UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

JAILAN DA SILVA SOUSA

PREDIÇÃO DE PEPTÍDEOS INIBIDORES DO CANAL DE SÓDIO DEPENDENTE DE VOLTAGEM DA *DROSOPHILA SUZUKII* UTILIZANDO APRENDIZADO DE MÁQUINA, DOCKING E DINÂMICA MOLECULAR

BELO HORIZONTE/MG
2025

JAILAN DA SILVA SOUSA

PREDIÇÃO DE PEPTÍDEOS INIBIDORES DO CANAL DE SÓDIO DEPENDENTE DE VOLTAGEM DA *DROSOPHILA SUZUKII* UTILIZANDO APRENDIZADO DE MÁQUINA, DOCKING E DINÂMICA MOLECULAR

Dissertação apresentada ao Programa de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais como requerimento parcial para obtenção do título de Mestra em Bioinformática.

Orientador: Prof. Dr. Bruno Silva Andrade

Co-orientadora: Profª. Drª. Joicymara Santos Xavier

BELO HORIZONTE/MG
2025

043 Sousa, Jailan da Silva.
Predição de peptídeos inibidores do canal de sódio dependente de voltagem da Drosophila suzukii utilizando aprendizado de máquina, docking e dinâmica molecular [manuscrito] / Jailan da Silva Sousa. – 2025.
74 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Bruno Silva Andrade. Co-orientadora: Profª. Drª. Joicymara Santos Xavier.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Inseticidas. 3. Peptídeos. 4. Controle de Pragas. 5. Drosophila. 6. Canais de Sódio Disparados por Voltagem. 7. Aprendizado de Máquina. I. Andrade, Bruno Silva. II. Joicymara Santos Xavier. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS
FOLHA DE APROVAÇÃO

Jailan da Silva Sousa

"Predição de peptídeos inibidores do canal de sódio dependente de voltagem da *Drosophila suzukii* utilizando aprendizado de máquina, docking e dinâmica molecular"

Dissertação aprovada pela banca examinadora constituída pelos Professores:

Prof. Bruno Silva Andrade - Orientador
UESB

Prof. Diego César Batista Mariano
UFMG

Prof. Eugenio Eduardo de Oliveira
UFV

Belo Horizonte, 30 de julho de 2025.

Documento assinado eletronicamente por **Diego César Batista Mariano**, **Usuário Externo**, em 05/11/2025, às 09:54, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Bruno Silva Andrade**, **Usuário Externo**, em 05/11/2025, às 10:45, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Eugenio Eduardo de Oliveira**, **Usuário Externo**, em 05/11/2025, às 12:07, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4709253** e o código CRC **0592B09A**.

Dedico esse trabalho à minha mãe, Elisvânia Nascimento da Silva, minha irmã Tailana Nascimento da Silva, e meu irmão Marcos Paulo. Aos meus orientadores, Dr. Bruno Silva Andrade e Dra. Joicimara Santos Xavier.

AGRADECIMENTOS

Expresso minha profunda gratidão à minha mãe, Elisvânia Nascimento da Silva, à minha irmã, Tailana Nascimento da Silva, e ao meu irmão, Marcos Paulo, pelo apoio incondicional, paciência e carinho que me sustentaram ao longo desta jornada.

Aos meus orientadores, Dr. Bruno Silva Andrade e Dra. Joicimara Santos Xavier, agradeço pela orientação dedicada, pelo rigor acadêmico e pelos valiosos ensinamentos que moldaram este trabalho.

Aos colegas do LGCM - Laboratório de Genética Celular e Molecular e do LBQC - Laboratório de Bioinformática e Química Computacional, minha gratidão pela colaboração, troca de ideias e apoio constante, que tornaram os desafios mais leves.

À Universidade Federal de Minas Gerais (UFMG), às agências de fomento CAPES e CNPq, pelo essencial suporte estrutural e financeiro, e ao programa interunidades de pós graduação em bioinformática, pela oportunidade de realizar esta pesquisa e dedicação.

Aos meus amigos, em especial Anderson Oliveira e Janaíne Aparecida de Paula, pelo apoio e companheirismo nos momentos difíceis, e a todos que, de alguma forma, cruzaram meu caminho e contribuíram para esta conquista, meu sincero obrigado.

“We are all conscious today that we are drowning in a sea of data and starving for knowledge.” (Sydney Brenner, 2001).

RESUMO

Os peptídeos representam alternativas ambientalmente mais seguras e específicas em relação aos inseticidas convencionais, oferecendo vantagens como biodegradabilidade e redução de efeitos fora do alvo. No entanto, sua aplicação prática ainda é limitada pela complexidade e pelo custo associados à síntese química e à triagem em larga escala. Neste estudo, desenvolvemos uma estrutura de aprendizado de máquina (ML) para prever peptídeos inibidores de canais de sódio voltagem-dependentes (VGSCs) de insetos, alvos fundamentais na sinalização neuronal e no controle de pragas. Seis algoritmos de aprendizado de máquina amplamente utilizados foram avaliados sistematicamente, sendo o Support Vector Classifier (SVC) aquele que apresentou o melhor desempenho preditivo. A análise de interpretabilidade do modelo por meio do método SHAP revelou que descritores físico-químicos, especialmente aqueles que descrevem padrões estruturais e a relações entre aminoácidos, foram os mais influentes nas previsões do modelo, em concordância com os determinantes estruturais conhecidos das interações entre toxinas e os VGSCs. Os peptídeos derivados de plantas com melhor classificação previstos pelo modelo foram posteriormente validados por meio do *docking* e simulações de dinâmica molecular com o VGSC de *Drosophila suzukii*, uma das principais pragas de frutos de casca fina, confirmando interações estáveis nas regiões do poro e do domínio sensor de voltagem. Esses peptídeos, classificados como defensinas ricas em cisteína, apresentaram padrões estruturais compatíveis com moduladores conhecidos de canais iônicos. Embora a disponibilidade limitada de peptídeos inibidores de VGSCs de insetos validados experimentalmente restringe a generalização do modelo, a abordagem proposta demonstra o potencial da análise de sequências orientadas por aprendizado de máquina para acelerar a descoberta de peptídeos bioativos. Ao integrar modelagem preditiva e simulações moleculares, este trabalho oferece uma estratégia computacionalmente eficiente e biologicamente relevante para a identificação de novos peptídeos bioativos voltados ao manejo sustentável de pragas.

Palavras-chave: Inseticidas baseados em peptídeos; Manejo de pragas; *Drosophila suzukii*; Canais de sódio dependentes de voltagem; Aprendizado de máquina.

ABSTRACT

Peptides represent environmentally safer and target-specific alternatives to conventional insecticides, offering advantages such as biodegradability and reduced off-target effects. However, their practical application remains limited by the complexity and cost of chemical synthesis and large-scale screening. In this study, we developed a machine learning (ML) framework to predict peptide inhibitors of insect voltage-gated sodium channels (VGSCs), which are key targets in neuronal signaling and pest control. Six well-established ML algorithms were systematically evaluated, with the Support Vector Classifier (SVC) achieving the best predictive performance. Model interpretability analysis using SHAP revealed that physicochemical descriptors, particularly those describing structural relationships and amino acid interactions, were the most influential for model predictions, consistent with the structural determinants of VGSC-toxin interactions. The top-ranked plant-derived peptides predicted by the model were further validated through molecular docking and molecular dynamics simulations with the Drosophila suzukii VGSC, a major pest of soft-skinned fruits, confirming stable interactions at the pore and voltage-sensing domains. These peptides, classified as cysteine-rich defensins, exhibited structural patterns compatible with known ion channel modulators. Although the limited availability of experimentally validated insect VGSC inhibitors constrains model generalization, the proposed approach demonstrates the potential of ML-driven sequence analysis to accelerate the discovery of insecticidal peptide candidates. By integrating predictive modeling with molecular simulations, this work provides a computationally efficient and biologically meaningful strategy for identifying novel bioactive peptides for sustainable pest management.

Keywords: Peptide-based insecticides; Pest management; *Drosophila suzukii*; Voltage-gated sodium channels; Machine learning.

LISTA DE ILUSTRAÇÕES

LISTAS DE TABELAS

LISTA DE ABREVIATURAS

Cryo-em	Microscopia crioeletrônica
DM	Dinâmica Molecular
DI-DIV	Domínio I-IV
DsNav	Canais de Sódio Dependentes de Voltagem da *D.suzukii*
GBMs	*Gradient boosting machines*
GPU	*Graphics Processing Unit*
IG	Portão de Inativação
IFM	Domínio Isoleucina-Fenilalanina-Metionina
LightGBM	*Light Gradient Boosting Machines*
ML	*Machine Learning*
PM	Módulo do Poro
RMN	Ressonância Magnética Nuclear
SVC	*Support Vector Classifier*
SNC	Sistema Nervoso Central
S1-S6	Hélice 1-6
TipE	*Temperature-induced paralytic E*
TEH	*TipE Homologous proteins*
VGSCs	Canais de Sódio Dependentes de Voltagem
VSD	Domínio Sensor de Voltagem
XGBoost	*eXtreme Gradient Boosting*

SUMÁRIO

1 INTRODUÇÃO

1.1 *DROSOPHILA SUZUKII*

1.1.1 Origem e distribuição

A *Drosophila suzukii* (Matsumura, 1931) é uma praga quarentenária de natureza polífaga, nativa do leste da Ásia (Hauser, 2011), que tem causado grandes perdas econômicas em diversas partes do mundo. Detectada pela primeira vez na Europa e na América do Norte em 2008 (Walsh et al., 2011, Rota-Stabelli et al., 2020), na América do Sul em 2013 (Andreazza et al., 2017; Deprá et al., 2014) e no norte da África em 2017 (Hassani et al., 2020) com potencial expansão para a o resto da África e para a Austrália devido às condições climática favoráveis (Boughdad et al., 2021; Kwadha et al., 2021). Descrita por Matsumura em 1931 no Japão, a *D. suzukii* tornou-se invasora na segunda metade do século XX, sendo registrada nos Estados Unidos a partir de 1980 no Havaí e, em 2008 na Califórnia, espalhando-se pelas costas oeste e leste do país até o Canadá (Hauser, 2011). A espécie também foi reportada na Espanha em 2008 e na França em 2009, espalhando-se então para a Europa Central e outros países da costa mediterrânea, como Eslovênia e Croácia (Cini et al., 2014).

1.1.2 Adaptação

O sucesso da invasão da *D. suzukii* pode ser parcialmente explicado pela tolerância a uma ampla gama de condições climáticas, ao hibernar por muitos meses e sobreviver ao transporte entre continentes nas fases de ovo, larva e adulto, dentro de frutas ou contêineres de transporte (Hoffmann et al., 2003; Rossi-Stacconi et al., 2016; Stockton; Brown; Loeb, 2019; Toxopeus et al., 2016). Além disso, esta espécie apresenta alta fecundidade (Emiljanowicz et al. 2014), ampla gama de hospedeiros (Lee et al. 2015, Kenis et al. 2016, Stockton et al. 2019, Thistlewood et al. 2019) e alto potencial de dispersão passiva e ativa. A *D. suzukii* é sazonalmente ativa da primavera no outono, mas persiste durante invernos frios, sobrevivendo principalmente às fêmeas adultas (Shearer et al., 2016). As flutuações populacionais são impulsionadas por uma combinação de fatores bióticos e abióticos, incluindo temperatura, umidade e disponibilidade de nutrientes (Evans; Toews; Sial, 2017; Little; Chapman; Hillier, 2020; Rendon et al., 2019). Ao contrário de outros drosofilídeos, que atacam frutas em

decomposição ou apodrecimento, as fêmeas de *D. suzukii* utilizam um ovipositor serrilhado para depositar ovos em frutas intactas e maduras (Goodhue et al., 2011; Walsh et al., 2011). Os danos decorrem diretamente de feridas na oviposição e da alimentação interna das larvas, e indiretamente de patógenos secundários, tornando as frutas infestadas impróprias para comercialização.

1.1.3 Dispersão e prejuízos econômicos no Brasil

No Brasil, *D. suzukii* foi relatada pela primeira vez nas florestas subtropicais da região Sul (Deprá et al., 2014), onde danificou morangos no município de Vacaria, no estado do Rio Grande do Sul (Andreazza et al., 2016). No estado de São Paulo, na região sudeste, foi encontrado em frutas comercializadas em um centro atacadista de frutas e hortaliças (Vilela; Mori, 2014). Além disso, espécimes de *D. suzukii* foram coletados no Cerrado brasileiro (vegetação savânica), em Brasília, DF (Paula et al., 2014) confirmando que a praga é capaz de se espalhar por longas distâncias (1.400 km) por ano (Calabria et al., 2012). O potencial de disseminação da praga no Brasil é muito significativo, pois mais de 80% das áreas de produção da maioria de seus hospedeiros estão localizadas em áreas com clima altamente favorável (Benito; Lopes-da-Silva; Santos, 2016). A maior parte dessa área está situada na região sudeste, incluindo os estados de São Paulo e Minas Gerais, que têm mais de 50% de sua área classificada como favorável ou altamente favorável com alta probabilidade de perdas econômicas. Dados dos relatórios mais recentes indicam uma perda média estimada de cerca de 30% na produção de morangos (Andreazza et al., 2016), e entre 20% e 30% nas culturas de figo e pêssego, o que representa um impacto econômico de aproximadamente US$ 40,7 milhões. No entanto, devido ao elevado potencial de dispersão da mosca em regiões do centro do país, onde as temperaturas são mais favoráveis (Viana et al., 2023), os prejuízos podem alcançar bilhões de dólares (Benito; Lopes-da-Silva; Santos, 2016).

1.2 CONTROLE QUÍMICO DE PRAGAS E TOXICIDADE PARA O HOME E O MEIO AMBIENTE

A estratégia mais eficaz de manejo de praga baseia-se no uso de compostos inseticidas pertencentes às classes dos piretroides, organofosforados, carbamatos e neonicotinoides (Bavithra et al., 2024; Kirst, 2010; Zhu et al., 2020). Com base em seu mecanismo de ação, os inseticidas mais amplamente utilizados agem em alvos nervosos e musculares como os inibidores da colinesterase (organofosforados e carbamatos) ativadores dos canais de sódio

(piretroides), mimetizadores da acetilcolina (neonicotinóides), e ativadores dos canais de cálcio (diamidas) (Rezende-Teixeira et al., 2022). No entanto, além de contribuírem para a poluição e contaminação do solo e dos recursos hídricos, essas substâncias também afetam organismos não alvo, incluindo polinizadores, comprometendo a sobrevivência desses animais e, consequentemente, a manutenção de espécies vegetais com as quais possuem relações coevolutivas de dependência (Beaumelle et al., 2023), como é o caso das abelhas, que são polinizadores com considerável valor econômico. Neste caso, a exposição ao excesso de pesticidas no ar ou na superfície das plantas durante a coleta de néctar e pólen representa um perigo considerável (Crenna et al., 2020). Os pesticidas usados no manejo de culturas diminuem a biodiversidade do solo e reduzem a população de minhocas, que são altamente suscetíveis a pesticidas que causam imobilidade e rigidez e interrompem diversas atividades fisiológicas (Miglani; Bisht, 2019). A intoxicação de mamíferos por organofosforados, carbamatos ou neonicotinóides pode promove o acúmulo de acetilcolina nas sinapses colinérgicas e a superestimulação dos receptores muscarínicos e nicotínicos (Kovacic, 2003). Isso leva à toxicidade aguda, com sintomas como náuseas, vômitos, fraqueza muscular, convulsões e depressão respiratória. A "síndrome colinérgica" resultante inclui efeitos como broncoconstrição, tremores e distúrbios do SNC. A exposição crônica pode danificar o sistema nervoso central e tem sido associada a doenças neurodegenerativas como Alzheimer, Parkinson e ELA (Costa, 2008; Richardson et al., 2019). Além disso, representam riscos ambientais que vão além da saúde humana, afetando diversas espécies não alvo como o bicho-da-seda *Philosamia ricini* que, por sua vez, é inseto economicamente valioso e até mesmo aves que quando expostas a doses subletais sofreram danos ao DNA e degeneração celular (Costa, 2008; Kalita; Haloi; Devi, 2016; Paracampo et al., 2015; Suliman et al., 2020; Tam; Berg; Van Cong, 2018).

Os piretróides, que são amplamente utilizados no controle de pragas domésticas, levaram ao aumento de resíduos ambientais e representam riscos à saúde devido à capacidade limitada dos humanos de metabolizá-los, devido à ausência de carboxilesterases séricas (Richardson et al., 2019). A exposição dérmica pode causar parestesia, como sensações de formigamento ou picadas (Costa et al., 2008), enquanto a exposição ocupacional tem sido associada a efeitos neurológicos, incluindo comprometimento cognitivo (Richardson et al., 2019). Os piretróides são classificados em dois tipos: Tipo I (síndrome T), que causa tremores, espasmos, coma e morte; e Tipo II (síndrome CS), que induz tremores severos, coreoatetose e convulsões. Os compostos do Tipo I afetam os potenciais de ação nervosa por uma duração mais curta do que os do Tipo II (Costa et al., 2008; Richardson et al., 2019).

Além disso, atinge e causa problemas para os organismos não alvo como abelhas (Decourtye et al., 2004), crustáceos (Gottardi et al., 2017; Hoffmann et al., 2016), peixes (Tu et al., 2016) e sapos (Radovanović et al., 2017). Em mamíferos causou danos cromossômicos e interrupção do ciclo celular em camundongos (Bhunya; Pati, 1988), puberdade precoce em camundongos fêmeas (Ye et al., 2017) e efeitos genotóxicos e citotóxicos em coelhos machos (Vardavas et al., 2016). Nas últimas duas décadas, o Brasil experimentou um aumento acentuado no uso de pesticidas, gerando preocupações globais devido ao papel significativo do país na produção e exportação agrícola (Braga et al., 2020). Somente em 2021, 499 novos pesticidas foram aprovados — 84,5% químicos e apenas 15,5% biológicos. Entre os inseticidas, todos os 12 ingredientes ativos recém-aprovados eram químicos, com piretróides (42%) e neonicotinoides (25%) sendo os mais comuns (Rezende-Teixeira et al., 2022). Apesar do aumento nas aprovações de pesticidas, o Brasil continua a favorecer compostos químicos tradicionais em detrimento de alternativas ecologicamente mais seguras, divergindo das tendências globais de sustentabilidade.

1.2.1. Resistência ao controle químico

Além do risco ao homem e ao meio ambiente, tem sido crescente as populações de pragas resistentes ao pesticidas comerciais comumente mais utilizados resultante do uso repetido e generalizado (Foster; Devine; Devonshire, 2017; Liang et al., 2025; Pu; Wang; Chung, 2020). Essa resistência pode surgir por meio de vários mecanismos, como aumento da desintoxicação, mutações no local-alvo, alterações comportamentais ou redução da penetração do inseticida (Hubbard; Murillo, 2024; Naqqash et al., 2016; Zalucki; Furlong, 2017). O desenvolvimento de resistência não apenas reduz a eficácia das estratégias de controle, mas também leva ao aumento das taxas de aplicação e à dependência de alternativas mais tóxicas ou caras, o que, em última análise, representa desafios ambientais e econômicos (Barathi et al., 2024). Adicionalmente, já foi observada resistência em subpopulações de *Drosophila suzukii* aos inseticidas comerciais mais utilizados (Deans; Hutchison, 2022; Disi; Sial, 2021; Ganjisaffar et al., 2022; Gress; Zalom, 2019; Smirle et al., 2017), com registros de aumentos significativos nas concentrações necessárias dos produtos, isto é, nos valores de concentração letal média (CL_{50}) e, consequentemente, a redução nas taxas máximas de mortalidade.

1.3 MECANISMO DE AÇÃO DOS INSETICIDAS EM CANAIS DE SÓDIO DEPENDENTES DE VOLTAGEM

Um dos principais alvos moleculares para a ação de inseticidas, sejam eles sintéticos ou de origem natural, são os canais de sódio dependentes de voltagem (VGSCs) (Ffrench-Constant et al., 2016). Os VGSCs são proteínas transmembranares responsáveis pela excitabilidade neuronal, mediando a alteração do potencial elétrico por meio da condução de íons de sódio através da membrana plasmática (Kasuya et al., 2019). Esses receptores são altamente conservados ao longo da evolução e desempenham um papel fundamental na iniciação e propagação de sinais elétricos em células excitáveis (Kasimova; Granata; Carnevale, 2016; Liebeskind; Hillis; Zakon, 2011). Sua estrutura e função centrais permaneceram estáveis desde bactérias até humanos, ressaltando seu papel essencial na fisiologia celular (Catterall; Wisedchaisri; Zheng, 2020). Dentre as regiões mais estruturalmente conservadas destacam-se especificamente o filtro de seletividade e os domínios sensores de voltagem (VSDs) que passaram por adaptações evolutivas para atender às demandas funcionais de diferentes organismos e tipos celulares (Zakon, 2012).

1.3.1. Estrutura e conservação dos VGSCs

A estrutura dos VGSCs, no geral, é composta por uma única cadeia polipeptídica que se dobra em quatro domínios repetidos (DI–DIV), cada um contendo seis hélices transmembranares (S1–S6) (Catterall, 2010) (Figura 1). Em cada domínio, as hélices S1–S4 formam o chamado domínio sensor de voltagem, (VSD) sendo a hélice S4 o principal sensor de voltagem. Já as hélices S5 e S6 contribuem para a formação do poro condutor de íons, consistindo em um laço P em forma de grampo que se projeta novamente para dentro da membrana, formando o filtro de seletividade iônica (Catterall; Wisedchaisri; Zheng, 2017, 2020; Clairfeuille et al., 2019). Durante a despolarização da membrana, o deslocamento das hélices S4, carregadas positivamente, para fora da membrana, gera a corrente de ativação que desencadeia a abertura do canal de sódio (Catterall; Wisedchaisri; Zheng, 2020). Uma função específica está associada ao movimento ascendente do segmento S4 no domínio IV (DIV), uma vez que ele está acoplado à comporta de inativação (IG), composta pelo motivo Isoleucina-Fenilalanina-Metionina (IFM). Quando as hélices S4 são deslocadas para fora, essa comporta bloqueia rapidamente a entrada de íons sódio no neurônio, em um processo conhecido como inativação rápida, que se completa em cerca de 1–2 milisegundos (Armstrong, 2006; Liu; Bezanilla, 2024). Esse mecanismo de bloqueio resulta em potenciais

de ação neuronais extremamente curtos, permitindo uma alta frequência na transmissão de sinais. A inibição da inativação rápida ou outras alterações na função dos VGSCs geralmente levam à despolarização prolongada, provocando disfunções cardíacas ou neurológicas graves e, em alguns casos, morte (Agbo et al., 2023).

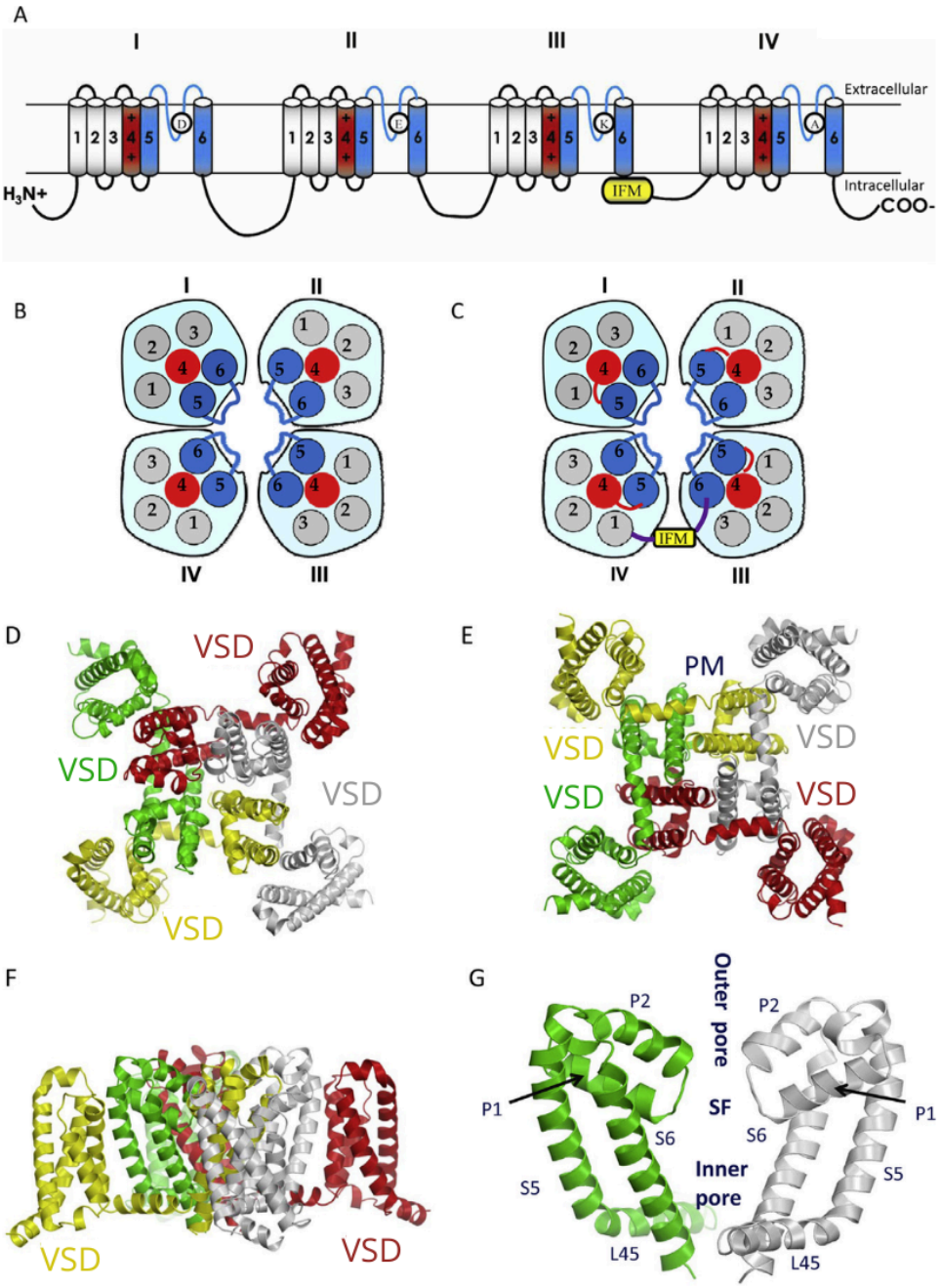

Figura 1. Estrutura do canal de sódio dependente de voltagem. (A) Topologia do canal de sódio, indicando as características da sequência que são críticas para a função do canal. A proteína do canal de sódio contém quatro repetições homólogas (I-IV), cada uma com seis segmentos transmembrana (1e6). A isoleucina no motivo IFM em canais de sódio de mamíferos é substituída por uma metionina em canais de sódio de insetos. (B e C) Representações esquemáticas de vistas extracelulares (B) e intracelulares (C) de canais de sódio. (D e G) Estrutura de raios X do canal de sódio NavAb fechada. Quatro subunidades de NavAb em amarelo, vermelho, verde e cinza, respectivamente, correspondem aos domínios I, II, III e IV em canais de sódio eucarióticos de quatro domínios. (D) Vista extracelular. (E) Vista intracelular indicando os quatro domínios sensores de voltagem (VSD) e o módulo de poro (PM). (F) Vista lateral. (G) Vista lateral expandida do módulo de poros, mostrando apenas duas subunidades para maior clareza. Os poros interno e externo são separados pela região do filtro de seletividade (SF). Fonte: Adaptada de Dong et al., 2014

As subunidades α do canal de sódio dependente de voltagem de insetos e mamíferos compartilham alta homologia de sequência e funções fisiológicas semelhantes (Dong et al., 2014). A principal diferença reside na presença de pelo menos cinco subunidades β distintas em mamíferos, que regulam a função do VGSC interagindo com a subunidade α de diversas maneiras (Brackenbury; Isom, 2011). Embora subunidades β verdadeiras não tenham sido identificadas em insetos, as famílias de proteínas TipE (*Temperature-induced paralytic E*) e TEH (*TipE Homologous proteins*) demonstraram desempenhar papéis regulatórios semelhantes (Bourdin et al., 2015; Wang et al., 2015). Ao contrário dos mamíferos, a maioria dos insetos possui um único gene de VGSC que codifica a subunidade α, que é altamente conservada entre as espécies (Silva; Scott, 2020). No entanto, o splicing alternativo extensivo e a edição de RNA contribuem significativamente para a modificação pós-transcricional, aumentando a diversidade estrutural e funcional dos VGSCs de insetos (Dong et al., 2014; Yuan et al., 2024).

1.4 POTENCIAL USO DE PEPTÍDEOS PARA O DESENVOLVIMENTO DE INSETICIDAS

Como alternativa aos inseticidas sintéticos, os biopesticidas e bioinseticidas, em especial os peptídeos, têm sido propostos como opções mais seguras para o meio ambiente, devido à sua biodegradabilidade, especificidade e potencial farmacológico (Ho et al., 2023; Jin et al., 2021; Sparks et al., 2020). Diversos peptídeos de origem natural apresentam propriedades neurotóxicas direcionadas aos VGSCs, entre os quais se destacam as toxinas alfa, beta/delta e as conotoxinas (Clairfeuille et al., 2019; Klint et al., 2012; Shen et al., 2018). Esses compostos atuam ligando-se, respectivamente, aos sítios 3 no VSD-IV, 4 no VSD-II e ao sítio E no poro dos VGSCs, inibindo sua atividade. As toxinas que se ligam ao sítio E

atuam como bloqueadoras dos canais de sódio, enquanto aquelas que interagem com os sítios 3 e 4 funcionam como moduladores da ativação dos VGSCs, inibindo o processo de inativação rápida (Li; Wu; Yan, 2024; Liu; Bezanilla, 2024). A maioria das toxinas alfa e beta/delta é encontrada nas peçonhas de artrópodes, como aranhas e escorpiões (Daly; Wilson, 2018), enquanto as conotoxinas são produzidas por gastrópodes do gênero Conus (Jin et al., 2019). Essas toxinas, geralmente ricas em resíduos de cisteína, constituem uma fonte valiosa para o estudo e descoberta de novos compostos com potencial aplicação como inseticidas. No entanto, a aplicação prática desses peptídeos no manejo de pragas enfrenta desafios, especialmente devido à sua baixa disponibilidade na natureza e aos processos complexos e custosos envolvidos em sua síntese química em laboratório (Clement et al., 2015; Wu et al., 2021). Nesse contexto, a identificação de novos peptídeos inseticidas representa uma estratégia promissora, capaz de impulsionar a descoberta e a futura comercialização de moléculas bioativas com aplicação no controle de pragas.

1.5 APRENDIZADO DE MÁQUINA

O aprendizado de máquina ou *machine learning* (ML) é um subcampo da inteligência artificial que busca o desenvolvimento de algoritmos que aprendem com dados para melhorar o desempenho em tarefas específicas sem serem explicitamente programados (Jiang; Gradus; Rosellini, 2020). Os métodos de ML são capazes de revelar padrões complexos em grandes conjuntos de dados e/ou de grandes dimensões como relação não lineares, estruturas latentes e/ou interações por meio de abordagens automatizadas, flexíveis e computacionalmente intensivas (Wani, 2025).

1.5.1 Aprendizado supervisionado

Entre as várias técnicas de aprendizado de máquina, o aprendizado supervisionado é um dos mais amplamente adotados e eficazes (Sarker et al., 2020). Ele envolve o treinamento de modelos em conjuntos de dados rotulados, onde cada entrada é pareada com uma saída correspondente, permitindo que o algoritmo aprenda a prever resultados para dados novos e não vistos (Awad; Khanna, 2015; Fabris; Magalhães; Freitas, 2017). O principal objetivo do aprendizado supervisionado é construir modelos que não apenas se ajustem aos dados de treinamento, mas também generalizem bem para observações futuras, fazendo previsões precisas sobre dados com características semelhantes (Barbiero; Squillero; Tonda, 2020).

1.5.2 Classificação binária

Tarefas de aprendizagem supervisionada são normalmente categorizadas em problemas de regressão e classificação para prever resultados contínuos (números com valores reais) e , principalmente, a atribuição de rótulos de classe, respectivamente (Sarker, 2021; Zhang et al., 2022). Dentro da classificação, uma distinção comum é feita entre classificação binária, onde a tarefa é diferenciar entre duas classes possíveis, e classificação multi-classe, que estende isso a mais de duas categorias (Grandini; Bagli; Visani, 2020). Na classificação binária, uma classe é frequentemente designada como a classe "positiva", representando o foco principal do estudo, como detectar e-mails de spam ou identificar a presença de doenças (Bekker; Davis, 2020; Janiesch; Zschech; Heinrich, 2021). Uma consideração importante no aprendizado supervisionado é a complexidade do modelo. Embora modelos mais complexos possam se ajustar melhor aos dados de treinamento, eles correm o risco de ajuste excessivo que, por sua vez, é o processo em que o modelo aprende ruído nos dados em vez de padrões subjacentes reduzindo, assim, sua capacidade de generalização para novos dados (Aftab et al., 2025; Yu et al., 2024). O nível apropriado de complexidade do modelo depende, em parte, da diversidade e do tamanho do conjunto de dados de treinamento com um conjuntos de dados mais ricos permitindo modelos mais sofisticados sem sacrificar a generalização (Althnian et al., 2021). Em tarefas de classificação, modelos de aprendizado de máquina visam aprender os limites de decisão que separam diferentes classes com base em dados de treinamento rotulados (Müller; Guido, 2016). Entre as abordagens mais amplamente utilizadas estão os algoritmos de Árvore de decisão (*Decision Tree*), Floresta Aleatória (*Random Forest*), Máquina de Vetor de Suporte ou *Support vector machine* (SVMs) e *Gradient boosting machines* (GBMs).

Um modelo de Árvore de Decisão classifica os dados dividindo recursivamente o conjunto de dados em subconjuntos com base em limite de valores das variáveis que compõem os dados para otimizar um critério de divisão, como impureza de Gini ou ganho de informação (Fürnkranz, 2011; Suthaharan, 2016). Em cada nó, o algoritmo seleciona a característica e o limite correspondente que melhor separam os dados em grupos mais homogêneos, resultando em uma estrutura de árvore onde cada nó (folha) representa um rótulo de classe. Essa estrutura é fácil de interpretar e pode capturar limites de decisão não lineares, mas é propensa a overfitting. Uma Floresta Aleatória é um método de aprendizado de conjunto que constrói uma coleção de árvores de decisão, cada uma treinada em uma

amostra (bootstrap) diferente do conjunto de dados original (amostragem com substituição) (Shaik; Srinivasan, 2019). Durante o treinamento de cada árvore, um subconjunto aleatório de variáveis é selecionado em cada divisão, o que introduz variabilidade adicional e ajuda a reduzir a correlação entre as árvores (Resende; Drummond, 2019). Esse processo melhora a generalização e reduz o sobreajuste em comparação com uma única árvore de decisão. A classificação final é feita agregando as previsões de todas as árvores individuais por meio de votação majoritária.

Os SVMs abordam a classificação encontrando o hiperplano ótimo que melhor separa pontos de dados de diferentes classes, maximizando a margem entre eles (Lorena; De Carvalho, 2007). Nos casos em que os dados não são linearmente separáveis, funções kernel são usadas para mapear os dados em espaços de dimensões superiores (hiperplanos) onde uma separação linear é possível. Modelos de GMB, incluídos os mais famosos como XGBoost (*eXtreme Gradient Boosting*) e LightGBM (*Light Gradient Boosting Machines*), também dependem de árvores de decisão mas diferem das Florestas Aleatórias em sua estratégia de aprendizado sequencial (Bentéjac; Csörgő; Martínez-Muñoz, 2021; Natekin; Knoll, 2013). As árvores são adicionadas uma de cada vez, e cada nova árvore é treinada para minimizar os erros (resíduos) cometidos pelo conjunto anterior usando gradiente descendente em uma função de perda definida. Isso leva a modelos altamente precisos que podem capturar padrões complexos, embora possam exigir ajustes cuidadosos para evitar sobreajuste.

1.5.3. Aplicações em descoberta e planejamento de drogas

Modelos de aprendizado de máquina são aplicados em uma ampla gama de domínios e usados em vários campos, como bioinformática (Shastry; Sanjay, 2020), medicina (Naik et al., 2024), finanças (Kanaparthi, 2024) e processamento imagens (Sungheetha, 2021) e sons (Jang et al., 2024). Na descoberta de medicamentos, acelerando a identificação e o desenvolvimento de novos compostos terapêuticos por meio da análise de vastos conjuntos de dados, incluindo estruturas químicas, perfis de atividade biológica, eficácia, toxicidade e farmacocinética de potenciais candidatos a medicamentos (Carracedo-Reboredo et al., 2021; Ugurlu, 2024). O uso de modelos de aprendizado de máquina aplicados a dados disponíveis em bases públicas tem se mostrado uma ferramenta valiosa para a triagem em larga escala na descoberta de novos peptídeos com atividade inseticida (Lee et al., 2021; Nambiar; Mitra; Dutta, 2023), antimicrobiana (Wang; Vaisman; Van Hoek, 2022), antibiofilme (Sharma et al., 2016), hemolítica (Plisson; Ramírez-Sánchez; Martínez-Hernández, 2020) e antiviral (Lin et al.,

2022; Xu et al., 2023). Nesse contexto, a identificação de peptídeos inseticidas auxiliada por técnicas de aprendizado de máquina configura-se como uma abordagem promissora para impulsionar a descoberta e eventual comercialização desses compostos bioativos. Tais modelos computacionais podem ser treinados para reconhecer padrões em bancos de dados existentes de peptídeos, permitindo, posteriormente, a predição da presença ou ausência de determinada atividade biológica em novas sequências (Nambiar; Mitra; Dutta, 2023).

1.6 BIOINFORMÁTICA ESTRUTURAL

Bioinformática é uma área de pesquisa que visa extrair conhecimento de dados biológicos, mais especificamente biomoléculas, por meio de modelos e algoritmos derivados da Ciência da Computação (Can, 2014). Envolve a coleta, o armazenamento, a recuperação, a manipulação e a modelagem de dados para fins de análise, visualização ou predição, utilizando algoritmos e estratégias computacionais. Além disso, incorpora conhecimentos de Física, Química, Estatística e Matemática para resolver problemas biológicos, contribuindo assim para o avanço de todas as disciplinas envolvidas (Bilotta; Cong, 2019; Luscombe; Greenbaum; Gerstein, 2001). Vários tipos de dados biológicos como sequências de nucleotídeos, expressão gênica, sequências de proteínas e estruturas de proteínas estão sendo gerados rapidamente, dando origem a dados ômicos. Esses conjuntos de dados exigem integração, organização e o desenvolvimento de estratégias computacionais confiáveis e precisas para aprimorar nossa compreensão da relação entre estrutura e função biomoleculares (Medema, 2021). Nesse contexto, surge a bioinformática estrutural que compreende organização de dados, algoritmos e ferramentas destinados a investigar, analisar, prever e interpretar estruturas biomacromoleculares (Cazals; Dreyfus, 2017). Um importante ponto de virada neste campo foi o rápido aumento na resolução de estruturas proteicas, principalmente por meio de cristalografia de raios X (Smyth, 2000) e, mais recentemente, por meio de espectroscopia de ressonância magnética nuclear (RMN) (Wüthrich, 1990), e microscopia crioeletrônica (Cryo-EM) (Yip et al., 2020). Esses dados estruturais tornaram-se publicamente disponíveis em bancos de dados, sendo o mais conhecido o Protein Data Bank, que atualmente possui 238.346 estruturas publicadas (data de acesso: 30/06/2025) (Burley et al., 2025). A bioinformática estrutural tem dois objetivos principais: o desenvolvimento de métodos de uso geral para manipular informações sobre macromoléculas biológicas e a aplicação desses métodos para resolver problemas biológicos e gerar novos conhecimentos (Bourne; Weissig, 2003).

1.6.1. Predição de Estruturas

Apesar do grande número de estruturas proteicas disponíveis em bancos de dados públicos, ainda existe uma lacuna significativa entre o número de sequências proteicas conhecidas e as estruturas resolvidas experimentalmente, destacando a necessidade de métodos precisos de predição de estruturas (Paiva et al., 2022). No entanto, predizer a estrutura tridimensional de uma proteína a partir de sua sequência de aminoácidos continua sendo um grande desafio não resolvido em Bioinformática. Para lidar com isso, vários métodos foram desenvolvidos, sendo a modelagem baseada em semelhança (modelagem comparativa) a abordagem mais amplamente utilizada antes do surgimento de técnicas baseadas em aprendizado de máquina, como o AlphaFold2 (Mufassirin; Newton; Sattar, 2023), considerado um método independente ou livre de templates que se baseia exclusivamente em informações de sequência de aminoácidos.O AlphaFold2, sendo o método mais utilizado, utiliza informações evolutivas, derivadas de relações coevolutivas, presentes nas sequências de aminoácidos das proteínas, capturadas por meio do multi-alinhamento de sequências, para definir restrições espaciais e construir modelos tridimensionais de proteínas (Jumper et al., 2021). Já a modelagem baseada em templates prevê a estrutura tridimensional de uma proteína utilizando estruturas já resolvidas experimentalmente que compartilham pelo menos 30% de identidade de sequência com a proteína-alvo (Arnold et al., 2006; Webb; Sali, 2017), ou empregando a predição de estrutura secundária de fragmentos das sequências seguida por uma busca por templates quando a identidade de sequência for inferior a 30% (Zhang, 2008). Normalmente, o processo envolve a seleção de um template estrutural apropriado, o alinhamento da sequência-alvo com um molde (t*emplate*) e a construção do modelo tridimensional com base nesse alinhamento. Outra abordagem também conhecida como livre de *template* são os métodos *ab initio*, que visam prever a estrutura da proteína utilizando apenas princípios físico-químicos empíricos ou teóricos, simulando o dobramento em nível atômico por meio da aplicação de forças físicas (Leman et al., 2020; Xu; Zhang, 2012). Entretanto, esses métodos são geralmente limitados pelo comprimento da proteína e pelos requisitos de recursos computacionais.

1.6.2. Acoplamento (*Docking*) molecular

A elucidação das interações entre proteínas e ligantes, ou entre as próprias proteínas, é altamente relevante em diversos campos científicos, incluindo a descoberta de fármacos e a compreensão dos processos de reconhecimento molecular (Sousa et al., 2013; Weng et al., 2020). O *docking* envolve o cálculo da orientação mais favorável que uma molécula pode adotar para formar um complexo estável com seu receptor. Esse processo se baseia em dois componentes fundamentais: o algoritmo de amostragem e a função de pontuação (Wang et al., 2016). O algoritmo de amostragem é responsável por gerar múltiplas posições de ligantes em todos os possíveis espaços de ligação do receptor. Esses algoritmos são tipicamente classificados em abordagens sistemáticas e estocásticas que, por sua vez, são estratégias heurísticas projetadas para reduzir o vasto e quase intratável espaço conformacional de busca (Agu et al., 2023). Algoritmos sistemáticos visam reconstruir o ligante passo a passo no sítio de ligação do receptor, estreitando progressivamente as possibilidades conformacionais (Ferreira et al., 2015). Em contraste, métodos estocásticos, como algoritmos genéticos ou simulações de Monte Carlo, realizam buscas de refinamento gerando um grande número de conformações, selecionando as de maior pontuação e iterando esse processo até que uma posição adequada seja identificada (Leonhart et al., 2019). Esses métodos podem não cobrir todo o espaço conformacional e são comumente empregados em docking local, quando o sítio de ligação é bem definido ou após uma fase de docking cego (Wang et al., 2016; Weng et al., 2020). Em simulações de Monte Carlo, conformações de ligantes são geradas aleatoriamente e aceitas com base no critério de Metropolis, permitindo a exploração eficiente de extensos espaços conformacionais (Zhou et al., 2023). A geração de posições de acoplamento (*poses*) é significativamente influenciada pelo espaço de simulação e pela flexibilidade molecular.

Avanços na tecnologia de GPU (*Graphics Processing Unit*) permitiram procedimentos de docking mais rápidos e flexíveis, melhorando a eficiência da previsão de poses de interação, especialmente com moléculas altamente flexíveis (Sousa et al., 2013).

A função de pontuação é o segundo componente fundamental do docking, ela visa classificar as conformações geradas pelo algoritmo de amostragem e estimar a afinidade de ligação entre o ligante e o receptor (Das et al., 2020). Essas funções são equações matemáticas projetadas para aproximar a estabilidade termodinâmica do complexo molecular e classificar as poses adequadamente. A pontuação continua sendo um grande gargalo no docking, visto que prever com precisão a afinidade de ligação de diversos grupos moleculares ainda é um desafio (Wang et al., 2016). Funções de pontuação são tipicamente categorizadas como empíricas, baseadas em campo de força ou baseadas em conhecimento (teóricas). Funções empíricas estimam a afinidade de ligação somando termos de interação específicos

(por exemplo, ligações de hidrogênio) cujos pesos são derivados de dados experimentais (Murray; Auton; Eldridge, 1998). Funções baseadas em campo de força calcula a afinidade usando parâmetros físicos, incorporando contribuições de ângulos de ligação, torções, forças de van der Waals, ligações de hidrogênio e interações eletrostáticas (Englebienne; Moitessier, 2009). Funções baseadas em conhecimento derivam potenciais estatísticos de complexos proteína-ligante conhecidos, atribuindo energias de interação com base na frequência de contatos de átomos ou grupos em bancos de dados estruturais (Huang; Zou, 2006). As distinções entre ferramentas de docking estão frequentemente relacionadas à complexidade estrutural dos sistemas envolvidos. A flexibilidade dos componentes do sistema é um fator-chave com os ligantes sendo moléculas pequenas, tipicamente apresentam maiores graus de liberdade translacional, rotacional e torcional do que proteínas, que são limitadas por sua estrutura terciária (Paiva et al., 2022). Consequentemente, funções de pontuação e protocolos de docking são adaptados para acomodar essas diferenças estruturais e otimizar a precisão da predição para cada cenário específico de docking.

1.6.3. Dinâmica Molecular

As simulações de dinâmica molecular (DM) visam prever os movimentos de cada átomo em uma proteína ou outro sistema molecular ao longo do tempo, com base em modelos físicos que descrevem interações interatômicas (Karplus; McCammon, 2002). Essas simulações capturam uma ampla gama de processos biomoleculares importantes, como mudanças conformacionais, interação de ligantes e enovelamento de proteínas, fornecendo dados com resolução atômica em escalas de tempo de femtossegundos (Hollingsworth; Dror, 2018). É importante ressaltar que as simulações de DM também podem prever como as biomoléculas respondem a perturbações incluindo mutações, fosforilação, protonação ou ligação/remoção de ligantes em nível atômico. As forças nas simulações de DM são calculadas usando campos de força da clássicos baseados na mecânica molecular ou derivados de cálculos da mecânica quântica e frequentemente refinados usando dados experimentais (Patodia, 2014). Um campo de força típico inclui termos para interações não ligadas como interações eletrostáticas (Coulomb) e de van der Waals, enquanto representam as ligações covalentes a partir de um potencial elástico semelhante a molas. Ainda, um campo de força contém informações detalhadas sobre o sistema em simulação, incluindo tipos de átomos, ângulos, ligações e diedros próprios e impróprios. Os campos de força mais utilizados em pesquisas acadêmicas incluem CHARMM (Guvench; MacKerell, 2008), AMBER (Tian et al.,

2020), GROMOS (Van Gunsteren; Daura; Mark, 1998) e OPLS-AA (Jorgensen; Maxwell; Tirado-Rives, 1996). Estes são implementados nos principais softwares de MD, com exceção do GROMOS, que foi desenvolvido especificamente para o GROMACS (Abraham et al., 2015). Embora esses campos de força tenham sido inicialmente projetados para proteínas, eles foram posteriormente estendidos para descrever ácidos nucleicos, lipídios, carboidratos e pequenas moléculas. As principais diferenças entre eles envolvem como eles tratam interações não ligadas, definem tipos de átomos e estimam diedros (particularmente importante para proteínas). Portanto, esses aspectos devem orientar a seleção de um campo de força para uma dada simulação (Paiva et al., 2022). Além disso, um limite de distância é frequentemente aplicado para limitar o intervalo de interações entre átomos, reduzindo o custo computacional (Goel et al., 2015).

Existem três tipos principais de simulações de MD: MD convencional (Case et al., 2005), amplamente utilizada para estudos como enovelamento de proteínas, estabilidade de complexo receptor-ligante e refinamento de estruturas; simulações QM/MM (mecânica quântica/mecânica molecular) (Kulkarni; Shah; Vyas, 2022), que permitem a modelagem de reações químicas; e metadinâmica (Barducci; Bonomi; Parrinello, 2011), que permite a estimativa de valores de energia livre. Em todos os casos, algoritmos são empregados para integrar as equações de movimento e energia, produzindo vetores para velocidade, posição e forças atômicas (Vlachakis et al., 2014). Esses vetores são atualizados em pequenos intervalos de tempo ao longo da simulação, permitindo o cálculo de propriedades físicas em cada etapa. Antes de executar qualquer simulação de MD, é necessária uma fase de preparação, incluindo minimização de energia e equilibração, para garantir a qualidade estrutural e a estabilidade do sistema (Lemkul, 2024). Um fator crítico para a precisão da simulação é a definição adequada das equações de movimento e energia, que depende diretamente do campo de força escolhido. Diversos campos de força foram desenvolvidos com diferentes níveis de resolução, desde representações de todos os átomos até modelos de átomo único que tratam grupos de átomos como uma única unidade (Dauber-Osguthorpe; Hagler, 2019; Riniker, 2018). A escolha do campo de força deve ser guiada pelas propriedades específicas a serem analisadas (por exemplo, interações atômicas e intercadeias), o nível de detalhe da interação necessário e a classe biomolecular em estudo. A seleção inadequada do campo de força pode comprometer a precisão e levar a erros que são percebidos, no geral, após o término da simulação resultando em gasto de tempo e recursos computacionais (Paiva et al., 2022).

2. JUSTIFICATIVA

A *Drosophila suzukii*, conhecida como mosca da asa manchada, tem se destacado como uma praga agrícola de grande relevância devido à sua capacidade de infestar frutas sadias ainda no campo, provocando consideráveis perdas econômicas em diversas culturas ao redor do mundo. No contexto brasileiro, sua presença representa uma ameaça potencial à fruticultura, setor de grande importância econômica e social. O controle dessa praga tem se baseado, majoritariamente, no uso de inseticidas convencionais, que, além de apresentarem alta toxicidade para o meio ambiente, também têm contribuído para o desenvolvimento de resistência por parte das populações de insetos, tornando o manejo menos eficaz ao longo do tempo.

Diante desse cenário, a busca por alternativas sustentáveis tem ganhado destaque, sendo os peptídeos bioativos uma promissora estratégia biotecnológica. Esses compostos apresentam potencial para atuar como inseticidas mais específicos e menos agressivos ao ambiente, oferecendo uma abordagem mais segura para o controle de pragas. Aliado a isso, o uso de técnicas de Inteligência Artificial tem se mostrado uma ferramenta poderosa na identificação rápida e precisa de novos peptídeos com atividade inseticida, acelerando o desenvolvimento e descobertas de drogas voltadas a terapias antivirais, antimicrobiana, anti-hemolítica, antibiofilm e ao manejo racional de pragas.

3. OBJETIVOS

3.1 Geral

Treinar um modelo supervisionado de aprendizado de máquina para predição de peptídeos bioativos contra o canal de sódio dependente de voltagem da *Drosophila suzukii*, tendo como ponto de partida peptídeos com atividade descrita na literatura.

3.2 Específicos

- Estruturar uma base de dados de peptídeos naturais que apresentem atividade em bancada para canais de sódio dependentes de voltagem de insetos;
- Treinar um modelo de aprendizado de máquina supervisionado a partir da base de dados dos peptídeos ativos em canais de sódio de insetos.
- Selecionar novos peptídeos a partir do modelo de aprendizado de máquina treinado.
- Modelar a estrutura 3D dos peptídeos selecionado e do canal de sódio dependente de voltagem da de *D. suzukii*
- Realizar a validação dos peptídeos selecionados a partir do docking e dinâmica molecular contra o canal de sódio dependente de voltagem da *D. suzukii;*

CAPÍTULO I

Journal of Molecular Graphics and Modelling

A Machine Learning framework for predicting peptide inhibitors of insect voltage-gated sodium channels

--Manuscript Draft--

Manuscript Number:	
Article Type:	Full Length Article
Keywords:	Peptide-based insecticides; Pest management; Drosophila suzukii; Voltage-gated sodium channels; Machine learning
Corresponding Author:	Bruno Andrade Universidade Estadual do Sudoeste da Bahia - Campus de Jequié Jequie, Bahia BRAZIL
First Author:	Jailan da Silva Sousa
Order of Authors:	Jailan da Silva Sousa
	Lucas Sousa Palmeira
	Fabrício Santos Barbosa
	Hugo Mauricio Peña Mercado
	Tarcisio Silva Melo
	Vasco Azevedo
	Aristóteles Góes-Neto
	Joicymara Santos Xavier
	Bruno Andrade
Abstract:	Peptides represent environmentally safer and target-specific alternatives to conventional insecticides, offering advantages such as biodegradability and reduced off-target effects. However, their practical application remains limited by the complexity and cost of chemical synthesis and large-scale screening. In this study, we developed a machine-learning (ML) framework to predict peptide inhibitors of insect voltage-gated sodium channels (VGSCs), key targets in neuronal signaling and pest control. Six well-established ML algorithms were systematically evaluated, with the Support Vector Classifier (SVC) achieving the best predictive performance. To enhance interpretability, SHAP analysis revealed that physicochemical descriptors, particularly those structural partners and reflecting side-chain hydrophobicity, were the most influential for model predictions consistent with the structural determinants of VGSC-toxin interactions. The top-ranked plant-derived peptides predicted by the model were further validated through molecular docking and molecular dynamics simulations with the Drosophila suzukii VGSC, confirming stable interactions at the pore and voltage-sensing domains. These peptides, classified as cysteine-rich defensins, showed structural patterns compatible with known ion channel modulators. Although the limited availability of experimentally validated insect VGSC inhibitors constrains model generalization, the proposed approach demonstrates the potential of ML-driven sequence analysis for accelerating peptide discovery. By integrating predictive modeling with molecular simulations, this work provides a computationally efficient and biologically meaningful strategy for identifying novel bioactive peptides for sustainable pest management.
Opposed Reviewers:	

4. ARTIGO CIENTÍFICO

A Machine Learning framework for predicting peptide inhibitors of insect voltage-gated sodium channels

Jailan da Silva Sousa[a,b], Lucas Sousa Palmeira[a,b], Fabrício Santos Barbosa[b], Hugo Mauricio Peña Mercado[a], Tarcisio Silva Melo[b], Vasco Azevedo[a], Aristóteles Góes-Neto[a,c], Joicymara Santos Xavier[a,d,c], Bruno Silva Andrade[b]*

[a]Graduate Program in Bioinformatics, Federal University of Minas Gerais, Belo Horizonte 31270-901, MG, Brasil.

[b]Laboratory of Bioinformatics and Computational Chemistry, State University of Southwest Bahia, Jequié, BA, Brazil.

[c]Department of Microbiology, Molecular and Computational Biology of Fungi Laboratory, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, MG, Brazil.

[d]Computer Science Division, Technological Institute of Aeronautics, São José dos Campos 89760-000, SP, Brazil.

[e]Center for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa.

KEYWORDS: Peptide-based insecticides, Pest management, *Drosophila suzukii*, Voltage-gated sodium channels, Machine learning.

ABSTRACT

Peptides represent environmentally safer and target-specific alternatives to conventional insecticides, offering advantages such as biodegradability and reduced off-target effects. However, their practical application remains limited by the complexity and cost of chemical synthesis and large-scale screening. In this study, we developed a machine learning (ML) framework to predict peptide inhibitors of insect voltage-gated sodium channels (VGSCs), which are key targets in neuronal signaling and pest control. Six well-established ML algorithms were systematically evaluated, with the Support Vector Classifier (SVC) achieving the best predictive performance. Model interpretability analysis using SHAP revealed that physicochemical descriptors, particularly those describing structural relationships and amino acid interactions, were the most influential for model predictions, consistent with the structural determinants of VGSC-toxin interactions. The top-ranked plant-derived peptides predicted by the model were further validated through molecular docking and molecular dynamics simulations with the *Drosophila suzukii* VGSC, a major pest of soft-skinned fruits, confirming stable interactions at the pore and voltage-sensing domains. These peptides, classified as cysteine-rich defensins, exhibited structural patterns compatible with known ion channel modulators. Although the limited availability of experimentally validated insect VGSC inhibitors constrains model generalization, the proposed approach demonstrates the potential of ML-driven sequence analysis to accelerate peptide discovery. By integrating predictive modeling with molecular simulations, this work provides a computationally efficient and biologically meaningful strategy for identifying novel bioactive peptides for sustainable pest management.

INTRODUCTION

Voltage-gated sodium channels (VGSCs) are among the principal molecular targets of both synthetic and naturally derived insecticides[1]. These membrane proteins play essential physiological roles, including the initiation and propagation of action potentials in excitable tissues such as neurons and cardiac muscle[2]. Structurally, a VGSC consists of a single polypeptide chain organized into four homologous domains (DI–DIV), each containing six transmembrane helices (S1–S6) [3]. Within each domain, helices S1–S4 form the voltage-sensing domain (VSD), in which S4 acts as the primary voltage sensor, while helices S5 and S6 contribute to the pore-forming region. The latter includes a re-entrant P-loop that folds back into the membrane to create the ion selectivity filter [4–6].

Several naturally occurring peptides are known to act on VGSCs by blocking or modulating their activity. These include α-, β-, and δ-toxins, which are typically rich in cysteine residues and predominantly found in the venoms of arthropods such as spiders and scorpions [7]. These toxins bind to specific receptor sites on VGSCs site 3 in VSD-IV, site 4 in VSD-II, and site E in the pore (P) domain, thereby inhibiting channel activity [8]. Toxins that bind to site E function as sodium channel blockers, whereas those interacting with sites 3 and 4 act as modulators of channel activation by interfering with the rapid inactivation process [8,9].

In pest management, peptides have been proposed as effective and environmentally safer alternatives to conventional chemical insecticides due to their biodegradability, target specificity, and favorable pharmacological properties [10–12]. However, despite these advantages, their practical application remains limited by the complexity and high cost of chemical synthesis and large-scale production [13,14]. Therefore, the discovery of novel insecticidal peptides represents a promising strategy to accelerate the identification and future commercialization of bioactive molecules for sustainable pest control.

While numerous machine-learning approaches have been developed to predict broad functional classes of bioactive peptides, including those with insecticidal [15,16], antimicrobial [17], antibiofilm [18], hemolytic [19], and antiviral [20,21] activities, most recent studies have shifted toward target-specific peptide prediction for ion channels and other protein targets. Notable examples include *PEP-PREDNa+*, a web server designed to predict Na^+ channel-blocking peptides [22]; *PrIMP*, which identifies ion channel modulating peptides across sodium, potassium, calcium, and nicotinic acetylcholine receptors [23]; and *MetaNaBP*, which identify mammalian voltage-gated sodium channel blocking peptides [24]. These developments demonstrate the feasibility and utility of target-level peptide prediction. However, existing tools vary in scope, organism focus, modeling framework, and interpretability, leaving room for complementary strategies.

Here, we present a Support Vector Machine (SVM) classifier trained on a curated dataset of insect VGSC inhibitor sequences using sequence-derived descriptors to provide an interpretable, target-specific predictor of potential VGSC inhibition. To assess its predictive capabilities, we applied the model to a set of plant-derived peptide sequences, selecting candidates with the highest probability of belonging to the active class. The affinity of these candidates for the VGSC from *Drosophila suzukii*, a major pest of soft-skinned fruits, was further evaluated using molecular docking and molecular dynamics simulations. This machine-learning guided approach streamlines virtual screening, reducing the complexity and time required for the discovery of novel bioactive peptides targeting insect VGSCs.

METHODOLOGY

Data collection

The dataset used in this study was composed of 248 peptide sequences, equally divided into a positive set (n = 124) and a negative set (n = 124).

Positive dataset

Peptides with experimentally validated insecticidal activity targeting voltage-gated sodium channels (VGSCs) were retrieved from the UniProt database [25] using the following keywords: “insecticidal peptides,” “neurotoxic peptides,” “ion channel inhibitor OR blocker peptides,” and “voltage-gated channel inhibitor OR blocker peptides.” Only reviewed entries were considered. Manual curation was performed to ensure that each selected peptide had experimentally confirmed inhibitory activity against insect VGSCs.

Negative dataset

The negative set consisted of peptides unrelated to insecticidal activity, including insect neuropeptides, neurohormones, hormones, and antimicrobial peptides. These sequences were retrieved from UniProt using the keywords “insect brain peptides,” “insect hormones,” “insect neurohormones,” “insect neuropeptides,” “antimicrobial peptides,” and “non-insecticidal peptides,” also restricted to reviewed entries. Because no specific database of peptides with confirmed lack of VGSC inhibitory activity exists, insect neuropeptides were prioritized as negative samples, given that they coexist with VGSCs in both the central and peripheral nervous systems of insects [26,27] but do not act on these channels. To minimize potential biases, sequences containing well-defined motifs (FMRFamides, LRLRFamides, tachykinins [26]) were excluded. Additionally, peptides longer than 20 amino acids and cysteine-rich sequences (≥ 7 cysteine residues) were prioritized to reduce bias arising from differences in sequence length and cysteine content relative to the positive set.

Sequence representation methods

Feature extraction was performed using the iFeature 1.0 [28], modlAMP 4.3.2 [29], and Biopython 1.85 [30] Python packages. A comprehensive set of descriptors (Table S1) was calculated for all peptide sequences in both the positive and negative datasets. For the Pseudo K-tuple Reduced Amino Acid Composition (PseKRAAC) descriptors implemented in iFeature, the parameters were set to ktuple = 2 and $\lambda = 3$. All available amino acid grouping schemes were employed. This configuration encodes the frequencies of reduced dipeptides (based on amino acid groupings) separated by three residues along the peptide chain, thereby capturing both reduced compositional and sequence-order information. The combination of these features resulted in a high-dimensional dataset comprising 38194 variables.

Machine Learning Algorithms

Given the limited size of our dataset, machine learning algorithms were selected based on their reported robustness and performance under data scarcity conditions [31,32]. Six different algorithms were evaluated: Decision Tree, Random Forest, Support Vector Classifier (SVC), Gradient Boosting, eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). These algorithms were implemented using the scikit-learn package (version 1.5.1) in Python, except for XGBoost and LightGBM, which were utilized through their respective Python libraries. Detailed information about the algorithms is provided in the Supplementary Data.

Preprocessing and model selection

After removing redundant features and features with zero variance, the dataset comprised 8252 features, it was randomly split into 80% for training and 20% for testing

using stratified sampling to preserve class proportions. Features submitted to the SVC were standardized to have a mean of 0 and a variance of 1. Hyperparameter optimization for each algorithm was performed using a Bayesian-inspired search in the Optuna framework 4.2.0 [33]. Each trial was evaluated via 5-fold cross-validation to identify the hyperparameter combination that maximized accuracy on the training set (Table S2). The best-performing algorithm, according to evaluation metrics, was selected for further refinement.

Model performance optimization and explainability

Feature selection was performed using Recursive Feature Elimination with Cross-Validation (RFECV) [34], implemented in scikit-learn 1.5.1, to identify the most relevant features for model predictions. To improve the reliability of predicted probabilities, the selected model was calibrated using isotonic regression method [35] via the CalibratedClassifierCV module in scikit-learn, to align predicted probabilities with the true likelihood that a peptide is an insect VGSC inhibitor. Finally, model interpretability was assessed using the SHapley Additive exPlanations (SHAP) method [36] . SHAP assigns an importance value to each feature for a given prediction, allowing identification of the most influential features driving the model's predictions.

Performance measures

The models were evaluated using the following metrics: recall, precision, average precision, accuracy [37], area under the ROC curve (AUC) [38], and Matthews correlation coefficient (MCC) [39]. These were calculated as follows:

$$recall = \frac{VP}{VP + FN}$$

$$precision = \frac{VP}{VP + FP}$$

$$average\ precision = \sum_{n}(R_n - R_{n-1})P_n$$

$$accuracy = \frac{VP + VN}{VP + VN + FP + FN}$$

$$MCC = \frac{VP \times VN - FP \times FN}{\sqrt{(VN+FN)(FP+VP)(VN+FP)(FN+VP)}}$$

where TP represents the number of true positive predictions, TN is the number of true negative predictions, FP is the number of false positive predictions and FN is the number of false negative predictions, R_n and P_n are the precision and recall at the nth threshold.

Statistical significance tests for amino acid composition and model performance

To evaluate differences in amino acid composition between active (positive) and inactive (negative) peptide sequences in the training set, a two-tailed Student's t-test was applied. The significance threshold was set at $\alpha = 0.05$. Prior to testing, the data were assessed for normality using the Shapiro-Wilk test and for homogeneity of variance using Levene's test.

To compare the predictive performance of the six evaluated machine learning models across the performance metrics, a Friedman test was conducted, with $\alpha = 0.05$. When the Friedman test indicated significant differences among models, a post hoc Conover test with Holm correction was performed to adjust for multiple comparisons and identify specific pairs of models with statistically significant differences. All statistical analyses were performed using Python 3.12 with the SciPy and scikit-posthocs packages, and results were visualized using Matplotlib and Seaborn.

Prediction novel insect VGSC inhibitor peptides

A total of 365 plant-derived peptides were retrieved from the UniProt database using the search term “plant peptide OR peptides”, restricted to reviewed entries that were submitted to classification by the selected model. Plant-derived peptides were chosen for classification because many of them are plant defensins, which have potential insecticidal activity. Peptides were selected for further consideration if their predicted probability of belonging to the positive class, corresponding to potential insect VGSC inhibitors, was equal to 1. The three-dimensional structures of the peptides selected by the trained machine learning model were retrieved from the Protein Data Bank (PDB) [40] and AlphaFoldDB [41]. Structures obtained from AlphaFoldDB were refined using ModRefiner (version 2018)[42] and subsequently assessed for structural quality using MolProbity 4.5[43] and Qualitative Model Energy ANalysis (QMEAN) via the SWISS-MODEL web server [44].

Modeling and refinement of Drosophila suzukii VGSC

The three-dimensional structure of the voltage-gated sodium channel of *Drosophila suzukii* (DsNav) [45] was predicted using AlphaFold2 [46] and subjected to a 500-nanosecond molecular dynamics simulation in explicit solvent with GROMACS 2023.1 [47]. The protein-membrane system was prepared using the CHARMM-GUI web server [48] with the CHARMM36 force field (July 2022 version) [49]. Simulations were performed under conditions approximating standard *in vitro* experimental setups, including a temperature of 298.15 K, pH 7.0, and pressure of 1 atm (detailed simulation parameters are provided in the Supplementary Data). Trajectory analyses, including root mean square deviation (RMSD), root mean square fluctuation (RMSF), and radius of gyration (Rg) of the protein backbone, were performed using GROMACS routines. All simulation plots were generated with the Python packages Matplotlib 3.10.1 and Seaborn 0.13.2.

Docking molecular and dynamic simulation of predicted peptides to DsNav

Peptides identified by the trained machine learning model were subjected to molecular docking against the voltage-gated sodium channel (VGSC) of *Drosophila suzukii* (DsNav) at each of the three experimentally characterized binding sites for naturally occurring toxins. The binding sites were identified by the crystal structures 6J8E [50], 6NT4 [6] , and 6A91 [64], corresponding to site E (pore domain), site 3 (voltage-sensing domain IV, VDSIV), and site 4 (voltage-sensing domain II, VDSII), respectively. Docking was performed using HADDOCK 2.4 [53], following the standard protocol for protein–peptide docking. Binding sites were defined based on a comprehensive literature review and structural data of voltage-gated sodium channels available in the PDB, focusing on complexes with well-characterized inhibitory toxins corresponding to each site. Three-dimensional representations of protein–peptide complexes and docking analyses were generated using PyMOL 2.5.0 (The PyMOL Molecular Graphics System, version 1.2r3pre, Schrödinger, LLC).

Molecular dynamics (MD) simulations of the DsNav–peptide complexes selected from molecular docking were performed for 100 nanoseconds, focusing on the peptides with the most favorable (most negative) docking scores at each of the three tested binding sites. The system preparation and simulation conditions were identical to those used for the APO DsNav simulations. Interactions between the receptor and peptides were analyzed across 800 frames of the MD trajectories, with the first 200 frames (corresponding to 20 ns) discarded as equilibration. Contacts were calculated at every 20th frame and ranked according to the accumulated contact score using PyContact 1.0.5 [54]. Binding free energy calculations were performed using gmx_MMPBSA 1.6.4 [55], with detailed parameters provided in the Supplementary Data.

RESULTS

Comparative analysis of amino acid composition in the training dataset

The Student's t-test revealed subtle but noteworthy differences in amino acid composition between the active (positive) and inactive (negative) peptide sets (Figure 1). Active peptides exhibited slightly higher frequencies of hydrophobic residues (Trp, Tyr), polar uncharged residues (Cys, Asn), and negatively charged residues (Asp, Glu). In contrast, the negative set contained relatively higher proportions of positively charged residues (Arg, His), as well as Phe, Leu, and Gln. The sequence length in the positive dataset ranged from 13 to 121 amino acid residues, with most sequences between 40 and 80 residues (Supplementary Figure S1A). In the negative dataset, sequence lengths varied from 14 to 109 residues, with the majority centered around 60 residues. Both datasets exhibited similar median lengths, indicating no major bias in sequence size between classes.

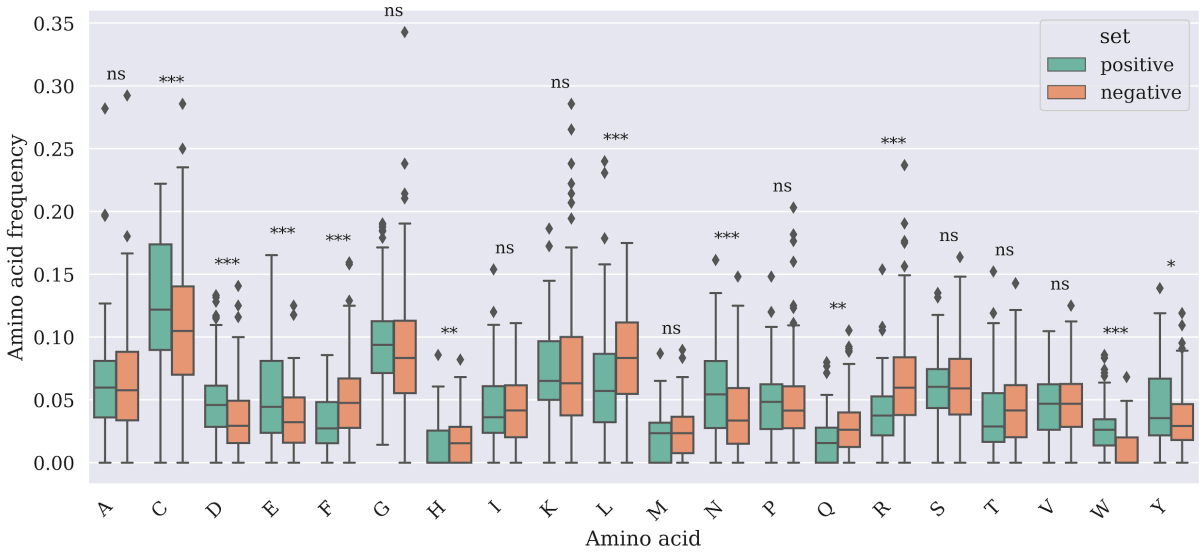

Figure 2. Comparison of amino acid frequency between positive and negative data sets. Box plots display the distribution of amino acid frequencies for each residue across both sets. Statistical significance between groups was determined using a t-test. Significance levels are indicated as $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***), and non-significant (ns).

Performance evaluation of ML algorithms

Among the tested algorithms, the SVC achieved the highest validation performance, stabilizing above 0.90 with low variance, which indicates strong generalization capacity (Figure 2). The SVC also exhibited the most stable learning process, with consistent validation accuracy and minimal signs of overfitting, as reflected by the small gap between training and validation curves. Ensemble-based models such as Random Forest, XGBoost, and Gradient Boosting showed similar patterns, with steady improvement and gradual convergence between training and validation curves as the training set size increased. In contrast, the Decision Tree model demonstrated persistent overfitting, maintaining near-perfect training accuracy but considerably lower validation accuracy. LightGBM and XGBoost displayed higher variance at smaller training sizes but improved performance with larger datasets. Overall, the SVC outperformed the other models, benefiting most from the increase in training data and achieving a superior bias-variance balance.

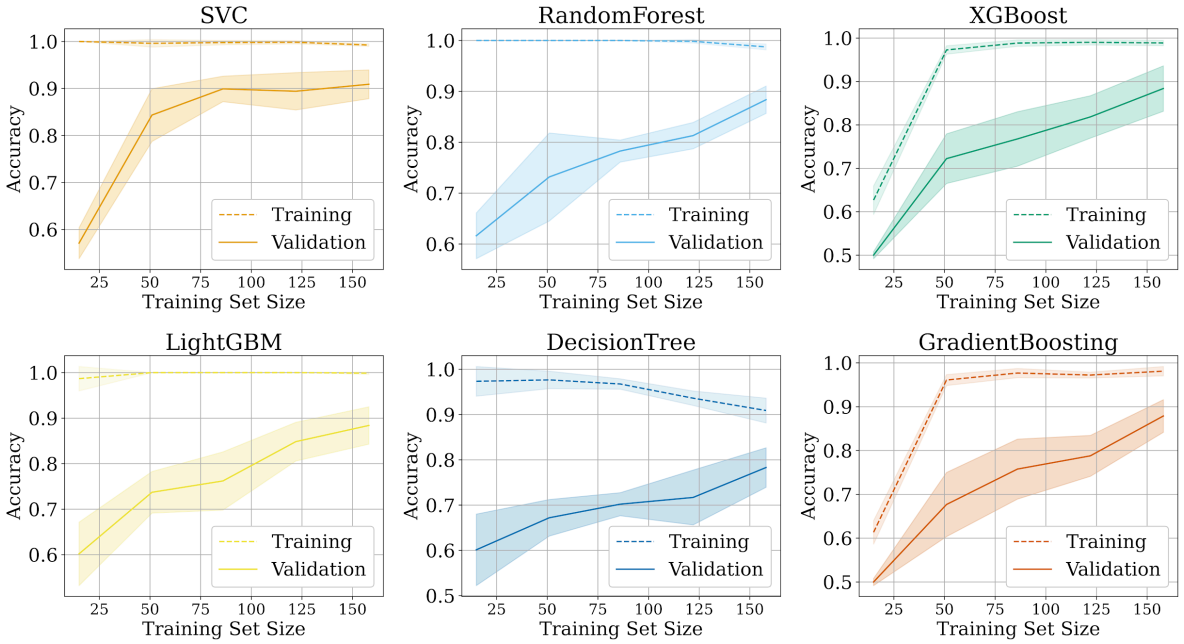

Figure 3. Learning curves of six supervised machine learning models used for peptide classification. Each plot shows training (dashed line) and validation (solid line) accuracy as a function of the training set size.

The SVC model achieved the best overall performance across all evaluated metrics in the 5-fold cross-validation (Table 1), slightly outperforming the other algorithms. The low variance across folds indicates consistent predictive behavior among the models, with the SVC showing notable better results in precision, average precision, and AUC. These metrics highlight the model's ability to effectively prioritize active sequences over negatives, an essential property in drug discovery tasks. Additionally, the higher MCC value for the SVC indicates stronger overall agreement between true and predicted classifications, reinforcing its superior generalization performance.

Table 1. Results of model evaluation in the 5-fold cross-validation

Models	Accuracy	Recall	Precision	Avr.Pre.	AUC	MCC
SVC*	**0.91 ± 0.03**	**0.93 ± 0.03**	**0.9 ± 0.05**	**0.97 ± 0.02**	**0.97 ± 0.02**	**0.82 ± 0.06**
RandomForest	0.88 ± 0.03	0.91 ± 0.04	0.87 ± 0.05	0.93 ± 0.03	0.93 ± 0.02	0.77 ± 0.05
XGBoost	0.88 ± 0.05	**0.93 ± 0.02**	0.85 ± 0.07	0.93 ± 0.04	0.93 ± 0.04	0.76 ± 0.09
LightGBM	0.88 ± 0.04	0.9 ± 0.04	0.87 ± 0.06	0.94 ± 0.02	0.94 ± 0.02	0.76 ± 0.07
DecisionTree	0.75 ± 0.04	0.83 ± 0.07	0.72 ± 0.05	0.7 ± 0.02	0.76 ± 0.03	0.52 ± 0.08
GB*	0.9 ± 0.04	**0.93 ± 0.05**	0.88 ± 0.04	0.93 ± 0.03	0.93 ± 0.03	0.8 ± 0.08

*SVC = Support Vector Machine; *GB = Gradient Boosting; Avr.Pre.= Average Precision

The Friedman test revealed significant differences among models for all metrics except recall. The Post hoc analysis using the Conover-Friedman test confirmed that the SVC achieved the best overall ranking, performing significantly better only for AUC, indicating a superior ability to correctly prioritize active sequences (Figure 3). Despite similar results for precision and average precision compared with ensemble methods in the test set (Supplementary Figure S2), the SVC showed the most stable and reliable performance, supporting its selection for subsequent optimization and explainability analyses.

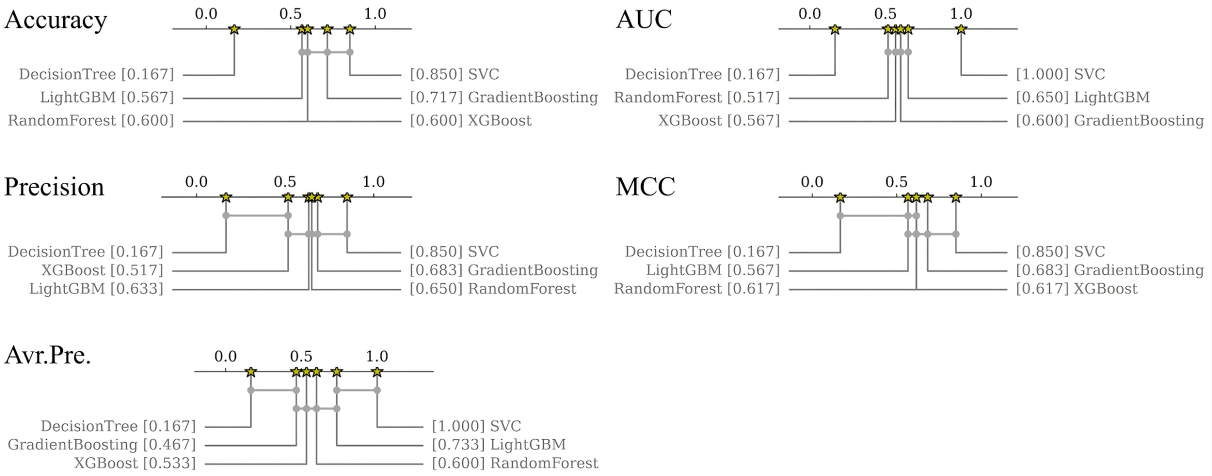

Figure 4. Critical difference diagrams showing model rankings across performance metrics based on the Conover–Friedman test with Holm correction (α = 0.05). The SVC achieved the highest overall rank, performing significantly better only for AUC, indicating its superior ability to distinguish active from inactive sequences.

Feature selection and calibration

From an initial set of 8252 features, a subset of 4130 features was identified as most relevant for classification by the SVC model. When the selected features were projected into a two-dimensional UMAP space, a clear separation between the active and negative datasets was observed (Supplementary Figure S3). To evaluate the reliability of the SVC's probabilistic predictions, we performed a calibration analysis using isotonic and sigmoid calibration methods (Supplementary Figure S4). The isotonic calibration yielded probabilities that closely followed the perfect calibration line, indicating that the predicted probabilities accurately represent the true fraction of positives. In contrast, both the uncalibrated and sigmoid-calibrated models deviated substantially, particularly in the mid-probability range, reflecting over- and under-confidence in predictions. Following optimization, the SVC model exhibited consistent performance gains across all evaluation metrics (Table 2).

Table 2. SVC performance metrics after feature selection and calibration.

Model	Accuracy	Recall	Precision	Av.precision	AUC	MCC
SVC	**0.98 ± 0.03**	**0.98 ± 0.02**	**0.97 ± 0.04**	**0.99 ± 0.01**	**0.99 ± 0.01**	**0.95 ± 0.05**

Model interpretation

SHAP analysis of feature importance revealed that the molar extinction coefficients for cystine (epsilon_prot_disu), reduced cysteine (epsilon_prot_cys), and gGranthamlag3 were the most influential features in the model prediction (Figure 4). The molar extinction coefficient features capture the overall absorbance properties of the sequences, which are influenced by the presence of cysteine, cystine, tyrosine, and tryptophan residues[56]. The SHAP summary plot shows that active sequences (positive class) display higher values for these features compared with the negative ones, indicating that increased extinction coefficients are associated with the likelihood of being an active peptide. This trend aligns with a slightly higher proportion of cysteine and tryptophs, which directly contribute to absorbance, in the active peptides compared with the inactive ones. The gGranthamlag3 feature also exhibits higher values in the active peptides, reflecting its contribution to the model's prediction. This feature represents the sequence-order-coupling (SOCNumber)

numbers, which describe the relationships between amino acids at various positions along the protein sequence[57]. Consequently, the gGranthamlag3[58] feature captures how the pattern of amino acid distribution and spatial relationships within the sequence contributes to the distinction between active and inactive peptides.

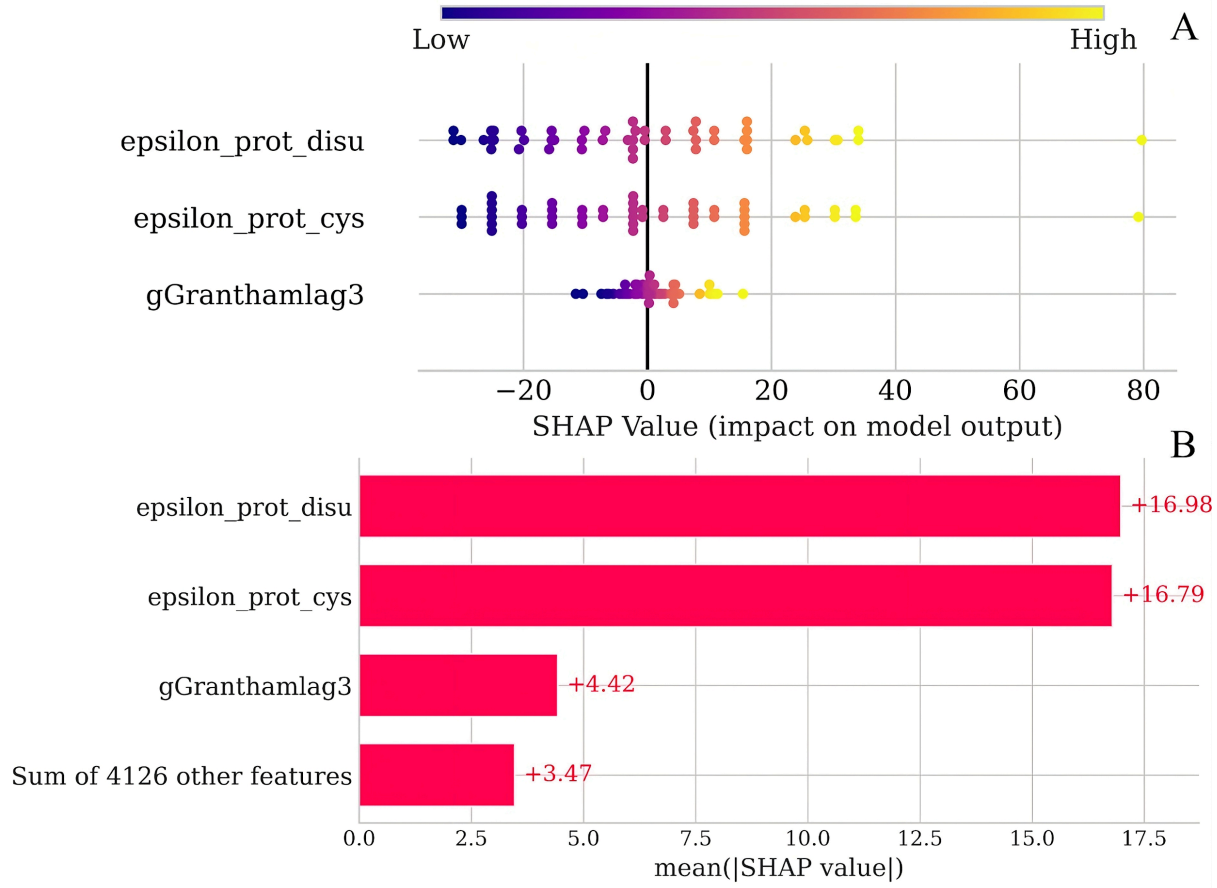

Figure 5. Most impactful features for the SVC model predictions on the test dataset, as measured by SHAP values. (A) The summary plot ranks the top 3 features from most to least impactful based on their SHAP values, illustrating the relationship between feature values (represented by color). (B) Bar plot displays the mean absolute SHAP value for each feature, highlighting their overall contribution to the model's predictive performance.

Structural validation and stability assessment of the DsNav model and selected peptides

The voltage-gated sodium channel of *Drosophila suzukii* (DsNav), exhibited three regions with pLDDT scores below 50, suggesting local disorder (Supplementary Figure S5 A). The 500-nanosecond molecular dynamics simulation in explicit solvent revealed that DsNav structure maintained its overall functional conformation, and showing no loss or inversion of secondary structural elements as well (Supplementary Figure S5 E and F). The elevated RMSD (~15 Å) observed (Supplementary Figure S5 C) resulted from residues within intrinsically disordered regions exhibiting RMSF values ≥10 Å (Supplementary Figure S5 D), consistent with the previously identified disordered segments. Structural validation indicated that 89% of the residues were located in the most favorable regions and 98% in allowed regions of the Ramachandran plot (Supplementary Figure S5 B). Notably, these disordered segments were located on the intracellular side of the membrane and did not overlap with any of the binding sites analyzed in this study. Because proper folding of these flexible loops would likely occur on the microsecond-to-minute timescale, extending beyond the simulated period, further sampling was deemed computationally prohibitive in our case. Collectively,

these findings confirmed that the DsNav structure met the quality criteria for downstream *in silico* interaction analyses.

From the 365 plant-derived peptides retrieved from the UniProt database and classified by the SVC model, 38 peptides with a predicted probability of 1.0 were selected, indicating high confidence in their classification as active candidates. Based on QMEAN scores, the structural quality of the models was comparable to that of experimentally determined structures resolved by X-ray crystallography (Supplementary Figure S6). Additionally, over 90% of the residues were located in the most favorable regions and 100% in allowed regions of the Ramachandran plot.

Molecular docking or dynamics analysis

For binding sites 3 and 4, the peptides with UniProt accession codes P56552 and P81930, respectively, were selected based on their most favorable (most negative) docking scores obtained with HADDOCK, corresponding to −126.7 for site 3 and −113.5 for site 4. For site E, the peptide P0DKH7 (second-best candidate, −123.5) was chosen, since the top-ranked peptide for this site (P81930) had already been selected as the best binder for site 4. This approach allowed the selection of a unique representative peptide for each of the three known binding sites targeted by natural toxins in voltage-gated sodium channels (VGSCs).

Molecular dynamics (MD) simulations of the DsNav–peptide complexes demonstrated that all screened peptides remained bound to the channel throughout the entire simulation period (Figure 5). The RMSD values for P0DKH7 at site E (Figure 5A-I) and P56552 at site 3 (Figure 5B-I) remained below 3 Å, consistent with conformational stability of the complexes. In contrast, P81930 at site 4 exhibited an RMSD of approximately 7 Å (Figure 5C-I), which can be attributed to the higher flexibility of its loop regions. For all complexes, the number of hydrogen bonds during the trajectories, indicating persistent intermolecular interactions (Figure 5 A-II, B-II, C-II).

All three complexes displayed binding free energy values compatible with stable interactions under physiological conditions. For the DsNav–P0DKH7 complex, the calculated enthalpy (ΔH) was −44.23 ± 14.75 kcal mol^{-1}, with −TΔS = 31.34 ± 1.92 kcal mol^{-1}, resulting in ΔG = 12.89 ± 14.87 kcal mol^{-1} (Figure 5 A-III). The DsNav–P56552 complex showed ΔH = −68.07 ± 17.28 kcal mol^{-1}, −TΔS = 35.40 ± 0.05 kcal mol^{-1}, and ΔG = −32.67 ± 17.28 kcal mol^{-1} (Figure 5 B-III). For DsNav–P81930, ΔH = −50.29 ± 9.47 kcal mol^{-1}, −TΔS = 28.48 ± 0.05 kcal mol^{-1}, and ΔG = −21.81 ± 9.47 kcal mol^{-1}(Figure 5 C-III).

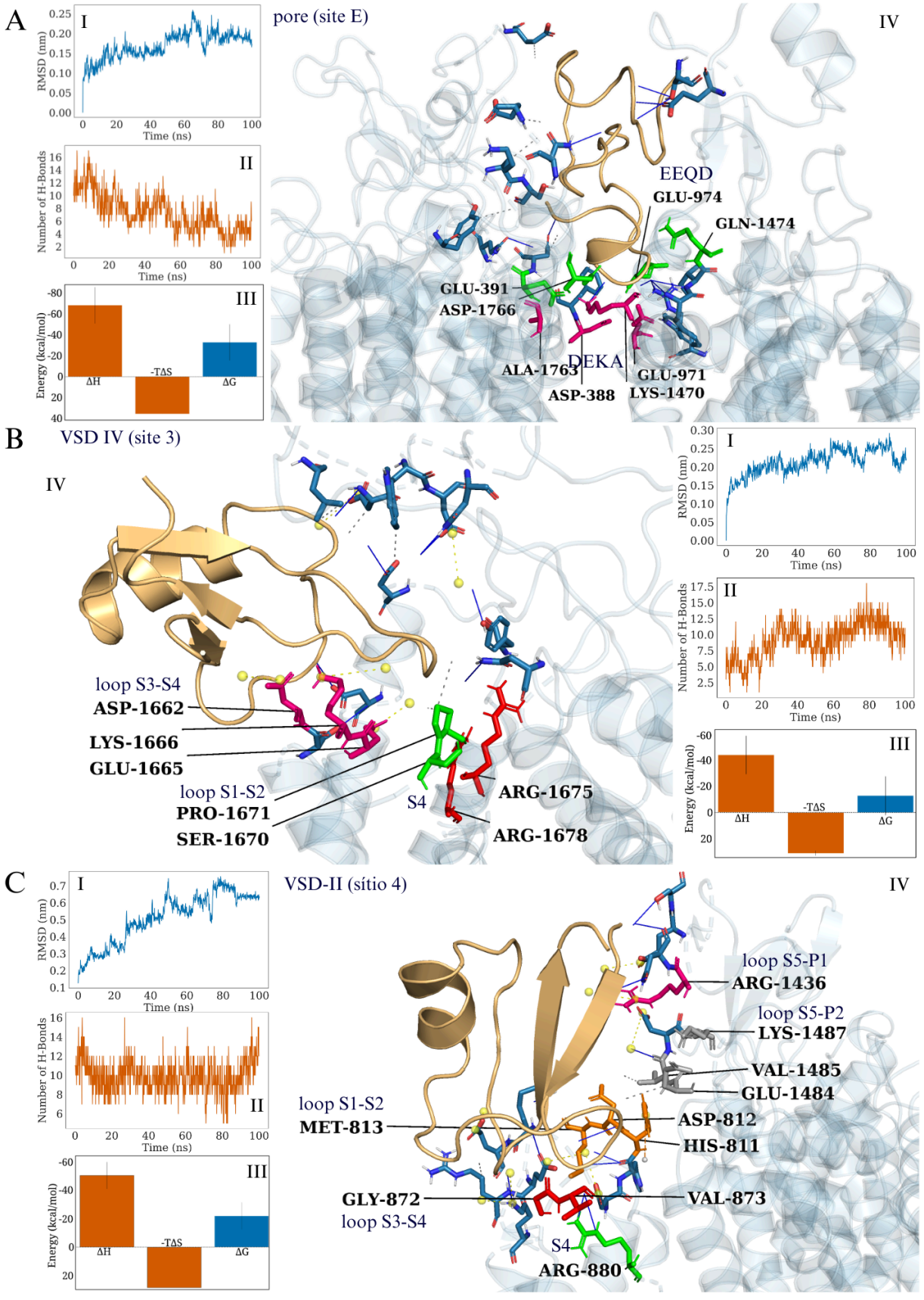
A
I
RMSD (nm)
Time (ns)
II
Number of H-Bonds
Time (ns)
III
Energy (kcal/mol)
ΔH
-TΔS
ΔG
pore (site E)
IV
EEQD
GLU-974
GLN-1474
GLU-391
ASP-1766
DEKA
ALA-1763
ASP-388
GLU-971
LYS-1470
B
VSD IV (site 3)
IV
loop S3-S4
ASP-1662
LYS-1666
GLU-1665
loop S1-S2
PRO-1671
SER-1670
S4
ARG-1675
ARG-1678
I
RMSD (nm)
Time (ns)
II
Number of H-Bonds
Time (ns)
III
Energy (kcal/mol)
ΔH
-TΔS
ΔG
C
I
RMSD (nm)
Time (ns)
II
Number of H-Bonds
Time (ns)
III
Energy (kcal/mol)
ΔH
-TΔS
ΔG
VSD-II (sítio 4)
IV
loop S5-P1
ARG-1436
loop S5-P2
LYS-1487
VAL-1485
GLU-1484
loop S1-S2
MET-813
ASP-812
HIS-811
GLY-872
loop S3-S4
VAL-873
S4
ARG-880

Figure 6. In silico affinity evaluation of three peptide candidates as DsNav inhibitor. (A) Interaction analysis of the DsNav–P0DKH7 complex, P0DKH7 a candidate predicted to block the E-site (pore region) showing: (I) Root Mean Square Deviation (RMSD) of P0DKH7 after least-squares fitting to its backbone amino acids residues; (II) Number of hydrogen bonds formed between DsNav and P0DKH7 during the molecular dynamics simulation; (III) Variation in binding enthalpy, entropy, and free energy throughout the simulation (IV) receptor-ligand interaction highlighting key residues involved in interaction. (B) and (C) same analyses as in (A), applied to the DsNav–P56552 and DsNav–P81930 complexes, where P56552 is a candidate for site-3 modulation and P81930 for site-4 modulation.

Peptide-receptor interactions

Analysis of the molecular interactions revealed that complex stabilization was primarily mediated by electrostatic contacts between positively charged residues (e.g., lysine and arginine) and negatively charged residues (e.g., aspartate and glutamate) (Supplementary Figure S7). The dominant interaction types across all complexes were salt bridges and hydrogen bonds.

For the P0DKH7 ligand (site E), key stabilizing residues included Lys29, Ala1, Pro13, and Pro28, which formed interactions mainly with residues belonging to the EEQD and DEKA motifs in the pore domain of DsNav (Figure 5 A-IV; Supplementary Figure S7 A). At site 3, the P56552 ligand established strong hydrogen bonds and salt bridges through Lys3, Lys5, Lys6, Lys42, and Arg43, primarily interacting with residues located in the loops connecting segments S1–S2 and S3–S4 of VSD IV (Figure 5 B-IV; Supplementary Figure S7 B). For the site 4 ligand P81930, the main contributing residues Lys1, Lys11, Arg36, Arg40, and Pro13, engaged residues from segments S1–S2, S3–S4, S5–P1, and S5–P2 of VSD II (Figure 5 C-IV; Supplementary Figure S7 C).

In the complexes formed at sites 3 and 4 (Supplementary Figure S7), negatively charged residues, mainly aspartate and glutamate, exhibited unfavorable energetic contributions to complex stabilization, likely due to electrostatic repulsion with negatively charged residues in the DsNav receptor. Nevertheless, these repulsive effects were effectively compensated by strong favorable interactions, primarily involving those residues that most contributed to complex stabilization.

DISCUSSION

Peptides offer environmentally safer and target-specific alternatives to conventional insecticides, but their practical application is often constrained by the complexity and cost of chemical synthesis. To address this challenge, we developed a machine-learning framework for predicting insect VGSC inhibitory peptides, enabling the identification of promising candidates. By systematically evaluating six widely used algorithms, we selected the Support Vector Classifier (SVC) as the most effective model, combining high predictive performance with interpretable insights via SHAP analysis. The practical utility of this approach was demonstrated through molecular docking and molecular dynamics simulations of top plant-derived candidates against the *Drosophila suzukii* VGSC. Collectively, these results illustrate that our integrated strategy provides a streamlined approach for accelerating the discovery of novel bioactive peptides for pest management.

Our strategy to minimize bias caused by extreme differences in sequence length and composition between the positive and negative datasets ensures that the model distinguishes active from inactive peptides based on structural partners that reflect physico-chemistry features. By preventing sequence length from becoming a dominant discriminative characteristic, we avoid a classification task that could be statistically straightforward but biologically uninformative, which might otherwise lead to misleading results [59]. A target-specific machine learning approach should instead capitalize on structural and

physicochemical attributes, as these are fundamental determinants of receptor–ligand interactions.

In a related study, Shoombuatong et al. (2024) applied a machine-learning approach to identify VGSC blockers for therapeutic purposes and reported that the sequence profiles of the positive and negative sets were comparable to those observed in our work. Similarly, mammalian VGSC inhibitors targeting domain IV of the voltage sensor exhibit patterns of hydrophobic and negatively charged residues [60], suggesting that side-chain hydrophobicity plays a central role in peptide-voltage sensor interactions. Voltage-gated sodium channels are generally highly conserved, particularly in the extracellular loop regions that form the pore domain, the selectivity filter, the voltage-sensing segments, and the inactivation loop [61] [62]. Among these, the pore domain and voltage sensor regions are especially significant, as they serve as binding sites for naturally occurring toxins that allosterically modulate VGSC function [8]. The S4 segments of all four domains act as voltage sensors and typically contain four to eight positively charged amino acids (Arg and Lys) interspersed with two hydrophobic residues [63,64]. This arrangement imparts basic characteristics to the region, facilitating interactions with peptides containing acidic side chains (Asp and Glu), which may be critical for binding to the S4 segment of VGSCs [24].

The peptide P0DKH7, identified in this study as a potential blocker of the DsNav pore (site E), may act similarly to other toxins known to obstruct VGSC pores, such as μ-conotoxins [51], by interacting with critical residues within the selectivity filter. Specifically, P0DKH7 forms interactions with Glu971 of the DEKA motif and with Glu391, Gln1474, and Asp1766 of the EEQD motif [45], thereby blocking the pore entrance.

The peptide P56552, predicted as a site 3 inhibitor targeting the VSD-IV domain, established strong hydrogen bonds and salt bridges, particularly with Glu1665 one of the residues contributing most to the stability of the DsNav-P56552 complex. This residue is located in the loop connecting the S3-S4 segments, a region that also contains Arg1675 and Arg1678, both positively charged residues in the S4 segment. These interactions are consistent with previous studies reporting the binding of scorpion α-toxins at this same site [6,65,66].

In the case of peptide P81930, which was identified as a site 4 inhibitor candidate with primary interactions with residues located in the loops connecting the S1–S2 segments of VSD-II, particularly His811, Asp812, and Met813—as well as Gln872 and Val873 in the S3–S4 loop, Arg1436 in the S5–P1 loop, and Glu1484, Val1485, and Lys1487 in the S6–P2 region, these regions have been reported as critical for maintaining sodium channel opening [51,52]. Throughout the simulations, these residues maintained strong interactions with P81930, with Asp812 and Glu1484 contributing most to the formation and stability of the DsNav–P81930 complex.

The three candidate DsNav inhibitors identified in this study are cysteine-rich plant-derived peptides classified as defensins [67]. The peptide P0DKH7 (Fa-AMP1), consisting of 40 amino acid residues, was isolated from Fagopyrum esculentum and exhibits antimicrobial activity against both Gram-positive and Gram-negative bacteria[68]. The peptide P56552 (brazzein), composed of 54 residues, was purified from Pentadiplandra brazzeana and is known for its sweet taste as well as its antimicrobial activity against bacteria and fungi [69,70]. The third peptide, P81930 (antifungal protein Psd2), isolated from Pisum sativum, has 47 residues and displays antifungal activity [71].

All three peptides predicted to act against DsNav possess four disulfide bridges, a feature directly associated with structural stability and biological activity. Fa-AMP1 and brazzein have been reported to disrupt the plasma membranes of target microorganisms, whereas Psd2 has been shown to interact with calcium channels in fungal membranes. However, no previous reports describe their insecticidal activity or their effects on insect

VGSCs. Interestingly, these peptides share structural motifs with animal-derived antimicrobial peptides, particularly cysteine patterns in which the residues are preceded or followed by polar amino acids every three positions [72]. This structural arrangement was among the most influential patterns identified by the SVC model, indicating that the algorithm successfully learned sequence features associated with bioactive peptides known to modulate insect voltage-gated sodium channels.

Despite the promising results, this study presents several limitations that should be considered. The dataset of known insect VGSC inhibitors remains limited in both size and diversity, which may restrict the generalization capacity of the machine learning model. Moreover, information regarding the potency of active peptides and their experimentally confirmed binding sites is largely unavailable, limiting the precision of predictions within a binding site–specific framework. In addition, the absence of a well-characterized negative dataset experimentally validated against insect VGSCs constrains the identification of features that most effectively distinguish active from inactive peptides.

CONCLUSIONS

This study demonstrates the potential of a target-specific machine learning approach to identify peptide inhibitors of the *D. suzukii* voltage-gated sodium channel (DsNav). By minimizing bias associated with sequence length and composition, our model effectively learned physicochemical and structural patterns that distinguish active from inactive peptides. The identified candidate peptides Fa-AMP1, brazzein, and Psd2 exhibited strong and stable interactions with distinct functional domains of DsNav, suggesting mechanisms of inhibition consistent with known toxin–channel interactions. These findings provide a valuable starting point for the discovery of novel bioinsecticidal peptides and highlight the importance of integrating target-specific machine learning predictions with molecular modeling for mechanistic interpretation. Future experimental validation will be essential to confirm the predicted activities and to refine computational strategies for peptide-based insecticidal design.

Selected peptides presented promising potential *in silico*, highlighting them as strong candidates for future *in vivo* experimental investigations as insecticidal agents against *Drosophila suzukii*. Furthermore, integrating machine learning into virtual screening pipelines offers significant advantages, including a substantial reduction in computational cost and time. Unlike conventional structure-based methods that rely on the availability of high-quality 3D models of ligands and receptors, sequence-based machine learning models can perform accurate predictions directly from amino acid sequences. This capability enables large-scale and rapid screening of potential ligands and targets, accelerating early-stage bioinsecticide discovery and making the overall process more efficient and scalable. Together, these results underscore the power of combining machine learning and structural modeling to uncover novel peptide inhibitors and pave the way for rational design of next-generation bioinsecticidal agents targeting insect sodium channels.

Data Availability

All the data and code used for developing this study is available on https://github.com/lbqc-uesb/Insect-VGSC-inhibitor-ML-predictor-framework.

Supplementary Data.

Additional methodology and results details, including tables and figures of the machine learning models' performance, and for *in silico* experiments.

REFERENCES

[1] ffrench-Constant RH, Williamson MS, Davies TGE, Bass C. Ion channels as insecticide targets. J Neurogenet 2016;30:163–77. https://doi.org/10.1080/01677063.2016.1229781.
[2] Kasuya J, Iyengar A, Chen H-L, Lansdon P, Wu C-F, Kitamoto T. Milk-whey diet substantially suppresses seizure-like phenotypes of *para*^{Shu}, a *Drosophila* voltage-gated sodium channel mutant. J Neurogenet 2019;33:164–78. https://doi.org/10.1080/01677063.2019.1597082.
[3] Catterall WA. Ion Channel Voltage Sensors: Structure, Function, and Pathophysiology. Neuron 2010;67:915–28. https://doi.org/10.1016/j.neuron.2010.08.021.
[4] Catterall WA. Structure and function of voltage‐gated sodium channels at atomic resolution. Exp Physiol 2014;99:35–51. https://doi.org/10.1113/expphysiol.2013.071969.
[5] Catterall WA, Wisedchaisri G, Zheng N. The chemical basis for electrical signaling. Nat Chem Biol 2017;13:455–63. https://doi.org/10.1038/nchembio.2353.
[6] Clairfeuille T, Cloake A, Infield DT, Llongueras JP, Arthur CP, Li ZR, et al. Structural basis of α-scorpion toxin action on Na_v channels. Science 2019;363:eaav8573. https://doi.org/10.1126/science.aav8573.
[7] Daly NL, Wilson D. Structural diversity of arthropod venom toxins. Toxicon 2018;152:46–56. https://doi.org/10.1016/j.toxicon.2018.07.018.
[8] Li Z, Wu Q, Yan N. A structural atlas of druggable sites on Na_v channels. Channels 2024;18:2287832. https://doi.org/10.1080/19336950.2023.2287832.
[9] Liu Y, Bezanilla F. Comparison of two fast-inactivation deficient mutants in voltage-gated sodium channel. Biophys J 2024;123:106a–7a. https://doi.org/10.1016/j.bpj.2023.11.765.
[10] Ho TNT, Turner A, Pham SH, Nguyen HT, Nguyen LTT, Nguyen LT, et al. Cysteine-rich peptides: From bioactivity to bioinsecticide applications. Toxicon 2023;230:107173. https://doi.org/10.1016/j.toxicon.2023.107173.
[11] Jin Y, Wang Z, Dong A-Y, Huang Y-Q, Hao G-F, Song B-A. Web repositories of natural agents promote pests and pathogenic microbes management. Brief Bioinform 2021;22:bbab205. https://doi.org/10.1093/bib/bbab205.
[12] Sparks TC, Crossthwaite AJ, Nauen R, Banba S, Cordova D, Earley F, et al. Insecticides, biologics and nematicides: Updates to IRAC's mode of action classification - a tool for resistance management. Pestic Biochem Physiol 2020;167:104587. https://doi.org/10.1016/j.pestbp.2020.104587.
[13] Wu Z, Li Y, Zhang L, Ding Z, Shi G. Microbial production of small peptide: pathway engineering and synthetic biology. Microb Biotechnol 2021;14:2257–78. https://doi.org/10.1111/1751-7915.13743.
[14] Clement H, Flores V, Diego-Garcia E, Corrales-Garcia L, Villegas E, Corzo G. A comparison between the recombinant expression and chemical synthesis of a short cysteine-rich insecticidal spider peptide. J Venom Anim Toxins Trop Dis 2015;21:19. https://doi.org/10.1186/s40409-015-0018-7.
[15] Lee B, Shin MK, Hwang I-W, Jung J, Shim YJ, Kim GW, et al. A Deep Learning Approach with Data Augmentation to Predict Novel Spider Neurotoxic Peptides. Int J Mol Sci 2021;22:12291. https://doi.org/10.3390/ijms222212291.
[16] Nambiar P, Mitra D, Dutta A. Machine learning assisted screening framework for insecticidal peptides. Mater Today Proc 2023;72:41–6. https://doi.org/10.1016/j.matpr.2022.05.455.
[17] Wang G, Vaisman II, Van Hoek ML. Machine Learning Prediction of Antimicrobial Peptides. In: Simonson T, editor. Comput. Pept. Sci., vol. 2405, New York, NY: Springer US; 2022, p. 1–37. https://doi.org/10.1007/978-1-0716-1855-4_1.
[18] Sharma A, Gupta P, Kumar R, Bhardwaj A. dPABBs: A Novel in silico Approach for Predicting and Designing Anti-biofilm Peptides. Sci Rep 2016;6:21839. https://doi.org/10.1038/srep21839.
[19] Plisson F, Ramírez-Sánchez O, Martínez-Hernández C. Machine learning-guided discovery and design of non-hemolytic peptides. Sci Rep 2020;10:16581. https://doi.org/10.1038/s41598-020-73644-6.
[20] Lin T-T, Sun Y-Y, Wang C-T, Cheng W-C, Lu I-H, Lin C-Y, et al. AI4AVP: an antiviral peptides predictor in deep learning approach with generative adversarial network data augmentation. Bioinforma Adv 2022;2:vbac080. https://doi.org/10.1093/bioadv/vbac080.
[21] Xu J, Xu C, Cao R, He Y, Bin Y, Zheng C-H. Generative Adversarial Network-Based Data Augmentation Method for Anti-coronavirus Peptides Prediction. In: Huang D-S, Premaratne P,

Jin B, Qu B, Jo K-H, Hussain A, editors. Adv. Intell. Comput. Technol. Appl., vol. 14088, Singapore: Springer Nature Singapore; 2023, p. 67–76. https://doi.org/10.1007/978-981-99-4749-2_6.

[22] Herrera-Bravo J, Farías JG, Contreras FP, Herrera-Belén L, Beltrán JF. PEP-PREDNa+: A web server for prediction of highly specific peptides targeting voltage-gated Na+ channels using machine learning techniques. Comput Biol Med 2022;145:105414. https://doi.org/10.1016/j.compbiomed.2022.105414.

[23] Lee B, Shin MK, Kim T, Shim YJ, Joo JWJ, Sung J-S, et al. Prediction Models for Identifying Ion Channel-Modulating Peptides via Knowledge Transfer Approaches. IEEE J Biomed Health Inform 2022;26:6150–60. https://doi.org/10.1109/JBHI.2022.3204776.

[24] Shoombuatong W, Homdee N, Schaduangrat N, Chumnanpuen P. Leveraging a meta-learning approach to advance the accuracy of Nav blocking peptides prediction. Sci Rep 2024;14:4463. https://doi.org/10.1038/s41598-024-55160-z.

[25] The UniProt Consortium, Bateman A, Martin M-J, Orchard S, Magrane M, Adesina A, et al. UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res 2025;53:D609–17. https://doi.org/10.1093/nar/gkae1010.

[26] Nässel DR. Neuropeptides in the nervous system of Drosophila and other insects: multiple roles as neuromodulators and neurohormones. Prog Neurobiol 2002;68:1–84. https://doi.org/10.1016/S0301-0082(02)00057-6.

[27] Wicher D. Peptidergic Modulation of an Insect Na^+ Current: Role of Protein Kinase A and Protein Kinase C. J Neurophysiol 2001;85:374–83. https://doi.org/10.1152/jn.2001.85.1.374.

[28] Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Wang Y, et al. *iFeature* : a Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 2018;34:2499–502. https://doi.org/10.1093/bioinformatics/bty140.

[29] Müller AT, Gabernet G, Hiss JA, Schneider G. modlAMP: Python for antimicrobial peptides. Bioinformatics 2017;33:2753–5. https://doi.org/10.1093/bioinformatics/btx285.

[30] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25:1422–3. https://doi.org/10.1093/bioinformatics/btp163.

[31] Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. Npj Comput Mater 2023;9:42. https://doi.org/10.1038/s41524-023-01000-z.

[32] Zantvoort K, Nacke B, Görlich D, Hornstein S, Jacobi C, Funk B. Estimation of minimal data sets sizes for machine learning predictions in digital mental health interventions. Npj Digit Med 2024;7:361. https://doi.org/10.1038/s41746-024-01360-w.

[33] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., Anchorage AK USA: ACM; 2019, p. 2623–31. https://doi.org/10.1145/3292500.3330701.

[34] Chen X, Jeong JC. Enhanced recursive feature elimination. Sixth Int. Conf. Mach. Learn. Appl. ICMLA 2007, Cincinnati, OH, USA: IEEE; 2007, p. 429–35. https://doi.org/10.1109/ICMLA.2007.35.

[35] Jiang X, Osl M, Kim J, Ohno-Machado L. Smooth isotonic regression: a new method to calibrate predictive models. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci 2011;2011:16–20.

[36] Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst 2014;41:647–65. https://doi.org/10.1007/s10115-013-0679-x.

[37] Dehmer M, Basak SC, editors. Statistical and Machine Learning Approaches for Network Analysis. 1st ed. Wiley; 2012. https://doi.org/10.1002/9781118346990.

[38] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 1997;30:1145–59. https://doi.org/10.1016/S0031-3203(96)00142-2.

[39] Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLOS ONE 2017;12:e0177678. https://doi.org/10.1371/journal.pone.0177678.

[40] wwPDB consortium, Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, et al. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res 2019;47:D520–8. https://doi.org/10.1093/nar/gky949.

[41] Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res 2024;52:D368–75. https://doi.org/10.1093/nar/gkad1011.
[42] Xu D, Zhang Y. Improving the Physical Realism and Structural Accuracy of Protein Models by a Two-Step Atomic-Level Energy Minimization. Biophys J 2011;101:2525–34. https://doi.org/10.1016/j.bpj.2011.10.024.
[43] Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, et al. MolProbity: More and better reference data for improved all‐atom structure validation. Protein Sci 2018;27:293–315. https://doi.org/10.1002/pro.3330.
[44] Benkert P, Tosatto SCE, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins Struct Funct Bioinforma 2008;71:261–77. https://doi.org/10.1002/prot.21715.
[45] Yuan L, Zhang K, Wang Z, Xian L, Liu K, Wu S. Functional diversity of voltage‐gated sodium channel in *Drosophila suzukii* (Matsumura). Pest Manag Sci 2024;80:592–601. https://doi.org/10.1002/ps.7786.
[46] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2.
[47] Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. J Comput Chem 2005;26:1701–18. https://doi.org/10.1002/jcc.20291.
[48] Jo S, Kim T, Iyer VG, Im W. CHARMM‐GUI: A web‐based graphical user interface for CHARMM. J Comput Chem 2008;29:1859–65. https://doi.org/10.1002/jcc.20945.
[49] Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, De Groot BL, et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat Methods 2017;14:71–3. https://doi.org/10.1038/nmeth.4067.
[50] Pan Y, Wang S, Zhang Q, Lu Q, Su D, Zuo Y, et al. Analysis and prediction of animal toxins by various Chou's pseudo components and reduced amino acid compositions. J Theor Biol 2019;462:221–9. https://doi.org/10.1016/j.jtbi.2018.11.010.
[51] Shen H, Li Z, Jiang Y, Pan X, Wu J, Cristofori-Armstrong B, et al. Structural basis for the modulation of voltage-gated sodium channels by animal toxins. Science 2018;362:eaau2596. https://doi.org/10.1126/science.aau2596.
[52] Zhu S, Gao B, Peigneur S, Tytgat J. How a Scorpion Toxin Selectively Captures a Prey Sodium Channel: The Molecular and Evolutionary Basis Uncovered. Mol Biol Evol 2020;37:3149–64. https://doi.org/10.1093/molbev/msaa152.
[53] Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. J Am Chem Soc 2003;125:1731–7. https://doi.org/10.1021/ja026939x.
[54] Scheurer M, Rodenkirch P, Siggel M, Bernardi RC, Schulten K, Tajkhorshid E, et al. PyContact: Rapid, Customizable, and Visual Analysis of Noncovalent Interactions in MD Simulations. Biophys J 2018;114:577–83. https://doi.org/10.1016/j.bpj.2017.12.003.
[55] Valdés-Tresanco MS, Valdés-Tresanco ME, Valiente PA, Moreno E. gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS. J Chem Theory Comput 2021;17:6281–91. https://doi.org/10.1021/acs.jctc.1c00645.
[56] Gill SC, Von Hippel PH. Calculation of protein extinction coefficients from amino acid sequence data. Anal Biochem 1989;182:319–26. https://doi.org/10.1016/0003-2697(89)90602-7.
[57] Chou K-C. Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect. Biochem Biophys Res Commun 2000;278:477–83. https://doi.org/10.1006/bbrc.2000.3815.
[58] Grantham R. Amino Acid Difference Formula to Help Explain Protein Evolution. Science 1974;185:862–4. https://doi.org/10.1126/science.185.4154.862.
[59] González M, Durán RE, Seeger M, Araya M, Jara N. Negative dataset selection impacts machine learning-based predictors for multiple bacterial species promoters. Bioinformatics 2025;41:btaf135. https://doi.org/10.1093/bioinformatics/btaf135.
[60] Murray JK, Long J, Zou A, Ligutti J, Andrews KL, Poppe L, et al. Single Residue Substitutions That Confer Voltage-Gated Sodium Ion Channel Subtype Selectivity in the Na_V 1.7 Inhibitory

Peptide GpTx-1. J Med Chem 2016;59:2704–17. https://doi.org/10.1021/acs.jmedchem.5b01947.
[61] Duclohier H. Structure–function studies on the voltage-gated sodium channel. Biochim Biophys Acta BBA - Biomembr 2009;1788:2374–9. https://doi.org/10.1016/j.bbamem.2009.08.017.
[62] Silva JJ, Scott JG. Conservation of the voltage‐sensitive sodium channel protein within the Insecta. Insect Mol Biol 2020;29:9–18. https://doi.org/10.1111/imb.12605.
[63] Davies TGE, Field LM, Usherwood PNR, Williamson MS. A comparative study of voltage‐gated sodium channels in the Insecta: implications for pyrethroid resistance in Anopheline and other Neopteran species. Insect Mol Biol 2007;16:361–75. https://doi.org/10.1111/j.1365-2583.2007.00733.x.
[64] Payandeh J, Scheuer T, Zheng N, Catterall WA. The crystal structure of a voltage-gated sodium channel. Nature 2011;475:353–8. https://doi.org/10.1038/nature10238.
[65] Phulera S, Dickson CJ, Schwalen CJ, Khoshouei M, Cassell SJ, Sun Y, et al. Scorpion α-toxin LqhαIT specifically interacts with a glycan at the pore domain of voltage-gated sodium channels. Structure 2024;32:1611-1620.e4. https://doi.org/10.1016/j.str.2024.07.021.
[66] Wang J, Yarov-Yarovoy V, Kahn R, Gordon D, Gurevitz M, Scheuer T, et al. Mapping the receptor site for α-scorpion toxins on a Na^+ channel voltage sensor. Proc Natl Acad Sci 2011;108:15426–31. https://doi.org/10.1073/pnas.1112320108.
[67] González‐Castro R, Gómez‐Lim MA, Plisson F. Cysteine‐Rich Peptides: Hyperstable Scaffolds for Protein Engineering. ChemBioChem 2021;22:961–73. https://doi.org/10.1002/cbic.202000634.
[68] Fujimura M, Minami Y, Watanabe K, Tadera K. Purification, Characterization, and Sequencing of a Novel Type of Antimicrobial Peptides, *Fa* -AMP1 and *Fa* -AMP2, from Seeds of Buckwheat (*Fagopyrum esculentum* Moench.). Biosci Biotechnol Biochem 2003;67:1636–42. https://doi.org/10.1271/bbb.67.1636.
[69] Ming D, Hellekant G. Brazzein, a new high‐potency thermostable sweet protein from *Pentadiplandra brazzeana* B. FEBS Lett 1994;355:106–8. https://doi.org/10.1016/0014-5793(94)01184-2.
[70] Yount NY, Yeaman MR. Multidimensional signatures in antimicrobial peptides. Proc Natl Acad Sci 2004;101:7363–8. https://doi.org/10.1073/pnas.0401567101.
[71] Almeida MS, Cabral KMS, Zingali RB, Kurtenbach E. Characterization of Two Novel Defense Peptides from Pea (Pisum sativum) Seeds. Arch Biochem Biophys 2000;378:278–86. https://doi.org/10.1006/abbi.2000.1824.
[72] Yacoub T, Rima M, Karam M, Sabatier J-M, Fajloun Z. Antimicrobials from Venomous Animals: An Overview. Molecules 2020;25:2402. https://doi.org/10.3390/molecules25102402.

5. CONCLUSÃO

Este estudo treinou um modelo de *Support Vector Machine Classifier* para prever peptídeos inibidores de VGSCs de insetos, demonstrando desempenho superior em comparação com os de Árvore de Decisão, Floresta Aleatória, XGBoost, LightGBM e GMB. Os candidatos a peptídeos selecionados pelo modelo de SVC e validados *in silico* são peptídeos ricos em cisteína, derivados de plantas, com atividade antimicrobiana. Este é o primeiro estudo a considerar essas moléculas como candidatas a moduladores de VGSCs de insetos. O cuidadoso processo de curadoria da construção de conjuntos de dados positivos e negativos, combinado com análises estruturais usando *docking* e simulações de dinâmica molecular, juntamente com avaliações de energia livre de ligação, forneceu evidências promissoras do potencial bioativo dos candidatos a peptídeos. Os peptídeos selecionados apresentaram potencial promissor, o que os sugere para futuras investigações experimentais *in vivo* como agentes inseticidas contra *Drosophila suzukii*. Além disso, uma das principais vantagens de incorporar aprendizado de máquina em fluxos de trabalho de triagem virtual é a redução significativa de tempo e recursos computacionais necessários. Os próprios métodos tradicionais exigem a preparação de estruturas 3D de alta qualidade para ligantes e receptores-alvo, que requer muito tempo dedicado e recursos computacionais. Essa etapa pode ser demorada e tecnicamente exigente, especialmente quando estruturas experimentais não estão disponíveis e o número de ligantes a serem avaliados for grande. Por outro lado, modelos de aprendizado de máquina, particularmente aqueles treinados com descritores baseados em sequência, podem fazer previsões precisas usando apenas as sequências de aminoácidos das proteínas-alvo. Isso elimina a necessidade de modelagem estrutural, permitindo a triagem rápida e em larga escala de potenciais ligantes e alvos. Como resultado, a descoberta de fármacos em estágio inicial pode ser bastante acelerada, tornando o processo geral mais eficiente e escalável em comparação com abordagens convencionais baseadas apenas em estrutura. Por fim, nossos modelos, assim como os dados, estão disponíveis publicamente para a comunidade acadêmica que pode ser usado para avaliar novos peptídeos inseticidas para outras pragas além da *D.suzukii*.

6. REFERÊNCIAS BIBLIOGRÁFICAS

DAS, Dibya Ranjan, et al. Molecular docking and its application in search of antisickling agent from Carica papaya. **J. Appl. Biol. Biotechnol**, 2020, 8.01: 105-116.

FRIEDMAN, Jerome H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, 2001, 1189-1232.

PAULA, M. A.; LOPES, P. H. S.; TIDON, R. First record of *Drosophila suzukii* in the Brazilian Savanna. **Drosophila Information Service**, 2014, 97.0: 113-115.

ABRAHAM, Mark James *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. **SoftwareX**, v. 1–2, p. 19–25, set. 2015.

AFTAB, Muhammad *et al.* Hyper-parameter tuning through innovative designing to avoid over-fitting in machine learning modelling: a case study of small data sets. **Journal of Statistical Computation and Simulation**, v. 95, n. 7, p. 1595–1609, 3 maio 2025.

AGBO, John *et al.* Therapeutic efficacy of voltage-gated sodium channel inhibitors in epilepsy. **Acta Epileptologica**, v. 5, n. 1, p. 16, 28 jun. 2023.

AGU, P. C. *et al.* Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management. **Scientific Reports**, v. 13, n. 1, p. 13398, 17 ago. 2023.

ALTHNIAN, Alhanoof *et al.* Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. **Applied Sciences**, v. 11, n. 2, p. 796, 15 jan. 2021.

ANDREAZZA, F. *et al.* Drosophila suzukii in Southern Neotropical Region: Current Status and Future Perspectives. **Neotropical Entomology**, v. 46, n. 6, p. 591–605, dez. 2017.

ANDREAZZA, Felipe *et al. Drosophila suzukii* (Diptera: Drosophilidae) Arrives at Minas Gerais State, a Main Strawberry Production Region in Brazil. **Florida Entomologist**, v. 99, n. 4, p. 796–798, dez. 2016a.

ANDREAZZA, Felipe *et al. Drosophila suzukii* (Diptera: Drosophilidae) Arrives at Minas Gerais State, a Main Strawberry Production Region in Brazil. **Florida Entomologist**, v. 99, n. 4, p. 796–798, dez. 2016b.

ARMSTRONG, Clay M. Na channel inactivation from open and closed states. **Proceedings of the National Academy of Sciences**, v. 103, n. 47, p. 17991–17996, 21 nov. 2006.

ARNOLD, Konstantin *et al.* The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. **Bioinformatics**, v. 22, n. 2, p. 195–201, 15 jan. 2006.

AWAD, Mariette; KHANNA, Rahul. Machine Learning. *In*: AWAD, Mariette; KHANNA, Rahul (Eds.). **Efficient Learning Machines**. Berkeley, CA: Apress, 2015. p. 1–18.

BARATHI, Selvaraj *et al.* Present status of insecticide impacts and eco-friendly approaches for remediation-a review. **Environmental Research**, v. 240, p. 117432, jan. 2024.

BARBIERO, Pietro; SQUILLERO, Giovanni; TONDA, Alberto. **Modeling Generalization in Machine Learning: A Methodological and Computational Study**. arXiv, , 2020. Disponível em: <https://arxiv.org/abs/2006.15680>. Acesso em: 27 jun. 2025

BARDUCCI, Alessandro; BONOMI, Massimiliano; PARRINELLO, Michele. Metadynamics. **WIREs Computational Molecular Science**, v. 1, n. 5, p. 826–843, set. 2011.

BAVITHRA, Chandra Mohan Muthu Lakshmi *et al.* Baseline susceptibility of an A1 quarantine pest - the South American tomato pinworm Tuta absoluta (Lepidoptera: Gelechiidae) to insecticides: past incidents and future probabilities in line to implementing successful pest management. **Frontiers in Plant Science**, v. 15, p. 1404250, 26 ago. 2024.

BEAUMELLE, Léa *et al.* Pesticide effects on soil fauna communities—A meta‐analysis. **Journal of Applied Ecology**, v. 60, n. 7, p. 1239–1253, jul. 2023.

BEKKER, Jessa; DAVIS, Jesse. Learning from positive and unlabeled data: a survey. **Machine Learning**, v. 109, n. 4, p. 719–760, abr. 2020.

BENITO, Norton Polo; LOPES-DA-SILVA, Marcelo; SANTOS, Régis Sivori Silva Dos. Potential spread and economic impact of invasive Drosophila suzukii in Brazil. **Pesquisa Agropecuária Brasileira**, v. 51, n. 5, p. 571–578, maio 2016.

BENTÉJAC, Candice; CSÖRGŐ, Anna; MARTÍNEZ-MUÑOZ, Gonzalo. A comparative analysis of gradient boosting algorithms. **Artificial Intelligence Review**, v. 54, n. 3, p. 1937–1967, mar. 2021.

BHUNYA, S. P.; PATI, P. C. Genotoxic effects of a synthetic pyrethroid insecticide, cypermethrin, in mice in vivo. **Toxicology Letters**, v. 41, n. 3, p. 223–230, jun. 1988.

BILOTTA, Anthony J.; CONG, Yingzi. Gut microbiota metabolite regulation of host defenses at mucosal surfaces: implication in precision medicine. **Precision Clinical Medicine**, v. 2, n. 2, p. 110–119, 1 jun. 2019.

BOUGHDAD, Ahmed *et al.* First record of the invasive spotted wing Drosophila infesting berry crops in Africa. **Journal of Pest Science**, v. 94, n. 2, p. 261–271, mar. 2021.

BOURDIN, Céline M. *et al.* Molecular and functional characterization of a novel sodium channel TipE-like auxiliary subunit from the American cockroach Periplaneta americana. **Insect Biochemistry and Molecular Biology**, v. 66, p. 136–144, nov. 2015.

BOURNE, Philip E.; WEISSIG, Helge (ORGS.). **Structural Bioinformatics**. 1. ed. *[S.l.]*: Wiley, 2003. v. 44

BRACKENBURY, William J.; ISOM, Lori L. Na+ Channel ? Subunits: Overachievers of the Ion Channel Family. **Frontiers in Pharmacology**, v. 2, 2011.

BRAGA, Anna Rafaela Cavalcante *et al.* Global health risks from pesticide use in Brazil. **Nature Food**, v. 1, n. 6, p. 312–314, 17 jun. 2020.

BURLEY, Stephen K. *et al.* Updated resources for exploring experimentally-determined PDB structures and Computed Structure Models at the RCSB Protein Data Bank. **Nucleic Acids Research**, v. 53, n. D1, p. D564–D574, 6 jan. 2025.

CALABRIA, G. *et al.* First records of the potential pest species *Drosophila suzukii* (Diptera: Drosophilidae) in Europe. **Journal of Applied Entomology**, v. 136, n. 1–2, p. 139–147, fev. 2012.

CAN, Tolga. Introduction to Bioinformatics. *In*: YOUSEF, Malik; ALLMER, Jens (Orgs.). **miRNomics: MicroRNA Biology and Computational Analysis**. Methods in Molecular Biology. Totowa, NJ: Humana Press, 2014. v. 1107 p. 51–71.

CARRACEDO-REBOREDO, Paula *et al.* A review on machine learning approaches and trends in

drug discovery. **Computational and Structural Biotechnology Journal**, v. 19, p. 4538–4558, 2021.

CASE, David A. *et al.* The Amber biomolecular simulation programs. **Journal of Computational Chemistry**, v. 26, n. 16, p. 1668–1688, dez. 2005.

CATTERALL, William A. Ion Channel Voltage Sensors: Structure, Function, and Pathophysiology. **Neuron**, v. 67, n. 6, p. 915–928, set. 2010.

CATTERALL, William A.; WISEDCHAISRI, Goragot; ZHENG, Ning. The chemical basis for electrical signaling. **Nature Chemical Biology**, v. 13, n. 5, p. 455–463, maio 2017.

CATTERALL, William A.; WISEDCHAISRI, Goragot; ZHENG, Ning. The conformational cycle of a prototypical voltage-gated sodium channel. **Nature Chemical Biology**, v. 16, n. 12, p. 1314–1320, dez. 2020.

CAZALS, Frédéric; DREYFUS, Tom. The structural bioinformatics library: modeling in biomolecular science and beyond. **Bioinformatics**, v. 33, n. 7, p. 997–1004, 1 abr. 2017.

CLAIRFEUILLE, Thomas *et al.* Structural basis of α-scorpion toxin action on Na_v channels. **Science**, v. 363, n. 6433, p. eaav8573, 22 mar. 2019a.

CLAIRFEUILLE, Thomas *et al.* Structural basis of α-scorpion toxin action on Na_v channels. **Science**, v. 363, n. 6433, p. eaav8573, 22 mar. 2019b.

CLEMENT, Herlinda *et al.* A comparison between the recombinant expression and chemical synthesis of a short cysteine-rich insecticidal spider peptide. **Journal of Venomous Animals and Toxins including Tropical Diseases**, v. 21, n. 1, p. 19, dez. 2015.

COSTA, Lucio, G. Neurotoxicity of pesticides: a brief review. **Frontiers in Bioscience**, v. 13, n. 13, p. 1240, 2008.

CRENNA, Eleonora *et al.* Characterizing honey bee exposure and effects from pesticides for chemical prioritization and life cycle assessment. **Environment International**, v. 138, p. 105642, maio 2020.

DALY, Norelle L.; WILSON, David. Structural diversity of arthropod venom toxins. **Toxicon**, v. 152, p. 46–56, set. 2018.

DAUBER-OSGUTHORPE, Pnina; HAGLER, A. T. Biomolecular force fields: where have we been, where are we now, where do we need to go and how do we get there? **Journal of Computer-Aided Molecular Design**, v. 33, n. 2, p. 133–203, fev. 2019.

DEANS, Carrie; HUTCHISON, William D. Propensity for resistance development in the invasive berry pest, spotted‐wing drosophila (*Drosophila suzukii*), under laboratory selection. **Pest Management Science**, v. 78, n. 12, p. 5203–5212, dez. 2022.

DECOURTYE, Axel *et al.* Effects of imidacloprid and deltamethrin on associative learning in honeybees under semi-field and laboratory conditions. **Ecotoxicology and Environmental Safety**, v. 57, n. 3, p. 410–419, mar. 2004.

DEPRÁ, Maríndia *et al.* The first records of the invasive pest Drosophila suzukii in the South American continent. **Journal of Pest Science**, v. 87, n. 3, p. 379–383, set. 2014.

DISI, Joseph Onwusemu; SIAL, Ashfaq A. Laboratory Selection and Assessment of Resistance Risk in Drosophila suzukii (Diptera: Drosophilidae) to Spinosad and Malathion. **Insects**, v. 12, n. 9, p. 794, 4 set. 2021.

DONG, Ke *et al.* Molecular biology of insect sodium channels and pyrethroid resistance. **Insect Biochemistry and Molecular Biology**, v. 50, p. 1–17, jul. 2014a.

DONG, Ke *et al.* Molecular biology of insect sodium channels and pyrethroid resistance. **Insect Biochemistry and Molecular Biology**, v. 50, p. 1–17, jul. 2014b.

DONG, Ke *et al.* Molecular biology of insect sodium channels and pyrethroid resistance. **Insect Biochemistry and Molecular Biology**, v. 50, p. 1–17, jul. 2014c.

ENGLEBIENNE, Pablo; MOITESSIER, Nicolas. Docking Ligands into Flexible and Solvated Macromolecules. 5. Force-Field-Based Prediction of Binding Affinities of Ligands to Proteins. **Journal of Chemical Information and Modeling**, v. 49, n. 11, p. 2564–2571, 23 nov. 2009.

EVANS, Richard K.; TOEWS, Michael D.; SIAL, Ashfaq A. Diel periodicity of Drosophila suzukii (Diptera: Drosophilidae) under field conditions. **PloS One**, v. 12, n. 2, p. e0171718, 2017.

FABRIS, Fabio; MAGALHÃES, João Pedro De; FREITAS, Alex A. A review of supervised machine learning applied to ageing research. **Biogerontology**, v. 18, n. 2, p. 171–188, abr. 2017.

FERREIRA, Leonardo *et al.* Molecular Docking and Structure-Based Drug Design Strategies. **Molecules**, v. 20, n. 7, p. 13384–13421, 22 jul. 2015.

FFRENCH-CONSTANT, Richard H. *et al.* Ion channels as insecticide targets. **Journal of Neurogenetics**, v. 30, n. 3–4, p. 163–177, 1 out. 2016.

FOSTER, S. P.; DEVINE, G.; DEVONSHIRE, A. L. Insecticide resistance. *In*: EMDEN, H. F. Van; HARRINGTON, R. (Orgs.). **Aphids as crop pests**. 2. ed. UK: CABI, 2017. p. 426–447.

FÜRNKRANZ, Johannes. Decision Tree. *In*: SAMMUT, Claude; WEBB, Geoffrey I. (Orgs.). **Encyclopedia of Machine Learning**. Boston, MA: Springer US, 2011. p. 263–267.

GANJISAFFAR, Fatemeh *et al.* Spatio-temporal Variation of Spinosad Susceptibility in *Drosophila suzukii* (Diptera: Drosophilidae), a Three-year Study in California's Monterey Bay Region. **Journal of Economic Entomology**, v. 115, n. 4, p. 972–980, 10 ago. 2022.

GOEL, Saurav *et al.* Diamond machining of silicon: A review of advances in molecular dynamics simulation. **International Journal of Machine Tools and Manufacture**, v. 88, p. 131–164, jan. 2015.

GOODHUE, Rachael E. *et al.* Spotted wing drosophila infestation of California strawberries and raspberries: economic analysis of potential revenue losses and control costs. **Pest Management Science**, v. 67, n. 11, p. 1396–1402, nov. 2011.

GOTTARDI, Michele *et al.* The effects of epoxiconazole and α-cypermethrin on *Daphnia magna* growth, reproduction, and offspring size. **Environmental Toxicology and Chemistry**, v. 36, n. 8, p. 2155–2166, 1 fev. 2017.

GRANDINI, Margherita; BAGLI, Enrico; VISANI, Giorgio. **Metrics for Multi-Class Classification: an Overview**. arXiv, , 2020. Disponível em: <https://arxiv.org/abs/2008.05756>. Acesso em: 27 jun. 2025

GRESS, Brian E.; ZALOM, Frank G. Identification and risk assessment of spinosad resistance in a California population of *Drosophila suzukii*. **Pest Management Science**, v. 75, n. 5, p. 1270–1276, maio 2019.

GUVENCH, Olgun; MACKERELL, Alexander D. Comparison of Protein Force Fields for Molecular Dynamics Simulations. *In*: KUKOL, Andreas (Org.). **Molecular Modeling of Proteins**. Methods in

Molecular Biology. Totowa, NJ: Humana Press, 2008. v. 443 p. 63–88.

HASSANI, I. M. *et al.* First occurrence of the pest Drosophila suzukii (Diptera: Drosophilidae) in the Comoros Archipelago (Western Indian Ocean). **African Entomology**, v. 28, n. 1, p. 78, 4 jun. 2020.

HAUSER, Martin. A historic account of the invasion of *Drosophila suzukii* (Matsumura) (Diptera: Drosophilidae) in the continental United States, with remarks on their identification. **Pest Management Science**, v. 67, n. 11, p. 1352–1357, nov. 2011.

HO, Thao N. T. *et al.* Cysteine-rich peptides: From bioactivity to bioinsecticide applications. **Toxicon**, v. 230, p. 107173, jul. 2023.

HOFFMANN, A. A. *et al.* Overwintering in Drosophila melanogaster: outdoor field cage experiments on clinal and laboratory selected populations help to elucidate traits under selection. **Journal of Evolutionary Biology**, v. 16, n. 4, p. 614–623, jul. 2003.

HOFFMANN, Krista Callinan *et al.* An analysis of lethal and sublethal interactions among type I and type II pyrethroid pesticide mixtures using standard *Hyalella azteca* water column toxicity tests. **Environmental Toxicology and Chemistry**, v. 35, n. 10, p. 2542–2549, 7 mar. 2016.

HOLLINGSWORTH, Scott A.; DROR, Ron O. Molecular Dynamics Simulation for All. **Neuron**, v. 99, n. 6, p. 1129–1143, set. 2018.

HUANG, Sheng‐You; ZOU, Xiaoqin. An iterative knowledge‐based scoring function to predict protein–ligand interactions: I. Derivation of interaction potentials. **Journal of Computational Chemistry**, v. 27, n. 15, p. 1866–1875, 30 nov. 2006.

HUBBARD, Caleb B.; MURILLO, Amy C. Behavioral resistance to insecticides: current understanding, challenges, and future directions. **Current Opinion in Insect Science**, v. 63, p. 101177, jun. 2024.

JANG, Seungbeom *et al.* Machine learning-based weld porosity detection using frequency analysis of arc sound in the pulsed gas tungsten arc welding process. **Journal of Advanced Joining Processes**, v. 10, p. 100231, nov. 2024.

JANIESCH, Christian; ZSCHECH, Patrick; HEINRICH, Kai. Machine learning and deep learning. **Electronic Markets**, v. 31, n. 3, p. 685–695, set. 2021.

JIANG, Tammy; GRADUS, Jaimie L.; ROSELLINI, Anthony J. Supervised Machine Learning: A Brief Primer. **Behavior Therapy**, v. 51, n. 5, p. 675–687, set. 2020.

JIN, Ai-Hua *et al.* Conotoxins: Chemistry and Biology. **Chemical Reviews**, v. 119, n. 21, p. 11510–11549, 13 nov. 2019.

JIN, Yin *et al.* Web repositories of natural agents promote pests and pathogenic microbes management. **Briefings in Bioinformatics**, v. 22, n. 6, p. bbab205, 5 nov. 2021.

JORGENSEN, William L.; MAXWELL, David S.; TIRADO-RIVES, Julian. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. **Journal of the American Chemical Society**, v. 118, n. 45, p. 11225–11236, 13 nov. 1996.

JUMPER, John *et al.* Highly accurate protein structure prediction with AlphaFold. **Nature**, v. 596, n. 7873, p. 583–589, 26 ago. 2021.

KALITA, Moni K.; HALOI, Kishor; DEVI, Dipali. Larval Exposure to Chlorpyrifos Affects Nutritional Physiology and Induces Genotoxicity in Silkworm Philosamia ricini (Lepidoptera:

Saturniidae). **Frontiers in Physiology**, v. 7, 15 nov. 2016.

KARPLUS, Martin; MCCAMMON, J. Andrew. Molecular dynamics simulations of biomolecules. **Nature Structural Biology**, v. 9, n. 9, p. 646–652, set. 2002.

KASIMOVA, M. A.; GRANATA, D.; CARNEVALE, V. Voltage-Gated Sodium Channels. *In*: **Current Topics in Membranes**. *[S.l.]*: Elsevier, 2016. v. 78 p. 261–286.

KASUYA, Junko *et al.* Milk-whey diet substantially suppresses seizure-like phenotypes of $para^{Shu}$, a *Drosophila* voltage-gated sodium channel mutant. **Journal of Neurogenetics**, v. 33, n. 3, p. 164–178, 3 jul. 2019.

KIRST, Herbert A. The spinosyn family of insecticides: realizing the potential of natural products research. **The Journal of Antibiotics**, v. 63, n. 3, p. 101–111, mar. 2010.

KLINT, Julie K. *et al.* Spider-venom peptides that target voltage-gated sodium channels: Pharmacological tools and potential therapeutic leads. **Toxicon**, v. 60, n. 4, p. 478–491, set. 2012.

KOVACIC, Peter. Mechanism of Organophosphates (Nerve Gases and Pesticides) and Antidotes: Electron Transfer and Oxidative Stress. **Current Medicinal Chemistry**, v. 10, n. 24, p. 2705–2709, 1 dez. 2003.

KULKARNI, Prajakta U.; SHAH, Harshil; VYAS, Vivek K. Hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) Simulation:A Tool for Structure-Based Drug Design and Discovery. **Mini-Reviews in Medicinal Chemistry**, v. 22, n. 8, p. 1096–1107, maio 2022.

KWADHA, Charles A. *et al.* Detection of the spotted wing drosophila, Drosophila suzukii, in continental sub-Saharan Africa. **Journal of Pest Science**, v. 94, n. 2, p. 251–259, mar. 2021.

LEE, Byungjo *et al.* A Deep Learning Approach with Data Augmentation to Predict Novel Spider Neurotoxic Peptides. **International Journal of Molecular Sciences**, v. 22, n. 22, p. 12291, 13 nov. 2021.

LEMAN, Julia Koehler *et al.* Macromolecular modeling and design in Rosetta: recent methods and frameworks. **Nature Methods**, v. 17, n. 7, p. 665–680, jul. 2020.

LEMKUL, Justin A. Introductory Tutorials for Simulating Protein Dynamics with GROMACS. **The Journal of Physical Chemistry B**, v. 128, n. 39, p. 9418–9435, 3 out. 2024.

LEONHART, Pablo Felipe *et al.* A biased random key genetic algorithm for the protein–ligand docking problem. **Soft Computing**, v. 23, n. 12, p. 4155–4176, jun. 2019.

LI, Zhangqiang; WU, Qiurong; YAN, Nieng. A structural atlas of druggable sites on Na_v channels. **Channels**, v. 18, n. 1, p. 2287832, 31 dez. 2024.

LIANG, Jiyun *et al.* Insect Resistance to Insecticides: Causes, Mechanisms, and Exploring Potential Solutions. **Archives of Insect Biochemistry and Physiology**, v. 118, n. 2, p. e70045, fev. 2025.

LIEBESKIND, Benjamin J.; HILLIS, David M.; ZAKON, Harold H. Evolution of sodium channels predates the origin of nervous systems in animals. **Proceedings of the National Academy of Sciences**, v. 108, n. 22, p. 9154–9159, 31 maio 2011.

LIN, Tzu-Tang *et al.* AI4AVP: an antiviral peptides predictor in deep learning approach with generative adversarial network data augmentation. **Bioinformatics Advances**, v. 2, n. 1, p. vbac080, 10 jan. 2022.

LITTLE, Catherine M.; CHAPMAN, Thomas W.; HILLIER, N. Kirk. Plasticity Is Key to Success of Drosophila suzukii (Diptera: Drosophilidae) Invasion. **Journal of Insect Science (Online)**, v. 20, n. 3, p. 5, 1 maio 2020.

LIU, Yichen; BEZANILLA, Francisco. Comparison of two fast-inactivation deficient mutants in voltage-gated sodium channel. **Biophysical Journal**, v. 123, n. 3, p. 106a–107a, fev. 2024a.

LIU, Yichen; BEZANILLA, Francisco. Comparison of two fast-inactivation deficient mutants in voltage-gated sodium channel. **Biophysical Journal**, v. 123, n. 3, p. 106a–107a, fev. 2024b.

LORENA, Ana Carolina; DE CARVALHO, André C. P. L. F. Uma Introdução às Support Vector Machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 20 dez. 2007.

LUSCOMBE, N. M.; GREENBAUM, D.; GERSTEIN, M. What is Bioinformatics? A Proposed Definition and Overview of the Field. **Methods of Information in Medicine**, v. 40, n. 04, p. 346–358, 2001.

MEDEMA, Marnix H. The year 2020 in natural product bioinformatics: an overview of the latest tools and databases. **Natural Product Reports**, v. 38, n. 2, p. 301–306, 2021.

MIGLANI, Rashi; BISHT, Satpal Singh. World of earthworms with pesticides and insecticides. **Interdisciplinary Toxicology**, v. 12, n. 2, p. 71–82, 1 out. 2019.

MUFASSIRIN, M. M. Mohamed; NEWTON, M. A. Hakim; SATTAR, Abdul. Artificial intelligence for template-free protein structure prediction: a comprehensive review. **Artificial Intelligence Review**, v. 56, n. 8, p. 7665–7732, ago. 2023.

MURRAY, Christopher W.; AUTON, Timothy R.; ELDRIDGE, Matthew D. Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian regression to improve the quality of the model. **Journal of Computer-Aided Molecular Design**, v. 12, n. 5, p. 503–519, set. 1998.

NAIK, Kunal *et al.* Current Status and Future Directions: The Application of Artificial Intelligence/Machine Learning for Precision Medicine. **Clinical Pharmacology & Therapeutics**, v. 115, n. 4, p. 673–686, abr. 2024.

NAMBIAR, Pranav; MITRA, Debirupa; DUTTA, Arnab. Machine learning assisted screening framework for insecticidal peptides. **Materials Today: Proceedings**, v. 72, p. 41–46, 2023.

NAQQASH, Muhammad Nadir *et al.* Insecticide resistance and its molecular basis in urban insect pests. **Parasitology Research**, v. 115, n. 4, p. 1363–1373, abr. 2016.

NATEKIN, Alexey; KNOLL, Alois. Gradient boosting machines, a tutorial. **Frontiers in Neurorobotics**, v. 7, 2013.

PAIVA, Vinícius De Almeida *et al.* Protein structural bioinformatics: An overview. **Computers in Biology and Medicine**, v. 147, p. 105695, ago. 2022.

PARACAMPO, Ariel *et al.* Acute toxicity of chlorpyrifos to the non-target organism *Cnesterodon decemmaculatus*. **International Journal of Environmental Health Research**, v. 25, n. 1, p. 96–103, 2 jan. 2015.

PATODIA, Sachin. Molecular Dynamics Simulation of Proteins: A Brief Overview. **Journal of Physical Chemistry & Biophysics**, v. 4, n. 6, 2014.

PLISSON, Fabien; RAMÍREZ-SÁNCHEZ, Obed; MARTÍNEZ-HERNÁNDEZ, Cristina. Machine

learning-guided discovery and design of non-hemolytic peptides. **Scientific Reports**, v. 10, n. 1, p. 16581, 6 out. 2020.

PU, Jian; WANG, Zinan; CHUNG, Henry. Climate change and the genetics of insecticide resistance. **Pest Management Science**, v. 76, n. 3, p. 846–852, mar. 2020.

RADOVANOVIĆ, Tijana B. *et al.* Sublethal effects of the pyrethroid insecticide deltamethrin on oxidative stress parameters in green toad (*Bufotes viridis* L.). **Environmental Toxicology and Chemistry**, v. 36, n. 10, p. 2814–2822, 5 maio 2017.

RENDON, Dalila *et al.* Interactions among morphotype, nutrition, and temperature impact fitness of an invasive fly. **Ecology and Evolution**, v. 9, n. 5, p. 2615–2628, mar. 2019.

RESENDE, Paulo Angelo Alves; DRUMMOND, André Costa. A Survey of Random Forest Based Methods for Intrusion Detection Systems. **ACM Computing Surveys**, v. 51, n. 3, p. 1–36, 31 maio 2019.

REZENDE-TEIXEIRA, Paula *et al.* What can we learn from commercial insecticides? Efficacy, toxicity, environmental impacts, and future developments. **Environmental Pollution**, v. 300, p. 118983, maio 2022.

RICHARDSON, Jason R. *et al.* Neurotoxicity of pesticides. **Acta Neuropathologica**, v. 138, n. 3, p. 343–362, set. 2019.

RINIKER, Sereina. Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview. **Journal of Chemical Information and Modeling**, v. 58, n. 3, p. 565–578, 26 mar. 2018.

ROSSI-STACCONI, Marco Valerio *et al.* Multiple lines of evidence for reproductive winter diapause in the invasive pest Drosophila suzukii: useful clues for control strategies. **Journal of Pest Science**, v. 89, n. 3, p. 689–700, jul. 2016.

ROTA-STABELLI, Omar *et al.* Distinct genotypes and phenotypes in European and American strains of Drosophila suzukii: implications for biology and management of an invasive organism. **Journal of Pest Science**, v. 93, n. 1, p. 77–89, jan. 2020.

SARKER, Iqbal H. *et al.* Cybersecurity data science: an overview from machine learning perspective. **Journal of Big Data**, v. 7, n. 1, p. 41, dez. 2020.

SARKER, Iqbal H. Machine Learning: Algorithms, Real-World Applications and Research Directions. **SN Computer Science**, v. 2, n. 3, p. 160, maio 2021.

SENIOR SOFTWARE ENGINEERING, MICROSOFT, NORTHLAKE, TEXAS, USA.; KANAPARTHI, Vijaya. Transformational Application of Artificial Intelligence and Machine Learning in Financial Technologies and Financial Services: A Bibliometric Review. **International Journal of Engineering and Advanced Technology**, v. 13, n. 3, p. 71–77, 28 fev. 2024.

SHAIK, Anjaneyulu Babu; SRINIVASAN, Sujatha. A Brief Survey on Random Forest Ensembles in Classification Model. *In*: BHATTACHARYYA, Siddhartha *et al.* (Orgs.). **International Conference on Innovative Computing and Communications**. Lecture Notes in Networks and Systems. Singapore: Springer Singapore, 2019. v. 56 p. 253–260.

SHARMA, Arun *et al.* dPABBs: A Novel in silico Approach for Predicting and Designing Anti-biofilm Peptides. **Scientific Reports**, v. 6, n. 1, p. 21839, 25 fev. 2016.

SHASTRY, K. Aditya; SANJAY, H. A. Machine Learning for Bioinformatics. *In*: SRINIVASA, K. G.;

SIDDESH, G. M.; MANISEKHAR, S. R. (Orgs.). **Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications**. Algorithms for Intelligent Systems. Singapore: Springer Singapore, 2020. p. 25–39.

SHEARER, Peter W. *et al.* Seasonal cues induce phenotypic plasticity of Drosophila suzukii to enhance winter survival. **BMC ecology**, v. 16, p. 11, 22 mar. 2016.

SHEN, Huaizong *et al.* Structural basis for the modulation of voltage-gated sodium channels by animal toxins. **Science**, v. 362, n. 6412, p. eaau2596, 19 out. 2018.

SILVA, Juan J.; SCOTT, Jeffrey G. Conservation of the voltage‐sensitive sodium channel protein within the Insecta. **Insect Molecular Biology**, v. 29, n. 1, p. 9–18, fev. 2020.

SMIRLE, Michael J. *et al.* Laboratory studies of insecticide efficacy and resistance in *Drosophila suzukii* (Matsumura) (Diptera: Drosophilidae) populations from British Columbia, Canada: The susceptibility of *D. suzukii* populations from BC to selected insecticides. **Pest Management Science**, v. 73, n. 1, p. 130–137, jan. 2017.

SMYTH, M. S. x Ray crystallography. **Molecular Pathology**, v. 53, n. 1, p. 8–14, 1 fev. 2000.

SOUSA, S. F. *et al.* Protein-Ligand Docking in the New Millennium – A Retrospective of 10 Years in the Field. **Current Medicinal Chemistry**, v. 20, n. 18, p. 2296–2314, 1 abr. 2013.

SPARKS, Thomas C. *et al.* Insecticides, biologics and nematicides: Updates to IRAC's mode of action classification - a tool for resistance management. **Pesticide Biochemistry and Physiology**, v. 167, p. 104587, jul. 2020.

STOCKTON, Dara G.; BROWN, Rachael; LOEB, Gregory M. Not berry hungry? Discovering the hidden food sources of a small fruit specialist, *Drosophila suzukii*. **Ecological Entomology**, v. 44, n. 6, p. 810–822, dez. 2019.

SULIMAN *et al.* Toxicity evaluation of pesticide chlorpyrifos in male Japanese quails (Coturnix japonica). **Environmental Science and Pollution Research**, v. 27, n. 20, p. 25353–25362, jul. 2020.

SUNGHEETHA, Akey. 3D Image Processing using Machine Learning based Input Processing for Man-Machine Interaction. **Journal of Innovative Image Processing**, v. 3, n. 1, p. 1–6, 22 fev. 2021.

SUTHAHARAN, Shan. Decision Tree Learning. *In*: SUTHAHARAN, Shan (Ed.). **Machine Learning Models and Algorithms for Big Data Classification**. Integrated Series in Information Systems. Boston, MA: Springer US, 2016. v. 36 p. 237–269.

TAM, Nguyen Thanh; BERG, Håkan; VAN CONG, Nguyen. Evaluation of the joint toxicity of chlorpyrifos ethyl and fenobucarb on climbing perch (Anabas testudineus) from rice fields in the Mekong Delta, Vietnam. **Environmental Science and Pollution Research**, v. 25, n. 14, p. 13226–13234, maio 2018.

TIAN, Chuan *et al.* ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. **Journal of Chemical Theory and Computation**, v. 16, n. 1, p. 528–552, 14 jan. 2020.

TOXOPEUS, Jantina *et al.* Reproductive arrest and stress resistance in winter-acclimated Drosophila suzukii. **Journal of Insect Physiology**, v. 89, p. 37–51, jun. 2016.

TU, Wenqing *et al.* Permethrin is a potential thyroid-disrupting chemical: In vivo and in silico envidence. **Aquatic Toxicology**, v. 175, p. 39–46, jun. 2016.

UGURLU, Sadettin Y. **Machine Learning Applications in Drug Discovery**. Chemistry, , 26 set. 2024. Disponível em: <https://chemrxiv.org/engage/chemrxiv/article-details/66f5321812ff75c3a18553fd>. Acesso em: 27 jun. 2025

VAN GUNSTEREN, Wilfred F.; DAURA, Xavier; MARK, Alan E. GROMOS Force Field. *In*: VON RAGUÉ SCHLEYER, Paul *et al.* (Orgs.). **Encyclopedia of Computational Chemistry**. 1. ed. *[S.l.]*: Wiley, 1998.

VARDAVAS, Alexander I. *et al.* Long-term exposure to cypermethrin and piperonyl butoxide cause liver and kidney inflammation and induce genotoxicity in New Zealand white male rabbits. **Food and Chemical Toxicology**, v. 94, p. 250–259, ago. 2016.

VIANA, José Pedro Cavalcante *et al.* Establishment and Expansion Scenario of Drosophila suzukii (Diptera: Drosophilidae) in Central Brazil. **Neotropical Entomology**, v. 52, n. 6, p. 975–985, 1 maio 2023.

VILELA, Carlos Ribeiro; MORI, Lyria. The invasive spotted-wing Drosophila (Diptera, Drosophilidae) has been found in the city of São Paulo (Brazil). **Revista Brasileira de Entomologia**, v. 58, n. 4, p. 371–375, dez. 2014.

VLACHAKIS, Dimitrios *et al.* Current State-of-the-Art Molecular Dynamics Methods and Applications. *In*: **Advances in Protein Chemistry and Structural Biology**. *[S.l.]*: Elsevier, 2014. v. 94 p. 269–313.

WALSH, Douglas B. *et al.* Drosophila suzukii (Diptera: Drosophilidae): Invasive Pest of Ripening Soft Fruit Expanding its Geographic Range and Damage Potential. **Journal of Integrated Pest Management**, v. 2, n. 1, p. G1–G7, 1 abr. 2011.

WANG, Guangshun; VAISMAN, Iosif I.; VAN HOEK, Monique L. Machine Learning Prediction of Antimicrobial Peptides. *In*: SIMONSON, Thomas (Org.). **Computational Peptide Science**. Methods in Molecular Biology. New York, NY: Springer US, 2022. v. 2405 p. 1–37.

WANG, Lingxin *et al.* Distinct modulating effects of TipE-homologs 2–4 on Drosophila sodium channel splice variants. **Insect Biochemistry and Molecular Biology**, v. 60, p. 24–32, maio 2015.

WANG, Zhe *et al.* Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. **Physical Chemistry Chemical Physics**, v. 18, n. 18, p. 12964–12975, 2016.

WANI, Aasim Ayaz. Advancing Material Stability Prediction: Leveraging Machine Learning and High-Dimensional Data for Improved Accuracy. **Materials Sciences and Applications**, v. 16, n. 02, p. 79–105, 2025.

WEBB, Benjamin; SALI, Andrej. Protein Structure Modeling with MODELLER. *In*: KAUFMANN, Michael; KLINGER, Claudia; SAVELSBERGH, Andreas (Orgs.). **Functional Genomics**. Methods in Molecular Biology. New York, NY: Springer New York, 2017. v. 1654 p. 39–54.

WENG, Gaoqi *et al.* Comprehensive Evaluation of Fourteen Docking Programs on Protein–Peptide Complexes. **Journal of Chemical Theory and Computation**, v. 16, n. 6, p. 3959–3969, 9 jun. 2020.

WU, Zhiyong *et al.* Microbial production of small peptide: pathway engineering and synthetic biology. **Microbial Biotechnology**, v. 14, n. 6, p. 2257–2278, nov. 2021.

WÜTHRICH, K. Protein structure determination in solution by NMR spectroscopy. **Journal of Biological Chemistry**, v. 265, n. 36, p. 22059–22062, dez. 1990.

XU, Dong; ZHANG, Yang. *Ab initio* protein structure assembly using continuous structure fragments and optimized knowledge‐based force field. **Proteins: Structure, Function, and Bioinformatics**, v. 80, n. 7, p. 1715–1735, jul. 2012.

XU, Jiliang *et al.* Generative Adversarial Network-Based Data Augmentation Method for Anti-coronavirus Peptides Prediction. *In*: HUANG, De-Shuang *et al.* (Orgs.). **Advanced Intelligent Computing Technology and Applications**. Lecture Notes in Computer Science. Singapore: Springer Nature Singapore, 2023. v. 14088 p. 67–76.

YE, Xiaoqing *et al.* Pyrethroid Insecticide Cypermethrin Accelerates Pubertal Onset in Male Mice via Disrupting Hypothalamic–Pituitary–Gonadal Axis. **Environmental Science & Technology**, v. 51, n. 17, p. 10212–10221, 5 set. 2017.

YIP, Ka Man *et al.* Atomic-resolution protein structure determination by cryo-EM. **Nature**, v. 587, n. 7832, p. 157–161, 5 nov. 2020.

YU, Sungduk *et al.* Two-Step Hyperparameter Optimization Method: Accelerating Hyperparameter Search by Using a Fraction of a Training Dataset. **Artificial Intelligence for the Earth Systems**, v. 3, n. 1, p. e230013, jan. 2024.

YUAN, Linlin *et al.* Functional diversity of voltage‐gated sodium channel in *Drosophila suzukii* (Matsumura). **Pest Management Science**, v. 80, n. 2, p. 592–601, fev. 2024.

ZAKON, Harold H. Adaptive evolution of voltage-gated sodium channels: The first 800 million years. **Proceedings of the National Academy of Sciences**, v. 109, n. supplement_1, p. 10619–10625, 26 jun. 2012.

ZALUCKI, Mp; FURLONG, Mj. Behavior as a mechanism of insecticide resistance: evaluation of the evidence. **Current Opinion in Insect Science**, v. 21, p. 19–25, jun. 2017.

ZHANG, Jieyu *et al.* **Binary Classification with Positive Labeling Sources**. arXiv, , 2022. Disponível em: <https://arxiv.org/abs/2208.01704>. Acesso em: 27 jun. 2025

ZHANG, Yang. I-TASSER server for protein 3D structure prediction. **BMC Bioinformatics**, v. 9, n. 1, p. 40, dez. 2008.

ZHOU, Jin *et al.* A novel molecular docking program based on a multi-swarm competitive algorithm. **Swarm and Evolutionary Computation**, v. 78, p. 101292, abr. 2023.

ZHU, Qiuyan *et al.* Synthesis, insecticidal activity, resistance, photodegradation and toxicity of pyrethroids (A review). **Chemosphere**, v. 254, p. 126779, set. 2020.

7. APÊNDICE - Material suplementar

Supplementary Data

A Machine Learning framework for predicting peptide inhibitors of insect voltage-gated sodium channels

Jailan da Silva Sousa[a,b], Lucas Sousa Palmeira[a,b], Fabrício Santos Barbosa[b], Hugo Mauricio Peña Mercado[a], Tarcisio Silva Melo[b], Vasco Azevedo[a], Aristóteles Góes-Neto[a,c], Joicymara Santos Xavier[a,d,c], Bruno Silva Andrade[b]*

[a]Graduate Program in Bioinformatics, Federal University of Minas Gerais, Belo Horizonte 31270-901, MG, Brasil.

[b]Laboratory of Bioinformatics and Computational Chemistry, State University of Southwest Bahia, Jequié, BA, Brazil.

[c]Department of Microbiology, Molecular and Computational Biology of Fungi Laboratory, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, MG, Brazil.

[d]Computer Science Division, Technological Institute of Aeronautics, São José dos Campos 89760-000, SP, Brazil.

Methodology

Sequence representation methods

Table S1. Calculated descriptors for feature extraction from amino acid sequences in the training dataset.

IFeature	
Amino Acid Composition	AAC, DPC, DDE, TPC[1,2]
Grouped Amino Acid Composition	GAAC, GDPC, GTPC[3]
Sequence Order	SOCNumber, QSOrder[4–6]
Pseudo-Amino Acid Composition	PAAC, APAAC[7,8]
Composition/Transition/Distribution	CTDC, CTDT, CTDD[9,10]
Autocorrelation	Moran, Geary[11]
Pseudo K-tuple reduced amino acids composition	PseKRAAC[12]
modlAMP[13,14]	
charge, charge density, aliphatic index, boman index, and hydrophobic ratio.	
	Amino acid descriptor scales

Mean hydrophobic moment; hydrophobic profile	"ABHPRK", "AASI", "bulkiness", "charge_phys", "charge_acid", "cougar", "eisenberg", "Ez", "flexibility", "hopp-woods", "ISAECI", "janin", "kytedoolittle", "levitt_alpha", "MSS", "MSW", "pepArc", "pepcats", "polarity", "PPCALI", "refractivity", "t_scale", "TM_tend", "z3", "z5"
Maximum hydrophobic moment	"AASI", "bulkiness", "charge_phys", "charge_acid", "eisenberg", "flexibility", "hopp-woods", "janin", "kytedoolittle", "levitt_alpha", "MSS", "polarity", "refractivity", "TM_tend"
Hydrophobic moment profile	"AASI", "bulkiness", "charge_phys", "charge_acid", "eisenberg", "flexibility", "hopp-woods", "janin", "kytedoolittle", "levitt_alpha", "MSS", "polarity","refractivity", "TM_tend"
Maximum property arcs	"ABHPRK", "charge_phys", "charge_acid", "pepArc", "pepcats"
Biopython	
molecular weight, aromaticity, instability index, isoelectric point, secondary structure fraction, molar extinction coefficient, gravy[15]	
Total combined dimensions	38194

Machine learning algorithms

Decision Tree algorithms classify data by recursively partitioning the dataset into subsets based on threshold values of predictor variables to optimize a splitting criterion, such as Gini impurity or information gain [16,17]. At each node, the algorithm selects the feature and corresponding threshold that most effectively separate the data into more homogeneous groups, producing a hierarchical tree structure in which each terminal node (leaf) represents a class label. Although this structure is easily interpretable and capable of modeling non-linear decision boundaries, it is susceptible to overfitting.

A Random Forest is an ensemble learning method that constructs a collection of decision trees, each trained on a different bootstrap sample (sampling with replacement) of the original dataset [18]. During the training of each tree, a random subset of features is selected at each split, introducing additional variability and reducing correlation among trees [19]. This procedure enhances model generalization and mitigates overfitting relative to a single decision tree. The final classification is obtained by aggregating the predictions of all individual trees through majority voting.

Support Vector Machines (SVMs) perform classification by identifying the optimal hyperplane that maximizes the margin separating data points from distinct classes [20]. When the data are not linearly separable, kernel functions are employed to map the data into higher-dimensional feature spaces where a linear separation becomes feasible.

Gradient Boosting Models (GBMs), including widely used implementations such as XGBoost (eXtreme Gradient Boosting) and LightGBM (Light Gradient Boosting Machine), also rely on decision trees but differ from Random Forests in their sequential learning strategy [21,22]. Trees are added iteratively, with each new tree trained to minimize the residual errors of the ensemble using gradient descent on a predefined loss function. This approach yields highly accurate models capable of capturing complex data patterns, although careful hyperparameter tuning is often required to prevent overfitting.

Hyperparameter search

Table S2. Hyperparameter optimization focused on parameters that control the model's fit to the training data.

models	default	search space	final parameter
Decision Tree			
max_depth	None	2-4	3
min_samples_split	2	2-4	2
min_samples_leaf	1	2-4	2
ccp_alpha	0.0	0-1	0.0388
Random Forest			
n_estimators	100	100-300	300
max_depth	None	2-4	4
min_samples_split	2	2-4	2
min_samples_leaf	1	1-4	1
ccp_alpha	0	0-1	0.0194
Gradient Boosting			
learning_rate	0.1	0.001-0.005	0.0042
n_estimators	100	100-300	150
max_depth	3	2-4	4
min_samples_split	2	2-4	4
min_samples_leaf	1	1-4	4
subsample	1	0-1	0.3311
LightGBM			
learning_rate	0.1	0.001-0.005	0.0046
n_estimators	100	100-300	300
max_depth	-1	2-4	4
num_leaves	31	10-31	30
min_data_in_leaf	20	5-20	6
subsample	1	0-1	0.6086
colsample_bytree	1	0-1	0.0167
reg_alpha	0	0-1	0.0362
reg_lambda	0	0-1	0.2310
XGBoost			
n_estimators	100	100-300	300
max_depth	6	2-4	3
learning_rate	0.3	0.001-0.005	0.0027
subsample	1	0.5-1.0	0.5960
colsample_bytree	1	0.5-1.0	0.7516
gamma	0	0.5-1.0	0.9397
SVC*			
C	1.0	0.0001-0.0005	0.0001794
degree	3	1-5	5
kernel	rbf	linear; rbf	linear
probability	False	True	True

SVC*: Support Vector Classifier.

Parameters used in the simulation of the voltage-gated sodium channel (VGSC) from Drosophila suzukii

The time step adopted was 2 fs, with the leap-frog integrator. Short-range unbound interactions were treated with the Verlet cutoff scheme, with a cutoff radius of 1.2 nm. Lennard-Jones interactions were smoothly switched to zero between 1.0 nm and 1.2 nm, using a force-switch modifier. Long-range electrostatic interactions were treated using the Particle Mesh Ewald (PME) method, with a real-space cutoff of 1.2 nm. The temperature was kept constant at 298.15 K using the velocity-rescale thermostat, with a coupling constant of 1.0 ps. The system was partitioned into three thermal coupling groups: protein, membrane, and

solvent. Pressure was controlled semi-isotropically at 1.0 bar using the C-rescale barostat, with a coupling constant of 5.0 ps and compressibility of 4.5×10^{-5} bar^{-1} in the xy planes and z axis. All bonds involving hydrogen atoms were constrained with the LINCS algorithm, allowing the use of a 2 fs integration step. Periodic boundary conditions were applied in all directions. The center of mass motion was removed every 100 steps using a linear scheme, with independent groups for removing the center of mass motion of the solute-membrane complex and the solvent.

Binding free energy calculation

Binding free energy calculations were performed with the gmx_MMPBSA tool 1.6.4[23], considering protein–peptide complexes incorporated into an implicit membrane bilayer, under physiological salt concentration (0.15 M). The PB radii set specific for the CHARMM force field were used (PBRadii = 7). An implicit membrane model was applied, with a heterogeneous dielectric plate along the bilayer normal. The free energy components included the Van der Waals and electrostatic energies (E_MM), the polar solvation contribution (G_PB), and the nonpolar solvation contribution (G_SA), calculated from 800 sampled frames. The final binding free energy results (ΔG_bind) and their conclusion were calculated as means, with the respective standard deviations calculated over the set, and analyzed with the gmx_MMPBSA_ana program. For these calculations, only the interaction regions between the peptides and DsNav were considered, disregarding the first 20 ns of the simulation, to make computational processing viable, since the inclusion of the lipid bilayer in the estimates contributes significantly to RAM consumption[24].

Results

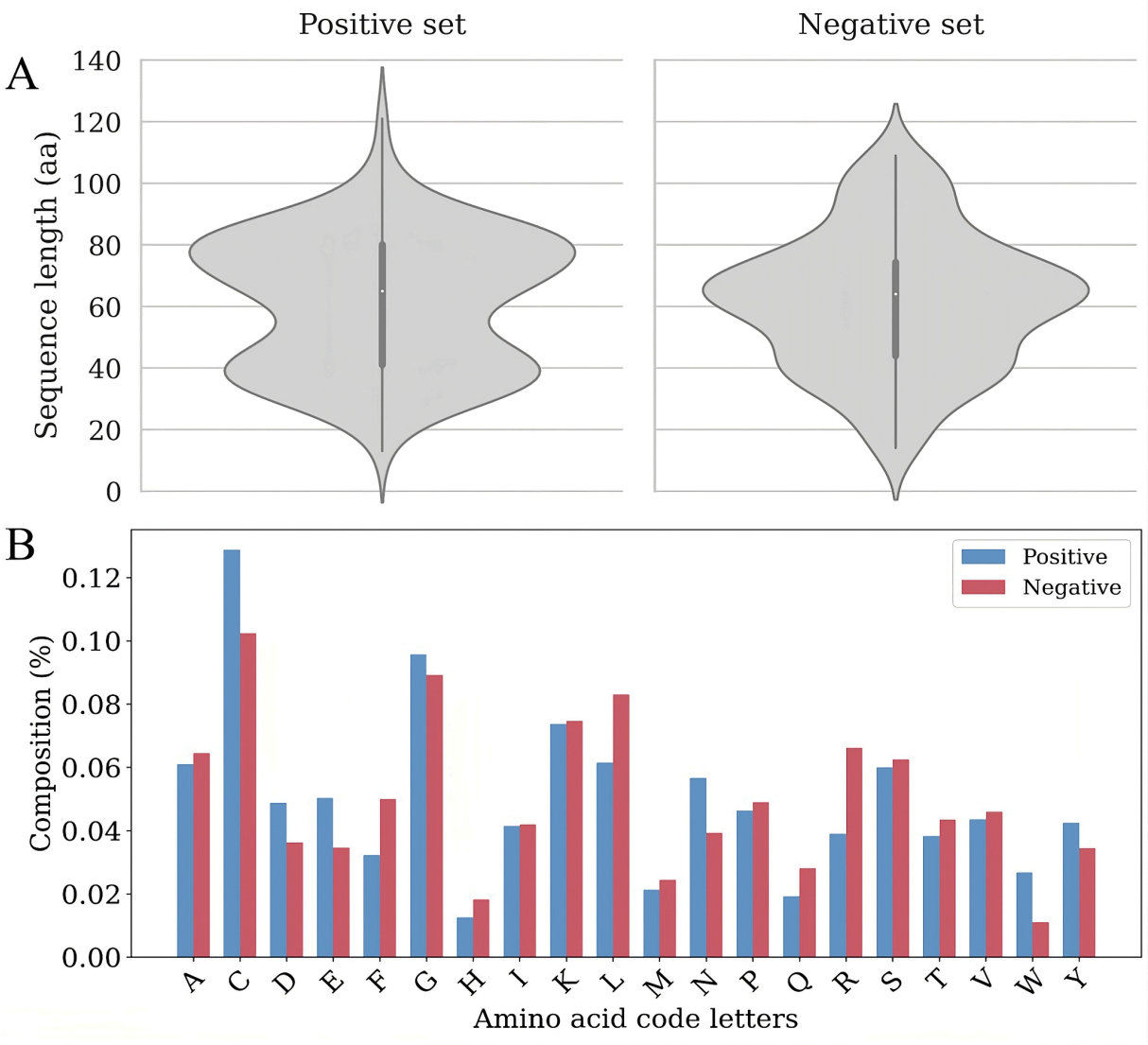

Figure S1. Sequence length distribution and amino acid composition of the positive and negative peptide datasets. (A) Violin plots showing the distribution of peptide sequence lengths in the training dataset. (B) Mean amino acid composition (%) for each residue type in positive and negative datasets.

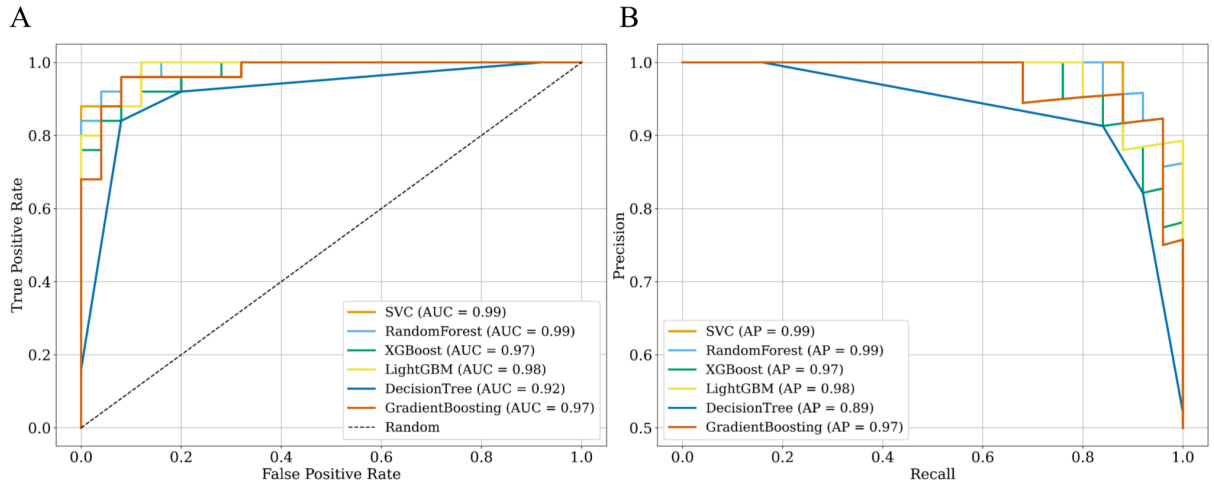

Figure S2. Performance comparison of machine learning models for peptide classification on the test set. (A) Receiver Operating Characteristic (ROC) curves showing the True Positive

Rate versus False Positive Rate and the area under the curve (AUC) for each model. (B) Precision–Recall (PR) curves displaying model precision as a function of recall, with the average precision (AP).

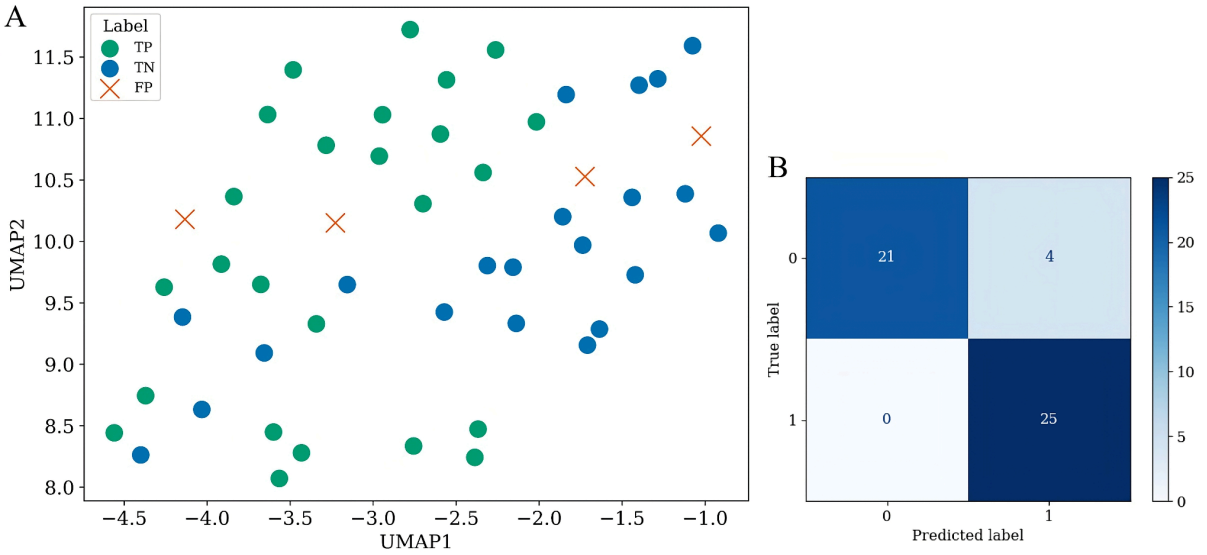

Figure S3. Visualization of model predictions on the test set. (A) UMAP projection of the feature space obtained from RFECV-selected features, showing clustering of true positives (TP), true negatives (TN), and false positives (FP). (B) Confusion matrix illustrating the classification performance of the SVC model on the test set, showing four false positives and no false negatives.

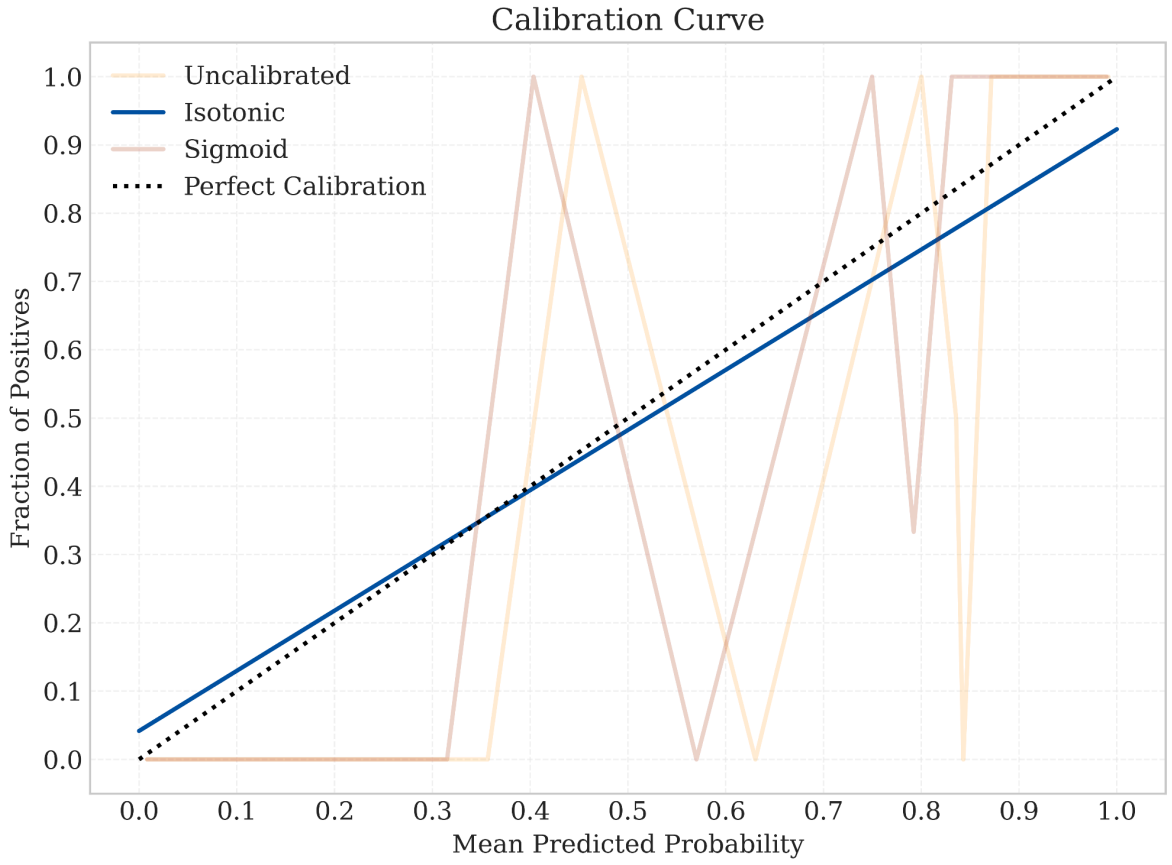

Figure S4. Calibration curves comparing uncalibrated, isotonic, and sigmoid-calibrated SVC models. The isotonic calibration shows the closest alignment with the perfect calibration line, indicating more reliable probability estimates.

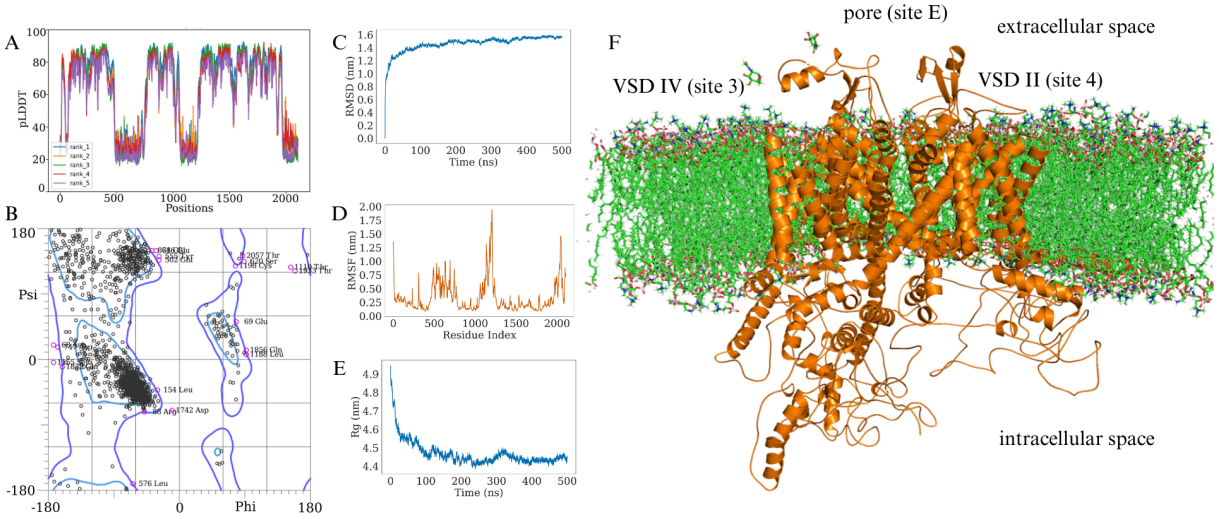

Figure S5. DsNav model quality assessment and validation. (A) AlphaFold2 pLDDT scores indicate confidence levels in the predicted protein structure. (B) Ramachandran plot showing amino acid residues in favorable and allowed regions after a 500-nanosecond molecular dynamics simulation in explicit solvent. (C) RMSD plot of the protein backbone following least-squares fitting. (D) RMSF plot of the protein backbone residues. (E) The radius of gyration of the protein throughout the simulation. (F) Final Frame of DsNav-membrane complex after 500 nanoseconds simulation.

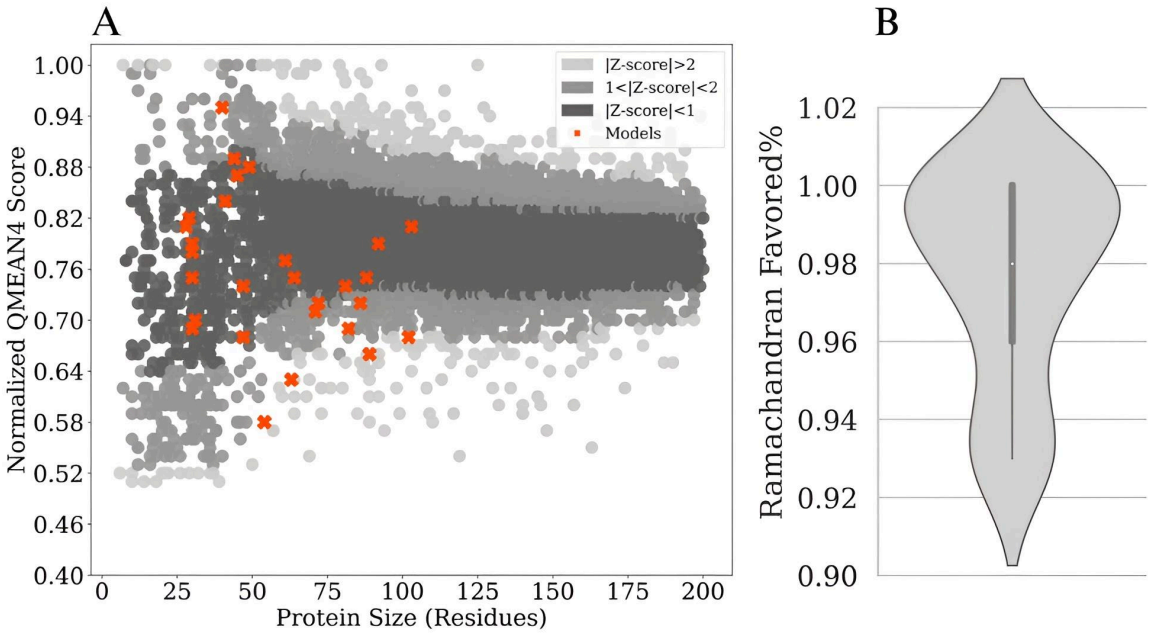

Figure S6. Plant peptides model quality. (A) QMEAN comparison of the PDB reference dataset of high-quality resolved structures with X-ray method (Resolution ≤ 3 Angstroms) and peptides structure resolved by Alphafold2. (B) Percentage of plant peptides with amino acid residues in favorable positions on the Ramachandran plot.

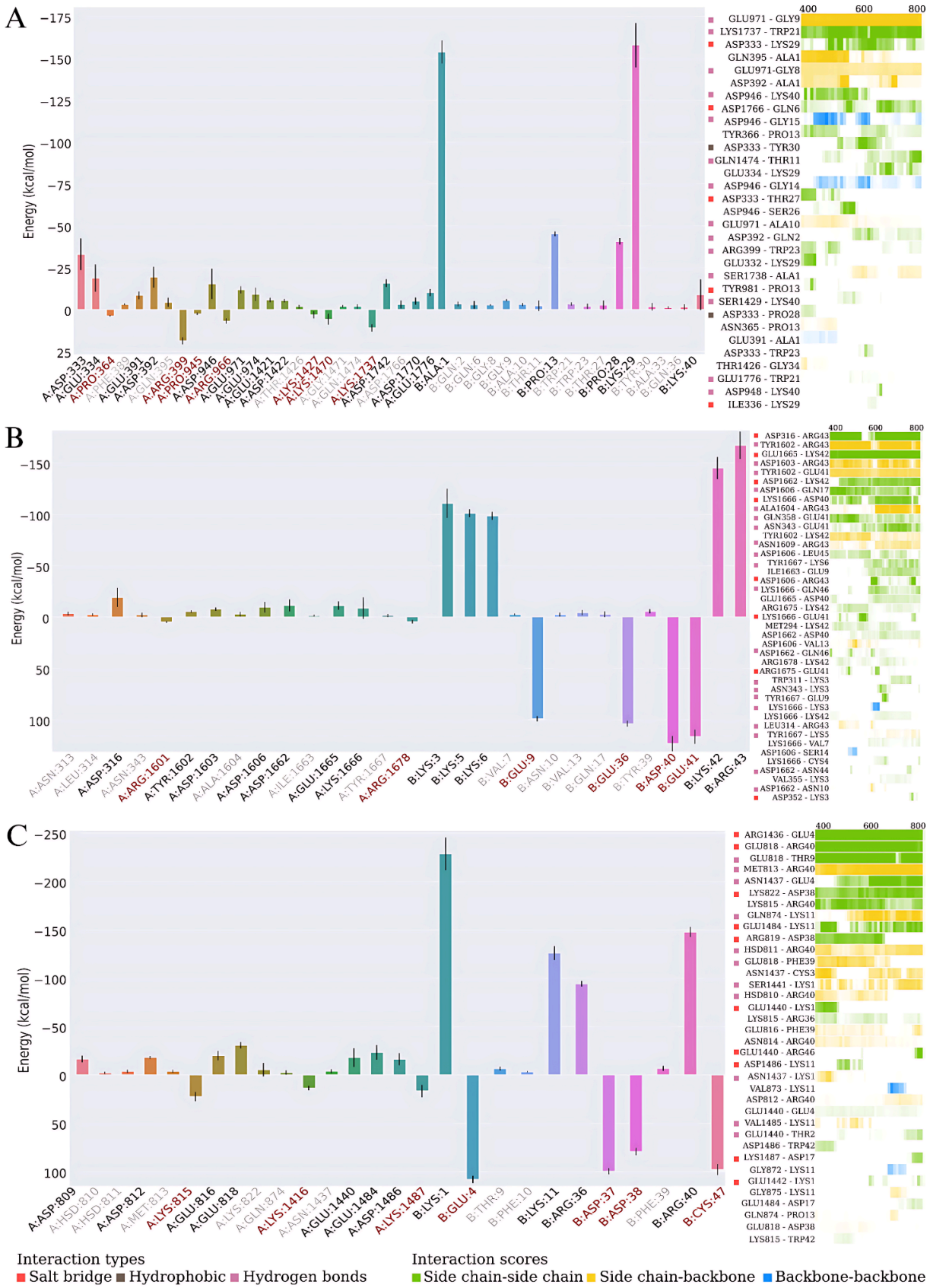
A
Energy (kcal/mol)
−175
−150
−125
−100
−75
−50
−25
0
25
A:ASP:333
A:GLU:334
A:PRO:364
A:GLU:391
A:ASP:392
A:ARG:399
A:PRO:945
A:ASP:946
A:ARG:966
A:GLU:971
A:GLU:974
A:GLU:1421
A:ASP:1422
A:LYS:1427
A:LYS:1470
A:LYS:1737
A:ASP:1742
A:ASP:1770
A:GLU:1776
B:ALA:1
B:PRO:13
B:PRO:28
B:LYS:29
B:LYS:40
400
600
800
GLU971 - GLY9
LYS1737 - TRP21
ASP333 - LYS29
GLN395 - ALA1
GLU971-GLY8
ASP392 - ALA1
ASP946 - LYS40
ASP1766 - GLN6
ASP946 - GLY15
TYR366 - PRO13
ASP333 - TYR30
GLN1474 - THR11
GLU334 - LYS29
ASP946 - GLY14
ASP333 - THR27
ASP946 - SER26
GLU971 - ALA10
ASP392 - GLN2
ARG399 - TRP23
GLU332 - LYS29
SER1738 - ALA1
TYR981 - PRO13
SER1429 - LYS40
ASP333 - PRO28
ASN365 - PRO13
GLU391 - ALA1
ASP333 - TRP23
THR1426 - GLY34
GLU1776 - TRP21
ASP948 - LYS40
ILE336 - LYS29
B
Energy (kcal/mol)
−150
−100
−50
0
50
100
A:ASN:313
A:LEU:314
A:ASP:316
A:ASN:343
A:ARG:1601
A:TYR:1602
A:ASP:1603
A:ALA:1604
A:ASP:1606
A:ASP:1662
A:ILE:1663
A:GLU:1665
A:LYS:1666
A:TYR:1667
A:ARG:1678
B:LYS:3
B:LYS:5
B:LYS:6
B:VAL:7
B:GLU:9
B:ASN:10
B:VAL:13
B:GLN:17
B:GLU:36
B:TYR:39
B:ASP:40
B:GLU:41
B:LYS:42
B:ARG:43
400
600
800
ASP316 - ARG43
TYR1602 - ARG43
GLU1665 - LYS42
ASP1603 - ARG43
TYR1602 - GLU41
ASP1662 - LYS42
ASP1606 - GLN17
LYS1666 - ASP40
ALA1604 - ARG43
GLN358 - GLU41
ASN343 - GLU41
TYR1602 - LYS42
ASN1609 - ARG43
ASP1606 - LEU45
TYR1667 - LYS6
ILE1663 - GLU9
ASP1606 - ARG43
LYS1666 - GLN46
GLU1665 - ASP40
ARG1675 - LYS42
LYS1666 - GLU41
MET294 - LYS42
ASP1662 - ASP40
ASP1606 - VAL13
ASP1662 - GLN46
ARG1678 - LYS42
ARG1675 - GLU41
TRP311 - LYS3
ASN343 - LYS3
TYR1667 - GLU9
LYS1666 - LYS3
LYS1666 - LYS42
LEU314 - ARG43
TYR1667 - LYS5
LYS1666 - VAL7
ASP1606 - SER14
LYS1666 - CYS4
ASP1662 - ASN44
VAL355 - LYS3
ASP1662 - ASN10
ASP352 - LYS3
C
Energy (kcal/mol)
−250
−200
−150
−100
−50
0
50
100
A:ASP:809
A:HSD:810
A:HSD:811
A:ASP:812
A:MET:813
A:LYS:815
A:GLU:816
A:GLU:818
A:LYS:822
A:GLN:874
A:LYS:1416
A:ASN:1437
A:GLU:1440
A:GLU:1484
A:ASP:1486
A:LYS:1487
B:LYS:1
B:GLU:4
B:THR:9
B:PHE:10
B:LYS:11
B:ARG:36
B:ASP:37
B:ASP:38
B:PHE:39
B:ARG:40
B:CYS:47
400
600
800
ARG1436 - GLU4
GLU818 - ARG40
GLU818 - THR9
MET813 - ARG40
ASN1437 - GLU4
LYS822 - ASP38
LYS815 - ARG40
GLN874 - LYS11
GLU1484 - LYS11
ARG819 - ASP38
HSD811 - ARG40
GLU818 - PHE39
ASN1437 - CYS3
SER1441 - LYS1
HSD810 - ARG40
GLU1440 - LYS1
LYS815 - ARG36
GLU816 - PHE39
ASN814 - ARG40
GLU1440 - ARG46
ASP1486 - LYS11
ASN1437 - LYS1
VAL873 - LYS11
ASP812 - ARG40
GLU1440 - GLU4
VAL1485 - LYS11
GLU1440 - THR2
ASP1486 - TRP42
LYS1487 - ASP17
GLY872 - LYS11
GLU1442 - LYS1
GLY875 - LYS11
GLU1484 - ASP17
GLN874 - PRO13
GLU818 - ASP38
LYS815 - TRP42
Interaction types
Salt bridge
Hydrophobic
Hydrogen bonds
Interaction scores
Side chain-side chain
Side chain-backbone
Backbone-backbone

Figure S7. Per-residue interaction profiles of DsNav–plant peptide complexes over 100-nanosecond molecular dynamics simulations. (A) Key interactions in the DsNav–P0DKH7 complex. (B) Key interactions in the DsNav–P56552 complex. (C) Key interactions in the DsNav–P81930 complex. Color intensity indicates the frequency and strength of residue interactions throughout the simulation.

References

[1] Bhasin M, Raghava GPS. Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition. J Biol Chem 2004;279:23262–6. https://doi.org/10.1074/jbc.M401932200.

[2] Saravanan V, Gautham N. Harnessing Computational Biology for Exact Linear B-Cell Epitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor. OMICS J Integr Biol 2015;19:648–58. https://doi.org/10.1089/omi.2015.0095.

[3] Lee T-Y, Lin Z-Q, Hsieh S-J, Bretaña NA, Lu C-T. Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences. Bioinformatics 2011;27:1780–7. https://doi.org/10.1093/bioinformatics/btr291.

[4] Ong SA, Lin HH, Chen YZ, Li ZR, Cao Z. Efficacy of different protein descriptors in predicting protein functional families. BMC Bioinformatics 2007;8:300. https://doi.org/10.1186/1471-2105-8-300.

[5] Grantham R. Amino Acid Difference Formula to Help Explain Protein Evolution. Science 1974;185:862–4. https://doi.org/10.1126/science.185.4154.862.

[6] Schneider G, Wrede P. The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. Biophys J 1994;66:335–44. https://doi.org/10.1016/S0006-3495(94)80782-9.

[7] Chou K. Prediction of protein cellular attributes using pseudo‐amino acid composition. Proteins Struct Funct Bioinforma 2001;43:246–55. https://doi.org/10.1002/prot.1035.

[8] Chou K-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 2005;21:10–9. https://doi.org/10.1093/bioinformatics/bth466.

[9] Dubchak I, Muchnik I, Holbrook SR, Kim SH. Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci 1995;92:8700–4. https://doi.org/10.1073/pnas.92.19.8700.

[10] Dubchak I, Muchnik I, Mayor C, Dralyuk I, Kim SH. Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. Proteins 1999;35:401–7.

[11] Sokal RR, Thomson BA. Population structure inferred by local spatial autocorrelation: An example from an Amerindian tribal population. Am J Phys Anthropol 2006;129:121–31. https://doi.org/10.1002/ajpa.20250.

[12] Li J, Wang W. Grouping of amino acids and recognition of protein structurally conserved regions by reduced alphabets of amino acids. Sci China C Life Sci 2007;50:392–402. https://doi.org/10.1007/s11427-007-0023-3.

[13] Dathe M, Wieprecht T, Nikolenko H, Handel L, Maloy WL, MacDonald DL, et al. Hydrophobicity, hydrophobic moment and angle subtended by charged residues modulate antibacterial and haemolytic activity of amphipathic helical peptides. FEBS Lett 1997;403:208–12. https://doi.org/10.1016/S0014-5793(97)00055-0.

[14] Hellberg S, Sjoestroem M, Skagerberg B, Wold S. Peptide quantitative structure-activity relationships, a multivariate approach. J Med Chem 1987;30:1126–35. https://doi.org/10.1021/jm00390a003.

[15] Walker JM, editor. The Proteomics Protocols Handbook. Totowa, NJ: Humana Press; 2005. https://doi.org/10.1385/1592598900.

[16] Fürnkranz J. Decision Tree. In: Sammut C, Webb GI, editors. Encycl. Mach. Learn., Boston, MA: Springer US; 2011, p. 263–7. https://doi.org/10.1007/978-0-387-30164-8_204.

[17] Suthaharan S. Decision Tree Learning. Mach. Learn. Models Algorithms Big Data Classif., vol. 36, Boston, MA: Springer US; 2016, p. 237–69. https://doi.org/10.1007/978-1-4899-7641-3_10.

[18] Shaik AB, Srinivasan S. A Brief Survey on Random Forest Ensembles in Classification Model. In: Bhattacharyya S, Hassanien AE, Gupta D, Khanna A, Pan I, editors. Int. Conf. Innov. Comput. Commun., vol. 56, Singapore: Springer Singapore; 2019, p. 253–60.

https://doi.org/10.1007/978-981-13-2354-6_27.
[19] Resende PAA, Drummond AC. A Survey of Random Forest Based Methods for Intrusion Detection Systems. ACM Comput Surv 2019;51:1–36. https://doi.org/10.1145/3178582.
[20] Lorena AC, De Carvalho ACPLF. Uma Introdução às Support Vector Machines. Rev Informática Teórica E Apl 2007;14:43–67. https://doi.org/10.22456/2175-2745.5690.
[21] Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev 2021;54:1937–67. https://doi.org/10.1007/s10462-020-09896-5.
[22] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobotics 2013;7. https://doi.org/10.3389/fnbot.2013.00021.
[23] Valdés-Tresanco MS, Valdés-Tresanco ME, Valiente PA, Moreno E. gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS. J Chem Theory Comput 2021;17:6281–91. https://doi.org/10.1021/acs.jctc.1c00645.
[24] Xiao L, Diao J, Greene D, Wang J, Luo R. A Continuum Poisson–Boltzmann Model for Membrane Channel Proteins. J Chem Theory Comput 2017;13:3398–412. https://doi.org/10.1021/acs.jctc.7b00382.