Computational and Structural Biotechnology Journal 19 (2021) 424–438

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Exploiting cheminformatic and machine learning to navigate the available chemical space of potential small molecule inhibitors of SARS-CoV-2

Abhinit Kumar [a], Saurabh Loharch [a], Sunil Kumar [a], Rajesh P. Ringe [a], Raman Parkesh [a,b,*]

[a] GNRPC, CSIR – Institute of Microbial Technology, Chandigarh - 160036, India
[b] Academy of Scientific and Innovation Research (AcSIR), Ghaziabad - 201002, India

ARTICLE INFO

Article history:
Received 1 October 2020
Received in revised form 19 December 2020
Accepted 20 December 2020
Available online 29 December 2020

Keywords:
COVID-19
SARS-CoV-2
Repurpose drugs
Chemical space
Gini coefficient

ABSTRACT

The current life-threatening and tenacious pandemic eruption of coronavirus disease in 2019 (COVID-19) has posed a significant global hazard concerning high mortality rate, economic meltdown, and everyday life distress. The rapid spread of COVID-19 demands countermeasures to combat this deadly virus. Currently, there are no drugs approved by the FDA to treat COVID-19. Therefore, discovering small molecule therapeutics for treating COVID-19 infection is essential. So far, only a few small molecule inhibitors are reported for coronaviruses. There is a need to expand the small chemical space of coronaviruses inhibitors by adding potent and selective scaffolds with anti-COVID activity. In this context, the huge antiviral chemical space already available can be analysed using cheminformatic and machine learning to unearth new scaffolds. We created three specific datasets called "antiviral dataset" (N = 38,428) "drug-like antiviral dataset" (N = 20,963) and "anticorona dataset" (N = 433) for this purpose. We analyzed the 433 molecules of "anticorona dataset" for their scaffold diversity, physicochemical distributions, principal component analysis, activity cliffs, R-group decomposition, and scaffold mapping. The scaffold diversity of the "anticorona dataset" in terms of Murcko scaffold analysis demonstrates a thorough representation of diverse chemical scaffolds. However, physicochemical descriptor analysis and principal component analysis demonstrated negligible drug-like features for the "anticorona dataset" molecules. The "antiviral dataset" and "drug-like antiviral dataset" showed low scaffold diversity as measured by the Gini coefficient. The hierarchical clustering of the "antiviral dataset" against the "anticorona dataset" demonstrated little molecular similarity. We generated a library of frequent fragments and polypharmacological ligands targeting various essential viral proteins such as main protease, helicase, papain-like protease, and replicase polyprotein 1ab. Further structural and chemical features of the "anticorona dataset" were compared with SARS-CoV-2 repurposed drugs, FDA-approved drugs, natural products, and drugs currently in clinical trials. Using machine learning tool DCA (DMax Chemistry Assistant), we converted the "anticorona dataset" into an elegant hypothesis with significant functional biological relevance. Machine learning analysis uncovered that FDA approved drugs, Tizanidine HCl, Cefazolin, Raltegravir, Azilsartan, Acalabrutinib, Luliconazole, Sitagliptin, Meloxicam (Mobic), Succinyl sulfathiazole, Fluconazole, and Pranlukast could be repurposed as effective drugs for COVID-19. Fragment-based scaffold analysis and R-group decomposition uncovered pyrrolidine and the indole molecular scaffolds as the potent fragments for designing and synthesizing the novel drug-like molecules for targeting SARS-CoV-2. This comprehensive and systematic assessment of small-molecule viral therapeutics' entire chemical space realised critical insights to potentially privileged scaffolds that could aid in enrichment and rapid discovery of efficacious antiviral drugs for COVID-19.

1. Introduction

Human Coronaviruses (HCoVs), including Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV), Middle-East Respiratory

Abbreviations: WHO, World Health Organization; COVID, COronaVIrus Disease; SARS-CoV-2, Severe Acute Respiratory Syndrome CoronaVirus-2; FDA, Food and Drug Administration.

* Corresponding author.
E-mail address: rparkesh@imtech.res.in (R. Parkesh).

https://doi.org/10.1016/j.csbj.2020.12.028

Syndrome Coronavirus (MERS-CoV), and now Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), all originated from animal sources have led to a worldwide outbreak of viral infection with high mortality rate and morbidity. The current pandemic caused by SARS-CoV-2 has been defying human health [1,2]. Currently, COVID-19 has infected more than thirty million people globally and resulted in around one-million deaths. The high fatality rate with rapid community spread establishes COVID-19 as a major global threat that humankind is facing in the 21st century. Therefore, considering the severity of the disease as well as the pandemic nature, WHO declared SARS-CoV-2 as a Very High Priority Pathogen [3,4]. SARS-CoV-2 was first reported from Wuhan city of China in December 2019 from pneumonia patients [5]. The WHO has estimated that SARS-CoV-2 is ten times more infectious than the typical flu caused by H1N1 [6]. COVID-19 represents an unprecedented "unmet medical need" and hence repositioning of current FDA-approved drugs for COVID-19is very appropriate to achieve timely cure.

Currently, there are no drugs approved by the FDA to treat COVID-19. The FDA's emergency approval to drug Remdesivir has been revoked [7,8] due to toxicity and efficacy issues. COVID-19 disease affects the respiratory system, gastrointestinal system, central nervous system, liver, heart, and kidney and causes multiple organ failure. It has become apparent that due to the complexity of COVID-19 infection, it is crucial to unearth therapeutics that are safe as well as potent. To aid the delivery of potential small-molecule therapeutics for SARS-CoV-2, it is critical to scrutinize and exploit the chemical space of the ligands and repurposed drugs reported for SARS-CoV, MERS-CoV, and other viruses. The extensive cheminformatic understanding of the chemical space of ligands and drugs reported for various coronaviruses and other viral pathogens can create the well-judged design of small-molecule therapeutics and the identification of chemical modulators of SARS-CoV-2. For example, similar ligands are known to exhibit similar affinity for the analogous binding sites. Vital information derived from cheminformatic analysis about the various chemical scaffolds, fragments, and polypharmacological ligands can offer critical insight into the potential extrapolation of bioactivity and structural data for the design of useful molecules.

The present study provides a computational chemistry perspective on effective chemical space development of small molecule inhibitors that target various coronaviruses with significant emphasis on SARS-CoV-2. The data covers all currently known coronavirus related experimental information such as the binding affinity of their protein inhibitors and the various protein targets associated. To assist rational and viable antiviral drug development, effective curation, standardization, simulation, mining, and transformation is applied to known data on coronavirus pharmacology [9] using three distinct datasets called "antiviral dataset", "drug-like antiviral dataset" and "anticorona dataset". We also report the clustering of "anticorona dataset" with "antiviral dataset", based on ECFP6 (extended connectivity fingerprints). This study leads to the identification of the most frequent core fragments and potential polypharmacological ligands for targeting multiple coronavirus family proteins. The ECFP6 belongs to the novel class of topological fingerprints that can be used to extract the structure–activity relationship [10]. The analysis identified antiviral compounds that show high similarity to highly active COVID-19 inhibitors.

In addition to chemical space analysis based on physical properties such as chemical descriptors, we applied the machine learning tool DCA to determine the chemical patterns based on the inductive learning algorithm. We identified chemical patterns that might play an important role in conferring high inhibitory activity against the SARS-CoV-2 main protease. This unique data generated can be highly advantageous towards the design and development of highly potent anti-COVID molecules in a short time. Overall, the study aims to introduce a cheminformatic analysis of coronavirus ligands and their importance in understanding the chemical space occupied by the ligands. The structure–function relationship generated is expected to supply cues to discovering new small-molecule therapeutics, which are expected to have a high impact on antiviral drug development.

2. Materials and methods

2.1. Database generation

We have collated all the molecular data (N = 52,356) that have reported antiviral activity from Chemical Abstract Services and Elsevier's Reaxys Medicinal Chemistry databases. Significant curation, standardization, and transformation are applied to this data set to create three distinct datasets, named as "antiviral dataset", "drug-like antiviral dataset" and "anticorona dataset". The database compounds (N = 52,356) were curated by removing duplicates and complex molecules, resulting in the "antiviral dataset" of 38,428 unique molecules. The "antiviral dataset" was subjected to an oral-bioavailability filter using Lipinski rules, resulting in the elimination of 14,819 molecules. The remaining 23,609 molecules were passed through the "PAINS" filter, which resulted in 20,963 molecules devoid of "PAINS" moiety. We called this dataset as "drug-like antiviral dataset". We then carefully separated molecules with experimentally reported pIC_{50} values that show inhibition against various coronaviruses such as SARS-CoV, MERS-CoV, HCoV 229E, Coronaviridae, and Coronavirinae. We have obtained 433 molecules in this category, and we called this dataset as "anticorona dataset". We created these three distinct datasets, especially to unearth potential chemical entities with the desired drug-like and antiviral properties against Covid-19 to expand the current anticorona data set. The antiviral data set has scaffolds of the highest probability of being exploited as a potential resource with viral inhibitory activities. We used "antiviral dataset," "drug-like antiviral dataset" and "anticorona dataset" for QSAR modelling and machine learning to identify potential molecular scaffolds. The "anticorona dataset" was used for scaffold analysis, activity cliff analysis, fragment analysis, SAR analysis, physicochemical analysis *etc*. We also collated additional datasets including "repurposed drugs-COVID-19 pipeline (N = 42)" [11] "natural products and drugs (N = 83)" [12], "repurposed drugs-Covid-19-clinical trials (N = 19)" [13], and "FDA-approved drugs (N = 2692)" [14] for principal component analysis, as reported in Section 3.3

Various resources were used to collect, analyze, and interpret the cheminformatic analysis data in this study. The resources include Instant JChem [15], CDK [16], RDKiT [17], alvadesc [18], Data Warrior [19], Scaffold hunter [20], MACCS [21], ECFP6 [22] and cheminformatics tools integrated with the KNIME analytical platform [23]. Instant JChem was used for structure database management, search, and prediction, Instant JChem 19.21.5, 2020, ChemAxon. Mona [24] was used for curation and compound library preparation.

2.2. Scaffold analysis

We generated the scaffolds from "anticorona dataset" (N = 433) using the various filters such as atom count, fingerprint, molecular weight, and pIC_{50} to explore fragments, molecular scaffolds, virtual scaffold and their relationship using Scaffold Hunter [20]. Scaffold Hunter first reads the scaffold data from an SQL database and automatically constructs and displays the scaffolds as a tree using various properties like atom count, fingerprints, *etc*. The scaffold tree

shows the relationship between the parent and child molecular scaffolds. The chemical scaffolds - both parent and child and other intermediate virtual scaffolds derived from the parent scaffolds are stored in the database. They can be retrieved as scalable vector graphics (SVG) images for further analysis. The fragment scaffold and the virtual scaffolds derived from the fragments were manually retrieved from the scaffold tree analysis.

We generated Murcko scaffolds by excluding the exocyclic double bonds and the α attached atoms [25] of both the "drug-like antiviral dataset" and "anticorona dataset". The Murcko scaffold was further used to create the skeleton scaffold. The skeleton analysis includes only the ring and replaced the heteroatoms by a carbon atom. We also analyzed structures in "drug-like antiviral dataset" and "anticorona dataset" using the scaffold representation proposed by Bemis and Murcko [26]. In this method, the molecule is dissected into ring systems, linkers, side-chain atoms, and the framework.

2.3. Activity cliff analysis

The structure–activity landscape index (SALI) calculated by the activity cliff analysis supplies a measure between activity (pIC_{50}) and chemical diversity (1–similarity) for each compound [27]. The analysis was carried out using the Skeleton Sphere descriptor, as given by Eq. (1).

$$\boldsymbol{SALI} = \frac{\left|\boldsymbol{A}_{A\pm} - \boldsymbol{A}_{\boldsymbol{j}}\right|}{1 - \boldsymbol{sim}\left(A\ddot{\pm}, \boldsymbol{j}\right)} \tag{1}$$

In this equation, A$A\pm$ and A$\boldsymbol{j}$ represent the biological activity measure of the individual $A\pm^{\text{th}}$ and the $\boldsymbol{j}^{\text{th}}$ molecule, and sim ($A\pm\boldsymbol{j}$) is the similitude among the molecule $A\pm$ and $\boldsymbol{j}$.

2.4. Physicochemical parameter

Physicochemical properties, which include descriptors such as cLogP, H-acceptor, H-donors, molecular weight, polar surface area (PSA), rotatable bond count (RB), relative polar surface area (RPSA), topological polar surface area (TPSA), total surface area, were evaluated using various tools such as CDK, Instant Jchem, *etc.*

2.5. Chemical network visualization

Hierarchical clustering was performed by ECFP6 fingerprint similarity [16,23]. The CDK nodes of the Knime Analytics Platform (KNIME 4.1.2) were used to calculate the ECFP6 fingerprints for "antiviral dataset" and "anticorona dataset". The fingerprints were compared between these two datasets. The edges and nodes were generated for the related compounds and were further used to represent these compounds' chemical networks. We used Gephi 0.9.2 for visualization using variousalgorithmic configurations, for example, Force Atlas, Fruchterman Reingold, Open Ord, Contraction, Force Atlas2, and Yifan, and Yifan Hu Proportional [28]. Binning clustering of the compounds in the database was done for "anticorona dataset" using the ChemMine web server [29], as the experimental value for this dataset is known.

2.6. QSAR modeling by machine learning

We applied DCA (DMax Chemistry Assistant) software [30] to derive the hypotheses and determine the relationship within the morphological and structural features of the anti-COVID ligands and their bioactivities. DCA is an ILP (inductive logic programming)-based software that allows it to develop Prolog rules hierarchically. Such a rule can form the basis for splitting the molecules into two alternative arguments, one that satisfies the hypothesis and the other that does not. DCA can generate the rules based on the individual functional groups and rings and also by incorporating the background knowledge. The background knowledge in DCA is defined by electrostatic, type of elements (e.g., carbon, sulfur, nitrogen), functional group and rings, and linkage (fused, linked, the positional topology of the ring, how different functional groups and rings are connected) between the substructures of the chemical molecule. DCA can relate this background knowledge of the chemical molecules to correlate with their experimental biological activities, such as inhibition or activation. In this study, we used DCA to construct the hypotheses to relate the standard structural features of the "anticorona dataset" with their pIC_{50} values.

2.7. Data visualization

R studio was used to process, analyze, and visualize the plots. The boxplots were generated using the boxplot package of the R language. To generate the 3D-PCA scatter plot, R studio with the plot3d package was used [31]. All images were prepared using Adobe Photoshop CS6 version 13.0 × 64 and Inkscape version 0.92.

3. Results and discussion

3.1. Scaffold analysis.

We analyzed the structures of "anticorona dataset" and "drug-like antiviral dataset" using the scaffold representation as proposed by Bemis and Murcko. In this method, the molecule is dissected into ring systems, linkers, side-chain atoms, and the framework. We analyzed the scaffolds with reported experimental biological activity (pIC_{50} value). Murcko scaffold analysis revealed 227 and 4779 unique scaffolds from "anticorona dataset" and "drug-like antiviral dataset" respectively, with varying degrees of frequencies. The scaffolds benzyl (1-oxo-1-((2-oxo-2-((2-(2-oxopyrrolidin-3-yl) ethyl)amino)ethyl)amino)-3-(pyridin-2-yl)propan-2-yl)carbamate and 1-(tetrahydrofuran-2-yl)pyrimidine-2,4(1H,3H)-dione were observed to have the highest frequencies of 13 and 13.02 respectively (Fig. 1). Singleton scaffold frequency was observed to be 152 in the "anticorona dataset" and 3031 in the "drug-like antivirial dataset". Subsequent Murcko Skeleton analysis resulted in further identification of 143 and 3034 skeleton scaffolds in the "anticorona dataset" and "drug-like antiviral dataset", respectively.

To calculate if the "anticorona dataset" and "drug-like antiviral dataset" are diverse, we used the well-known inequality distribution metrics, Gini coefficient [32,33]. The Gini coefficient was originally used to describe statistical distribution of income amongst a population [32] and is widely used in many different fields [33–36]. The Gini coefficient's value varies between 0 and 1, where 0 implies an equal distribution of income, whereas 1 implies complete inequality, that is, few wealthy individuals represent the major percentage of the total income of the population. In drug discovery, the Gini coefficient has been used to evaluate the diversity of compounds from a sizeable dataset [37]. As in wealth, a lower Gini coefficient shows perfect equality, that is, various molecular scaffolds are important, indicating high scaffold diversity. A higher Gini coefficient means activity is concentrated in a few molecular scaffolds and hence indicating low scaffold diversity. It is particularly advantageous as it only needs the molecules' structure without prior knowledge about the composition. The Gini coefficient computed for the compounds from both the "antiviral dataset" and "drug-like antiviral dataset" showed that both the dataset's compounds show low diversity. For example, "drug-like antiviral dataset" (N = 20,963) shows a Gini index of 0.846762 and entropy of 0.53381, confirming that this data set is of low diversity. This is a

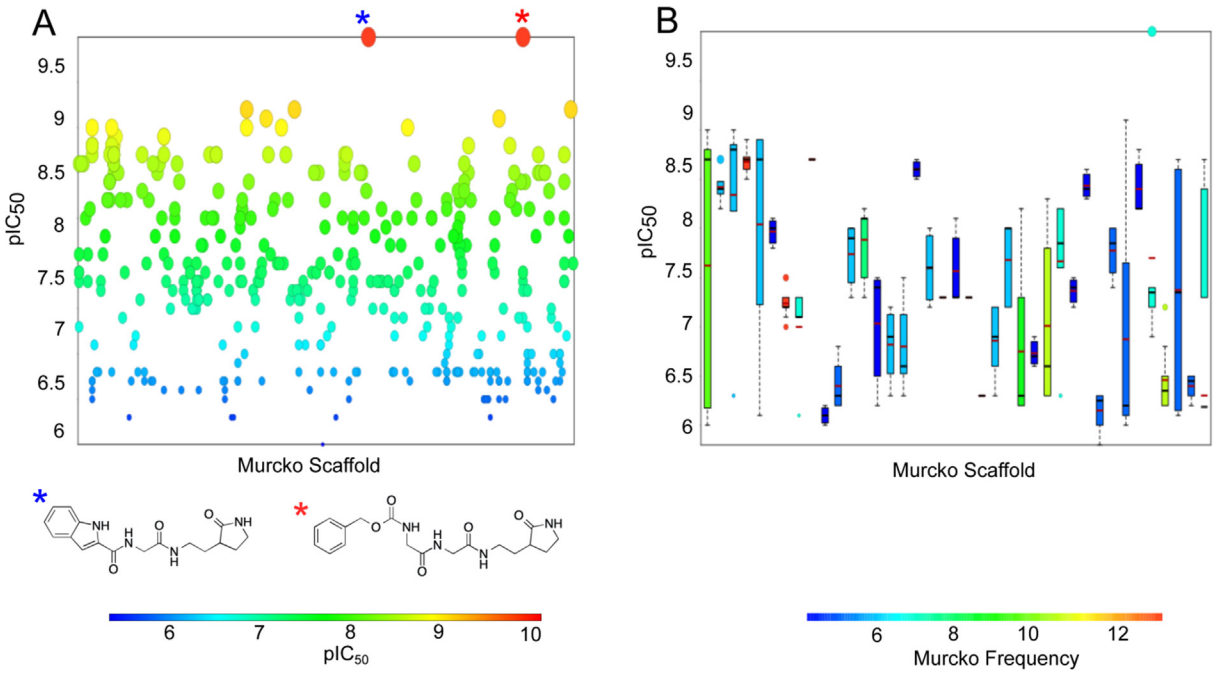

Fig. 1. A) Murcko vs. pIC_{50} value presented as a scatter plot for the "anticorona dataset". Colors indicate pIC_{50} values, with higher and lower values represented by red and blue, respectively. The arrows depict the structures of the corresponding scaffold with the highest pIC_{50}. B) Murcko vs. pIC_{50} value presented as a box plot. Colors depict the Murcko frequency, ranging from blue (lower) to red (higher). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

disadvantage in terms of unearthing unique scaffolds for COVID −19 drug discovery. We could not calculate the Gini coefficient for the "anticorona dataset" as the dataset size (N = 433) is not sufficient for statistical analysis. The Murcko scaffold diversity calculated in terms of the scaffold ratio and total molecules (Ns/M) [38] was computed for both datasets (Table 1). The scaffold diversity analysis of the "anticorona dataset" (Murcko scaffold (0.52), singleton scaffold (0.35), and skeleton scaffold (0.33) suggest diverse chemical representation. It will still require efforts to populate this library with more novel and unique scaffolds to increase diversity further. In contrast, scaffold diversity analysis of "drug-like antiviral dataset" (Murcko scaffolds (0.23), singleton scaffold (0.16), and skeleton scaffold (0.14)) suggests low chemical diversity, as confirmed by Gini coefficient. Interestingly, we were able to identify some promising scaffolds from the "anticorona dataset" with favorable characteristics for designing novel derivatives by SAR studies (Fig. S1). For instance, the scaffolds of N-(2-oxo-2-((2-(2-oxopyrrolidin-3-yl)ethyl)amino)ethyl)-1H-indole-2-carboxamide and benzyl (2-oxo-2-((2-oxo-2-((2-(2-oxopyrrolidin-3-yl)ethyl)amino)ethyl)amino)ethyl)carbamate showed highest biological activity (Fig. 1A).

A scaffold tree of the compounds in the "anticorona dataset" was generated to visualize the standard core structure or scaffold. We first isolated the most active chemical scaffold from the database, then generated the entire possible parent scaffold, followed by selecting one parent–child pair. This process continued until all of the possible successive parent–child scaffold pairs of the "anticorona dataset" were exhausted. The scaffolds were achieved by cutting all of the side chains but keeping the double bonds connected directly to a ring [39]. All 433 molecules of the "anticorona dataset" were pruned until a single ring was attained. We identified oxopyrrolidine, indoline, cyclopropylbenzene, thiophene, indole, dioxole, cyclobutylbenzene, azaspiro, pyranone, and phenylsulfane as some of the most frequent fragment by this scaffold analysis (Fig. 2). Fragment-based analysis of the "anticorona dataset" inhibitors revealed that spiro compounds represent an interesting scaffold-point to develop potent coronaviruses inhibitors. However, so far, only a few spiro compounds had been explored to target coronaviruses (Fig. 2, azaspiro). Further, spiro compounds have inherent three-dimensionality and structural diversity [40]. Therefore, it will be promising to include novel spiro scaffolds for targeting coronaviruses, incredibly challenging to treat SARS-CoV-2 infection.

We then looked for common single ring scaffolds, that are common to molecules representing coronaviruses targets like main protease, papain-like protease, replicase polyprotein 1ab, and helicase. The results depicted in Fig. 3 shows that oxopyrrolidine is the standard basic structure core, present in all these molecules. By scaffold hopping, we were able to construct the promising virtual scaffolds, which can be used as polypharmacological ligands for

Table 1
Scaffold diversity analysis of "anticorona dataset" (N = 433) and "drug-like antiviral dataset" (N = 20,963).

	Dataset size (M)	Murcko scaffolds (N_s)	Singleton Murcko scaffolds (N_{ss})	Skeleton scaffold (N_{sc})	N_{sc}/M	N_s /M	N_{ss}/M	N_{ss}/N_s
"Anticorona dataset"	433	227	152	143	0.33	0.52	0.35	0.67
"Drug-like antiviral dataset"	20,963	4779	3301	3034	0.14	0.23	0.16	0.69

(Nsc/M) shows proportion of Skeleton scaffolds (Nsc) to that of the either "anticorona dataset" or "drug-like antiviral dataset" (M), Ns/M.
Shows the ratio of Murcko scaffolds (Ns) to that of either "anticorona dataset" or "drug-like antiviral dataset" (M), (Nss/M), show the ratio of singleton Murcko scaffolds to that of either "anticorona dataset" or "drug-like antiviral dataset" (M), and (Nss/N) shows the ratio of singleton Murcko scaffold (Nss) to that of Murcko scaffold (Ns).

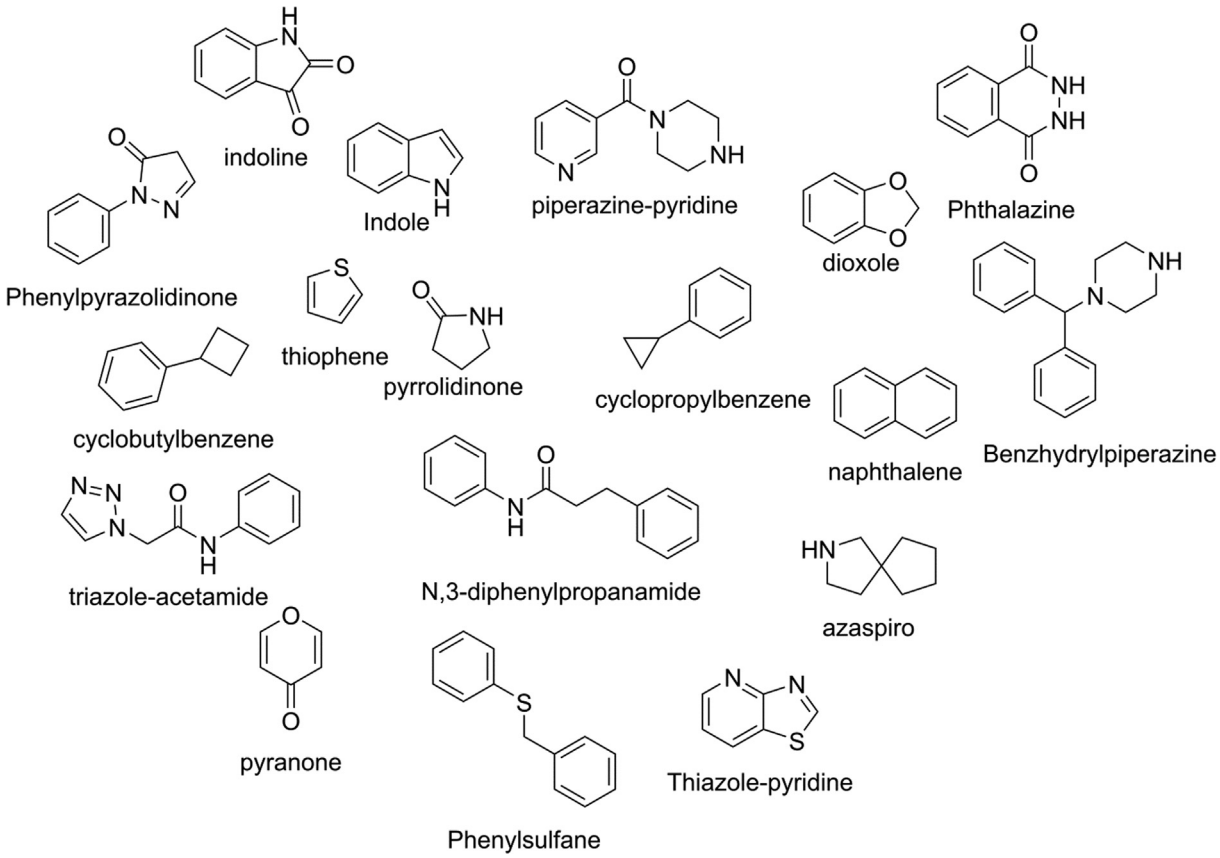

Fig. 2. Representative examples of frequent fragments identified from the "anticorona dataset" using scaffold hopping.

targeting these proteins. These fragments are the ideal starting point for the fragment-based drug discovery for targeting essential SARS-CoV-2 proteins such as main protease, replicase polyprotein 1b, helicase, *etc.* The fragments and virtual scaffolds identified in the present study (Fig. 3) could serve as a possible starting point for further derivatization. These virtual scaffolds can also be used directly as query molecules in high throughput three-dimensional shape-based screening of commercial libraries to identify novel and unique molecules for further biological testing. Understanding and identifying scaffolds will result in the synthesis of new diverse analogs for antiviral drug discovery, leading to the generation of good quality, highly diverse database of small molecules for targeting SARS-CoV-2.

3.2. Physicochemical properties

A drug is expected to specifically act on a biological target and exert therapeutic effects by modulating its function [41]. The bioavailability and efficacy of a drug mainly depends on its physicochemical properties such as absorption, distribution, metabolism, and elimination (ADME) [42,43]. We calculated six pharmaceutically relevant physicochemical descriptors of "anticorona dataset" namely octanol–water partition coefficient (cLogP), molecular weight (MW), hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), rotatable bonds (nRotB), and topological polar surface area using N, O, S and P polar contributions (PSA) to generate ADME profiles.

The distribution of each physicochemical descriptor for inhibitors of the most studied protein targets, including helicase [44], replicase polyprotein 1ab [45], main protease [46,47], and papain-like protease [48], are represented in the form of boxplots in Fig. 4. Each box represents the values between the first and third quartiles; the bold black line depicts the middle value of the dataset, i.e., the median, the 'whisker lines' indicate the top quarter and bottom quarter of the data, and the circles represent the outliers. The shaded area represents the parameters off-limit to the rules for favorable oral bioavailability [49]. The analysis revealed that the almost all inhibitors of replicase polyprotein 1ab exhibit MW value 'greater than 500′, nRotB count 'greater than 10′, PSA 'more significant than 120 Å' and HBA count 'more than 10′ (Fig. 4A–C, E) suggesting negligible oral bioavailability. Similarly, the majority of main protease inhibitors display MW value 'greater than 500′, nRotB count 'greater than 10′, and PSA 'greater than 120 Å' (Fig. 4A–C). Whereas, more than ~ 50% inhibitors of helicase have PSA 'greater than 120 Å'. Few outliers were observed in the case of the HBA and HBD distribution for the main protease inhibitors (Fig. 4E and F). In general, it can be concluded that most coronavirus inhibitors so far unearthed lack the general criteria for oral bioavailability and, thus, are not drug-like.

3.3. PCA plot analysis

To assess the molecular diversity of the compounds in "anticorona dataset" we performed principal component analysis (PCA) based on six 2D descriptors, namely, molecular weight (MW), the logarithm of partition coefficient of a compound between n-octanol and water (cLogP), number of hydrogen bond acceptors (HBA), number of hydrogen bond donors (HBD), topological polar surface area (PSA) and number of rotatable bonds (nRotB). The PCs (principal components) were calculated for "anticorona dataset" against "repurposed drugs-COVID-19 pipeline (N = 42)" [11] "natural products and drugs (N = 83)" [12], "repurposed drfugs-Covid-19-clinical trials (N = 19)" [13], and "FDA-approved drugs (N = 2692)" [14] to represent a comparison of

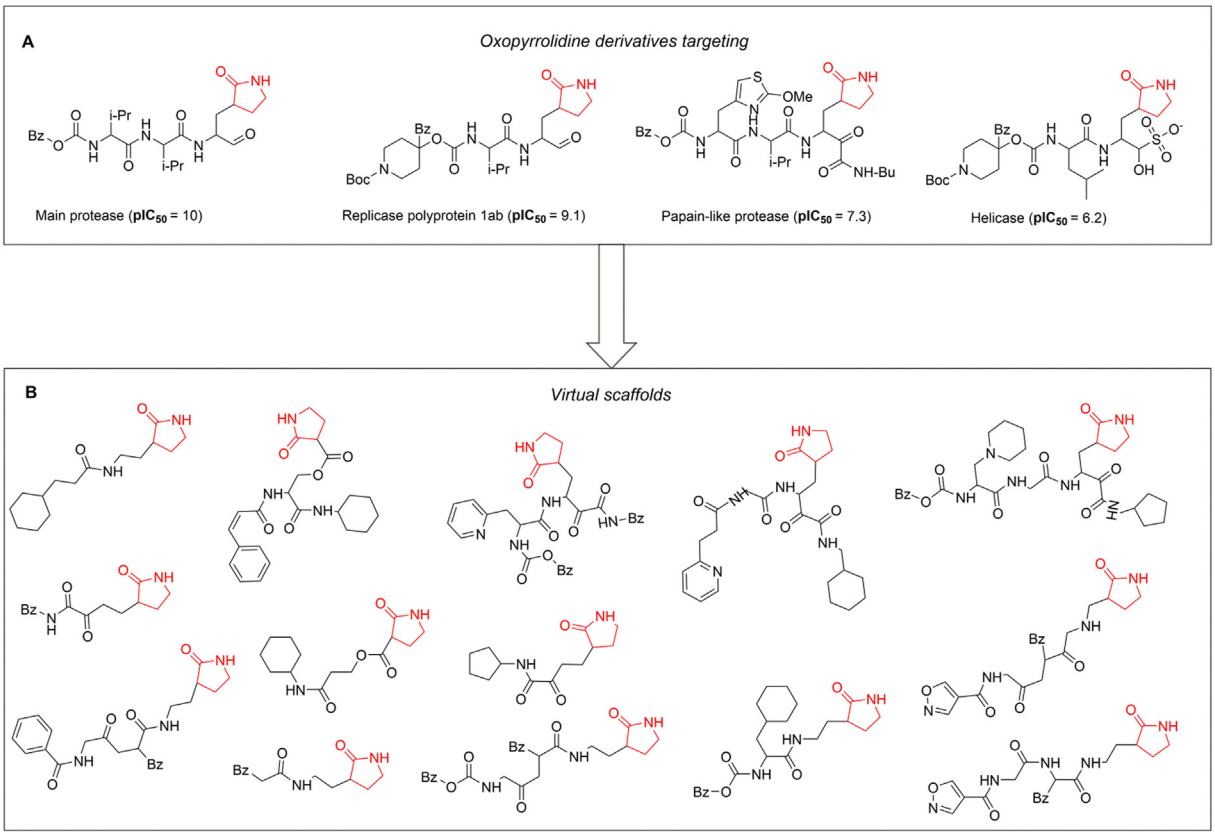

Fig. 3. A. Oxopyrrolidine derivatives targeting SARS-CoV-2 main protease, SARS-CoV-2 replicase polyprotein 1ab, papain-like protease and helicase. B. Representative examples of virtual scaffolds of oxopyrrolidine generated using scaffold hunter.

the property space (Fig. 5). As summarized in Table 2, the first three PCs capture 94.94% of the covariance, implying that the first three PCs are sufficient to define the property space, and thus, it is plausible to represent the property space in the form of a three-dimensional PCA plot. The first PC has the highest loadings by cLogP (0.26) and almost equal loadings by MW, HBA, HBD, PSA, and nRotB. The second PC is primarily contributed by cLogP and secondarily by nRotB and MW. These results indicate that the property space is governed by different molecular descriptors and differs substantially among "anticorona dataset". Fig. 5 shows a three-dimensional illustration of the property space scatterplot of 'anticorona dataset" (yellow spheres), 'repurposed drugs-COVID-19 pipeline' (red spheres), 'natural products and drugs' (blue spheres), 'repurposed drugs-COVID-19-clinical trials' (black spheres) and 'FDA-approved drugs' (green spheres). As observed, many of the 'anticorona dataset (yellow spheres) expand along both the PC1 and PC3 axes, indicating that they vary significantly from the 'approved drugs' (green spheres). Some of 'natural products and drugs' expand majorly along the PC1 axis and PC3 axis, implying that they differ marginally in their property space.

Evaluation of the molecular diversity of molecules targeting coronaviruses by PCA showed that the "anticorona dataset" ligands differ significantly in chemical space from traditional medicinal property space. Therefore, there is an urgent need to modify potent scaffolds selected from the "anticorona dataset" into the drug-like space using fragment-based drug design, computational medicinal chemistry, and scaffold optimization.

3.4. Cluster analysis

Clustering is a powerful resource for medicinal and computational chemists and is extensively utilized to identify chemical scaffolds with similar structural features and further correlate structural properties with biological activity profile [29]. We performed a binning clustering analysis of 433 molecules belonging to "anticorona dataset" for which experimental data were available, to understand the scaffold diversity of the chemical space of the compounds targeting coronaviruses main protease (n = 364), replicase polyprotein 1ab (n = 15), papain-like protease (n = 24), and helicase (n = 20). Clustering was performed using a Tanimoto similarity score of 0.4.

The Tanimoto coefficient is a chemical fingerprint or feature based similarity metrics to measure the chemical similarity between pairs of the molecules [29]. In this study, the similarity is measured between the reference and the database structure. The bin clustering partition grouped the molecules into various similarity groups (Table S4). For main protease inhibitors, the largest bin cluster (n = 275), represented by oxopyrrolidine scaffold, was observed (Fig. 6). Additionally, singleton bin clusters of disulfuram, ebselen, and shikonin scaffolds were also noticed (Fig. 6). For papain-like protease, 12 bin clusters were observed, where naphthalene-based scaffold constitutes the largest bin cluster. Furthermore, a singleton bin cluster comprising nitrophenyl-piperazine, nitropyridine-amine, and ethyl (phenyl) carbamodithioate scaffold was also observed. For helicase, four bin clusters were observed, with the chromenone scaffold representing the largest bin cluster. Additionally, singleton bin clusters, including triazole and benzothiazole scaffold, were detected. For replicase polyprotein 1ab, four bin clusters were observed with an oxopyrrolidine scaffold representing the largest bin cluster. The phenanthro-furan scaffold represented singleton bin clusters. The clusters identified by this analysis can be further explored to design scaffold derivatives by using medicinal chemistry and SAR information. For example, for targeting main protease of

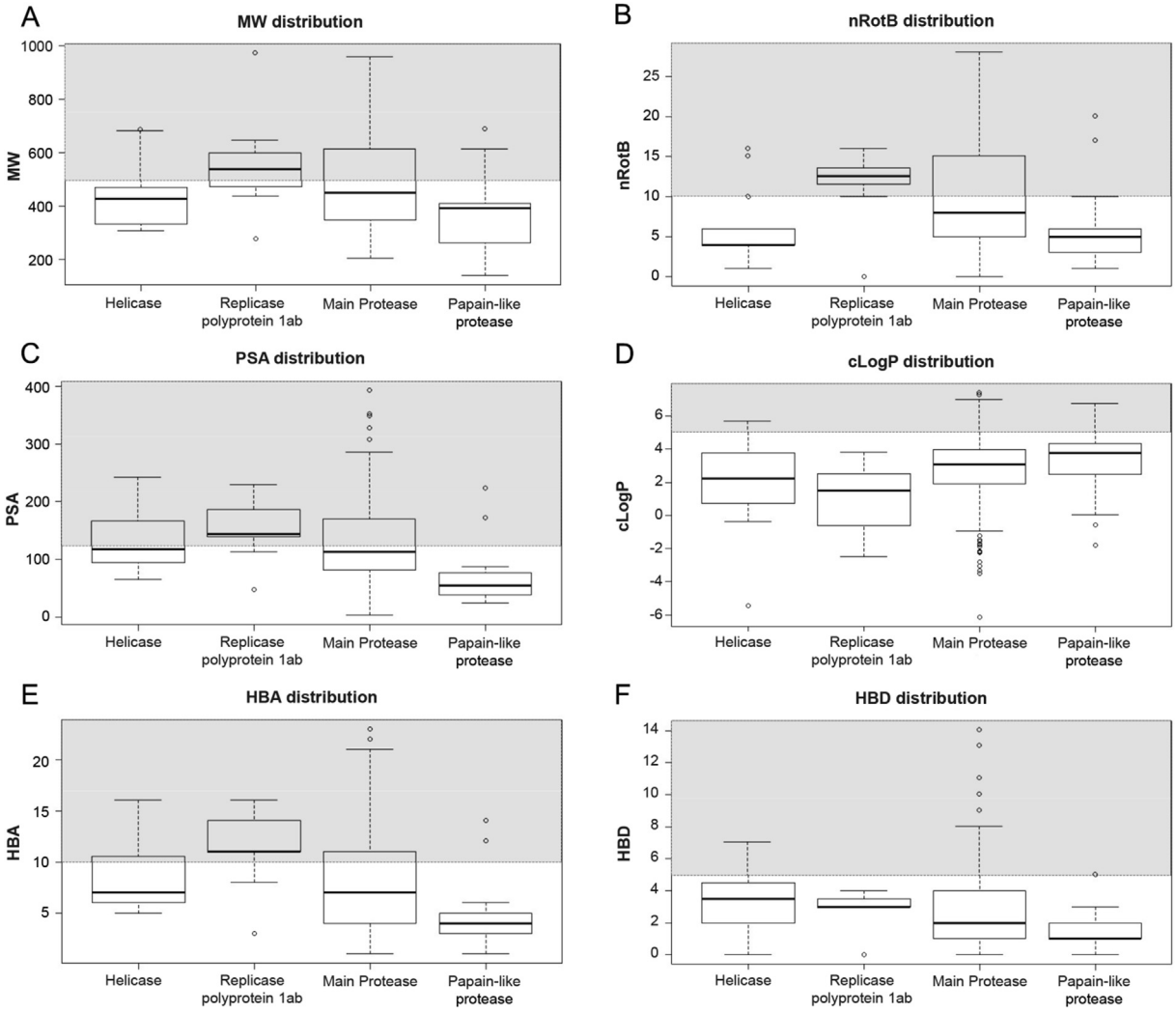

Fig. 4. Distribution of physicochemical properties of major protein targets of coronaviruses: A) MW, B) nRotB, C) PSA, D) cLogP, E) HBA, and F) HBD. The shaded area represents the region of the parameters deviating from the accepted rules of oral bioavailability.

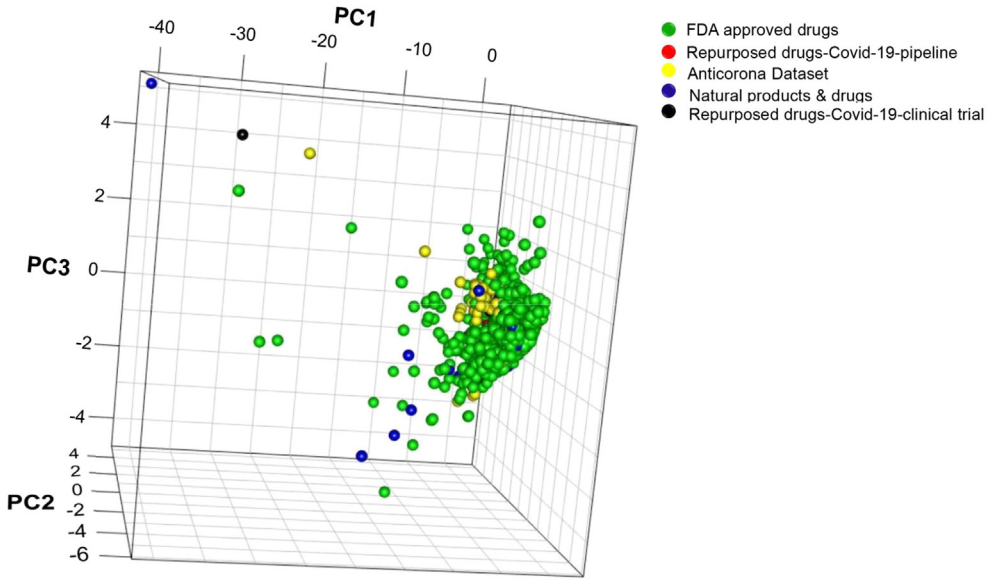

Fig. 5. PCA based three-dimensional property space analysis of coronavirus inhibitors showing spatial representation of six physicochemical properties namely MW, cLogP, HBA, HBD, PSA, and nRotB.

Table 2
Comparative principal component analysis (PCA) of physicochemical properties for COVID inhibitors to FDA approved drugs.

Property	PC1	PC2	PC3	PC4	PC5	PC6
MW	−0.41	0.37	−0.29	0.61	−0.49	−0.08
cLogP	0.26	0.83	−0.32	−0.29	0.24	−0.02
HBA	−0.46	0.00	−0.14	0.14	0.53	0.68
HBD	−0.43	−0.11	−0.33	−0.71	−0.42	0.11
PSA	−0.46	−0.09	−0.16	−0.03	0.49	−0.72
nRotB	−0.39	0.39	0.82	−0.17	−0.06	−0.01
Cumulative proportion%	74.78	90.47	94.94	98.15	99.65	100.00

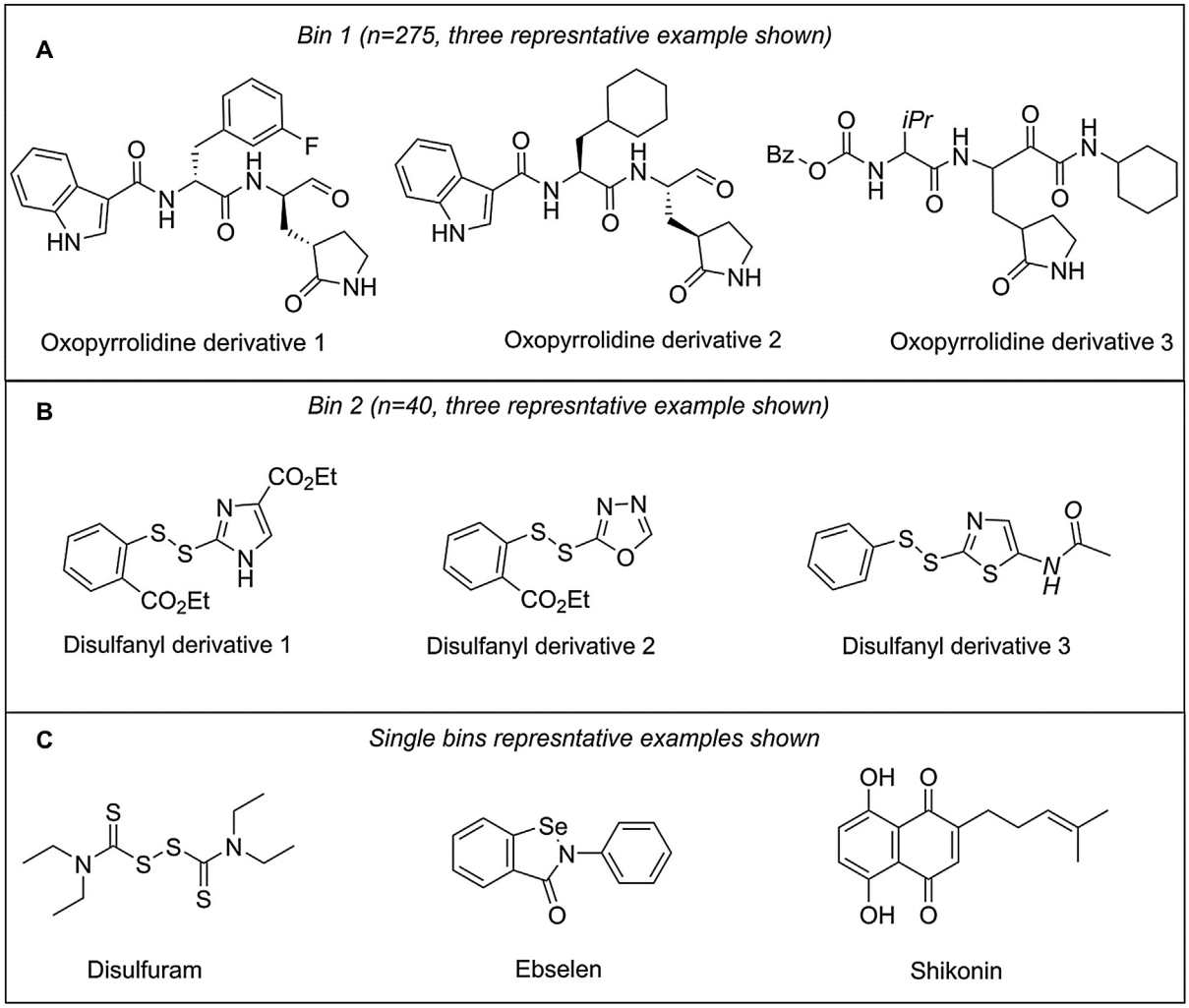

Fig. 6. Binning cluster of main protease inhibitors: (A) a few representative examples of largest bin cluster (n = 275); (B) a few representative examples of second bin cluster (n = 4); (C) examples of singlet bin clusters.

SARS-Cov-2, it will be good idea to design and synthesize molecules, based on the disulfuram, ebselen and shikonin based molecular scaffolds, as the population of these molecules is under-represented in the already know inhibitors of coronaviruses. On the other hand, many different derivatives of oxopyrrolidine are well known for targeting main protease. However, the advantage in terms of designing of novel derivatives of oxopyrrolidine is that oxopyrrolidine scaffold shows activity for other essential targets of coronaviruses family, so it can be exploited as an excellent polypharmacological ligand for coronaviruse family.

Further, we performed the hierarchical clustering of "anti-corona dataset" versus "antiviral dataset" to inspect the common substructures and their anticipated role in the structure–activity relationship. We used ECFP6 (extended-connectivity fingerprints) to cluster the molecules [16,50]. ECFP6 are circular fingerprints that not only determine the substructure and similarity but can be generated quickly. They represent the novel structural classes, including stereochemical information, and define both favorable and unfavorable structural information for molecular activity [51]. The ECFP6 fingerprints were determined by exploiting the CDK nodes for KNIME. To visualize the similarity network map of the related compounds, Gephi was utilized [52] (Fig. 7). Each node represents an inhibitor compound, where the yellow-colored nodes represent a set of "antiviral dataset" compounds. The remaining nodes belong to the "anticorona dataset" with color variations based on different protein targets. For example, ligands

that inhibit the Main protease of coronaviruses are depicted in blue color nodes (Fig. 8). The nodes are sized according to their reported pIC_{50} values, and the width of the edge (lines connecting the nodes) is proportional to the Tanimoto Coefficient of related compounds. Thus, the nodes connected by thicker edges represent the most similar compounds, and vice-versa. The analysis revealed the similarity of the "anticorona dataset" to a total of 245 compounds from the "antiviral dataset" (Supplementary Table S1 and S2). We identified analogs from the antiviral dataset that share similar structures and chemical features to the "anticorona dataset" (Fig. 8, Supplementary Table S1). For example, "antiviral dataset" compound AV18985 showed similarity to "anticorona dataset" compounds 54, 83, and 138, with the Tanimoto coefficient of 0.43, 0.42, and 0.34, respectively. These three compounds (compounds 54, 83, and 138) of the "anticorona dataset" show high inhibitory activity against coronavirus main protease and replicase polyprotein 1ab' proteins. Similarly, AV18965 from the "antiviral dataset" exhibits similarity to "anticorona dataset" compounds 248 and 350, which are active against main protease and spike glycoprotein proteins. Further, we also generated chemical networks to investigate the similarity of inhibitors, individually for each protein target such as, "Main Protease", "Helicase", "Papain-like protease", "Replicase polyprotein 1ab", "Replicase polyprotein 1a", "Spike glycoprotein" and "Non-structural protein 5" (Supplementary Figs. S2–S8). Many of the potential inhibitors (with high pIC_{50} values) of Main protease show significant similarity to already reported antiviral compounds in literature, such as AV18982, AV18952, AV44780, AV21576, AV22574, AV18985, and AV18954. In comparison, most inhibitors for other protein targets possess similarity to antiviral compounds though not that high.

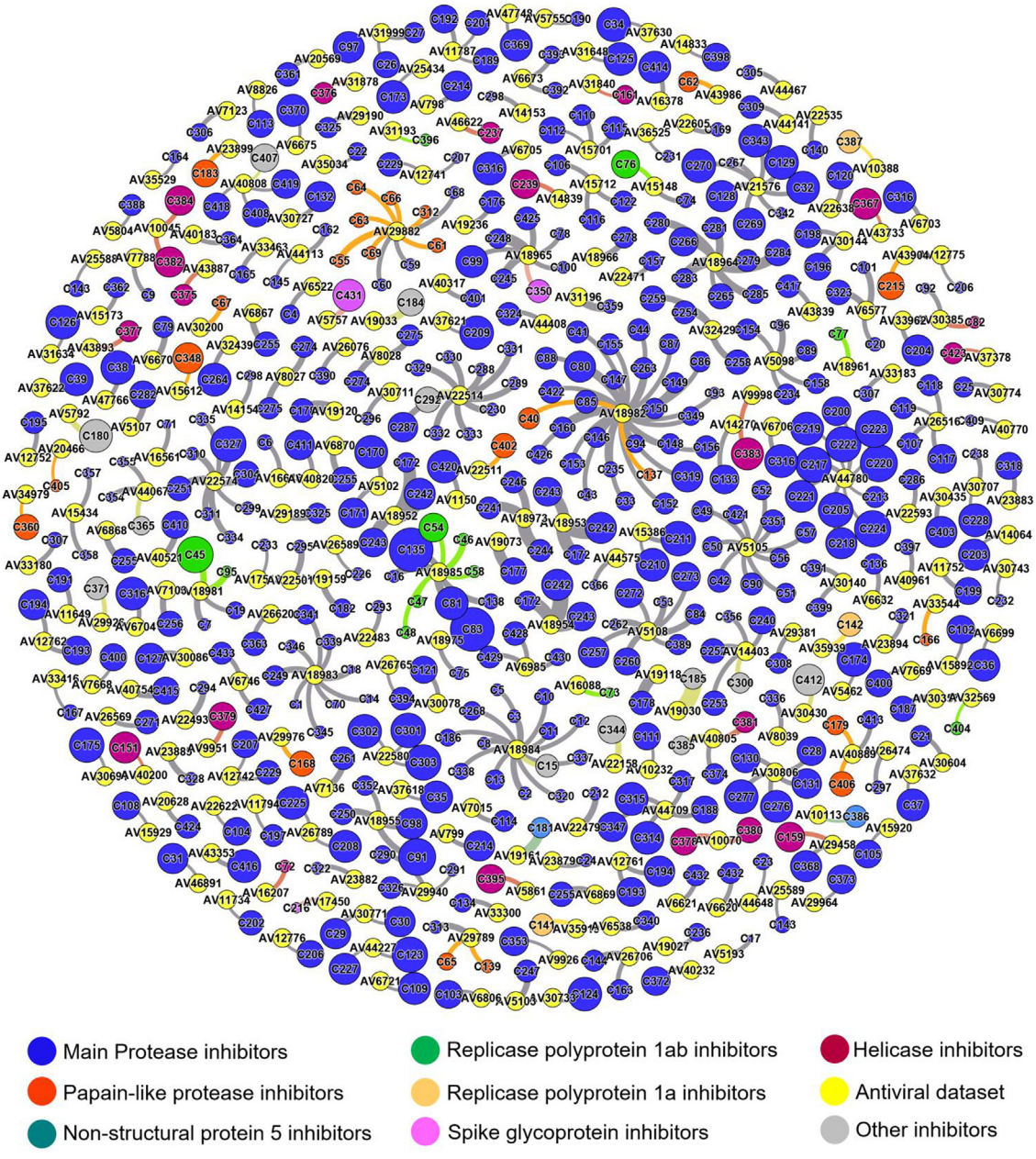

Fig. 7. Chemical network visualization of inhibitors of "anticorona dataset" in conjunction with antiviral compounds. The node color represents the different protein targets of Coronaviruses, and node size (except for yellow nodes) complies with the pIC_{50} values. The edge thickness is in proportion with the Tanimoto coefficient. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

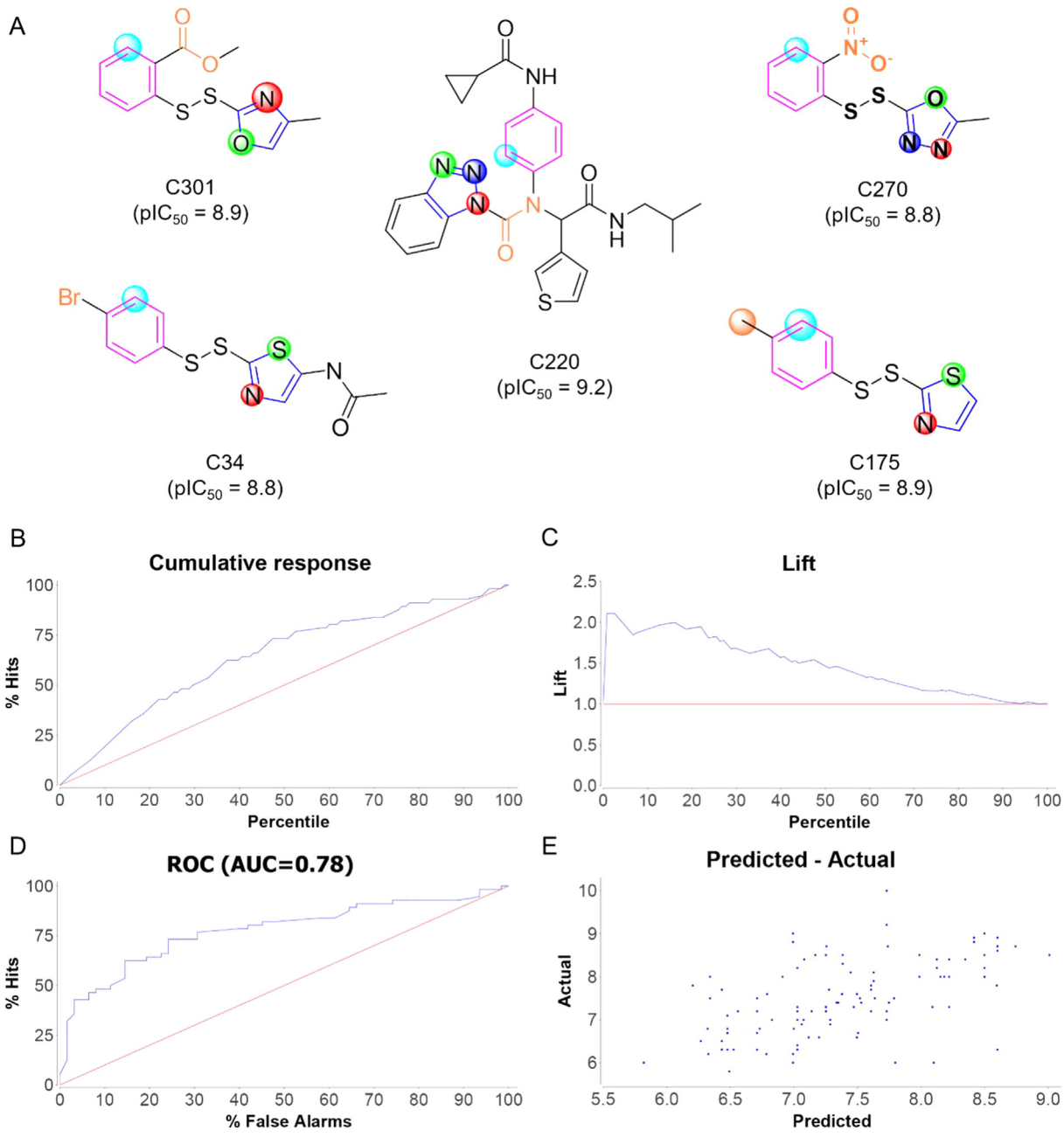

Fig. 8. A) Representative examples of main protease inhibitors with high pIC_{50} values, highlighting the characteristic patterns of hypothesis. The hypothesis states that inhibitors with high pIC_{50} may have a hetero-aromatic-5-ring (shown in blue) that contains two heteroatoms (shown in green and red at a distance 2), an aromatic ring (shown in purple), and a general function group (shown in orange). The aromatic ring (shown in purple) is connected to the general function group (shown in orange) by a single bond. On the aromatic ring (shown in purple) substituent general function group (shown in orange) and unsubstituted atoms (shown in cyan) are at a distance 1. B) Cumulative response plot of the inhibitor model, representing the relationship between the percentage of hits (y-axis) and percentiles (x-axis). C) Lift plot of the inhibitor model, to evaluate the performance of the model. The observation suggests that the top 30% of data will outperform a random model by ~ 2 times (y-axis). D) ROC plot of the inhibitor model clearly distinguishes the non-hits (x-axis) from hits. AUC of 0.78 represents a good measure of separation between hits and non-hits E) Predicted– actual scatter plot of the inhibitor model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

However, some of the prominent examples are AV43733, AV10045, AV9951, AV14839, and AV9998 for Helicase; AV19030, AV15612, AV29882, and AV40889 for Papain-like protease; AV18985 and AV15148 for Replicase polyprotein 1ab. We also identified many other known antiviral compounds that showed significant similarity to "anticorona dataset" compounds but exhibited different chemical features (Supplementary Table S2). Collectively, these antiviral compounds identified from the "antiviral dataset" hint towards their potential use against COVID. The similarity of multiple inhibitors from different protein targets to a single antiviral compound inspires to utilize the multi-targeted drug approach to combat coronaviruse, especially to address the problem of drug resistance due to mutation [53,54].

3.5. Identification of potent chemical features by machine learning

Machine learning is a rising field of artificial intelligence that offers automated learning and enhances prediction quality from various data types. This can be used seamlessly to predict the clinical usefulness of a particular chemical dataset. To determine and

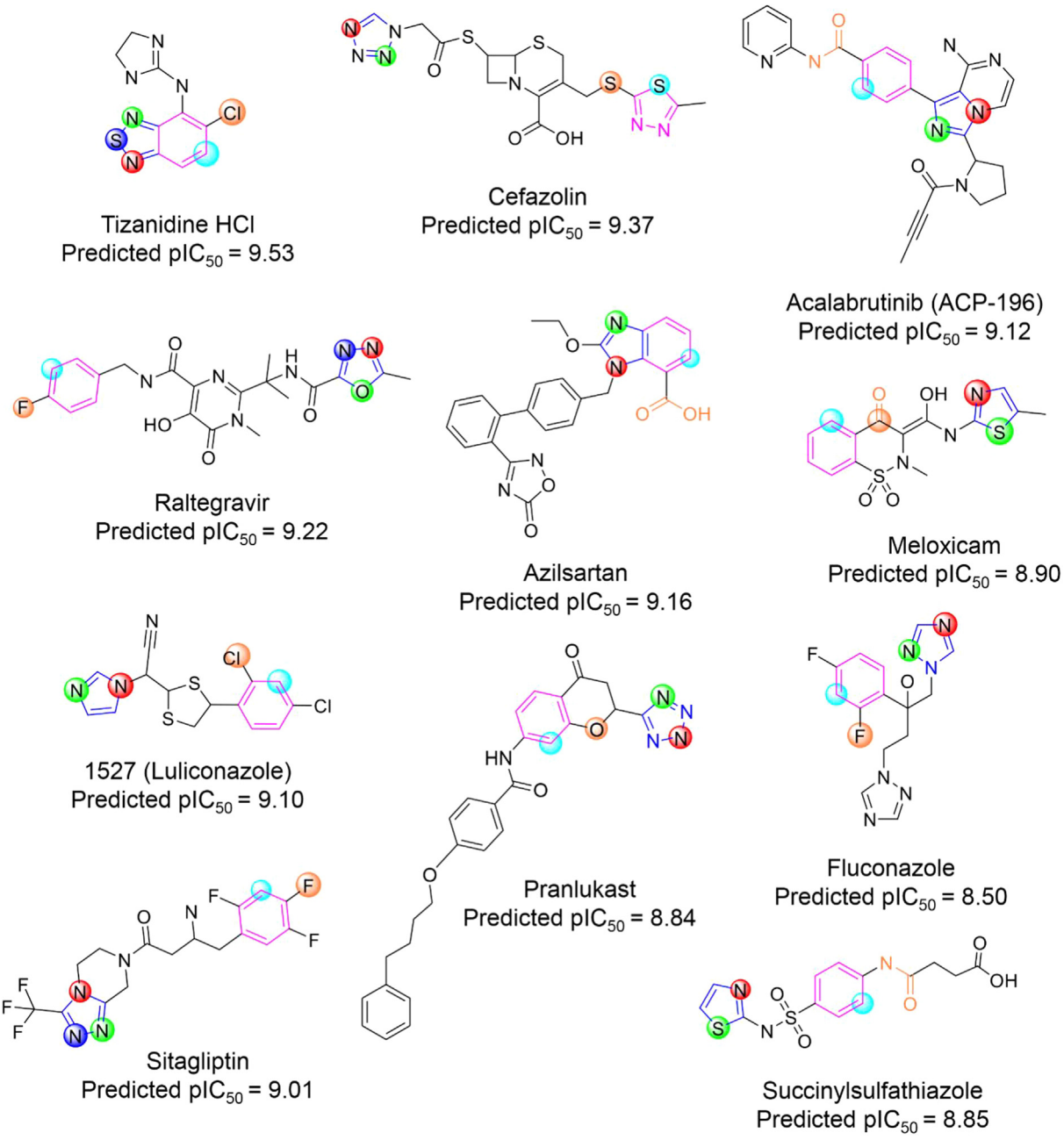

Fig. 9. Selected examples of FDA approved drugs that showed high predicted pIC_{50} values in agreement with the hypothesis.

relate the common chemical patterns of the "anticorona dataset" inhibitors and their bioactivities, we used the machine learning tool DCA (DMax Chemistry Assistant) [30]. DCA is an ILP (Inductive Logic Programming) based approach that uses the existing knowledge such as electron flow, element, moiety, and substructure relationship, to generate the hypotheses that best corroborates with the given data. It starts by reading the functional groups and rings and then constructs the hypotheses to determine the building blocks and their relation to each other. To deduce a significant outcome, we needed a reasonable number of inhibitors with varying high and low activity measurements. Thus, we selected the coronaviruses main protease inhibitors for this investigation. We were successful in generating the inhibitor model hypothesis. The hypothesis suggested that inhibitors with specific patterns containing a benzene ring and a five-membered hetero-aromatic rings such as imidazole, thiazole, triazole, thiadiazole, tetrazole, *etc.* may possess high pIC_{50} (P-value = 7.06×10^{-4}). The representative examples of compounds with high pIC_{50} are shown in Fig. 8A. The generated model was satisfactory with a 'rank high' cut-off of 7.426, and RMSE (Root Mean Square Error) of 0.74. The inhibitor model's cumulative response plot shows the coverage of ~ 75% of hits at the 50th percentile (Fig. 8B), and the lift curve suggests that the top 30% of data will outperform a random model by ~ 2 times (Fig. 8C). The ROC curve (receiver operating characteristic curve) clearly distinguishes the non-hits (x-axis) from hits, and an AUC (Area Under the Curve) value of 0.78 marks a good measure of separation between hits and non-hits (Fig. 8D). The predicted versus actual scatter plot of the inhibitor model is shown in Fig. 8E.

To verify the analysis, we applied the hypothesis to screen FDA approved drugs library [8,55]. The results showed an aromatic ring and a five-membered hetero-aromatic ring, in agreement with the generated hypothesis model for "anticorona dataset". The selected

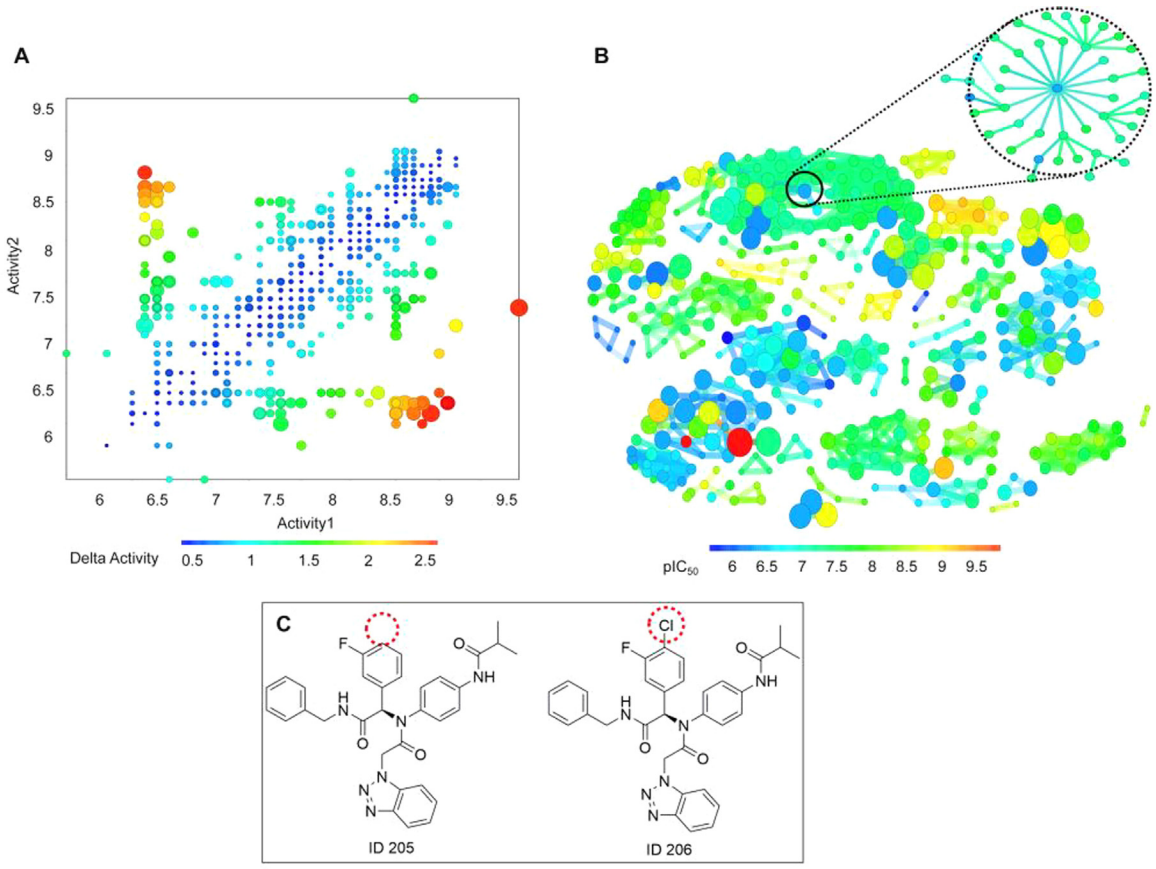

Fig. 10. (A) SALI plot of compound pairs generated from the "anticorona dataset". X- and Y-axis represents activity values; color indicates the delta activity; higher and lower values are indicated by red and blue, respectively. The size of the scatters is indicated by the SALI value. (B) The activity cliff set was grouped based on neighborhood similarity relationships. Colors indicate pIC_{50} value with the higher value represented by red color and lower value represented by blue high. The scatters size suggests a max of SALI pIC_{50} value/SkeleSpheres. (C) An example of a compound pair from the "anticorona dataset" (ID: 205 and ID: 206) represents the 'activity cliff'. The red circle shows the structural variation in the 'activity cliff' pair. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

examples of FDA approved drugs that showed high predicted pIC_{50} values are displayed in Fig. 9. Our results showed Tizanidine HCl, Cefazolin, Raltegravir (MK-0518), Azilsartan (TAK-536), Acalabrutinib (ACP-196), Luliconazole, Sitagliptin, Meloxicam (Mobic), Succinyl sulfathiazole Fluconazole and Pranlukast as some of the exciting candidates for further development of SARS-CoV-2 inhibitors. Interestingly, Raltegravir has been used to prevent viral replication of HIV-1 by inhibiting the Integrase protein [56] and has recently been proposed as a lead candidate to target coronavirus by Khan *et al.* [57]. Similarly, Sitagliptin, which is used traditionally to treat diabetes, has been proposed by another group to reduce the severity of COVID-19 patients [58]. The findings of this analysis are quite intriguing and can be widely applied to repurpose the existing drugs. Considering the hour of need, our machine learning results, together with established literature, recommends the immediate need for a detailed study on these antiviral drugs to understand their mechanisms about protein targets of SARS-CoV-2 and repurpose them for COVID-19 treatment until new drugs or vaccines are developed.

3.6. Structure-activity relationship and activity cliff analysis

Activity cliffs can be evaluated by investigating the biological landscape using similarity metrics that work under the evidence that structurally similar compounds are inclined to have a similar biological response. We generated affinity scatters plot (pIC_{50} value) for each molecule of the "anticorona dataset" (Fig. 10 A). This similarity and activity analysis represent all 957 pairwise comparisons between the 433 compounds of the "anticorona dataset, identified using the cut-off similarity threshold of 86%. The vast amount of "activity cliff" pairs provides vital information on QSAR models and is also useful for designing virtual molecules libraries to be employed for COVID-19 screening.

For the paired comparison between pair ID 205 and ID 206 (Fig. 10C), the determined SALI value is 69.707, the similarity is 0.959, and the activity values are 8.9 and 6.1, respectively. A higher SALI value indicates a significant difference between the biological activities of two structurally similar compounds (Table S5). Fig. 10B represents groups based on neighbouring similarity and their respective SALI values. A representative example of activity cliff in phenyl cinnamamide derivatives is shown in Fig. 11. These results confirm that small changes in structures can induce significant potency variations. The design of new focused libraries for targeting SARS-CoV-2 is possible by incorporating potent moieties identified by "activity cliff" analysis as displayed in Fig. 11.

Using the "anticorona dataset" of 433 compounds with reported pIC_{50} values, SAR based on R-group decomposition [59] was performed, which generated the R group and core fragment using the most central ring system. Different core fragments and six R-groups were generated for the "anticorona dataset". The analysis revealed that compounds with the same core fragment showed different activity (Fig. 12), implying that different R-groups influence biological activity differentially. The analysis revealed that pyrrolidine and the indole core fragment yielded the highest biological activity in the dataset. Pyrrolidine derivatives have been identified to interact with viral main protease protein family [60,61]. While,

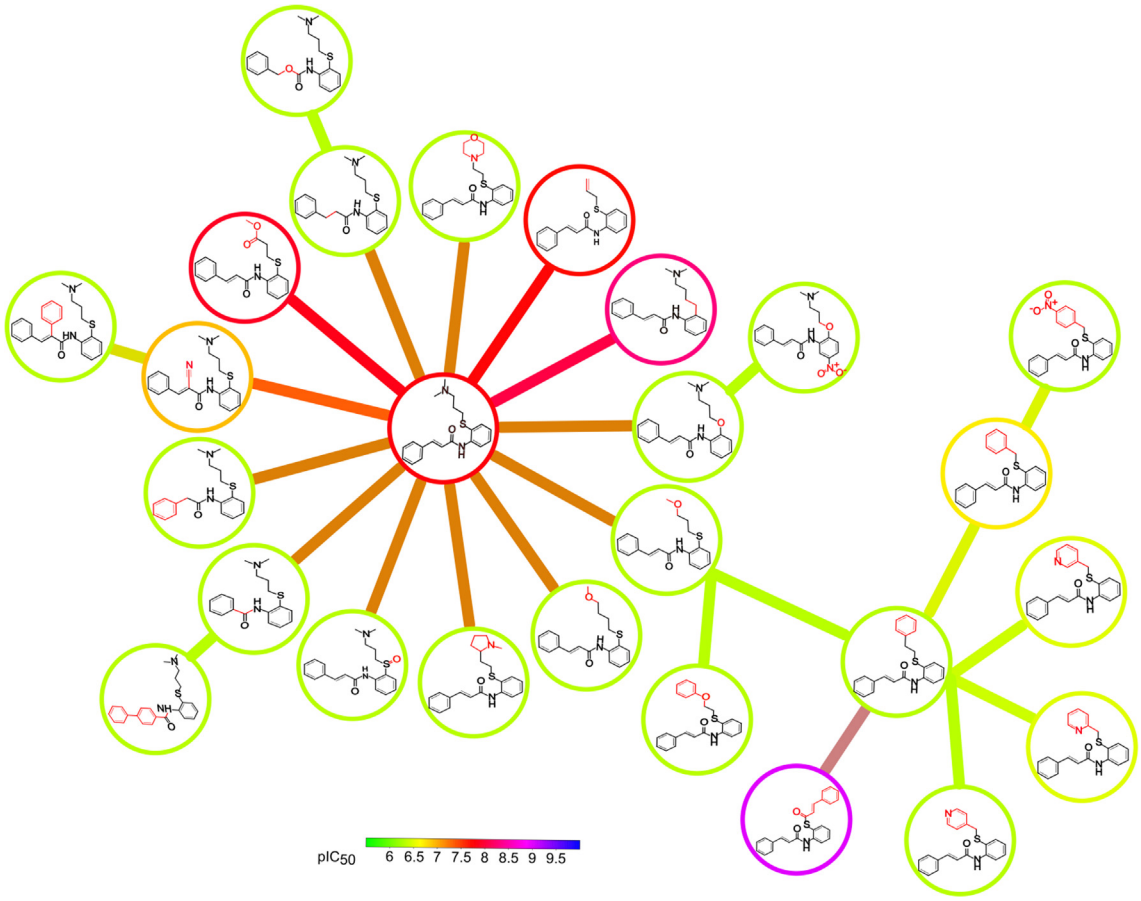

Fig. 11. "Activity cliff" analysis of phenyl cinnamamaide derivatives. The value of pIC_{50} is color-coded with green color denoting lower activity and blue color denoting the highest activity. The structural variation is highlighted in the red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

indole-containing molecules have been widely implicated in antiviral drug research [62].

The results from the SAR study are expected to help design different derivatives with the desired activity against COVID-19. These core fragments are structurally similar to the fragments generated based on scaffold-hoping, as shown in Fig. 2. This corroboration further proposes that pyrrolidine and the indole are the potent fragments for designing and synthesizing the novel drug-like molecules for targeting SARS-CoV-2.

4. Conclusion and perspective

The current coronavirus pandemic has severely impacted the world, causing more than 961 K deaths and 31.1 million coronavirus cases and still increasing. Although researchers are working diligently to find a cure or vaccine for this deadly virus, no successes have been found. Furthermore, because vaccine development might take a long time to enter the market, finding a drug or inhibitor is optimistic and can impede the further spread of the virus. Keeping this in mind, we conducted this study to assess the chemical space using the available knowledge base of closely related coronavirus inhibitors.

This report has clearly defined the molecules' chemical space, which can potentially be used for targeting SARS-CoV-2. We have determined common fragments and generated promising virtual scaffolds which have not been described before and can be further explored for targeting SARS-CoV-2. We also identified oxopyrrolidine based scaffolds that can be used as polypharmacological ligands [63].We have generated dataset of curated ~20,000 antiviral compounds, among which, 245 molecules show structural similarity with "anticorona" compounds, therefore can enrich the "anticorona dataset" chemical space and can further be used for structure or ligand-based drug discovery to target important SARS-CoV-2 targets. Among the "anticorona dataset", indole and pyrrolidine core fragments show the highest biological activity and may be used as a framework to design novel SARS-CoV-2 inhibitors. Indole and pyrrolidine based scaffolds have been extensively used in drug discovery and have resulted in the development of many approved drugs. Additionally indole scaffold is widely used in the design and synthesis of the antiviral inhibitors. A few examples of marketed indole-containing antiviral drugs include Arbidol and Delavirdine. Currently, a number of indole derivatives are actively undergoing different phases of clinical evaluation, such as Atevirdine, GSK2248761 (IDX-12899), Golotimod, Panobinostat (LBH589), BILB 1941, BMS-791325, MK-8742 and Enfuvirtide. Commercial availability of indole and pyrroldine based building blocks and their significant interactions within the active site of the Mpro protein suggest [61] these pyrrolidine-based derivatives as promising candidates for further investigation. Experimental validation of our results will be really useful and we are currently synthesizing the molecules to prove our hypothesis that indole and pyrroldine based chemical scaffolds will be highly active.

Previous studies indicate that COVID viruses are notably diverse and mutate rapidly [64], so it is complicated and challenging for an

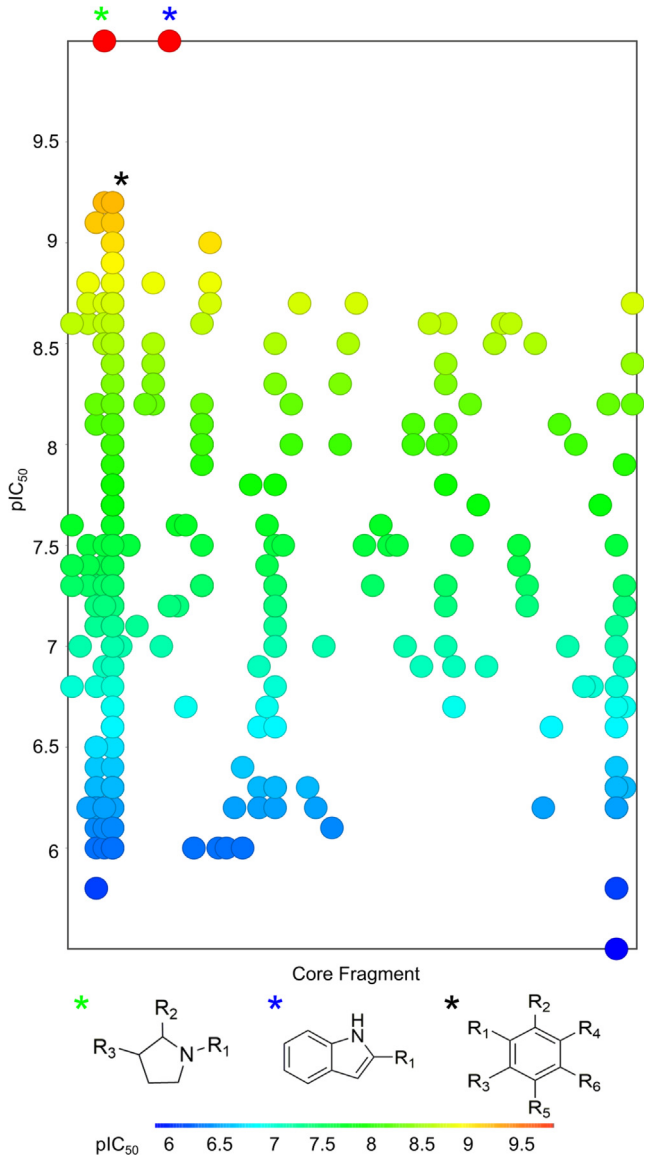

Fig. 12. Scatter plot of Core fragments vs. pIC_{50} generated from "anticorona dataset". The value of the pIC_{50} is color coded, with red and blue colors showing higher and lower pIC_{50} values, respectively. The structure of the core fragment marked by asterisk is indicated at the bottom of the illustration. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

anti-COVID drug to work against the virus. Therefore, in our study, we focused on essential targets of the COVID virus and conducted machine learning and cheminformatics based research to predict the specific scaffold responsible for inhibiting a target. Machine learning resulted in the hypothesis generation of structural patterns and identification of FDA-approved drugs that can be repositioned for COVID-19.

In conclusion, this analysis provides the groundwork for designing diverse chemical libraries, fragment libraries, virtual scaffolds for shape and ligand-based screening, and identification of essential FDA drugs. It thus will be useful for the discovery of small molecule therapeutics for COVID-19.

5. Funding sources

RP thanks CSIR-IMTECH for funds (Project OLP0136).

CRediT authorship contribution statement

Abhinit Kumar: Methodology, Formal analysis, Writing - review & editing. **Saurabh Loharch:** Methodology, Software, Investigation, Data curation, Conceptualization, Formal analysis, Writing - original draft. **Sunil Kumar:** Methodology, Formal analysis, Visualization, Writing - original draft. **Rajesh P. Ringe:** Methodology, Investigation, Formal analysis. **Raman Parkesh:** Conceptualization, Supervision, Funding acquisition, Project administration, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We would like to thank Dr. Mandar Bodas for helpful discussion.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.12.028.

References

[1] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z. Clinical features of patients infected with 2019 novel coronavirus in Wuhan China. Lancet 2020;395(10223):497–506.
[2] Du L, He Y, Zhou Y, Liu S, Zheng BJ, Jiang S. The spike protein of SARS-CoV – A target for vaccine and therapeutic development. Nat Rev Microbiol 2009;7(3):226–36.
[3] Zaki AM, Van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med 2012;367(19):1814–20.
[4] Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020. https://doi.org/10.1056/nejmoa2001017.
[5] Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579(7798):270–3.
[6] Su S, Wong G, Shi W, Liu J, Lai AC, Zhou J, Liu W, Bi Y, Gao GF. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol 2016;24(6):490–502.
[7] Elfiky AA. Ribavirin, Remdesivir, Sofosbuvir, Galidesivir, and Tenofovir against SARS-CoV-2 RNA dependent RNA polymerase (RdRp): A molecular docking study. Life Sci 2020. https://doi.org/10.1016/j.lfs.2020.117592.
[8] https://www.fda.gov/news-events/pressannouncements/coronavirus-covid-19-update-fda-issues-emergency use-authorization-potential-covid-19-treatment. Accessed on April 28, 2020.
[9] Noe MC. The modern drug discovery process. Handbook Med Chem 2014:456–85.
[10] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50(5):742–54.
[11] Huang J, Tao G, Liu J, Cai J, Huang Z, Chen JX. Current prevention of COVID-19: Natural products and herbal medicine. Front Pharmacol 2020;11. https://doi.org/10.3389/fphar.2020.588508.
[12] Singh TU, Parida S, Lingaraju MC, Kesavan M, Kumar D, Singh RK. Drug repurposing approach to fight COVID-19. Pharmacol Rep 2020:1–30.
[13] Rosa, S.G.V., Santos, W.C. Clinical trials on drug repositioning for COVID-19 treatment. Revista Panamericana de Salud Pública 2020, 44, DOI.org/10.26633/RPSP.2020.40
[14] https://www.selleckchem.com/screening/fda-approved-drug-library.html Accessed on December 16, 2020.
[15] Instant JChem was used for structure database management, search, and prediction. Instant J Chem 19.21.5, 2020, ChemAxon.
[16] Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): An open-source Java library for chemo-and bioinformatics. J Chem Inf Comput Sci 2003;43(2):493–500.
[17] Landrum, G. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. http://www.rdkit.org/RDKit_Overview.pdf
[18] Mauri A. alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. Ecotoxicol QSARs 2020:801–20.

[19] Sander T, Freyss J, von Korff M, Rufener C. DataWarrior: an open-source program for chemistry aware data visualization and analysis. J Chem Inf Model 2015;55(2):460–73.
[20] Wetzel S, Klein K, Renner S, Rauh D, Oprea TI, Mutzel P, Waldmann H. Interactive exploration of chemical space with Scaffold Hunter. Nat Chem Biol 2009;5(8):581.
[21] Schuffenhauer A, Varin T. Rule-based classification of chemical structures by scaffold. Mol Inf 2011;30(8):646–64.
[22] Willett, P., Barnard, J. M., Downs, G. M. Chemical similarity searching.Journal of chemical information and computer sciences 1998, 38(6), 983-996
[23] P Mazanetz, M., J Marmon, R., BT Reisser, C., Morao, I. Drug discovery applications for KNIME: an open source data mining platform. Current topics in medicinal chemistry 2012, 12(18), 1965-1979.
[24] Hilbig M, Rarey M. MONA 2: a light cheminformatics platform for interactive compound library processing. J Chem Inf Model 2015;55(10):2071–8.
[25] Nitulescu G, Zanfirescu A, Olaru OT, Nicorescu IM, Nitulescu GM, Margina D. Structural analysis of sortase A inhibitors. Molecules 2016;21:1591.
[26] Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. J Med Chem 1996;39(15):2887–93.
[27] Guha R, Van Drie JH. Structure-activity landscape index: identifying and quantifying activity cliffs. J Chem Inf Model 2008;48(3):646–58.
[28] Bastian, M., Heymann, S., Jacomy, M. March. Gephi: an open-source software for exploring and manipulating networks. In Third international AAAI conference on weblogs and social media 2009.
[29] Backman TW, Cao Y, Girke T. ChemMine tools: an online service for analyzing and clustering small molecules. Nucleic Acids Res 2011;39(2):W486–91.
[30] Ando HY, Dehaspe L, Luyten W, Van Craenenbroeck E, Vandecasteele H, Van Meervelt L. Discovering H-bonding rules in crystals with inductive logic programming. Mol Pharm 2006;3(6):665–74.
[31] Loharch S, Parkesh R. Epigenetic drug discovery: systematic assessment of chemical space. Future Med Chem 2019;11(21):2803–19.
[32] Lorenz, M.O. Methods of measuring the concentration of wealth. Publications of the American statistical association, 1905, 9(70), pp.209-219
[33] Gini C. Variabilità e mutabilità. Vamu 1912.
[34] Nishi A, Shirado H, Rand DG, Christakis NA. Inequality and visibility of wealth in experimental social networks. Nature 2015;526(7573):426–9.
[35] Lee SB, Lee SM, Lee KY. A Gini coefficient based evaluation on the reliability of travel time forecasting. J King Saud Univ-Eng Sci 2019;31(4):314–9.
[36] Cai YM, Chatelet DS, Howlin RP, Wang ZZ, Webb JS. A novel application of Gini coefficient for the quantitative measurement of bacterial aggregation. Sci Rep 2019;9(1):1–12.
[37] Weidlich IE, Filippov IV. Using the Gini coefficient to measure the chemical diversity of small-molecule libraries. J Comput Chem 2016;37(22):2091–7.
[38] Langdon SR, Brown N, Blagg J. Scaffold diversity of exemplified medicinal chemistry space. J Chem Inf Model 2011;51(9):2174–85.
[39] Lipkus AH, Yuan Q, Lucas KA, Funk SA, Bartelt Ii WF, Schenck RJ, Trippe AJ. Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry. J Org Chem 2008;73(12):4443–51.
[40] Zheng Y, Tice CM, Singh SB. The use of spirocyclic scaffolds in drug discovery. Bioorg Med Chem Lett 2014;24(16):3673–82.
[41] Damião MCFCB, Pasqualoto KFM, Polli MC, PariseFilho R. To be drug or prodrug: structure-property exploratory approach regarding oral bioavailability. J Pharm Pharm Sci 2014;17(4):532–40.
[42] 42. Loharch, S., Karmahapatra, V., Gupta, P., Madathil, R. and Parkesh, R. Integrated cheminformatics approaches toward epigenetic drug discovery. In Structural Bioinformatics: Applications in Preclinical Drug Discovery Process 2019, 247-269. Springer, Nature.
[43] Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 2002;45(12):2615–23.
[44] Tanner JA, Watt RM, Chai YB, Lu LY, Lin MC, Peiris JM, Poon LL, Kung HF, Huang JD. The severe acute respiratory syndrome (SARS) coronavirus NTPase/helicase belongs to a distinct class of 5′ to 3′ viral helicases. J Biol Chem 2003;278(41):39578–82.
[45] Prentice E, McAuliffe J, Lu X, Subbarao K, Denison MR. Identification and characterization of severe acute respiratory syndrome coronavirus replicase proteins. J Virol 2004;78(18):9977–86.
[46] Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, Becker S, Rox K, Hilgenfeld R. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. Science 2020;368(6489):409–12.
[47] Jin Z, Du X, Xu Y, Deng Y, Liu M, Zhao Y, Zhang B, Li X, Zhang L, Peng C, Duan Y. Structure of M pro from SARS-CoV-2 and discovery of its inhibitors. Nature 2020:1–5.
[48] Barretto N, Jukneliene D, Ratia K, Chen Z, Mesecar AD, Baker SC. The papain-like protease of severe acute respiratory syndrome coronavirus has deubiquitinating activity. J Virol 2005;79(24):15189–98.
[49] Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. J Pharmacol Toxicol Methods 2000;44(1):235–49.
[50] Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular de-novo design through deep reinforcement learning. J Cheminf 2017;9(1):48.
[51] Rogers D, Hahn M. (2010) Extended-connectivity fingerprints. J Chem Inf Model 2010;50(5):742–54.
[52] Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. Icwsm 2009;2009(8):361–2.
[53] de Castro S, Camarasa MJ. Polypharmacology in HIV inhibition: can a drug with simultaneous action against two relevant targets be an alternative to combination therapy?. Eur J Med Chem 2018;150:206–27.
[54] Chopra, G., Samudrala, R. Exploring Polypharmacology in Drug Discovery and Repurposing Using the CANDO Platform. Curr Pharm Des. 2016, 22, 3109-23
[55] De Wilde, A.H., Jochmans, D., Posthuma, C.C., Zevenhoven-Dobbe, J.C., Van Nieuwkoop, S., Bestebroer, T.M., Van Den Hoogen, B.G., Neyts, J., Snijder, E.J. Screening of an FDA-approved compound library identifies four small-molecule inhibitors of Middle East respiratory syndrome coronavirus replication in cell culture. Antimicrobial agents and chemotherapy 2014, 58 (8), 4875-4884.
[56] Pandey KK. Raltegravir in HIV-1 infection: safety and efficacy in treatment-naive patients. Clin Med Rev Therapeut 2011;2012(4):13.
[57] Khan RJ, Jha RK, Amera GM, Jain M, Singh E, Pathak A, Singh RP, Muthukumaran J, Singh AK. Targeting SARS-CoV-2: a systematic drug repurposing approach to identify promising inhibitors against 3C-like proteinase and 2′-O-ribose methyltransferase. J Biomol Struct Dyn 2020:1–14.
[58] Mozafari N, Azadi S, Mehdi-Alamdarlou S, Ashrafi H, Azadi A. Inflammation: A bridge between diabetes and COVID-19, and possible management with sitagliptin. Med Hypotheses 2020;143.
[59] López-López E, Naveja JJ, Medina-Franco JL. DataWarrior: An evaluation of the open-source drug discovery tool. Expert Opin Drug Discov 2019;14(4):335–41.
[60] Kanhed AM, Patel DV, Teli DM, Patel NR, Chhabria MT, Yadav MR. Identification of potential Mpro inhibitors for the treatment of COVID-19 by using systematic virtual screening approach. Mol Diversity 2020:1–19.
[61] Sacco MD, Ma C, Lagarias P, Gao A, Townsend JA, Meng X, Dube P, Zhang X, Hu Y, Kitamura N, Hurst B. Structure and inhibition of the SARS-CoV-2 main protease reveal strategy for developing dual inhibitors against Mpro and cathepsin L. Sci Adv 2020;6(50):0751.
[62] Naim MJ, Alam O, Alam J, Bano F, Alam P, Shrivastava N. Recent Review on Indole: A Privileged Structure Scaffold. Int J Pharm Sci Res 2016;7:51–62.
[63] Reddy AS, Zhang S. Polypharmacology: drug discovery for the future. Expert Rev Clin Pharmacol 2013;6(1):41–7.
[64] van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CC, Boshier FA, Ortiz AT. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol 2020. https://doi.org/10.1016/j.meegid.2020.104351.