UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

GABRIELA SILVA DE OLIVEIRA

Estudos de modelagem molecular para a descoberta de potenciais inibidores da enzima PI4KIIIβ de *Plasmodium falciparum*: triagem virtual e métodos de aprendizado de máquina

São Carlos

2022

GABRIELA SILVA DE OLIVEIRA

Estudos de modelagem molecular para a descoberta de potenciais inibidores da enzima PI4KIIIβ de *Plasmodium falciparum*: triagem virtual e métodos de aprendizado de máquina

Trabalho de conclusão de curso apresentado ao Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de Bacharel em Ciências Físicas e Biomoleculares.

Orientador: Prof. Dr. Rafael Victório Carvalho Guido

São Carlos

2022

Oliveira, Gabriela Silva de
Estudos de modelagem molecular para a descoberta de potenciais inibidores da enzima PI4KIIIBeta de *Plasmodium falciparum*: triagem virtual e métodos de aprendizado de máquina / Gabriela Silva de Oliveira; orientador Rafael Victório Carvalho Guido -- São Carlos, 2022.
33 p.

Trabalho de Conclusão de Curso (Bacharel em Ciências Físicas e Biomoleculares) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2022.

1. Malária. 2. PI4K. 3. In sílico. I. Guido, Rafael Victório Carvalho , orient. II. Título.

RESUMO

A malária é uma doença infecciosa causada por protozoários do gênero Plasmodium ssp. e apesar do sucesso no combate dessa doença, em 2020 foram reportados 241 milhões de casos e 627 mil óbitos. É de extrema importância a descoberta de novos antimaláricos que, além de eficientes, tenham mecanismos de ação diferentes daqueles já utilizados para evitar mecanismos de resistência cruzada. Para isso, o objetivo deste trabalho é descobrir potenciais inibidores da enzima fosfatidilinositol-4-quinase do tipo III ß do *Plasmodium falciparum* (PfPI4KIIIß). Essa enzima é um alvo macromolecular atrativo devido a sua importância no desenvolvimento do parasito e representa um novo mecanismo de ação diferente dos antimaláricos usuais. A partir da estrutura tridimensional da PfPI4KIIIß e de ligantes das suas proteínas homólogas humanas, integramos técnicas de docagem molecular e triagem virtual com métodos de aprendizado de máquina. Assim, visamos identificar e classificar potenciais moléculas capazes de interagir com o seu sítio ativo da PfPI4KIIIß. Em vista disso, construímos um modelo da estrutura tridimensional da enzima através do *AlphaFold colab notebook*, recuperamos moléculas com atividade conhecida contra enzimas homólogas da PI4K no banco de dados ChEMBL e moléculas com atividade desconhecida contra esse alvo pelo banco de dados ZINC. Para iniciar o estudo, validamos a estrutura por alinhamento estrutural com a enzima PI4K humana, utilizando a métrica pLDDT e simulação de dinâmica molecular. Apesar da docagem molecular ter sido validada pela redocagem do ligante cristalográfico da enzima hPI4K, a abordagem aplicada a triagem virtual não foi validada. Diante disso, foram treinados modelos de aprendizado de máquina para predizer compostos potentes contra as proteínas homólogas humanas. Esses tiveram uma acurácia de aproximadamente 75%, vislumbrando a possibilidade de realizar a triagem virtal de compostos com atividade desconhecida contra essas proteínas.

Palavras-chave: Malária. PI4K. *In sílico*.

SUMÁRIO

1 INTRODUÇÃO

A malária é uma doença infecciosa causada por protozoários do gênero *Plasmodium ssp*. e transmitida aos humanos pela picada de fêmeas do mosquito *Anopheles spp.* Apesar do sucesso significativo do combate à essa doença nas duas últimas décadas, em 2020, quase metade da população mundial corria o risco de contrair a malária e houve uma estimativa de 241 milhões de casos e 627.000 óbitos devido à malária. (1)

A persistência desses quadros endêmicos tem como principais causas a resistência do mosquito à inseticidas e o surgimento de parasitos resistentes aos antimaláricos mais utilizados, incluindo as terapias recomendadas pela Organização Mundial da Saúde (OMS) baseadas na combinação de derivados de artemisinina (ACTs). (2–5) Portanto, torna-se de extrema importância a descoberta de novos fármacos antimaláricos inovadores que evitem os mecanismos de resistência adquiridos, possuindo modos de ação diferentes daqueles convencionais. (6-7)

Devido ao seu papel central nas vias metabólicas essenciais aos parasitos, as enzimas são alvos biológicos importantes para os estudos de planejamento de fármacos. (8-9) Tendo em vista a resistência à medicamentos, é interessante a busca por alvos ainda pouco explorados do *P. falciparum,* parasito responsável pela forma mais letal da doença (1,5), que direcionem a descoberta de moléculas bioativas com mecanismos de inibição inovadores.

Dentre esses alvos, destaca-se a proteína quinase lipídica fosfatidilinositol-4-OH quinase tipo III beta do *P. falciparum* (*Pf*PI4KIIIβ) que, além de ser uma enzima essencial ao desenvolvimento do parasito em diferentes estágios, pode ser inibida por moléculas pequenas (ligantes), como os derivados imidazopirazinícos. (10-11) Os ortólogos de PI4K fosforilam os lipídios do tipo fosfoinositídeos, que em sua forma difosfatada torna-se um importante regulador de diversas vias de sinalização celular, incluindo o processo de citocinese (12-13) Atualmente, existe um inibidor seletivo da *Pf*PI4K, denominado MMV3900048, que se encontra em estudos clínicos de fase 2. (6)

Nesse contexto, métodos computacionais em bioinformática associados ao planejamento de fármacos são altamente promissores, visto que aceleram, direcionam e contribuem significativamente desde a identificação de *hits* até a otimização de *leads.* (14) Dentre as estratégias de planejamento, destacam-se os métodos de SBDD (do inglês, *Structure-Based Drug Design*), que se baseiam no conhecimento da estrutura tridimensional de alvos biológicos, como proteínas ou ácidos nucleicos, adquirido por técnicas de cristalografia, ressonância magnética nuclear (RMN), criomicroscopia eletrônica (Cryo-EM), modelagem por

homologia ou de aprendizado de máquina. Um dos principais métodos SBDD é a docagem molecular, uma estratégia que realiza a previsão do modo de ligação entre o alvo molecular e seu ligante, que, em conjunto com a triagem virtual de várias moléculas de um banco de dados, permite a seleção de compostos com atividade biológica promissora. (9)

A partir da estrutura tridimensional do alvo, é possível integrar técnicas de docagem molecular e triagem virtual com métodos de aprendizagem de máquina para realizar a predição de propriedades físico-químicas, afinidade de ligação, seletividade e classificação de moléculas. (14–16) No contexto de classificação de moléculas na triagem virtual, a abordagem clássica de aprendizado de máquina consiste em treinar modelos computacionais utilizando-se como entrada um conjunto de complexos proteína-ligante já conhecidos que serão utilizados na classificação de atividade biológica de compostos desconhecidos. (17)

Uma segunda estratégia muito utilizada para a descoberta de novos *hits* é através dos métodos de LBDD (do inglês, *Ligand-Based Drug Design*). Esses, por sua vez, baseiam-se na modelagem e análise computacional de ligantes conhecidos, cujas informações estruturais e atividade biológica podem ser encontrados em bancos de dados públicos. Dessa forma, é possível correlacionar esses dados e realizar uma triagem virtual em larga escala para recuperar compostos promissores. Dentre tais técnicas, as principais são: modelos farmacofóricos, QSAR (do inglês, *Quantitative Structure Activity Relationship*) e cálculos de similaridade baseados em propriedades físico-químicas das moléculas. (18)

Portanto, considerando a necessidade de encontrar novos candidatos a inibidores dos parasitos resistentes aos medicamentos convencionais, esse projeto baseou-se no estudo da enzima *Pf*PI4KIIIβ do *P. falciparum* por meio de técnicas de modelagem por homologia. Além disso, também integramos a estratégia de LBDD com técnicas de aprendizado de máquina para construir modelos capazes de predizer possíveis inibidores dessa proteína a partir de informações de moléculas que conhecidamente interagem com homólogas humanas dessa enzima. Dessa forma, além de analisar e estudar a estrutura de um alvo ainda pouco explorado, propomos a descoberta de novos potenciais inibidores para a *Pf*PI4KIIIβ cujo mecanismo de ação seja inovador e capaz de trazer luz ao problema de resistência atualmente enfrentado com os tratamentos convencionais. Finalmente, será sugerida a validação experimental de potenciais inibidores da enzima *Pf*PI4KIIIβ.

2 MATERIAIS E MÉTODOS

2.1 Visão geral do projeto

Para encontrar potenciais inibidores da enzima *Pf*PI4KIIIβ, foram planejadas duas triagens virtuais em larga escala concomitantes. A Figura 1 (**A**) mostra a primeira abordagem, que se baseou em modelos de aprendizado de máquina, os quais foram treinados a partir de dois tipos de *fingerprints*, *Extended Connectivity Fingerprint* (ECFP) (19) e *fingerprints* de interação. (20) Já a segunda, representada pela Figura 1 (**B**), se baseia apenas no ranqueamento obtido com a docagem molecular entre as moléculas e a enzima PfPI4KIIIβ utilizando-se diferentes funções de pontuação.

A predição da potência dos compostos contra a proteína de interesse vai depender da informação contida nos *fingerprints*, que serão os atributos utilizados pelos métodos de aprendizado de máquina. Esses *fingerprints* são vetores que representam a estrutura química ou propriedades que definem das moléculas em estudo. (17) No caso dos ECFP, temos a informação estrutural e dos grupos químicos da molécula determinada por funções de simetria circular. (17,19) Já os *fingerprints* de interação carregam a informação das interações intermoleculares entre a molécula analisada e os resíduos do sítio de ligação da proteína que ela interage, dependendo da docagem molecular. (21) Dessa forma, temos dois tipos de informações diferentes para realizar o aprendizado de máquina e poder escolher o modelo com melhor capacidade de predição.

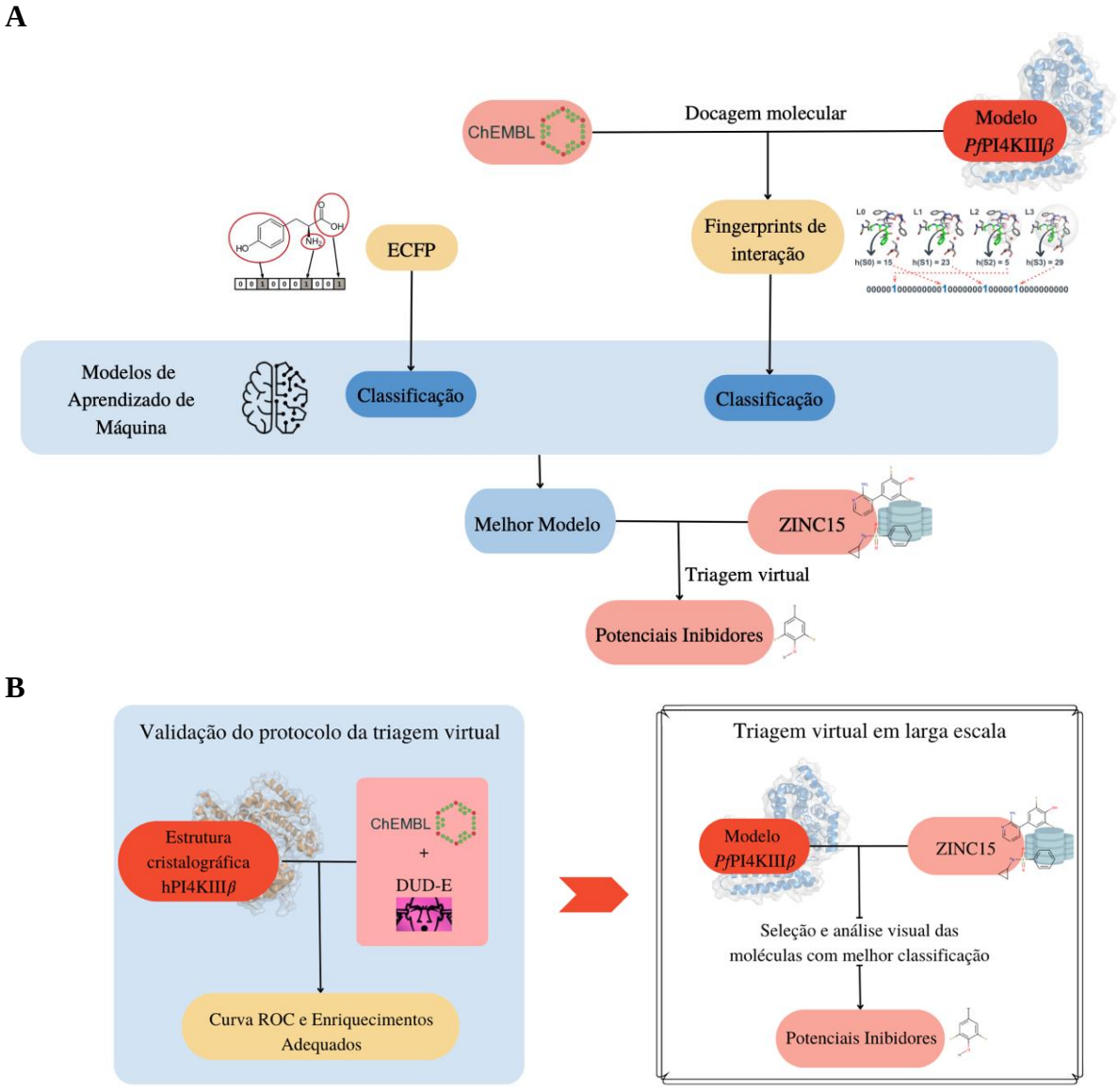


Figura 1 - (**A**) Fluxograma representando a triagem virtual baseada em modelos de aprendizado de máquina para predizer potenciais inibidores da enzima PfPI4KIIIβ. (**B**) Esquema simplificado das etapas necessárias para realizar a triagem virtual em larga com base no ranqueamento obtido através da docagem molecular entre os ligantes e a enzima *Pf*PI4KIIIβ.
Fonte: Elaborada pela autora.

2.2 Obtenção do modelo estrutural da enzima PfPI4KIIIβ

Inicialmente, verificamos a existência de um modelo estrutural da proteína *Pf*PI4KIIIβ disponível nos bancos de dados *SWISS-MODEL* (19) e no *ModBase* (23), obtidos via modelagem comparativa de alta resolução, e no banco de dados *AlphaFoldDB* (24), obtido através de um modelo de aprendizado de máquina baseado em redes neurais. Além disso, como as estruturas encontradas possuem regiões de baixa confiabilidade, também foi predita a estrutura dessa enzima através do *AlphaFold colab notebook* (24) com modificações na

sequência primária para obter um modelo melhor a fim de aplicar as técnicas de modelagem *in silico*. O *notebook* usado disponibiliza um modelo de aprendizado de máquina treinado com uma parte selecionada do banco de dados do *AlphaFold,* porém, possui a precisão similar ao sistema completo do código-fonte *AlphaFold*.

Para a predição da estrutura, utilizou-se a sequência do gene *PF3D7_0509800*, presente no banco de dados *PlasmoDB* (25), os preditores de domínio, famílias e locais funcionais *Pfam* (26) e *PROSITE* (26), a ferramenta de alinhamento local *BLAST* (27) e a ferramenta de alinhamento múltiplo *Clustal*. (28) Através do alinhamento local da proteína expressa pelo gene *PF3D7_059800* e o seu alinhamento múltiplo com as enzimas humanas homólogas à *Pf*PI4KIII*β*, que possuem maior resolução no banco de dados RCSB PDB (29), foram selecionados domínios importantes para a proteína. Além disso, tais domínios também foram escolhidos considerando atribuições sobre a sua importância na atividade catalítica da enzima, resultando em dois domínios principais: n-lobe proximal e domínio catalítico. (30) Outra métrica usada para a construção do modelo estrutural *Pf*PI4KIII*β* foi a comparação das sequências desses domínios com a estrutura encontrada no *AlphaFoldDB,* levando em conta a sua acurácia na predição em cada um desses trechos. Para isso, utilizou-se o índice *pLDDT* (do inglês, *predicted Local-Distance Difference Test*) (24) que indica o quão confiável foi a previsão do modelo gerado. Verificou-se que, entre os dois domínios citados, existe uma região de 437 resíduos (843-1279) com baixo *pLDDT*. No entanto, como estruturalmente esses domínios estão próximos na proteína humana (PDB ID: 4D0L), substituímos essa região por uma alça de 10 glicinas. A sequência final da proteína *Pf*PI4KIII*β* está indicada no Apêndice A.

2.3 Dinâmica molecular para validar a modelagem da estrutura tridimensional

Realizamos 100 ns de simulações de dinâmica molecular usando o pacote de programas AMBER19 (31) para avaliar a estabilidade da estrutura predita pelo *AlphaFold colab notebook*. Inicialmente, todos os hidrogênios foram adicionados com o software *reduce* (32), pertencente ao pacote AmberTools. (31) A proteína foi parametrizada utilizando o campo de força AMBER FF14SB. (33) A estrutura foi envolvida por uma caixa cúbica de água (*e.g.*, TIP3P (34)) com 18 Å de distância entre a proteína e a borda da caixa. Além disso, um íon de sódio foi adicionado ao sistema para mantê-lo eletronicamente estável.

2.4 Recuperação de moléculas com atividade biológica conhecida contra homólogos de pfpi4k e análise do espaço químico

A partir da base de dados ChEMBL, selecionamos compostos com atividade inibitória contra as enzimas ortólogas da *Pf*PI4K. Os filtros utilizados foram: as atividades biológicas foram medidas e reportadas como valores de IC_{50} ou K_i; o operador de atividade biológica foi "="; a atividade biológica foi padronizada como "nM"; o tipo do alvo foi "SINGLE PROTEIN"; não houve comentários em relação à validade da atividade biológica nem sinalizadores de duplicação de dados; não conter as palavras-chaves '*inconclusive'*, '*not determined'*, '*undetermined'*, ou '*approximate value'* no campo de comentários para a atividade biológica ('*activity comment'*). Para determinar a potência de cada molécula, usada como classe no aprendizado de máquina, foram usados os dados presentes no operador atividade.

A análise do espaço químico foi feita a partir da biblioteca Python *ChemPlot*. (35) A partir de um conjunto de moléculas e suas características estruturais, esse pacote permite a visualização bidimensional do espaço químico utilizando três tipos de algoritmos de redução de dimensionalidade: *PCA*, *t-SNE* e *UMAP*. A fim de comparar a dimensão do espaço químico delimitado pelas moléculas recuperadas, também fizemos um banco de moléculas que interagem com proteínas diversas. Para conferir diversidade estrutural às moléculas, escolhemos enzimas de funções diferentes, e, para isso, buscamos no *ChEMBL* alvos que possuam o primeiro digito do código EC (do inglês, *Enzyme Commission Number*) diferentes uns dos outros. Também utilizamos o coeficiente de Tanimoto para calcular a similaridade molecular entre os compostos do banco de dados e agrupamos as moléculas em *clusters* utilizando-se o algoritmo Butina da biblioteca Python RDKit. (36)

2.5 Docagem molecular de inibidores de hPI4KIIIβ

A docagem molecular foi realizada utilizando-se o *DOCK 6* (versão 6.9). Essa é uma extensão do *DOCK 5* com amostragem e pontuação aprimoradas que também utiliza algoritmo de busca sistemática para encontrar a melhor pose do ligante, explorando a flexibilidade do ligante através da sua fragmentação. (37)

Antes de realizarmos a docagem molecular, as estruturas da proteína e dos ligantes foram preparadas removendo os íons, outros cofatores e moléculas de água não estruturais, ou seja, não conservadas em outras estruturas cristalográficas e localizadas fora do sítio de ligação; foram adicionados átomos de hidrogênios com a ferramenta *AddH* do *software Chimera* (38),

considerando o pH fisiológico 7,4; foram atribuídas cargas parciais através do modelos *AMBER FF14SB* (33) e *AMI-BCC* (39) aos resíduos da proteína e aos ligantes, respectivamente.

Em seguida, realizamos a minimização de energia da estrutura utilizando-se *o software Open Babel*. (40) Por fim, geramos esferas dentro do sítio de ligação do receptor com o programa *sphgen* do *DOCK 6*. (41) O sítio de ligação foi definido com base no ligante cristalográfico da estrutura (PDB ID: 4D0L). Além disso, utilizamos o programa *Showbox* do *DOCK 6* para delimitar uma caixa ao redor das esferas com uma margem de 10 Å em todas as direções e utilizamos o programa *GRID* do *DOCK 6* (42) para calcular a energia das interações entre um átomo fictício e o receptor dentro da caixa definida anteriormente. Os ligantes foram tratados como flexíveis de acordo com o protocolo padrão de acoplamento flexível (FLX) descrito em. (37)

O protocolo de docagem foi então avaliado a partir da estratégia de redocagem (*redocking*), que consiste em reposicionar o ligante cristalográfico do complexo receptor-ligante nas coordenadas originais do receptor usando um programa de docagem. Esse experimento é comumente empregado com a finalidade de avaliar a precisão de uma função de pontuação de um programa de docagem em relação às suas capacidades de reprodução de poses conhecidas. Geralmente, as poses preditas são comparadas às suas respectivas poses cristalográficas com base no desvio quadrático médio (RMSD) entre elas.

Para realizar a docagem também foi utilizado o software *GOLD*. (43) Diferente do *DOCK 6*, o *GOLD* (*Genetic Optimization for Ligand Docking*) prediz o modo de ligação dos ligantes com base em algoritmo genético, explorando a flexibilidade do ligante através de populações conformacionais das moléculas analisadas. (44) A validação do protocolo de docagem foi feito através da redocagem do ligante cristalográfico disponível na estrutura do homólogo humano da *Pf*PI4KIIIβ (PDB: 4D0L), cujas poses foram avaliadas com base em todas as funções de pontuação disponíveis no GOLD (*Goldscore*, *Chemscore*, *CHEMPLP* e *ASP*). Além disso, também foi utilizada a função *Chemscore* parametrizada para docagem molecular em quinases.

Neste trabalho, o estudo de redocagem foi realizado utilizando-se a estrutura cristalográfica a enzima hPI4KIIIβ (PDB ID: 4D0L). As poses preditas pelo programa de docagem foram comparadas com as poses cristalográficas utilizando-se o valor de RMSD corrigido por simetria. Tal como em (37), caracterizamos os resultados da docagem tal como se segue: se a pose de maior pontuação estiver até 2 Å de distância (medida em RMSD) da pose cristalográfica, consideramos o resultado válido; no entanto, se a pose correta (próxima à pose cristalográfica) for amostrada, mas não pontuada como a melhor, consideramos que houve uma

falha de pontuação da pose; finalmente, se a pose correta não tiver sido amostrada dentre as 100 conformações geradas durante a docagem, consideramos que houve uma falha de amostragem. A soma desses três casos possíveis será sempre igual a 100%.

2.6 Validação do protocolo para a triagem virtual

Primeiramente realizamos as docagens moleculares entre moléculas com atividade definida contra as enzimas hPI4KIIIβ recuperadas do banco de dados *ChEMBL*. Para tanto, utilizamos a estrutura tridimensional dessa enzima resolvida por cristalografia de raios X (PDB ID: 4D0L). Para cada uma das moléculas recuperadas, também foram geradas 50 *decoys* pelo servidor *DUD-e* (45), que são moléculas com características físico-químicas semelhantes às moléculas de atividade conhecida, porém, que são topologicamente diferentes. Isto é, são moléculas que potencialmente não se ligam ao sítio ativo da enzima hPI4KIIIβ, apesar de suas similaridades físico-químicas às moléculas conhecidamente ativas. (45) Além disso, foram feitas as docagens moleculares entre os *decoys* e a enzima hPI4KIIIβ.

Dessa forma, a partir da função de pontuação das moléculas docadas, foram feitos os gráficos de curva *ROC* (do inglês, *receiver operation characteristic*) e calculou-se a taxa de enriquecimento entre as moléculas recuperadas, consideradas ligantes da enzima hPI4KIIIβ, e os seus *decoys*. Para avaliar se o protocolo de docagem foi capaz de priorizar os ligantes com atividade biológica definida em detrimento dos *decoys*, utilizamos valores maiores ou iguais a 0.8 como ideais para as métricas *AUC* (do inglês, *area under the curve*) e enriquecimento. Essa última métrica calculada tem como objetivo identificar se, dentre as milhares de moléculas analisadas, encontramos aquelas que realmente se ligam ao sítio ativo nas primeiras posições. (46) Isto, pois, ao final de uma triagem virtual, normalmente, selecionam-se as primeiras moléculas para avaliação experimental. Logo, idealmente, espera-se que todas as moléculas ativas sejam ranqueadas nas primeiras posições, o que configuraria um sucesso na triagem virtual. Assim, o enriquecimento é normalmente avaliado nas primeiras frações do conjunto de moléculas (por exemplo, 0.01% ou 1% do total de compostos). Os gráficos de curva *ROC* e enriquecimento foram gerados pelo *Rstudio*, utilizando a linguagem de programação *R*, e através das bibliotecas ROCR e *enrichvs*.

2.7 Aprendizado de máquina

As técnicas de aprendizado de máquina *Random Forest*, *Gradient Tree Boosting* e *SVM* (47) foram utilizadas neste projeto para predizer a atividade de compostos desconhecidos em

relação às proteínas homólogas da *Pf*PI4KIIIβ. O treinamento foi realizado com o método de validação cruzada *K-fold*, em que *K* é o número de subconjuntos nos quais o conjunto total de dados foram subdivididos. Neste trabalho, utilizamos *K* igual a 5. Para todas as três técnicas utilizamos suas implementações disponíveis na biblioteca *scikit-learn* [47] com os parâmetros mantidos como padrão, exceto o número de estimadores/interação que foram definidos como 500. Os modelos foram treinados utilizando como entrada os *fingerprints ECFP4* das moléculas e foram avaliados por meio da métrica acurácia (porcentagem de acertos do modelo).

3 RESULTADOS

3.1 *Modelagem estrutural da enzima Pf*PI4KIII*β*

Por meio de estudos que correlacionam a sequência, estrutura e função proteica dos homólogos da enzima PI4KIII*β* (Figura 2, delimitamos dois domínios principais para a atividade dessa proteína, domínio N-lobe e C-catalítico. Dessa forma, como ambos são necessários para a constituição do sítio de ligação ao ATP, as duas regiões foram escolhidas para compor o modelo.

Figura 2 - Representação dos domínios presentes em PI4KIII*β* de *Plasmodium falciparum* e de humano com base em alinhamento de sequência e estrutura. Domínio helicoidal em verde, N-lobe em vermelho e C-catalítico em amarelo.
Fonte: STERNBERG; ROEPE. (30)

Como o algoritmo do *AlphaFold* utiliza MSA (do inglês, *Multiple Sequence Alignment*) nas primeiras camadas da rede neural para fazer a predição da estrutura terciária através da estrutura primária (24), existe uma tendência a estabelecer uma relação estrutural entre sequências homólogas. Então, como há uma semelhança sequencial de 75 % entre os resíduos dos domínios de interesse da proteína PI4K humana e plasmodial, podemos usar como referência a estrutura cristalográfica humana (PDB ID: 4D0L) e comparar com o modelo gerado para corroborar com a sua validação e iniciar os estudos. (30) Os alinhamentos de sequência e estrutural (RMSD de 1,043 Å) entre os domínios citados são mostrados nas Figuras 4 e 5, respectivamente.

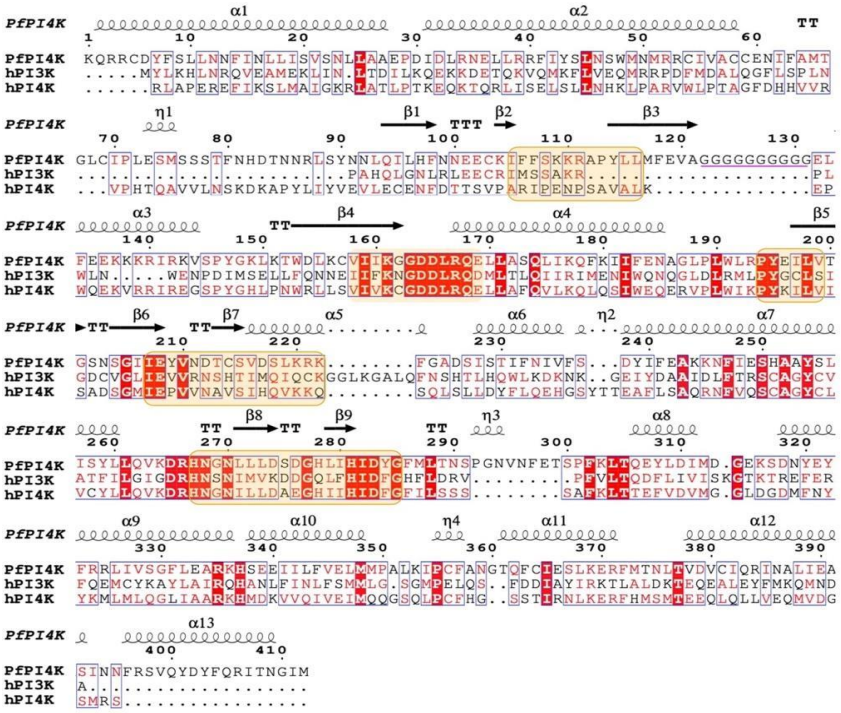
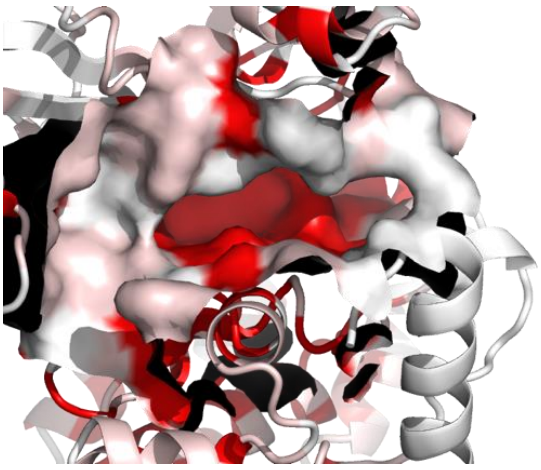


Figura 3 - Alinhamento das sequências primárias dos domínios N-lobe e C-catalítico das proteínas *Pf*PI4K, hPI4K (sequência obtida do PDB 4D0L) e hPI3K (sequência obtida do PDB 7MLK) gerado pelo Clustal (28) e imagem gerada pelo servidor ESPript. (48) A representação da estrutura secundária da *Pf*PI4K é baseada na predição feita pelo *AlphaFold*. As regiões conservadas estão delimitadas pelos retângulos azuis e as regiões idênticas estão coloridas em vermelho. As regiões que correspondem ao sítio ativo estão destacadas por retângulos amarelados.

Fonte: Elaborada pela autora.

A



B

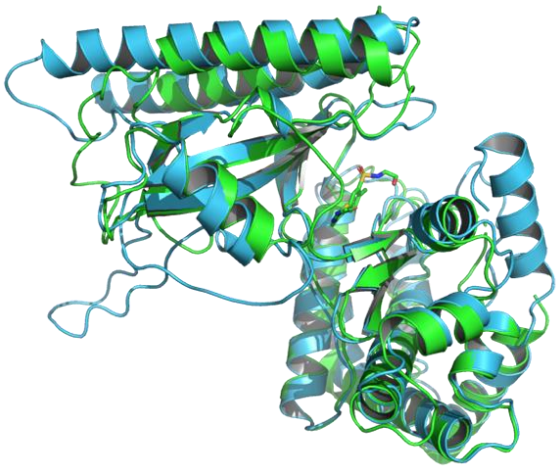


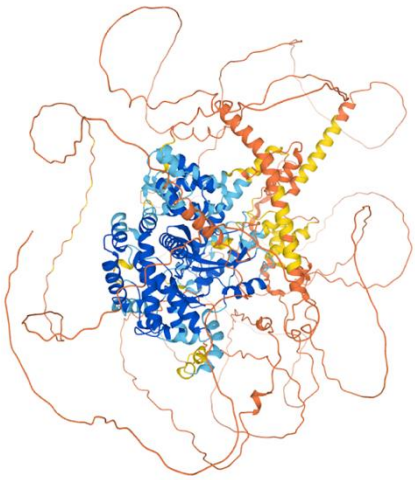
Figura 4 - Comparação do modelo estrutural gerado para os domínios da enzima *Pf*PI4K com seus homólogos humanos. (**A**) Representação da cavidade delimitada pelo sítio de ligação ao ATP do modelo da proteína *Pf*PI4K, cujo código de cores é determinado pela similaridade do alinhamento linear com as homólogas PI4K e PI3K humanas visto na Figura 3, isto é, em vermelho estão destacados os resíduos idênticos, rosa com similaridade entre 0.7 - 1.0 e branco com similaridade menor que 0.7. (**B**) Alinhamento estrutural entre o modelo gerado em azul e os domínios N-lobe e C-catalítico da enzima hPI4K (PDB: 4D0L) (RMSD de 1.043 Å).

Fonte: Elaborada pela autora.

Além disso, como uma das métricas utilizadas para validar o modelo estrutural é o valor pLDDT de cada resíduo predito, também analisamos a confiabilidade das regiões da proteína predita para o gene PF3D7_0509800, encontrada no banco de dados *AlphaFoldDB*. Dessa forma, selecionamos os resíduos próximos dos domínios n-lobe e c-cat que estivessem preditos com alto pLDDT. Para unir os domínios, foi utilizado uma alça formada por 10 resíduos de glicina, por ser um aminoácido flexível que não influenciaria significativamente os campos eletrostáticos ao seu redor. Ao final, foi obtido um modelo para a estrutura da *Pf*PI4KIII*β* com RMSD de 0,78 Å em relação ao modelo encontrado no *AlphaFoldDB* e com altos valores de *pLDDT*, sendo válido para modelagens computacionais como docagem e dinâmica molecular.

Para confirmar que o modelo predito não se encontra em um mínimo local de energia, também foi feita uma análise da estabilidade estrutural do modelo. Na Figura 6 constatamos flutuações significativas da proteína (1 – 4 Å) durante a simulação. Porém, quando analisamos o gráfico de RMSF (do inglês, *root-mean-square fluctuation*), observamos que as regiões de *loop* e alças foram as que apresentaram maior flexibilidade, ao passo que as regiões correspondentes à cavidade de ligação (resíduos 200 a 290) se mostraram mais estáveis. A única exceção, como esperado, foi a região do motivo p-loop presente no sítio ativo (resíduos 103 ao 113) que apresentou maior flexibilidade ao longo da simulação.

A



B

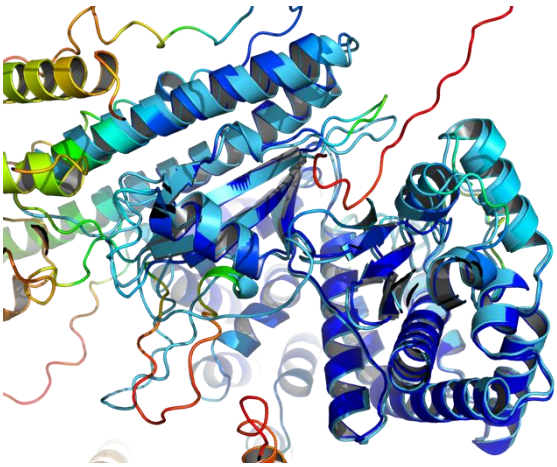


Figura 5 - Modelos da enzima *Pf*PI4KIII*β*. (A) Modelo da *Pf*PI4KIII*β* disponível no *AlphaFoldDB* colorido de acordo com seu *pLDDT* (per-residue confidence score). Azul escuro = pLDDT > 90 (alta confiabilidade no modelo predito); azul claro = 90 > pLDDT < 70 (modelo com confiabilidade reduzida); amarelo = 70 > pLDDT < 50 (baixa confiabilidade); vermelho = pLDDT < 50 (baixíssima confiabilidade). (B) Alinhamento estrutural do modelo disponível no AlphaFoldDB e do modelo predito através da sequência presente no APÊNDICE A.

Fonte: Elaborada pela autora.

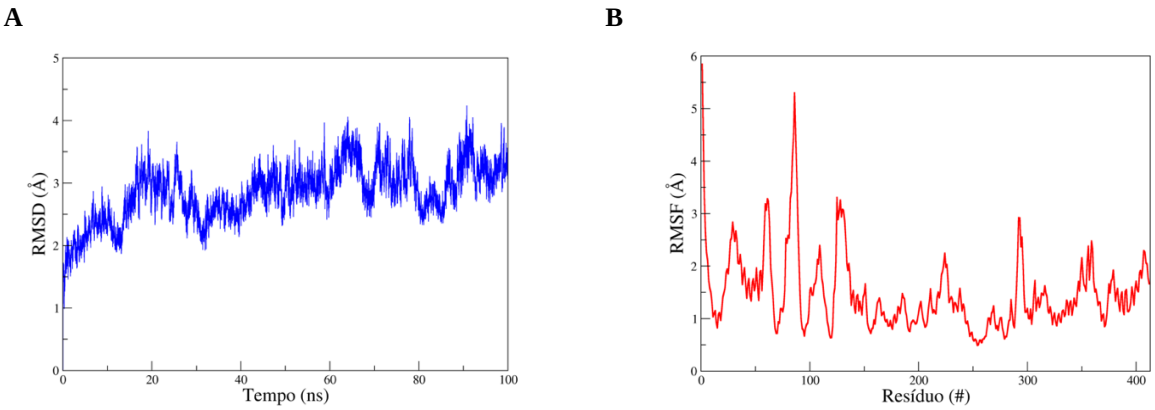


Figura 6 - Resultado da simulação de dinâmica molecular do modelo predito. (A) Gráfico de RMSD. (B) Gráfico de RMSF (flutuação média de cada resíduo durante a simulação).
Fonte: Elaborada pela autora.

3.2 *Recuperação de moléculas com atividade biológica conhecida contra homólogos de* hPI4K *e análise do espaço químico*

A fim de construir um modelo de aprendizado de máquina para predizer inibidores promissores para a enzima *Pf*PI4K buscamos por inibidores conhecidos dessa enzima na base de dados *ChEMBL. (*49) No entanto, na data de consulta, não foi encontrada nenhuma molécula. Assim, dada a similaridade de sequência entre a enzima *Pf*PI4K e suas homólogas humanas, decidimos por assumir que os inibidores conhecidos para as proteínas homólogas humanas da PI4K também seriam válidos para o alvo desejado. No total, foram recuperadas 5.083 moléculas da base de dados *ChEMBL* com atividade biológica (IC_{50} ou K_i) avaliada contra os alvos humanos: PI3K ('CHEMBL1075102', 'CHEMBL3268', 'CHEMBL5554' e 'CHEMBL1163120'), PI4K ('CHEMBL1770034', 'CHEMBL2251', 'CHEMBL3667', 'CHEMBL5667' e 'CHEMBL1795194') e PI5K ('CHEMBL1908383' e 'CHEMBL5969').

Para a construção de modelos de classificação, selecionamos apenas as moléculas cujo parâmetro '*Standard relation*' era igual a '<', '>', '≤' ou '≥' a fim de classificá-las como potentes e pouco potentes com base em suas atividades biológicas (IC_{50} ou K_i). Além disso, filtramos as moléculas cujo parâmetro '*Standard Unit*' foi igual a 'nM' e excluímos as moléculas que possuíam o parâmetro '*Standard relation*' vazio. Após a aplicação destes filtros restaram 1.888 das 5.083 moléculas iniciais.

Considerando que moléculas ativas e inativas próximas ao limiar de 10.000 nM (pActivity = 5) podem ser estruturalmente similares, o que acarretaria um enviesamento dos modelos de aprendizado de máquina, também criamos um banco de dados em que excluímos

as moléculas com atividade considerada intermediária. Para isso, consideramos como potentes, as moléculas com '*Standard value*' ≤ 1.000 nM (pActivity ≥ 6), intermediárias com 1.000 nM < '*Standard value*' ≤ 10.000 nM (5 ≤ pActivity < 6) e não potentes com '*Standard value*' > 10.000 nM (pActivity < 5). Desse modo, obtivemos 386 moléculas potentes, 486 não potentes e 1.016 intermediárias. Logo, obtivemos um conjunto de dados cuja proporção de moléculas potentes e não potentes é 44 % e 56 %, respectivamente.

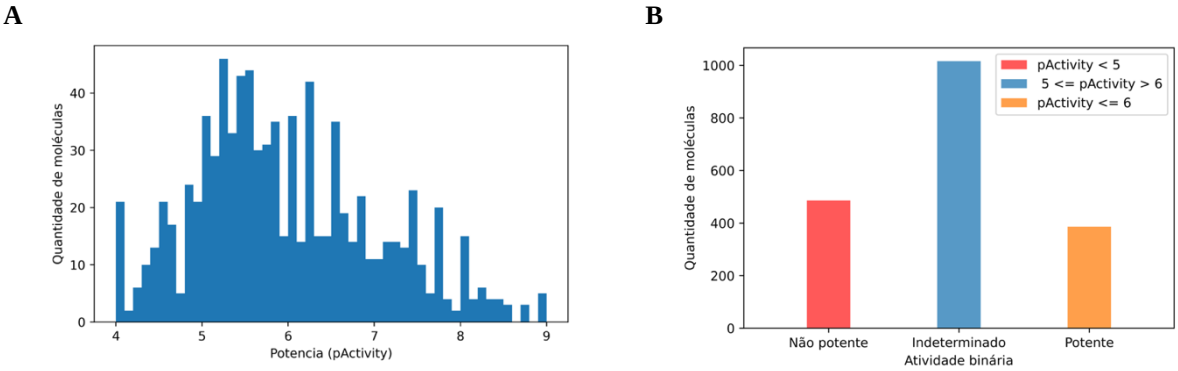


Figura 7 - Análise da atividade do banco de dados de moléculas que interagem com homólogos humanos de PI4K. (**A**) Histograma do -log(IC_{50} ou K_i). (**B**) Histograma da classificação das moléculas em não potentes, intermediárias e potentes.
Fonte: Elaborada pela autora.

Para verificar se essas moléculas são representativas do conjunto de dados, analisamos o espaço químico ocupado pelas moléculas com o auxílio de métodos de redução de dimensionalidade. (35) Além disso, para confirmar se esse espaço delimitado pelas moléculas recuperadas é amplo o suficiente, também analisamos o espaço químico com moléculas estruturalmente diversas que interagem com proteínas diferentes. Assim, a partir da Figura 8, observamos que as moléculas de interesse estão bem distribuídas no espaço químico.

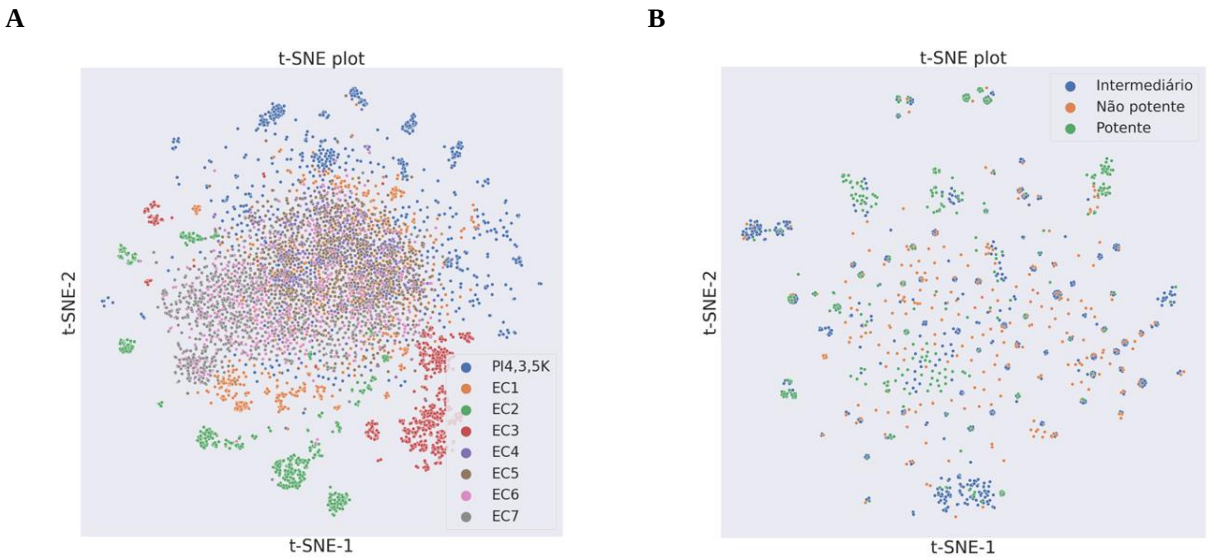


Figura 8 - Representação do espaço químico das moléculas do bando de dados construído. (**A**) Espaço químico delimitado por moléculas com atividade biológica determinada para diferentes enzimas. (**B**) Espaço químico delimitado pelas moléculas do banco de dados que interagem somente com as proteínas homólogas à PI4K.
Fonte: Elaborada pela autora.

Quando analisamos o coeficiente de Tanimoto entre as moléculas (Figura 9), vemos também que as moléculas não são estruturalmente similares nessa métrica, em que valores distantes de 1 representam moléculas diferentes.

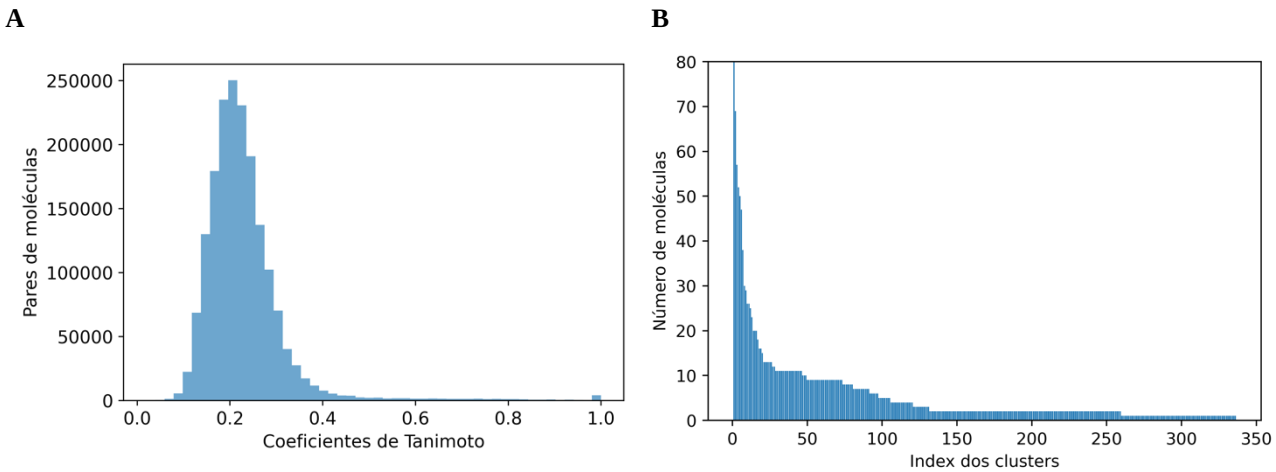


Figura 9 - Representação da similaridade entre as moléculas por dois métodos diferentes. (**A**) O histograma apresenta os coeficientes de Tanimoto de cada par de moléculas no banco de dados, cuja média é 0,23. (**B**) Quantidade de moléculas que compõe cada um dos 336 clusters formados por moléculas que possuem distância mínima de 0,2 através do algoritmo Butina da biblioteca Python RDKit.
Fonte: Elaborada pela autora.

Utilizamos também o algoritmo Butina para agrupar as moléculas de acordo com um limiar que delimita a distância euclidiana mínima entre as moléculas e os representantes de cada

cluster (moléculas). Foram testados vários limiares para o máximo de dissimilaridade entre as moléculas e aquela que melhor separou os clusters foi o limiar de 0,2. Assim, obtivemos 336 clusters: 77 com 1 molécula, 97 com mais de 5 moléculas, 11 com mais de 25 moléculas e 1 com mais de 100 moléculas. Tal diversidade química é crucial para a construção de modelos de aprendizado de máquina com alta capacidade de generalização. Assim, é possível ver que as moléculas presentes no banco de dados são diversas, contribuindo para um melhor aprendizado do modelo de classificação para a atividade.

3.3 Docagem molecular de inibidores de PI4KIII*β e seus decoys*

Antes de realizar a triagem virtual, validamos o protocolo de docagem com poses e ligantes conhecidos. Para isso, foram feitas a re-docagem do ligante cristalográfico e a análise da curva ROC e do enriquecimento feitos a partir de funções de classificação das docagens de ligantes conhecidos da enzima PI4KIII*β* (319 moléculas) e 50 *decoys* para cada um deles. Como algumas moléculas são da mesma classe, tiveram *decoys* gerados que foram repetidos. Assim, foram gerados somente 14729 *decoys*, totalizando uma média de, aproximadamente, 46 *decoys* por ligante.

As redocagens foram feitas utilizando-se os programas DOCK 6 e GOLD, que possuem estratégias diferentes para a identificação da melhor pose. O DOCK 6 utiliza o método de busca sistemática, que fragmenta o ligante e explora os graus de liberdade da molécula no sítio ativo de forma a combinar esses fragmentos até a reconstrução total da molécula. Já o GOLD, utiliza algoritmo genético como método de busca para encontrar de forma estocástica as melhores poses a partir de populações de conformações, mutações e combinações dos graus de liberdade.

Para ambos os algoritmos, encontramos a pose cristalográfica do ligante nas docagens (RMSD < 2 Å). Porém, as funções de pontuação do GOLD (e.g., chemscore kinase, chemPLP, chemscore) tiveram maior sucesso em classificar poses com menor valor de RMSD do que a função do DOCK 6 (*dockscore*). Todas as funções de pontuação utilizadas no GOLD classificaram a pose com o menor RMSD como a melhor pontuação, mas a função chemscore.kinase.params foi escolhida para continuar os estudos para a triagem virtual. Isso porque esta função foi parametrizada especificamente para triagens virtuais de quinases (Figura 10).

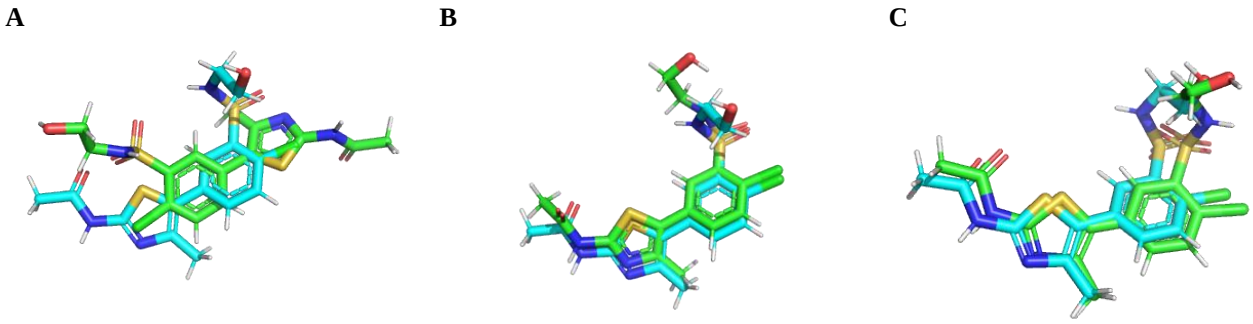


Figura 10 - Redocagem molecular do ligante cristalográfico da proteína 4D0L (N-(5-(4-cloro-3-(2-hidroxi-etilsulfamoil) - feniltiazol-2-il) -acetamida). (**A**) Melhor pose obtida com o DOCK 6 (dockscore de -54 e RMSD de 7.9 Å). (**B**) Segunda melhor pose obtida com o DOCK 6 (dockscore de -48.4 e RMSD de 1.0 Å). (**C**) Melhor pose obtida com o GOLD a partir da função de pontuação chemscore.kinase.params com RMSD de 1.1 Å.
Fonte: Elaborada pela autora.

Para a validação final do protocolo, foram utilizadas 319 moléculas recuperadas do ChEMBL que apresentam atividade biológica contra o alvo CHEMBL3286. Nessa seleção, aplicamos os seguintes filtros: '*Standard relation*' igual a '=' e '*Standard Unit*' igual a 'nM'. Também utilizamos o servidor DUD-E para gerar no máximo 50 *decoys* para cada uma dessas moléculas. Como resultado, obtivemos um total de 14.732 *decoys* únicos. No entanto, a área sob a curva ROC (AUC) foi de 0,47 (Figura 11, ou seja, a função de pontuação teve uma acurácia similar a um algoritmo aleatório (AUC = 0,5). Além disso, o enriquecimento de aproximadamente 12 moléculas (0,5% do total de moléculas docadas – 753 moléculas) também confirma a dificuldade de se encontrar possíveis inibidores da PI4K através desse método. Esse resultado com baixa acurácia e poder de classificação das moléculas pode ser devido a falhas no protocolo ou até mesmo a baixa resolução da estrutura da enzima PDB ID 4D0L, de 2,94 Å. Essa última variável pode estar relacionada com a incerteza da localização de resíduos chaves para a interação de alguns dos ligantes, que adquirem poses desfavoráveis na docagem molecular quando em relação a sua pose real no sítio de ligação. Assim, *decoys* podem ter poses mais favoráveis, obtendo melhor classificação com as funções de pontuação.

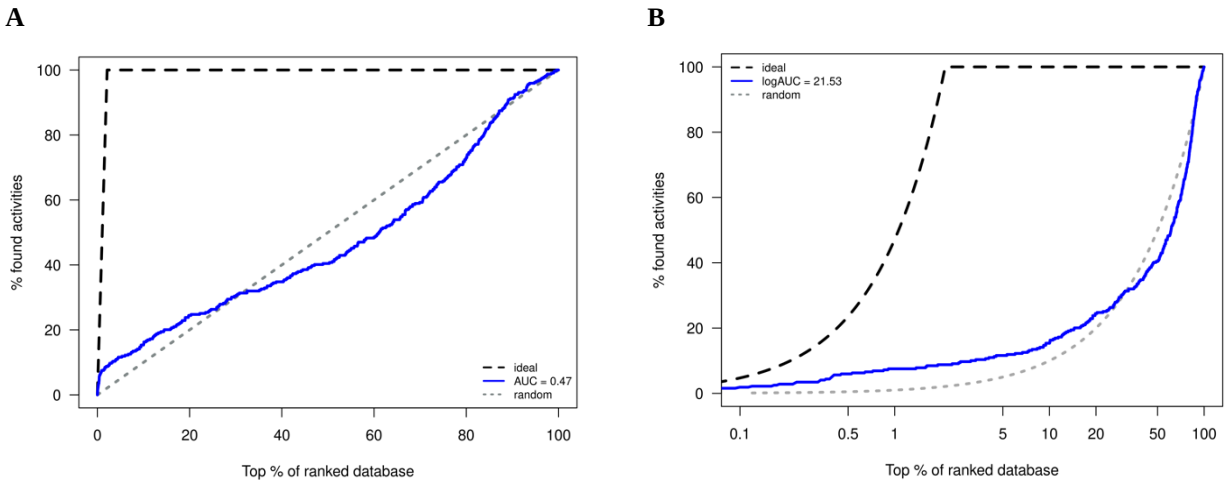


Figura 11 - Resultado da validação para a triagem virtual. (**A**) Curva ROC. (**B**) Curva ROC semilogarítmica.
Fonte: Elaborada pela autora.

3.4 Aprendizado de máquina com fingerprints gerados

Para treinar os modelos de predição de atividade (Activity ou pActivity) das moléculas em relação às proteínas homólogas de hPI4K, foi utilizado o banco de dados da seção 3.2 e os algoritmos *Random Forest*, *Gradient Tree Boosting* (GBoost) e *Support Vector Machines* (SVM) com 4 funções diferentes de kernel. Os resultados de cada algoritmo de classificação são apresentados na Tabela 1. Como métrica de qualidade utilizamos: acurácia, representando o desempenho geral do modelo; especificidade, mostrando a cobertura das amostras negativas; e AUC.

Tabela 1 - Resultados dos algoritmos de aprendizado de máquina para a classificação de compostos ativos e inativos contra a enzima hPI3K. Os modelos foram treinados utilizando-se *fingerprints* ECFP4 com 2.048 bits.

Algoritmo	**Acurácia**	**Sensibilidade**	**AUC**
Random Forest	0,75 ± 0,06	0,6 ± 0,2	0,82 ± 0,09
Gradient Tree Boosting	0,74 ±0,08	0,7 ± 0,2	0,80 ± 0,07
SVM (kernel = rbf)	0,75 ± 0,09	0,7 ± 0,2	0,85 ± 0,07
SVM (kernel = linear)	0,74 ± 0,08	0,7 ± 0,1	0,82 ± 0,07
SVM (kernel = poly)	0,74 ± 0,09	0,7 ± 0,2	0,85 ± 0,04
SVM (kernel = sigmoid)	0,8 ± 0,1	0,7 ± 0,2	0,83 ± 0,08

Fonte: Elaborada pela autora.

Observamos que as moléculas foram bem classificadas pelos três tipos de algoritmos e que os resultados foram equivalentes entre si, não apresentando diferença significativa. Além

disso, a sensibilidade mede a proporção de moléculas preditas como ativas que são de fato ativas contra as proteínas, sendo extremamente importante para a triagem de diversos ligantes com atividade desconhecida. Nesse sentido, tivemos uma predição de aproximadamente 75% dos ligantes.

4 CONCLUSÕES E CONSIDERAÇÕES FINAIS

O estudo de moléculas com mecanismos de ação diferentes dos fármacos com atividade antiplasmodial disponíveis é de extrema importância para a eliminação da malária. Nesse trabalho, procuramos explorar a estrutura de um alvo validado para procurar novos inibidores com estruturas diversas que potencialmente apresente modos de ligação inovadores. Para isso, foi construído um modelo estrutural da proteína *Pf*PI4KIIIβ através do algoritmo AlphaFold v2.1.0. A sequência primária utilizada para fazer a predição foi definida de acordo com a análise dos domínios funcionais e a estrutura resolvida da proteína homóloga humana PI4KIIIβ. Para validar essa estrutura, utilizamos a métrica pLDDT, a comparação por alinhamento estrutural com a proteína hPI4KIIIβ e dinâmica molecular. Todas as estratégias de validação mostraram que, além de estável, a estrutura predita possui uma alta confiabilidade.

Como não possuímos bancos de dados de moléculas inibidoras da PI4K de *Plasmodium spp.*, utilizamos as moléculas recuperadas do ChEMBL que possuem atividade biológica definida contra a enzima hPI4KIIIβ (PDB ID: 4D0L). Apesar de conseguirmos validar a docagem molecular através da reprodução (*redocking*) da pose original do ligante cristalográfico, não foi possível validar a triagem virtual com base nos ligantes conhecidos e *decoys*. Tal erro de validação pode ter ocorrido devido à baixa capacidade das funções de pontuação em priorizar moléculas ativas em detrimento de *decoys* ou ao protocolo de triagem virtual estabelecido. Assim, em trabalhos futuros, planejamos realizar novos experimentos para determinar se o erro de validação ocorreu devido ao protocolo utilizado.

Já os modelos de aprendizado de máquina treinados utilizando-se somente a estrutura das moléculas que interagem com os homólogos da enzima PI4K apresentaram alta acurácia. Considerando que as moléculas utilizadas para o treinamento apresentaram diversidade estrutural, acreditamos que o modelo obtido será útil para a descoberta de novos inibidores a partir de um banco de moléculas cuja atividade contra a enzima PI4K sejam desconhecidas. Portanto, os próximos passos serão realizar uma triagem virtual com os modelos de aprendizado de máquina agrupar as moléculas ativas por similaridade estrutural e selecionar as moléculas representantes de cada cluster que possuam maior diversidade estrutural em relação às moléculas obtidas da base de dados ChEMBL para validação experimental em ensaios contra o parasita padronizados em nosso laboratório. Além disso, pode-se ainda realizar uma análise visual do modo de ligação dos compostos selecionados via docagem molecular e uma análise de estabilidade dos complexos proteína-ligante preditos através de dinâmica molecular. Projetos futuros utilizarão esse modelo estrutural para aplicar diferentes técnicas SBDD para

encontrar novos inibidores. Também pretendemos expressar de forma heteróloga a proteína plasmodial, obter a estrutura experimental e padronizar um ensaio de atividade com a PI4K para possibilitar a triagem *in vitro* de moléculas encontradas computacionalmente.

REFERENCIAS

1 WORLD HEALTH ORGANIZATION. *Monitoring progress on universal health coverage and the health-related sustainable development Goals in the WHO South-East Asia Region*: 2021 update. New Delhi: World Health Organization, 2021. Disponível em: https://apps.who.int/iris/bitstream/handle/10665/344764/9789290228936-eng.pdf?sequence=1&isAllowed=y. Acesso em: 04 jun. 2022.

2 ASHLEY, E. A. *et al*. Spread of artemisinin resistance in Plasmodium falciparum malaria. *New England Journal of Medicine*, v. 371, n. 5, p. 411–423, July 2014. DOI: 10.1056/NEJMoa1314981.

3 COLLINS, K. A. *et al*. Enhancing protective immunity to malaria with a highly immunogenic virus-like particle vaccine. *Science Reports*, v. 7, n. 1, Apr. 2017. DOI: 10.1038/srep46621.

4 GUYANT, P. *et al*. Past and new challenges for malaria control and elimination: the role of operational research for innovation in designing interventions. *Malaria Journal*, v. 14, n. 1, p. 279, July 2015. DOI: 10.1186/s12936-015-0802-4.

5 PHILLIPS, M. A. *et al*. Malaria. *Nature Reviews Disease Primers*, v. 3, n. 1, Aug. 2017. DOI: 10.1038/nrdp.2017.50.

6 SINXADI, P. *et al*. Safety, tolerability, pharmacokinetics, and antimalarial activity of the novel Plasmodium phosphatidylinositol 4-Kinase inhibitor MMV390048 in healthy volunteers. *Antimicrobial Agents and Chemotherapy*, v. 64, n. 4, p. e01896-19. DOI: 10.1128/AAC.01896-19.

7 VERLINDEN, B. K.; LOUW, A.; BIRKHOLTZ, L.-M. Resisting resistance: is there a solution for malaria? *Expert Opinion on Drug Discovery*, v. 11, n. 4, p. 395–406, Apr. 2016. DOI: 10.1517/17460441.2016.1154037.

8 GUIDO, R. V. C.; ANDRICOPULO, A. D.; OLIVA, G. Planejamento de fármacos, biotecnologia e química medicinal: aplicações em doenças infecciosas. *Estudos Avançados*, v. 24, p. 81–98, 2010. DOI: 10.1590/S0103-40142010000300006.

9 GUIDO, R. V. C.; OLIVA, G. Structure-based drug discovery for tropical diseases. *Current Topics in Medicinal Chemistry*, v. 9, n. 9, p. 824–843, 2009. DOI: 10.2174/156802609789207064.

10 AGUIAR, A. C. *et al*. New molecular targets and strategies for antimalarial discovery. Current Medicinal Chemistry, v. 26, n. 23, p. 4380–4402, July 2019. DOI: 10.2174/0929867324666170830103003.

11 MCNAMARA, C. W. *et al*. Targeting Plasmodium PI(4)K to eliminate malaria. *Nature*, v. 504, n. 7479, p. 248-253, Dec. 2013. DOI: 10.1038/nature12782.

12 BURKE, J. E. *et al*. Structures of PI4KIIIβ complexes show simultaneous recruitment of Rab11 and its effectors. *Science*, v. 344, n. 6187, p. 1035–1038, May 2014. DOI: 10.1126/science.1253397.

13 D'ANGELO, G. *et al*. Phosphoinositides in Golgi complex function. *In*: BALLA, T.; WYMANN, M.; YORK, J. D. (ed.). *Phosphoinositides II*: the diverse biological functions. Dordrecht: Springer, 2012, p. 255–270. DOI: 10.1007/978-94-007-3015-1_8.

14 KITCHEN, D. B. *et al*. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, v. 3, n. 11, p. 935–949, Nov. 2004. DOI: 10.1038/nrd1549.

15 LENSELINK, E. B. *et al*. Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *Journal of Cheminformatics*, v. 9, n. 1, p. 45, Dec. 2017. DOI: 10.1186/s13321-017-0232-0.

16 LYNE, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today*, v. 7, n. 20, p. 1047–1055, Oct. 2002. DOI: 10.1016/S1359-6446(02)02483-2.

17 VAMATHEVAN, J. *et al*. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, v. 18, n. 6, June 2019. DOI: 10.1038/s41573-019-0024-5.

18 DRWAL, M. N. GRIFFITH, R. Combination of ligand- and structure-based methods in virtual screening. *Drug Discovery Today*: technologies, v. 10, n. 3, p. e395–e401, Sept. 2013. DOI: 10.1016/j.ddtec.2013.02.002.

19 ROGERS, D.; HAHN, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, v. 50, n. 5, p. 742–754, May 2010. DOI: 10.1021/ci100050t.

20 FASSIO, A. V. *et al*. *Prioritizing virtual screening with interpretable interaction fingerprints*. 2022. Disponível em: https://www.biorxiv.org/content/10.1101/2022.05.25.493419v1.full.pdf. Acesso em: 15 jun. 2022. DOI: 10.1101/2022.05.25.493419.

21 FASSIO, A. V. *et al*. nAPOLI: a graph-based strategy to detect and visualize conserved protein-ligand interactions in large-scale. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, v. 7, n. 4, p. 1317-1328, 2019. DOI: 10.1109/TCBB.2019.2892099.

22 BIENERT, S. *et al*. The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Research*, v. 45, n. D1, p. D313–D319, Jan. 2017. DOI: 10.1093/nar/gkw1132.

23 PIEPER, U. *et al*. ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research*, v. 42, n. D1, p. D336–D346, Jan. 2014. DOI: 10.1093/nar/gkt1144.

24 JUMPER, J. *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature*, v. 596, n. 7873, p. 583–589, Aug. 2021. DOI: 10.1038/s41586-021-03819-2.

25 *PlasmoDB*. Disponível em: https://plasmodb.org/plasmo/app/. Acesso em: 26 fev. 2022.

26 MISTRY, J. *et al*. Pfam: the protein families database in 2021. *Nucleic Acids Research*, v. 49, n. D1, p. D412–D419, Jan. 2021. DOI: 10.1093/nar/gkaa913.

27 BORATYN, G. M. *et al*. BLAST: a more efficient report with usability improvements. *Nucleic Acids Research*, v. 41, n. W1, p. W29–W33, July 2013. DOI: 10.1093/nar/gkt282.

28 SIEVERS, F. *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, v. 7, n. 1, p. 539, Jan. 2011. DOI: 10.1038/msb.2011.75.

29 BERMAN, H. M. *et al*. The protein data bank. *Nucleic Acids Research*, v. 28, n. 1, p. 235–242, Jan. 2000. DOI: 10.1093/nar/28.1.235.

30 STERNBERG, A. R.; ROEPE, P. D. Heterologous expression, purification, and functional analysis of the Plasmodium falciparum phosphatidylinositol 4-Kinase IIIβ. *Biochemistry*, v. 59, n. 27, p. 2494–2506, July 2020. DOI: 10.1021/acs.biochem.0c00259.

31 CASE, D. A. *Amber 2019*. 2019.Disponível em: http://ambermd.org/contributors.html. Acesso em: 26 fev. 2022.

32 WORD, J. M. *et al*. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation11Edited by J. Thornton. *Journal of Molecular Biology*, v. 285, n. 4, p. 1735–1747, Jan. 1999. DOI: 10.1006/jmbi.1998.2401.

33 MAIER, J. A. *et al*. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of Chemical Theory and Computation*, v. 11, n. 8, p. 3696–3713, Aug. 2015. DOI: 10.1021/acs.jctc.5b00255.

34 PRICE, D. J.; BROOKS, C. L. A modified TIP3P water potential for simulation with Ewald summation. *Journal of Chemical Physics*, v. 121, n. 20, p. 10096–10103, Nov. 2004. DOI: 10.1063/1.1808117.

35 SORKUN, M. C. *et al*. ChemPlot, a Python library for chemical space visualization. *Chemistry Methods*, v. 2, n. 7, p. e202200005, 2022. DOI: 10.1002/cmtd.202200005.

36 LANDRUM, G. *et al*. *rdkit/rdkit*: 2022_03_3 (Q1 2022) release. June 2022. DOI: 10.5281/zenodo.6605135.

37 ALLEN, W. J. *et al*. DOCK 6: impact of new features and current docking performance. *Journal of Computational Chemistry*, v. 36, n. 15, p. 1132–1156, 2015. DOI: 10.1002/jcc.23905.

38 PETTERSEN, E. F. *et al*. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, v. 25, n. 13, p. 1605–1612, 2004. DOI: 10.1002/jcc.20084.

39 JAKALIAN, A. *et al*. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. method. *Journal of Computational Chemistry*, v. 21, n. 2, p. 132–146, 2000.

40 O'BOYLE, N. M. *et al*. Open Babel: an open chemical toolbox. *Journal of Cheminformatics*, v. 3, n. 1, p. 33, Oct. 2011, DOI: 10.1186/1758-2946-3-33.

41 DESJARLAIS, R. L. *et al*. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *Journal of Medicinal Chemistry*, v. 31, n. 4, p. 722–729, Apr. 1988. DOI: 10.1021/jm00399a006.

42 MENG, E. C.; SHOICHET, B. K.; KUNTZ, I. D. Automated docking with grid-based energy evaluation. *Journal of Computational Chemistry*, v. 13, n. 4, p. 505–524, 1992. DOI: 10.1002/jcc.540130412.

43 LIEBESCHUETZ, J. W.; COLE, J. C.; KORB, O. Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test. *Journal of Computer-Aided Molecular*, v. 26, n. 6, p. 737–748, June 2012. DOI: 10.1007/s10822-012-9551-4.

44 JONES, G. *et al*. Development and validation of a genetic algorithm for flexible docking11 edited by F. E. Cohen. *Journal of Molecular Biology*, v. 267, n. 3, p. 727–748, Apr. 1997. DOI: 10.1006/jmbi.1996.0897.

45 MYSINGER, M. M. *et al*. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, v. 55, n. 14, p. 6582–6594, July 2012. DOI: 10.1021/jm300687e.

46 BENDER, B. J. *et al*. A practical guide to large-scale docking. *Nature Protocols*, v. 16, n. 10, p. 4799–4832, Oct. 2021. DOI: 10.1038/s41596-021-00597-z.

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

48 ROBERT, X.; GOUET, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Research*, v. 42, n. W1, p. W320–W324, July 2014. DOI: 10.1093/nar/gku316.

49 *ChEMBL Database*. 2021. Disponível em: https://www.ebi.ac.uk/chembl/. Acesso em: 13 mar. 2022.

APÊNDICE A

Sequência usada para construir o modelo.

>PfPI4K_nlobe_catalitico

KQRRCDYFSLLNNFINLLISVSNLLAAEPDIDLRNELLRRFIYSLNSWMNMRRCIVACC
ENIFAMTGLCIPLESMSSSTFNHDTNNRLSYNNLQILHFNNEECKIFFSKKRAPYLLMF
EVAGGGGGGGGGELFEEKKKRIRKVSPYGKLKTWDLKCVIIKGGDDLRQELLASQ
LIKQFKIIFENAGLPLWLRPYEILVTGSNSGIIEYVNDTCSVDSLKRKFGADSISTIFNIV
FSDYIFEAKKNFIESHAAYSLISYLLQVKDRHNGNLLLDSDGHLIHIDYGFMLTNSPGN
VNFETSPFKLTQEYLDIMDGEKSDNYEYFRRLIVSGFLEARKHSEEIILFVELMMPALK
IPCFANGTQFCIESLKERFMTNLTVDVCIQRINALIEASINNFRSVQYDYFQRITNGIM