Infection and Drug Resistance

Dovepress
open access to scientific and medical research

Open Access Full Text Article

REVIEW

In-Silico Approaches for the Screening and Discovery of Broad-Spectrum Marine Natural Product Antiviral Agents Against Coronaviruses

Zachary Boswell[1], Jacopo Umberto Verga[1,2], James Mackle[1], Karen Guerrero-Vazquez[2], Olivier P Thomas[3], James Cray[4], Bethany J Wolf[5], Yeun-Mun Choo[6], Peter Croot[7], Mark T Hamann[8], Gary Hardiman[1,5,9]

[1]School of Biological Sciences and Institute for Global Security, Queen's University, Belfast, Northern Ireland, UK; [2]Genomic Data Science, University of Galway, Galway, Ireland; [3]School of Biological and Chemical Sciences, Ryan Institute, University of Galway, Galway, H91TK33 Ireland; [4]Department of Biomedical Education and Anatomy, College of Medicine and Division of Biosciences, College of Dentistry, Ohio State University, Columbus, OH, USA; [5]Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, USA; [6]Department of Chemistry, University of Malaya, Kuala Lumpur, Malaysia; [7]Irish Centre for Research in Applied Geoscience, Earth and Ocean Sciences and Ryan Institute, School of Natural Sciences, University of Galway, Galway, Ireland; [8]Departments of Drug Discovery and Biomedical Sciences and Public Health, Colleges of Pharmacy and Medicine, Medical University of South Carolina, Charleston, SC, USA; [9]Department of Medicine, Medical University of South Carolina, Charleston, SC, USA

Correspondence: Gary Hardiman, School of Biological Sciences and Institute for Global Food Security, Faculty of Medicine, Health, and Life Sciences, Queen's University Belfast, 19 Chlorine Gardens, Belfast, BT9 5DL, Northern Ireland, UK, Email g.hardiman@qub.ac.uk

Abstract: The urgent need for SARS-CoV-2 controls has led to a reassessment of approaches to identify and develop natural product inhibitors of zoonotic, highly virulent, and rapidly emerging viruses. There are yet no clinically approved broad-spectrum antivirals available for beta-coronaviruses. Discovery pipelines for pan-virus medications against a broad range of betacoronaviruses are therefore a priority. A variety of marine natural product (MNP) small molecules have shown inhibitory activity against viral species. Access to large data caches of small molecule structural information is vital to finding new pharmaceuticals. Increasingly, molecular docking simulations are being used to narrow the space of possibilities and generate drug leads. Combining in-silico methods, augmented by metaheuristic optimization and machine learning (ML) allows the generation of hits from within a virtual MNP library to narrow screens for novel targets against coronaviruses. In this review article, we explore current insights and techniques that can be leveraged to generate broad-spectrum antivirals against betacoronaviruses using in-silico optimization and ML. ML approaches are capable of simultaneously evaluating different features for predicting inhibitory activity. Many also provide a semi-quantitative measure of feature relevance and can guide in selecting a subset of features relevant for inhibition of SARS-CoV-2.

Keywords: SARS-CoV-2, natural products, protease, methyl transferases, RNA dependent polymerases, viral transcription, genome replication, betacoronavirus

Plain Language Summary

Coronaviruses (CoVs) are a family of viruses that cause lung and intestinal illnesses in humans and animals. Generally, these are mild illnesses characterized by cold-like symptoms. However, viruses that cause severe disease have emerged over the past twenty years, first with the severe acute respiratory syndrome (SARS) epidemic in China in 2002–2003 and subsequently the Middle East respiratory syndrome (MERS) on the Arabian Peninsula in 2012. The novel coronavirus that emerged in Wuhan at the tail end of 2019 has to date killed >6.01 million people (JHU-CSSE, 2022) and collectively cost the world's economy >16 trillion dollars. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causative virus responsible for this coronavirus disease (COVID-19). The urgent need to control this family of viruses has led to a reassessment of approaches to identify and new antiviral drugs. One area of investigation has been marine natural products, compounds or substance produced by living organisms present in the marine environment. Access to collections of these natural products is vital to

Received: 18 December 2022
Accepted: 16 March 2023
Published: 19 April 2023

finding new pharmaceuticals to combat coronavirus diseases. Increasingly, advances in artificial intelligence are aiding this drug discovery process.

Introduction

The novel coronavirus that emerged in Wuhan at the tail end of 2019 has to date killed >6.01 million people (JHU-CSSE, 2022) and collectively cost the world's economy >16 trillion dollars.[1] Despite the massive research response to the pandemic,[2] there are yet no clinically approved broad-spectrum antivirals available for betacoronaviruses. The difficulty with relying solely on immunological agents for protection against coronavirus, is the high mutation rate and loss of epitope specificity between family members and variants of the same species of coronavirus.[3] This means that, in the event of a future outbreak of coronavirus, a bespoke vaccine must be developed and pass regulatory testing before it can be used to prevent the spread of the disease and treat those infected.[4]

Discovery pipelines for pan-virus medications against a broad range of betacoronaviruses are therefore a priority for preventing high mortality rates in future outbreaks.[5] Natural metabolites of plants, fungi and bacteria have long been known to have antiviral activity.[6] These organisms lack the complex adaptive immune system of animals and rely on the production of broad-spectrum, small molecule inhibitors to keep pathogenic viruses at bay.[7] Currently, however, the focus of drug discovery in relation to coronaviruses focuses on the repurposing of already characterized pharmaceuticals or using existing pharmaceuticals as structural leads.[8–11] Existing drug structural information is, however, a much smaller sample space than the population of known molecular structures, with only around 12,000 characterized members on the online database DrugBank (as of March 2022), accounting for only about 0.5% of structures available off ChEMBL. Conversely, there are currently around 400,000 natural products, which are curated in online databases and are accessible for molecular docking simulations.[12] A variety of NP small molecules isolated from photosynthetic algae (eg, phlorotannins, sulfated polysaccharides), marine bacteria (e.g. lactones), and sponges (e.g., nucleosides, sesquiterpene hydroquinones, cyclic depsipeptides, alkaloids, etc.) have shown inhibitory activity against viral species including human immunodeficiency virus-1 (HIV-1), HCV, influenza, and herpes simplex virus.[13] Examination of marine natural products from algae that are active against SARS-CoV-2 proteins presents a novel and viable approach.

Naturally occurring bioactive compounds represent a viable approach for the development of antiviral agents. Flavonoids, for example, exhibit broad antiviral and immunomodulatory activities against coronaviruses. Flavonoids are key secondary plant metabolites that have been the subject of much study for their therapeutic potential in inflammatory diseases owing to their cytokine-modulatory effects. The antiviral activity of flavonoids is realized via enzymatic inhibition of the 3C-like protease (3CLpro) the primary protease found in coronaviruses. Recently, five compounds obtained from *Camellia reticulata* and *Anastatica hierochuntica* (plants) and *Kermia aegyptiaca* (a marine gastropod mollusk), namely taxifolin, pectolinarigenin, tangeretin, gardenin B, and hispidulin, were examined for activity against SARS-CoV-2 and represent promising candidates for, for COVID-19 management.[14] In a separate study thirty-three focused marine NPs related to the pederins, mycalamides, onnamides and theopederins polyketide families were assessed using computational approaches including molecular docking and molecular dynamics simulations studies for their affinity for the dimeric form of 3Clpro. This revealed that the majority of the marine NPs examined had favorable binding scores, in particular dihydro-onnamide A, onnamide C, and pseudo-onnamide A.[15]

Access to large data caches of small molecule structural information is vital to finding new pharmaceuticals. This is because, in general, in-vitro-only de novo drug discovery, reliant on vast chemical libraries and expensive and extensive robotics, is a laborious and low yielding strategy.[16,17] Increasingly, in the age of big data and deep learning, molecular docking simulations are being used to narrow the space of possibilities and generate drug leads.[11,18,19] It would therefore be worthwhile to combine in-silico methods, augmented by metaheuristic optimization and machine learning (ML), to generate hits from within a virtual library of natural products in order to narrow down high throughput in vitro screens for new drug targets against coronaviruses. In this article, we will explore current insights and techniques that can be leveraged to generate broad-spectrum antivirals against betacoronaviruses using in-silico optimization and ML.

Drug Targets

Members of the genus betacoronaviridae are enveloped (surrounded by a bilipid membrane) and positive sense RNA viruses (which means the genome can be used directly in translation), that share a unique lifecycle and 50–80% sequence homology (between SARS-CoV, SARS-CoV-2 and MERS-CoV).[20] Of the 29 open reading frames in its small genome, four encode for structural proteins: S (Spike), important for host recognition and attachment, M (membrane) and E (envelope), which mediate bilipid fusion in entry and release of the virus, and N (nucleocapsid), which forms the protective protein shell around the viral genome and mediates assembly of the final virion after replication.[21] Structural proteins are essential to the entry and assembly of viruses and can therefore be targeted by antiviral therapies (Figure 1). Despite the important role of these proteins in the viral life cycle and their utility as targets for extracellular drug therapy, M, E and N proteins are nonetheless unattractive targets for the development of a broad-spectrum antiviral due to the high variability in protein sequence across the phylum.[21,22] On the other hand, the ACE2 binding domain of the S protein, as determined by a team at Tsinghua University, Beijing, shows high structural conservation across the family of coronaviruses most related to SARS-CoV and SARS-CoV-2, making it a promising target for inhibitory binding.[23]

Due to the fact that structural proteins are most exposed to host immune recognition and are therefore subject to stronger selection pressure, intracellular viral proteins are far more conserved.[22] The most important non-structural proteins (NSP) to the betacoronavirus life cycle are an RNA dependent RNA polymerase (RdRp) and two selective proteases, a serine-type protease (M^{PRO}) and a papaine like protease (PL^{PRO}).[22] RdRp is essential for the reproduction of all non-retroviral RNA viruses that infect animals (animals do not express endogenous RdRp) and many that infect plants, and in positive sense RNA viruses it is used to first create a template negative strand and then replicate the genome aided by an RNA helicase.[24] The RdRp protein is highly conserved in betacoronavirus with similarity of 96% between SARS-CoV and SARS-CoV-2, and 70% between MERS-CoV and SARS-CoV/SARS-CoV-2.[25] The genome of most plant viruses are single positive strand RNA, which means there are potentially many phytochemical inhibitors of RdRp available from natural sources.[8,26]

Proteases M^{PRO} and PL^{PRO} are essential for post-translational processing in the viruses replicative cycle.[22] Figure 2 shows the despite a low percentage identity between PL^{PRO} protein sequences in SARS and MERS coronaviruses (30–28%), structure-based multiple protein alignment revealed a conserved homology in the proteolytic active site (core root mean square deviation (RMSD) of 1.47). A further search of this conserved sequence on the Conserved Domain Database, showed that the catalytic domain preserves the cysteine protease catalytic triad (Cys, His, Asp) even with the most diverse members of the group.[27] Many plant viruses use a papain-like cysteine protease for reproductive

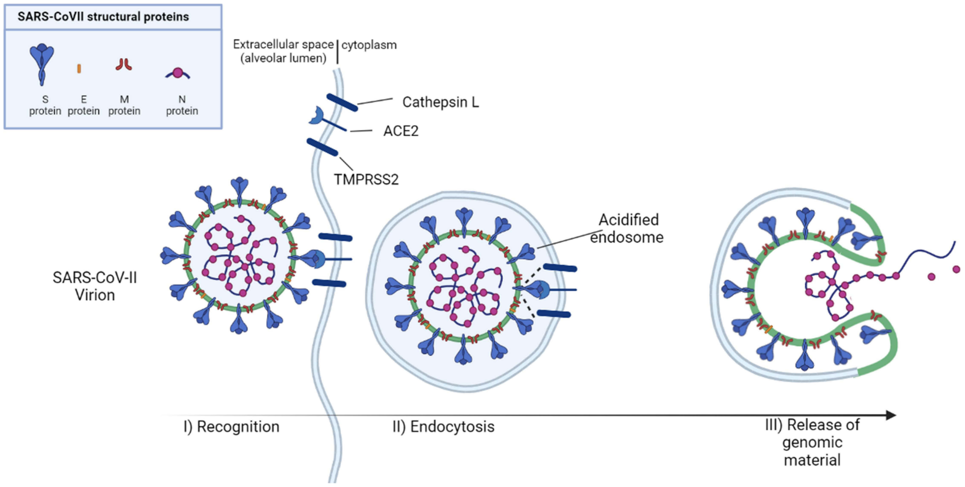

Figure 1 Function of structural proteins S, E, M and N in the mode of entry of SARS-CoV-2. I The S protein binds to ACE2 on the host cell membrane initiating endocytosis. II: host acidifies endosome, causing TMPRSS2 or Cathepsin L mediated proteolysis of S protein. III: Proteolysis of S protein causes membrane fusion, facilitated by E and M proteins, and release of viral genome. Nucleocapsid disassembles. (Su et al, 2021). Figure made using BioRender© at Biorender.com.

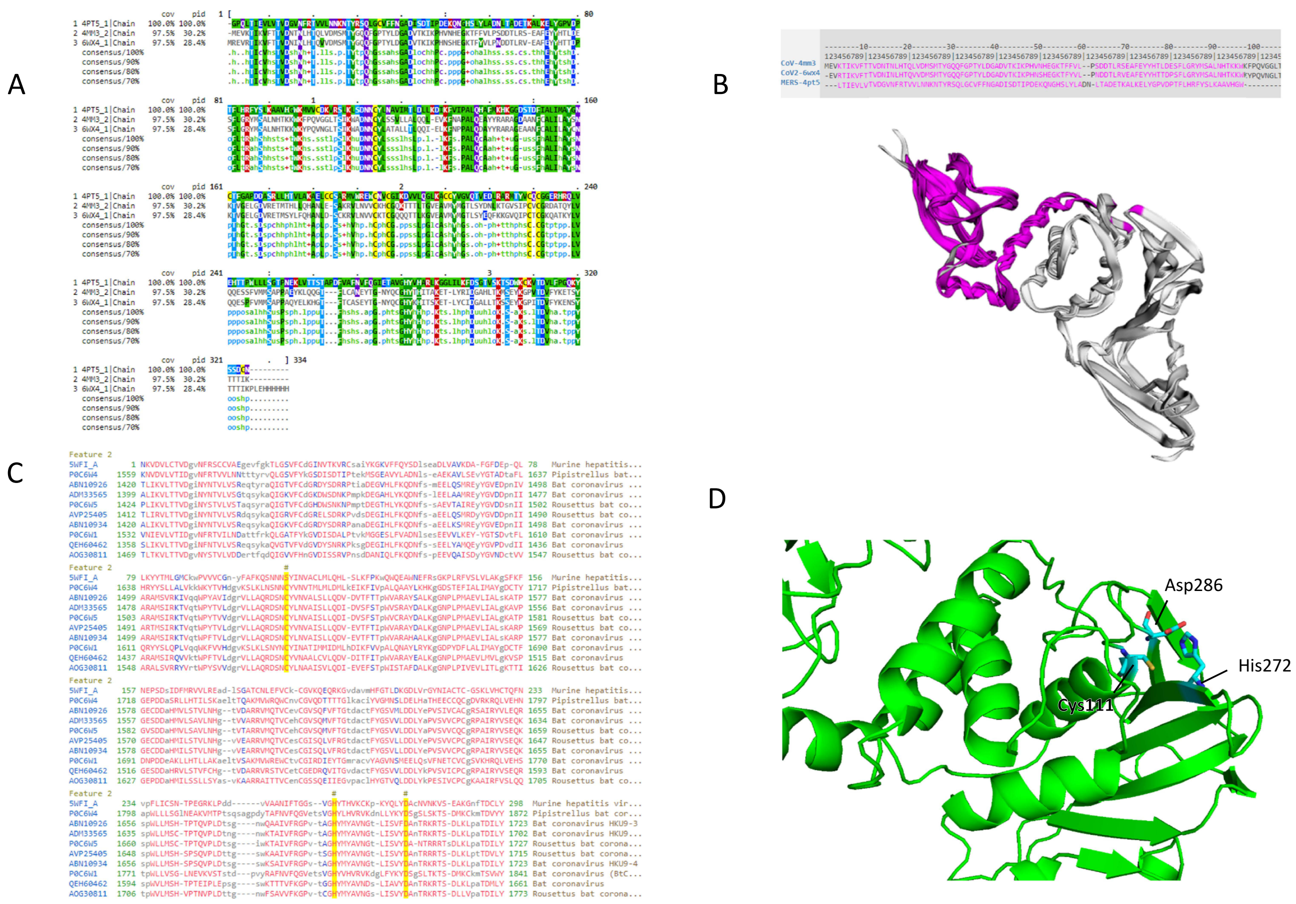

Figure 2 Important structural features of PL^{PRO} are conserved across the betacoronavirus family. (**A**) ClustalOmega alignment of PL^{PRO} peptide sequence from MERS-CoV (first sequence, PDB: 4PT5), SARS-CoV (second sequence, PDB: 4MM3) and SARS-CoV-2 (third sequence, PDB: 6wx4), showing low sequence homology. (**B**) Protein 3D structure alignment with mTm-align (Dong et al, 2018) showing a common core with an RMSD of 1.48 (shown in magenta in both sequence view and 3D view) around the catalytic and peptide binding site. (**C**) Conserved Domain Database (NCBI) entry for *betacoronavirus* PL^{PRO} conserved catalytic domain, showing alignment of most diverse members. Conserved catalytic triad, Cys, His and Asp, are highlighted in yellow (Lu et al, 2020). (**D**) SARS-CoV-2PL^{PRO} (PDB: 6wx4) showing conserved catalytic triad. Figure made using Pymol© by Tubiana; Open-Source PyMOL is Copyright © Schrodinger, LLC. All Rights Reserved.

cleavage[28] which indicates there may be phytochemical inhibitors of PL^{PRO}.[29] The M^{PRO} protease is similarly a popular target of competitive inhibition due to its conserved serine-like proteolytic domain.[27]

2‘-O-MTase protein is another promising betacoronavirus target, which binds to S-adenosylmethionine (SAM) to facilitate 2‘-O-ribose methylation. It exists as a heterodimer made of two non-structural protein units, ie NSP16 and NSP10. NSP10 assists in the stabilization of NSP16 and SAM binding sites.[22,30] A previous study has shown that the NSP16 and NSP10 protein sequence is highly conserved in SARS-CoV and SARS-CoV-2 with similarity values of 99.7% and 99.3%, respectively. In the case of MERS-CoV, the similarity with SARS-CoV/SARS-CoV-2 is lower at 53.9% and 63.8%, for NSP16 and NSP10, respectively. Despite the low similarity between SARS-CoV/SARS-CoV-2 and Mers-CoV, the 2‘-O-MTase protein is the highly conserved SAM binding pocket with similar interacting residues in betacoronavirus.[31]

The highly conserved binding pocket of the betacoronavirus protein targets (such as in M^{PRO}, PL^{PRO}, 2‘-O-MTase, and RdRp; Figure 3) opens up the potential for discovering a compound, which can bind effectively to a similar target protein of different species or variants from this genus to inhibit the protein’s activity. This strategy paves the way for the development of a broad-spectrum antiviral agent as a treatment for coronavirus infection.

Three strategies currently exist for developing new drugs targeted against SARS-CoV-2. The first is focused on existing broad-spectrum anti-virals including interferons, ribavirin, and cyclophilin inhibitors, which are all employed to treat pneumonia caused by coronavirus. The limitations with these are that they are too “broad-spectrum” and cannot neutralize coronaviruses in a targeted fashion. The second approach exploits existing molecular databases to screen for

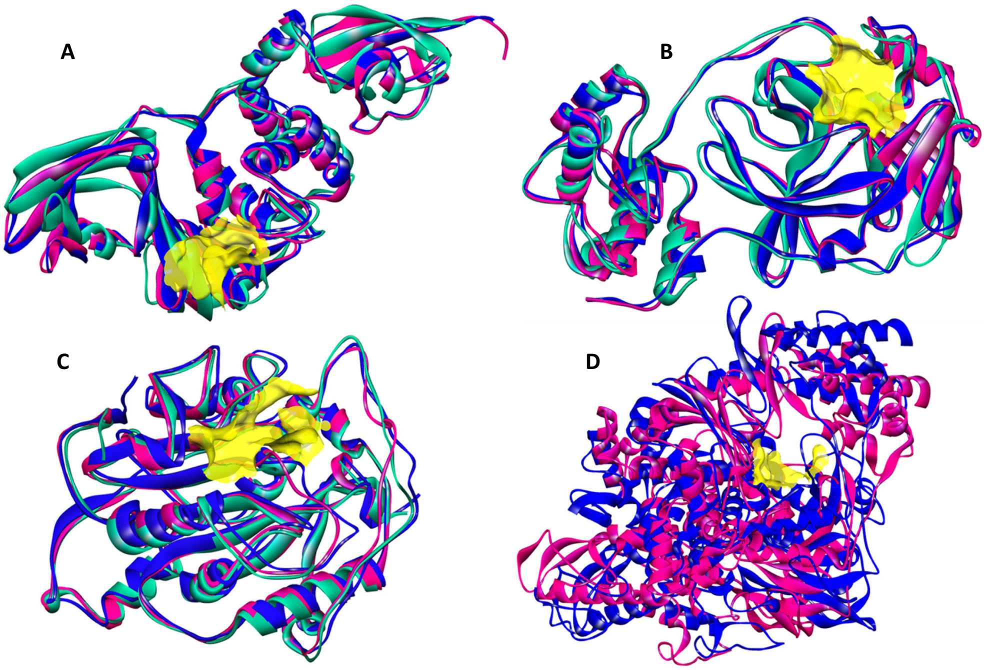

Figure 3 Ligand binding sites (yellow) of betacoronavirus proteins. (**A**) Superposition of PL^{PRO} of SARS-CoV-2 (PDB: 7cjm, blue), SARS-CoV (PDB: 3e9s, red), and MERS-CoV (PDB: 5w8u, green). (**B**) Superposition of M^{PRO} of SARS-CoV-2 (PDB: 6luv, blue), SARS-CoV (PDB: 3v3m, red), and MERS-CoV (PDB: 4rsp, green). (**C**) Superposition of 2'-O-MTase of SARS-CoV-2 (PDB: 6w4h, blue), SARS-CoV (PDB: 3r24, red), and MERS-CoV (PDB: 5ynb, green). (**D**) Superposition of RdRp of SARS-CoV-2 (PDB: 7bv2, blue) and RS-CoV (PDB: 6nur, red). Figure made by YMChoo using BIOVIA, Dassault Systèmes, Discovery Studio Visualizer, v21.1.0.20298 software program was used to perform the calculations and to generate the graphical results.

molecules that may have therapeutic effects on coronavirus. The third strategy leverages genomic information and the pathological characteristics of different coronaviruses to develop novel targeted drugs. Theoretically, these therapeutics would exhibit better anti-coronavirus effects, however the research and development required could take greater than a decade.[32]

Molecular Docking Algorithms

Molecular docking is a widely used technique to determine a preliminary assessment of how well a ligand interacts with a drug target.[33,34] The general approach of such algorithms is to use a predetermined structure of the macromolecular target (generally formatted as a PDB file) and 1) subdivide the entire structure or predetermined binding pockets into a search grid, 2) randomly place the ligand structure inside the grid, 3) assign the binding energy of the ligand-target pose based on a scoring function, 4) repeat steps 2 +3 exhaustively or until a threshold is achieved using heuristic optimization and 5) report this threshold. Several approaches can be used to shrink the search space as a non-parametric examination, otherwise it would become too computationally expensive for a library search.[33] The number of poses that must be calculated is proportional to the size of the assigned grid and this dimension can usually be reduced by knowing predetermined binding sites for substrates (for example, in order to inhibit the S protein an inhibitor must bind to the ACE2 binding site).[23] If important sites are not known or if allosteric binding is desired, "hot spots" can be determined before the search, by using cavity search algorithms (such as the one used by MolDock)[34] or by using small molecular fragment probes to search to entire protein surface (small molecules have fewer poses and have, theoretically, better binding energies than more complex ligands).[35] BetaCoV PL^{PRO} has 3 conserved sites: a peptide binding site/catalytic domain, a zinc binding site and an ubiquitin-binding domain.[27] All three appear to be important for the function of this protease and the life cycle of the virus.[29] Establishing the hotspots of these domains and the drugability of each binding site might be beneficial for guiding a library search. Figure 4 shows three binding hotspots found by Fragment Hotspot Maps,[35] showing overlap with all three conserved domains.

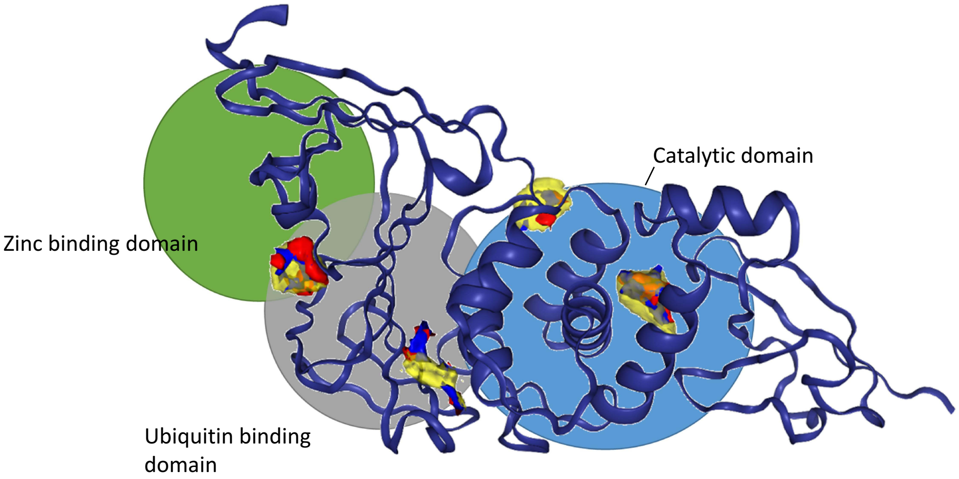

Figure 4 SARS-CoV-2 PL^{PRO} showing drugability hotspots in conserved domains. Fragment Hotspot Maps (Radoux et al, 2016) returned 4 binding hotspots (hydrophobic binding in yellow, hydrogen bond donor in red and hydrogen bond acceptor in blue) on top of the three active domains (blue circle – catalytic domain, grey circle – ubiquitin binding domain, green circle – zinc binding domain). Figure made using Pymol© by Tubiana; Open-Source PyMOL is Copyright © Schrodinger, LLC. All Rights Reserved.

Although systematic, exhaustive search strategies are sometimes employed by molecular dynamics (MD) algorithms, heuristic methods are preferential with ligands with more degrees of freedom.[33] Different meta-heuristics are used to explore the search space efficiently, including a MonteCarlo approach (SMINA and ISO software), genetic algorithms (DockThor, GAsDock and GOLD) and particle swarm optimization (Plants). In general, Monte Carlo approaches randomly sample poses, identify clusters of high scoring poses and iteratively search inside the clusters for optimums.[36] This algorithm works excellently for finding global optima while avoiding "trapping" in local optima, however it requires a large amount of sampling, scaling exponentially with the degrees of freedom of the ligand. Genetic algorithms (the most common approach) improve on Monte Carlo based approaches by reducing the amount of stochastic sampling. These algorithms 1) select the highest scoring poses in a random sampling, 2) "mutate" the samples by randomly altering their pose, 3) shuffle the bond rotations and average the position between two "mating" pairs (also called recombination), 4) retest, select the highest scoring poses and repeat 2–4 until a local maximum is found or a pre-set number of generations have elapsed.[37]

Database Searching

There are around 400,000 structures in the COlleCtion of Open Natural ProdUcTs (COCONUT),[12] which means that an iterative search for a new inhibitor of a target protein, if each search took 2 minutes, would take 91 years and 2 months to complete (of course only if searches were run consecutively). Even with concurrent searches, iterative strategies are computationally expensive and would require a reduction of accuracy by using simpler approximations.[33]

A popular non-systematic approach is to use a fragment library, made from a library of structures, to iteratively build an "ideal inhibitor". This approach can use a genetic algorithm or a deterministic meta-heuristic model, such as simulated annealing to build a stochastically optimized molecule.[38,39] Each candidate can then be used to search the library of natural products for "real candidates" based on 3D similarity. Supplementary Table S1 provides a summation of the metrics that can be used to compare an ideal molecule to a library of real molecules. Based on these metrics, the top candidates from the library can then be tested by molecular docking in a small number of similar poses. This process can be repeated until a list of high scoring molecules are available.[39]

Alternatively, a meta-heuristic approach can be used to search an organized library. We note that the similarity scores based on metrics from Supplementary Table S1 and information from Table 1 can be used to arrange the library of molecules as nodes in n-dimensional space with edge weights being proportional to 3 dimensional similarity (Supplementary Table S1) and similarity of attributes (Table 2; Figure 5A).[40]

Table 1 Quantifiable Molecular Attributes That Can Be Used to Classify Molecular Similarity

Molecular Attributes	Description	Parameters and Assumptions
Quantitative Structure Activity Relationship, QSAR[96]	Based on Lasso Regression of attributes of characterized drugs versus their quantified activity.	Attribute list can be generated based on observation or generated by machine learning.
Differential solubility[97]	LogP = log base 10 of partition coefficient in a n-octanol- water system. Log P must not differ much from zero for differential solubility to be high. DS = |logP|Where |x| is the absolute value.	Assumption that drug must be taken up (hydrophilicity) and diffuse across barriers (lipophilicity).
Diffusability[97]	Inversely proportional to molecular weight.	Assumption that molecular size is inversely proportional to uptake by cells or diffusion across biological barriers.

Table 2 Databases Containing Natural Marine Compounds

Database	Marine Compounds	Link	Note
Seaweed Metabolite Database	1,110	https://www.swmd.co.in/	Only from algae.
MarinChem3D	30,117	http://mc3d.qnlm.ac/	All marine compounds.
Comprehensive Marine Natural Products Database	32,000	https://www.cmnpd.org/	All marine compounds.
Dictionary of Marine Natural Products		http://dmnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml	Paid access database.
MarinLit	35,790	http://pubs.rsc.org/marinlit	Paid access database.
MetaboLights	Not specified	https://www.ebi.ac.uk/metabolights/index	Generic Metabolites database, it is possible to select single species.

We propose that this network be explored by a particle swarm algorithm (Figure 5B and C) where only local optima are recorded to avoid over-representing structural homologues and a tabu list be maintained to ensure that the same molecule is not re-tested.[41] After a list of high scoring molecules is extracted from the organized library, molecular docking of homologues surrounding the optimum in the network can be used to extract a pharmacore or to identify plant extracts for in-vitro validation.

In-Vitro Validation

The main weakness of in-silico drug design is the unrealistic approximations that must be used to shrink the computational requirements of a given search. Often, quantum approximations and empirical scoring functions are used to

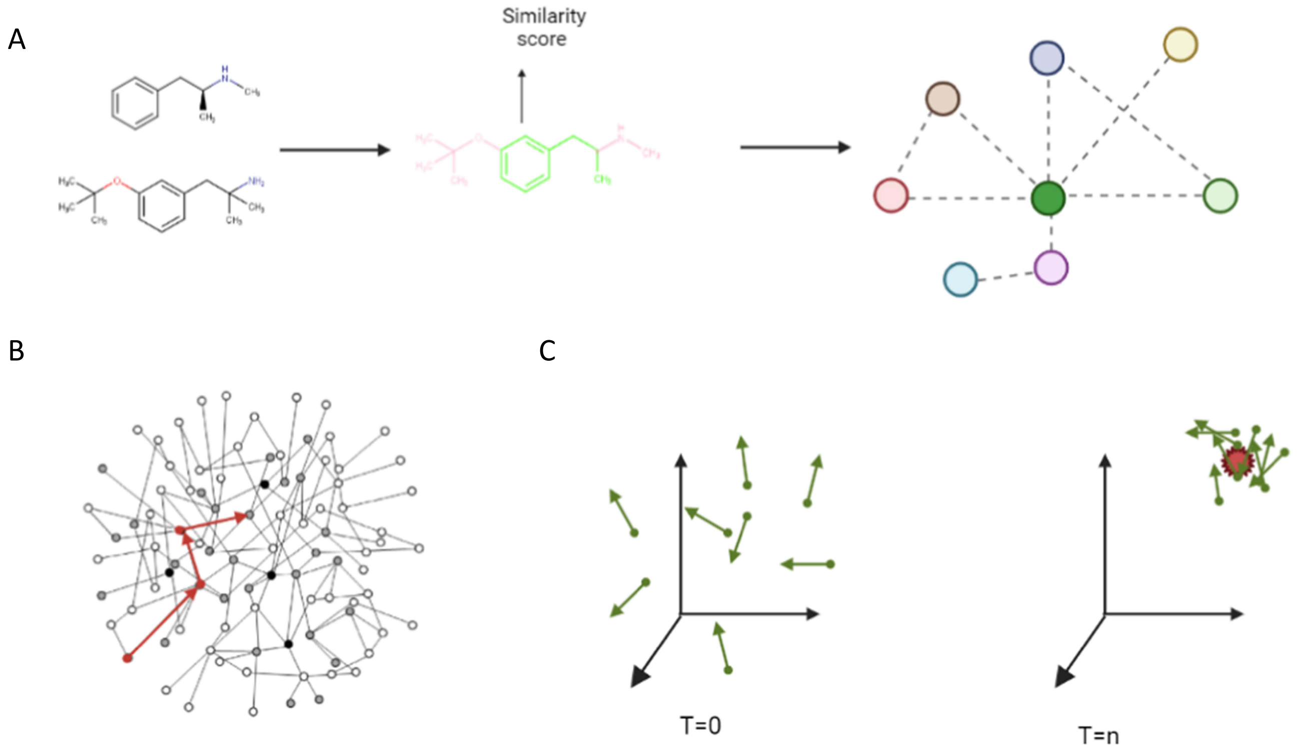

Figure 5 A meta-heuristic strategy for searching a large chemical database for molecular docking. (**A**) Construction of the Network: For each pair of molecules, common structures are selected using a similarity score based on predetermined metrics. If the maximum similarity score is higher than a threshold, it is marked as a connection represented as an edge in the network constructed using a "push and pull" technique.[40] (**B**) Swarm particle approach: A n number of nodes are selected to 'host' a particle, where the vector that the particle will take is defined by the adjacent neighbors of the node where the particle is a given momentum t based on an initial random velocity.[41] Each node the particle travels through is a structure that is computed by molecular docking. (**C**) Swarm particle convergence: The swarm particle algorithm starts at time 0 with many random particles exploring the space through random vectors with their own magnitude, direction, and sense. With each iteration, the particle's vector will change with a velocity that is influenced by the optimal ligand that it has uncovered. Eventually, the particles will converge on an optimum at time.[41] Figure made using BioRender© at Biorender.com.

simplify that calculation of binding affinities. Furthermore, receptor flexibility is generally not allowed in docking simulations due to the added search space entailed in moving both the ligand and the target.[42] As a result, in-silico evaluation is best suited to a "hit to lead" approach, where a wide list of potential leads (termed "hits") are assayed in vitro to narrow down the "true leads".[10,11,18,19,21,29,33,39,43]

Selected natural extracts from in-silico analysis can then undergo high throughput screening for inhibitor activity. Techniques for finding PL^{PRO} or M^{PRO} inhibitors can utilize the proteolytic rate constant as an indicator of activity,[44] by for example, using fluorescent polarization of fluorescently labelled peptides.[45] Alternatively, a technique such as surface plasmon resonance can be used to directly measure KD dissociation constants.[11,33] Once an extract is confirmed as being inhibitory, chromatographic fractions can be tested in another round of high throughput screening to identify the molecular formula of the inhibitor.[29]

In 2021, this group proposed that in-silico drug discovery could be exploited to generate a series of hits that would be traced back to natural product extracts, and that subsequently these extracts would be assayed using a mass spectrometry technique that allows the detection of ligands via mass signatures. This type of high throughput workflow allows for the rapid creation of natural, antiviral drug leads and could effectively lead to a broad-spectrum prophylactic/therapeutic for future betacoronavirus pathogens.[46]

Machine Learning and Drug Discovery

The terms machine learning (ML), artificial intelligence (AI) and deep learning (DL) are often used interchangeably but have distinctly different meanings, AI is the field of computer science that simulate intelligent behavior utilizing their environment as input. ML is a branch of artificial intelligence based on the idea that systems can learn from data, identify

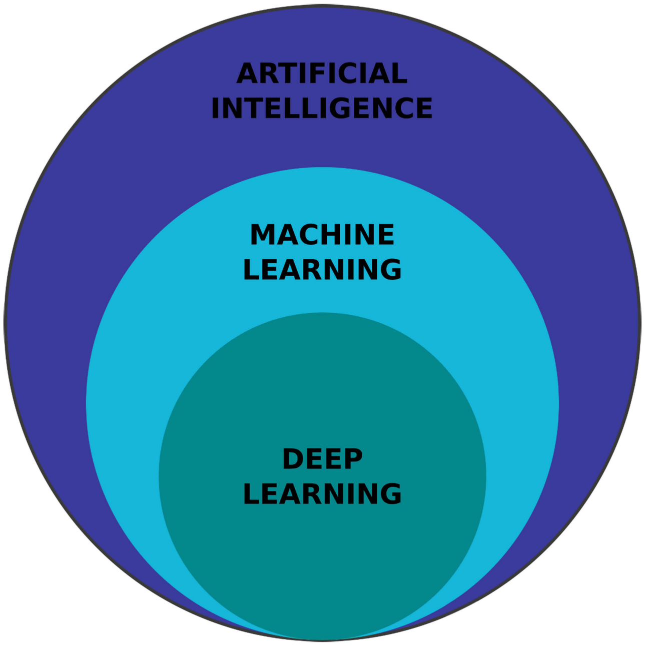

Figure 6 Graphical representation of the relationship between Artificial Intelligence, Machine Learning and Deep Learning.

patterns, recognize behaviors, and make decisions with minimal human intervention,[47] while DL are the implementation of representation-learning methods with multiple levels of representation,[48] being then a subset of ML and, it is commonly found implemented on neuronal networks (Figure 6). ML Is not a new term, it was coined in 1959 by Arthur Samuel an IBM engineer describing a checkers game.[49] The capabilities of ML algorithms have matured in the past 20 years as computers and computational resources have gained in sophistication. Historically, the utility of ML algorithms has been limited by data availability and computational power. Since the 2000s ML tools have become more accessible, with several open-source libraries providing powerful and accessible methods. These libraries have enabled faster and wider application and implementation of ML algorithms to various tasks. The recent uptake in ML has seen it applied to a variety of tasks, including stochastic prediction and data classification.[50] ML models have proven useful on a myriad of data types and tasks.

A range of open-source ML libraries available in various programming languages have been developed. The most common is Python, which enables the use of Scikit-learn,[51] Keras,[52] TensorFlow,[42] Facebook AI Similarity Search (Faiss)[53] and PyTorch.[54] Their aptitude for pattern recognition has made these libraries useful in biomedical applications including neuroimaging, general pathology, and protein-folding.[55–57] Just as there are many applications for ML, there are also many different machine models to choose from when performing a task. Types of ML can be broadly and not strictly categorised by their functionality, such as classification, prediction, or clustering or by the degree of user input, such as supervised, semi-supervised and unsupervised.[58,59]

ML algorithms do not always provide the optimal solution to a task as some data sets are simply not large enough to train an algorithm sufficiently. Sometimes, it is not feasible to analyze all the possible patterns one by one, in this cases, instead of getting an absolute deterministic solution, the ML algorithm will apply shortcuts to reduce the space search area and provide a good enough solution, like in the case of heuristics.[60] Choosing an algorithm(s) for a given application is also an exhaustive process, due to the fact that no algorithm will be better than the other in every application and it is highly dependent on the particular case. Once the algorithm is selected, it is crucial to select the model that better generalizes the data,[60] one way to do so is the tuning of the parameters. The influence of the refining can introduce bias if not done carefully.

In addition to these selecting and refining steps the hyperparameters of the algorithms can further introduce bias. The field of automated ML (auto ML) offers a potential solution to these problems. It utilizes ML techniques to automate data selection and preparation, model selection and hyperparameter optimisation, potentially reducing bias and allowing effective comparisons exploiting many algorithms and configurations.[61] ML algorithms can result in overfitting, where the models produced fit too closely to the training data and are not representative of the reality. ML practitioners implement techniques such as the holdout method and cross validation to reduce the risk of overfitting.[62–64] ML models are further limited by the available data, the quality of data pre-processing, selection, or reduction, the hyperparameters chosen and time constraints.

A common limitation when implementing ML, especially when considering biological data, is the high dimensionality of them and a low number of observations. Too many features can exponentially increase the computational strain involved in training, testing, and implementation. An excess of features can also reduce the effectiveness of an algorithm by introducing noise. To reduce the number of features, feature selection, and feature reduction techniques can be implemented. Feature selection aims to extract the most relevant features from a dataset.[65] This can be manual or automatic, with many of the automatic feature selection methods relying on statistical methods. Feature extraction aims to reduce the number of features by using the existing features to create new ones, lower in number and more representative, and discarding the old ones. Many feature extraction methods are in themselves unsupervised ML algorithms, examples include principal component analysis (PCA), Uniform Manifold Approximation and Projection (UMAP) and independent component analysis (ICA) (18).

In supervised learning, data is divided into training and test datasets. The performance of algorithms is assessed using the test data, which can inform changes to the hyperparameters and features. Testing data cannot be used to improve the model performance.[50] In the drug discovery field ML algorithms enable searches of large databases for potential therapeutic compounds of interest, the results of these algorithms are then used to inform further studies, streamlining searches for potential drugs. In drug discovery, there are various possible features, which are often categorised by the dimension they describe. For example, 0D descriptors include molecular weight and counts of specific atom types such as heavy atoms. 1D descriptors describe the 1-dimensional features of a molecule such as acetyl or hydroxyl functional groups; 2-dimensional descriptors include topological features such as polarity number and Wiener index; 3-dimensional descriptors include geometrical molecular descriptors and steric properties.[66] Other descriptors include stats on their simplified molecular-input line-entry system (SMILES).[67–69]

One of the first tasks when developing an ML algorithm for drug discovery is determining which combination of descriptors work most effectively with given datasets. These data are available for both ligands and the receptors they bind to, to train the ML algorithm(s) to find any underlying patterns between which molecules are likely to bind to each other, and then to test the efficacy of the trained algorithm using methods such as cross validation. ML algorithms have found applications alongside classical approaches from the early stages involving screening of compounds libraries, to the later phases including clinical trials, enabling improvements in the accuracy and speed of the drug discovery process.

One key factor used to identify suitable candidates is the estimation of the pharmacokinetics properties, ie, absorption, distribution, metabolism, excretion, and toxicity (ADMET). ADMET provides important insights into the behaviours of the compound in the living organism, including bioavailability and toxicity.[70] Since the high complexity of biological organisms and chemical reactions, the production of a mathematical model capable of correctly taking properties represents a very challenging problem. Quantitative structure–activity relationship (QSAR) models, that try to relate chemical data to biological properties, and propose mechanisms of interaction among compounds and the target protein, were widely used to estimate ADMET properties. Both techniques, usually based on molecular descriptors (including fingerprint, 2D and 3D descriptors and steric parameters), have benefited from the introduction of ML approaches. Supervised ML methods predict single or multiple properties of the selected compounds from independent variables (single and multitask models) to reduce the number of experimental failures. Unsupervised ML models using unlabeled data, can be used to identify new scaffolds (scaffold-hopping) or compounds with query-like properties.[70,71] Several ML algorithms are in use in QSAR and ADMET estimations, including k-nearest neighbour, random forest (RF), support vector machines (SVM), principal component analysis and deep networks, to both perform properties predictions and similarity analysis to obtain compounds with similar features from large datasets.[67,70]

Multitask Deep Neural Networks (DNNs) are used to infer small-molecules properties and activities and make predictions about the readout of a molecule in a new experimental setup.[72,73] These methods significantly boosted the classical statistical methods used before, due to the fact that they are faster and can efficiently predict relevant pharmacological parameters from large datasets, reducing the number of compounds to be tested in virtual screening (VS). The scoring functions in molecular docking programs must be able to position the ligands in the best available pose (docking power). Subsequently, the scoring function must correctly estimate the binding affinities based on the poses obtained (scoring power). Furthermore, it must classify the different compounds as good or bad ligands according to the obtained poses (ranking power).[71] The classical scoring functions are based on force-fields, empirical-based, or knowledge-based. Here, ML approaches are used to improve the accuracy of binding energy predictions in various ways: SVM and RF trained on ligand-protein complexes described as geometrical features or chemical descriptors, RF-score based on features deriving from different docking programs and SVM to predict IC_{50} of protein inhibitors.[71] Also, given a set of bound ligands to a target protein, the ranking power should correctly find the top hits among ligand poses. Here, for each experimentally obtained protein-ligand complex and relative binding affinity, the features from different scoring functions have been retrieved, the set of ligand-protein featured has been used to train 6 ML models able to rank the complexes and predict the highest, median and lowest binding affinity.[71] Also, a non-parametric ML approach was proposed to build target-specific scoring functions, this method is considered useful in lead optimization and prediction of the best novel ligands. The docking power identifies the best binding pose of a ligand given a set of poses. The most stable complexes reside in the minimum of energy; different methods to search in the space are used to reach these configurations: Steepest descendant optimization, simulated annealing, stochastic approaches (Monte Carlo) and Molecular Dynamics. In this field SVM and RF insensitive to the docking pose accuracy has been proposed.[71] Further improvements in ML approaches in docking power will improve the effectiveness and accuracy of docking simulations. Finally, the screening power is the ability to identify molecules that can effectively bind the target among random molecules. Here, scaffold hopping techniques are used to identify novel ligands based on similar chemical positioning. In ML aided VS approaches, an interesting procedure was used to build a discriminating SVM model, based on extended-connectivity fingerprints (ECFPs) and physical properties, to discern putative inhibitors of c-Met to other compounds from large libraries, the virtual screening carried only on the putative inhibitors revealed the better performance and more accurate compound selection for downstream analysis.[74]

ML has also been applied to compound synthesis after dereplication. Reaction condition recommendation remains an indispensable aspect for achieving computer-assisted synthesis. Accurate reaction conditions are needed for experimental validation and exert a significant effect on the success or failure of an attempted transformation. De novo condition selection has traditionally relied on chemists' background knowledge and expertise. Neural-network models are being exploited to predict the chemical context (ie, catalyst(s), solvent(s), reagent(s)), in addition to the optimal temperature for any given organic reaction.[75,76] Pairing chemistry and ML, via data-driven analyses, neural network predictions and monitoring of chemical systems, is adding to (i) our understanding of the complexity of chemical data and (ii) experimental design and streamlining.[76]

Integration of ML in Screening of Marine Natural Products

Examination of marine compounds active against SARS-CoV-2 proteins via virtual screening with ML-integrated approaches presents a novel and viable approach. Firstly, we need to consider available databases containing marine compounds to screen. Genomic databases could be considered in the future if a family of related compounds are identified. An example of how this might be achieved for natural products, ie, identifying biosynthetic signatures in genomic data, predictions of what structures will be created from those genomic signatures, and the types of activity one might expect from those molecules is reviewed by Prihoda et al.[77]

Today, available resources for marine compounds are limited to five commercial or free of charge databases of marine natural compounds. The two commercial databases are MarinLit[78] and the Dictionary of Marine Natural Products.[79] Both databases are the most comprehensive marine natural compounds databases available, containing data from publications, synthesis, organisms, and biological activities.[80] According to MarinLit, the database contains >35,000 articles. The open databases available are highlighted in Table 2.

- The Seaweed Metabolite Database (SWMD) contains 1110 compounds from Brown, Green and Red algae.[81]
- MarinChem3D contains more than 30,000 compounds with defined 3D structures and molecular descriptors.[82]
- The Comprehensive Marine Natural Products Database (CMNPD) contains more than 32,000 marine compounds with various physicochemical and pharmacokinetic properties and biological activity data.[38]

In addition to these databases, MetaboLights[83] for metabolomics experiments, contains compounds from many sources, reducing the research only to the species of interest. Structural databases including ChEMBL, ZINC,[84] PubChem,[84] DrugBank[85] and ChemSpider[86] include several molecules including already patented drugs, natural compounds, bioactive compounds and related properties. In addition, MoleculeNet[87] is a benchmark designed for testing ML methods of molecular properties. It contains features for >700,000 molecules. MoleculeNet, developed the python library DeepChem,[87] containing several ML and DL algorithms, and tools to retrieve descriptors and allow preprocessing of data. The direct connection of DeepChem to the MoleculeNet datasets, including toxicology, solubility and biological activities makes DeepChem an important tool in ML aided drug-design. Considering retrieving data from one of the two largest free databases (MarinChem3D or CMNPD), a dataset of >30,000 compounds will be available that can reach higher numbers if added with inhibitors of the SARS-CoV-2 target-proteins (ie, 159 for the 3CL^{pro}), according to ChEMBL, ID: CHEMBL3927 (Figure 7).

To reduce virtual screening times, and the number of false hits due to compounds with bad pharmacokinetic properties or toxic effects, integrating supervised and unsupervised ML approaches offers an advantage. Screens can focus on those compounds that demonstrate structural or pharmacokinetic similarities with already known

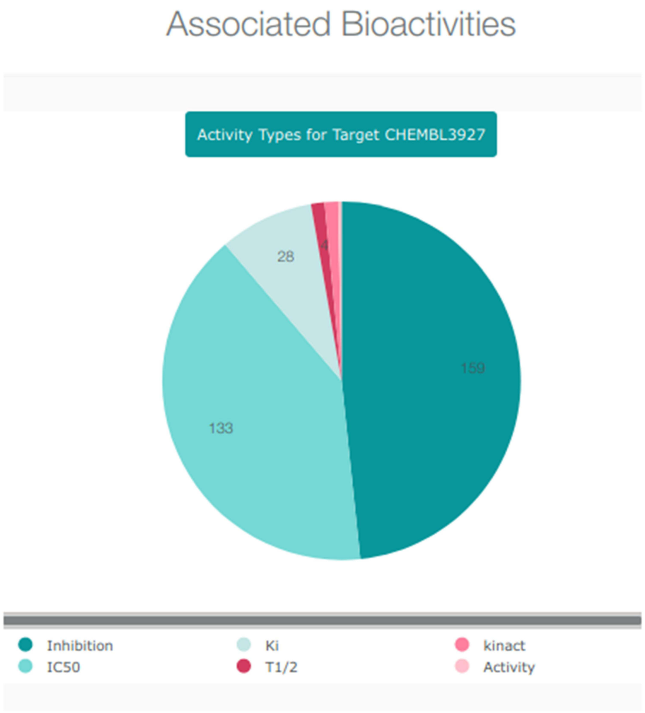

Figure 7 Compounds with associated activities specifically against 3CL^{pro} of SARS-CoV-2. ID: CHEMBL3927. Creative commons. Available from: https://www.ebi.ac.uk/chembl/g/#browse/activities/filter/target_chembl_id%3ACHEMBL3927.

drugs. To focus research on new molecular scaffolds with suitable pharmacophoric properties, a multitask DNN could be utilized to learn a model predicting multiple pharmacokinetics properties (including absorption, distribution, metabolism, excretion, and toxicity) from physicochemical properties.[70,72]

The prediction of important pharmacokinetic properties will reduce the number of molecules to be screened and focus the analysis on those compounds with optimal pharmacophoric characteristics, reducing the risk of finding high affinity but unsuitable compounds. Furthermore, the model trained based on physicochemical properties will allow research unaffected by pre-existing structural patterns and scaffolds. On the other hand, including already known inhibitors and descriptors (pharmacophore and structural fingerprints) allows for unsupervised approaches to group the datasets based on the similarity among the compounds and the inhibitors.[69] A fuzzy clustering will provide overlapping clusters, the molecules can be classified in multiple clusters, to better represent the similarity relationships in the dataset. The selection of the clusters containing the inhibitors will make the virtual screening focused only on molecules that share similarities and, probably, similar binding affinities.

Datasets of protein-ligand complexes, obtained from PDB (https://www.rcsb.org/), can be used to train ML models to better estimate ligand–protein interactions, binding energies and rankings, improving performance and accuracy of the molecular docking analysis.[71] Once obtained the top hits from the molecular docking process, ligand-protein complexes should be validated using MD simulations. These analyses are based on force fields to approximate the intra- and inter-molecular interactions among solvent, protein, and ligand. Despite the high computational costs, MD makes it possible to study numerous factors such as optimization of the poses, the stability of the complexes, interactions, and binding energies. Since MD simulations require high computational costs, ML approaches have been proposed to replace or work with MD simulations based on force-fields. Starting from a dataset containing well-defined ligand and protein sequences, added with the fingerprints relative to the ligands, some ML models have been trained to predict the binding pose and energy of the ligand-protein complexes.[88–90]

These models are designed to work in the very initial stages of the drug-design process, the robustness highly depends on the variety and quality of the training datasets. Similar ML approaches can be used to reduce the number of compounds to be screened and eventually, depending on the robustness of the model, be used to skip docking and MD simulations. Also, several ML approaches have been proposed to enhance the speed and accuracy of force-fields and scoring functions, cutting the time required for MD simulations.[91–93] These ML models have been trained on 3D properties (X-Ray, *ab-initio* MD) and features deriving from already existing scoring functions. Finally, to reduce the number of MD runs needed to retrieve the best pose, a probabilistic ML model based on the Best Arm Identification (BAI) has been proposed.[94] This approach does not replace or enhance the MD analyses but selects the most promising poses to reduce the number of runs and therefore the calculation costs where these are limited. Figure 8 summarizes the options for integrating ML approaches into the drug-design workflow. The choice of the optimal route depends on the available data (number of compounds to be screened, known inhibitors, available features), characteristics of the target proteins (similarity to other known viral proteins, the distinctiveness of the active site and finally the computational resources).

Information captured for different natural product including spectra (eg, MS, NMR, etc.), binding affinities with protease targets, presence/absence component molecules and their respective anti-viral activity, are potentially relevant features for discovering natural products with a high probability of SARS-CoV-2 inhibition. Although such features when considered simultaneously may provide insight into candidates with the greatest likelihood of success for inhibition. Furthermore, identification of key features or combinations of features likely to have strong inhibitory effects can be used in screening newly discovered natural products.

A recent success was a study aiming to discover potential inhibitor(s) of Transmembrane protease, serine 2 (TMPRSS2) via virtual screening against a homology model of TMPRSS2 using the library of marine natural products (MNPs).[95] Molecular docking, binding affinity analysis using MM-GBSA and ADME evaluations were carried out to explore the inhibitory activity of MNPs against TMPRSS2 and determine their pharmacokinetic properties. Seven MNPs inhibited TMPRSS2 and one in particular, MNP-10 or Watasenia β-D- Preluciferyl glucopyrasoiuronic acid, was the optimal inhibitor of TMPRSS2 with acceptable pharmacokinetic properties. This MNP holds promise as a novel TMPRSS2 blocker to combat SARS-CoV-2.

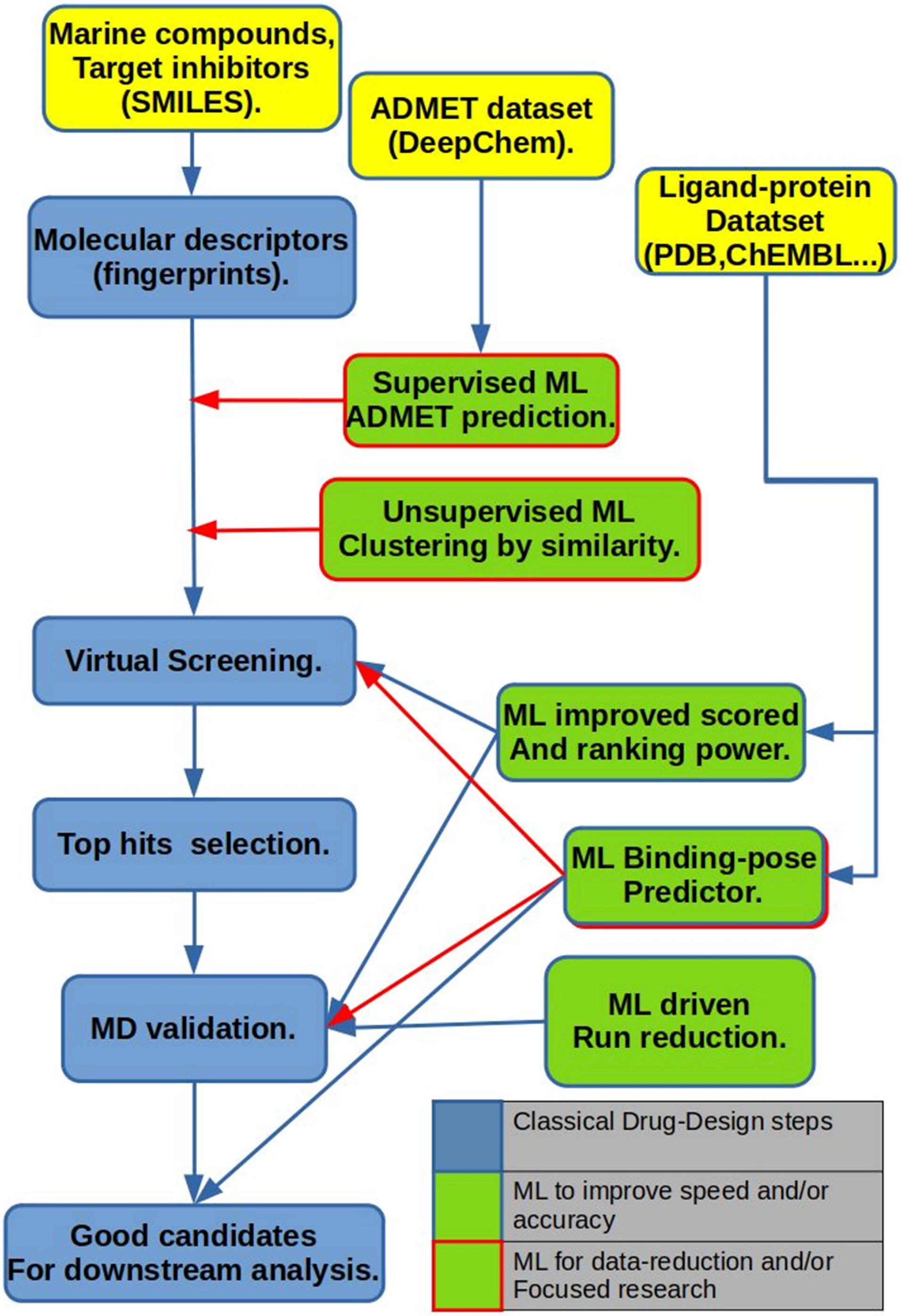
Marine compounds, Target inhibitors (SMILES).
ADMET dataset (DeepChem).
Ligand-protein Datatset (PDB,ChEMBL...)
Molecular descriptors (fingerprints).
Supervised ML ADMET prediction.
Unsupervised ML Clustering by similarity.
Virtual Screening.
ML improved scored And ranking power.
Top hits selection.
ML Binding-pose Predictor.
MD validation.
ML driven Run reduction.
Good candidates For downstream analysis.
Classical Drug-Design steps
ML to improve speed and/or accuracy
ML for data-reduction and/or Focused research

Figure 8 Typical in-silico drug-design workflow. In green are highlighted ML approaches that can be exploited to improve speed, accuracy, and performances of the analyses.

Conclusion

Viral structural proteins are most vulnerable to the host immune defense and are therefore subject to stronger selection pressure and mutate more frequently than intracellular viral proteins. From the perspective of therapeutic intervention the fact that intracellular viral proteins are better conserved makes them better targets for broad-spectrum therapeutic intervention. Molecular docking determines the interaction of a ligand with a drug target. The powerful computational ability afforded by molecular docking, ML and databases of marine natural products will help the identification of broad-spectrum inhibitors for SARS-CoV-2 and other betacoronaviruses. ML algorithms facilitate searches of large databases for potential therapeutic compounds of interest, the results subsequently inform further studies, and streamlining searches for potential drugs. The natural world is likely to be a source of useful small molecule antivirals against RNA viruses like coronaviruses that utilize papaine-like proteases and RdRp for replication. In-vitro-only high throughput screening of natural product extracts is an inefficient and expensive strategy. In-silico augmentation of drug discovery is at the cutting edge of SARS-CoV research and pharmaceutical research in general. ML algorithms have found applications alongside traditional approaches from the initial stages involving compounds library screening, to the later phases including clinical trials, enabling improvements in the accuracy and speed of the drug discovery process.

Disclosure

BJW, MH and GH report grants from the National Institutes of Health during the conduct of the study. G.H acknowledges support from National Institute of Health (NIH) U54MD010706, U01DA045300 and QUB start-up funds. G.H. is a founder of Altomics Datamation Ltd. and a member of its scientific advisory board. The authors report no other conflicts of interest in this work.

References

1. Cutler DM, Summers LH. The COVID-19 pandemic and the $16 trillion virus. *JAMA*. 2020;324(15):1495–1496. doi:10.1001/jama.2020.19759
2. Henderson LR, McShane H, Kiparoglou V. Rapid research response to the COVID-19 pandemic: perspectives from a national institute for health biomedical research centre. *Health Res Policy Syst*. 2022;20(1):1–4. doi:10.1186/s12961-022-00827-0
3. Gupta AM, Chakrabarti J, Mandal S. Non-synonymous mutations of SARS-CoV-2 leads epitope loss and segregates its variants. *Microbes Infect*. 2020;22(10):598–607. doi:10.1016/j.micinf.2020.10.004
4. Pardi N, Weissman D. Development of vaccines and antivirals for combating viral pandemics. *Nat Biomed Eng*. 2020;4(12):1128–1133. doi:10.1038/s41551-020-00658-w
5. Mohammad T, Shamsi A, Anwar S, et al. Identification of high-affinity inhibitors of SARS-CoV-2 main protease: towards the development of effective COVID-19 therapy. *Virus Res*. 2020;288:198102. doi:10.1016/j.virusres.2020.198102
6. Ghildiyal R, Prakash V, Chaudhary V, Gupta V, Gabrani R. Phytochemicals as antiviral agents: recent updates. In: *Plant-Derived Bioactives*. Springer; 2020:279–295.
7. Omrani M, Keshavarz M, Nejad Ebrahimi S, et al. Potential natural products against respiratory viruses: a perspective to develop anti-COVID-19 medicines. *Front Pharmacol*. 2021;2021:2115.
8. Singh TU, Parida S, Lingaraju MC, Kesavan M, Kumar D, Singh RK. Drug repurposing approach to fight COVID-19. *Pharmacol Rep*. 2020;72 (6):1479–1508. doi:10.1007/s43440-020-00155-6
9. Totura AL, Bavari S. Broad-spectrum coronavirus antiviral drug discovery. *Expert Opin Drug Discov*. 2019;14(4):397–412. doi:10.1080/17460441.2019.1581171
10. Sun C, Zhang J, Wei J, et al. Screening, simulation, and optimization design of small molecule inhibitors of the SARS-CoV-2 spike glycoprotein. *PLoS One*. 2021;16(1):e0245975. doi:10.1371/journal.pone.0245975
11. Zhou Y, Wang F, Tang J, Nussinov R, Cheng F. Artificial intelligence in COVID-19 drug repurposing. *Lancet Digit Health*. 2020;2(12):e667–e676. doi:10.1016/S2589-7500(20)30192-8
12. Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: collection of open natural products database. *J Cheminform*. 2021;13(1):1–13. doi:10.1186/s13321-020-00478-9
13. Gogineni V, Schinazi RF, Hamann MT. Role of marine natural products in the genesis of antiviral agents. *Chem Rev*. 2015;115(18):9655–9706. doi:10.1021/cr4006318
14. Al-Karmalawy AA, Farid MM, Mostafa A, et al. Naturally available flavonoid aglycones as potential antiviral drug candidates against SARS-CoV-2. *Molecules*. 2021;26(21):6559. doi:10.3390/molecules26216559
15. El-Demerdash A, Al-Karmalawy AA, Abdel-Aziz TM, Elhady SS, Darwish KM, Hassan AHE. Investigating the structure-activity relationship of marine natural polyketides as promising SARS-CoV-2 main protease inhibitors. *RSC Adv*. 2021;11(50):31339–31363. doi:10.1039/d1ra05817g
16. Pauli I, Macedo Timmers LF, Caceres RA, Pereira Soares MB, de Azevedo J, Filgueira W. In silico and in vitro: identifying new drugs. *Curr Drug Targets*. 2008;9(12):1054–1061. doi:10.2174/138945008786949397
17. Macarron R, Banks MN, Bojanic D, et al. Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov*. 2011;10(3):188–195. doi:10.1038/nrd3368
18. An G, Bartels J, Vodovotz Y. In silico augmentation of the drug development pipeline: examples from the study of acute inflammation. *Drug Dev Res*. 2011;72(2):187–200. doi:10.1002/ddr.20415
19. Yu R, Chen L, Lan R, Shen R, Li P. Computational screening of antagonists against the SARS-CoV-2 (COVID-19) coronavirus by molecular docking. *Int J Antimicrob Agents*. 2020;56(2):106012. doi:10.1016/j.ijantimicag.2020.106012
20. Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. The architecture of SARS-CoV-2 transcriptome. *Cell*. 2020;181(4):914–921. e10. doi:10.1016/j.cell.2020.04.011
21. Su H, Zhou F, Huang Z, et al. Molecular Insights into Small-Molecule Drug Discovery for SARS-CoV-2. *Adv Chem*. 2021;133(18):9873–9886. doi:10.1002/ange.202008835
22. Yadav R, Chaudhary JK, Jain N, et al. Role of structural and non-structural proteins and therapeutic targets of SARS-CoV-2 for COVID-19. *Cells*. 2021;10(4):821. doi:10.3390/cells10040821
23. Lan J, Ge J, Yu J, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*. 2020;581 (7807):215–220. doi:10.1038/s41586-020-2180-5

24. Gao Y, Yan L, Huang Y, et al. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science*. 2020;368(6492):779–782. doi:10.1126/science.abb7498
25. Picarazzi FA, Vicenti IA, Saladini F, Zazzi MA, Mori MA. Targeting the RdRp of emerging RNA viruses: the structure-based drug design challenge. *Molecules*. 2020;25(23):5695. doi:10.3390/molecules25235695
26. Altabella T, Ramirez-Estrada K, Ferrer A. Phytosterol metabolism in plant positive-strand RNA virus replication. *Plant Cell Rep*. 2021;2021:1–11.
27. Lu S, Wang J, Chitsaz F, et al. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res*. 2020;48(D1):D265–D268. doi:10.1093/nar/gkz991
28. Mann KS, Sanfaçon H. Expanding repertoire of plant positive-strand RNA virus proteases. *Viruses*. 2019;11(1):66. doi:10.3390/v11010066
29. Ul Qamar MT, Alqahtani SM, Alamri MA, Chen -L-L. Structural basis of SARS-CoV-2 3CLpro and anti-COVID-19 drug discovery from medicinal plants. *J Pharm Anal*. 2020;10(4):313–319. doi:10.1016/j.jpha.2020.03.009
30. Gorkhali R, Koirala P, Rijal S, Mainali A, Baral A, Bhattarai HK. Structure and function of major SARS-CoV-2 and SARS-CoV Proteins. *Bioinform Biol Insights*. 2021;15:11779322211025876. doi:10.1177/11779322211025876
31. Chang LJ, Chen TH. NSP16 2'-O-MTase in Coronavirus Pathogenesis: possible Prevention and Treatments Strategies. *Viruses*. 2021;13(4):538. doi:10.3390/v13040538
32. Wu C, Liu Y, Yang Y, et al. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm Sin B*. 2020;10(5):766–788. doi:10.1016/j.apsb.2020.02.008
33. Torres PH, Sodero AC, Jofily P, Silva-Jr FP. Key topics in molecular docking for drug design. *Int J Mol Sci*. 2019;20(18):4574. doi:10.3390/ijms20184574
34. Thomsen R, Christensen MH. MolDock: a new technique for high-accuracy molecular docking. *J Med Chem*. 2006;49(11):3315–3321. doi:10.1021/jm051197e
35. Radoux CJ, Olsson TS, Pitt WR, Groom CR, Blundell TL. Identifying interactions that determine fragment binding at protein hotspots. *J Med Chem*. 2016;59(9):4314–4325. doi:10.1021/acs.jmedchem.5b01980
36. Brooijmans N, Kuntz ID. Molecular recognition and docking algorithms. *Annu Rev Biophys Biomol Struct*. 2003;32(1):335–373. doi:10.1146/annurev.biophys.32.110601.142532
37. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*. 1997;267(3):727–748. doi:10.1006/jmbi.1996.0897
38. Lyu C, Chen T, Qiang B, et al. CMNPD: a comprehensive marine natural products database towards facilitating drug discovery from the ocean. journal article. *Nucleic Acids Res*. 2020. doi:10.1093/nar/gkaa763
39. Lyu J, Wang S, Balius TE, et al. Ultra-large library docking for discovering new chemotypes. *Nature*. 2019;566(7743):224–229. doi:10.1038/s41586-019-0917-9
40. Probst D, Reymond J-L. A probabilistic molecular fingerprint for big data settings. *J Cheminform*. 2018;10(1):1–12. doi:10.1186/s13321-018-0321-8
41. Zheng X, Zhou D, Li N, Wu T, Lei Y, Shi J. Self-regulated particle swarm multi-task optimization. *Sensors*. 2021;21(22):7499. doi:10.3390/s21227499
42. Abadi M, Barham P, Chen J, et al. TensorFlow: a system for large-scale machine learning. Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation; 2016; GA, USA: Savannah.
43. Singh S, Sk MF, Sonawane A, Kar P, Sadhukhan S. Plant-derived natural polyphenols as potential antiviral drugs against SARS-CoV-2 via RNA-dependent RNA polymerase (RdRp) inhibition: an in-silico analysis. *J Biomol Struct Dyn*. 2021;39(16):6249–6264. doi:10.1080/07391102.2020.1796810
44. Zhao Y, Du X, Duan Y, et al. High-throughput screening identifies established drugs as SARS-CoV-2 PLpro inhibitors. *Protein Cell*. 2021;12(11):877–888. doi:10.1007/s13238-021-00836-9
45. Leissring MA, Lu A, Condron MM, et al. Kinetics of amyloid β-protein degradation determined by novel fluorescence-and fluorescence polarization-based assays. *J Biol Chem*. 2003;278(39):37314–37320. doi:10.1074/jbc.M305627200
46. Hanna GS, Choo Y-M, Harbit R, et al. Contemporary approaches to the discovery and development of broad-spectrum natural product prototypes for the control of coronaviruses. *J Nat Prod*. 2021;84(11):3001–3007. doi:10.1021/acs.jnatprod.1c00625
47. Gutierrez DD. *Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R*. Technics Publications; 2015.
48. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444. doi:10.1038/nature14539
49. Koza JR, Bennett FH, Andre D, Keane MA. Automated design of both the topology and sizing of analog electrical circuits using genetic programming. In: *Artificial Intelligence in Design'96*. Springer; 1996:151–170.
50. Bishop CM, Nasrabadi NM. *Pattern Recognition and Machine Learning*. Vol. 4. Springer; 2006.
51. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825–2830.
52. Chollet F. Keras. Available from: https://keras.io. Accessed March 16, 2023.
53. Hervé Jegou MD, Johnson J. Faiss: a library for efficient similarity search. Facebook; 2020. Available from: https://engineering.fb.com/2017/03/29/data-infrastructure/faiss-a-library-for-efficient-similarity-search/. Accessed March 16, 2023.
54. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst*. 2019;32:8024–8035.
55. Winkler JK, Sies K, Fink C, et al. Melanoma recognition by a deep learning convolutional neural network-Performance in different melanoma subtypes and localisations. *Eur J Cancer*. 2020;127:21–29. doi:10.1016/j.ejca.2019.11.020
56. Abraham A, Pedregosa F, Eickenberg M, et al. Machine learning for neuroimaging with scikit-learn. Methods. *Front Neuroinform*. 2014;8(14). doi:10.3389/fninf.2014.00014
57. Noé F, De Fabritiis G, Clementi C. Machine learning for protein folding and dynamics. *Curr Opin Struct Biol*. 2020;60:77–84. doi:10.1016/j.sbi.2019.12.005
58. Zhang Y. *New Advances in Machine Learning*. IntechOpen; 2010.
59. Serra A, Galdi P, Tagliaferri R. Machine learning for bioinformatics and neuroimaging. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2018;8(5):e1248. doi:10.1002/widm.1248
60. Lantz B. *Machine Learning with R: Expert Techniques for Predictive Modeling*. Packt publishing ltd; 2019.
61. Feurer M, Klein A, Eggensperger K, Springenberg JT, Blum M, Hutter F. Auto-sklearn: efficient and Robust Automated Machine Learning. In: Hutter F, Kotthoff L, Vanschoren J, editors. *Automated Machine Learning: Methods, Systems, Challenges*. Springer International Publishing; 2019:113–134.
62. Roelofs R, Shankar V, Recht B, et al. A meta-analysis of overfitting in machine learning. *Adv Neural Inf Process Syst*. 2019;32:9179–9189.
63. Ng AY. Preventing "overfitting" of cross-validation data. Proceedings of the Fourteenth International Conference on Machine Learning; 1997.

64. Ng VW, Breiman L. *Bivariate Variable Selection for Classification Problem*. Technical report, Department of Statistics, University of California-Berkeley; 2005. Citeseer.
65. Nanga S, Bawah AT, Acquaye BA, et al. Review of dimension reduction methods. *J Data Analysis Inf Process*. 2021;9(3):189–231. doi:10.4236/jdaip.2021.93013
66. Todeschini R, Ballabio D, Consonni V, Mauri A, Pavan M. CAIMAN (Classification And Influence Matrix Analysis): a new approach to the classification based on leverage-scaled functions. *Chemometr Intell Lab Syst*. 2007;87(1):3–17. doi:10.1016/j.chemolab.2005.11.001
67. Lo Y-C, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. *Drug Discov Today*. 2018;23(8):1538–1546. doi:10.1016/j.drudis.2018.05.010
68. Mapari S, Camarda KV. Use of three-dimensional descriptors in molecular design for biologically active compounds. *Curr Opin Chem Eng*. 2020;27:60–64. doi:10.1016/j.coche.2019.11.011
69. Rifaioglu AS, Atas H, Martin MJ, Cetin-Atalay R, Atalay V, Doğan T. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Brief Bioinform*. 2019;20(5):1878–1912. doi:10.1093/bib/bby061
70. Maltarollo VG, Gertrudes JC, Oliveira PR, Honorio KM. Applying machine learning techniques for ADME-Tox prediction: a review. journal article. *Expert Opin Drug Metab Toxicol*. 2014;11(2):259–271. doi:10.1517/17425255.2015.980814
71. Khamis MA, Gomaa W, Ahmed WF. Machine learning in computational docking. *Artif Intell Med*. 2015;63(3):135–152. doi:10.1016/j.artmed.2015.02.002
72. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. journal article. *Nat Rev Drug Discov*. 2019;18(6):463–477. doi:10.1038/s41573-019-0024-5
73. Cheng F, Desai RJ, Handy DE, et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat Commun*. 2018;9(1):1–12. doi:10.1038/s41467-018-05116-5
74. Xie -Q-Q, Zhong L, Pan Y-L, et al. Combined SVM-based and docking-based virtual screening for retrieving novel inhibitors of c-Met. *Eur J Med Chem*. 2011;46(9):3675–3680. doi:10.1016/j.ejmech.2011.05.031
75. Gao H, Struble TJ, Coley CW, Wang Y, Green WH, Jensen KF. Using machine learning to predict suitable conditions for organic reactions. *ACS Cent Sci*. 2018;4(11):1465–1476. doi:10.1021/acscentsci.8b00357
76. Cova TF, Pais AA. Deep learning for deep chemistry: optimizing the prediction of chemical patterns. *Front Chem*. 2019;7:809. doi:10.3389/fchem.2019.00809
77. Prihoda D, Maritz JM, Klempir O, et al. The application potential of machine learning and genomics for understanding natural product diversity, chemistry, and therapeutic translatability. *Nat Prod Rep*. 2021;38(6):1100–1108. doi:10.1039/D0NP00055H
78. Reschke M, Mihic-Probst D, van der Horst EH, et al. HER3 is a determinant for poor prognosis in melanoma. *Clin Cancer Res*. 2008;14(16):5188–5197. doi:10.1158/1078-0432.CCR-08-0186
79. Gallardo A, Lerma E, Escuin D, et al. Increased signalling of EGFR and IGF1R, and deregulation of PTEN/PI3K/Akt pathway are related with trastuzumab resistance in HER2 breast carcinomas. *Br J Cancer*. 2012;106(8):1367–1373. doi:10.1038/bjc.2012.85
80. Barbosa AJM, Roque ACA. Free marine natural products databases for biotechnology and bioengineering. journal article. *Biotechnol J*. 2019;14(11):1800607. doi:10.1002/biot.201800607
81. Davis GDJ, Vasanthi AHR. Seaweed metabolite database (SWMD): a database of natural compounds from marine algae. *Bioinformation*. 2011;5(8):361–364. doi:10.6026/97320630005361
82. Fagan DH, Yee D. Crosstalk between IGF1R and estrogen receptor signaling in breast cancer. *J Mammary Gland Biol Neoplasia*. 2008;13(4):423–429. doi:10.1007/s10911-008-9098-0
83. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. *Nucleic Acids Res*. 2016;45(D1):D945–D954. doi:10.1093/nar/gkw1074
84. Irwin JJ, Shoichet BK. ZINC − A free database of commercially available compounds for virtual screening. *J Chem Inf Model*. 2005;45(1):177–182. doi:10.1021/ci049714
85. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*. 2017;46(D1):D1074–D1082. doi:10.1093/nar/gkx1037
86. Pence HE, Williams A. ChemSpider: an online chemical information resource. *J Chem Educ*. 2010;87(11):1123–1124. doi:10.1021/ed100697w
87. Wu Z, Ramsundar B, Feinberg Evan N, et al. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci*. 2018;9(2):513–530. doi:10.1039/C7SC02664A
88. Parks C, Gaieb Z, Amaro RE. An analysis of proteochemometric and conformal prediction machine learning protein-ligand binding affinity models. *Front Mol Biosci*. 2020;7. doi:10.3389/fmolb.2020.00093
89. Kundu I, Paul G, Banerjee R. A machine learning approach towards the prediction of protein–ligand binding affinity based on fundamental molecular properties. journal article. *RSC Adv*. 2018;8(22):12127–12137. doi:10.1039/c8ra00003d
90. Lahey S-LJ, Rowley CN. Simulating protein–ligand binding with neural network potentials. journal article. *Chem Sci*. 2020;11(9):2362–2368. doi:10.1039/c9sc06017k
91. Li H, Sze K-H, Lu G, Ballester PJ. Machine-learning scoring functions for structure-based drug lead optimization. *WIREs Comput Mol Sci*. 2020;10(5). doi:10.1002/wcms.1465
92. Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*. 2018;34(21):3666–3674. doi:10.1093/bioinformatics/bty374
93. Chmiela S, Sauceda HE, Müller KR, Tkatchenko A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat Commun*. 2018;9(1). doi:10.1038/s41467-018-06169-2
94. Terayama K, Iwata H, Araki M, Okuno Y, Tsuda K. Machine learning accelerates MD-based binding pose prediction between ligands and proteins. *Bioinformatics*. 2017;34(5):770–778. doi:10.1093/bioinformatics/btx638
95. Mahmudpour M, Nabipour I, Keshavarz M, Farrokhnia M. Virtual screening on marine natural products for discovering TMPRSS2 inhibitors. *Front Chem*. 2021;9:722633. doi:10.3389/fchem.2021.722633
96. Kraker JJ, Hawkins DM, Basak SC, Natarajan R, Mills D. Quantitative Structure–Activity Relationship (QSAR) modeling of juvenile hormone activity: comparison of validation procedures. *Chemometr Intell Lab Syst*. 2007;87(1):33–42. doi:10.1016/j.chemolab.2006.03.001
97. Dearden JC. Partitioning and lipophilicity in quantitative structure-activity relationships. *Environ Health Perspect*. 1985;61:203–228. doi:10.1289/ehp.8561203