



OPEN Analytical and machine learning approaches identify a sea star steroid with promising activity for COVID-19 therapeutic development

Mohamed S. M. Abd El Hafez^{1✉}, Aya I. Maiyza², Hanan A. Hassan², Sohila Osama³, Mohamed G. Seadawy⁴, Maha A. El Demellawy^{5,6} & Doaa A. Ghareeb^{6,7,8}

The pressing demand for safe and efficient COVID-19 treatments has intensified interest in Natural products, especially those derived from marine organisms. In this study, a bioactive steroidal compound, 5 α -cholesta-9(11)-en-3 β ,20 β -diol, was successfully isolated from the starfish *Acanthaster planci*. Structural elucidation was achieved using HREIMS, FTIR, and advanced 1D/2D NMR spectroscopy, confirming a molecular formula of C₂₇H₄₆O₂ and characteristic functionalities including hydroxyl and double bond moieties. The compound demonstrated notable anti-SARS-CoV-2 activity, attaining 85% viral inhibition at 5 ng/ μ l with an IC₅₀ of 5.86 μ M, as demonstrated by plaque reduction assays. Molecular docking studies demonstrated significant binding affinities toward key viral targets Mpro, NSP10, and RNA-dependent RNA polymerase with free binding energies of -26.85, -27.59, and -35.08 kcal/mol, respectively. These affinities surpassed those of their respective co-crystallized reference ligands. In-silico ADMET profiling indicated favorable pharmacokinetic properties, including high BBB penetration, moderate intestinal absorption, and non-hepatotoxicity. Toxicity assessments predicted low carcinogenic risk, a high rat MTD, and minimal ocular and dermal irritancy. Additionally, we developed a predictive web application based on machine learning to estimate IC₅₀ values of SARS-CoV-2 inhibitors, streamlining the drug discovery process. The forecasted values nearly matched the experimental outcomes, demonstrating the model's reliability and its potential to reduce time, cost, and risk in early-stage drug development. Moreover, machine learning models, particularly XGBoost, demonstrated excellent performance in predicting pIC₅₀ values (RMSE = 0.1357, MAE = 0.1022), supporting the development of a web-based IC₅₀ prediction application. The bioactivity prediction platform ENHPCG further validated the compound's antiviral potential, estimating an IC₅₀ of 5.95 μ M. Overall, these integrated analytical, biological, and computational approaches highlight 5 α -cholesta-9(11)-en-3 β ,20 β -diol as a potential SARS-CoV-2 inhibitor and a candidate for further pharmacological development.

Keywords COVID-19 inhibitor, Marine natural products, Machine learning, Molecular docking, Drug discovery

¹National Institute of Oceanography and Fisheries, NIOF, Cairo, Egypt. ²Informatics Research Institute (IRI), City of Scientific Research and Technological Applications (SRTA-City), Alexandria, Egypt. ³Institute of Graduate Studies and Research, Alexandria University, Alexandria, Egypt. ⁴Biodefense center for infectious and emerging diseases, MOD, Cairo, Egypt. ⁵Medical Biotechnology Department (MBD), Genetic Engineering and Biotechnology Research Institute (GEBRI), City of Scientific Research and Technological Applications (SRTA-City), New Borg El Arab, Alexandria, Egypt. ⁶Center of Excellence for Drug Preclinical Studies (CE-DPS), Pharmaceutical and Fermentation Industries Development Center, City of Scientific Research and Technological Applications (SRTA-City), New Borg El Arab, Alexandria, Egypt. ⁷Bio-screening and preclinical trial lab, Biochemistry Department, Faculty of Science, Alexandria University, Alexandria, Egypt. ⁸Research Projects unit, Pharos University in Alexandria, Canal El Mahmoudia Street, Beside Green Plaza Complex 21648 Alexandria, Egypt. ✉email: mohamedsaid80@yahoo.com

In the post-genomic era, computational methodologies play a pivotal role in analyzing large-scale biological data, facilitating the discovery of novel natural-product-derived drugs. A comprehensive understanding of molecular structures and chemical properties is essential for elucidating how physicochemical characteristics influence biological activity. Computer-aided drug design (CADD) encompasses various computational techniques that have significantly contributed to chemical biology and the study of structure-activity relationships (SARs)¹. These approaches evaluate bioactivity based on ligand properties, known as chemical descriptors². A key computational approach, quantitative structure-activity relationship (QSAR) modeling, is widely utilized to develop predictive models that identify critical molecular fingerprints influencing bioactivity and pharmacological properties³. QSAR models have been effectively utilized for diverse therapeutic targets, including antibacterial⁴, anticancer⁵, and antiviral⁶ agents. Furthermore, machine learning (ML) techniques have been leveraged to predict IC_{50} values⁷ and classify chemo-types with potential activity against viral proteases⁸.

There has been a sudden rise in viral diseases in recent years, posing significant threats to human health⁹. Natural products have been crucial in drug discovery and development during the past decades. Many bioactive compounds from plants and marine origins have been explored as effective alternatives for treating various diseases^{10,11}. Sea stars are rich sources of structurally diverse natural products with broad biological activities¹². Among these, marine steroids are particularly notable for their potential in drug development. These compounds exhibit various pharmacological activities, including anticancer, antiviral, antioxidant, antibacterial, anticoagulant, and immune-modulatory actions⁹. Marine steroids possess unique structural characteristics that contribute to their antiviral properties, either by directly inactivating virions prior to infection or by obstructing viral multiplication within host cells. Therefore, sea stars, which are rich in marine steroids, are pivotal in facilitating the identification and development of antiviral medicines.

A central challenge in antiviral drug discovery, particularly in response to emerging viral threats such as SARS-CoV-2, is the development of fast, cost-effective, and accurate approaches to screen and validate bioactive compounds. Traditional experimental methods for determining IC_{50} values are resource-intensive and often impractical for high-throughput screening. Therefore, integrating machine learning (ML)-based prediction with experimental validation presents a promising solution. This study investigates whether a combined machine learning and experimental approach can accurately predict and validate the antiviral activity (IC_{50}) of a marine-derived steroid compound against Covid-19. To explore this, we selected *Acanthaster planci* (crown-of-thorns starfish) as a bioresource due to its rich repertoire of marine steroidal metabolites.

Echinoderms significantly impact marine ecosystems by affecting the reproductive success of benthic species and acting as primary consumers of plant matter and detritus¹³. Due to their diverse bioactive compounds, echinoderms hold significant nutritional and medicinal value, contributing to various physiological functions in humans¹⁴. Many echinoderm species serve as model organisms in studies related to cell biology, developmental biology, and immunology. Their defense mechanisms largely depend on a robust innate immune system, which enables them to recognize and eliminate pathogens while facilitating wound healing¹³.

Echinoderms possess a diverse array of bioactive compounds, including saponins, peptides, polysaccharides, lipids, and proteins¹⁵. These compounds exhibit significant physiological functions in humans, with many lacking structural or functional analogs among terrestrial secondary metabolites. Current research primarily focuses on these marine-derived secondary metabolites owing to their significant toxicological and pharmacological attributes, as well as their ecological roles within marine environments¹³. Numerous studies have explored the antimicrobial, antifungal, and antitumor actions of echinoderm-derived metabolites; however, their antiviral potential remains largely unexplored. This study investigates the antiviral properties of recently isolated bioactive substances from echinoderms, potentially providing significant insights into the antiviral immunological processes of these creatures. Among echinoderms, starfish are particularly rich in polyhydroxy steroids, steroidal glycosides (sulfated), and asterosaponins, with sulfation occurring either on the carbohydrate moiety or the steroidal structure. Sea stars primarily contain saponins and polyhydroxy steroids, which exhibit cytotoxic, hemolytic, anti-inflammatory, and antimicrobial activities¹³.

In the ongoing pursuit of effective treatments for Covid-19, accurately predicting IC_{50} values for potential therapeutic compounds is essential for evaluating their effectiveness. Traditional experimental methods for determining IC_{50} values are often time-consuming, costly, and hazardous. This study introduces an innovative amalgamation of machine learning approaches and experimental validation to optimize the drug discovery process. Several ML-based platforms and models have recently been proposed for predicting anti-SARS-CoV-2 compound efficacy, including deep learning-based QSAR models and virtual screening tools³. However, many of these approaches lack direct experimental verification, limiting their translational utility. Our study builds upon this foundation by combining a robust ML-based predictive model for estimating IC_{50} values with experimental in vitro validation, offering a more comprehensive and practical workflow. Additionally, we introduce a user-friendly web application that allows researchers to input molecular structures and receive real-time IC_{50} predictions. This integration of in-silico prediction with experimental confirmation enhances both the speed and reliability of early-phase antiviral drug discovery.

However, the antiviral potential of these marine steroids, especially from *Acanthaster planci*, against SARS-CoV-2 remains poorly investigated. Moreover, while machine learning (ML) techniques and cheminformatics have greatly enhanced drug screening, few studies have applied these tools to marine sterols with subsequent experimental validation. This study combines in-vitro assays, molecular docking, and ADMET analysis to evaluate an *A. planci*-derived steroid toward SARS-CoV-2. In the current research, a steroid compound was isolated and structurally characterized from *A. planci*. Its antiviral activity against Covid-19 was evaluated using in-vitro plaque reduction assays, and its molecular structure was determined through advanced spectroscopic techniques.

Concurrently, we developed a machine learning-based model to predict IC_{50} values and performed molecular docking and ADMET study to evaluate its drug-likeness and interaction with viral targets. What distinguishes

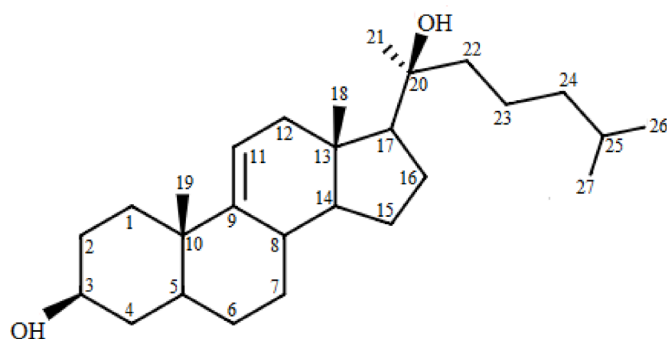


Fig. 1. The elucidated structure of the compound obtained “5α-cholesta-9(11)-en-3β, 20β-diol”.

Sample	Conc. (ng/μl)	Initial viral count	Viral count (PFU/ml)	Inhibition %
		(PFU/ml)		
The isolated Compound	5	9 × 10 ⁵	1.35 × 10 ⁵	85
	2.5		3.96 × 10 ⁵	56
	1.25		7.2 × 10 ⁵	20
	0.625		8.37 × 10 ⁵	7

Table 1. Plaque reduction assay (SARS-CoV2).

this work is the novel integration of ML-based IC₅₀ prediction with experimental validation, enabling a robust assessment of the compound's therapeutic potential. We further developed a user-friendly web application that allows researchers to input molecular structures and receive real-time IC₅₀ predictions, an innovation aimed at reducing the cost and time associated with early-stage drug screening. This integrative strategy demonstrates a comprehensive framework for antiviral drug discovery by combining marine biotechnology with advanced computational tools. It also underscores the untapped potential of echinoderm-derived metabolites in contributing to the development of next-generation antiviral agents.

Results

Extraction and isolation of the compound

The process for extracting and isolating the pure compound from *Acanthaster planci* (*A. planci*) is outlined in the flowchart (Fig. S1).

The isolated compound was a white crystalline substance with the molecular formula C₂₇H₄₆O₂, indicating five degrees of unsaturation. FTIR analysis revealed bands for the hydroxyl (3423 cm⁻¹) and double bond (1737 cm⁻¹) functional groups. ¹H NMR revealed a methine proton at δ 3.55, indicating a hydroxyl at C-3. Additionally, five methyl signals were detected, supporting the structure's sterol-like characteristics.

Carbon-13 NMR analysis revealed 27 carbon signals, including five methyl, eleven methylene, and four quaternary carbons, based on APT data. Key chemical shifts were noted at δ 72.0 (C-3) and δ 41.7 (C-20), with methyl carbons at δ 14.1, 18.0, 27.8, 19.4, and 19.6. COSY-45°, HSQC, and HMBC spectra enabled full proton and carbon assignments. A vicinal coupling between H-11 and H-12 in the COSY-45° spectrum confirmed a Δ^{9,11} double bond (Table S1, Figs. S2–S8). Based on these spectroscopic findings, the compound was identified as 5α-cholesta-9(11)-en-3β,20β-diol, as illustrated in Fig. 1.

Anti-SARS-CoV-2 activity determination

Table 1 shows that the isolated compound inhibited SARS-CoV-2 by 85% at 5 ng/μl, with an IC₅₀ of 5.86 μM.

Molecular Docking

To validate the docking protocol, native co-crystallized ligands were re-docked into their respective active sites. The resulting RMSD values (3.10 Å for M^{pro}, 1.07 Å for nsp10, and 1.34 Å for RdRp) confirmed acceptable alignment and method reliability. These RMSD outcomes confirm the robustness and precision of the docking methodology, indicating its suitability for accurately reproducing native ligand binding modes (Fig. 2). We acknowledge the visible structural deviation between the co-crystallized ligand and the most stable re-docked pose of PRD_002214, which resulted in an RMSD of 3.10 Å. While this value slightly exceeds the commonly accepted threshold of 2.0 Å, it may be attributed to the inherent flexibility of the ligand or limitations in the docking algorithm's conformational sampling¹⁶.

Comprehensive molecular docking simulations were achieved using MOE 14.0 software to explore the binding relations among the candidate compounds and selected target proteins. The calculated binding free energy (ΔG) values, summarized in Table S2, reflect the affinity and stability of these molecular complexes. The docking results indicated that the isolated compound consistently exhibited low ΔG values, suggestive of high

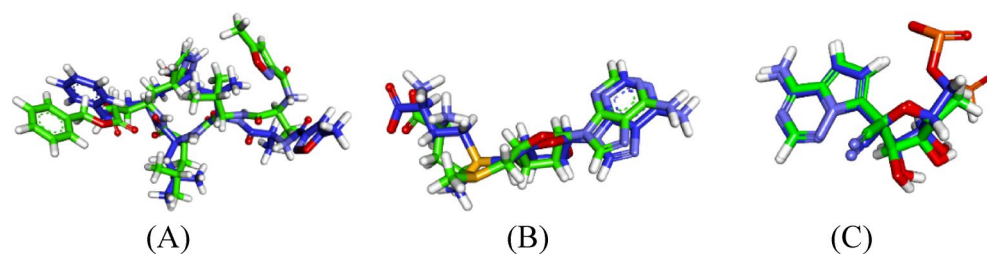


Fig. 2. Structural alignment of co-crystallized and docked conformations for ligands: (A) PRD_002214, (B) SAM, and (C) F86.

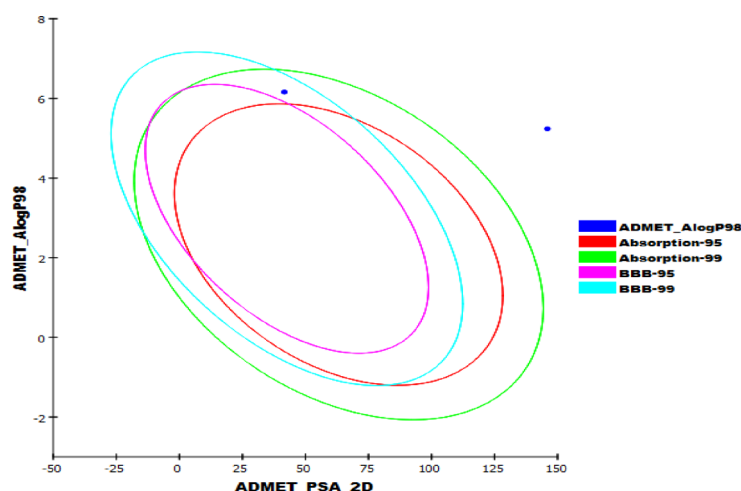


Fig. 3. The expected ADMET profile of the isolated compound.

binding affinity. The isolated compound exhibited notably strong binding affinities toward key SARS-CoV-2 targets. These binding energies were considerably lower than those of the respective co-crystallized reference ligands, highlighting the compound's enhanced interaction potential with the viral proteins (Table S2).

In-silico ADMET analysis

The ADMET characteristics of the isolated compound were evaluated in comparison to ritonavir as a reference, with key pharmacokinetic and toxicity properties predicted through Discovery Studio 4.0 (Table S3). Remarkably, the isolated compound exhibited substantial BBB penetration, whereas ritonavir displayed limited permeability.

The isolated compound showed very high blood-brain barrier (BBB) penetration (level 0), suggesting its potential to reach the central nervous system, while ritonavir exhibited very low penetration (level 4). In terms of aqueous solubility, the isolated compound was predicted to have very low solubility (level 1), which may affect its formulation and bioavailability, although this is not uncommon for lipophilic drug candidates. Intestinal absorption was rated as moderate (level 1) for the isolated compound, slightly better than ritonavir, which had poor absorption (level 2). Importantly, the isolated compound was not predicted to inhibit the CYP2D6 enzyme, indicating a lower risk of drug-drug interactions, unlike ritonavir, which was classified as a CYP2D6 inhibitor. Moreover, the compound was Non-hepatotoxic, in contrast to ritonavir, which showed potential for Liver toxicity. Finally, the compound displayed high plasma protein binding exceeding 90%, meaning it is likely to bind strongly to blood proteins, which can influence its distribution and half-life in the body (Fig. 3).

Toxicity studies

The toxicity assessment of the isolated compound was conducted using validated in silico predictive models implemented in Discovery Studio 4.0 software^{17,18}. Key safety parameters were assessed, including FDA rodent carcinogenicity predictions, which estimated the compound's carcinogenic potential based on its molecular structure. Carcinogenic potency (TD50) was predicted to estimate the tumorigenic dose rate at which 50% of test rodents develop tumors in chronic exposure studies¹⁹. The rat maximum tolerated dose (MTD) was estimated to identify the highest dose that induces no significant toxic effects^{20,21}, while the oral LD50 was calculated to determine the acute lethal dose causing mortality in 50% of test animals²². The chronic lowest observed adverse effect level (LOAEL) was also predicted to establish the minimum dose eliciting adverse effects upon long-term exposure^{23,24}. Additionally, ocular and dermal irritancy were assessed using the Draize eye test and rabbit skin irritation assay, respectively, to evaluate the compound's potential to cause eye and skin irritation²⁵.

ML model	Training		Testing	
	RMSE	MAE	RMSE	MAE
XGB	0.0370	0.0202	0.13574	0.10219
RF	0.0547	0.0392	0.14095	0.10255
SVR	0.0407	0.0272	0.7605	0.5578

Table 2. Summary of the predictive performance of machine learning models for estimating the pIC₅₀ values against SARS-CoV-2.

Model	Best parameters
XGB	Colsample_bytree: 0.8954454315573364 Learning_rate: 0.07200530796302361 Max_depth: 8 N_estimators: 496 Reg_alpha: 0.21458911970837413 Reg_lambda: 0.37725908322797586 Subsample: 0.71168893969307
RF	bootstrap: True max_depth: 44 min_samples_leaf: 1 min_samples_split: 2 n_estimators: 355
SVR	C: 546.8102793432797 degree: 3 epsilon: 0.025656646227779294 kernel: poly

Table 3. Best parameters of ML models for predicting pIC₅₀ of SARS-CoV-2.

The in-silico toxicity assessment revealed that the isolated compound exhibited a favorable safety profile relative to Ritonavir. It was anticipated to be non-carcinogenic and showed a higher TD₅₀ (2.651 mg/kg/day) and MTD (0.047 g/kg). The compound also demonstrated a significantly higher oral LD₅₀ (2.855 mg/kg/day) than Ritonavir (0.477 mg/kg/day). Although its LOAEL (0.005 g/kg) was lower than Ritonavir's. These comprehensive toxicity predictions provide essential insights into the compound's safety, supporting its potential applicability in further pharmacological and biomedical research.

Machine learning modeling

This section evaluates the prediction efficacy of machine learning models for pIC₅₀ estimation. The overall accuracy of different models following hyperparameter tuning is summarized in Table 2, while Table 3 presents the optimal parameters identified through the random search method. Among the evaluated models, the XGB algorithm demonstrated superior predictive performance, yielding the lowest root mean square error (RMSE) and mean absolute error (MAE) across both training and testing datasets (Table 2).

Accordingly, this model was selected for the deployment of the IC₅₀ prediction web application, utilizing the optimized parameters listed in Table 3. Furthermore, Fig. 4 illustrates the relationship between experimentally determined and predicted pIC₅₀ values generated by the model employing MACCS (Molecular ACCess System) fingerprint-based descriptors.

Roadmap for novel compounds in predicting IC₅₀ for SARS-CoV-2

To facilitate the prediction of IC₅₀ values for SARS-CoV-2, we employed DECIMER (Deep Learning for Chemical Image Recognition)²⁶, an advanced AI-driven tool designed for the automated extraction of chemical structures from images. This web-based platform utilizes deep learning algorithms to accurately recognize and convert chemical structure depictions into machine-readable formats such as SMILES. By streamlining the digitization of chemical information from research articles, patents, and other image-based sources, DECIMER enhances the efficiency of cheminformatics workflows.

In our study, we uploaded the 2D structural representation of the target compound (Fig. 5a) onto the DECIMER platform. The tool successfully generated the corresponding 3D molecular structure (Fig. 5b) and provided its SMILES notation, enabling further computational analysis and predictive modeling. The SMILES notation for the compound is as follows:

CC(C)CCC[C@](C)(C1CCC2C3CCC4C[C@H](CC[C@]4(C)C3=CC[C@@]21C)O)O

This representation encodes the molecular structure in a linear, machine-readable format, facilitating computational modeling and cheminformatics analyses.

The IC₅₀ value of the isolated compound toward SARS-CoV-2 M^{Pro} was predicted using the bioactivity prediction module integrated within the ENHPCG platform. Through the proposed web-based application for bioactivity assessment (<https://enhpcgic50prediction.streamlit.app/>), the compound yielded a predicted IC₅₀ value of 5.9488 μM, as illustrated in Fig. 6.

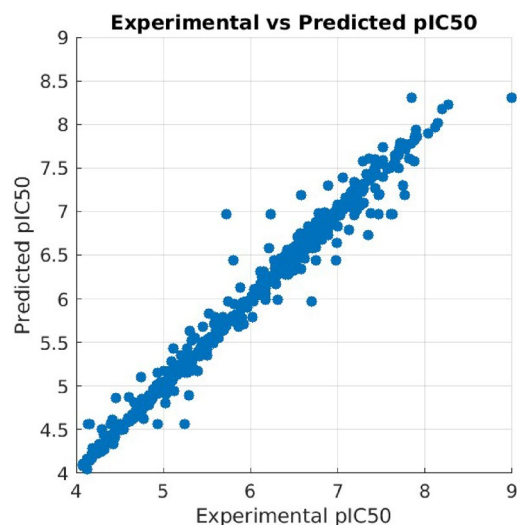


Fig. 4. Comparison of experimental and predicted pIC₅₀ values generated by the machine learning model utilizing MACCS fingerprint descriptors.

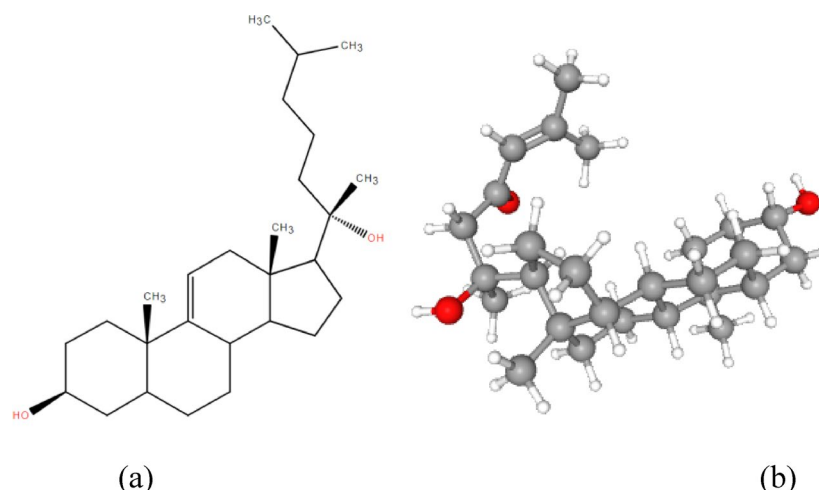


Fig. 5. (a) 2d and (b) 3d structure of the isolated compound.

Validation of the proposed XGB model for IC₅₀ prediction in SARS-CoV-2

To assess the predictive performance of our proposed XGB model, we conducted a validation study by comparing its IC₅₀ predictions with experimentally determined values for 32 bicyclopamine derivatives targeting SARS-CoV-2 main protease (Mpro). Table S4 presents a comparative analysis between the experimental IC₅₀ values reported by²⁷ and the predicted values by our machine learning model. The results indicate a strong correlation between the predicted and experimental data, underscoring the accuracy and reliability of our approach in estimating the bioactivity of potential SARS-CoV-2 inhibitors.

To evaluate the predictive accuracy of our XGB model for lower-activity compounds against Mpro, we compared the experimentally determined IC₅₀ values reported by²⁸ with the predicted IC₅₀ values for selected compounds targeting the Mpro of SARS-CoV-2 (Table S5). The findings indicate that the model effectively predicts IC₅₀ values with reasonable accuracy, demonstrating a strong correlation with the experimental data.

For our isolated compound, experimental IC₅₀ determination was carried out using the logistic regression method, as shown in Fig. 7. The experimentally derived IC₅₀ was 2.23 ng/μL. Using the RDKit cheminformatics library, we computed the molecular weight (402.66 g/mol) and converted this value into μM units to enable direct comparison with model predictions as follows:

$$IC_{50} (\mu M) = \frac{IC_{50} (ng/\mu l)}{Molecular\ Weight (g/mol)} \times 10^3 = \frac{2.23}{402.66} \times 10^3 = 5.5469 \mu M$$

Predict IC50 Value

Instructions:

- To convert your compound to a Simplified Molecular Input Line Entry System (SMILES), please visit this website: decimer.ai

Enter SMILES String

CC(C)CCC[C@@](C)(C1CCC2C3CCC4C[C@H](CC[C@]4(C)C3=CC[C@@]21C)O)O

Predict and Classify

Predicted IC₅₀: 5.9488 μ M

Predicted IC₅₀: 5948.7998046875 nM

Extracted Features:

	MolecularWeight	LogP	HydrogenBondDonors	HydrogenBondAcceptors	TopologicalPolarSurfaceAr
0	402.663	6.5036	2	2	40.

Download Features as CSV

Fig. 6. The Predicted IC₅₀ value for the isolated compound using ENHPCG bioactivity service.

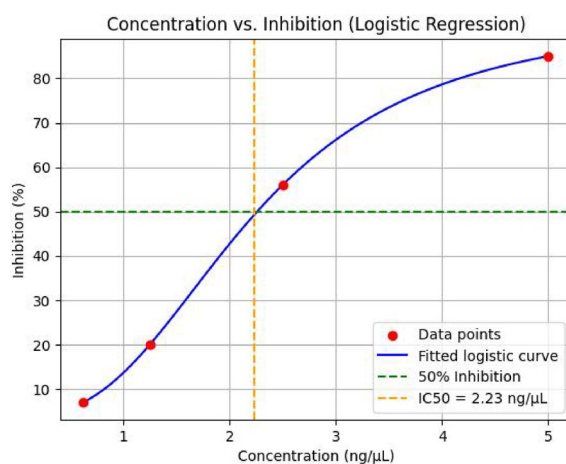


Fig. 7. Dose-Response Curve to calculate IC₅₀ experimentally.

This case study demonstrates how combining experimental assays with machine learning predictions provides a robust, dual-layer validation approach that strengthens confidence in candidate compound selection.

Discussion

In this study, we isolate a novel compound from the *Acanthaster planci* starfish with potential inhibition of COVID-19 infection. Comprehensive structural characterization, molecular docking analyses, and in-silico ADMET evaluations were achieved. Our primary focus is on developing a robust model for predicting the IC₅₀ values of compounds with activity against SARS-CoV-2. To improve accessibility, we have also developed a web application that enables researchers to input compounds and efficiently obtain predicted IC₅₀ values. The model's validity was assessed by comparing the predicted IC₅₀ values with experimental measurements obtained through in-vitro bioassays, demonstrating its effectiveness in reducing both the cost and risk associated with IC₅₀ determination.

The urgent demand for a safe and effective drug to combat COVID-19 has intensified research into natural products, which have long served as fundamental components of traditional medicine. These bioactive compounds represent a key resource for developing novel therapeutic agents, making their exploration a crucial step in drug discovery. In the treatment of retroviral infections, several molecular targets can be inhibited.

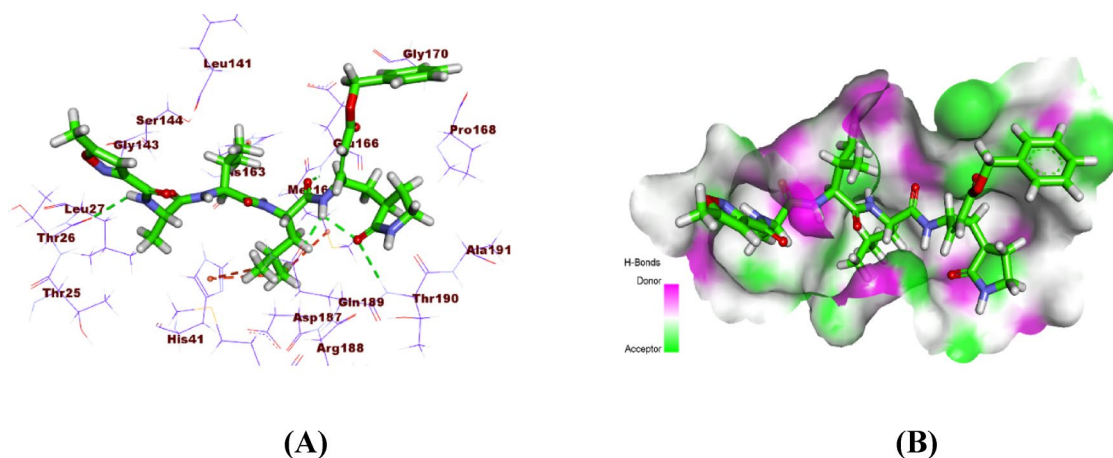


Fig. 8. (A) 3D docking and (B) Surface mapping of the co-crystallized ligand (PRD_002214) within the active site of the COVID-19 main protease.

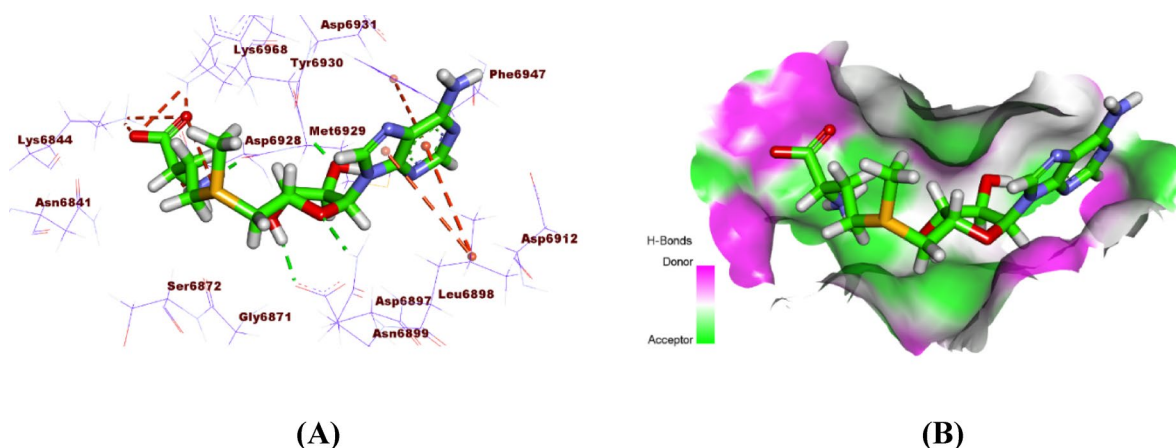


Fig. 9. (A) 3D structure, and (B) surface mapping of SAM bound to the active site of SARS-CoV-2 NSP10.

In-silico target identification offers a rapid, cost-effective method for screening natural products. Key computational tools include QSAR modeling and molecular docking.

Docking helps reveal interactions between natural compounds and target biomacromolecules, aiding mechanism-of-action studies. This strategy supports drug discovery and has been central to developing natural product-based therapeutics.

The Mpro protein active site contains four distinct pockets, each interacting with specific ligand moieties. The co-crystallized ligand (PRD_002214) forms four hydrogen bonds and three hydrophobic interactions. Key interactions include hydrogen bonds with Thr190, Gln189, and Thr26, and hydrophobic contacts with His41 and Met165. These interactions collectively stabilize the ligand within the active site (Fig. 8).

The co-crystallized ligand SAM interacts with SARS-CoV-2 nsp10 through three hydrogen bonds, along with several hydrophobic and electrostatic interactions, thereby guaranteeing stable occupancy of the active site. The critical residues implicated are Asn6899, Tyr6930, Asp6928, Phe6947, and Leu6898 (Fig. 9).

The co-crystallized ligand F86 binds firmly to the active site of SARS-CoV-2 RdRp via many interactions. This comprises three hydrogen bonds, six hydrophobic relations, and two electrostatic contacts. The critical residues implicated are Arg555, Val557, Asp623, Asp760, and Cys622. This interaction network improves the ligand's binding affinity and conformational stability (Fig. 10).

The isolated compound bound to three key pockets of SARS-CoV-2 Mpro, forming seven hydrophobic contacts and one hydrogen bond. Its binding orientation closely resembled that of the co-crystallized ligand, indicating similar active site interactions. The (1R,4R)-4-methylcyclohexan-1-ol and indene moieties interacted with Met49, His41, and Cys145 across the first two pockets. The side chain reached the third pocket, producing a hydrogen bond with Gln189 and hydrophobic contacts with Met165 and His41. (Fig. 11).

The isolated compound exhibits stable binding within the NSP10 active site via four hydrophobic connections and one hydrogen bond. The (1r,4r)-4-methylcyclohexan-1-ol moiety forms a hydrogen bond with Asp6912 and hydrophobic contacts with Leu6898 and Phe6947. Additional hydrophobic interactions with Pro6932 from two separate moieties further stabilize the ligand (Fig. S9). The isolated compound showed strong binding

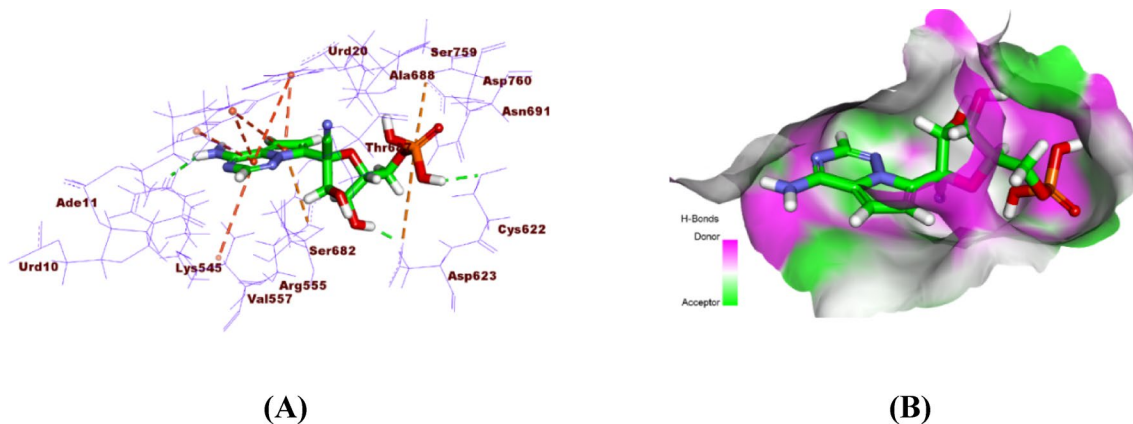


Fig. 10. A) 3D binding, and B) surface mapping of the co-crystallized ligand F86 within the active site of COVID-19 RdRp.

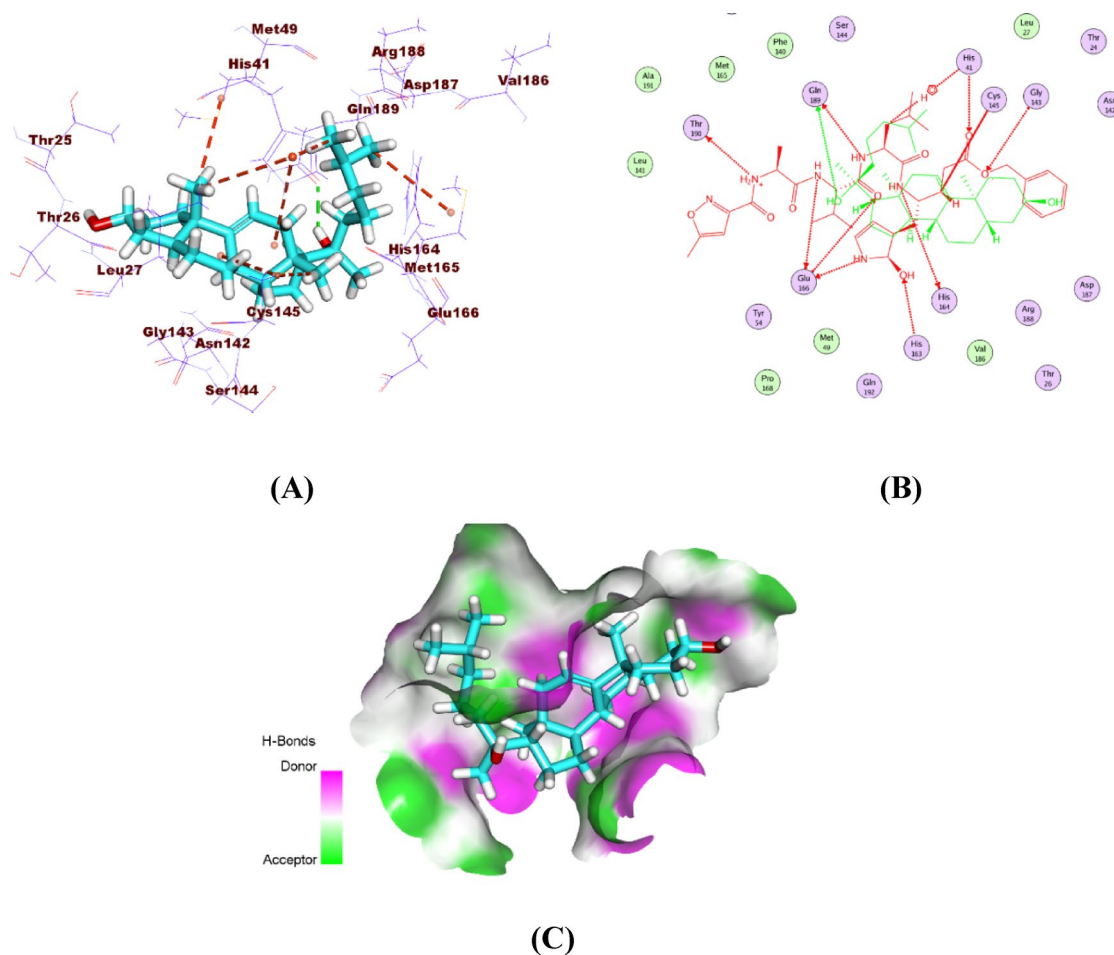


Fig. 11. Binding of the isolated compound to the SARS-CoV-2 main protease: (A) 3D binding pose, (B) 2D overlay with the co-crystallized ligand, and (C) surface mapping within the active pocket.

with RdRp, forming ten hydrophobic relations and two hydrogen bonds within the active site. Key fragments interacted with residues such as Lys545, Val557, and Urd20, enhancing binding stability. Additionally, the side chain formed a hydrogen bond with Asp623, further supporting its inhibitory potential (Fig. 12).

A detailed analysis of the docking results revealed that only specific functional groups within the isolated compound actively participate in key interactions with the target binding site. The steroidal core structure of the compound establishes hydrophobic contacts with non-polar residues lining the binding pocket, contributing

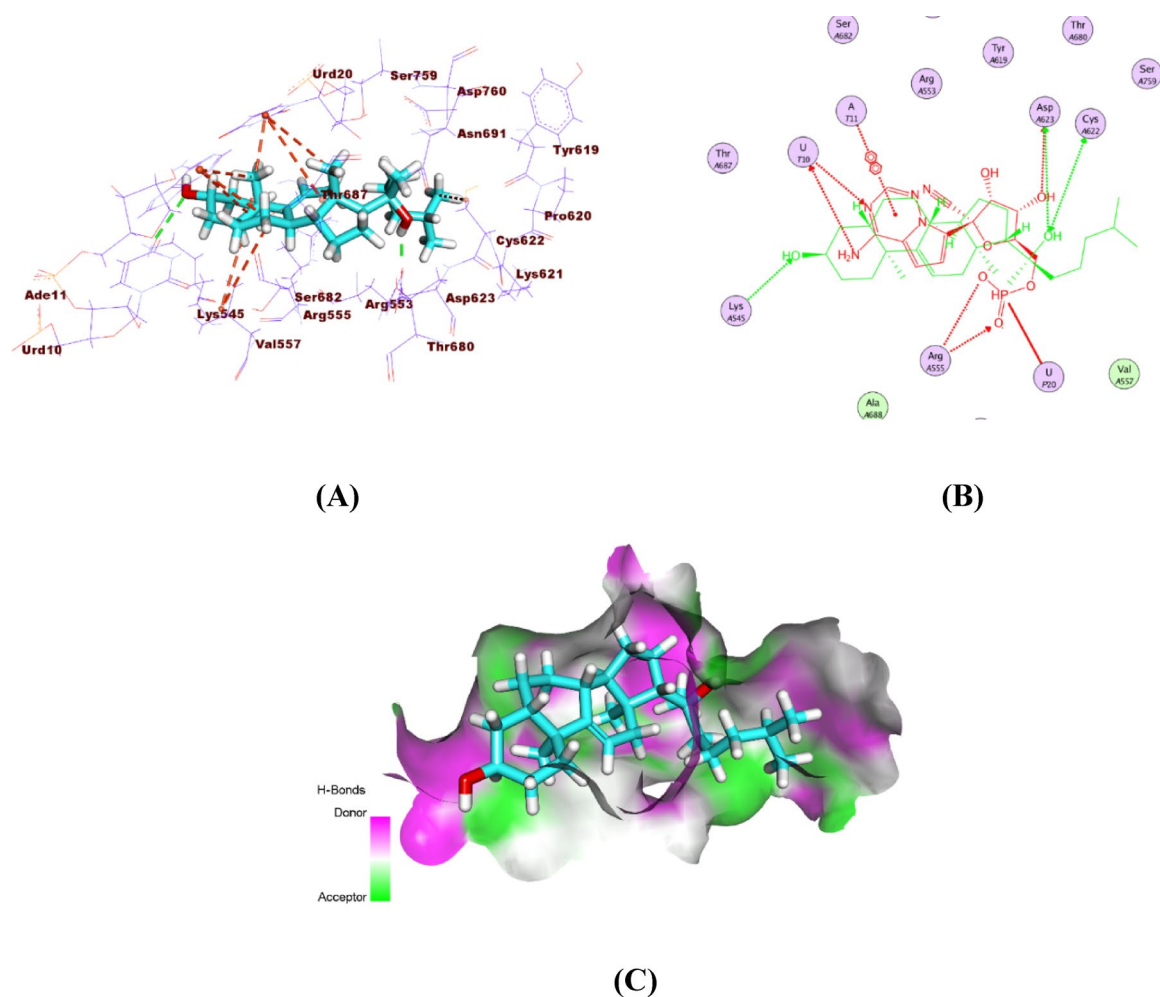


Fig. 12. Binding interaction of the isolated compound with SARS-CoV-2 RdRp, showing (A) 3D, (B) 2D, and (C) surface views. The compound aligns with the co-crystallized ligand, confirming its fit within the enzyme's active site.

to the stabilization of the ligand within the active site. Additionally, the presence of hydroxyl groups enables hydrogen bonding interactions with polar residues such as Ser144 and His163, which are known to be critical for ligand recognition in the main protease active site. These polar interactions contribute substantially to the improvement of binding affinity and specificity. Furthermore, the spatial orientation of the functional groups dictates the ligand's ability to fit properly within the pocket, influencing the overall docking score and binding mode. This selective engagement of chemical moieties underscores the importance of molecular complementarity in structure-based drug design and supports the observed docking performance of the isolated compound.

As presented in Table S6, in-silico toxicity assessments indicated that the isolated compound exhibited low toxicity across multiple predictive models. It was classified as non-carcinogenic according to the FDA rodent carcinogenicity model. The compound displayed a higher TD_{50} value (2.651 mg/kg/day) and a greater predicted maximum tolerated dose (0.047 g/kg) compared to ritonavir (1.369 mg/kg/day and 0.038 g/kg, respectively), indicating a favorable toxicity profile.

Furthermore, the compound demonstrated an oral LD_{50} of 2.855 mg/kg body weight/day, significantly exceeding that of ritonavir (0.477 mg/kg body weight/day), indicating a lower potential for acute toxicity. The lowest observed adverse effect level (LOAEL) was identified as 0.005 g/kg body weight, which is lower than ritonavir's LOAEL of 0.014 g/kg body weight, suggesting enhanced safety at low-dose exposure. Additionally, predictive models identified the compound as a potential ocular irritant while showing no skin irritation potential, reinforcing its favorable safety characteristics.

The comparison between the isolated steroid and known antiviral agents, such as ritonavir, underscores its therapeutic potential. Notably, the compound exhibits favorable binding affinity, suggesting a strong interaction with the target viral protein, which is essential for effective inhibition. This computational prediction is further supported by molecular docking and dynamics studies, positioning the compound as a viable antiviral candidate. Additionally, its ADMET profile reveals lower acute toxicity (higher LD_{50}) and a higher maximum tolerated dose (MTD) compared to ritonavir, indicating that it may be better tolerated in-vivo. These pharmacokinetic

advantages, combined with predicted non-carcinogenicity and skin safety, enhance its attractiveness for further development.

What distinguishes the isolated compound most significantly is its unique chemical structure and biological origin. As a novel marine-derived steroid, it represents a new chemotype not commonly found among current antiviral drugs, which are often synthetic or peptide-based. This structural novelty may reduce the risk of cross-resistance with existing therapies and provide new mechanisms of action. Moreover, the compound's balanced profile of efficacy (via binding affinity) and safety (via in-silico toxicity studies) suggests a promising starting point for lead optimization and preclinical investigation. Its potential as a safe and effective antiviral agent merits further validation through in-vitro assays and mechanistic studies.

For the majority of compounds analyzed, the predicted IC_{50} values exhibited a strong correlation with experimentally determined values, underscoring the model's ability to capture key molecular determinants of antiviral activity. This correlation suggests that the machine learning approach, particularly the XGB model, is well-suited for early-stage screening and prioritization of SARS-CoV-2 inhibitors. For example, compounds such as Ebselen, Disulfiram, and Tideglusib were predicted to have low IC_{50} values, consistent with previously reported experimental data supporting their inhibitory potential.

Nonetheless, certain discrepancies were observed. In cases like Carmofur and PX-12, the model tended to overestimate the IC_{50} values, potentially due to factors such as off-target effects, conformational flexibility, or specific interactions not captured by the chosen molecular descriptors. These deviations highlight a key limitation of the current model: it relies solely on the MACCS fingerprint, which, while efficient and interpretable, may not fully represent complex bioactive features or three-dimensional interactions. Furthermore, the training dataset size, though curated and specific, may limit generalizability to structurally novel compounds or those with underrepresented scaffolds.

Despite these constraints, the model demonstrates substantial utility as a computational filter to reduce experimental workload and resource consumption. This is particularly beneficial when focusing on highly active compounds (e.g., $IC_{50} < 1 \mu M$), where early prediction can direct attention to the most promising leads, as summarized in Table S4.

Conclusion

This study successfully isolated and characterized a steroid compound from starfish, demonstrating its potential as a COVID-19 inhibitor. Using a combined analytical approach, we identified the structural and functional properties of the compound, which exhibited promising in-vitro antiviral activity. Furthermore, machine learning models provided predictive insights into the compound's inhibitory potential, reinforcing its relevance as a candidate for further drug development.

Our findings highlight the significance of marine-derived natural products in antiviral research and underscore the power of integrating computational and experimental methodologies for drug discovery. Future studies should focus on in-vivo evaluations and structural modifications to enhance the compound's efficacy and pharmacokinetic properties, paving the way for novel marine-based therapeutics against COVID-19 and other viral infections.

This study exposed that *A. planci* is a promising reservoir of bioactive compounds with significant potential for the development of new therapeutic agents. The development of our predictive model for IC_{50} values has proven effective in estimating the potency of compounds against SARS-CoV-2, with validation against experimental data confirming its accuracy. The integrated web application enhances accessibility, allowing researchers to efficiently obtain predictions and streamline the drug discovery process.

This model significantly reduces the cost and risk associated with traditional IC_{50} measurements. The strong correlation observed between predicted and experimental IC_{50} values affirms the model's applicability in real-world drug screening scenarios. This serves as a foundation for incorporating ML-driven prediction into antiviral discovery pipelines. Future investigations will aim to enhance the model's predictive performance by integrating expanded datasets and employing more sophisticated machine learning algorithms.

Further validation with a broader range of experimental datasets is planned to enhance the model's robustness. Additionally, we aim to expand the web application's features to include predictive capabilities for other viral targets and integrate it with existing drug discovery pipelines. This study represents a preliminary exploration of the antiviral potential of the isolated compound. While the selection of the three targets was guided by existing literature and their known relevance to SARS-CoV-2, we acknowledge the limitation of not employing computational target prediction tools such as PharmMapper or SEA. In future work, we intend to incorporate such in silico target fishing approaches to ensure a more comprehensive and unbiased identification of potential molecular targets.

Methods

Sample collection

Live *Acanthaster planci* specimens (100–300 g) were collected in July 2022 from the Red Sea near Hurghada, Egypt (27.28°N, 33.77°E), at a depth of 4–10 m. Identification was based on morphological traits. Specimens were washed, sealed in bags, frozen, and transported under refrigeration. Samples were stored at $-20^{\circ}C$ for later analysis. Taxonomic classification was confirmed according to the World Register of Marine Species. The starfish belongs to Phylum Echinodermata, Class Asteroidea, Order Valvatida, Family Acanthasteridae. All procedures complied with NIOF ethical standards and ARRIVE guidelines for marine invertebrate research. All methods were performed in accordance with the relevant guidelines and regulations. All the experimental protocols were approved by the Institutional Animal Care and Use Committee (IACUC), National Institute of Oceanography and Fisheries.

Extraction and isolation

Whole-body tissue of *A. planici* (1.5 kg) was extracted using chloroform/methanol at room temperature, repeated three times. The pooled extracts were dried under reduced pressure and stored at -20°C . Crude extract underwent purification using silica gel column chromatography.

A stepwise elution gradient from hexane and dichloromethane to methanol was applied for fractionation (Fig. S1).

Spectroscopic characterization of the isolated compound

The structural elucidation of the isolated compound was accomplished through a comprehensive set of spectroscopic techniques, including NMR, FT-IR, and mass spectrometry.

NMR spectral analysis

Proton (^1H) and carbon (^{13}C) nuclear magnetic resonance (NMR) spectra were obtained using a Bruker AVANCE 400 MHz spectrometer at the University of Winnipeg, Canada, with deuterated chloroform (CDCl_3) as the solvent. Structural elucidation was further supported by two-dimensional NMR techniques, including COSY, HSQC, HMBC, and NOESY. Chemical shifts (δ) are reported in ppm, and coupling constants (J) are given in Hz.

FT-IR analysis

FT-IR spectroscopy was utilized to determine the functional groups within the isolated compound. The sample was finely ground and uniformly blended with an equal amount of dried KBr, then pressed into a pellet. Spectral data were acquired in transmission mode using a Bruker Alpha FT-IR spectrometer, spanning the wavenumber range of 4000 to 400 cm^{-1} .

GC-MS analysis

GC-MS analysis was performed to recognize and verify the structure of the isolated bioactive molecule. A high-resolution GC-MS system, the Agilent 7890 A-5975 C, was employed, utilizing helium as the carrier gas. The temperature gradient was set from 90°C to 300°C . Compound identification relied on spectral matching with known reference databases through automated techniques.

Assessment of Anti-SARS-CoV-2 activity

The antiviral activity of the isolated compound toward SARS-CoV-2 was assessed through a plaque reduction assay, performed according to the methodology outlined by²⁹. This experiment was carried out at the central laboratories of the Chemical Warfare Department in Egypt. The compound's efficacy was determined by calculating the percentage reduction in viral plaque formation relative to untreated control wells, providing a measure of its antiviral potential as follows:

$$\% \text{ inhibition} = \frac{\text{viral count (untreated)} - \text{viral count (treated)}}{\text{viral count (untreated)}} \times 100$$

Molecular docking studies

The three-dimensional crystal structures of the selected target proteins: (i) COVID-19 main protease (M^{pro}) (PDB ID: 6lu7, resolution: $2.16\text{ \AA}$), (ii) non-structural protein (nsp10) (PDB ID: 6W4H, resolution: $1.80\text{ \AA}$), and (iii) RNA-dependent RNA polymerase (PDB ID: 7BV2, resolution: $2.50\text{ \AA}$) were retrieved from the Protein Data Bank (<http://www.pdb.org>). Molecular docking was performed using the Molecular Operating Environment (MOE) software platform³⁰ to evaluate the binding affinity and interaction profiles of the test compounds with the active sites of the target proteins.

Prior to docking, water molecules were removed from the protein structures, and a single, functionally relevant protein chain was retained for each target. The co-crystallized ligands served as reference molecules for validation of the docking protocol. The protein structures were protonated and energy-minimized using the MMFF94x force field after hiding hydrogen atoms. Binding sites were identified and defined based on previously reported methodologies^{31–33}, enabling accurate docking simulations and comparative analysis of binding energies and poses.

The chemical structures of the investigated compound and the co-crystallized Ligands were initially constructed using ChemBioDraw Ultra 14.0 and exported in Structure Data File (SDF) format. These files were subsequently imported into Molecular Operating Environment (MOE) software, where the 3D structures were protonated, and energy minimization was performed using the MMFF94x force field.

To validate the docking protocol, redocking of each co-crystallized ligand into its corresponding binding site was conducted. Successful validation was indicated by low root-mean-square deviation (RMSD) values between the docked poses and the original crystallographic conformations, following established methodologies with acceptable cutoff: $\text{RMSD} \leq 2.0\text{ \AA}$ ^{34,35}. Docking simulations were executed using the default settings in MOE, generating 30 docking poses per Ligand via genetic algorithm-based conformational searches. The docked complexes were subsequently examined and visualized via Discovery Studio 4.0 software^{35–39}. Figure (S10) indicate a schematic diagram for the docking workflow.

In-Silico ADMET prediction

The compound was structurally optimized using the CHARMM force field, then analyzed using the ADMET Descriptors protocol in Discovery Studio 4.0. This evaluation predicted its pharmacokinetic and toxicity-related properties^{34–37}.

In-Silico toxicity assessment

Toxicological profiling of the compound was also conducted using Discovery Studio 4.0. Ritonavir served as the reference control drug⁴⁰. Ritonavir was selected as a reference compound due to its well-established antiviral activity and its reported inhibitory effects on SARS-CoV-2 main protease in several in-silico and in-vitro studies. It has been widely used in molecular docking and ADMET studies as a benchmark for evaluating the pharmacokinetic and toxicity profiles of potential antiviral candidates. After structural optimization using the CHARMM force field, toxicity endpoints were evaluated via the Toxicity Prediction (Extensible) protocol. Multiple predictive parameters were evaluated to determine the compound's safety profile^{41–43}.

Machine learning model development

A dataset comprising 758 small-molecule inhibitors of the SARS-CoV-2 main protease (Mpro) was obtained from a publicly available and curated dataset³. The compounds span a broad range of chemical scaffolds and activity values, making them suitable for training robust regression models for IC₅₀ prediction. To represent each molecule's structural and physicochemical features, we computed molecular fingerprints, which are widely used in QSAR modeling to encode relevant chemical information. Among the various fingerprint types available, such as PubChem, ECFP, CDK, Substructure, and MACCS, we selected the 166-bit MACCS (Molecular Access System) keys, generated using the RDKit cheminformatics library.

MACCS fingerprints were chosen for several reasons. First, they offer broad coverage of substructural features, enabling an informative representation of bioactivity-relevant chemical motifs⁴⁴. Second, their low dimensionality helps mitigate the risk of overfitting and enhances model interpretability, particularly in datasets of limited size. Third, they are computationally efficient, making them well-suited for large-scale virtual screening and seamless integration into web-based predictive platforms. Although more complex and higher-dimensional fingerprints, such as Klekota-Roth or ECFP, are available and commonly used in cheminformatics applications, preliminary experiments revealed that these alternatives provided only marginal improvements in predictive performance while significantly increasing computational overhead. Given this tradeoff, MACCS fingerprints were deemed appropriate and optimal for the goals of the current study.

Feature selection and model training

The MACCS fingerprints served as input features for model training. No additional dimensionality reduction or filtering was applied due to the relatively low feature count (166). The dataset was randomly split into 80% training and 20% testing subsets. To validate model generalization and prevent overfitting, we employed 4-fold cross-validation during hyperparameter tuning. Three regression algorithms were evaluated: Random Forest (RF), Extreme Gradient Boosting (XGB), and Support Vector Regression (SVR). Performance was assessed using standard regression metrics (e.g., RMSE, MAE).

Subsequently, several widely adopted regression algorithms were trained and evaluated, including Extreme Gradient Boosting Regressor (XGB), Random Forest (RF), and Support Vector Regression (SVR). These models were assessed for their predictive performance on the prepared dataset. The XGB model, recognized for its computational efficiency and high predictive accuracy when handling large-scale datasets, was evaluated for its capability to address real-world regression tasks using optimized resource utilization⁴⁵. The RF model, which constructs an ensemble of decision trees to enhance predictive robustness and reduce variance, was tested for its reliability and generalization performance³. SVR, known for its effectiveness in managing high-dimensional feature spaces and its flexibility in non-linear regression, was also examined for its applicability to the current dataset⁴⁶.

Model performance was quantitatively assessed using standard regression evaluation metrics to determine suitability for the intended QSAR application. The biological activity values (IC₅₀, in μM) were transformed into their logarithmic form (pIC₅₀), where $\text{pIC}_{50} = -\log_{10}(\text{IC}_{50})$, a commonly used transformation in QSAR studies to normalize the data and improve model interpretability⁴⁷. Hyperparameter optimization for each model was performed using MATLAB's *fitrensemble* function, integrated with a custom random search strategy, evaluating combinations across predefined hyperparameter ranges. This ensured robust exploration of the hyperparameter space.

Computational framework and model optimization

All analyses were conducted on the Egyptian National High-Performance Computing Grid (ENHPCG)⁴⁸, using MATLAB R2022a Update 4. The grid offers scalable computing infrastructure to handle large datasets and intensive training processes⁴⁹. To optimize model performance, we utilized MATLAB's *fitrensemble* function, integrated with a custom random search algorithm for hyperparameter tuning. This method involved specifying predefined ranges for each hyperparameter and systematically evaluating various combinations to identify those yielding the highest predictive accuracy. Hyperparameter optimization was performed using a 4-fold cross-validation strategy to ensure robustness and generalizability in predicting pIC₅₀ activity. The optimal hyperparameter configurations for each regression model were as follows:

Random Forest (RF): *max_depth*, *min_samples_leaf*, *min_samples_split*, and *n_estimators*.

Extreme Gradient Boosting (XGB): *colsample_bytree*, *learning_rate*, *max_depth*, *n_estimators*, along with regularization parameters *reg_alpha*, *reg_lambda*, and *subsample*.

Support Vector Regression (SVR): *C*, *degree*, *epsilon*, and a polynomial (*poly*) kernel.

By combining *fitrensemble* with the randomized search approach, we aimed to maximize the predictive performance and efficiency of the regression models in forecasting pIC₅₀ values.

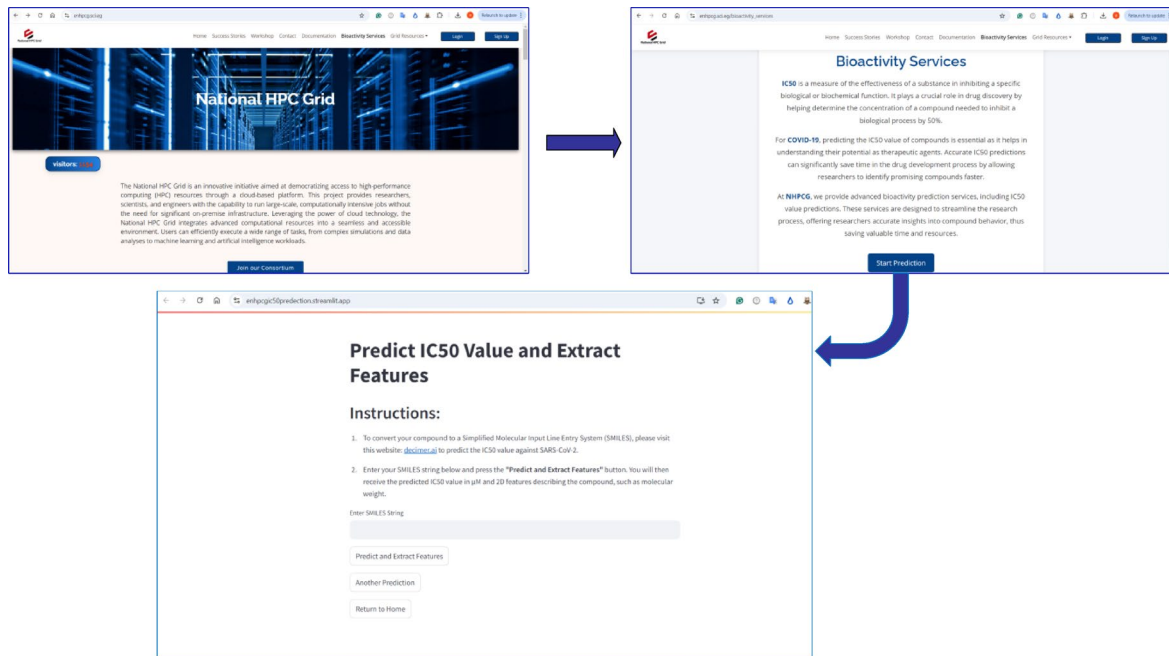


Fig. 13. IC₅₀ Prediction Application Hosted on the ENHPCG Platform.

Image-to-SMILES conversion with DECIMER

Several bioactive compounds from recent literature were available only in image (2D structure) format, without associated SMILES strings. To incorporate these into our dataset and prediction pipeline, we used DECIMER, a deep-learning-based tool for image-to-SMILES conversion. This integration was essential for digitizing chemical structures from scientific figures and enabling their use in cheminformatics workflows, especially when the textual representations were not provided.

Deployment of machine learning model as a Web-Based application

After model optimization, the best-performing model was deployed as a web application on the ENHPCG platform (<https://enhpcg.sci.ehpcg.org/bioactivity-services>), (<https://enhpcg.sci.ehpcg.org/bioactivity-services>) (<https://enhpcg50prediction.streamlit.app/>).

As illustrated in Fig. 13, users can input molecules as SMILES, and the tool outputs predicted IC_{50} values by transforming predicted pIC_{50} via $IC_{50} = 10^{-pIC_{50}}$. This user-friendly system enables rapid in-silico screening of antiviral compounds.

Experimental validation

To validate predictive accuracy, we compared predicted IC_{50} values of newly isolated echinoderm-derived compounds against their experimentally measured IC_{50} values from in-vitro assays. This comparison confirmed the model's reliability and potential applicability in real-world antiviral drug discovery pipelines.

Statistical analysis

Statistical comparisons were conducted utilizing one-way analysis of variance (ANOVA) in SPSS version 15, with a significance level established at $p < 0.05$. Results are presented as mean $\pm$ standard deviation (SD).

Data availability

All data supporting the conclusions of this investigation are included within the main manuscript and Supplementary Information. Additional raw datasets are available from the corresponding author upon reasonable request. Source data are included.

Received: 22 April 2025; Accepted: 15 September 2025

Published online: 07 October 2025

References

1. Prachayasittikul, V. et al. Computer-Aided drug design of bioactive natural products. *Curr Top. Med. Chem* **15**, 1780–1800 (2015).
2. Nantasenamat, C., Isarakura-Na-Ayudhya, C. & Prachayasittikul, V. Advances in computational methods to predict the biological activity of compounds. *Expert Opinion on Drug Discovery* vol. 5 at (2010). <https://doi.org/10.1517/17460441.2010.492827>
3. Ferdous, N. et al. MproPred: A machine learning (ML) driven Web-App for bioactivity prediction of SARS-CoV-2 main protease (Mpro) antagonists. *PLoS One* **18** (6), e2087179 (2023).
4. Pir, M., Agirbas, H., Budak, F. & Ilter, M. Synthesis, characterization, antimicrobial activity, and QSAR studies on substituted Oxadiazaboreals. *Med Chem. Res* **25**, 1794–1812 (2016).

5. Suvannang, N. et al. Probing the origin of Estrogen receptor alpha inhibition: via large-scale QSAR study. *RSC Adv* **8**, 11344–11356 (2018).
6. Worachartcheewan, A., Prachayasittikul, V., Toropova, A. P., Toropov, A. A. & Nantasenamat, C. Large-scale structure-activity relationship study of hepatitis C virus NS5B polymerase Inhibition using SMILES-based descriptors. *Mol Divers* **19**, 955–964 (2015).
7. Nand, M. et al. Virtual screening of anti-HIV1 compounds against SARS-CoV-2: machine learning modeling, chemoinformatics and molecular dynamics simulation based analysis. *Sci Rep* **10** (1), 20397 (2020).
8. Mekni, N., Coronello, C., Langer, T., De Rosa, M. & Perricone, U. Support vector machine as a supervised learning for the prioritization of novel potential sars-cov-2 main protease inhibitors. *Int J. Mol. Sci* **22** (14), 7714 (2021).
9. Hans, N., Malik, A. & Naik, S. Antiviral activity of sulfated polysaccharides from marine algae and its application in combating COVID-19: Mini review. *Bioresource Technology Reports* vol. 13 at (2021). <https://doi.org/10.1016/j.biteb.2020.100623>
10. Abd, E., Hafez, M. S. M., El-Wahab, A., Abdel-Hamid, M. G. & Ghareeb, A. S. A. El Demellawy, M. A. Biological activities of secondary metabolites from turbinaria triquetra (Phaeophyceae), hypnea cornuta (Florideophyceae), and Ulva prolifera (Ulvophyceae) methanolic extracts. *Egypt. J. Aquat. Biol. Fish.* **26**, 1227–1246 (2022).
11. El-Din, N. G. S., Hafez, M. S. M. A., El, El-Wahab, M. G. A. & Ibrahim, H. A. H. Biological activities of derived pigments and polyphenols from the newly recorded Alga phyllymenia gibbesii. *Sci Rep* **14**, 21284 (2024).
12. Goswami, L., Pakshirajan, K. & Pugazhenth, G. Biological treatment of biomass gasification wastewater using hydrocarbonoclastic bacterium *Rhodococcus opacus* in an up-flow packed bed bioreactor with a novel waste-derived nano-biochar based bio-support material. *J Clean. Prod* **256**, 120253 (2020).
13. Guo, M. & Li, C. Current progress on identification of virus pathogens and the antiviral effectors in echinoderms. *Developmental and Comparative Immunology* vol. 116 at (2021). <https://doi.org/10.1016/j.dci.2020.103912>
14. Courtney Smith, L. et al. Echinodermata: the complex immune system in echinoderms. *Adv. Comp. Immunol.* https://doi.org/10.1007/978-3-319-76768-0_13 (2018).
15. Bordbar, S., Anwar, F. & Saari, N. High-value components and bioactives from sea cucumbers for functional foods - A review. *Marine Drugs* vol. 9 at (2011). <https://doi.org/10.3390/md9101761>
16. Poli, G., Martinelli, A. & Tuccinardi, T. Reliability analysis and optimization of the consensus Docking approach for the development of virtual screening studies. *J. Enzyme Inhib. Med. Chem.* **31**, 167–173 (2016).
17. Xia, X., Maliski, E. G., Gallant, P. & Rogers, D. Classification of kinase inhibitors using a bayesian model. *J Med. Chem* **47**, 4463–4470 (2004).
18. BIOVIA. QSAR, ADMET and Predictive Toxicology. (2020). <https://www.3dsbiovia.com/products/collaborative-science/biovia-discovery-studio/qsar-admet-and-predictive-toxicology.html>
19. Venkatapathy, R., Wang, N. C. Y., Martin, T. M., Harten, P. F. & Young, D. Structure-Activity relationships for carcinogenic potential. *Gen. Appl. Toxicol.* <https://doi.org/10.1002/9780470744307.gat079> (2009).
20. Goodman, G. & Wilson, R. Comparison of the dependence of the TD50 on maximum tolerated dose for mutagens and nonmutagens. *Risk Anal* **12** (4), 525–533 (1992).
21. Krewski, D., Gaylor, D. W., Soms, A. P. & Szyszkowicz, M. An overview of the report: correlation between carcinogenic potency and the maximum tolerated dose: implications for risk assessment. *Risk Anal.* **13**, 383–398 (1993).
22. Diaz, R. G. et al. Comparison of in Silico tools for evaluating rat oral acute toxicity. *SAR QSAR Environ. Res* **26**, 1–27 (2015).
23. Pizzo, F. & Benfenati, E. In Silico models for repeated-Dose toxicity (RDT): prediction of the no observed adverse effect level (NOAEL) and lowest observed adverse effect level (LOAEL) for drugs. in *Methods Mol. Biology* **1425**, 163–176 (2016).
24. Venkatapathy, R., Moudgal, C. J. & Bruce, R. M. Assessment of the oral rat chronic lowest observed adverse effect level model in TOPKAT, a QSAR software package for toxicity prediction. *J Chem. Inf. Comput. Sci* **44** (5), 1623–1629 (2004).
25. Wilhelmus, K. R. The Draize eye test. *Survey of ophthalmology* vol. 45 at (2001). [https://doi.org/10.1016/S0039-6257\(01\)00211-9](https://doi.org/10.1016/S0039-6257(01)00211-9)
26. DECIMER. <https://decimer.ai/>
27. Costa, A. S., Martins, J. P. A. & de Melo, E. B. SMILES-based 2D-QSAR and similarity search for identification of potential new scaffolds for development of SARS-CoV-2 MPRO inhibitors. *Struct Chem* **33** (5), 1691–1706 (2022).
28. Jin, Z. et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **582**, 289–293 (2020).
29. Mostafa, A. et al. FDA-approved drugs with potent in vitro antiviral activity against severe acute respiratory syndrome coronavirus 2. *Pharmaceuticals* **13**, (2020).
30. Martins, N., Petropoulos, S. & Ferreira, I. C. F. R. Chemical composition and bioactive compounds of garlic (*Allium sativum* L.) as affected by pre- and post-harvest conditions: A review. *Food Chemistry* vol. 211 at (2016). <https://doi.org/10.1016/j.foodchem.2016.05.029>
31. El-Gamal, K. M., El-Morsy, A. M., Saad, A. M., Eissa, I. H. & Alswah, M. Synthesis, docking, QSAR, ADMET and antimicrobial evaluation of new quinoline-3-carbonitrile derivatives as potential DNA-gyrase inhibitors. *J. Mol. Struct.* **1166**, 15–33 (2018).
32. Li, N. et al. Screening of Some Sulfonamide and Sulfonylurea Derivatives as Anti-Alzheimer's Agents Targeting BACE1 and PPAR γ . *J. Chem.* (2020). (2020).
33. Hagra, M. et al. Discovery of new Quinolines as potent Colchicine binding site inhibitors: design, synthesis, Docking studies, and anti-proliferative evaluation. *J Enzyme Inhib. Med. Chem* **36** (1), 640–658 (2021).
34. Ibrahim, M. K. et al. Design, synthesis, molecular modeling and anti-hyperglycemic evaluation of quinazolin-4(3H)-one derivatives as potential PPAR γ and SUR agonists. *Bioorganic Med. Chem* **25** (17), 4723–4744 (2017).
35. Elmetwally, S. A., Saied, K. F., Eissa, I. H. & Elkade, E. B. Design, synthesis and anticancer evaluation of thieno[2,3-d]pyrimidine derivatives as dual EGFR/HER2 inhibitors and apoptosis inducers. *Bioorg Chem* **88**, 102944 (2019).
36. Mahdy, H. A. et al. Design, synthesis, molecular modeling, in vivo studies and anticancer evaluation of quinazolin-4(3H)-one derivatives as potential VEGFR-2 inhibitors and apoptosis inducers. *Bioorg Chem* **94**, 103422–103446 (2020).
37. El-Zahabi, M. A. et al. Design, synthesis, molecular modeling and anti-hyperglycemic evaluation of phthalimide-sulfonylurea hybrids as PPAR γ and SUR agonists. *Bioorg Chem* **91**, 103115 (2019).
38. El-Naggar, A. M., Eissa, I. H., Belal, A. & El-Sayed, A. A. Design, eco-friendly synthesis, molecular modeling and anticancer evaluation of thiazol-5(4-H)-ones as potential tubulin polymerization inhibitors targeting the Colchicine binding site. *RSC Adv* **10**, 2791–2811 (2020).
39. Nasser, A. A. et al. Discovery of new pyrimidine-5-carbonitrile derivatives as anticancer agents targeting EGFR/Tand EGFR/T790M. *Org. Biomol. Chem* **18**, 7608–7634 (2020).
40. Mahdian, S., Ebrahim-Habibi, A. & Zarrabi, M. Drug repurposing using computational methods to identify therapeutic options for COVID-19. *J Diabetes Metab. Disord* **19**, 691–699 (2020).
41. El-Demerdash, A. et al. Comprehensive virtual screening of the antiviral potentialities of marine polycyclic guanidine alkaloids against sars-cov-2 (Covid-19). *Biomolecules* **11**, (2021).
42. Alanazi, M. M. et al. New bis([1,2,4]triazolo)[4,3-a:3',4'-c]quinoxaline derivatives as VEGFR-2 inhibitors and apoptosis inducers: Design, synthesis, in Silico studies, and anticancer evaluation. *Bioorg Chem* **112**, 104949 (2021).
43. Alesawy, M. S. et al. In Silico studies of some isoflavonoids as potential candidates against COVID-19 targeting human ACE2 (hACE2) and viral main protease (Mpro). *Molecules* **26**, 2806 (2021).
44. Orosz, Á., Héberger, K. & Rácz, A. Comparison of Descriptor- and fingerprint sets in machine learning models for ADME-Tox targets. *Front Chem* **10**, 852893 (2022).
45. CHEN, C. & XGBoost, T. G. A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. *San Fr. Calif.* (2016).

46. Zhang, F. & O'Donnell, L. J. Support vector regression. in *Machine Learning: Methods and Applications to Brain Disorders* (2019). <https://doi.org/10.1016/B978-0-12-815739-8.00007-9>
47. Hammoudan, I. & Chafi, M. QSAR modeling of pyrazoline derivative as carbonic anhydrase inhibitors. *Environ. Sci. Pollut. Res.* <https://doi.org/10.1007/s11356-023-28277-3> (2023).
48. ENHPCG. (2024). <https://enhpcg.sci.eg/>
49. Elshamakey, M., Maiya, A. I., Kashkoush, M. S., Fathy, G. M. & Hassan, H. A. The Egyptian National HPC grid (EN-HPCG): open-source slurm implementation from cluster to grid approach. *J. Supercomput.* **80**, 16795–16823 (2024).

Acknowledgements

We would like to thank the Department of Chemistry, University of Winnipeg, Canada, where the elucidation of structures was completely achieved under the supervision of Prof. Athar Ata.

We would like to express our gratitude to the Egyptian National HPC Grid (En-HPCg) project, funded by ASRT, for providing essential MATLAB resources. The availability of these advanced computing facilities played a crucial role in the success of this research, significantly accelerating the processing of complex algorithms and enabling a more comprehensive exploration of the proposed methodologies.

Author contributions

M.S.M.A.E.H., M.A.E.D., and D.A.G. were responsible for the study's conceptualization and design. M.S.M.A.E.H. performed the compound isolation and structural characterization. The anti-SARS-CoV-2 bioassay was conducted by M.G.S. The machine learning analyses were carried out by A.I.M., H.A.H., and S.O. The initial manuscript draft was prepared by M.S.M.A.E.H. All authors contributed to manuscript revision and editing, and approved the final version for submission.

Funding

Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-20443-6>.

Correspondence and requests for materials should be addressed to M.S.M.A.E.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025