Digital Innovations & Contemporary Research in SCIENCE, ENGINEERING & TECHNOLOGY
Vol. 12. No. 4, 2024 Series

Article Citation Format
Obi, E.D., Yentumi, J.A., Ajayi O.O., Omotuyi, O.I. & Ashimolowo, S.B. (2024): SkyNet For Drugs: A Novel User-Centric Approach to Drug Discovery. Journal of Digital Innovations & Contemporary Research in Science, Engineering & Technology. Vol. 12, No. 4. Pp 1 9-36. www.isteams.net/digitaljournal. dx.doi.org/10.22624/AIMS/DIGITAL/V12N4P3

Article Progress Time Stamps
Article Type: Research Article
Manuscript Received: 22nd August, 2024
Review Type: Blind Peer
Final Acceptance:22nd September, 2024

SkyNet For Drugs: A Novel User-Centric Approach to Drug Discovery

[1]*Obi, E.D., [1]Yentumi, J.A., [2]Ajayi O.O., [3]Omotuyi, O.I. & [1]Ashimolowo, S.B.
[1] Autogon Inc. 3002 Falls at Fairdale, 77057, USA
[2]Department of Computer Science, Adekunle Ajasin University, Nigeria
[3]Department of Pharmacology and Toxicology, College of Pharmacy, Afe Babalola University, Ado-Ekiti, Ekiti State, Nigeria
E-mail: eobi@autogon.ai, joshua@autogon.ai, olusola.ajayi@aaua.edu.ng, olaposi.omotuyi@abuad.edu.ng, bolu@autogon.ai;
Phone: [1*]+1 832 925 1036, [1]+233 54 100 6410, [2]+27 64 0499906, [3]+234 807 094 3256, [1]+234 803 527 8071;

ABSTRACT

Traditional drug discovery is an expensive and laborious multi-step process that requires a detailed understanding of disease pathobiology, potential drug target characterization, synthesis, experimental evaluation, and optimization of putative drug candidates as a pretext for clinical evaluation which often does not translate into success. With the advent of whole genome sequencing, machine learning, and artificial intelligence, drug discovery and development are now enhanced both in speed and precision. SkyNet For Drugs (Skynet4D) is an AI-driven platform that automates the preclinical processes by integrating drug-target retrieval based on patient's diagnostic medical reports, and chemical database/ADMETox screening leading to the selection of a potential drug candidate. Skynet4D framework utilizes a combination of Natural Language Processing (NLP-GenAI), custom molecular docking tools (DiffDock®-GNINA®) for high-throughput virtual screening (HTVS), and an AI-enabled ADMETox (ADMETAi®) tool to generate drug candidates with a high chance of being effective within clinical settings. In this paper, the proof-of-concept for Skynet4D was presented with a medical diagnostic report of a COVID-19 patient; leading to the prediction of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) main protease, betaCoV_Nsp5_Mpro as putative targets. The target sequences retrieval, 3D structures retrieval/generation, putative ligand search (COCONUT, LOTUS, ChEMBL, etc.), and ranking were done autonomously within the Skynet4D pipeline. DiffDock docking confidence score of ≥ 1 and a GNINA binding affinity score of ≤ -6.0 kcal/mol, signaled suitable ligands for selection within the DiffDock®-GNINA®. The best-ranked compounds were filtered with ADMETAi® which ultimately gave Lansoprazole as the candidate compound.

Keywords: Drug discovery, Computer-aided drug discovery, Structure-based drug design, Skynet4D, High-throughput virtual screening, Patient-centric drug discovery

1. INTRODUCTION

Traditional drug development methods are painstakingly long, costly, and often have uncertain outcomes. This process usually includes disease identification and selection, validating disease target(s), discovering, and optimizing lead molecule(s), and performing preclinical and clinical trials.[1-3] This process often differs between drugs; for example, according to[1], drugs that provide only minor improvements in disease conditions compared with existing marketed drug solutions usually have a much longer review process than those urgently needed. The cost of taking a drug through the traditional discovery process exceeds US $ 2 billion.[4] The development of computer-assisted drug discovery/design (CADD) has significantly affected the creation of small therapeutic molecules over the past three decades.[5] Computational or *in silico* drug discovery is a modern approach to drug discovery that utilizes computer-based methods to identify, design, and optimize potential drug candidates.[6] In the post-genomic era, there has been a significant increase in the abundance of information available on biomacromolecules and small molecules. This wealth of data has revolutionized the CADD approach and led to tremendous benefits.

Computational drug discovery approaches can be grouped into structure-based drug design (SBDD) and ligand-based drug design (LBDD).[7] SBDD methods, such as molecular docking and de novo drug design, rely on understanding target macromolecules' structure. This information is commonly derived from crystal structures, nuclear magnetic resonance (NMR) data, and homology models, which are readily accessible in protein databases, such as the Protein Data Bank (PDB) and UniProt.[5] In SBDD, a potential therapeutic target and its associated ligands are identified, and using high-throughput screening techniques, these ligands are docked at the binding site of the target protein to identify the most suitable ligands with potential drug-like properties.[8] The top hits ranked by a scoring system using the electrostatic and steric interaction properties of the binding site with the ligand, such as binding affinity, are then selected and synthesized. These top compounds are further tested *in vitro.*[8]

Numerous drugs have been identified through this process. A notable example is the thymidylate synthase inhibitor Raltitrexed. Amprenavir, a potential HIV protease inhibitor, was discovered through protein modeling and molecular docking simulations.[9]

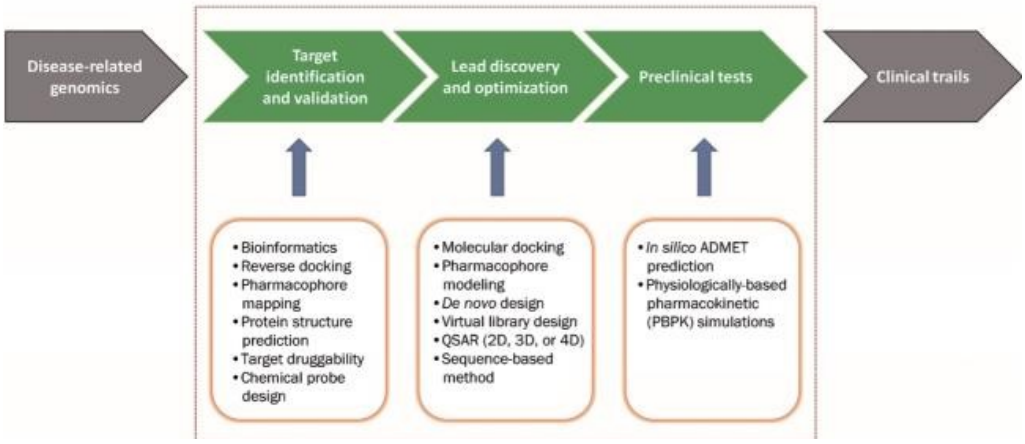


Fig. 1: The Computationally Aided Drug Discovery Pipeline

When potential drug targets lack three-dimensional (3D) structures, LBDD tools, such as quantitative structure-activity relationship (QSAR), pharmacophore modeling, molecular field analysis, and 2D or 3D similarity assessment, are essential for gaining insights into the interactions between drug targets and ligands. This facilitates the construction of predictive models appropriate for lead molecule discovery and optimization.[6] An example of LBDD in action is the discovery of a novel family of Peroxisome Proliferator-Activated Receptor ɣ agonists (PPAR-ɣ) which are receptors for the antidiabetic drug thiazolidinedione.[10,11]
tools.

These computational methods' efficiency, accuracy, and speed rely enormously on various technical settings, such as scoring functions, molecular similarity calculations, conformation generation, and sampling, amongst others.[12–15]Fig. 1 shows an atypical drug discovery pipeline using computational

1.1. RELATED WORKS

The COVID-19 pandemic presented the opportunity to test the efficiency and applicability of CADD techniques in rapidly generating novel targets for new and unknown disease conditions. Using SBDD techniques, virtual screening, and HTS, the authors in (Jin et. al., 2020)[16] identified a mechanism-based inhibitor and through HTS six compounds that inhibit the SARS-CoV-2 M protein. Out of the six compounds, docking scores and *in vitro* antiviral assays revealed that the Michael acceptor inhibitor known as N3 and ebselen as the strongest antiviral inhibitors to the target protein. The study acknowledges that while cell-based phenotypic screening is valuable, its complexity makes it less compatible with high-throughput pipelines. This limitation hinders the ability to effectively identify specific molecular targets or mechanisms of action. Additionally, the high-throughput screening identified hits that may covalently bond to the catalytic cysteine of the SARS-CoV-2 main protease (M pro).

However, these compounds are likely to be promiscuous binders, which may limit their potential as viable drug leads. This raises concerns about the specificity and efficacy of the identified compounds. While the study presents promising leads, further validation through additional experimental studies is necessary to confirm the effectiveness and safety of the identified compounds in clinical settings. This step is crucial for translating laboratory findings into therapeutic options. The study results give further credence to the impact computational drug discovery techniques can provide to the timely development of drug molecules in a pandemic scenario. Yazdani et al, (2022)[17] employed virtual screening and CADD approaches to identify dual-function inhibitors for isoforms of the human inosine monophosphate dehydrogenase (IMDPH).

This enzyme plays a key role as an immunosuppressant in heart and kidney transplants. The methodology involved retrieving the 3D structures of IMDPH isoforms type II (1NF7) and type I (1JCN) and re-docking these compounds using two search algorithms (MolDock Optimizer and MolDock Simplex Evolution) with two scoring functions (PLANTS score and PLANTS score grid) in the Molegro Virtual Docker v6.0 program. ZINC15 was utilized as the ligand screening database with selected ligands downloaded in their 3D format for virtual high-throughput screening followed by a double-step docking process in the binding pockets of both protein isoforms. The authors used SwissADME and DruLi software as the ADMET prediction tools followed by molecular dynamics simulations using the GROMACS software.

Results of the study showed twelve ligand molecules with potential activity against the IMDPH enzyme with MD simulations resulting in the selection of only a single molecule with high inhibitory activity. Structural analogs can be generated using SBDD approaches and tested for antiviral and anticancer abilities as IMDPH inhibition seems to play a role in improving outcomes in patients with Covid-19. In Delre et al, (2022)[18], the authors investigated potential molecules capable of reducing or preventing human-ether-a-go-go-related (hERG)-mediated drug-induced cardiotoxicity. Their methodology involved the use of a ligand-based (QSAR)-machine learning approach involving the development of six classification algorithms, random forest, K-nearest neighbors, gradient boosting, extreme gradient boosting, multilayer perceptron, and support vector machines to predict molecules capable of inhibiting the hERG potassium channel. ChEMBL and hERG-DB were used as the source of bioactivity data.

Ligand molecules retrieved were sanitized using an in-house procedure and converted to a standardized QSAR-ready format using Obabel on the KNIME Analytics platform, allowing for uniformity in ligand molecule representation. Additional filtering methods were employed such as employing a specified IC_{50} threshold. These methods resulted in a final dataset of 792 compounds for classifier testing. The dataset was split into test and validation sets and using the DRAGON software, molecular descriptors were obtained for each molecule. The Synthetic Minority Oversampling Technique (SMOTE) was used to balance the number of blocker and non-blocker samples in the test dataset. A five-fold CV was used in training all six classification models which accurately predicted three compounds from the potential drug pool as hERG-blockers.

From the study, the authors demonstrate the applicability of incorporating machine learning models in the drug discovery pipeline. The models can be combined with SBDD approaches for drug molecule classification and identification. Despite this, further studies should investigate the applicability and efficiency of this ligand-based prediction tool against other toxicity-induced disease conditions. Tuberculosis (TB) caused by *Mycobacterium tuberculosis* remains a significant global health threat, exacerbated by drug-resistant strains and their synergy with HIV. In Aina et al, (2024),[19] the authors explored the potential of quinoline derivatives as new anti-TB agents, leveraging computational chemistry to design and screen novel compounds. 2-chloroquinoline-3-carbaldehyde was selected as a lead compound based on its drug-likeness and favorable bioavailability parameters and then structurally modified to generate thirty-two (32) hypothetical compounds with potential inhibitory effects on *M. tuberculosis*.

These compounds were then analyzed for drug-likeness using ADME criteria in SwissADME, resulting in fourteen compounds showing potential as drug candidates. Toxicity assessments indicated that five selected compounds exhibited no significant toxicity, making them suitable for further evaluation. Twenty-seven (27) standard drugs for some selected diseases were downloaded from PubChem to serve as reference drugs to compare the performance of these hypothetical compounds. Docking simulations using Pyrx AutoDock indicated that two of the five compounds exhibited higher binding energies against various disease proteins than standard anti-TB drugs. The results suggest that the designed compounds could outperform existing treatments, warranting further investigation.

The findings highlight the effectiveness of *in-silico* methods in drug design, particularly for addressing challenges posed by drug-resistant TB strains. Future studies should further explore the pharmacological properties and therapeutic potential of these compounds in clinical settings. In Avilés-Alía et al, (2024),[20] the authors highlighted the urgent need for effective antivirals against SARS-CoV-2, especially given the limitations of current treatments and the challenges posed by vaccine hesitancy and viral evolution.

The study emphasized the potential of drug repurposing and computational methods to expedite the identification of antiviral compounds targeting the SARS-CoV-2 spike protein, particularly the receptor binding domain (RBD). The target protein receptor's crystal structure was retrieved from the PDB and used as the initial structure for molecular dynamics simulations, performed using the AMBER20 suite of programs.

A comprehensive virtual screening process was conducted using FDA-approved drugs and natural products from the Selleck FDA-approved drugs and Selleck database of Natural Products, employing ensemble docking and pharmacophore-guided approaches to identify potential inhibitors. The screening resulted in a shortlist of 48 compounds, which underwent further evaluation through molecular dynamics simulations to determine their binding efficacy. Ten compounds showed significant binding efficacy and were procured for further biological validation.

The selected compounds were tested *in vitro* using pseudotyped vesicular stomatitis virus assays to evaluate their antiviral activity against SARS-CoV-2, leading to the identification and selection of two compounds that showed no toxic effects in cells. To conclude, the research successfully identified two lead compounds as promising candidates for further development as antiviral agents against SARS-CoV-2, with distinct mechanisms of action. The findings underscore the effectiveness of the computational drug repurposing approach in rapidly identifying potential treatments for emerging viral threats. Further studies could employ additional computational approaches in structurally improving these compounds to enhance their efficacy against SARS-CoV-2.

In Ayodele et al, (2024),[21] the authors aimed to identify the deleterious non-synonymous single nucleotide polymorphisms (nsSNPs) in the O-linked N-acetylglucosamine transferase (OGT) gene that could serve as therapeutic targets for diabetes. One hundred and fifty-nine (159) SNPs were retrieved from the National Centre for Biotechnology Information (NCBI) dnSNPs server, and their effects were investigated using PhD-SNP, SNPs&Go, PROVEAN, and Polyphen software tools. This resulted in identifying seven (7) SNPs that generated consistent deleterious effects across the four tools.

Here, we present SkyNet For Drugs (Skynet4D) as an AI-driven platform that automates the preclinical processes by integrating drug-target retrieval based on patient's diagnostic medical reports, and chemical database/ADMETox screening leading to the selection of a potential drug candidate. Skynet4D framework utilizes a combination of Natural Language Processing (NLP-GenAI), custom molecular docking tools (DiffDock®-GNINA®) for high-throughput virtual screening (HTVS), and an AI-enabled ADMETox (ADMETAi®) tool to generate drug candidates with a high chance of being effective within clinical settings.

2. METHODS

2.1 SkyNet for Drugs Application Pipeline

To improve the drug discovery process via automation, Skynet for Drugs (SkyNet4D) uses a novel computational approach to perform SBDD. First, a medical diagnostic report is scanned, the diagnosed disease conditions are retrieved, and a high-throughput database search for the target protein sequence and the possible ligands known to bind to the sequence are retrieved. The SBDD process begins with the molecular docking of all retrieved proteins and their ligands. Using a combination of binding affinity and confidence scores from our docking tools, the best ligands were selected and passed through a chemical absorption, distribution, metabolism, excretion, and toxicity (ADMET) screen to identify molecules with the best drug-likeness scores (Figure 2).

SkyNet4D was the first to use patient medical reports to generate small-molecule drug targets, significantly improving the drug discovery process for research scientists and pharmaceutical companies

Fig 2: SkyNet for Drugs (SkyNet4D) Computer-Assisted Drug Discovery Process

2.1.1 Patient Medical Report Screening

The first step in using SkyNet for Drugs (SkyNet4D) is uploading a patient medical report for further downstream processing. We employed an NLP pipeline that used GenAI to read patient reports and identify patient-reported medical conditions. Using OpenAI GPT model 3.5, the NLP pipeline was primed to identify specific diseases or conditions stated in the medical report, determine the types of compounds related to the identified disease condition, and identify the biological target in the body that is compromised by the disease condition.

Owing to its vast training data collection, the pipeline is further primed to retrieve all genetic, pharmacological, and medical information from relevant databases concerning the identified disease conditions and biological targets. This information included the target disease treatment, causal organism of the condition, and available treatment options, among other specified search terms.

2.1.2. Protein Sequence and Bioactivity/Ligand Database Search

After obtaining the target causal organism, we performed a global search of the Protein Data Bank (PDB) using the organism's name and the target causal protein identified from the NLP-GenAI pipeline. To verify that the protein sequence obtained from the PDB was the correct target sequence, cross-validation of the sequence was performed using its unique accession number from UniProt. PDB IDs and their matching UniProt accession numbers confirmed the target protein sequences. This process was repeated for each disease-related protein sequence. Other data annotations retrieved from UniProt included protein sequence length, molecular weight, common name, *taxid, uniProtkbId,* and *nucleotideId*.

A ligand search was performed in the Chemistry Database for Bioactive Molecules (ChEMBL) using the protein accession number. The target hits obtained were filtered using a single protein identifier and score ID. An activity search was performed on all filtered queries that returned results using the *target_chembl_id* as the search term, and further filtering for inhibitory activity using the IC50 search term. This search returned a data frame containing the target protein, its IC50 values, and all known ligands with activity against the target protein. The ligand data frame was filtered to ensure that only the best ligands were selected for downstream processing. All inactive or partially active ligands (with IC50 threshold values greater than 1000) were excluded. The ligand structures were sanitized by removing duplicates and null values. To widen the ligand search space, natural databases, such as COCONUT and LOTUS, were searched using the same search criteria described for ChEMBL. All results obtained were stored in a local database.

2.2 DiffDock®-GNINA®

Following the protein-ligand search, each protein sequence and its associated ligands were appropriately prepared into respective structured data file (SDF) and protein data bank (pdb) formats. SMILES was checked for molecular validity using the RDKit Cheminformatics tool before converting to SDF using the Open Babel toolkit. The PDB format for each protein sequence was obtained from the Protein Data Bank repository. Protein-ligand pairs in their respective file formats were then fed to the DiffDock Molecular Docking tool using twenty (20) inference steps, batch size of ten (10) and ten (10) samples per protein-ligand complex, to undergo molecular docking.

This step facilitated the identification of ligands with strong binding affinities to the target disease protein sequence. The results from the DiffDock molecular docking process were then fed into the GNINA docking protocol to establish protein-ligand pairs with the strongest binding affinities. Using a filter score of DiffDock confidence score and GNINA minimized affinity score, the molecules generated from the DiffDock process were filtered, and only the best ligands were selected.

2.3 ADMETAi®

The ADMET score is an important metric for measuring the drug-likeness of a molecule. The chemical absorption, distribution, metabolism, excretion, and toxicity (ADMET) score measures several parameters that contribute to determining whether a proposed molecule is suitable as a drug candidate during *in silico* drug discovery (Guan et al, 2019). Owing to the vast nature of these molecular predictors, we used an industry-based standard model, ADMETAi, which takes a SMILE molecule in SDF format and returns a list of ADMET descriptors, such as molecular weight of the molecule, solubility (logP), hydrogen bond acceptor and donor values, quantitative estimation of drug-likeness (QED), bioavailability, and Lipinski score amongst many others.

3. RESULTS

3.1 SkyNet4D Implementation

3.1.1 NLP report screening and bioactivity database query

In developing our application, our first approach involved using an NLP technique combined with API calls to OpenAI's GPT-4 to screen and interpret the provided patient medical report. This approach provides an efficient way to identify the correct diseases and their targets. The results were set up under sections, including the name of infection/disease, causal organism, known target treatment options, disease target mechanism of action, type of target nucleic acid (DNA/RNA), and disease/non-disease, among other selected headings.

Using the Causal organism, name of disease/condition, and name of the target protein, a database search on Protein Data Bank (PDB) and UniProt was performed to retrieve the protein sequence of the disease condition. UniProt ID was then used to perform a Chemistry Database for Bioactive Molecules (ChEMBL) search to retrieve all ligands for that protein sequence.

The ligand search was further expanded to include natural ligand derivatives from natural product databases, such as COCONUT and LOTUS. Using the selection criterion, only ligands with strong inhibitory concentration (IC50) values as provided by ChEMBL were selected for further downstream processing. Figure 3A shows the results of the NLP screening pipeline when run on a test medical report. Figure 3B shows the results of the protein sequence database query before performing the bioactivity query. SkyNet4D was tested using a simulated patient medical data record of Patient John Doe to demonstrate this process. SkyNet4D successfully recognized that the patient had been diagnosed with COVID-19 pneumonia from the NLP screening pipeline.


```
[{'primaryAccession': 'P0DTD1',
  'uniProtkbId': 'R1AB_SARS2',
  'organism': 'Severe acute respiratory syndrome coronavirus 2',
  'taxid': 2697049,
  'sequence': {'value':
'MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQHLKDGTCGLVEVEKGVLPQLEQPYVFIKRSDARTAPHGHMMVELVAELEGIQYGRSGETLGVLVPHVGEIPVAYRKVLLRKNGNKGAGGHSYGAI
   'length': 7096,
   'molWeight': 794058,
   'crc64': 'A4E62D971508B8CC',
   'md5': 'E6608B50FCD6E004708A875615DDF2D9'},
  'protein': 'Replicase polyprotein 1ab',
  'nucleotideId': None},
 {'primaryAccession': 'P0DTC1',
  'uniProtkbId': 'R1A_SARS2',
  'organism': 'Severe acute respiratory syndrome coronavirus 2',
  'taxid': 2697049,
  'sequence': {'value':
'MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQHLKDGTCGLVEVEKGVLPQLEQPYVFIKRSDARTAPHGHMMVELVAELEGIQYGRSGETLGVLVPHVGEIPVAYRKVLLRKNGNKGAGGHSYGAI
   'length': 4405,
   'molWeight': 489989,
   'crc64': '7F8A21148A7A7E2A',
   'md5': 'E781B58591B8D8DD15F84DCBDEC82105'},
  'protein': 'Replicase polyprotein 1a',
  'nucleotideId': None},
 {'primaryAccession': 'P52292',
  'uniProtkbId': 'IMA1_HUMAN',
  'organism': 'Homo sapiens',
```


Fig. 3: (A) Sample results from protein sequence search on UniProt

```
{'COVID-19_pneumonia': {'Target_Treatment': ['Small Molecule Drugs',
   'Biologics',
   'Supportive Care'],
  'Causal_organism': 'SARS-CoV-2',
  'Target_protein_for_drug_action': ['Viral RNA polymerase (for Remdesivir)',
   'ACE2 receptor (for ACE Inhibitors)',
   'N/A for supportive care'],
  'mechanism': {'Remdesivir': 'Inhibits viral RNA polymerase',
   'ACE Inhibitors': 'Blocks the conversion of angiotensin I to angiotensin II',
   'Statins': 'Inhibits HMG-CoA reductase',
   'Antipyretics': 'Reduces fever',
   'Monoclonal Antibodies': 'Neutralizes virus',
   'Vaccines': 'Induces immunity'},
  'sources': {'FDA': 'https://www.fda.gov/',
   'PubChem': 'https://pubchem.ncbi.nlm.nih.gov/',
   'WHO': 'https://www.who.int/',
   'AHA': 'https://www.heart.org/',
   'ESC': 'https://www.escardio.org/',
   'DrugBank': 'https://go.drugbank.com/',
   'NIH': 'https://www.nih.gov/',
   'CDC': 'https://www.cdc.gov/',
   'VAERS': 'https://vaers.hhs.gov/'},
  'DNAorRNA': 'RNA',
  'Disease_or_NonDisease': 'Disease',
  'Needs_Gene_Editing_boolean': False,
  'Needs_Drug_Fix_boolean': True,
  'hasExistingApprovedDrug_Boolean': True,
  'isExistingDrugTreatableOrCurableOrNotApplicable': 'Treatable'}}
```


Fig. 3(B): NLP Screening results of Patient Doe.

The protein sequence retrieved was then passed (using the accession number) to the ChEMBL database for bioactivity search using the ChEMBL SDK as described earlier. This search yielded 1538 ligand records with inhibitory activity against our target protein (Figure 4A) which was further filtered using the standard IC50 value and canonical smile columns by dropping null and duplicate values. This filtering yielded 1136 valid ligand molecules that were further filtered by selecting only active ligands with strong inhibitory potential against the target protein sequence (IC50 < 1000). This final filtering yielded 581 ligands (Figure 4B).

	action_type	activity_comment	activity_id	activity_properties	assay_chembl_id	assay_description	assay_type
0	None	Dtt Insensitive	19964199	[]	CHEMBL4495583	SARS-CoV-2 3CL-Pro protease inhibition IC50 de...	F
1	None	Dtt Insensitive	19964200	[]	CHEMBL4495583	SARS-CoV-2 3CL-Pro protease inhibition IC50 de...	F
2	None	Dtt Insensitive	19964201	[]	CHEMBL4495583	SARS-CoV-2 3CL-Pro protease inhibition IC50 de...	F
3	None	Dtt Insensitive	19964202	[]	CHEMBL4495583	SARS-CoV-2 3CL-Pro protease inhibition IC50 de...	F
4	None	Dtt Insensitive	19964203	[]	CHEMBL4495583	SARS-CoV-2 3CL-Pro protease inhibition IC50 de...	F
...	...	...	...	...	...	...	...
1533	{'action_type': 'INHIBITOR', 'description': 'N...	None	25099146	[{'comments': None, 'relation': '=', 'result_f...	CHEMBL5260693	Inhibition of MBP tagged recombinant SARS-CoV-...	B
1534	{'action_type': 'INHIBITOR', 'description': 'N...	None	25099147	[{'comments': None, 'relation': '=', 'result_f...	CHEMBL5260694	Inhibition of SARS-CoV-2 RdRP using ATP substr...	B

Fig. 4: (A) Initial bioactivity search results using target protein sequence accession number on ChEMBL SDK.

	molecule_chembl_id	canonical_smiles	standard_value
0	CHEMBL480	Cc1c(OCC(F)(F)F)ccnc1C[S+]([O-])c1nc2ccccc2[nH]1	390.00
1	CHEMBL178459	Cc1c(-c2cnccn2)ssc1=S	210.00
2	CHEMBL3545157	O=c1sn(-c2cccc3ccccc23)c(=O)n1Cc1ccccc1	80.00
4	CHEMBL4303595	O=C1C=Cc2cc(Br)ccc2C1=O	40.00
6	CHEMBL55400	Nc1ccc2cc3ccc(N)cc3nc2c1	360.00
...	...	...	...
1526	CHEMBL5279748	CC(C)(C)[C@H](NC(=O)C(F)(F)F)C(=O)N1CC2(CC2)C[...	36.82
1528	CHEMBL5283975	CC(C)(C)[C@H](NC(=O)C(F)(F)F)C(=O)N1[C@@H]2CC[...	54.64
1529	CHEMBL5266964	CC(C)(C)[C@H](NS(=O)(=O)C1CC1)C(=O)N1C[C@H]2[C...	22.83
1530	CHEMBL5286307	CC(C)(C)[C@H](NC(=O)C(F)(F)F)C(=O)N1[C@@H]2CCC...	18.06
1531	CHEMBL5282079	CC(C)(C)[C@H](NS(=O)(=O)C(F)(F)F)C(=O)N1C[C@H]...	22.42

4(B) Final ligand results containing only ligands with strong and active inhibitory properties against the target protein sequence

3.2 Docking

The protein sequences and ligands obtained from the screening and bioactivity searches were stored in the required PDB and sdf formats, respectively, in preparation for high-throughput screening through docking (Figure 5). For the docking process, we used a dual approach that performed an initial docking with DiffDock and a second docking with GNINA. The results of the DiffDock docking process then served as input for the GNINA docking application. This dual process was important as DiffDock provided a docking confidence score, while GNINA provided a binding affinity score, both of which provided a robust mechanism to allow for selecting only the best ligands for the drug-likeness check.

Using a DiffDock docking confidence score of ≥ 1 and a GNINA binding affinity score of ≤ -6kcal/mol, suitable ligands were selected. We used the COVID-19 main protease, betaCoV_Nsp5_Mpro against selected ligands to test our docking pipeline (Figure 6A). Molecular docking/high-throughput screening is usually a time-intensive operation, however, a full run (using our test protein sequence from John Doe against 581 ligands obtained from the CHEMBL database. Based on docking scores, three hundred and thirty-six molecules (336) showed strong binding affinity for the target protein and were selected for the AdmeTox pipeline.

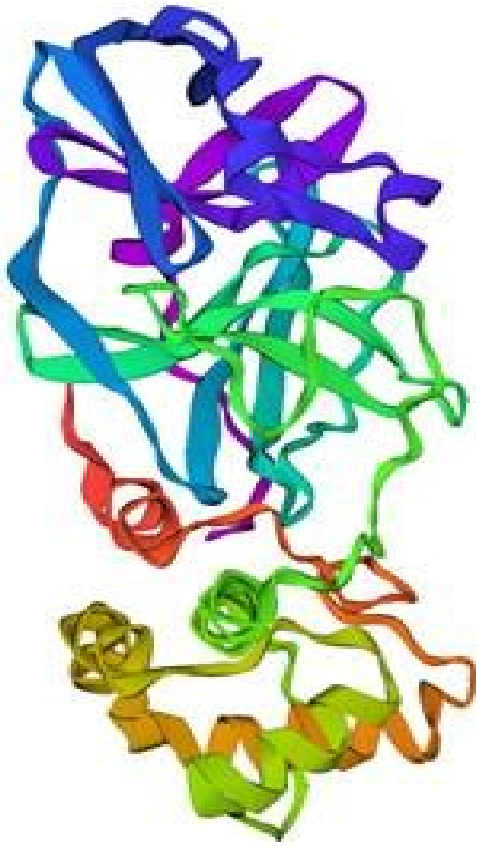
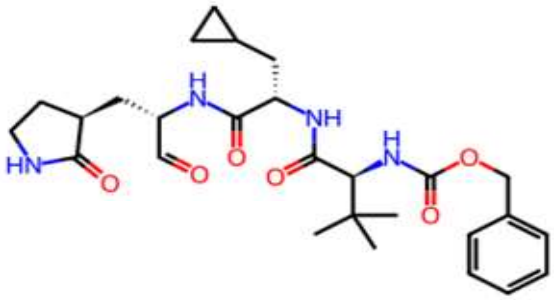
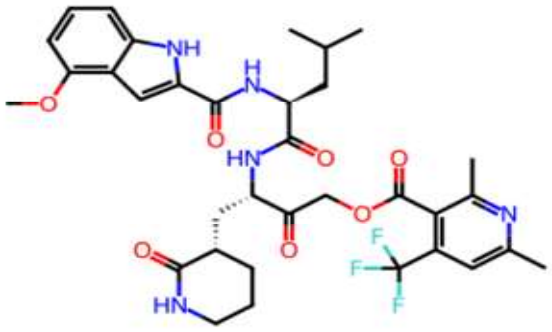


Fig. 5: 3D structure of SARS-CoV-2 main protease betaCoV_Nsp5_Mpro



SMILE:
CC(C)(C)[C@H](NC(=O)OCc1ccccc1)C(=O)N[C@@H](CC1CC1)C(=O)N[C@H](C=O)C[C@@H]1CCNC1=O
Molecule formula: C27H38N4O6
Molecular weight: 514.28



SMILE:
COc1cccc2[nH]c(C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](C[C@@H]3CCCNC3=O)C(=O)COC(=O)c3c(C(F)(F)F)cc(C)nc3C)cc12
Molecule formula: C34H40F3N5O7
Molecular weight: 687.29

Figure 6: Some 2D structures of hit ligands. The best conformation was selected based on the binding affinity and docking confidence scores.

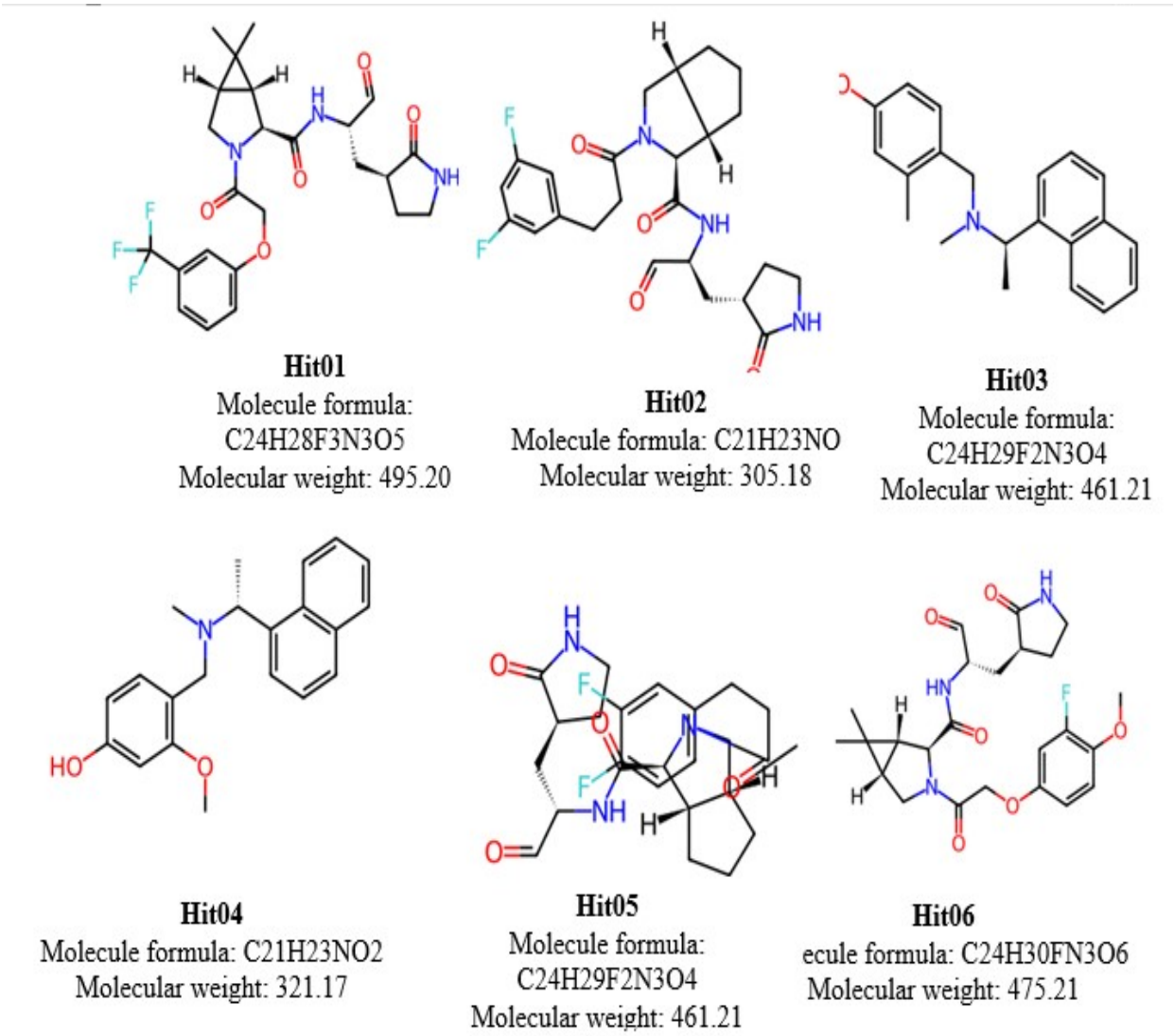
3.3 ADMET Screening

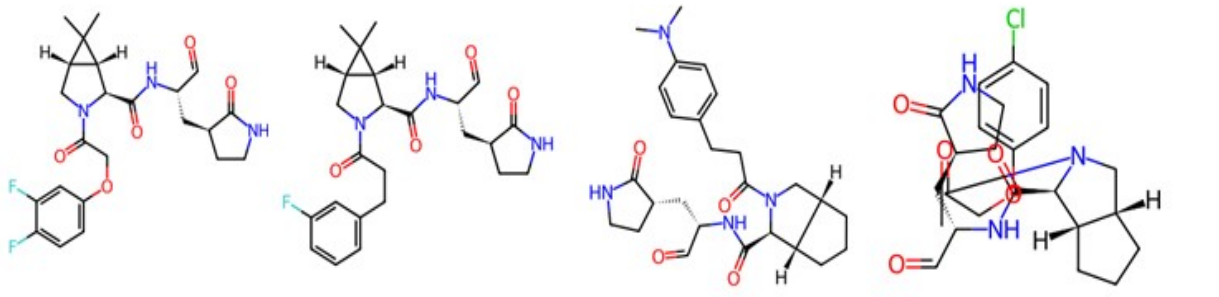
To assess the drug-likeness of the selected ligand molecules obtained from the docking process, we used ADMET analysis to evaluate their absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties.

```
'molecular_weight': 490.6230000000003,
'logP': 4.568400000000003,
'hydrogen_bond_acceptors': 4.0,
'hydrogen_bond_donors': 2.0,
'Lipinski': 4.0,
'QED': 0.4856863182535359,
'stereo_centers': 1.0,
'tpsa': 64.68,
'AMES': 0.1966819792985916,
'BBB_Martins': 0.9502562284469604,
'Bioavailability_Ma': 0.6308143496513366,
'CYP1A2_Veith': 0.12445070445537568,
'CYP2C19_Veith': 0.25390447676181793,
'CYP2C9_Substrate_CarbonMangels': 0.1568199008703232,
'CYP2C9_Veith': 0.09424271062016487,
'CYP2D6_Substrate_CarbonMangels': 0.4720202684402466,
'CYP2D6_Veith': 0.865126633644104,
'CYP3A4_Substrate_CarbonMangels': 0.809829831123352,
'CYP3A4_Veith': 0.5985177457332611,
'Carcinogens_Lagunin': 0.24701070487499238,
'ClinTox': 0.7014313697814941,
'DILI': 0.3578716337680817,
```

Fig 7: Some ADMET Descriptors generated for some selected inhibitory ligands of SARS-CoV-2 main protease

This step is vital in the drug discovery process, as it establishes, *in silico*, the potential of a generated molecule as a drug candidate. ADMETAi simplified the ADMET check process, providing several descriptors of the generated molecules such as molecular weight, Lipinski, logP, ClinTox, AMES scores, etc. (Figure 7). The descriptors also contained additional descriptors generated from the DrugBank repository for benchmarking results. For the test target protein, 336 hit ligands obtained from the docking process were passed to the ADMET pipeline, and using selected ADMET descriptors and their scores, ten (10) molecules were selected as potential drug molecules (Figure 8A and Figure 8B). The descriptors used were the Lipinski (>=2.0), AMES (<=0.7), Bioavailabity_Ma (>=0.4), BBB_Martins (>=0.2), ClinTox (<=0.4), Carcinogens_lagunin (<=0.3), QED (>=0.5), DILI (<=0.6) and HIA_Hou (>=0.7) (Table 1).





Hit07
Molecule formula: C23H27F2N3O5
Molecular weight: 463.19

Hit08
Molecule formula: C24H30FN3O4
Molecular weight: 443.22

Hit09
Molecule formula: C26H36N4O4
Molecular weight: 468.27

Hit10
Molecule formula: C23H28ClN3O5
Molecular weight: 461.17

Figure 8A (above) and Figure 8B (below): Top hits generated from the ADMETox Pipeline

Table 1: Hit ligands and their ADMETox Properties DISCUSSIONS

Ligands	LogP	QED	Lipinski	AMES	Bioavailability	BBB_Martins	ClinTox	DILI	HIA_Hou	Carcinogens_Lagunins
Hit01	1.777	0.535	4	0.616	0.715	0.874	0.378	0.392	0.996	0.237
Hit02	5.046	0.729	3	0.179	0.659	0.895	0.193	0.174	0.999	0.220
Hit03	1.734	0.576	4	0.516	0.703	0.701	0.301	0.331	0.989	0.151
Hit04	4.747	0.736	4	0.324	0.525	0.926	0.297	0.129	0.998	0.064
Hit05	1.734	0.576	4	0.541	0.716	0.790	0.308	0.338	0.996	0.162
Hit06	0.906	0.513	4	0.653	0.777	0.828	0.328	0.569	0.996	0.115
Hit07	1.036	0.560	4	0.602	0.687	0.852	0.308	0.439	0.993	0.146
Hit08	1.451	0.594	4	0.603	0.619	0.890	0.362	0.210	0.996	0.132
Hit09	1.522	0.535	4	0.399	0.509	0.519	0.211	0.397	0.966	0.124
Hit10	1.555	0.571	4	0.320	0.808	0.469	0.289	0.589	0.993	0.169

Since its inception decades ago, computer-aided drug discovery (CADD) has increased the number of clinical drug candidates available. Using advanced physics-based molecular modeling, deep learning, and artificial intelligence techniques coupled with the explosion of genomic and chemical data, some campaigns have shown promising disease target-to-drug molecule time as low as one (1) to two (2) months,[22] or target-to-clinic under one (1) year.[23] The recent development of ligand databases requiring universal methods for the virtual representation of small molecules has led to several reproducible methods for representing small molecules.

Representations such as the Simplified Molecular Input Line System (SMILES), SMILES Arbitrary Target Specification (SMARTS), and International Chemical Identifier (InChI) are the most common representations.[24] SMILES were designed for good human readability in a molecular file format. In contrast, SMARTS allow for variability in the represented molecular structures, allowing for the addition of substructure search functionality to SMILES. InChI provides a non-proprietary machine-readable code unique for all chemical structures.[24,25]

Structure-based CADD (SB-CADD) focuses on modeling the ability of a molecule to interact with a binding site efficiently and favorably on a target protein. This binding site is often strongly involved in the biological functions of the protein.[26] Novel compounds can then be identified by carefully analyzing this protein-binding site. Ligand-based Computer-Assisted Drug Discovery/Design (LB-CADD) analyzes known ligand molecules and their interactions with a target of interest. This is an alternative method for drug discovery that does not rely on an understanding or prior knowledge of the target protein structure. LB-CADD analyses the physicochemical properties and activities of known ligands and generates alternative structural designs of new compounds with desired drug-like properties).[1] The first step in CADD involves identifying drug targets from a large repertoire of candidate macromolecules, a method that is often time-consuming, challenging, and important.[27]

In developing SkyNet for Drugs, we streamlined this process by starting from the patient's medical report and retrieving relevant information such as disease names and target protein sequences. This significantly reduces the time required to identify the appropriate disease target and its associated ligands or binding molecules. Our high-throughput screening (HTS) method involves a multifaceted approach to obtaining ligands. This approach aims to increase the hits generated through HTS by combining the ligand search space to include chemical and natural databases. To ensure the efficiency and appropriateness of the generated ligand hits, we opted for a two-prong molecular docking screening approach, utilizing two industry-standard docking applications. This allowed the selection of the right molecules by combining the docking scores from both applications.

AdMeTox screening of these hit ligands obtained from the docking process produced a considerable number of molecules that could serve as potential drug candidates and would require further testing *in vitro* and *in vivo* to establish the clinical efficacy of these identified targets. In SkyNet for Drugs, we aimed to create a complete system in which new molecules are generated and screened *in silico* for drug-likeness before moving the hit molecules for wet lab or clinical validation. We view this step as crucial as it reduces the number of clinical validation tests required on generated ligand molecules, thus providing significant advantages and improvements in the drug discovery process, especially for resource-poor establishments and regions.

4. CONCLUSIONS

Traditional drug discovery approaches are no longer viable due to exorbitant costs, substantial risk of failure, and time-consuming processes. However, computational approaches, recent advancements in machine learning, and deep learning present significant opportunities to lower costs, improve efficiency, and save time during drug discovery and development. Furthermore, these computational tools increase the discovery rate of new drugs and pave the way for innovative therapeutic interventions.

Machine learning algorithms can pinpoint novel drug targets, unveil hidden correlations between diseases and molecular pathways, and anticipate potential repurposing opportunities for existing drugs. This comprehensive approach to drug discovery, driven by computational methods, holds promise for tackling complex diseases that have historically eluded conventional approaches. The synergy between computational approaches and experimental methods will be crucial in addressing global health challenges. These innovative technologies, in tandem with human expertise, will herald a new epoch of drug discovery characterized by heightened innovation and reduced costs. The continued development of these computational approaches will contribute significantly to shaping the future of pharmaceutical research and improving patient outcomes worldwide.

4.1 Future Perspective

SkyNet For Drugs represents the first phase in our efforts to enhance drug discovery using computational approaches. The next step involves a generative artificial intelligence model that can generate ligand structures by examining protein sequences. This is an effort to address the biological problem of undruggable protein targets.

Funding

This work was funded by Autogon Inc.

Credit author statement

Conceptualization: Obi, ED, Yentumi, JA. Data curation: Obi, ED, Yentumi, JA. Formal analysis: Obi, ED, Yentumi, JA. Funding acquisition: N/A. Investigation: Obi, ED, Yentumi, JA Methodology: Obi, ED, Yentumi, JA. Project administration: Obi, ED, Yentumi, JA, Ashimolowo, B. Resources: Obi, ED, Yentumi, JA. Software: Obi, ED, Yentumi, JA. Supervision: Obi, ED, Ashimolowo, B. Validation: Obi, ED, Yentumi, JA, Ajayi, OO. Visualization: Yentumi, JA. Writing – original draft: Yentumi, JA. Writing – review and editing: Yentumi, JA, Ajayi, OO, Obi, DE, Omotuyi, OI.

Declaration of competing interest

The authors declare that they are in partnership with Autogon Inc.

Data availability

All data used in this process were obtained from the public databases Protein Data Bank, UniProt, ChEMBL, COCONUT, and LOTUS. ADMET AI can be accessed here: https://github.com/swansonk14/admet_ai

Code availability

The code employed in this application development is a proprietary asset and is not available to the public.

Acknowledgments

The support from the Autogon management and the entire team is appreciated.

REFERENCES

1. Berdigaliyev, N., & Aljofan, M. (2020). An overview of drug discovery and development. *Future Medicinal Chemistry*, 12(10), 939-947. https://doi.org/10.4155/fmc-2019-0307
2. Blay, V., Tolani, B., Ho, S. P., & Arkin, M. R. (2020). High-throughput screening: Today's biochemical and cell-based approaches. *Drug Discovery Today*, 25(10), 1807-1821. https://doi.org/10.1016/j.drudis.2020.07.024
3. Guan, L., Yang, H., Cai, Y., et al. (2019). ADMET-score—A comprehensive scoring function for evaluation of chemical drug-likeness. *MedChemComm*, 10(1), 148-157. https://doi.org/10.1039/C8MD00472B
4. Wouters, O. J., McKee, M., & Luyten, J. (2020). Estimated research and development investment needed to bring a new medicine to market, 2009–2018. *JAMA*, 323(9), 844-853.
5. Vemula, D., Jayasurya, P., Sushmitha, V., Kumar, Y. N., & Bhandari, V. (2023). CADD, AI, and ML in drug discovery: A comprehensive review. *European Journal of Pharmaceutical Sciences*, 181, 106324. https://doi.org/10.1016/j.ejps.2022.106324

6. Chang, Y., Hawkins, B. A., Du, J. J., Groundwater, P. W., Hibbs, D. E., & Lai, F. (2023). A guide to in silico drug design. *Pharmaceutics*, 15(1), 10049. https://doi.org/10.3390/pharmaceutics15010049
7. Ece, A. (2023). Computer-aided drug design. *BMC Chemistry*, 17(1), 939. https://doi.org/10.1186/s13065-023-00939-w
8. Batool, M., Ahmad, B., & Choi, S. (2019). A structure-based drug discovery paradigm. *International Journal of Molecular Sciences*, 20(11), 2783. https://doi.org/10.3390/ijms20112783
9. Ejalonibu, M. A., Ogundare, S. A., Elrashedy, A. A., et al. (2021). Drug discovery for Mycobacterium tuberculosis using structure-based computer-aided drug design approach. *International Journal of Molecular Sciences*, 22(24), 13259. https://doi.org/10.3390/ijms222413259
10. Salunke, R. D., Rathod, S. P., Bansode, S. S., & Rao, A. R. (2024). Molecular docking study of 2,4-thiazolidinedione for anti-diabetic by using CADD tools. *World Journal of Pharmacy and Pharmaceutical Sciences*, 13, 769. https://doi.org/10.20959/wjpps20246-27481
11. Da'adoosh, B., Marcus, D., Rayan, A., King, F., Che, J., & Goldblum, A. (2019). Discovering highly selective and diverse PPAR-delta agonists by ligand-based machine learning and structural modeling. *Scientific Reports*, 9(1), 38508. https://doi.org/10.1038/s41598-019-38508-8
12. Bekono, B. D., Sona, A. N., Eni, D. B., Owono, L. C., Megnassan, E., & Ntie-Kang, F. (2021). Molecular mechanics approaches for rational drug design: Forcefields and solvation models. *Physical Sciences Reviews*, 20190128.
13. Sohraby, F., Bagheri, M., & Aryapour, H. (2019). Performing an in-silico repurposing of existing drugs by combining virtual screening and molecular dynamics simulation. In Computational Methods for Drug Repurposing (pp. 23-43). Springer.
14. Li, J., Fu, A., & Zhang, L. (2019). An overview of scoring functions used for protein-ligand interactions in molecular docking. Interdisciplinary Sciences: *Computational Life Sciences*, 11(2), 320-328. https://doi.org/10.1007/s12539-019-00327-w
15. Klambauer, G., Hochreiter, S., & Rarey, M. (2019). Machine learning in drug discovery. *Journal of Chemical Information and Modeling*, 59, 945-946.
16. Jin, Z., Du, X., Xu, Y., et al. (2020). Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature*, 582, 289-295. https://doi.org/10.1101/2020.02.26.964882
17. Yazdani, M., Zamani, J., & Fatemi, S. S. A. (2022). Identification of a potent dual-function inhibitor for hIMPDH isoforms by computer-aided drug discovery approaches. *Frontiers in Pharmacology*, 13, 977568. https://doi.org/10.3389/fphar.2022.977568
18. Delre, P., Lavado, G. J., Lamanna, G., et al. (2022). Ligand-based prediction of hERG-mediated cardiotoxicity based on the integration of different machine learning techniques. *Frontiers in Pharmacology*, 13, 951083. https://doi.org/10.3389/fphar.2022.951083
19. Aina, O. S., Rofiu, M. O., Oloba-Whenu, O. A., et al. (2024). Drug design and in-silico study of 2-alkoxylatedquinoline-3-carbaldehyde compounds: Inhibitors of Mycobacterium tuberculosis. *ScienceDirect*, 23, e01985. https://doi.org/10.1016/j.sciaf.2023.e01985
20. Avilés-Alía, A. I., Zulaica, J., Perez, J. J., Rubio-Martínez, J., Geller, R., & Granadino-Roldán, J. M. (2024). The discovery of inhibitors of the SARS-CoV-2 S protein through computational drug repurposing. *Computers in Biology and Medicine*, 171, 108163. https://doi.org/10.1016/j.compbiomed.2024.108163
21. Ayodele, A. O., Udosen, B., Oluwagbemi, O. O., et al. (2024). An in-silico analysis of OGT gene association with diabetes mellitus. *BMC Research Notes*, 17(1), 6744. https://doi.org/10.1186/s13104-024-06744-5

22. Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*, 37, 1038-1040.
23. Schrödinger. (2022). Schrödinger announces FDA clearance of investigational new drug application for SGR-1505, a MALT1 inhibitor. https://ir.schrodinger.com/node/8621/pdf
24. Wigh, D. S., Goodman, J. M., & Lapkin, A. A. (2022). A review of molecular representation in the age of machine learning. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5), 1603. https://doi.org/10.1002/wcms.1603
25. Choi, S., Lee, J., Seo, J., Han, S. W., Lee, S. H., Seo, J. H., & Seok, J. (2024). Automated BigSMILES conversion workflow and dataset for homopolymeric macromolecules. *Scientific data*, 11(1), 371. https://doi.org/10.1038/s41597-024-03212-4
26. Tang, Y., Moretti, R., & Meiler, J. (2024). Recent advances in automated structure-based de novo drug design. *Journal of Chemical Information and Modeling*, 64(6), 1794-1805. https://doi.org/10.1021/acs.jcim.4c00247
27. Chen, L., Fan, Z., Chang, J., et al. (2023). Sequence-based drug design as a concept in computational drug design. *Nature Communications*, 14(1), 39856. https://doi.org/10.1038/s41467-023-39856-w