ISSN 0974 - 9373

Vol. 26 No. 3 (2022) **J. Int. Acad. Phys. Sci.** pp. 339-362

An Integrated Computational Approach for Drug Discovery Against SARS-CoV-2

Sibi Raj and Dhruv Kumar*
School of Health Sciences and Technology, UPES University
Dehradun, Uttarakhand, India
*Corresponding author- dhruvbhu@gmail.com

Ramesh Choudhari
Center of Emphasis in Cancer, Paul L. Foster School of Medicine
Department of Molecular and Translation Medicine,
Texas Tech University Health Sciences Center, El Paso, TX, USA

Brijesh Rathi
Laboratory for Translational Chemistry and Drug Discovery,
Department of Chemistry, Hansraj College, University of Delhi, Delhi, India

Jayaramulu Kolleboyina
Department of Chemistry, Indian Institute of Technology Jammu, Room No. 642
Jagti, NH-44, Nagrota, Jammu - 181221, J&K, India

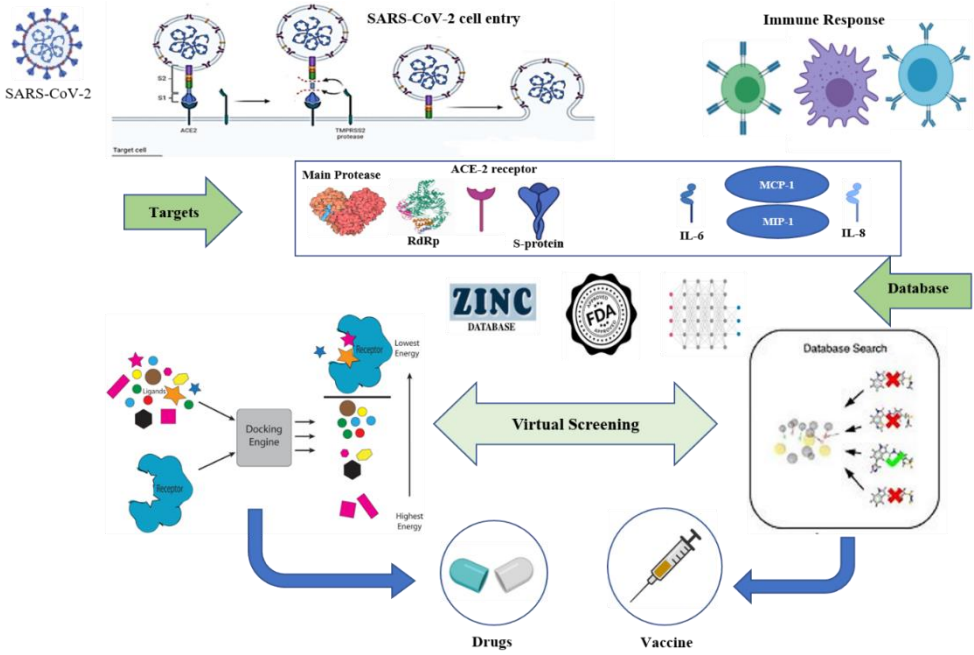
(Received August 5, 2022, Accepted September 2, 2022)

Abstract: The novel coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has killed thousands and infected millions of people worldwide. Urgent development of potential drugs against SARS-CoV-2 could be helpful to save millions lives around the world. In the last few years, significant development has been made in virtual screening (VS) and drug development. Drug screening through virtual method has evolved from traditional similarity searching, which utilizes advanced application domain like similarity search, data mining, and machine- learning approaches, which require large and representative training-set compounds to learn robust decision rules. Tremendous growth of public domain-available chemical databases including FDA -approved drugs and structural database of druggable targets of SARS-CoV-2, has generated huge effort to design, analyze, and apply novel learning methodologies to develop novel drug molecules against SARS-CoV-2. In this review, we focus on machine-learning techniques within the

context of ligand-based virtual screening to develop potential drugs against SARS-CoV-2.

Keywords: SARS-CoV-2, virtual screening, drug discovery, machine-learning, artificial intelligence.

Graphical Abstract



1. Introduction

The outbreak of SARS-CoV-2 disease in 2019 December has caused a pandemic situation worldwide[1]. At the beginning of the major outbreak of coronavirus infection, palm cats had been reported to be the major source for SARS-CoV-2 and camels for the MERS CoV. Later more advanced solutions, reported bats to be the host for SARS-CoV-2, spreading to other responsible intermediate hosts before infecting humans. It has been reported that most of the bat CoVs are the gene source for alpha-CoV and beta CoV. While the gamma and delta CoVs are reported to be originated from birds. The transmission of this novel coronavirus has been reported via the close human to human contact[2]. Scientists across the globe are searching for new technologies to screen the patients and find effective medicines at a faster rate to control viral spread. SARS-CoV-2 is a type of coronaviruses that has a crown shape structure due to the spike protein covering its outer surface[3].

These are single-stranded positive sense RNA viruses having 29,903 nucleotides and two untranslated sequences of 254 and 229 nucleotides at 5' and 3' ends. Capsid is made up of 4 structural proteins named spike protein, envelope protein, membrane protein and the nucleocapsid[4]. The spike protein is majorly responsible for the attachment to human cells and its entry for viral infection. This process majorly takes place through the angiotensin converting enzyme-2 receptor, which is present on the surface of the host epithelial cells. The SARS-CoV-2 initially infects the lower airways and binds to ACE-2 on alveolar epithelial cells[5]. As, the virus is a potent inducer of cytokines, the cytokine storm or cytokine cascade is the major mechanism suggested for organ damage by the viral infection.

Furthermore, the virus activates the immune cells triggering the secretion of inflammatory cytokines and chemokines into pulmonary vascular endothelial cells. The transmission primarily occurs through the respiratory droplets produced when an infected person sneeze. When inhaled, these droplets can settle in the lungs, nasal mucosa or in the mouth of the people. Like most of other respiratory viruses, SARS-CoV-2 is most contagious when people are most symptomatic. The basic reproduction number (R0) of the novel coronal virus indicated the transmissibility of the virus to be 4.71, but now the viral reproducibility has been reported to be declined to 2.08[6]. This trend predicts that over time there should be a gradual decline in the spread of the disease. The current global aim is to prevent the pandemic spread and minimize the transmission wherever possible.

Several vaccine candidates for coronavirus, Oxford Institute's Covishield, which is being developed by the Pune-based Serum Institute, and Bharat Biotech's Covaxin, received emergency approval from drug regulators. In addition, several other vaccines are under process to get approval globally. However, not many vaccines or treatments available against coronavirus disease. At present, people have been advised to follow rules such as social distancing and hygiene measures to avoid the infection. There is a collective world-wide effort in developing effective vaccines against COVID-19. Currently, 200 vaccines are in various stages of development with over 30 vaccines in clinical trials. Many different angles have been explored to develop COVID-19 vaccines including RNA, DNA, non-replicating viral vectors and inactivated vaccines.

Machine learning approach has emerged as most recent trend in the development of drugs against SARS-CoV-2, because it promises better scale-up, speed-up processing, which can also, at times, outperform humans in specific healthcare tasks[7]. Various studies have provided enough proof showing machine learning impact on assisting different clinicians in tackling

various communicable and non-communicable diseases[9]. Machine learning approaches improves the screening of diseases with much higher accuracy MYCIN was the first expert system developed in 1976 to create a relationship with the healthcare area[8]. This system suggested antibiotics to patients suffering from bacterial infection by following 450 rules collected from a medical expert. Early detection marks the importance of early treatment and to saves many lives. Fast diagnosis is critically essential to overcome pandemic diseases such as SARS-CoV-2 infection to prevent the spread, making the process cost-effective and speed up the diagnosis. Machine learning approaches can also augment the screening process of the suspected patient along with traditional diagnosis methodologies such as with radio imaging technology and computed tomography. A recent study by Ardakani and team reported that employing machine learning approaches on 1020 CT images of 108 SARS-CoV-2 infected patients and viral pneumonia of 86 patients gave an accuracy of 86.27% and specificity of 83.33% [10].

In the race for SARS-CoV-2 vaccine development and accurate diagnosis, a deep learning algorithm has been used for automatic SARS-CoV-2 detection. Designing models with a combination of clinical features provides an expert system to enhance the accuracy of disease prediction, which reserve more time for doctors to save more lives and stop the spread. WHO has reported that SARS-CoV-2 infection spreads from individual to individual majorly through saliva, droplets, or discharge from the nose through contact transmission[11]. One of the major steps to prevent viral spread is to break the chain of virus transmission. Machine learning has equally contributed to contact tracing of the viral spread and helped in managing to identify people susceptible to viral infection[12]. Therefore, the to further undergo self-isolation for 14 days since the exposure to reduces the magnitude of the recent pandemic[13]. With this approach, many countries struggling with the pandemic have come up with several digital applications using mobile and technologies such as Bluetooth, a global positioning system (GPS), social graph, contact details, network-based API, mobile tracking data, card transaction data and system physical address[14]. Digital based application performs faster than manual processes. These technologies collect suspected person's personal data is analyzed by machine learning tools to trace a Covid-19 positive person and their recent contact. Many countries such as India, China, Australia, Italy, and Malaysia have been effectively using these machine learning-based applications to trace the affected persons to minimize the viral spread.

Developing a drug or vaccine for the SARS-CoV-2 infection remains the most critically required to tackle pandemic. Machine learning approaches have provided enough opportunities to fasten up the race for vaccine or drug

development. Repurposing of existing drugs to effectively fight the viral infection has been in trend to combat the pandemic[15]. Machine learning approaches in this field have enabled virtual screening of drugs against the main structural proteins of the virus. Auto dock like platforms has enabled to screen effective drugs against SARS-CoV-2, especially targeting their structural proteins such as Mpro, Spike protein, and nsp-9 (Fig. 1)[16]. Many of the drugs such as remdisivir, ritonavir, darunavir were outlined to combat the viral proteinases. Also, drug toxicity prediction and further analysis can be made using machine-based approaches. These virtual platforms provide a faster time for drug development against the virus. Usual drug discovery needs a huge amount of time to develop and come into the market. These machine learning-based technologies avoid the time loss at the pre-clinical study stage, which effectively repurpose drugs for SARS-CoV-2 infection, that are normally used to treat other diseases.

2. Homology modeling of target proteins

The genome sequence of newly discovered SARS-CoV-2 was initially released by several countries, denoting its similarity to that of other coronaviruses. The novel coronavirus genomes consist of an open reading frame containing ORF1 ab, which is common in all beta coronaviruses; this ORF codes for different structural proteins of the SARS-CoV-2[17]. The primary sequence of ten important proteins of SARS-CoV-2, which are ORF1 ab polyprotein, ORF6, ORF7a, ORF8, nucleocapsid phosphoprotein, and ORF10 were resolved and published in the National center for biotechnology information (NCBI)[18]. With the increase of viral spread, the viral pneumonia was also in consideration to urgently tackle the disease. A study by Tao et al., reported that SARS-CoV-2 resembled the bat coronavirus HKU9-1[19]. Also, reported the significance of S protein in SARS-CoV-2 viral infection. In the same study they pointed out the important discovery of S protein interaction with the ACE-2 receptor, which possessed significant health risk via human transmission[20]. With the knowledge on the structure of 3CL hydrolase of SARS-CoV-2, combined the machine learning-based approaches of virtual screening and enzymology test to screen several FDA approved drugs, and built a database of high-yield compounds from medicinal plants. The study is successful in establishing 30 effective drugs for the SARS-CoV-2 treatment. Similarly, team of scientists pointed out papain-like protease (PLP) that is encoded by SARS-CoV-2 non-structural protein-3 played an important role in viral replication and also helped in escaping the natural immunity of the host[21]. PLP also has deubiquitinase activity, which in combination evades the

host immune response and inhibits the expression of interferons through a series of pathways. To find effective targets against SARS-CoV-2, scientists Irwin and his team built the PLP protein structure of SARS-CoV-2 using homology modeling method, and defined its drug-binding pocket. It enabled the screening of active molecules from the existing drug library of ZINC, a database from Chinese medicine and natural products using computer-based structure-based virtual screening method[22] done with computational biology and bioinformatics, promises to screen effective drugs against SARS-CoV-2 infection.

Computational tools such as the Basic Local Alignment Search Tool (BLAST) help find similarity regions between the sequences. This program compares nucleotide or protein sequences with database sequences and provides the statistical significance matches. Similarly, SWISS-MODEL is a completely automated protein structure homology-modeling server that is easily accessible. To build a homology modeling there are four major steps that involve identification of structural templates, alignment of a target sequence and template structure, model building and evaluation of model quality[23]. Another tool is clustal omega, a multiple sequence alignment program that utilizes the seeded guide trees and HMM profile-profile methods to create alignments between multiple sequences[24]. The protein models created through such computational tools are very useful for molecular docking-based studies and molecular dynamics-based simulations, which would provide more meaningful structure-based virtual screening and related computer-aided drug design (Figure-2).

3. Virtual screening for SARS-COV-2 drug discovery

Virtual screening (VS) is a computational based technique which is majorly used in the process of drug discovery. VS majorly helps in searching small molecules from huge libraries to identify potential structures that can most likely bind to the target drug, which can mostly be a protein receptor or enzyme[25]. This computational approach screens many libraries using computer-based programs and helps choose effective compounds that can be purchased to choose against a particular disease. VS is broadly categorized into two main screening techniques: ligand-based and structure-based VS[26]. In ligand-based VS based upon the collective information contained in the sets of structurally different ligands binding to a receptor, the model of the receptor is built. These models are referred to as pharmacophore models, which can be compared with the candidate ligand to determine the compatibility of each other when they bind together. 2-D chemical similarity

analysis methods are used as another approach to ligand-based VS. Structure-based VS revolves around the docking of candidate ligands into a target protein, which upon binding, provides a scoring function that estimates the affinity rate of ligand binding to the protein[27].

Virtual screening protocols are mostly used for the early steps in the whole drug discovery process to outline the initial library with active compounds. Before processing with virtual screening, the data given should be completely analyzed to know the target and screen out the possible methods incorporated in the virtual screening workflow. The first most crucial step is the bibliographic research, which focuses on receptor function, ligand characteristics, and catalytic mechanism. Databases such as UniProt or Brenda can easily provide such information[28]. This also helps acquire the previous knowledge of drug development against the target compound, which can provide an insight into current challenges and limitations regarding their mechanism of actions. Studies based on structure-activity relationship also help provide useful information to design inhibitors against the target of interest, providing a better understanding on several parts of the compound where the compound can be modified based on better binding and also increased activity to the target. Molecular visualization studies using Flare could provide an advantage to observe SAR studies to understand the nature of the interactions between the ligand and the protein[29]. The retrieving of previously reported or already established data for inhibitor structures is very important for virtual screening studies. Databases such as ChEMBL, Reaxys, Binding DB, or PubChem provide easy access to retrieve such information[30]. An equally important aspect is the availability of the three-dimensional structure of the receptor quantity and quality of their crystallographic structures. Databases such as Protein data bank provide crystallographic models of such proteins and also the protein-ligand complexes. The final collection of compounds used for the virtual screening process is referred to as a virtual screening library. The generation of a virtual screening library can be done using various databases such as ZINC, Reaxys, or from any other source of compound supplier. Most of the structures could be available in a two-dimensional format. Still, for many virtual screening software, three-dimensional conformations are essentially required based on atoms arranged in space. In a three-dimensional structure, molecules are analyzed based on bond lengths, bond angles, and torsion angles, which are then arranged based on low-energy conformations. There are various software present that provides the three-dimensional conformation of molecules such as LigPrep, DecoyFinder, OMEGA (Table-1)[31].

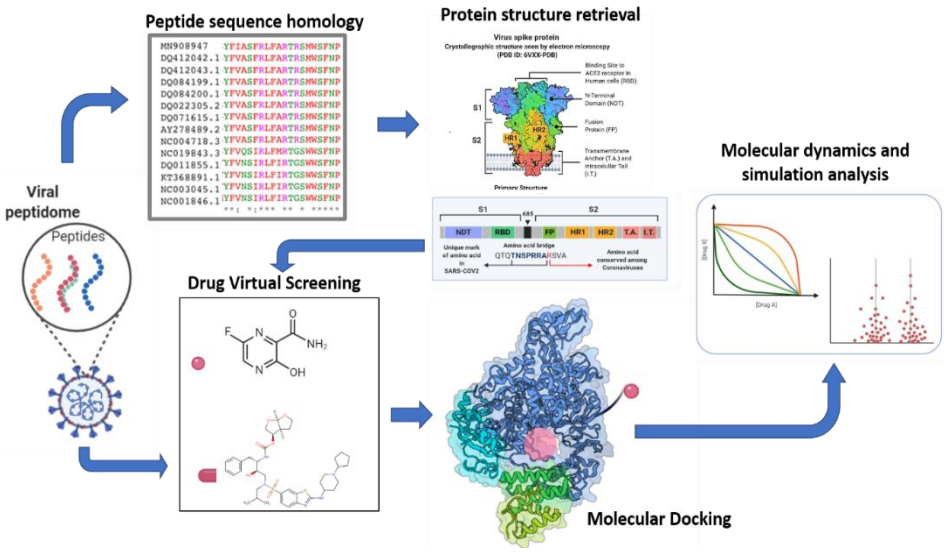


Figure 1. Drug discovery for SARS-CoV-2 through computational

Table 1. List of various software utilized for virtual screening for drug development

Software	Function	Developer
Flare Maestro VIDA	Graphical user interface	Cresset Schrodinger, LLC OpenEye Scientific Software Inc.
DecoyFinder	Decoy set preparation	Cheminformatics and Nutrition Research Group (Universitat Rovira I Virgili)
VHELIBS	Crystal structure validation	Cheminformatics and Nutrition Research Group (Universitat Rovira I Virgili)
Standardizer LigPrep MolVS	Molecular standardization	ChemAxon Schrödinger, LLC RDKit
Conformer generation	OMEGA ConfGen Distance Geometry ETKDG	OpenEye Scientific Software Inc. Schrödinger, LLC Distance Geometry RDKit
ADME	QikProp	Schrödinger, LLC

property prediction	SwissADME FAFDrugs4	Swiss Institute of Bioinformatics UMRS Paris Diderot-Inserm 973
Shape screening ROCS	Shape similarity	OpenEye Scientific Software Inc. Schrödinger, LLC
Phase Ligandscout	Pharmacophore	Schrödinger, LLC Inte:Ligand GmbH
Glide GOLD DOCK Autodock	Docking	Schrödinger, LLC The Cambridge Crystallographic Data Centre University of California San Francisco The Scripps Research Institute

Computational based methodologies after a series of virtual screening require validation of the data obtained both Insilco and invitro to prove their correctness. This Insilco based validation is done through virtual screened active compounds in parallel to inactive or decoy compounds. Active compounds re molecules having increased activity towards the protein of interest. The threshold above which compounds are referred to active if their IC50, Ki, or EC50 values range in micromoles or nanomoles[32]. Although compounds with higher threshold have increased activity towards the target, but their mechanism of action might differ as compared to that of in virtual screening. Inactive compounds are molecules with lesser threshold values and have low activity towards the target of interest. This threshold differentiation sets the bar high to demand better hits through virtual screening. Databases such as PubChem and ChEMBL include chemical compounds that are inactive[33]. Decoy compounds act similar to active compounds but their activity towards the target of interest has not been reported and is presumed to be inactive. Decoys can be usually found from databases such as DUD-E or using tools such as DecoyFinder[34].

As in the case of SARS-COV-2 drug discovery, several anti-fungal, anti-malarial, phytochemicals, antibacterial drugs have been screened overtime against different structural proteins to discover a potential drug. Papain like protease is the largest replicase unit and a multi-domain protein. It is mainly responsible for the cleavage of replicase polyprotein. 3CLpro is another important structural protein that has been widely used for docking studies[35]. It cleaves the polyprotein at 11 distinct sites to produce other essential non-structural proteins required for viral replication. Thus, there are studies focused on to target inhibiting the virus. RNA-directed RNA polymerase (RdRp) has an active site that is highly conserved and contains aspartate

residues that play major roles in viral genome replication. RdRp is considered as the molecular target for the development of new drugs against SARS-CoV-2 (Fig. 2)[36]. Non-structural protein 15 is an endoribonuclease that cleaves at the 3' of uridylates through a ribonuclease A and majorly has a role in the inhibition of the host immune system. The S protein is another major viral target that is considered for drug discovery as it has a major role in host interaction and entry through the ACE-2 receptor.

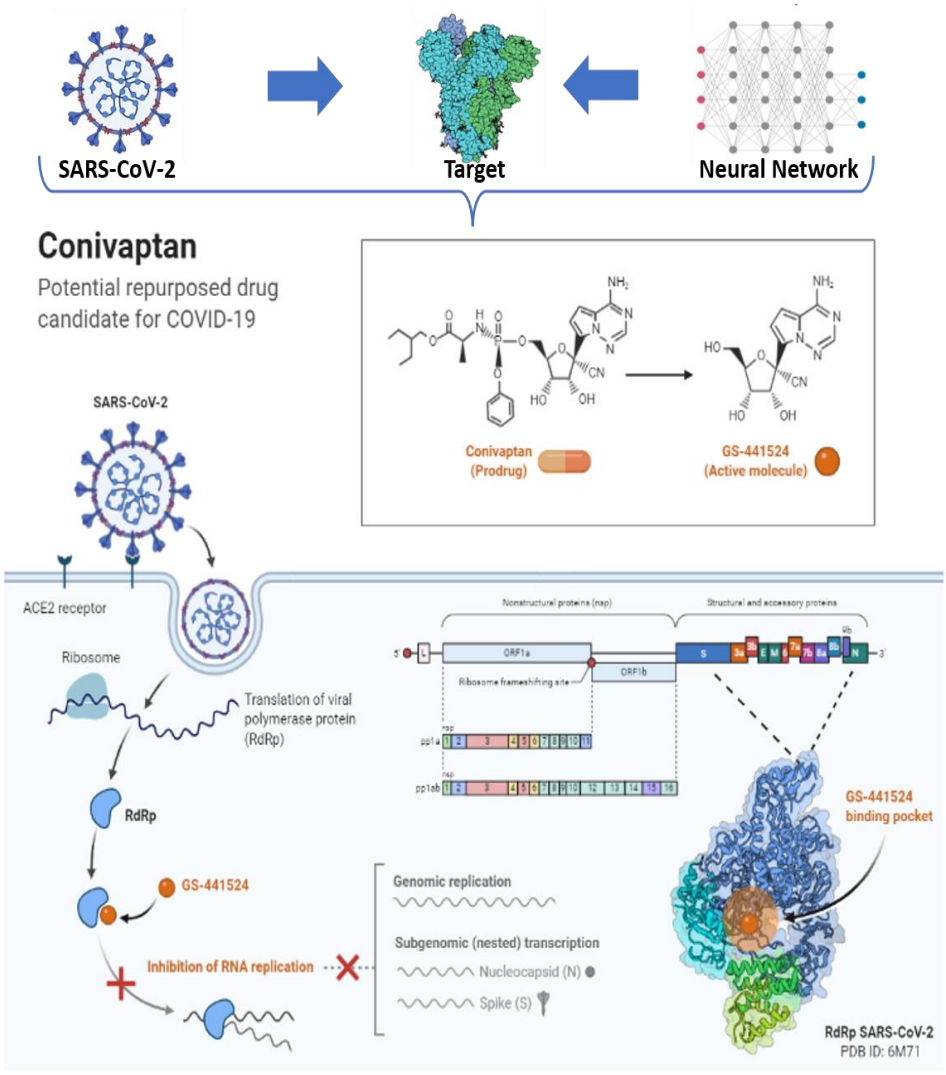


Figure 2. Drug repurposing approach against SARS-CoV-2. Repurposing of Conivaptan against SARS-CoV-2 through targeting RdRp using computational approach.

4. Molecular docking

Molecular docking is bioinformatics-based modeling that involves the interaction of two or more molecules to form a stable complex when bound to each other. The interaction of both ligand and target determines the three-dimensional structure of the complex, which is also quite stable. Molecular docking of ligand to a target and ligand creates all different possible complexes ranked and grouped using a scoring function in the software[37]. Docking simulations predict the most effective docking conformer based upon the total energy of the system. This system enables us to understand the mechanism of interaction between the drug and the protein, thus leading to rational drug design and discovery. With the information obtained from molecular docking, the binding energy, free energy, and the stability of the ligand-protein binding can be obtained. The main purpose of molecular docking is to attain a strong drug-protein binding complex with lower binding free energy. The total free binding energy, which is denoted as ΔG_{bind} is estimated based on several parameters such as the hydrogen bond, electrostatic energy, torsion-free energy, dispersion and repulsion, desolvation, total internal energy, and unbound system's energy. Thus, a proper understanding of basic principles affecting binding energy provides information about the different aspects of interactions, which leads to molecular docking. To begin with, 3D structure of target is a must for molecular docking, which is usually stored and handled in PDB file format. It can be downloaded for known-targets from PDB database or can be generated using Swiss-model. This can be done using various software such as Discovery studio, which makes the ligand available in PDB format, and this particular tool help to organize the various ligands based on their ability to interact with the target proteins. Docking with small molecules for a definite target requires few pre-defined samplings to obtain an optimized conformation of the complex as ligand interacts with the particular groove of the target. This kind of optimization is possibly done using the scoring function of the software. The used methodologies to investigate and establish any three-dimensional structure of a biomolecule are nuclear magnetic resonance spectroscopy (NMR), X-ray crystallography, and Cryo Electron Microscopy[38]. But, applications like homology modeling techniques make it easy to understand the possible structure of proteins with unknown structure with high sequence homology to known structure. This approach, in general, also provides a substitute target structure and initiates the beginning of *insilico* drug discovery approaches. The different databases that provide possible information on small ligand molecules are Cambridge Structural Database (CSD), Available Chemical Directory (ACD), MDL Drug Data

Report (MDDR), and National Cancer Institute Database (NCI)[39]. In the docking process, conformers with different interaction are created and compared with each other. The continuous search for a perfect conformer continues until accepted. The conformers after docking are preferred based on their binding affinities and free energy binding rather than their binding orientation. False positives are a major concern during the docking procedure; this is usually avoided based on different scoring functions to the same docked conformer. Molecular docking is performed mainly based upon two approaches i.e. the simulation approach and the second shape complementary approach. The simulation approach is made by ensuring a physical distance between the ligand and the target, so here, binding occurs only after certain times of moves in its conformational space. At definite time movement of the ligand and target impacts on the structure of the ligand either internally (torsion angle rotations) or externally (rotations and translations)[40]. The ligand movement within the conformational limit releases energy as total energy. This approach provides an advantage towards accepting the flexibility of the ligand molecule and also can easily assess the molecular recognition between ligand and the target. However, process of docking takes a larger time to estimate optimal docked conformer due to huge amount of energy release for each conformation. Still, recent developments in molecular docking has successfully developed fast optimization methods and grid-based tools to make the simulation approach more user-friendly. The shape complementary approach uses the surface structural characteristics of the ligand and target to provide information about molecular interaction. The surface of the target is featured on this approach with the solvent-accessible surface area, whereas the ligand surface is featured based on its matching surface illustration. The complementarity is established between two surfaces based on the matching of shapes, which helps find the complementarity groove for a ligand on the surface of the target. In the target protein molecule, the hydrophobicity can be calculated by measuring the number of turns in the main-chain atoms. This approach of molecular docking is relatively fast. It involves fast scanning of a large number of ligand molecules within seconds to find out possible binding properties of ligand on the target protein surface.

Docking tools are mostly used based on search algorithms such as genetic algorithm, fragment-based algorithms, Monte Carlo algorithm, and molecular dynamics algorithms. High throughput docking studies and simulations are done using tools such as DOCK, GOLD, FlexX, and ICM[41]. Based on the objectives of docking simulations, different types of docking procedures are present, which involves either ligand/target flexibility or rigidity. Lead optimization hit identifications, drug-DNA interactions are some of the basic

applications of molecular docking[42]. Before caring out at the experimental level, the feasibility of any chemical reaction can be easily obtained through molecular docking. It can also predict an optimized orientation of ligand on its protein target. Also, it provides the different binding aspects of a ligand to the groove of its target. This methodology can pave the way to develop more potent and efficient drug candidates. The scoring function in molecular docking helps to evaluate large databases to rule out potent drug candidates targeting the protein of interest *in silico*.

Molecular docking also has an important role in the initial prediction of drug interaction to the nucleic acids, information obtained here is critically essential to assess the cytotoxic potential of the drug. This approach is generally used by medicinal chemists to elucidate the anti-cancer potency of drugs at the molecular level by assessing the drug interaction with the nucleic-acids. Some of the basic challenges that remain with molecular docking are for ligand preparation, receptor flexibility, and scoring function[43]. Preparation of ligand is the key step in molecular docking as recognition of ligand by any target biomolecule depends on the three-dimensional conformation and its electrostatic interaction. All relevant protonation state should be taken into account considering pKa of ligands titratable groups and the fact, that pKa can be slightly shifted after ligand to receptor binding. The majority of the databases, keep the molecules in their neutral form, but upon physiological conditions, the molecules are in an ionized state. Therefore, making it compulsory to ionize the molecules before docking. The handling of a flexible protein molecule is yet another major challenge as it adopts different conformations based upon different ligands used for binding.

Different conformational states of a protein are the most neglected aspect in most of the docking studies as it accounts for the better affinity between the target protein and the ligand molecule. Another important factor for target flexibility is the water molecules present at the active site of the target. Water molecules must can be eliminated to avoid artifacts in the docking process, on the other hand, some water molecules can significantly contribute to ligand binding and such should be kept in the structure and taken into account in drug design[44]. Couple of approaches has been develop to identify water molecules, which may play important structural role in ligand binding, e.g., WaterMap. Imperfection in the scoring function is yet another challenge faced in molecular docking approaches. Scoring functions should potentially be able to differentiate true binding modes from the rest of the parallel modes. Despite all the progress the scoring functions face still many challenges and their reliability is limited. The lack of scoring functions based on accuracy and speed remains a major drawback in molecular docking programming.

The current scenario of the SARS-CoV-2 pandemic marks the need for such approaches for the rapid drug discovery. The molecular-based studies and target-based virtual screening approaches have moved at a much faster pace since the consideration of the first ligand-bound SAR-CoV-2 3CLpro, main proteases crystal structure in February 2020[45]. Later many of the virtual compound libraries were in demand and have achieved some successes. Many of the crystal structures of SARS-CoV-2 were generated and were utilized by scientists across the globe for virtual docking studies with already existed drugs for a drug repurposing approach. Spike-protein, nsp-9, nsp-5, Mpro, and other polyproteins and proteases provided a way to be analyzed for different docking approaches against different FDA drugs[46]. Many drugs, such as the hydroxychloroquine, talampicillin, rubitecan, loprazolam, etc, have been already reported as an effective drug against different targets in SARS-CoV-2[47].

5. Access data and computational Resources to address SARS-CoV-2

There have been immense efforts made by several federal agencies such as NIH, public consortia, and private entities to come up with open access data and computational resources to bring out fast solutions against SARS-CoV-2[48]. Some of the resources listed are Amazon Web Services (AWS) data lake for analysis of COVID-19 data. This is a centralized repository that includes updated and curated datasets on the spread and characteristics of SARS-CoV-2. It includes several visualization tools, epidemiology healthcare resources, and literature. Broad Terra cloud commons for pathogen surveillance is mostly associated with the genomic data of SARS-CoV-2. Resource includes raw SARS-CoV-2 sequencing data from NCBI sequence Read Archive (SRA), workflows for genome assembly, quality control, metagenomic classification, and aggregate statistics. CAS COVID-19 antiviral candidate compounds dataset is an open-source dataset that consists of 50,000 chemical substances subdivided into antiviral drugs and related compounds that are similar in structure to already existing antiviral drugs as to use it for research, data mining, machine learning, and analytics purposes. ClinicalTrials.gov COVID-19 related studies include a list of clinical studies submitted in a structured format directly by the investigators performing the study. The collection of 3D print models of SARS-CoV-2 virions and proteins enables to access of chemical structure data of SARS-CoV-2 proteins. CORD-19 (COVID-19 open Research Dataset and AI challenge is another approach driven to make available a free dataset of 45,000 research articles on SARS-CoV-2 and related coronaviruses to enable the application of natural language

processing and other AI techniques. COVID Digital Pathology Resource (COVID-DPR) is another resource that provides whole slide images of histopathologic samples related to SARS-CoV-2 that mainly focuses on tissues from the lungs, heart, liver, and kidney. EMBL-EBI's COVID-19 Data Portal includes the facility to share and analyze data related to SARS-CoV-2.

GenBank Nucleotide sequences is a platform that provides fast access to viral nucleotide sequences. The other one is the GenBank Protein sequences, which similarly provide rapid, open access to virus conceptually translated protein sequences. Google cloud platform (GCP) datasets for SARS-CoV-2 offer free hosting of a repository of public datasets and queries of the COVID dataset. NCBI Virus: SARS-CoV-2 data hub includes links, search filters for the contents focused upon SARS-CoV-2. These could be downloaded from the repository and contains the most updated nucleotide and protein sequences from GenBank and RefSeq (taxid 2697049). PubChem contains small molecule compounds, bioactivity data, biological targets, bioassays, chemical substances, patents, and pathways. RCSB Protein Data Bank COVID-19/SARS-CoV-2 Resources offers access to SARS-CoV-2 related PDB structures to research and related images and videos for education. Reactome is a free open source database that provides intuitive bioinformatics tools to visualize, interpret, and analyze the pathways to support basic research, genome analysis, modeling, systems biology, and education. SARS-CoV-2 related structures is a database that carefully validates the viral protein structure and models that are being deposited to the protein data bank. Sequence Read Archive (SRA) is another resource that provides fast access to viral nucleotide or metagenomic sequences. several computational based approaches that have been actively taking part to address SARS-CoV-2 issues. Atrio is one such resource that offers easy access to huge numbers of high-performing GPU and CPU resources. Beta coronavirus BLAST is a database that contains sequences from Beta coronavirus, including the SARS-CoV-2 sequences in GenBank and RefSeq. One other freely available high-performing computing resource for SARS-CoV-2 related research is Cloud resources for SARS-CoV-2 research.

6. ADME

Successfully finding out the hit compound that binds the target of interest i.e., to a specific protein molecule and confirming its activity in vitro does not justify its actual effect in *in-vivo* conditions. Further, a better understanding of how the compound is properly absorbed by the organism? And how the drug is distributed to the site of interest without the metabolic loss and

excretion? will be helpful[49]. The influence of compound in animal models through modulation of one of these stages is ADME, which abbreviates Absorption, Distribution, Metabolism, and Excretion. These properties determine the drug-effectiveness and how the drug would be processed inside the animal models. ADME generally determines the compound's physical properties or vice versa, such as hydrophobicity and its solubility. An ideal drug needs to be soluble enough to enter into the bloodstream to reach the target and also equally should be lipophilic to pass the lipid bilayer that composes the cell membrane. These critical factors are important functions of a particular structure of the compound, which can mostly be determined using In silco platforms with mathematical algorithms. The ADME properties determining the drug solubility based on skin, gut-blood, and brain-based barriers can also be determined using computational approaches based on the existing database for known drugs[50]. Bioavailability or the proper absorption of the drug and its amount reaching the bloodstream is an essential factor to determine the effectiveness of the drug apart from its potency. Compounds having lesser bioavailability but show a higher activity towards the protein of interest remains a not so effective drug for the target. Lipinski's rule of 5, which was developed by Lipinski et al.; remains the most popular method for determining the bioavailability of a compound[51]. This rule is based on the ADME properties of known compounds, and an orally active drug should fulfill three out of four criteria from the stated rule to be stated as most active. The four criteria are:

1. A maximum of 5 hydrogen bond donors
2. A maximum of 10 hydrogen bond acceptors
3. Molecular weight of fewer than 500 daltons
4. An octanol-water partition coefficient not greater than 5

These criteria involved in Lipinski's rule of 5 determine how active the oral drug is among the screened compounds and can be filtered at the early stages of virtual screening. However, Lipinski's rule is only applied for orally-active biomolecules and not for drugs administered through other routes.

7. SARS-CoV-2 drug discovery through machine learning approaches

With the fast spread of the coronavirus disease, it has been a major concern to develop therapeutic approaches through drug repurposing to enable rapid clinical trials and regulatory views. The lifecycle of the coronavirus has emerged as the most important step to figure out potential targets against SAS-CoV-2. This includes the endocytic entry of the viruses to the host cell

with the help of ACE2 receptor and the transmembrane protease serine 2 (TMPRSS2), RNA replication and transcription, which involves the helicase and RNA-dependent RNA polymerase (RdRp), translation and proteolytic processing of viral proteins such as the chymotrypsin-like and papain-like proteases, virion assembly and the viral release through exocytic systems[52]. Although viral targets are of major focus some of the host targets stand equally important for viral replication and disease progression. SARS-CoV-2 entry is mediated through the ACE-2 receptor. Recombinant ACE2, namely rhACE2, APN01, has been developing to treat acute lung injury and pulmonary arterial hypertension and has been successfully proven in phase 1 trial in healthy volunteers[53]. RhACE2 has been successfully reported to reduce the viral entry into the host cells derived organoids. This study has led to further clinical trials based on the blockade of viral entry with APN01 for COVID-19 patients. Another important player in a viral entry is the spike glycoprotein whose proteolytic cleavage is done by TMPRSS2[54]. Camostat, an inhibitor of TMPRSS2, is already approved in Japan for the treatment of chronic pancreatitis and postoperative gastric reflux. Camostat, along with nefamostat has been reported to block the viral infection in a mouse model (Fig. 3)[55].

Coronaviruses, before uncoating themselves in the host surface, utilizes the endolysosomal pathway for their entry to the host cell. The drugs, chloroquine and hydroxychloroquine have been in the news for SARS-CoV-2 drug discovery. These anti-malarial drugs, affects the endosomal function by blocking autophagosome-lysosome fusion. Both these drugs were repurposed to target SARS-CoV-2, and have shown to interrupt the replication process in cellular models. Azithromycin is another drug for zika virus and influenza virus, which blocks autophagosome clearance in human cells. However, these drugs didn't last for a long time as a successful candidate for the coronavirus inhibition. Both hydroxychloroquine and azithromycin had potential cardiotoxicity, which could lead to fatal arrhythmia and hydroxychloroquine also had negative effects on the eye of the patients. This emphasizes the importance of clinical study-based trials on humans to test the efficacy and safety of the existing drugs for COVID-19 patients. After the uncoating of viral particles, the RNA of the virus mediates cap-dependent translation to generate two polypeptides, which are cleaved proteolytically to generate viral proteins such as RdRp and two proteases[56]. These proteases can act as effective targets to the various protease inhibitors that are already FDA approved for HIV and other viruses. Repurposing of drugs lopinavir and ritonavir, a known HIV protease inhibitors are used to treat Covid19 patients, showed effectiveness in the clinical approach. The

importance of helicase and RdRp proteins for viral replication complex is a good target to develop virus helicase inhibitors. RdRp is a protein responsible for viral replication and transcription, clearly opening a way to discover molecules for blocking the viral life cycle. RdRp is an important protein present among mostly all viruses, making it possible to repurpose broad spectrum RdRp inhibitor drugs like remdesivir and favipiravir to be approved in clinical trials[57]. Remdesivir is a drug developed against filoviruses that cause Ebola and Marburg diseases. In studies, remdesivir was shown to have an effect against SARS-CoV-1 in animal models (Fig. 2).

Figure 3. Role of artificial intelligence and machine learning against SARS-CoV-2.

Whereas, favipiravir developed for influenza also showed a positive effect against SAS-CoV-2 human cells in laboratory conditions. Hydroxychloroquine, azithromycin, Camostat, nefamostat, and viral RdRp inhibitors such as remdesivir and favipiravir are best justified drugs repurposed to treat COVID-19 patients. Additionally, machine learning-based approaches are well utilized based on viral entry or replication process to screen approved drug candidates much more broadly. This justifies the current strategy to find drugs against the viral spread is through a machine learning approach, which enables in identifying molecules at a very short time. Many of the active phytochemicals, anti-malarial, anti-cancer, anti-tuberculosis drugs have been repurposed through such methods. High throughput virtual screening among large databases is considered to be the fastest approach to find suitable targets. PyRx software helps screen databases along with the dock, Vina, and Autodock software's for docking[58].

Autodock utilizes tools such as MGL, which consists of computer-aided drug discovery (CADD) pipeline for high throughput virtual screening of large databases to find an efficient target drug. High throughput virtual screening allows the docking of multiple ligands on a single target protein. One of the

freely available high throughput virtual screening software is PyRx. SARS-CoV-2 based studies require the crystal structure of the main protease or any of its structural proteins to perform molecular screening. Many viral structures are already available in the protein data bank, and can be downloaded freely. The databanks for drug molecules are DrugBank, Zinc Natural Product database, PubChem, where the structures can easily be downloaded. The Zinc Natural Product database contains most of the natural derivatives used against different diseases (Fig. 1 & 2).

It mostly takes more than a decade to develop safe, effective anti-viral medicines to be in the market for public use. One major way considered for fast viral treatment is through drug repurposing with some already existing drugs. One of the studies on drug repurposing screened an already existing chemical library containing 12,000 drug compounds having the potential against SARS-CoV-2. This approach gave 21 effective drugs having the potential against this virus in both non-human primate and human cell lines grown in laboratory conditions. The majority of these drugs had the purpose of treating diseases such as HIV, autoimmune diseases, osteoporosis, and other conditions. Scientist S. Chandra had reported the latest findings with the help of a small molecule drug library called ReFRAME, which was created in 2018 by Calibr, which is a non-profit drug discovery division. They initially developed a high-throughput method for rapid screening of the library of almost 11,987 drugs to find out potential candidates against SARS-CoV-2[59]. Many of such drugs repurposed against SARS-CoV-2 include hydroxychloroquine, Remdesivir, Apilimod. These drugs, in general, serve other purposes such as anti-HIV, anti-hepatitis C, and much more.

8. Concusion

The requirement for increased productivity in the drug development industry to tackle complex diseases such as SARS-CoV-2 has led in the exploration of machine learning-based techniques on a cost-effective basis. The availability of data remains the key source for successful machine learning-based methods. Thousands of data based on online platforms provides the effort to integrate itself into the modeling approaches and modulate the drug development process. The use of mathematical algorithms is one kind of challenge and remains the field that should be handled with enough care. Machine learning-based approaches have been widely used to predict compounds with pharmacological activity, specific pharmacodynamic, and ADME properties with the aim of rapid drug development and its evaluation. Also, artificial intelligence-based approaches such as neural networks,

support vector machines, and genetic reprogramming have been widely utilized for predicting inhibitors, antagonists, blockers, agonists, activators, and protein substrates required as therapeutic agents against a specific target. Tools such as quantitative structure-activity relationship modeling actively helped identify potential active biomolecules from a huge number of candidate compounds rapidly and in a cost-effective manner. approach of machine learning stands a very important aspect in the current situation for vaccine development against SARS-CoV-2.

Acknowledgment: The author acknowledged Dhruv Kumar & Jayaramulu Kolleboyina who designed study; Ramesh Chaudhary, Brijesh Rathi, Jayaramulu Kolleboyina & Dhruv Kumar for manuscript review and revision.

References

1. https://covid19.who.int/
2. M. A. Shereen, S. Khan, A.Kazmi, N.Bashir and R. Siddique; COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. *Journal of advanced research*, 24 (2020) 91-98.
3. Fangli, Structure, Function, and Evolution of Coronavirus Spike Proteins, *Annu Rev Virol*. 3 (1), (2017) 237-261.
4. S. Satarker, M. Nampoothiri; Structural Proteins in Severe Acute Respiratory Syndrome Coronavirus-2, *Arch Med Res*. 51(6), (2020), 482-491
5. S Raj, V Chandel, B Rathi, D Kumar. Understanding the molecular mechanism(s) of SARS-COV2 infection and propagation in human to discover potential preventive and therapeutic approach. *Bentham Science*, 2020.
6. Y. Li, L Wang, Z. Peng, *et al.*; Basic reproduction number and predicted trends of coronavirus disease 2019 epidemic in the mainland of China. *Infect Dis Poverty*, 9 (2020) 94.
7. A. Chandra Kaushik and U. Raj; AI-driven drug discovery: A boon against COVID-19?. *AI Open*, 1 (2020) 1–4.
8. EH Shortliffe; *Computer-Based Medical Consultations: MYCIN*. *Artificial Intelligence Series*. New York: Elsevier Computer Science Library, 1976.
9. C.J. Kelly, A. Karthikesalingam, M. Suleyman *et al.*; Key challenges for delivering clinical impact with artificial intelligence, *BMC Med*, 17 (2019) 195.
10. A. A. Ardakani, A. R. Kanafi, U. R. Acharya, N. Khadem and A. Mohammadi; Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. *Computers in biology and medicine,* 121 (2020) 103795.

11. who.int/news-room/commentaries/detail/transmission-of-sars-cov-2 implications-for-infection-prevention-precautions.
12. H. Chen, B. Yang, H. Pei and J. Liu; Next Generation Technology for Epidemic Prevention and Control: Data-Driven Contact Tracking. *IEEE access : practical innovations, open solutions*, 7 (2018) 2633-2642.
13. M Cetron, S Maloney, R Koppaka *et al.*; *Isolation And Quarantine: Containment Strategies For Sars*, 2003.
14. E. Mbunge; Integrating emerging technologies into COVID-19 contact tracing: Opportunities, challenges and pitfalls, *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(6) (2020) 1631-1636.
15. B. Mercorelli, G. Palù and A. Loregian; Drug Repurposing for Viral Infectious Diseases: How Far Are We?, *Trends in microbiology*, 26(10) (2018) 865-876.
16. C. Vaishali, P.P. Sharma, S Raj, B. Rathi, D. Kumar, Structure Based Drug Repurposing Through Targeting Nsp9 Replicase and Spike Proteins of SARS-CoV-2, *Journal of Biomolecular structure and dynamics*, (2020).
17. R. A. Khailany, M. Safdar and M Ozaslan; Genomic characterization of a novel SARS-CoV-2. *Gene reports*, 19 (2020).
18. S. Dong, J. Sun, Z. Mao, L. Wang, Y L Lu and J Li; A guideline for homology modeling of the proteins from newly discovered betacoronavirus, 2019 novel coronavirus (2019-nCoV), *Journal of medical virology*, (2020).
19. Y. Tao, M. Shi, C. Chommanard, K. Queen, J. Zhang, W. Markotter, I.V. Kuzmin, E.C. Holmes, S. Tong; Surveillance of bat coronaviruses in Kenya identifies relatives of human coronaviruses NL63 and 229E and their recombination history, *J. Virol*., 91 (2017) e01953-16.
20. M. Gheblawi, K. Wang, A. Viveiros, Q. Nguyen, J. C. Zhong, A. J. Turner, M. K. Raizada, M. B. Grant and G Y Oudit; Angiotensin-Converting Enzyme 2: SARS-CoV-2 Receptor and Regulator of the Renin-Angiotensin System: Celebrating the 20th Anniversary of the Discovery of ACE2. *Circulation research*, 126(10) (2020) 1456-1474.
21. Báez-Santos, Y. M., St John, S. E., & Mesecar, A. D., The SARS-coronavirus papain-like protease: structure, function and inhibition by designed antiviral compounds. *Antiviral research*,2016, *115*, pp.21–38.
22. Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., & Coleman, R. G., ZINC: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 2017, *52*(7), pp.1757–1768.
23. A. Fiser, Template-based protein structure modeling. *Methods in molecular biology (Clifton, N.J.)*, 2010, *673*, pp.73–94.

24. D Kumar, V Chandel, S Raj, B Rathi, In silico identification of potent FDA approved drugs against coronavirus COVID-19 main protease: A drug repurposing approach, *Chemical Biology Letters* 2020.
25. E. Lionta, G. Spyrou, D. K. Vassilatis and Z. Cournia, Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Current topics in medicinal chemistry*, 2014, *14*(16), 1923-1938.
26. L. G. Ferreira, R. N. Dos Santos, G. Oliva and A. D. Andricopulo, Molecular docking and structure-based drug design strategies, *Molecules (Basel, Switzerland)*, 2015, *20*(7), 13384-13421.
27. G. Klebe, Virtual ligand screening: strategies, perspectives and limitations. *Drug discovery today*, 2006, *11*(13-14), 580-594.
28. D. Xu and Y Xu, Protein databases on the internet. *Current protocols in molecular biology*, 2006, *Chapter 19*, 19.
29. A. Gimeno, M. Ojeda-Montes, S. Tomás-Hernández, A. Cereto-Massagué, Beltrán-Debón, R., Mulero, M., Garcia-Vallvé, S., *The Light and Dark Sides of Virtual Screening: What Is There to Know? International Journal of Molecular Sciences, 20(6), 1375.*
30. George Papadatos, Mark Davies, Nathan Dedman, Jon Chambers, Anna Gaulton, James Siddle, Richard Koks, Sean A. Irvine, Joe Pettersson, Nicko Goncharoff, Anne Hersey, John P. Overington, SureChEMBL: a large-scale, chemically annotated patent document database, *Nucleic Acids Research*, 2016, 44(1), 1220-1228.
31. Tejashree Redij, Rajan Chaudhari, Zhiyu Li, Xianxin Hua, and Zhijun Li, Structural Modeling and in Silico Screening of Potential Small-Molecule Allosteric Agonists of a Glucagon-like Peptide 1 Receptor, *ACS Omega*, 2019, *4* (1), 961-970.
32. Jonathan Z. Sexton, Polina V. Danshina, David R. Lamson, Mark Hughes, Alan J. House1, Li-An Yeh, Deborah A. O'Brien, and Kevin P. Williams, Development and Implementation of a High Throughput Screen for the Human Sperm-Specific Isoform of Glyceraldehyde 3-Phosphate Dehydrogenase (GAPDHS), Current Chemical Genomics, 2011, 5, 30-41.
33. S. Kim, Getting the most out of PubChem for virtual screening. *Expert opinion on drug discovery*, 2016, *11*(9), 843-855.
34. A. Cereto-Massagué, L. Guasch, C. Valls, M. Mulero, G. Pujadas, S Garcia-Vallvé, DecoyFinder: An easy-to-use python GUI application for building target-specific decoy sets, *Bioinformatics*, 2012, 28, 1661-1662.
35. Tahir Ul Qamar, M., Alqahtani, S. M., Alamri, M. A., & Chen, L. L, Structural basis of SARS-CoV-2 $3CL^{pro}$ and anti-COVID-19 drug discovery from medicinal plants. *Journal of pharmaceutical analysis*, 2020, *10*(4), 313-319.

36. Wei Zhu, C.Z, chen, K. Gorshkov, Miao Xu, RNA-Dependent RNA Polymerase as a Target for COVID-19 Drug Discovery, *SLAS DISCOVERY Advancing the Science of Drug Discovery*, 2020.
37. Pagadala, N. S., Syed, K., & Tuszynski, J., Software for molecular docking: a review. *Biophysical reviews*, 2017, *9*(2), 91-102.
38. Marion D., An introduction to biological NMR spectroscopy. *Molecular & cellular proteomics : MCP*, 2013, *12*(11), 3006-3025.
39. Shoichet BK, McGovern SL, Wei B, Irwin JJ, Lead discovery using molecular docking. Curr Opin Chem Biol,2002, 6, 439-446.
40. Hart TN, Read RJ, A multiple-start Monte Carlo docking method. Proteins, 1992, 13(3), 206–222.
41. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD, *Proteins*, 2003, 52(4), 609-623.
42. Lamb ML, Jorgensen WL, Computational approaches to molecular recognition, Curr Opin Chem Biol, 1997, 1, 449-457.
43. I. A. Guedes, C. S. de Magalhães and L. E. Dardenne, Receptor-ligand molecular docking, *Biophysical reviews*, 6(1) (2014), 75-87.
44. K. M. Elokely and R. J Doerksen, Docking challenge: protein sampling and molecular docking performance, *Journal of chemical information and modeling*, 53(8) (2013), 1934-1945.
45. S. Kumar, P. P. Sharma, U. Shankar, D. Kumar, S. K. Joshi, L. Pena, R. Durvasula, A. Kumar, P. Kempaiah, Poonam and B. Rathi, Discovery of New Hydroxyethylamine Analogs against $3CL^{pro}$ Protein Target of SARS-CoV-2: Molecular Docking, Molecular Dynamics Simulation, and Structure-Activity Relationship Studies, *Journal of chemical information and modeling*, 2020.
46. H. Yang, W. Xie, X. Xue, K. Yang, J. Ma, W. Liang, et al., Design of wide-spectrum inhibitors targeting coronavirus main proteases, PLoS Biol, 3 (2005), e324.
47. https://datascience.nih.gov/covid-19-open-access-resources.
48. A. K. Singh, A, Singh, A. Shaikh, R. Singh and A. Misra, Chloroquine and hydroxychloroquine in the treatment of COVID-19 with or without diabetes: A systematic search and a narrative review with a special reference to India and other developing countries, *Diabetes & metabolic syndrome*, 14(3) (2020), 241-246.
49. M. P. Doogue and T. M. Polasek, The ABCD of clinical pharmacokinetics, *Therapeutic advances in drug safety*, 4(1) (2013), 5-7.
50. F. Zafar, A. Gupta, K. Thangavel, K. Khatana, A. A. Sani, A. Ghosal, P. Tandon and N Nishat, Physicochemical and Pharmacokinetic Analysis of Anacardic Acid Derivatives, *ACS omega*, 5(11) (2020), 6021-6030.

51. CA Lipinski, F Lombardo, BW Dominy, PJ Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv Drug Deliv Rev*., 46(1-3), (2001) 3-26.
52. R. Poduri, G. Joshi and G. Jagadeesh, Drugs targeting various stages of the SARS-CoV-2 life cycle: Exploring promising drugs for the treatment of Covid-19, *Cellular signalling*, 74 (2020), 109721.
53. N. Roshanravan, S. Ghaffari and M. Hedayati, Angiotensin converting enzyme-2 as therapeutic target in COVID-19, *Diabetes & metabolic syndrome*, 14(4) (2020), 637–639.
54. S. Belouzard, J. K. Millet, B. N. Licitra and G. R. Whittaker, Mechanisms of coronavirus cell entry mediated by the viral spike protein, *Viruses*, 4(6) (2012), 1011-1033.
55. M. Hoffmann, S. Schroeder, H. Kleine-Weber, M. A. Müller, C. Drosten, and S. Pöhlmann, Nafamostat Mesylate Blocks Activation of SARS-CoV-2: New Treatment Option for COVID-19, *Antimicrobial agents and chemotherapy*, 64(6) (2020), e00754-20.
56. D. P. Nayak, Virus Morphology, Replication, and Assembly, *Viral Ecology*, (2000), 63-124.
57. N. Bibi, S. Gul, J. Ali and M. A. Kamal, Viroinformatics approach to explore the inhibitory mechanism of existing drugs repurposed to fight against COVID-19, *European journal of pharmacology*, (2020), 173496.
58. E. Maia, L. R. Medaglia, A. M. da Silva and A. G. Taranto, Molecular Architect: A User-Friendly Workflow for Virtual Screening, *ACS omega*, 5(12) (2020), 6628-6640.
59. B. Naik, N. Gupta, R. Ojha, S, Singh, V K Prajapatiand D Prusty, High throughput virtual screening reveals SARS-CoV-2 multi-target binding natural compounds to lead instant therapy for COVID-19 treatment, *International journal of biological macromolecules*, 160 (2020), 1-17.