantes de interesse relacionados à família específica de proteínas que gerou o modelo.

5. RESULTADOS E DISCUSSÃO

Nesta seção são apresentados os resultados obtidos por meio de ambas as abordagens e os melhores modelos encontrados para predição de afinidade.

5.1. CONJUNTO DE DADOS

A busca no site PDB retornou inicialmente 156 estruturas de proteínas complexadas com ligantes. Após a filtragem, o conjunto de dados final resultou em 97 estruturas únicas (sem ligantes repetidos). Os Quadros 1 e 2, a seguir, apresentam as informações relativas a cada estrutura do *dataset,* a partir de seu código de acesso no PDB, para os subconjuntos de treino e teste, respectivamente.

Quadro 1. Dados referentes às proteínas e ligantes que compõem o conjunto de treino.

Código PDB	Proteína	Organismo	Ligante	Valor $\log(K_i)$	Resolução (Å)
1M2R	CK2	*Zea mays*	5,8-DI-AMINO-1,4-DIHYDROXY-ANTHRAQUINONE	-6.45593	1.70
1NVQ	CHK1	*Homo sapiens*	7-HYDROXYSTAUROSPORINE	-8.25181	2.00
1W1D	PDK1	*Homo sapiens*	INOSITOL-(1,3,4,5)-TETRAKISPHOSPHATE	-6.52288	1.50
1W1G	PDK1	*Homo sapiens*	(2R)-3-{[(S)-{[(2S,3R,5S,6S)-2,6-DIHYDROXY-3,4,5-TRIS(PHOSPHONOOXY)CYCLOHEXYL]OXY}(HYDROXY)PHOSPHORYL]OXY}-2-(1-HYDROXYBUTOXY)PROPYL BUTYRATE	-7.67778	1.45
1ZOE	CK2	*Zea mays*	4,5,6,7-TETRABROMO-N,N-DIMETHYL-1H-BENZIMIDAZOL-2-AMINE	-7.39794	1.77
1ZOH	CK2	*Zea mays*	5,6,7,8-TETRABROMO-1-METHYL-2,3-DIHYDRO-1H-IMIDAZO[1,2-A]BENZIMIDAZOLE	-7	1.81
2BR1	CHK1	*Homo sapiens*	2-[5,6-BIS-(4-METHOXY-PHENYL)-FURO[2,3-D]PYRIMIDIN-4-YLAMINO]-ETHANOL	-5.14267	2.00
2BRB	CHK1	*Homo sapiens*	2-[(5,6-DIPHENYLFURO[2,3-D]PYRIMIDIN-4-YL)AMINO]ETHANOL	-4.86328	2.10
2BRM	CHK1	*Homo sapiens*	3-AMINO-3-BENZYL-[4.3.0]BICYCLO-1,6-DIAZANONAN-2-ONE	-5.88606	2.20
2C3J	CHK1	*Homo sapiens*	DEBROMOHYMENIALDISINE	-6.18111	2.10
2C3L	CHK1	*Homo sapiens*	3-(1H-BENZIMIDAZOL-2-YL)-1H-INDAZOLE	-5.24195	2.35
2E9N	CHK1	*Homo sapiens*	3-(4'-HYDROXYBIPHENYL-4-YL)-N-(4-HYDROXYCYCLOHEXYL)-1,4-DIHYDROINDENO[1,2-C]PYRAZOLE-6-CARBOXAMIDE	-8.20066	2.50

2E9P	CHK1	*Homo sapiens*	1-(5-CHLORO-2-METHOXYPHENYL)-3-{6-[2-(DIMETHYLAMINO)-1-METHYLETHOXY]PYRAZIN-2-YL}UREA	-7.69897	2.60
2E9U	CHK1	*Homo sapiens*	18-CHLORO-11,12,13,14-TETRAHYDRO-1H,10H-8,4-(AZENO)-9,15,1,3,6-BENZODIOXATRIAZACYCLOHEPTADECIN-2-ONE	-8.10127	2.00
2E9V	CHK1	*Homo sapiens*	18-CHLORO-2-OXO-17-[(PYRIDIN-4-YLMETHYL)AMINO]-2,3,11,12,13,14-HEXAHYDRO-1H,10H-4,8-(AZENO)-9,15,1,3,6-BENZODIOXATRIAZACYCLOHEPTADECINE-7-CARBONITRILE	-7.89296	2.00
2ESM	ROCK1	*Homo sapiens*	5-(1,4-DIAZEPAN-1-SULFONYL)ISOQUINOLINE	-6.41266	3.20
2FAP	FKBP12	*Homo sapiens*	C49-METHYL RAPAMYCIN	-8.39794	2.20
2GHG	CHK1	*Homo sapiens*	5-{5-[(S)-2-AMINO-3-(1H-INDOL-3-YL)-PROPOXYL]-PYRIDIN-3-YL}-3-[1-(1H-PYRROL-2-YL)-METH-(Z)-YLIDENE]-1,3-DIHYDRO-INDOL-2-ONE	-6.89963	3.50
2NRU	IRAK-4	*Homo sapiens*	1-(3-HYDROXYPROPYL)-2-[(3-NITROBENZOYL)AMINO]-1H-BENZIMIDAZOL-5-YL PIVALATE	-8.92082	2.00
2OXD	CK2	*Zea mays*	4,5,6,7-TETRABROMO-1H,3H-BENZIMIDAZOL-2-ONE	-6.82391	2.30
2OXX	CK2	*Zea mays*	4,5,6,7-TETRABROMO-1H,3H-BENZIMIDAZOL-2-THIONE	-6.69897	2.30
2PVH	CK2	*Zea mays*	N,N'-DIPHENYLPYRAZOLO[1,5-A][1,3,5]TRIAZINE-2,4-DIAMINE	-6.58503	2.20
2PVJ	CK2	*Zea mays*	2-(CYCLOHEXYLMETHYLAMINO)-4-(PHENYLAMINO)PYRAZOLO[1,5-A][1,3,5]TRIAZINE-8-CARBONITRILE	-8.18709	1.70
2PVM	CK2	*Zea mays*	4-(2-(1H-IMIDAZOL-4-YL)ETHYLAMINO)-2-(PHENYLAMINO)PYRAZOLO[1,5-A][1,3,5]TRIAZINE-8-CARBONITRILE	-6.4437	2.00
2PVN	CK2	*Zea mays*	N-(3-(8-CYANO-4-(PHENYLAMINO)PYRAZOLO[1,5-A][1,3,5]TRIAZIN-2-YLAMINO)PHENYL)ACETAMIDE	-9.45593	2.00
2UVM	PKBalpha (PKB/Akt)	*Homo sapiens*	BENZENE-1,2,3,4-TETRAYL TETRAKIS[DIHYDROGEN (PHOSPHATE)]	-7.09691	1.94
2WTV	Aurora-A	*Homo sapiens*	4-{[9-CHLORO-7-(2,6-DIFLUOROPHENYL)-5H-PYRIMIDO[5,4-D][2]BENZAZEPIN-2-YL]AMINO}BENZOIC ACID	-8.45182	2.40
2ZJW	CK2	*Homo sapiens*	2,3,7,8-tetrahydroxychromeno[5,4,3-cde]chromene-5,10-dione	-7.69897	2.40
3BE9	CK2	*Zea mays*	19-(cyclopropylamino)-4,6,7,15-tetrahydro-5H-16,1-(azenometheno)-10,14-(metheno)pyrazolo[4,3-o][1,3,9]triazacyclohexadecin-8(9H)-one	-7.61979	2.00
3BGP	PIM-1	*Homo sapiens*	4-[3-(4-chlorophenyl)-2,1-benzisoxazol-5-yl]pyrimidin-2-amine	-7.04096	2.80
3BGQ	PIM-1	*Homo sapiens*	N-cyclohexyl-3-[3-(trifluoromethyl)phenyl][1,2,4]triazolo[4,3-b]pyridazin-6-amine	-7.95861	2.00
3BGZ	PIM-1	*Homo sapiens*	2,3-diphenyl-1H-indole-7-carboxylic acid	-6.25964	2.40

3BQC	CK2	*Homo sapiens*	3-METHYL-1,6,8-TRIHYDROXYANTHRAQUINONE	-5.73283	1.50
3C4C	B-RAF	*Homo sapiens*	N-{3-[(5-chloro-1H-pyrrolo[2,3-b]pyridin-3-yl)carbonyl]-2,4-difluorophenyl}propane-1-sulfonamide	-8.58503	2.57
3D0E	AKT2	*Homo sapiens*	4-{2-(4-amino-1,2,5-oxadiazol-3-yl)-1-ethyl-7-[(3S)-piperidin-3-ylmethoxy]-1H-imidazo[4,5-c]pyridin-4-yl}-2-methylbut-3-yn-2-ol	-8.52288	2.00
3E88	AKT2	*Homo sapiens*	4-[2-(4-amino-1,2,5-oxadiazol-3-yl)-6-{[(2R)-2-amino-3-phenylpropyl]oxy}-1-ethyl-1H-imidazo[4,5-c]pyridin-4-yl]-2-methylbut-3-yn-2-ol	-8.85387	2.50
3EQR	ACK1	*Homo sapiens*	N~3~-(2,6-dimethylphenyl)-1-(3-methoxy-3-methylbutyl)-N~6~-(4-piperazin-1-ylphenyl)-1H-pyrazolo[3,4-d]pyrimidine-3,6-diamine	-8.39794	2.00
3FL5	CK2	*Zea mays*	1,2,5,8-tetrahydroxyanthracene-9,10-dione	-7.284	2.30
3IDP	B-RAF	*Homo sapiens*	N~1~-(4-chlorophenyl)-6-methyl-N~5~-[3-(7H-purin-6-yl)pyridin-2-yl]isoquinoline-1,5-diamine	-9	2.70
3JVR	CHK1	*Homo sapiens*	(1S)-1-(1H-benzimidazol-2-yl)ethyl (3,4-dichlorophenyl)carbamate	-5.72354	1.76
3JVS	CHK1	*Homo sapiens*	2-[(4-tert-butyl-3-nitrophenyl)carbonyl]-N-naphthalen-1-ylhydrazinecarboxamide	-6.83565	1.90
3OWJ	CK2	*Homo sapiens*	9-hydroxy-5,11-dimethyl-4,6-dihydro-1H-pyrido[4,3-b]carbazol-1-one	-6.0655	1.85
3PE1	CK2	*Homo sapiens*	5-[(3-chlorophenyl)amino]benzo[c][2,6]naphthyridine-8-carboxylic acid	-9.52071	1.60
3PE2	CK2	*Homo sapiens*	5-[(3-ethynylphenyl)amino]pyrimido[4,5-c]quinoline-8-carboxylic acid	-9.75696	1.90
3PWD	CK2	*Zea mays*	8-hydroxy-4-methyl-9-nitro-2H-benzo[g]chromen-2-one	-6.65758	2.20
3U9C	CK2	*Homo sapiens*	7-hydroxy-3H-phenoxazin-3-one	-5.88606	3.20
4FBX	CK2	*Homo sapiens*	N-(6-oxohexyl)-2-(4,5,6,7-tetrabromo-1H-benzimidazol-1-yl)acetamide	-5.83863	2.33
4FSN	CHK1	*Homo sapiens*	4-(6-{[(4-METHYLCYCLOHEXYL)AMINO]METHYL}-1,4-DIHYDROINDENO[1,2-C]PYRAZOL-3-YL)BENZOIC ACID	-7.69897	2.10
4FST	CHK1	*Homo sapiens*	4-[(6,7-dimethoxy-2,4-dihydroindeno[1,2-c]pyrazol-3-yl)ethynyl]-2-methoxyphenol	-8.29158	1.90
4FSW	CHK1	*Homo sapiens*	8-chloro-5,10-dihydro-11H-dibenzo[b,e][1,4]diazepin-11-one	-4.78187	2.30
4FSY	CHK1	*Homo sapiens*	2-(11-oxo-10,11-dihydro-5H-dibenzo[b,e][1,4]diazepin-3-yl)benzamide	-6.40012	2.30
4FT0	CHK1	*Homo sapiens*	2-methoxy-4-(11-oxo-10,11-dihydro-5H-dibenzo[b,e][1,4]diazepin-3-yl)benzoic acid	-7.10846	2.30
4FT5	CHK1	*Homo sapiens*	1-{5-chloro-2-[(3R)-pyrrolidin-3-yloxy]phenyl}-3-(5-cyanopyrazin-2-yl)urea	-7.5817	2.40

4FT7	CHK1	*Homo sapiens*	1-{5-bromo-2-[(3R)-3-hydroxypiperidin-1-yl]phenyl}-3-(5-cyanopyrazin-2-yl)urea	-8.92812	2.20
4FTA	CHK1	*Homo sapiens*	6-{[(5-cyanopyrazin-2-yl)carbamoyl]amino}naphthalene-2-carboxylic acid	-7.07988	2.40
4FTT	CHK1	*Homo sapiens*	methyl [3-(1-methyl-1H-imidazol-5-yl)-11-oxo-10,11-dihydro-5H-dibenzo[b,e][1,4]diazepin-8-yl]acetate	-7.11182	2.30
4FTU	CHK1	*Homo sapiens*	methyl [11-oxo-3-(pyridin-4-ylamino)-10,11-dihydro-5H-dibenzo[b,e][1,4]diazepin-8-yl]acetate	-7.96859	2.10
4K0Y	PIM-1	*Homo sapiens*	N-(4-fluorophenyl)-7-hydroxy-5-(piperidin-4-yl)pyrazolo[1,5-a]pyrimidine-3-carboxamide	-7.82391	1.95
4K18	PIM-1	*Homo sapiens*	5-(4-cyanobenzyl)-N-(4-fluorophenyl)-7-hydroxypyrazolo[1,5-a]pyrimidine-3-carboxamide	-8.95861	2.05
4K1B	PIM-1	*Homo sapiens*	N-[5-(2-fluorophenyl)-1H-pyrrolo[2,3-b]pyridin-3-yl]-5-({[(3R,4R)-3-fluoropiperidin-4-yl]methyl}amino)pyrazolo[1,5-a]pyrimidine-3-carboxamide	-1.15229	2.08
4KWP	CK2	*Homo sapiens*	4,5,6,7-tetrabromo-1-(2-deoxy-beta-D-erythro-pentofuranosyl)-1H-benzimidazole	-7.39794	1.25
4MNF	B-RAF	*Homo sapiens*	2-{4-[(1E)-1-(hydroxyimino)-2,3-dihydro-1H-inden-5-yl]-3-(pyridin-4-yl)-1H-pyrazol-1-yl}ethanol	-9.88606	2.80
4N70	PIM-1	*Homo sapiens*	N-{4-[(3R,4R,5S)-3-amino-4-hydroxy-5-methylpiperidin-1-yl]pyridin-3-yl}-6-(2,6-difluorophenyl)-5-fluoropyridine-2-carboxamide	-9.60076	2.10
4O0T	P21-activated kinases (PAK)	*Homo sapiens*	1-({1-(2-aminopyrimidin-4-yl)-2-[(2-methoxyethyl)amino]-1H-benzimidazol-6-yl}ethynyl)cyclohexanol	-5.5376	2.60
4O0X	P21-activated kinases (PAK)	*Homo sapiens*	1-{[1-(4-amino-1,3,5-triazin-2-yl)-2-methyl-1H-benzimidazol-6-yl]ethynyl}cyclohexanol	-7.16749	2.48
4ZY4	P21-activated kinases (PAK)	*Homo sapiens*	2-(4-aminopiperidin-1-yl)-N-(5-cyclopropyl-1H-pyrazol-3-yl)thieno[3,2-d]pyrimidin-4-amine	-7.65758	2.60
5DIA	PIM-1	*Homo sapiens*	(1S,3S)-N-{6-[5-(pyridin-3-yl)-1H-pyrazolo[3,4-c]pyridin-3-yl]pyridin-2-yl}cyclohexane-1,3-diamine	-9.40121	1.96

Fonte: A autora (2020).

Quadro 2. Dados referentes às proteínas e ligantes que compõem o conjunto de teste.

Código PDB	Proteína	Organismo	Ligante	Valor log(K_i)	Resolução (Å)
1M2P	CK2	*Zea mays*	1,8-DI-HYDROXY-4-NITRO-ANTHRAQUINONE	-6.10791	2.00
1M2Q	CK2	*Zea mays*	1,8-DI-HYDROXY-4-NITRO-XANTHEN-9-ONE	-6.09691	1.79
1NVR	CHK1	*Homo sapiens*	STAUROSPORINE	-8.10791	1.80

1NVS	CHK1	*Homo sapiens*	REL-(9R,12S)-9,10,11,12-TETRAHYDRO-9,12-EPOXY-1H-DIINDOLO[1,2,3-FG:3',2',1'-KL]PYRROLO[3,4-I][1,6]BENZODIAZOCINE-1,3(2H)-DIONE	-7.82391	1.80
1OM1	CK2	*Zea mays*	(5-OXO-5,6-DIHYDRO-INDOLO[1,2-A]QUINAZOLIN-7-YL)-ACETIC ACID	-6.76955	1.68
1ZOG	CK2	*Zea mays*	4,5,6,7-TETRABROMO-2-(METHYLSULFANYL)-1H-BENZIMIDAZOLE	-7.1549	2.30
2C3K	CHK1	*Homo sapiens*	4-[3-(1H-BENZIMIDAZOL-2-YL)-1H-INDAZOL-6-YL]-2-METHOXYPHENOL	-7.58503	2.60
2CSN	CK1	*Schizosaccharomyces pombe*	N-(2-AMINOETHYL)-5-CHLOROISOQUINOLINE-8-SULFONAMIDE	-4.40894	2.50
2ETR	ROCK1	*Homo sapiens*	(R)-TRANS-4-(1-AMINOETHYL)-N-(4-PYRIDYL) CYCLOHEXANECARBOXAMIDE	-6.96508	2.60
2OXY	CK2	*Zea mays*	4,5,6,7-TETRABROMO-BENZIMIDAZOLE	-6.52288	1.81
2PVK	CK2	*Zea mays*	2-(4-CHLOROBENZYLAMINO)-4-(PHENYLAMINO)PYRAZOLO[1,5-A][1,3,5]TRIAZINE-8-CARBONITRILE	-8.18709	1.90
2PVL	CK2	*Zea mays*	2-(4-ETHYLPIPERAZIN-1-YL)-4-(PHENYLAMINO)PYRAZOLO[1,5-A][1,3,5]TRIAZINE-8-CARBONITRILE	-7.61979	1.90
3H30	CK2	*Homo sapiens*	5,6-dichloro-1-beta-D-ribofuranosyl-1H-benzimidazole	-4.67847	1.56
3O0G	CDK5	*Homo sapiens*	{4-amino-2-[(4-chlorophenyl)amino]-1,3-thiazol-5-yl}(3-nitrophenyl)methanone	-6.22185	1.95
3RWP	PDK1	*Homo sapiens*	[4-amino-7-(propan-2-yl)-7H-pyrrolo[2,3-d]pyrimidin-5-yl](6-{[(3S,4R)-4-(4-fluorophenyl)tetrahydrofuran-3-yl]amino}pyrazin-2-yl)methanone	-9.22185	1.92
3RWQ	PDK1	*Homo sapiens*	[4-amino-7-(propan-2-yl)-7H-pyrrolo[2,3-d]pyrimidin-5-yl](6-{[2-(pyridin-3-yl)ethyl]amino}pyrazin-2-yl)methanone	-7.45593	2.55
4APP	P21-activated kinases (PAK)	*Homo sapiens*	N-[6,6-dimethyl-5-[(2S)-4-methyl-2-(phenylmethyl)piperazin-1-yl]carbonyl-2,4-dihydropyrrolo[3,4-c]pyrazol-3-yl]-3-phenoxy-benzamide	-7.19382	2.20
4DRI	FKBP51	*Homo sapiens*	RAPAMYCIN IMMUNOSUPPRESSANT DRUG	-8.52288	1.45
4EWH	ACK1	*Homo sapiens*	6-{4-[2-(dimethylamino)ethoxy]phenyl}-N-(1,3-dithiolan-2-ylmethyl)-5-phenyl-7H-pyrrolo[2,3-d]pyrimidin-4-amine	-9.52288	2.50
4FSR	CHK1	*Homo sapiens*	6,7-dimethoxy-3-[4-(1H-tetrazol-5-yl)phenyl]-1,4-dihydroindeno[1,2-c]pyrazole	-7.88606	2.50
4FT3	CHK1	*Homo sapiens*	1-(5-chloro-2,4-dimethoxyphenyl)-3-pyrazin-2-ylurea	-6.19997	2.50
4FT9	CHK1	*Homo sapiens*	1-(5-cyanopyrazin-2-yl)-3-isoquinolin-3-ylurea	-7.6925	2.20
4FTR	CHK1	*Homo sapiens*	2-[3-(3-methoxy-4-nitrophenyl)-11-oxo-10,11-dihydro-5H-dibenzo[b,e][1,4]diazepin-8-yl]-N,N-dimethylacetamide	-8.2186	2.25

4N6Y	PIM-1	*Homo sapiens*	2-(acetylamino)-N-[2-(piperidin-1-yl)phenyl]-1,3-thiazole-4-carboxamide	-7.03937	2.60
4N6Z	PIM-1	*Homo sapiens*	3-amino-N-{4-[(3S)-3-aminopiperidin-1-yl]pyridin-3-yl}pyrazine-2-carboxamide	-7.79588	2.20
4O0Y	P21-activated kinases (PAK)	*Homo sapiens*	4-[1-(4-amino-1,3,5-triazin-2-yl)-2-(ethylamino)-1H-benzimidazol-6-yl]-2-methylbut-3-yn-2-ol	-7.22185	2.20
4OBO	MAP4K4	*Homo sapiens*	6-(3-chlorophenyl)quinazolin-4-amine	-6.96658	2.10
4UAL	MRCK	*Homo sapiens*	4-chloro-1-(piperidin-4-yl)-N-[3-(pyridin-2-yl)-1H-pyrazol-4-yl]-1H-pyrazole-3-carboxamide	-8.39794	1.71
4YVC	ROCK1	*Homo sapiens*	2-fluoro-N-[4-(pyridin-4-yl)-1,3-thiazol-2-yl]benzamide	-5.95861	3.20
5BML	ROCK1	*Homo sapiens*	N-[4-(2-fluoropyridin-4-yl)thiophen-2-yl]-2-{3-[(methylsulfonyl)amino]phenyl}acetamide	-8	2.95

Fonte: A autora (2020).

Das estruturas encontradas, 77 são provenientes de *Homo sapiens*, 19 tiveram origem no organismo *Zea mays* (milho) e 1 estrutura originou-se da levedura *Schizosaccharomyces pombe*. Esta grande disponibilidade de estruturas provenientes de humanos mostra-se compatível com a informação da literatura de que as cinases têm sido amplamente estudadas como opção para novos fármacos.

Observou-se que, aproximadamente 30% do conjunto de dados é composto pela enzima CHK1 (*Checkpoint Kinase 1*), cujo exemplo de estrutura pode ser localizado no PDB com o código 2E9N, e está representado juntamente com seu ligante na Figura 3. A inibição desta enzima, segundo Tong et al.(2007), oferece um mecanismo sensibilizante para várias terapias relacionas a danos de DNA e tem sido estudada para desenvolvimento de tratamentos anticâncer.

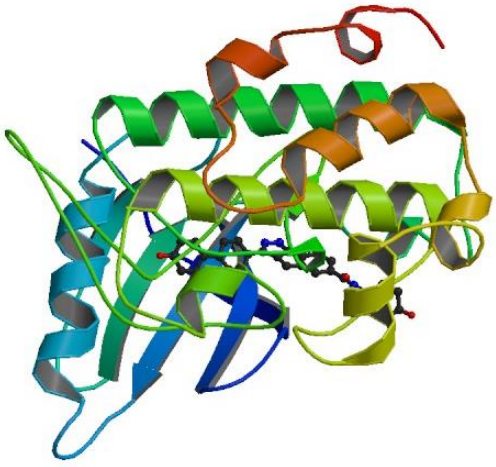


Figura 3. Estrutura cristalográfica da proteína cinase CHK1 complexada com 1,4-di-hidroindeno [1,2-c] pirazóis. Código de acesso no PDB: 2E9N. Fonte: Tong, et al. 2007.

Da mesma forma, pode-se observar que a cinase CK2 (Caseína Cinase II) também possui uma grande participação no conjunto de dados, com quase 29% do total de estruturas. A enzima PIM-1 foi a terceira em número de estruturas, conforme pode ser visto na Tabela 1.

A separação dos subconjuntos por meio de uma semente aleatória e executada com o programa SAnDReS, manteve quase totalmente a proporção dos diferentes tipos de proteína em cada subgrupo, com exceção de algumas, como a B-RAF cinase e a ROCK1 cinase, visto que nenhuma estrutura foi utilizada no conjunto teste para a primeira, e a maioria das enzimas ROCK1 foi utilizada no conjunto teste.

Com relação aos ligantes presentes no *dataset,* pode-se dizer que tratam-se todos de inibidores*,* competitivos para ligação no bolsão de ATP e todos possuem dados da Constante de Inibição (K_i).

Tabela 1. Distribuição de cada tipo de proteína nos conjuntos de treino e teste.

PROTEÍNA	QTDE DATASET	% DATASET	QTDE TREINO	% TREINO	QTDE TESTE	% TESTE
CHK1	30	30,9%	23	34,3%	7	23,3%
CK2	28	28,9%	20	29,9%	8	26,7%
PIM-1	10	10,3%	8	11,9%	2	6,7%
P21-activated kinases (PAK)	5	5,2%	3	4,5%	2	6,7%
PDK1	4	4,1%	2	3,0%	2	6,7%
ROCK1	4	4,1%	1	1,5%	3	10,0%
B-RAF	3	3,1%	3	4,5%	0	0 %
ACK1	2	2,1%	1	1,5%	1	3,3%
AKT2	2	2,1%	2	3,0%	0	0%
Aurora-A	1	1,0%	1	1,5%	0	0%
CDK5	1	1,0%	0	0%	1	3,3%
CK1	1	1,0%	0	0%	1	3,3%
FKBP12	1	1,0%	1	1,5%	0	0%
FKBP51	1	1,0%	0	0%	1	3,3%
IRAK-4	1	1,0%	1	1,5%	0	0%
MAP4K4	1	1,0%	0	0%	1	3,3%
MRCK	1	1,0%	0	0%	1	3,3%
PKBalpha (PKB/Akt)	1	1,0%	1	1,5%	0	0%
TOTAL:	97	100,0%	67	100%	30	100%

Fonte: A autora (2020).

5.2. FUNÇÕES ESCORE CLÁSSICAS

O Quadro 3 apresenta os resultados dos coeficientes de correlação entre as afinidades prevista e experimental log(K_i) para as estruturas do conjunto de dados de teste, após o cálculo dos termos de energia por meio das Funções Escore clássicas. As letras *a* e *b*, que aparecem ao lado de cada função, representam *AutoDock4* e *AutoDock Vina*, respectivamente.

O intervalo entre os resultados da correlação de Spearman variou de -0,668 a 0,422, e a correlação mais significativa (valor absoluto) observada ocorreu para a Função Escore *Gauss2* (ρ = -0,668 e p-value = $5{,}479.10^{-3}$). A análise da correlação quadrática (R_2) gerou uma correlação mais baixa, com $R_2 < 0.254$ para todos os termos.

Quadro 3. Capacidade de predição das Funções Escore clássicas para o conjunto teste.

Função Escore/Termos de Energia	ρ	p-value1	R^2	p-value2	SD	RMSE
Free Energy[a]	0,105	0,5816	0,013	0,5518	69769,604	$5{,}325.10^{+9}$
Final Intermolecular Energy[a]	0,325	0,0797	0,013	0,5518	69769,745	$5{,}325.10^{+9}$
vdW+Hbond+desolv Energy[a]	0,351	0,0570	0,013	0,5517	34892,442	$1{,}332.10^{+9}$
Electrostatic Energy[a]	-0,120	0,5264	0,013	0,5517	34890,364	$1{,}332.10^{+9}$
Final Total Internal Energy[a]	0,060	0,7525	0,006	0,6869	0,924	42,3986
Torsional Free Energy[a]	-0,420	0,0210	0,154	0,0320	0,812	77,7073
Affinity Score[b]	0,422	0,0200	0,112	0,0707	2,470	2,5395
Gauss1 Score[b]	-0,567	$1{,}097{,}10^{-3}$	0,254	$4{,}527.10^{-3}$	29,782	98,0079
Gauss2 Score[b]	-0,668	$5{,}479.10^{-3}$	0,182	0,0186	468,059	1458,13
Repulsion Score[b]	-0,213	0,2587	0,051	0,2316	1,104	10,2097
Hydrophobic Score[b]	-0,020	0,9154	0,013	0,5458	25,263	52,8671
Hydrogen Score[b]	-0,205	0,2772	0,046	0,2547	1,359	9,50158
Torsions[b]	-0,408	0,0210	0,154	0,0319	2,719	12,4399

Calculado utilizando [a] AD4. [b] Vina. SD: *standard deviation* (desvio padrão). RMSE: *root mean square error* (erro quadrático médio da raiz)

A Figura 4 apresenta o gráfico de dispersão para a Função Escore *Gauss2* em relação ao log(K_i), que demonstra uma correlação negativa, indicada pela direção da linha em torno da qual estão dispersos os valores experimentais e calculados. Embora sua correlação seja alta, a Função Escore *Gauss2* mostrou uma performance ruim de forma geral, considerando os altos valores observados para o RMSE e o SD.

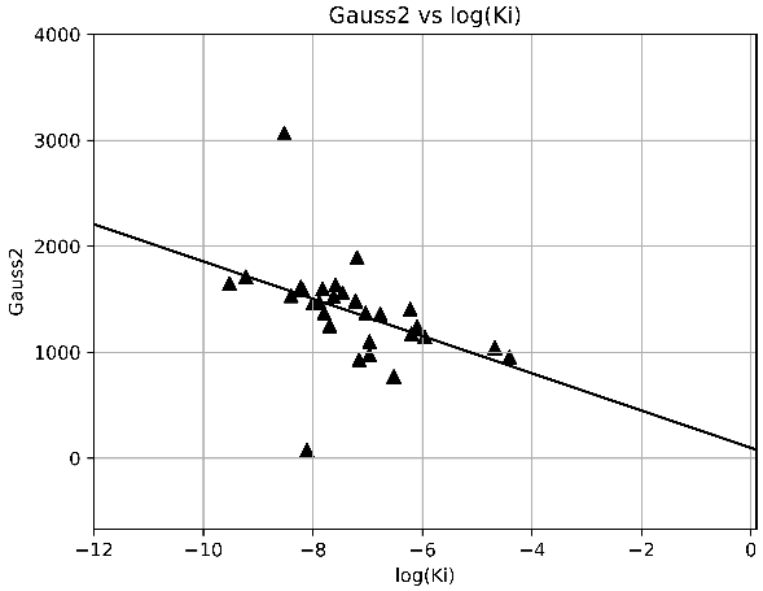


Figura 4. Gráfico de dispersão para a informação experimental log(K_i) e a afinidade prevista com a Função Escore *Gauss2* para um subconjunto de teste obtido a partir do conjunto de dados. A função *Gauss2 Score* está em unidades arbitrárias.

5.3. MODELOS DE APRENDIZADO DE MÁQUINA

Na aplicação da “Abordagem A”, foram gerados dois modelos de Funções Escore otimizadas. Algumas combinações foram testadas entre os termos calculados e o número de neurônios utilizados nas *hidden layers* do algoritmo NN. Dentre todos os gerados, o modelo que apresentou a melhor correlação para os termos do *AutoDock4* foi chamado de “Modelo 1”, e para os termos do *AutoDock Vina*, chamamos o melhor modelo de “Modelo 2”.

No “Modelo 1”, utilizaram-se os seguintes termos de energia calculados com o programa *AutoDock4*:

Final Intermolecular Energy (x_1), vdW+Hbond+desolv Energy (x_2), Electrostatic Energy (x_3), Final Total Internal Energy (x_4) e *Torsional Free Energy (x_5)*.

A Função Escore com a maior correlação (ρ) utilizando estes termos de energia resultou na seguinte equação, que foi gerada pelo algoritmo de Regressão PLS:

$$log(K_i) = -0.000191721 * FinalIntermolecularEnergy + 0.055672 * (vdW+Hbond+desolvEnergy) - 0.05529 * ElectrostaticEnergy + 0.362288 * FinalTotalInternalEnergy - 0.42399 * TorsionalFreeEnergy - 6.16521$$

Ainda dentro da "Abordagem A", o "Modelo 2" foi gerado com uso dos seguintes termos energéticos calculados com o programa *AutoDock Vina:*

Gauss1 Score (x_1), Gauss2 Score (x_2), Repulsion Score (x_3), Hydrogen Score (x4) e Torsions (x5).

O "Modelo 2" também foi gerado com uso do algoritmo de regressão PLS. A Figura 5 apresenta o gráfico de dispersão para este modelo, que demonstra uma boa correlação, levando-se em consideração o coeficiente de Spearman (ρ) em conjunto com o RMSE e o SD. A Função Escore com a mais alta correlação para este modelo resultou na seguinte expressão:

$$log(K_i) = 0.000512639 * Gauss1 - 0.00207336 * Gauss2 - 0.2778 * Repulsion - 0.0445979 * Hydrogen + 0.0725063 * Torsions - 4.05264$$

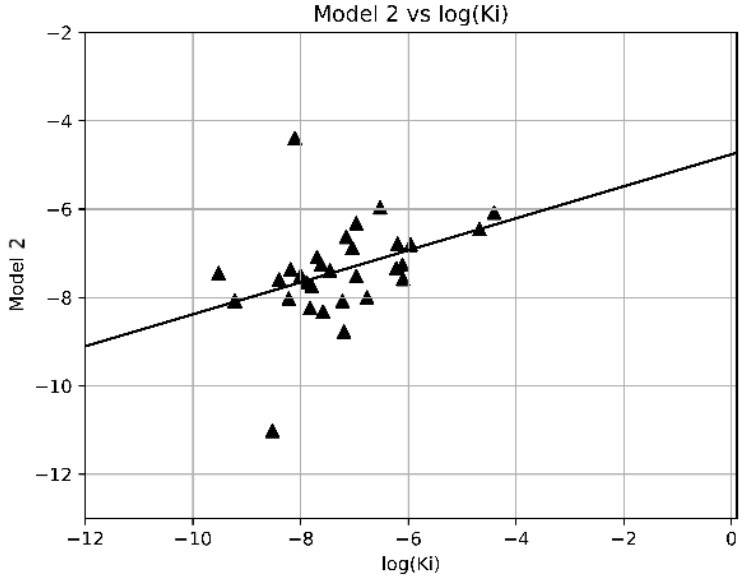


Figura 5. Gráfico de dispersão para o dado experimental log(K_i) e a afinidade prevista para "Modelo 2", para o subconjunto de teste obtido a partir do conjunto de dados.

Dentre todos os modelos gerados com o programa Taba na "Abordagem B", verificou-se que o melhor modelo de forma geral, foi o que obteve a maior correlação (ρ), e mostrou uma Função Escore com três variáveis explanatórias e uma distância de corte de 6.0 Å. Ao modelo gerado pelo Taba, chamamos de "Modelo 3".

Para o "Modelo 3", as constantes de regressão foram as seguintes: $\alpha_0 = -7{,}465237$; $\alpha_{C,N} = -20{,}732288$; $\alpha_{C,F} = 0{,}519849$; $\alpha_{N,N} = 20{,}823863$ e as distâncias de equilíbrio foram as seguintes: $d_{0,C,N} = 4{,}99029$; $d_{0,C,F} = 5{,}05205$; $d_{0,N,N} = 4{,}98988$ Å. Estes resultados foram obtidos utilizando-se o método de Regressão Multilinear padrão.

Após a análise estatística de todos os modelos gerados pela "Abordagem A" e pela "Abordagem B", chegou-se, por fim, a três Funções Escore que apresentaram as melhores correlações para o conjunto de teste. Elas são apresentadas no Quadro 4.

Analisando-se conjuntamente o RMSE e a correlação de Spearman dos 3 modelos finais, pode-se verificar que a melhor capacidade de predição foi obtida com o "Modelo 2", usando os termos de energia calculados pelo *software AutoDock Vina*.

Quadro 4. Capacidade de predição dos modelos de Aprendizado de Máquina (conjunto teste).

Funções Escore	ρ	p-value1	R^2	p-value2	SD	RMSE
Modelo 1[a]	0,302	0,1044	0,101	0,0878	0,595	1,126
Modelo 2[b]	0,484	$6{,}69.10^{-3}$	0,145	0,0382	1,082	1,2426
Modelo 3[c]	0,401	0,02822	0,070	0,15754	4,87795	13,85343
Calculado utilizando [a] AD4. [b] Vina [c] Taba. SD: *standard deviation* (desvio padrão). RMSE: *root mean square error* (erro quadrático médio da raiz)						

No entanto, apesar de não ter sido o modelo com melhor performance, o "Modelo 3" mostrou também uma correlação muito próxima, se verificado o coeficiente de Spearman (ρ). Contudo, os resultados de SD e RMSE mostraram-se insatisfatórios.

Ainda assim, mesmo com resultados piores para SD e RMSE, foi possível capturar com este modelo características interessantes da interação proteína-ligante para a classe enzimática específica sendo estudada (EC 2.7.11.1), uma vez que ele

possui três variáveis explanatórias envolvendo interações Carbono-Nitrogênio, Carbono-Flúor e Nitrogênio-Nitrogênio.

Estudos publicados anteriormente (De Ávila, et al., 2017) indicam a importância dos átomos halogênios na inibição das cinases e este modelo de AM capturou esta interação essencial na equação de regressão, com a participação das interações interatômicas envolvendo o Flúor.

Uma destas interações pode ser vista entre a PIM-1 cinase e o inibidor 5-(4-cianobenzil)-N-(4-fluorofenil)-7-hidroxipirazolo[1,5-a]pirimidina-3-carboxamida, conforme ilustrado na figura abaixo:

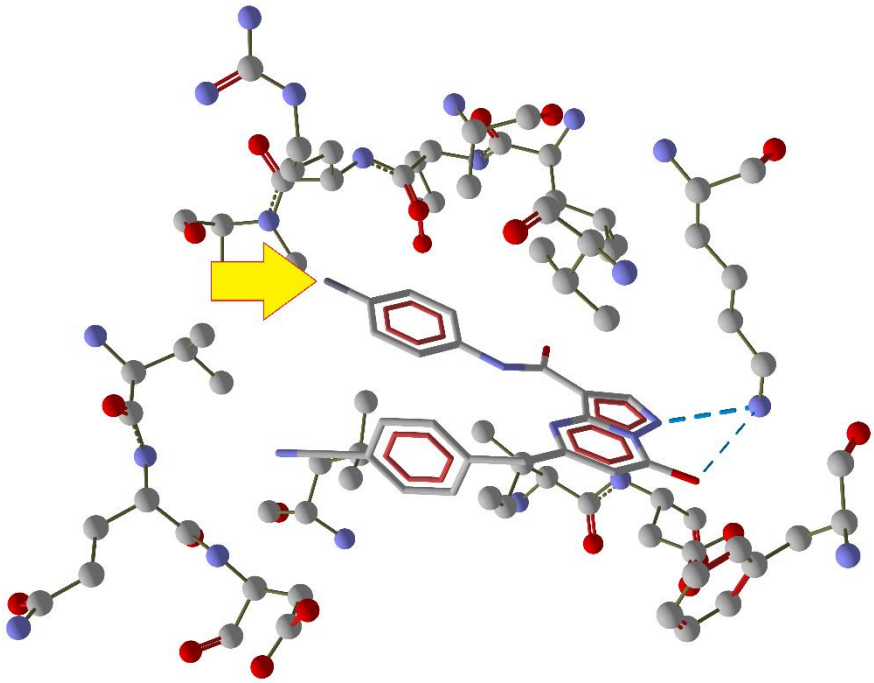


Figura 6. Estrutura da PIM-1 cinase complexada com 5-(4-cianobenzil)-N-(4-fluorofenil)-7-hidroxipirazolo[1,5-a]pirimidina-3-carboxamida. O átomo de flúor pode ser visto em cinza na extremidade do ligante. Os tracejados representam as ligações de hidrogênio. A figura foi gerada com uso do programa MVD (*Molegro Virtual Docker*), a partir da estrutura encontrada no PDB com o código 4K18.

O “Modelo 2”, além de ter se mostrado superior na correlação com os dados experimentais, também elucidou alguns pontos interessantes relacionados ao conhecimento estrutural das serino/treonino cinases não específicas. O termo de energia de repulsão possui uma participação significativa na equação do “Modelo 2”,

o que é consistente com as interações hidrofóbicas observadas nos inibidores deste tipo de cinase (YH et al., 2016).

Os demais resultados de correlação dos termos de energia e dos modelos gerados para ambas as abordagens, podem ser verificados com maiores detalhes no artigo publicado e nos materiais suplementares disponíveis nos Apêndices A e B.

6. CONSIDERAÇÕES FINAIS

O desenvolvimento de modelos de Aprendizado de Máquina utilizando os programas MDM e Taba mostrou capacidade de predição de afinidade superior em comparação com as Funções Escore clássicas. As Funções Escore otimizadas desenvolvidas com uso dos termos de energia do programa *AutoDock Vina* superaram em performance as Funções Escore clássicas e os demais modelos de Aprendizado de Máquina testados neste trabalho. Além disso, os modelos de Aprendizado de Máquina puderam identificar características estruturais responsáveis pela inibição das serino/treonino cinases não específicas, tais como a importância das interações hidrofóbicas e a participação dos átomos halogênicos na inibição.

Analisando conjuntamente, pode-se concluir que o uso da abstração do “Espaço de Funções Escore” e os algoritmos de regressão de Aprendizado de Máquina possuem um potencial de desenvolvimento de modelos computacionais com capacidade preditiva superior, assim como já demonstrado em estudos prévios para outras famílias de proteínas.

6.1. TRABALHOS FUTUROS

Como perspectivas de estudo para trabalhos futuros, podemos apontar o uso de uma abordagem alternativa para geração dos modelos de Aprendizado de Máquina utilizando diferentes termos energéticos, como os disponíveis em outros programas de docagem, a exemplo do *Molegro Virtual Docker* (MVD).

Pode-se ainda alternar o número de variáveis explanatórias nas abordagens utilizadas, bem como os termos utilizados nas equações geradas, a fim de avaliar se os resultados obtidos se mantêm iguais ao deste estudo ou se esta variação impacta significativamente na qualidade dos modelos gerados.

Além disso, os modelos gerados que obtiveram capacidade preditiva superior podem ser utilizados para avaliação de resultados de docagem, ao serem aplicados às proteínas específicas da mesma família daquelas contidas no *dataset* utilizado, mas que não possuam ainda ligantes complexados a elas e não tenham informações de afinidade.

6.2. FINANCIAMENTO DA PESQUISA

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

REFERÊNCIAS

ALBERTS B., BRAY D., HOPKIN K., JOHNSON A., LEWIS J., RAFF M., ROBERTS K., WALTER P. Fundamentos da Biologia Celular. 4ª ed. Porto Alegre: Artmed, 2017.

BARREIRO EJ., FRAGA CAM. Química Medicinal – As Bases Moleculares da Ação dos Fármacos. 3ª ed. Porto Alegre: Artmed, 2015.

BATTISTUTTA R., MAZZORANA M., SARNO S., KAZIMIERCZUK Z., ZANOTTI G. PINNA L.A. Inspecting the Structure-Activity Relationship of Protein Kinase CK2 Inhibitors Derived from Tetrabromo-Benzimidazole. Chemistry & Biology. 2005; Vol. 12, 1211–1219.

BELL J. Machine Learning – Hands-on For Developers And Technical Professionals. Indianapolis: John Wiley & Sons Inc., 2015.

BERMAN HM, et al. The Protein Data Bank. Nucleic Acids Research. 2000; 28:235-242.

BINDING MOAD. Disponível em: <http://bindingmoad.org/>. Acesso em: 17 dez. 2019.

BITENCOURT-FERREIRA G., DE AZEVEDO JR. WF. Development of a machine-learning model to predict Gibbs free energy of binding for protein-ligand complexes. Biophys Chem. 2018; 240: 63–69.

BITENCOURT-FERREIRA G., DA SILVA AD., DE AZEVEDO JR. WF. Application of Machine Learning Techniques to Predict Binding Affinity for Drug Targets. A Study of Cyclin-Dependent Kinase 2. Current Medical Chem. 2020; 27: 1-11.

BOHACEK R.S., MCMARTIN C., GUIDA WC. The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. Med. Res. Rev. 1996, 16(1), 3-50.

BÖHM HJ. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. J Comput Aided Mol Des. 1994; 8(3): 243-56.

BÖHM HJ. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. J Comput Aided Mol Des. 1998; 12(4): 309-23.
CLC BIO. Molegro Data Modeller User Manual – MDM 2013.3.0 for Windows, Linux, and Mac OS X. 2013; 8:84-90.

DA SILVA AD., BITENCOURT-FERREIRA G., DE AZEVEDO JR. WF. Taba: A Tool to Analyze the Binding Affinity. J. Comput. Chem. 2020.

DE ÁVILA MB., XAVIER MM., PINTRO VO., AZEVEDO WF. Supervised machine learning techiniques to predict binding affinity. A study fo cyclin-dependent kinase 2. Biochemical and Biophysical Research Communications. 2017, Epub. DOI:10.1016/j.bbrc.2017.10.035.

DE AZEVEDO WF JR., DIAS R. Evaluation of ligand-binding affinity using polynomial empirical scoring functions. Bioorganic & Medicinal Chemistry. 2008, 16(20): 9378-9382.

DOBSON CM. Chemical space and biology. Nature. 2004, 432(7019), 824–828.

FACELI K et al. Inteligência artificial: Uma Abordagem de Aprendizagem de Máquina. 1. ed. Rio de Janeiro: LTC, 2015

HEBERLÉ G., DE AZEVEDO WF. Bio–inspired algorithms applied to molecular docking simulations. Current Medicinal Chemistry, 2011, 18, 1339–1352.

HECK GS et al. Supervised machine learning methods applied to predict ligand binding affinity. Curr Med Chem. 2017; 24(23): 2459-2470.

HUANG SY., GRINTER SZ., ZOU X. Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. Phys Chem Chem Phys. 2010; 12: 899-908.

HUEY R., MORRIS GM., OLSON AJ., GOODSELL DS. A Semiempirical Free Energy Force Field with Charge-Based Desolvation. J. Comput. Chem., 2007, 28, 1145-1652.

JAIN A.N. Scoring functions for protein–ligand docking. Curr Protein Pept Sci. 2006 Oct;7(5):407–20.

LABUTE P. Methods of Exploring Protein–Ligand Interactions to Guide Medicinal Chemistry Efforts. Methods Mol. Biol. 2018, 1705, 159–177.

LEGENDRE AM. Nouvelle méthodes pour la déterminiation des orbites des comètes, Courcier, Paris. 1805.

LITCHFIELD DW. Protein kinase CK2: structure, regulation and role in cellular decisions of life and death. Biochem J. 2003 Jan 1; 369(Pt 1):1-15.

MIELECKI M., LESYNG B. Cinnamic Acid Derivatives as Inhibitors of Oncogenic Protein Kinases – Structure, Mechanisms and Biomedical Effects *Curr. Med. Chem*. 2016, *23*(10), 954–982.

MORRIS GM., GOODSELL DS., HALLIDAY RS., HUEY R., HART WE., BELEW RK., OLSON AJ. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. Journal of Computational Method. 1998; 19: 1639-1662.

MORRIS GM., HUEY R., LINDSTROM W., SANNER MF., BELEW RK., GOODSELL DS., OLSON AJ. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J. Comput. Chem. 2009, 30, 2785-2791.

NIEFIELD K., GUERRA B., ERMAKOWA I., ISSINGER OG. Crystal structure of human protein kinase CK2: Insights into basic properties of theCK2 holoenzyme. The EMBO Journal. 2001, Vol. 20, No.19, pp. 5320-5331.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M., DUCHESNAY E. Scikit–learn: machine learning in python. J. Mach. Learn. Res., 2011, 12, 2825–2830.

RCSB. Disponível em: <http://www.rcsb.org>. Acesso em: 17 dez. 2019.

SILVA BV., HORTA BAC., ALENCASTRO RBD., PINTO AC. Proteínas quinases: características estruturais e inibidores químicos. Química Nova. 2009; *32*(2), 453-462.

SIPPL MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol. 1990; 213(4): 859-83.

SLIWOSKI G., KOTHIWALE S., MEILER J, JR EWL. Computational Methods in Drug Discovery. Pharmacol Rev. 2014; 66:334–395.

THE BINDING DATABASE. Disponível em: <https://www.bindingdb.org/bind/index.jsp>. Acesso em: 17 dez. 2019.

THOMSEM R., CHRISTENSEN MH. MolDock: a new technique for high-accuracy molecular docking. J Med Chem. 2006; 49:3315-3321.

TIBSHIRANI R., Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Series B Stat. Methodol. 58 (1996) 267–288.

TIKHONOV AN., On the regularization of ill-posed problems, Dokl. Akad. Nauk SSSR 153 (1963) 49–52.

TONG Y., CLAIBORNE A., STEWART K.D., PARK C., KOVAR P., CHEN Z., CREDO R.B., GU WZ., GWALTNEY II S. L., JUDGE R.A., ZHANG H., ROSENBERG S.H., SHAM H.L., SOWIN T.J., LIN NH. Discovery of 1,4-dihydroindeno[1,2-c]pyrazoles as a novel class of potent and selective checkpoint kinase 1 inhibitors. Bioorganic & Medicinal Chemistry. 2007; Vol. 15: 2759–2767.

TROTT O., OLSON AJ., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, J. Comput. Chem. 31 (2010) 455–461

TUTONE M., ALMERICO, AM. Recent advances on CDK inhibitors: An insight by means of in silico methods Eur. J. Med. Chem. 2017, 142, 300–315.

WANG R., FANG X., LU Y., WANG S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. J Med Chem. 2004; 47(12): 2977-80.

XAVIER MM., HECL GS., DE AVILA MB., LEVIN NM., PINTRO VO., CARVALHO NL., DE AZEVEDO JR, WF. Sandres a computational tool for statistical analysis of docking results and development of scoring functions. Comb Chem High Throughput Screen. 2016; 19(10): 801–812.

ZAR JH. Significance Testing of the Spearman Rank Correlation Coefficient. J. Am. Statist. Assoc., 1972; 67(339): 578-580.

ZOU H., HASTIE T. Regularization and variable selection via the elastic net, J. R. Stat. Soc. Series B Stat. Methodol. 67 (2005) 301–320.

APÊNDICE A – ARTIGO CIENTÍFICO SUBMETIDO

Artigo com os resultados da pesquisa submetido ao periódico "*Journal of Computational Chemistry*" de responsabilidade da Wiley Periodicals, Inc.

Título: **"*Application of Machine Learning Methods to Predict Binding Affinity. A Study of Non-specific Serine/Threonine Protein Kinases*".**

Fator de Impacto: 3.224

Site do periódico: https://onlinelibrary.wiley.com/journal/1096987x

17-Dec-2019

Manuscript number: JCC-19-0686

Dear Prof. de Azevedo Jr.:

We are pleased to receive your manuscript entitled Application of Machine Learning Methods to Predict Binding Affinity. A Study of Non-specific Serine/Threonine Protein Kinases by Waszak, Rosana; Bitencourt-Ferreira, Gabriela; de Azevedo Jr., Walter. It will be assigned to an appropriate Associate Editor who manages the review process.

If your manuscript is a REVISION, it is automatically assigned to the Associate Editor who handled it previously. Depending on the amount of revision required, this editor will either have your revision re-reviewed or will make a decision based on your revisions in response to the previous reviews.

Please remember in any future correspondence regarding this article to always include its manuscript ID number JCC-19-0686.

If you experience problems associated with the submission web site, please contact the Wiley support staff directly at: jccadmins@umich.edu

Many thanks for submitting your manuscript,

Journal of Computational Chemistry Editorial Office

Application of Machine Learning Methods to Predict Binding Affinity. A Study of Non-specific Serine/Threonine Protein Kinases

Rosana da Silva Waszak,[1,2] Gabriela Bitencourt-Ferreira,[2] and Walter Filgueira de Azevedo, Jr.[1,2]

Correspondence to: Walter Filgueira de Azevedo (E-mail: *walter@azevedolab.net*)

[1] *Laboratory of Computational Systems Biology, School of Sciences, Pontifical Catholic University of Rio Grande do Sul (PUCRS), Ipiranga Avenue, 6681 Partenon - Porto Alegre/RS. Brazil. Zip Code: 90619-900*
[2] *Specialization Program in Bioinformatics, School of Sciences, Pontifical Catholic University of Rio Grande do Sul (PUCRS), Ipiranga Avenue, 6681 Partenon - Porto Alegre/RS. Brazil. Zip Code: 90619-900*

ABSTRACT

Kinases are the most intensively studied protein in drug design and development. Among kinases, non-specific serine/threonine protein kinase represents an interesting protein system for modeling purposes due to the availability of structural and functional experimental data. Non-specific serine/threonine protein kinase comprises an important class of protein targets used to develop drugs to treat cancer. Here, we describe the creation of machine learning models to predict protein-ligand binding affinity for this enzymatic class. We make use of energy terms available in classical scoring functions such as Autodock4 and AutoDock Vina. We use these terms to build a novel scoring function targeted to a dataset comprised of nearly 100 protein-ligand complexes for which experimental crystallographic structure and inhibition constant data are available. We also applies a hybrid mass-spring method to determine binding affinity using the program Taba. We carry out predictive performance analysis of all scoring functions. Our study clearly shows that machine learning models present superior predictive performance when compared with classical scoring functions. Also, our machine learning models can capture structural features responsible for the binding affinity against non-specific serine/threonine protein kinases.

Introduction

Protein kinases have been extensively studied since many members of this protein class are targets for drug development.[1,2] Considering protein kinases relevant for drug design, we may highlight the following enzymes of this group: non-specific serine/threonine protein kinases, cyclin-dependent kinases, casein kinase 1, and calmodulin/calcium-regulated kinase. Focusing specifically on non-specific serine/threonine protein kinase, we find inhibitors of this enzymatic class in clinical trials, such as Tozasertib and Alisertib.[3]

Due to the importance of this enzymatic class for drug discovery, the development of computational methods to assess protein-ligand interactions has a significant impact in the early stages of drug discovery. Assessment of intermolecular interactions using the atomic coordinates of protein-ligand complexes is of pivotal importance to understand the structural basis for the specificity of inhibitors for a target.[4] Also, the explosion in the number of three-dimensional structures makes it possible

to determine the binding affinity computationally.[5-7] We may estimate the binding energy using quantum physics.[8] Another approach to calculating binding energy is the molecular dynamics simulation.[9,10]

Molecular dynamics simulations and quantum physics can build computational models to calculate binding energy. But they have high computational cost when compared with classical scoring functions, being the quantum physics method the most demanding from the computational point of view.[8]

Furthermore, experimental information about inhibition constant and the crystallographic data about the protein-ligand complexes of members of this enzymatic class and small molecules for which binding affinity data is available paves the way to develop machine learning models to predict binding affinity. These computational models rely on robust experimental information derived from X-ray diffraction crystallography. We also use data obtained from functional techniques to determine binding affinity and thermodynamic parameters. Taking together this experimental information makes the raw material used to train machine learning models to predict binding affinity for enzymes.

The constant growth of the number of structures available in the protein data bank (PDB)[11], and the increasing experimental data available for binding affinity data and thermodynamic parameters stored in databases such as BindingDB,[12] MOAD,[13] and PDBbind[14] is the solid base to create targeted scoring functions, specific for protein systems of interest. Also, the development of machine learning libraries such as TensorFlow, scikit-learn[15], and Keras made available open-source computational tools to build programs to predict protein-ligand binding affinity.

This synergism involving increasing experimental data and the development of machine learning tools to treat this experimental data contributed to establishing the mathematical basis to construct a robust theoretical framework to interpret the relationships involving the protein[16] and chemical[17,18] spaces. Let`s consider a member of the protein space, for instance, a non-specific serine/threonine protein kinase (EC 2.7.11.1). Looking at the chemical space, we may consider a subspace comprised of inhibitors of this enzyme. Now we add a third space composed of mathematical equations that can calculate binding affinity based on the atomic coordinates of protein-ligand complexes. These expressions form the scoring function space.[19] Once we defined one element of the protein space and the subset of small molecules that bind to this element of the protein space, we may scan the scoring space function. Figure 1 illustrates the exploration of the scoring function space to find an adequate equation to predict binding affinity based on the atomic coordinates. The search of the scoring function space may find a computational model able to predict binding affinity based solemnly on the atomic coordinates of a protein-ligand complex.

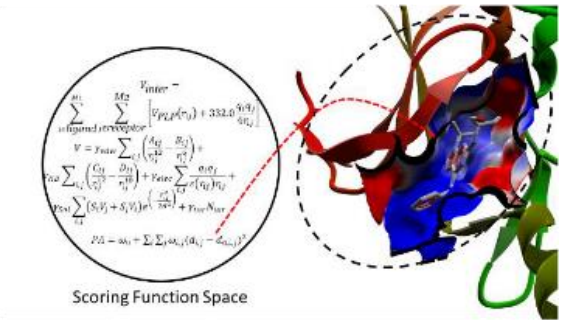


Figure 1. Relationship between scoring function space and one element of the protein space. We show this relationship with an arrow in the above figure. We propose here that our methodology can find an element of the scoring function space, which can assess the binding affinity of the chosen protein.

Machine learning techniques have great potential to explore the scoring function space[20,21], finding an adequate model to predict binding affinity. Our goal here is to develop targeted-scoring functions to predict the inhibition of non-specific serine/threonine protein kinases. We used machine learning techniques available in the programs Molegro Data Modeller (MDM)[22,23] and Taba.[24,25] We envisage this approach as a way to explore scoring function space, to find a computational model adequate to predict inhibition of non-specific serine/threonine protein kinases.

Methods

To generate machine learning models to predict the inhibition constant of non-specific serine/threonine protein kinases, we used two main approaches. The first method involves a combination of MDM[22,23] with SAnDReS[26], AutoDock4 (AD4)[27] and AudoDock Vina (Vina).[28] The second method uses only the program Taba. Figure 2 shows the flowchart used to generate machine learning models with MDM (first method). In the following, we describe each step of this flowchart.

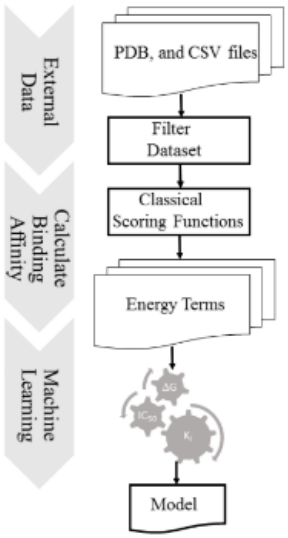


Figure 2. Schematic flowchart for the generation of machine learning models using MDM, SAnDReS, and docking programs (AD4 and Vina).

Dataset

We searched the PDB for crystallographic structures of non-specific serine/threonine protein kinases for which inhibition constant data were available. This dataset was filtered to eliminated repeated ligands to have diversity in the subset of the chemical space used to train our machine learning models. We divided this dataset into two subgroups, the training and test sets. We used the program SAnDReS for downloading and filtering of the dataset. SAnDReS integrates different techniques to carry out docking and to calculate binding affinity. SAnDReS analyzes data from any protein-ligand docking program. It is necessary to have structures in PDB format, ligands in Structure Data Format (SDF), docking and scoring function data in comma-separated values (CSV) format. PDB gathers experimental binding affinity data from three other databases: BindingDB, MOAD, and PDBbind.

Classical Scoring Functions

We calculated binding affinity using classical scoring functions available in the programs AD4 and Vina. We used the crystallographic positions of the ligands in the calculations of the energy terms and scoring functions. We did not use docked structures to evaluate binding affinity for the structures of our dataset. We prepared all ligands using default charge values for the programs AD4 and Vina. We defined protein atomic charges according to default parameters of the same programs. We used the docking programs to calculate also energy terms involving intermolecular and intramolecular interactions. We employ these energy terms as explanatory variables to train our machine learning models against the experimental data.

Machine Learning

The program SAnDReS merged the energy terms of each scoring function available in the previously mentioned docking programs to feed

the MDM. In the development of machine learning models, we take as explanatory variables the energy terms and scoring functions calculated by the programs AD4 and Vina. MDM has four techniques for regression analysis, namely multiple linear regression, partial least squares, support vector machines, neural networks. To generate the regression models with neural networks, we used two hidden layers, and the number of neurons in each hidden layer varies from 1 to 10. The algorithm used in MDM for training neural network models is called back-propagation. We used the model with the highest correlation with the experimental data.

We applied the energy terms to build a polynomial equation[29,30] to evaluate the log(K_i) for protein-ligand complexes in the dataset. We consider that we estimated the energy of the protein-ligand interactions through five terms named here x_1, x_2, x_3, x_4, and x_5. These terms are the explanatory variables of the equation below,

$$y = \lambda_0 + \lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 + \lambda_4 X_4 + \lambda_5 X_5 \quad (1)$$

where λ_0 in the regression constant, and λ's are the relative weights of each explanatory variable. The response variable is *y*, is the logarithm of the binding affinity ($log(K_i)$).

Taba

In addition to the machine learning workflow we described above, we also generated machine learning models with the program Taba. This program considers protein-ligand interactions as a mass-spring system, where we have equilibrium distances for springs taken from interatomic distances of pair of atoms, one from the ligand and the second from the protein. Taba uses the classical energy term for the mass-spring energy taking the equilibrium distances obtained from the training set. We used up to five energy terms of mass-spring energy as explanatory variables. These explanatory variables will have their relative weights determined by supervised machine learning techniques such as linear regression, least absolute shrinkage, and selection operator (Lasso)[31], ridge[32], and elastic net.[33] These last three techniques allow the application of cross-validation methods.

For a given protein-ligand structure, the predicted binding affinity (log(K_i)) is the machine-learning model expressed by the following equation:

$$log(K_i) = \alpha_0 + \sum_i \sum_j \alpha_{i,j} (d_{i,j} - d_{0,i,j})^2 \quad (2)$$

In the equation above, α_0 is the regression constant, $\alpha_{i,j}$ is the relative weight of each variable. The double summation is taken over all protein (i) and ligand atoms (j). The term $d_{0,i,j}$ is the average distance for a given pair of atoms i and j, which is calculated for all structures in the dataset. The terms (α_0, $\alpha_{i,j}$ and $d_{0,i,j}$) are calculated considering all structures in the training set. The term $d_{i,j}$ is the distance for a given pair of atoms for one specific structure (not averaged for all structures).

We have previously used Taba to generate machine learning models to predict binding affinity for CDK. The machine learning step of Taba makes use of seven regression classes implemented in Python and accessible in the scikit-learn library. We have one class for each method as follows: Ordinary Linear Regression (*sklearn.linear_model.LinearRegression*), Lasso (*sklearn.linear_model.Lasso*), Lasso with cross-validation (*sklearn.linear_model.LassoCV*), Ridge (*sklearn.linear_model.Ridge*), Ridge with cross-validation (*sklearn.linear_model.RidgeCV*), Elastic Net (*sklearn.linear_model.ElasticNet*), and Elastic Net with cross-validation (*sklearn.linear_model.ElasticNetCV*).

Statistical Analysis

We evaluate the predictive performance of the machine learning models, and classical scoring function through the calculation of root mean square error (RMSE), standard deviation (SD), Spearman's rank (ρ), and Pearson correlation coefficients and related p-values[34].

Results and Discussion

Dataset

We have selected three-dimensional structures of the enzymatic class EC 2.7.11.1 solved by X-ray diffraction crystallography for which inhibition constant (K_i) data was available (search carried out on the PDB on December 16, 2019). This search returned 156 complex structures. We further filtered structural information to eliminate entries without crystallographic water molecules or repeated ligands. After data filtering, we ended up with a dataset comprising of 97 unique (no repeated ligands) complex structures. Table 1 shows the PDB access codes for the training and test sets.

Table 1. PDB access codes for training and test sets.

Training set	1M2R,1NVQ,1W1D,1W1G,1ZOE,1ZOH,2BR1,2BRB,2BRM,2C3J,2C3L,2E9N,2E9P,2E9U,2E9V,2ESM,2FAP,2GHG,2NRU,2OXD,2OXX,2PVH,2PVJ,2PVM,2PVN,2UVM,2WTV,2ZJW,3BE9,3BGP,3BGQ,3BGZ,3BQC,3C4C,3D0E,3E88,3EQR,3FL5,3IDP,3JVR,3JVS,3OWJ,3PE1,3PE2,3PWD,3U9C,4FBX,4FSN,4FST,4FSW,4FSY,4FT0,4FT5,4FT7,4FTA,4FTT,4FTU,4K0Y,4K18,4K1B,4KWP,4MNF,4N70,4O0T,4O0X,4ZY4,5DIA
Test set	1M2P,1M2Q,1NVR,1NVS,1OM1,1ZOG,2C3K,2CSN,2ETR,2OXY,2PVK,2PVL,3H30,3O0G,3RWP,3RWQ,4APP,4DRI,4EWH,4FSR,4FT3,4FT9,4FTR,4N6Y,4N6Z,4O0Y,4OBO,4UAL,4YVC,5BML

Classical Scoring Functions

We used the crystallographic position of ligands for the structures in the dataset. We applied the scoring functions implemented in the programs AD4 and Vina to predict binding affinity (supplementary materials 1 and 2). Table 2 shows the results for correlation coefficients between predicted affinities and experimental log(K_i) for the structures in our dataset (test set). The Spearman rank correlation ranges from -0.668 to 0.422. We observed the most significant correlation (absolute value) for the Gauss2 Score (ρ= -0.668 and p-value = $5.479.10^{-3}$). Squared correlation (R2) analysis generated a lower correlation, with R2 < 0.254 for all terms. Nevertheless, the Gauss2 Score shows poor overall performance, considering the high values observed for RMSE and SD. Figure 3 shows the scatter plot for the Gauss2 Score in the ligand against the log(K_i).

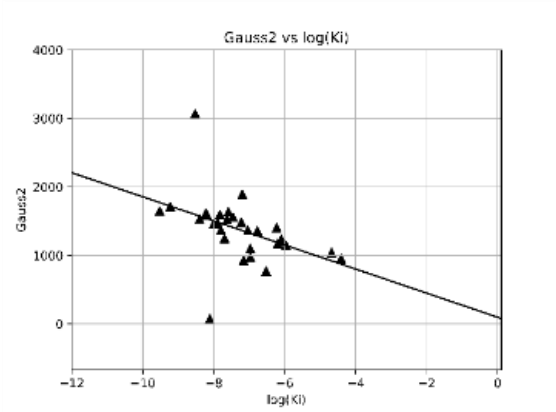


Figure 3. Scatter plot for experimental log(K_i) and predicted affinity (Gauss2 Score) for a test set taken from the dataset.

Machine Learning Models

We generated three machine learning models. In the first model, we used the following the energy terms calculated with AD4: Final Intermolecular Energy (x_1), vdW+Hbond+desolv Energy (x_2), Electrostatic Energy (x_3), Final Total Internal Energy (x_4), and Torsional Free Energy (x_5). The highest correlation (ρ) model using this energy terms has the following equation:

$$log(K_i) = -0.000191721 * FinalIntermolecularEnergy + 0.055672 * (vdW+Hbond+desolvEnergy) - 0.05529 * ElectrostaticEnergy + 0.362288 * FinalTotalInternalEnergy - 0.42399 * TorsionalFreeEnergy - 6.16521 \quad (3)$$

We obtained the above model using PLS as regression algorithm.

Using energy terms calculated with Vina (Gauss1 Score, Gauss2 Score, Repulsion Score, Hydrogen

Score, and Torsions), the highest correlation model has the following expression:

$$log(K_i) = 0.000512639 * Gauss1 - 0.00207336 * Gauss2 - 0.2778 * Repulsion - 0.0445979 * Hydrogen + 0.0725063 * Torsions - 4.05264 \quad (4)$$

We generated this model, also using PLS as a regression algorithm. Figure 4 shows the scatter plot for this model.

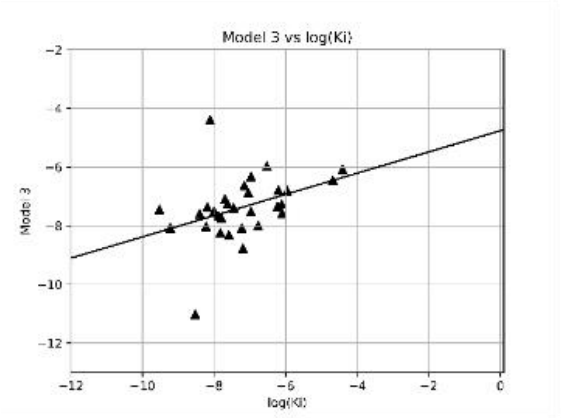


Figure 4. Scatter plot for experimental log(K_i) and predicted affinity (Model 3) for a test set taken from the dataset.

We also created machine learning models with Taba varying the number of explanatory variables from 2 to 5 and using the following distance cutoff: 3.5, 4.5, 6.0, 9.0, and 12.0 Å. Taking the test set results (Table 3), we considered the best overall model, the one with the highest correlation (ρ), which shows a scoring function with three explanatory variables and a cutoff distance of 6.0 Å. For this model, the regression constants are the following: α_0 = -7.465237; $\alpha_{C,N}$ = -20.732288; $\alpha_{C,F}$ = 0.519849; $\alpha_{N,N}$ = 20.823863. The equilibrium distances are the following: $d_{0,C,N}$ = 4.99029; $d_{0,C,F}$ = 5.05205; $d_{0,N,N}$ = 4.98988 Å. Taba obtained these results using the ordinary multilinear regression method.

Table 2. Predictive performance of classical scoring functions (test set).

Scoring Functions	ρ	p-value1	R^2	p-value2	SD	RMSE
Free Energy[a]	0.105	0.5816	0.013	0.5518	69769.604	$5.325.10^{+9}$
Final Intermolecular Energy	0.325	0.0797	0.013	0.5518	69769.745	$5.325.10^{+9}$
vdW+Hbond+desolv Energy[a]	0.351	0.0570	0.013	0.5517	34892.442	$1.332.10^{+9}$
Electrostatic Energy[a]	-0.120	0.5264	0.013	0.5517	34890.364	$1.332.10^{+9}$
Final Total Internal Energy[a]	0.060	0.7525	0.006	0.6869	0.924	42.3986
Torsional Free Energy[a]	-0.420	0.0210	0.154	0.0320	0.812	77.7073
Affinity Score[b]	0.422	0.0200	0.112	0.0707	2.470	2.5395
Gauss1 Score[b]	-0.567	$1.097.10^{-3}$	0.254	$4.527.10^{-3}$	29.782	98.0079
Gauss2 Score[b]	-0.668	$5.479.10^{-3}$	0.182	0.0186	468.059	1458.13
Repulsion Score[b]	-0.213	0.2587	0.051	0.2316	1.104	10.2097
Hydrophobic Score[b]	-0.020	0.9154	0.013	0.5458	25.263	52.8671
Hydrogen Score[b]	-0.205	0.2772	0.046	0.2547	1.359	9.50158
Torsions[b]	-0.408	0.0210	0.154	0.0319	2.719	12.4399

Calculated using [a] AD4. [b] Vina. SD: standard deviation. RMSE: root mean square error.

Taking together RMSE and correlation (ρ), we see that the best predictive performance for model 2 obtained using Vina energy terms. Model 3 shows close correlation (ρ) but with poor results for SD and RMSE. Even with poor performance in the SD and RMSE, model 3 was able to capture some interesting features of the protein-ligand interactions for this enzymatic class. Model 3 has three explanatory variables involving interactions of Carbon-Nitrogen, Carbon-Fluorine, and Nitrogen-Nitrogen. Previously published studies[35,36] indicated the importance of halogen atoms in the inhibition of kinases, and this machine learning model captured this essential interaction in the regression equation with the participation of interatomic interactions involving Fluorine. Model 2 also highlights some interesting points related to the structural knowledge of non-specific serine/threonine protein kinases. The repulsion energy term has significant participation in the model 2 equation, which is consistent with hydrophobic interactions observed in inhibitors of this type of kinase.[37]

Table 3. Predictive performance of machine learning models (test set).

Scoring Functions	ρ	p-value1	R^2	p-value2	SD	RMSE
Model 1[a]	0.302	0.1044	0.101	0.0878	0.595	1.126
Model 2[b]	0.484	$6.69_{-3}10$	0.145	0.0382	1.082	1.2426
Model 3[c]	0.401	0.02822	0.070	0.15754	4.87795	13.85343

Calculated using [a] AD4. [b] Vina.[c] Taba. SD: standard deviation. RMSE: root mean square error.

Conclusions

The development of machine learning models using MDM and Taba showed superior predictive performance when compared with classical scoring functions. The targeted-scoring function built using vina energy terms overperformed classical scoring functions and the machine learning models tested here. Also, machine learning models could identify structural features responsible for the inhibition of non-specific serine/threonine protein kinases, such as the importance of hydrophobic interactions and the participation of halogen atoms in the inhibition. Taking together, we may say the use of the abstraction of the scoring function space and machine learning regression algorithm has the potential to develop computational models with superior predictive performance.

Acknowledgments

RSW acknowledges the receipt of a fellowship from the Capes. GBF acknowledges the receipt of a scholarship from the Pontifical Catholic University of Rio Grande do Sul (PUCRS) (Programa de Bolsa/Pesquisa para Alunos da Graduação, BPA program). WFA is a researcher for CNPq (Brazil) (Process Number: 309029/2018-0). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001.

Keywords: Scoring function space, machine learning, kinase, drug design, Taba

((Additional Supporting Information may be found in the online version of this article.))

References and Notes

1. M. Tutone, A. M. Almerico, *Eur. J. Med. Chem*. **2017**, *142*, 300–315.
2. M. Mielecki, B. Lesyng, *Curr. Med. Chem*. **2016**, *23*(10), 954–982.
3. M. Michaelis, F. Selt, F; Rothweiler, M. Wiese, J. Cinatl Jr., *BMC Res. Notes*. **2015**, *8*, 484.
4. P. Labute. *Methods Mol. Biol*. **2018**, *1705*, 159–177.
5. V. G. Maltarollo, T. Kronenberger, B. Windshugel, C. Wrenger, G. H. G.

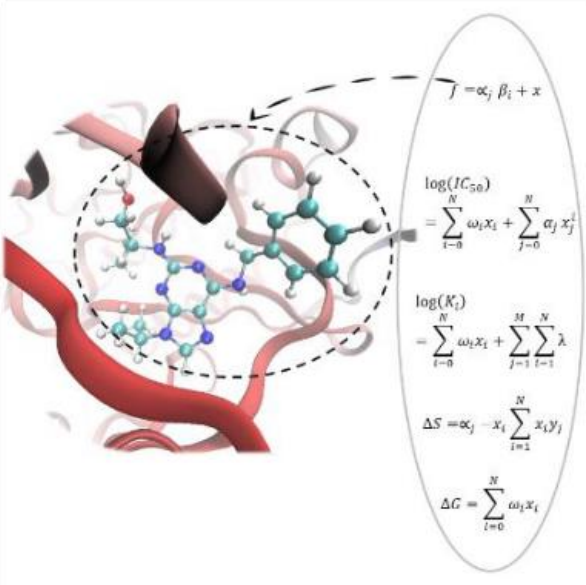
Trossini, K. M. Honorio, *Curr. Drug Targets.* **2018**, *19*(2), 144–154.
6. Q. Zhao, Y. Lu, Y. Zhao, R. Li, F. Luan, M. N; Cordeiro, *Curr. Drug Targets.* **2017**, *18*(5), 576–591.
7. M. Kontoyianni, B. Lacy, *Curr. Med. Chem*. **2018**, *25*(28), 3353–3373.
8. R. S. Rathore, M. Sumakanth, M. S. Reddy, P. Reddanna, A. A. Rao, M. D. Erion, M. R. Reddy, *Curr. Pharm. Des*. **2013**, *19*(26), 4674–4686.
9. W. F. de Azevedo Jr., *Curr. Med. Chem*. **2011**, *18*(9), 1353–1366.
10. G. Bitencourt-Ferreira, W. F. de Azevedo Jr., *Methods Mol. Biol*. **2019**, *2053*, 109–124.
11. J. Westbrook, Z. Feng, L. Chen, H. Yang, H. M. Berman, *Nucleic. Acids. Res*. **2003**, *31*, 489–491.
12. T. Liu, Y. Lin, X. Wen, R. N. Jorissen, M. K. Gilson, *Nucleic Acids Res*. **2007**, *35*, 198–201.
13. L. Hu, M. L. Benson, R. D. Smith, M. G. Lerner, H. A. Carlson, *Proteins*. **2005**, *60*, 333–340.
14. R. Wang, X. Fang, Y. Lu, S. Wang, *J. Med. Chem*. **2004**, *47*, 2977–2980.
15. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
16. J. M. Smith, *Nature.* **1970**, *225*(5232), 563–564.
17. R. S. Bohacek, C. McMartin, W. C. Guida, *Med. Res. Rev*. **1996**, *16*(1), 3-50.
18. C. M. Dobson. *Nature*. **2004**, *432*(7019), 824–828.
19. G. S. Heck, V. O. Pintro, R. R. Pereira, M. B. de Ávila, N. M. B. Levin, W. F. de Azevedo, Affinity. *Curr. Med. Chem*. **2017**, *24*, 2459–2470.
20. G. Bitencourt-Ferreira, W. F. de Azevedo Jr., *Methods Mol. Biol*. **2019**, *2053*, 251–273.
21. G. Bitencourt-Ferreira, W. F. de Azevedo Jr., *Methods Mol. Biol*. **2019**, *2053*, 275–281.
22. R. Thomsen, M. H. Christensen, *J. Med. Chem*. **2006**, *49*, 3315–3321.
23. G. Heberlé, W. F. de Azevedo Jr., *Curr. Med. Chem*. **2011**, *18*(9), 1339–1352.
24. A. D. da Silva, G. Bitencourt-Ferreira, W. F. de Azevedo Jr., *J. Comput. Chem*. **2020**, *41*(1), 69–73.
25. G. Bitencourt-Ferreira, A. D. da Silva, W. F. de Azevedo Jr., *Curr. Med. Chem*. **2020**, doi: 10.2174/2213275912666191102162959
26. M. M. Xavier, G. S. Heck, M. B. Avila, N. M. B. Levin, V. O. Pintro, N. L. Carvalho, W. F. Azevedo, *Comb. Chem. High Throughput Screen*. **2016**, *19*, 801–812.
27. G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, A. J. Olson, *J. Comput. Chem*. **1998**, *19*, 1639–1662.
28. O. Trott, A. J. Olson, *J. Comput. Chem*. **2010**, *31*, 455–461.
29. W. F. de Azevedo Jr., R. Dias, *Bioorg. Med. Chem*. **2008**, *16*(20), 9378–9382.
30. W. F. de Azevedo Jr., R. Dias, *Curr. Drug Targets*. **2008**, *9*(12), 1071–1076.
31. R. J. Tibshirani, R. *Stat. Soc. Series B Stat. Methodol*. **1996**, *58*(1), 267–288.
32. A. N. Tikhonov, *Dokl. Akad. Nauk SSSR*. **1963**, *153*, 49–52.
33. H. Zou, T. J. Hastie, *R. Stat. Soc. Series B Stat. Methodol*. **2005**, *67*(2), 301–220.
34. J. H. Zar, *J. Am. Stat. Assoc.* **1972**, *67*(339), 578–580.
35. M. Z. Hernandes, S. M. Cavalcanti, D. R. Moreira, W. F. de Azevedo, A. C. Leite, *Curr. Drug Targets*. **2010**, *11*, 303–314.
36. M. B. de Ávila, M. M. Xavier, V. O. Pintro, W. F. de Azevedo. *Biochem. Biophys. Res. Commun*. **2017**, *494*, 305–310.
37. M. Rask-Andersen, J. Zhang, D. Fabbro, H. B; Schiöth, *Trends Pharmacol. Sci*. **2014**, *35*(11), 604-620.

GRAPHICAL ABSTRACT

Rosana da Silva Waszak, Gabriela Bitencourt-Ferreira, and Walter Filgueira de Azevedo, Jr.

Application of Machine Learning Methods to Predict Binding Affinity. A Study of Non-specific Serine/Threonine Protein Kinases

This study describes the development of machine learning models to predict inhibition of non-specific Serine/Threonine Protein Kinase. This enzymatic class has protein targets, many of them with application in the development of anticancer drugs. Computational exploration of the scoring function space allows identifying adequate models to predict inhibition of this enzymatic class. The use of these tools can generate models to predict inhibition, which may contribute to the early stages of drug discovery.



APÊNDICE B – MATERIAIS SUPLEMENTARES

Nesta seção são apresentados os resultados dos cálculos dos termos de energia para o *dataset* utilizado, que foram enviados como materiais suplementares junto ao artigo científico submetido.

Supplementary material 1. Predicted binding affinities for all scores avaiable in the program AutoDock4 and experimental log(Ki) (training set). MLF: Multiple linear regression. PLS: Partial least square. SVM: support vector machine. ANN: Artificial neural networks. ANN (6-10-2-1) means an artificial neural networks with 5 explanatory variable + 1 response variable (6), 10 neuron in the first hidden layer, 2 neuros in the second hidden layer, and one output (1).

Ligand	FreeEnergy	FinalIntermolecularEnergy		vdW+Hbond+desolvEnergy		ElectrostaticEnergy	FinalTotalInternalEnergy		TorsionalFreeEnergy	Torsions	log(ki)	MLR-Train (5D)
	PLS-Train (5LC)	SVM-Train (51SV)	ANN (6-4-6-1)	ANN (6-7-2-1)	ANN (6-10-2-1)	ANN (6-8-3-1)	ANN (6-5-4-1)					
MNY_351 A	-8.81	-8.07	-7.99	-0.08	-1.34	0.6	2	-6.45593	-7.3477	-7.3477	-7.40707	-7.37144
	-7.34555	-7.35248	-7.36073	-7.35982								
UCN_400 A	-12.91	-13.28	-11.62	-1.66	-1.41	1.79	6	-8.25181	-7.99496	-7.99496	-7.90245	-7.89322
	-7.8762	-7.88977	-7.89766	-7.87613								
4IP_1552 A	-11.4	-16.13	-5.39	-10.75	1.75	2.98	10	-6.52288	-6.49617	-6.49617	-7.09734	-6.7858
	-6.88798	-6.83052	-6.81746	-6.86869								
4PT_1550 A	-8.49	-14.48	-3.87	-10.61	1.22	4.77	16	-7.67778	-7.37762	-7.37762	-7.5462	-7.76501
	-7.87589	-7.8038	-7.75489	-7.81577								
K25_501 A	-6.66	-6.85	-7.06	0.22	-0.11	0.3	1	-7.39794	-6.73445	-6.73445	-6.73985	-6.66573
	-6.63554	-6.66017	-6.65887	-6.64139								
K44_501 A	-7.32	-7.32	-7.39	0.07	0	0	0	-7	-6.57613	-6.57613	-6.58936	-6.52957
	-6.47786	-6.52662	-6.52608	-6.48164								
PFP_1277 A	-8.27	-8.36	-8.28	-0.08	-1.69	1.79	6	-5.14267	-7.99981	-7.99981	-8.03322	-8.02458
	-7.99397	-8.02087	-8.0328	-7.99778								
PFQ_1277 A	-7.05	-7.82	-7.81	0	-0.43	1.19	4	-4.86328	-7.26086	-7.26086	-7.22419	-7.14319
	-7.15642	-7.13696	-7.13529	-7.15825								
DFZ_1275 A	-6.72	-7.08	-7.07	-0.01	-1.13	1.49	5	-5.88606	-7.60353	-7.60353	-7.66291	-7.63246
	-7.62802	-7.62569	-7.63074	-7.62887								
DBQ_1271 A	-7.22	-7.7	-6.9	-0.8	0.19	0.3	1	-6.18111	-6.55917	-6.55917	-6.62604	-6.54801
	-6.51133	-6.54921	-6.54815	-6.51373								
IDZ_1274 A	-6.43	-6.35	-6.39	0.03	-0.37	0.3	1	-5.24195	-6.78192	-6.78192	-6.84569	-6.77626
	-6.74952	-6.76541	-6.76455	-6.75815								
76A_1001 A	1.41e+06	1.41e+06	704000	704000	-0.82	1.79	6	-8.20066	-8.51739	-8.51739	-7.86392	-8.23078
	-8.21926	-8.24202	-8.24454	-8.20371								
77A_1001 A	-11.23	-10.74	-8.76	-1.99	-2.58	2.09	7	-7.69897	-8.37431	-8.37431	-8.514	-8.45651
	-8.39113	-8.44844	-8.47027	-8.41318								
A25_1001 A	-9.03	-9.03	-8.78	-0.25	0	0	0	-8.10127	-6.63549	-6.63549	-6.58935	-6.52957
	-6.47786	-6.52662	-6.52608	-6.48164								
85A_1001 A	324000	324000	162000	162000	-0.48	1.49	5	-7.89296	-7.18275	-7.18275	-7.55674	-7.45635
	-7.48044	-7.45565	-7.45523	-7.47336								

M77_416 A	-9.69 -7.15256	-9.33 -7.15178	-7.39 -7.15503	-1.94 -7.16404	-0.95	0.6	2	-6.41266	-7.06845	-7.06845	-7.22275	-7.16873
RAD_108 A	215000 -8.82448	215000 -8.84935	108000 -8.87314	108000 -8.89608	-3.13	4.18	14	-8.39794	-9.0396	-9.0396	-8.73582	-8.90159
A53_1 A	-8.72 -8.09383	-10.25 -8.1068	-10.15 -8.10596	-0.09 -8.08389	-1.16	2.68	9	-6.89963	-8.28886	-8.28886	-8.04444	-8.10285
T12_600 A	-8.43 -8.3101	-10.46 -8.33408	-9.81 -8.33098	-0.64 -8.30355	-1.26	3.28	11	-8.92082	-8.53218	-8.53218	-8.19292	-8.33088
K32_338 A	-6.42 -6.47786	-6.42 -6.52662	-6.43 -6.52608	0.01 -6.48164	0	0	0	-6.82391	-6.51954	-6.51954	-6.58936	-6.52957
K22_501 A	-6.52 -6.64841	-6.67 -6.67189	-6.67 -6.67061	0 -6.65459	-0.14	0.3	1	-6.69897	-6.71159	-6.71159	-6.75174	-6.67809
P19_501 A	-7.39 -7.29897	-7.87 -7.28374	-7.99 -7.28452	0.13 -7.3019	-0.72	1.19	4	-6.58503	-7.38421	-7.38421	-7.35502	-7.29261
P44_501 A	-6.3 -7.31631	-7.87 -7.28282	-8.95 -7.2775	1.08 -7.30884	-0.22	1.79	6	-8.18709	-7.56325	-7.56325	-7.36154	-7.28107
P29_501 A	-4.97 -7.39682	-7.51 -7.34888	-7.49 -7.33742	-0.02 -7.37899	0.15	2.39	8	-6.4437	-7.54186	-7.54186	-7.41405	-7.33741
P63_501 A	-8.84 -7.19894	-10.65 -7.16208	-10.68 -7.15557	0.03 -7.19136	0.02	1.79	6	-9.45593	-7.51312	-7.51312	-7.26135	-7.15753
GVF_1116 A	-11.35 -6.25182	-16.58 -6.30043	-6.73 -6.3028	-9.85 -6.23851	2.85	2.39	8	-7.09691	-5.96597	-5.96597	-6.7579	-6.23636
ZZL_1390 A	-10.72 -7.56308	-10.65 -7.56512	-10.31 -7.57255	-0.34 -7.56898	-1.26	1.19	4	-8.45182	-7.68452	-7.68452	-7.60899	-7.57597
REF_336 A	-1.91 -8.30418	-1.15 -8.36204	-1.08 -8.38541	-0.07 -8.32445	-2.55	1.79	6	-7.69897	-7.91572	-7.91572	-8.41916	-8.36917
P04_501 A	-10.48 -6.78168	-10.91 -6.7875	-10.82 -6.78573	-0.08 -6.78743	-0.17	0.6	2	-7.61979	-7.07642	-7.07642	-6.87896	-6.79362
VX1_314 A	-7.7 -7.02577	-8.16 -7.015	-7.92 -7.01386	-0.23 -7.03108	-0.43	0.89	3	-7.04096	-7.12616	-7.12616	-7.10457	-7.02374
VX2_314 A	-8.36 -7.50503	-8.41 -7.50239	-8.28 -7.50814	-0.13 -7.51014	-1.14	1.19	4	-7.95861	-7.53962	-7.53962	-7.55176	-7.51308
VX3_314 A	-9.23 -7.20636	-9.32 -7.19751	-8.7 -7.19935	-0.63 -7.21361	-0.8	0.89	3	-6.25964	-7.28268	-7.28268	-7.27139	-7.21024
EMO_400 A	-6.02 -7.42538	-6.8 -7.40713	-6.65 -7.40713	-0.15 -7.42389	-0.71	1.49	5	-5.73283	-7.41872	-7.41872	-7.46717	-7.41249
324_2 A	-9.65 -7.7128	-10.96 -7.69911	-10.91 -7.69718	-0.05 -7.70287	-0.78	2.09	7	-8.58503	-7.94233	-7.94233	-7.71168	-7.69793
G93_1 A	-14.28 -8.32032	-15.14 -8.35909	-12.88 -8.36774	-2.26 -8.32577	-1.82	2.68	9	-8.52288	-8.56153	-8.56153	-8.32729	-8.36065
G96_1 A	-11.88 -8.59713	-13.3 -8.63873	-13.1 -8.64849	-0.2 -8.6228	-2.15	3.58	12	-8.85387	-9.19295	-9.19295	-8.5481	-8.65313
T74_1 A	97000 -8.28	97000 -8.31279	48500 -8.318	48500 -8.28004	-1.59	2.68	9	-8.39794	-7.95066	-7.95066	-8.25538	-8.31197

TXQ_338 A	-8.51 -7.54379	-8.49 -7.54423	-8.21 -7.55109	-0.27 -7.54942	-1.22	1.19	4	-7.284	-7.55725	-7.55725	-7.58988	-7.55504
L1E_1 B	87800 -7.78651	87800 -7.79741	43900 -7.80618	43900 -7.7886	-1.38	1.49	5	-9	-7.35995	-7.35995	-7.82146	-7.80302
AGX_901 A	-6.69 -7.34203	-7.63 -7.3191	-7.7 -7.31748	0.06 -7.33994	-0.54	1.49	5	-5.72354	-7.4264	-7.4264	-7.39004	-7.32333
AGY_900 A	-8.16 -7.59474	-8.59 -7.58942	-9.16 -7.59354	0.57 -7.59511	-1.06	1.49	5	-6.83565	-7.72604	-7.72604	-7.62991	-7.59607
1EL_332 A	-6.78 -6.57656	-7.11 -6.60699	-7.06 -6.60571	-0.05 -6.58084	0.03	0.3	1	-6.0655	-6.66823	-6.66823	-6.68556	-6.60947
3NG_338 A	-11.03 -7.0499	-11.45 -7.03898	-10.55 -7.03816	-0.89 -7.05548	-0.48	0.89	3	-9.52071	-7.25375	-7.25375	-7.12658	-7.04839
E1B_338 A	-10.86 -7.04024	-11.29 -7.02936	-10.32 -7.02841	-0.97 -7.04571	-0.46	0.89	3	-9.75696	-7.22924	-7.22924	-7.11775	-7.0385
CZ0_1 A	-5.44 -6.44916	-7.22 -6.48326	-6.39 -6.48212	-0.83 -6.44558	0.89	0.89	3	-6.65758	-6.52479	-6.52479	-6.60941	-6.46569
04G_404 A	-6.04 -6.68841	-6.68 -6.70095	-6.76 -6.69892	0.08 -6.6924	0.04	0.6	2	-5.88606	-6.78318	-6.78318	-6.79486	-6.70288
0TJ_1 B	-5.92 -7.08823	-6.82 -7.06822	-6.81 -7.06568	-0.02 -7.08959	-0.29	1.19	4	-5.83863	-7.15302	-7.15302	-7.16301	-7.07277
A58_301 A	270000 -7.67731	270000 -7.67138	135000 -7.67522	135000 -7.67346	-0.95	1.49	5	-7.69897	-7.31999	-7.31999	-7.71492	-7.67413
HK4_301 A	74000 -7.90534	74000 -7.92017	37000 -7.92796	37000 -7.90459	-1.4	1.79	6	-8.29158	-7.48642	-7.48642	-7.92881	-7.92266
HK6_301 A	-6.85 -6.47786	-6.85 -6.52662	-6.73 -6.52608	-0.12 -6.48164	0	0	0	-4.78187	-6.52897	-6.52897	-6.58936	-6.52957
HK7_301 A	-8.32 -7.16131	-9.07 -7.14193	-9.04 -7.14033	-0.03 -7.16318	-0.44	1.19	4	-6.40012	-7.3311	-7.3311	-7.22861	-7.14827
HK9_301 A	-8.66 -6.87891	-9.43 -6.87186	-9.32 -6.86935	-0.11 -6.88256	-0.12	0.89	3	-7.10846	-7.09701	-7.09701	-6.97241	-6.87582
H2K_300 A	-8.33 -7.76067	-8.97 -7.7616	-8.01 -7.76586	-0.96 -7.75777	-1.15	1.79	6	-7.5817	-7.73834	-7.73834	-7.78084	-7.76476
H3K_301 A	-8.61 -8.05779	-8.55 -8.09183	-8.76 -8.10605	0.21 -8.06417	-1.85	1.79	6	-8.92812	-8.1011	-8.1011	-8.10733	-8.09578
H5K_301 A	158000 -7.62647	158000 -7.61908	79100 -7.62261	79100 -7.62421	-0.96	1.49	5	-7.07988	-7.21324	-7.21324	-7.66236	-7.62349
6HK_301 A	151000 -7.26353	151000 -7.23262	75600 -7.2295	75600 -7.25874	-0.22	1.49	5	-7.11182	-6.93785	-6.93785	-7.3377	-7.23313
7HK_301 A	-8.18 -7.3261	-9.73 -7.29302	-9.5 -7.28783	-0.23 -7.31866	-0.24	1.79	6	-7.96859	-7.5284	-7.5284	-7.37008	-7.29147
1OA_401 A	-10.13 -7.42357	-9.79 -7.4253	-8.39 -7.43251	-1.39 -7.43343	-1.24	0.89	3	-7.82391	-7.38438	-7.38438	-7.47865	-7.44019
1OB_402 A	-10.4 -8.05388	-10.35 -8.08748	-9.24 -8.10157	-1.11 -8.06009	-1.84	1.79	6	-8.95861	-8.05084	-8.05084	-8.10271	-8.09142

1OC_402 A	-11.05 -8.12264	-10.82 -8.16368	-9.05 -8.18029	-1.77 -8.13202	-2.02	1.79	6	-11.5229	-8.06957	-8.06957	-8.18521	-8.16804
EXX_408 A	-7.22 -7.22018	-7.85 -7.20212	-7.82 -7.20144	-0.03 -7.22248	-0.56	1.19	4	-7.39794	-7.30734	-7.30734	-7.28221	-7.20967
29L_801 A	-7.93 -7.64494	-8.82 -7.63429	-8.89 -7.63533	0.07 -7.64001	-0.9	1.79	6	-9.88606	-7.7528	-7.7528	-7.66481	-7.63688
2HX_401 A	-10.54 -8.51224	-9.01 -8.57939	-8.73 -8.61028	-0.28 -8.55091	-3.32	1.79	6	-9.60076	-8.61035	-8.61035	-8.72132	-8.59575
2OL_601 A	-3.41 -7.97947	-5.22 -7.97819	-5 -7.97337	-0.22 -7.96441	-0.87	2.68	9	-5.5376	-7.88977	-7.88977	-7.9193	-7.97242
2OQ_601 A	-6.64 -7.19481	-7.9 -7.16683	-7.82 -7.16303	-0.08 -7.19201	-0.24	1.49	5	-7.16749	-7.31547	-7.31547	-7.25814	-7.16816
4T3_601 A	-6.98 -7.4693	-7.67 -7.45399	-7.3 -7.45495	-0.37 -7.4682	-0.8	1.49	5	-7.65758	-7.47552	-7.47552	-7.50857	-7.45981
5E6_401 A	-10.88 -7.82131	-10.83 -7.83848	-10.69 -7.84957	-0.13 -7.82602	-1.55	1.49	5	-9.40121	-7.95145	-7.95145	-7.86228	-7.84531

Supplementary material 2. Predicted binding affinities for all scores avaiable in the program AutoDock4 and experimental log(K_i) (test set). MLF: Multiple linear regression. PLS: Partial least square. SVM: support vector machine. ANN: Artificial neural networks. ANN (6-10-2-1) means an artificial neural networks with 5 explanatory variable + 1 response variable (6), 10 neuron in the first hidden layer, 2 neuros in the second hidden layer, and one output (1).

Ligand	FreeEnergy	FinalIntermolecularEnergy		vdW+Hbond+desolvEnergy		ElectrostaticEnergy	FinalTotalInternalEnergy		TorsionalFreeEnergy	Torsions	log(ki)	MLR (5D)
	PLS (5LC)	SVM (51SV)	ANN (6-4-6-1)	ANN (6-7-2-1)	ANN (6-10-2-1)	ANN (6-8-3-1)	ANN (6-5-4-1)					
HNA_351 A	-7.88	-6.91	-6.68	-0.23	-1.87	0.89	3	-6.10791	-7.58448	-7.57788	-7.78188	-7.76729
	-7.72583	-7.75321	-7.77041	-7.7408								
MNX_351 A	-7.45	-7.39	-7.22	-0.17	-0.66	0.6	2	-6.09691	-7.05107	-7.04985	-7.09007	-7.02322
	-7.01115	-7.00929	-7.00988	-7.0206								
STU_400 A	-1.34	-2.24	-0.58	-1.66	0	0.89	3	-8.10791	-6.48219	-6.48264	-6.9235	-6.82107
	-6.8236	-6.81927	-6.8165	-6.82657								
UCM_400 A	-5.96	-6.31	-6.02	-0.28	-0.55	0.89	3	-7.82391	-7.0619	-7.06027	-7.1577	-7.08322
	-7.08387	-7.07296	-7.07263	-7.08981								
IQA_338 A	-7.6	-8.07	-6.98	-1.09	-0.13	0.6	2	-6.76955	-6.7927	-6.79348	-6.8626	-6.77596
	-6.76365	-6.77061	-6.76876	-6.76908								
K37_501 A	-7.27	-7.51	-7.52	0.01	-0.06	0.3	1	-7.1549	-6.73002	-6.73191	-6.72022	-6.64538
	-6.61429	-6.6409	-6.6396	-6.61958								
ABO_1271 A	-11.61	-11.34	-11.5	0.16	-1.16	0.89	3	-7.58503	-7.61364	-7.60971	-7.44046	-7.39809
	-7.38417	-7.38339	-7.3895	-7.39351								
CKI_300 A	-6.74	-6.68	-5.84	-0.83	-1.56	1.49	5	-4.40894	-7.6472	-7.64007	-7.86702	-7.85023
	-7.82577	-7.84341	-7.85465	-7.83058								
Y27_416 A	-5.74	-6.49	-6.42	-0.07	-0.45	1.19	4	-6.96508	-7.18718	-7.18508	-7.23303	-7.15337
	-7.16622	-7.14691	-7.14539	-7.16812								
K17_1001 A	-6.14	-6.14	-6.13	0	0	0	0	-6.52288	-6.50234	-6.5053	-6.58935	-6.52958
	-6.47787	-6.52662	-6.52609	-6.48165								
P45_501 A	-6.41	-8.19	-9.12	0.93	-0.3	2.09	7	-8.18709	-7.72169	-7.71761	-7.50065	-7.44803
	-7.48552	-7.45207	-7.4455	-7.47322								
P55_501 A	-5.34	-7.64	-7.88	0.23	0.52	1.79	6	-7.61979	-7.18585	-7.18571	-7.06878	-6.91232
	-6.9591	-6.92481	-6.91764	-6.95184								
RFZ_336 A	-3.91	-4.77	-4.81	0.04	-0.63	1.49	5	-4.67847	-7.2979	-7.29427	-7.43067	-7.37048
	-7.38621	-7.36561	-7.36481	-7.38441								
3O0_293 B	-8.89	-10.02	-9.24	-0.78	-0.95	2.09	7	-6.22185	-7.87141	-7.86488	-7.78855	-7.78407
	-7.79014	-7.78464	-7.78477	-7.78171								
ABQ_360 A	-6.8	-9.97	-9.68	-0.28	0.78	2.39	8	-9.22185	-7.41833	-7.41747	-7.18617	-7.01725
	-7.09176	-7.03835	-7.02661	-7.07598								
3RW_1 A	-9.05	-9.62	-9.29	-0.32	-2.12	2.68	9	-7.45593	-8.5798	-8.56721	-8.45142	-8.45878
	-8.40726	-8.45364	-8.46641	-8.42092								

N53_601 A	336000	336000	168000	168000	-0.9	1.79	6	-7.19382	-7.47779	-7.4852	-7.81522	-7.80195
	-7.80737	-7.80357	-7.80558	-7.79834								
RAP_201 A	148000	148000	74000	74000	-3.41	3.88	13	-8.52288	-9.16436	-9.1492	-8.86974	-8.89947
	-8.81978	-8.84718	-8.87348	-8.8927								
T77_401 B	-9.92	-10.24	-10.2	-0.04	-2.37	2.68	9	-9.52288	-8.73734	-8.7238	-8.55118	-8.53166
	-8.47228	-8.5229	-8.53891	-8.49311								
HKC_300 A	158000	158000	79100	79100	-0.84	1.19	4	-7.88606	-7.04387	-7.04637	-7.49892	-7.43573
	-7.44027	-7.42784	-7.43083	-7.44175								
H1K_301 A	-7.34	-7.15	-7.01	-0.14	-1.39	1.19	4	-6.19997	-7.56012	-7.55448	-7.67125	-7.64363
	-7.62525	-7.63278	-7.64211	-7.63214								
H4K_300 A	-7.32	-7.67	-7.42	-0.26	-0.85	1.19	4	-7.6925	-7.37855	-7.37494	-7.41515	-7.36063
	-7.36303	-7.35095	-7.35311	-7.36654								
5HK_301 A	-9.23	-10.17	-10.31	0.14	-0.55	1.49	5	-8.2186	-7.5793	-7.57598	-7.39452	-7.32857
	-7.34694	-7.32426	-7.32272	-7.34488								
2HV_401 A	-9.29	-8.77	-8.67	-0.09	-1.71	1.19	4	-7.03937	-7.77213	-7.76528	-7.82484	-7.80655
	-7.77407	-7.79616	-7.81045	-7.78388								
2HW_401 A	-9.32	-8.58	-8.33	-0.25	-2.23	1.49	5	-7.79588	-8.06279	-8.05313	-8.18049	-8.15641
	-8.10233	-8.14971	-8.17078	-8.11671								
2OO_601 A	-6.88	-9.13	-8.82	-0.31	0.16	2.09	7	-7.22185	-7.46786	-7.46552	-7.31145	-7.20824
	-7.26153	-7.21722	-7.20823	-7.24907								
2QV_403 A	-8.72	-8.71	-8.64	-0.08	-0.31	0.3	1	-6.96658	-6.87868	-6.87963	-6.82071	-6.75008
	-6.72274	-6.74039	-6.73937	-6.73074								
3FV_501 A	-11.37	-10.29	-8.7	-1.59	-1.97	0.89	3	-8.39794	-7.6577	-7.65073	-7.82978	-7.81738
	-7.77172	-7.80366	-7.82254	-7.78771								
4KH_501 A	-6.58	-7.08	-6.95	-0.13	-0.39	0.89	3	-5.95861	-7.06325	-7.06223	-7.08707	-7.00418
	-7.00655	-6.99598	-6.99461	-7.01165								
4TW_501 A	-9.85	-10.7	-10.54	-0.16	-0.93	1.79	6	-8	-7.84256	-7.83696	-7.67864	-7.65241
	-7.65904	-7.64972	-7.65112	-7.65432								

Supplementary material 3. Predicted binding affinities for all scores avaiable in the program AutoDock Vina and experimental log(K_i) (training set). MLF: Multiple linear regression. PLS: Partial least square. SVM: support vector machine. ANN: Artificial neural networks. ANN (6-10-2-1) means an artificial neural networks with 5 explanatory variable + 1 response variable (6), 10 neuron in the first hidden layer, 2 neuros in the second hidden layer, and one output (1).

Ligand	Affinity	Gauss1	Gauss2	Repulsion	Hydrophobic	Hydrogen	Torsions	log(ki)	MLR-Train (5D)	PLS-Train (5LC)	SVM-Train (51SV)	
	ANN (6-4-9-1)	ANN (6-6-9-1)	ANN (6-9-9-1)	ANN (6-10-6-1)	ANN (6-4-6-1)							
000MNY_351 A	-9.97395	87.3634	1276.25	2.74709	52.8681	2.25291	2	-6.45593	-7.37259	-7.37259	-7.169	
	-7.14602	-7.20101	-7.24076	-7.13184	-7.17826							
000UCN_400 A	-12.5576	110.817	1896.4	4.07525	63.8518	4.41865	6	-8.25181	-8.82189	-8.82189	-8.64646	
	-8.67732	-8.70246	-8.76577	-8.73308	-8.69055							
0004IP_1552 A	-4.82996	60.0788	828.153	5.92484	0	10.1108	10	-6.52288	-7.11069	-7.11069	-6.87957	
	-7.03271	-7.02099	-7.04743	-7.04001	-7.02422							
0004PT_1550 A	-4.04447	57.7193	851.194	6.15489	0	10.3476	16	-7.67778	-6.7991	-6.7991	-7.341	
	-6.88158	-6.85046	-6.89043	-6.88016	-6.83436							
000K25_501 A	-6.49823	55.9399	974.243	1.45906	31.1351	0	1	-7.39794	-6.37674	-6.37674	-6.32916	
	-5.90518	-5.87434	-5.83152	-5.94589	-5.86822							
000K44_501 A	-6.68922	60.5626	959.955	2.1018	38.5742	0	0	-7	-6.59581	-6.59581	-6.43524	
	-5.97848	-5.99779	-5.97884	-5.98915	-5.96669							
000PFP_1277 A	-8.17057	82.7315	1433.13	1.5417	37.3751	0.75812	6	-5.14267	-7.00867	-7.00867	-7.11494	
	-6.87713	-6.89905	-6.97181	-6.87086	-6.88797							
000PFQ_1277 A	-8.26417	75.4718	1278.46	1.42538	45.0373	0.50209	4	-4.86328	-6.79299	-6.79299	-6.81565	
	-6.45405	-6.49743	-6.55263	-6.41276	-6.47339							
000DFZ_1275 A	-7.15236	64.8289	1283.9	1.63401	34.3274	0.47018	5	-5.88606	-6.79376	-6.79376	-6.82751	
	-6.36437	-6.41501	-6.48054	-6.33738	-6.38621							
000DBQ_1271 A	-8.57964	74.1758	1117.48	2.83283	24.0014	3.34557	1	-6.18111	-7.19521	-7.19521	-6.68178	
	-6.78827	-6.8586	-6.88434	-6.73922	-6.83602							
000IDZ_1274 A	-4.55119	62.4264	1037.09	5.13539	32.1728	0.74713	1	-5.24195	-7.55833	-7.55833	-6.91738	
	-6.54737	-6.72416	-6.84982	-6.42243	-6.6121							
00076A_1001 A	-9.60912	101.513	1611.31	4.19666	69.9388	2.6797	6	-8.20066	-8.19174	-8.19174	-8.32473	
	-8.20897	-8.22818	-8.25433	-8.33573	-8.22398							
00077A_1001 A	-6.94532	97.442	1458.55	3.22535	20.1647	1.35649	7	-7.69897	-7.47575	-7.47575	-7.80487	
	-7.49364	-7.51069	-7.55245	-7.52918	-7.48199							
000A25_1001 A	-10.2831	73.8618	1423.17	1.46305	17.2314	1.60508	0	-8.10127	-7.44354	-7.44354	-7.08	
	-7.12629	-7.18929	-7.22189	-7.17377	-7.19575							
00085A_1001 A	-8.12386	78.8119	1677.52	3.00686	13.4342	1.47837	5	-7.89296	-8.02906	-8.02906	-8.01947	
	-7.88749	-7.93358	-7.98922	-8.04537	-7.92977							
000M77_416 A	-8.10671	69.844	1258.28	2.38374	33.3764	1.54683	2	-6.41266	-7.21188	-7.21188	-6.95698	
	-6.72794	-6.81509	-6.87501	-6.68572	-6.77624							

000RAD_108 A	-16.2404	167.729	3065.56	4.19468	144.164	4.94847	14	-8.39794	-10.6936	-10.6936	-8.1985
-8.99497	-9.09774	-9.14501	-8.88263	-8.93769							
000A53_1 A	-8.30211	101.688	1696.14	3.18376	46.815	1.85205	9	-6.89963	-7.83171	-7.83171	-8.17303	-8.05463
-8.02537	-8.00792	-8.18283	-8.05293								
000T12_600 A	-8.09817	113.052	1783.48	4.38896	45.5229	2.90848	11	-8.92082	-8.24388	-8.24388	-8.5832
-8.43928	-8.41623	-8.36248	-8.53223	-8.43513							
000K32_338 A	-5.11733	59.2277	897.146	3.37411	34.7239	0	0	-6.82391	-6.81971	-6.81971	-6.48629
-6.00867	-6.07574	-6.08309	-6.00097	-6.01582							
000K22_501 A	-5.77288	59.5055	861.197	2.04109	31.2154	0	1	-6.69897	-6.30222	-6.30222	-6.2419
-5.85652	-5.81285	-5.74934	-5.90737	-5.80753							
000P19_501 A	-7.93414	81.5891	1353	1.7673	35.3981	0.2701	4	-6.58503	-7.02905	-7.02905	-7.08833
-6.72091	-6.77844	-6.85221	-6.67426	-6.74363							
000P44_501 A	-8.55311	100.215	1524.15	1.99235	28.8208	0.92067	6	-8.18709	-7.32087	-7.32087	-7.5096
-7.43826	-7.4246	-7.45529	-7.49029	-7.4341							
000P29_501 A	-7.43911	96.9097	1494.65	2.00216	27.78	0.428	8	-6.4437	-7.09715	-7.09715	-7.40024
-7.16077	-7.1533	-7.21457	-7.16135	-7.14146							
000P63_501 A	-9.07494	121.103	1734.46	4.98494	45.2635	2.2942	6	-9.45593	-8.63882	-8.63882	-8.72072
-8.53438	-8.5717	-8.60662	-8.60731	-8.53941							
000GVF_1116 A	-5.08799	71.5872	734.931	6.54082	0.8508	11.2241	8	-7.09691	-7.27729	-7.27729	-6.75832
-7.28734	-7.24701	-7.26159	-7.30521	-7.27725							
000ZZL_1390 A	-10.0212	95.867	1763.31	3.29362	58.3161	1	4	-8.45182	-8.32902	-8.32902	-8.35533
-8.21761	-8.26275	-8.31156	-8.35803	-8.25306							
000REF_336 A	-7.47707	67.6346	1277.34	2.69888	33.4595	1.52609	6	-7.69897	-7.04913	-7.04913	-7.10322
-6.64153	-6.71334	-6.79814	-6.6009	-6.66637							
000P04_501 A	-11.792	108.902	1578.43	2.8155	64.4783	2.15087	2	-7.61979	-8.00254	-8.00254	-7.90405
-8.09623	-8.10562	-8.11653	-8.2244	-8.12663							
000VX1_314 A	-8.11823	59.9951	1273.28	1.89941	46.1253	1	3	-7.04096	-7.01659	-7.01659	-6.84725
-6.47039	-6.54847	-6.61261	-6.44081	-6.51327							
000VX2_314 A	-9.16559	68.388	1439.7	1.89867	70.7248	0.96198	4	-7.95861	-7.28293	-7.28293	-7.19601
-6.92097	-6.99024	-7.06966	-6.93477	-6.96517							
000VX3_314 A	-10.535	71.4478	1334.54	1.6386	94.0989	1.75822	3	-6.25964	-7.09911	-7.09911	-6.91472
-6.7843	-6.83527	-6.88619	-6.77847	-6.82513							
000EMO_400 A	-8.82962	75.8797	1187.52	1.76683	52.5849	1.59465	5	-5.73283	-6.67531	-6.67531	-6.69536
-6.40705	-6.43576	-6.47375	-6.36558	-6.41716							
000324_2 A	-8.953	99.5032	1580.74	3.81558	69.565	2.42185	7	-8.58503	-7.93951	-7.93951	-8.18912	-8.02993
-8.03443	-8.04669	-8.15374	-8.03575								

000G93_1 A	-9.55844	132.065	1888.69	5.69006	34.923	5.36285	9	-8.52288	-9.0682	-9.0682	-8.72087	-8.79293
	-8.82806	-8.87022	-8.78978	-8.78158								
000G96_1 A	-9.34873	129.235	2008.31	3.70187	40.7452	3.08694	12	-8.85387	-8.44633	-8.44633	-8.51731	-8.66331
	-8.62286	-8.52585	-8.69819	-8.66005								
000T74_1 A	-10.0376	123.284	2099.3	3.81956	50.8915	2.59893	9	-8.39794	-8.86648	-8.86648	-8.85387	-8.73272
	-8.75049	-8.74851	-8.75597	-8.73438								
000TXQ_338 A		-10.997	100.035	1324.7	2.52734	53.8064	3.62405	4	-7.284	-7.32163	-7.32163	-7.25134
	-7.48522	-7.45909	-7.457	-7.54908	-7.49104							
000L1E_1 B	-12.4142	141.389	2008.72	4.03007	69.3814	2.73882	5	-9	-9.02414	-9.02414	-8.81185	-8.78087
	-8.82251	-8.86686	-8.77966	-8.77656								
000AGX_901 A		-5.95677	89.391	881.011	1.91925	39.5815	0.0361	5	-5.72354	-6.00572	-6.00572	-6.31962
	-5.92177	-5.86887	-5.7984	-5.93861	-5.87391							
000AGY_900 A		-7.52033	103.13	1119.58	3.46359	54.3266	2.17807	5	-6.83565	-7.01787	-7.01787	-7.23855
	-6.90209	-6.9421	-6.98728	-6.79302	-6.88754							
0001EL_332 A		-9.43637	62.1923	1115.9	0.67278	55.4279	0.62679	1	-6.0655	-6.47677	-6.47677	-6.40097
	-6.07048	-6.06614	-6.04763	-6.07837	-6.06349							
0003NG_338 A		-10.8302	105.64	1340.47	2.3768	68.4392	2.82654	3	-9.52071	-7.34658	-7.34658	-7.30713
	-7.49224	-7.47587	-7.47394	-7.53931	-7.49915							
000E1B_338 A		-10.7942	106.352	1369.28	2.56172	66.5522	2.83464	3	-9.75696	-7.45767	-7.45767	-7.41899
	-7.61045	-7.59616	-7.59351	-7.67728	-7.61986							
000CZ0_1 A	-7.97661	64.639	1109.69	2.42333	53.385	1.79488	3	-6.65758	-6.85602	-6.85602	-6.64103	-6.34116
	-6.40658	-6.44248	-6.29633	-6.36952								
00004G_404 A		-7.54757	66.06	991.368	1.91495	40.5427	1.21897	2	-5.88606	-6.51556	-6.51556	-6.36266
	-6.06894	-6.08409	-6.06521	-6.06087	-6.06434							
0000TJ_1 B	-5.56235	54.8462	1014.52	3.12706	32.4098	1.44553	4	-5.83863	-6.77114	-6.77114	-6.59182	-6.13197
	-6.19465	-6.2234	-6.11011	-6.1487								
000A58_301 A		-9.16528	79.9052	1471.56	2.13214	63.345	1.6768	5	-7.69897	-7.36732	-7.36732	-7.36253
	-7.24591	-7.27512	-7.33111	-7.30356	-7.27181							
000HK4_301 A		-9.54149	105.97	1562.41	3.9019	65.9202	2.97892	6	-8.29158	-8.01953	-8.01953	-8.17925
	-8.14741	-8.14602	-8.15151	-8.2711	-8.15651							
000HK6_301 A		-7.53136	50.6226	908.22	1.25031	33.0266	1.60496	0	-4.78187	-6.32867	-6.32867	-6.09281
	-5.90418	-5.85976	-5.8004	-5.95197	-5.86344							
000HK7_301 A		-9.74542	84.5866	1262.84	1.55391	48.112	2.63751	4	-6.40012	-6.88688	-6.88688	-6.78518
	-6.80564	-6.81111	-6.83715	-6.78933	-6.82085							
000HK9_301 A		-10.2663	88.5372	1397.81	2.33229	45.8598	3.50751	3	-7.10846	-7.49224	-7.49224	-7.24327
	-7.5375	-7.53381	-7.54311	-7.64291	-7.56975							

000H2K_300 A	-7.49083	86.125	1366.63	2.0305	15.0233	1.65876	6	-7.5817	-7.04501	-7.04501	-7.16055
-6.96424	-6.98039	-7.03806	-6.95489	-6.96893							
000H3K_301 A	-8.57178	95.3316	1460.42	2.93436	16.7999	3.46942	6	-8.92812	-7.56661	-7.56661	-7.63014
-7.74537	-7.71942	-7.72381	-7.8648	-7.75251							
000H5K_301 A	-7.60731	83.5284	1350.9	3.0642	24.9266	2.32483	5	-7.07988	-7.40312	-7.40312	-7.41824
-7.24315	-7.28619	-7.34205	-7.26195	-7.25958							
0006HK_301 A	-8.46275	69.9389	1318.1	1.45489	32.7422	2.51864	5	-7.11182	-6.90364	-6.90364	-6.77876
-6.70926	-6.72858	-6.77431	-6.71657	-6.73555							
0007HK_301 A	-8.29184	90.7987	1421.18	3.50976	38.8304	3.37847	6	-7.96859	-7.64337	-7.64337	-7.70493
-7.69605	-7.69941	-7.72401	-7.79883	-7.70453							
0001OA_401 A	-9.781	70.0673	1474.85	1.34361	55.3488	1	3	-7.82391	-7.27494	-7.27494	-7.11731
-6.97374	-7.02819	-7.09246	-7.01089	-7.02446							
0001OB_402 A	-8.84382	93.3054	1577.81	2.83705	54.4413	1.19881	6	-8.95861	-7.68273	-7.68273	-7.87763
-7.69755	-7.71442	-7.74956	-7.79703	-7.70855							
0001OC_402 A	-8.81367	83.02	1829.68	3.54357	37.7128	2	6	-11.5229	-8.44223	-8.44223	-8.45064
-8.29457	-8.33551	-8.39349	-8.45086	-8.3334							
000EXX_408 A	-6.38131	62.9614	1010.58	1.7473	34.4354	0.51914	4	-7.39794	-6.33419	-6.33419	-6.3718
-5.97158	-5.95347	-5.92453	-5.99171	-5.94689							
00029L_801 A	-9.90301	89.3674	1533.21	2.211	68.6719	1.98464	6	-9.88606	-7.45341	-7.45341	-7.54005
-7.53194	-7.52283	-7.55259	-7.63286	-7.54439							
0002HX_401 A	-9.91022	92.1149	1807.28	4.28778	54.4162	3.23486	6	-9.60076	-8.65293	-8.65293	-8.57809
-8.47701	-8.51619	-8.58558	-8.59595	-8.50475							
0002OL_601 A	-8.68549	103.732	1668.78	3.20442	39.5341	2.99455	9	-5.5376	-7.83064	-7.83064	-8.12197
-8.13655	-8.08707	-8.05438	-8.27149	-8.13686							
0002OQ_601 A	-10.4889	105.841	1600.54	3.27741	43.2402	3.67744	5	-7.16749	-8.02882	-8.02882	-8.02053
-8.2256	-8.20641	-8.20492	-8.36419	-8.2461							
0004T3_601 A	-6.6345	59.3245	1222.12	2.71512	18.9548	2.6921	5	-7.65758	-7.06792	-7.06792	-6.87579
-6.6164	-6.68746	-6.75266	-6.59254	-6.65396							
0005E6_401 A	-10.1955	94.3306	1637.7	3.11595	65.4305	2.38855	5	-9.40121	-8.00943	-8.00943	-8.04575
-8.05933	-8.06924	-8.0918	-8.20796	-8.08729							

Supplementary material 4. Predicted binding affinities for all scores avaiable in the program AutoDock Vina and experimental log(K_i) (test set). MLF: Multiple linear regression. PLS: Partial least square. SVM: support vector machine. ANN: Artificial neural networks. ANN (6-10-2-1) means an artificial neural networks with 5 explanatory variable + 1 response variable (6), 10 neuron in the first hidden layer, 2 neuros in the second hidden layer, and one output (1).

Ligand	Affinity	Gauss1	Gauss2	Repulsion	Hydrophobic	Hydrogen	Torsions	log(ki)	MLR (5D)	PLS (5LC)	SVM (51SV)	ANN (6-4-9-1)	ANN (6-6-9-1)	ANN (6-9-9-1)	ANN (6-10-6-1)	ANN (6-4-6-1)
000HNA_351 A	-6.95638	-9.29026	75.4481	1231.41	2.78949	54.9978	2.98876	3	-6.10791	-7.25782	-7.25781	-6.96838	-7.01463	-7.05882	-6.93983	-6.99327
000MNX_351 A	-7.31392	-8.89382	86.6666	1243.88	3.63211	44.6608	2.84003	2	-6.09691	-7.57788	-7.57786	-7.26039	-7.37869	-7.42705	-7.31235	-7.34774
000STU_400 A	-5.55638	-0.79591	7.71561	77.7424	1.14046	0	1.83534	3	-8.10791	-4.39103	-4.39102	-5.38156	-5.05019	-4.92371	-5.75011	-5.31352
000UCM_400 A	-8.0006	-10.2696	83.9617	1599.24	3.74934	56.5867	1.86325	3	-7.82391	-8.23255	-8.23254	-8.03802	-8.06504	-8.13419	-8.14327	-8.04681
000IQA_338 A	-7.67358	-9.30706	91.2329	1359.68	4.46965	81.9496	1.73388	2	-6.76955	-7.99898	-7.99897	-7.73551	-7.7582	-7.83865	-7.71266	-7.70697
000K37_501 A	-5.95373	-5.55565	56.4228	927.635	2.70759	38.9011	0	1	-7.1549	-6.6267	-6.62669	-6.43667	-5.97919	-5.96539	-5.97024	-5.94207
000ABO_1271 A	-8.31417	-10.9166	108.035	1634.33	3.70736	50.6608	2.68972	3	-7.58503	-8.31815	-8.31814	-8.21452	-8.33885	-8.37585	-8.43698	-8.34178
000CKI_300 A	-5.825	-5.26232	49.2466	954.598	1.50333	29.0866	0.3711	5	-4.40894	-6.07827	-6.07826	-6.2009	-5.73765	-5.66729	-5.89877	-5.76259
000Y27_416 A	-6.00462	-6.74608	51.9931	976.714	1.55438	25.182	2.82405	4	-6.96508	-6.3188	-6.31879	-6.11587	-5.97809	-5.93573	-6.03333	-5.98568
000K17_1001 A	-5.73789	-5.77318	48.4216	770.389	1.20707	31.1624	0	0	-6.52288	-5.96044	-5.96043	-6.02953	-5.59211	-5.48844	-5.84159	-5.63058
000P45_501 A	-7.58634	-8.55165	105.877	1594.03	1.94744	31.2004	0.6115	7	-8.18709	-7.3641	-7.36408	-7.60077	-7.55042	-7.56433	-7.65634	-7.5735
000P55_501 A	-7.29232	-8.23238	98.2774	1523.38	1.87362	27.5113	0.23574	6	-7.61979	-7.25675	-7.25674	-7.45966	-7.29561	-7.34642	-7.31463	-7.28842
000RFZ_336 A	-5.99981	-6.08558	40.8889	1040.99	1.86997	20.6432	2.31781	5	-4.67847	-6.45035	-6.45034	-6.29831	-5.99092	-5.97659	-6.03535	-5.98959
0003O0_293 B	-7.26896	-7.62226	83.6981	1408.85	2.93107	34.324	2.21811	7	-6.22185	-7.33643	-7.33642	-7.51839	-7.2911	-7.34803	-7.30443	-7.27161
000ABQ_360 A	-8.38728	-9.02641	118.284	1712.68	3.47349	18.2732	3.33024	8	-9.22185	-8.07643	-8.07641	-8.30556	-8.34504	-8.29889	-8.49316	-8.39098
0003RW_1 A	-7.7161	-8.32095	101.282	1560.6	2.36227	16.2923	3.36394	9	-7.45593	-7.39011	-7.3901	-7.62071	-7.80628	-7.68942	-7.93589	-7.79282
000N53_601 A	-8.60915	-9.46654	120.678	1893.27	4.36901	40.1237	1.68856	6	-7.19382	-8.7702	-8.77018	-8.85142	-8.65037	-8.68691	-8.67002	-8.61729
000RAP_201 A	-9.00086	-16.6888	174.718	3067.61	4.95292	134.154	5.75287	13	-8.52288	-11.0133	-11.0132	-8.19619	-9.12329	-9.21253	-8.88367	-8.938
000T77_401 B	-7.72928	-8.0167	103.88	1648.15	2.30462	39.9727	0.98485	9	-9.52288	-7.4482	-7.44819	-7.76159	-7.67809	-7.67585	-7.82405	-7.711

000HKC_300 A	-9.71374	96.0987	1470.66	2.86261	60.0144	2.179	4	-7.88606	-7.65498	-7.65496	-7.67102	
-7.68031	-7.69453	-7.71881	-7.77351	-7.69906								
000H1K_301 A	-6.7293	69.3431	1173.31	1.95946	11.4494	1.74971	4	-6.19997	-6.78213	-6.78212	-6.68315	
-6.40156	-6.44768	-6.48606	-6.36226	-6.42311								
000H4K_300 A	-7.26922	77.2492	1250.33	2.48166	16.9276	1.79039	4	-7.6925	-7.08465	-7.08464	-7.01945	
-6.75452	-6.81662	-6.87715	-6.70767	-6.78063								
0005HK_301 A	-10.0598	103.954	1614.04	3.1772	44.522	3.54864	5	-8.2186	-8.0242	-8.02419	-8.0134	
-8.21081	-8.19288	-8.19328	-8.35317	-8.23322								
0002HV_401 A	-8.54163	82.2869	1371.83	1.1343	42.4805	0	4	-7.03937	-6.87984	-6.87983	-6.93863	
-6.62285	-6.66119	-6.72807	-6.58411	-6.64195								
0002HW_401 A	-7.81306	84.363	1372.22	3.81552	11.996	4	5	-7.79588	-7.73031	-7.7303	-7.56864	
-7.68716	-7.7065	-7.74186	-7.79582	-7.70906								
0002OO_601 A	-8.10514	108.728	1483.02	4.8224	31.555	4.03157	7	-7.22185	-8.08367	-8.08366	-8.21745	
-8.21605	-8.21232	-8.21768	-8.32933	-8.21564								
0002QV_403 A	-8.42826	89.4474	1102.21	4.39684	77.7474	1.76177	1	-6.96658	-7.51959	-7.51957	-7.17115	
-7.00285	-7.11661	-7.19694	-6.88515	-7.03737								
0003FV_501 A	-8.90483	81.77	1533.87	2.12374	29.4141	0.67813	3	-8.39794	-7.5937	-7.59369	-7.52915	
-7.38077	-7.43848	-7.49872	-7.44916	-7.42333								
0004KH_501 A	-7.37817	74.1055	1146.93	2.06842	33.0325	1.19247	3	-5.95861	-6.80292	-6.80291	-6.71874	
-6.3792	-6.43508	-6.47195	-6.32444	-6.40058								
0004TW_501 A	-8.81592	103.888	1462.79	2.80681	38.2926	2.85423	6	-8	-7.50427	-7.50425	-7.6677	
-7.73717	-7.70374	-7.69933	-7.83472	-7.73353								

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br