Molecular Informatics

WILEY-VCH

molecular informatics

RESEARCH ARTICLE OPEN ACCESS

The Chemical Space Spanned by Manually Curated Datasets of Natural and Synthetic Compounds with Activities against SARS-CoV-2

Jude Y. Betow[+1,2] | Gemma Turon[+3] | Clovis S. Metuge[1,2] | Simeon Akame[1,4] | Vanessa A. Shu[1,2] | Oyere T. Ebob[5] | Miquel Duran-Frigola[3] | Fidele Ntie-Kang[1,2,6]

[1]Center for Drug Discovery, Faculty of Science, University of Buea, Buea, Cameroon | [2]Department of Chemistry, Faculty of Science, University of Buea, Buea, Cameroon | [3]Ersilia Open Source Initiative, Barcelona, Spain | [4]Department of Clinical Microbiology, Faculty of Health Sciences, University of Buea, Buea, Cameroon | [5]Department of Chemistry and Forensics, School of Science and Technology (SST), Nottingham Trent University, Nottingham, UK | [6]Institute of Pharmacy, Martin-Luther University Halle-Wittenberg, Halle (Saale), Germany

Correspondence: Fidele Ntie-Kang (fidele.ntie-kang@ubuea.cm) | Miquel Duran-Frigola (miquel@ersilia.io)

Received: 1 October 2024 | **Revised:** 28 October 2024 | **Accepted:** 29 October 2024

Funding: Bill & Melinda Gates Foundation, Grant/Award Number: INV-036848; Bill & Melinda Gates Foundation, Grant/Award Number: INV-055897; LifeArc, Grant/Award Number: 10646; Alexander von Humboldt Foundation, Grant/Award Number: 3.4-1156361-CMR-IP

Keywords: chemical space | data curation | SARS-CoV-2

ABSTRACT

Diseases caused by viruses are challenging to contain, as their outbreak and spread could be very sudden, compounded by rapid mutations, making the development of drugs and vaccines a continued endeavour that requires fast discovery and preparedness. Targeting viral infections with small molecules remains one of the treatment options to reduce transmission and the disease burden. A lesson learned from the recent coronavirus disease (COVID-19) is to collect ready-to-screen small molecule libraries in preparation for the next viral outbreak, and potentially find a clinical candidate before it becomes a pandemic. Public availability of diverse compound libraries, well annotated in terms of chemical structures and scaffolds, modes of action, and bioactivities are, therefore, crucial to ensure the participation of academic laboratories in these screening efforts, especially in resource-limited settings where synthesis, testing and computing capacity are scarce. Here, we demonstrate a low-resource approach to populate the chemical space of naturally occurring and synthetic small molecules that have shown *in vitro* and/or *in vivo* activities against the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and its target proteins. We have manually curated two datasets of small molecules (naturally occurring and synthetically derived) by reading and collecting (hand-curating) the published literature. Information from the literature reveals that a majority of the reported SARS-CoV-2 compounds act by inhibiting the main protease, while 25% of the compounds currently have no known target. Scaffold analysis and principal component analysis revealed that the most common scaffolds in the datasets are quite distinct. We then expanded the initially manually curated dataset of over 1200 compounds *via* an ultra-large scale 2D and 3D similarity search, obtaining an expanded collection of over 150 k purchasable compounds. The spanned chemical space significantly extends beyond that of a commercially available coronavirus library of more than 20 k small molecules and constitutes a good starting collection for virtual screening campaigns given its manageable size and proximity to hand-curated compounds.

Jude Y. Betow and Gemma Turon contributed equally.

1 | Introduction

Chemical space is a well-known concept in cheminformatics, often defined simply as the set of molecules in a company's chemical inventory or vendor catalogue, and other times defined as the totality of molecules that can potentially be constructed using known reactions and building blocks within a certain range of properties [1,2]. When all possible molecules that abide by a given set of construction principles are characterised, their chemical space refers to the properties spanned by all these compounds [1–12] . With the rise of the popularity of ultra-large-scale chemical libraries, it becomes crucial to develop efficient ways to encircle small, manageable subsets of molecules with desired properties [3–12]. In their simplest form, chemical spaces are often limited by certain functional groups, chemotypes, or properties that are easy to calculate from a chemical structure alone [2]. "Drug-like chemical space" is used in the context of drug discovery to reflect the vast number of molecules with physical properties similar to those of existing small-molecule therapeutics. These properties are often encapsulated in "rules of thumb" like Lipinski's "rule of five" [3,7] and other well-known rules and metrics adhered to by most approved drugs, e.g. "Ghose rule" [8], "Veber's rule" [9], "Egans' rule" [10] and the quantitative estimate of drug-likeness (QED) metric, among others [11,12] . Even this relatively straightforward set of rules unveils significant complexity. For example, it has been shown that all currently known drugs only occupy a very minute portion of the available and/or explorable "synthetically accessible chemical space" [13–17]. This implies that current molecular libraries cover only a small fraction of the total possible drug-like chemical space if one were to enumerate the compounds resulting from an exhaustive combination of feasible chemical reactions and rules [18,19]. There have been several attempts to estimate the size of the realistic drug-like chemical space [14], including compounds that are directly available to be purchased and screened in biological assays [17]. A widely-cited estimate is that the number of possible Lipinski-compliant (i.e. with MW < 500 Da) molecules surpasses 10^{60}, which is far beyond the chemical space of bioactive compounds reported in literature-curated datasets like ChEMBL ($2{\cdot}10^{6}$) [14,19].

Given the intractable size of the drug-like chemical space, it is necessary to further constrain it with target- or disease-specific properties that will make subsequent (virtual) screening campaigns feasible, especially when resources are limited. Different ways of evaluating the chemical space include using molecular assembly trees, scaffold hopping, similarity search techniques, pharmacophore matching, quantum-based machine learning, and chemography [16,20–23]. To efficiently narrow down the drug-like space, datasets of compounds with desirable properties or bioactivities are frequently used as starting points. With the advent of artificial intelligence (AI), (deep) generative models are getting a lot of attention, since they have the potential to rapidly suggest new chemical matter in a property-constrained manner, often taking a known molecule as a seed. At its core, the approach is based on the idea that similar compounds tend to bind to similar targets, which is a guiding principle in chemoinformatics [24]. First, we get a set of starting molecules, and then we explore the surroundings of this set [25] to identify a much larger collection of molecules that are still within the space and could capture the relevant chemical features required for exhibiting the desired biological activity.

Today, a wide array of computational approaches for exploring chemical spaces exist, with significant improvements to ensure the synthetic feasibility of the compounds [26,27], leveraged by the incorporation of AI techniques to characterise and learn the plausible structures associated with target properties [27]. The huge improvement of computational power available to researchers, including cloud computing [27–33], has made it possible to generate large virtual collections of potentially interesting compounds, the challenge being now how to choose which of them to synthesise and test within a design-make-test cycle [33]. For small laboratories and drug discovery centres operating under strong resource limitations, as is the case of many computational drug discovery groups. Moreso, in the context of Africa where our facilities are located, rapid synthesis of the compounds is a true limiting factor. In this scenario, a more practical approach is to limit the search to the purchasable chemical space, which nowadays is far beyond the billion-scale. This is particularly relevant in the search for antiviral drugs, since diseases caused by viruses spread very quickly and are quite challenging to contain whenever there is a viral outbreak [34,35]. In many resource-limited countries, vaccine accessibility, and acceptability, have remained challenging, implying that the quick discovery of small molecules that target viral infections remains one of the ways forward, with computational methods playing a crucial role in the pipeline [36–38]. Thus, it becomes important to develop diverse and focused compound libraries that could be readily screened to keep pace with possible expected viral outbreaks or mutations.

This work aims to explore the chemical space of potential antiviral agents, beginning with a manually curated dataset of synthetically derived and naturally occurring compounds with activities against known severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) targets or with cell growth inhibitory properties against the virus. More specifically, we have analysed the properties of small molecules that have shown *in vitro* or *in vivo* activities against SARS-CoV-2 and its target proteins, as reported in the literature. We have then explored the chemical space associated with the starting library using a rapid search within the purchasable space of the Enamine REAL [39,40] and ZINC [41] libraries. The expanded dataset was compared with the Coronavirus Library available from ChemDiv [42] and drug molecules from DrugBank [43], to verify if the expanded dataset could be used as a reasonable sized, easy-to-test starting library for virtual screening efforts to target the disease, i.e. a library that could be further easily screened *in silico* via docking and molecular dynamics, followed by *in vitro* screening. With limited conventional computing capacity, the goal would be to look for a manageable dataset that does not require high performance computing, e.g. 100,000–500,000 compounds.

2 | Materials and Methods

2.1 | Data Collection

The hand-curated (natural and synthetic) compound libraries were obtained as follows. The electronic databases employed for the assortment of relevant information include Scopus, NISCAIR, SciFinder, PubMed, Springer Link, Science Direct, Google Scholar, Web of Science, and an exhaustive library search for keywords and combinations of keywords related to "COVID-19", "SARS-CoV-2", "compounds", "small molecules", etc., and a combination of these terms as previously described [44]. Each individual term and a sum of them, e.g. "COVID-19 + compound", "SARS-CoV-2 + compound", "small molecule=-COVID-19", etc were used in the search. This was carried out during the period from January–July 2024. The retrieved articles were checked and compounds showing activities against the virus and/or viral targets were selected. The authors then went ahead and double checked the published papers if there were reported bioactive compounds against SARS-CoV-2 in the retrieved literature sources. The compounds were classified into natural products (NPs) and synthetic derivatives (SDs) according to the information available from the literature sources. The chemical structures were downloaded from the PubChem database, when available [45]. Compounds not available in PubChem were drawn using the ChemDraw Ultra software (version 19.1). Additionally, PubChem and ChemSpider databases were used to check the IUPAC names of the compounds, as previously described [44]. Figure 1A provides a workflow of the manual curation procedure. In summary, in checking through the compounds available in the literature found from the various search engines, if a compound had been tested in clinical trials or had been repurposed for the treatment of COVID-19, it was automatically retained. Compounds that had shown activity in viral assays with $>50\%$ growth inhibition or had shown activity in a target-based assay were also kept. The retrieved articles were checked and compounds showing activities against the virus or viral targets (e.g. M^{pro}, PL^{pro}, Spike/ACE2, RdRp, etc., see Table 1) were selected. The selection criteria were based on the phenotypic and/or target assays (IC_{50}, EC_{50}) reported in the literature, with compounds with IC_{50} or EC_{50} $<50\ \mu M$ retained, while those not falling in this cutoff and those not repurposed for COVID-19 treatment were discarded. The mode of action was either determined from the available experimental *in vitro* assay results against specific viral enzyme targets or through molecular simulations, e.g. by docking and molecular dynamics/binding affinity calculations. Additional information on the modes of action of the compounds were found by searching the COVID-19 HELP [46] and MedChemExpress [47] databases. ChemDiv's Coronavirus Library (containing 21145 small molecule compounds) has been directly retrieved from the ChemDiv website [42], while the DrugBank dataset used was version 5.1.10 [43]. Both were downloaded in July 2024.

2.2 | Principal Components and Scaffold Diversity Analysis of Manually Curated Synthetic and Natural Product Libraries

The molecular descriptors of the compounds in the two datasets were calculated using the Molecular Operating Environment (MOE) software (version 2016.08, 2016) [48]. The computed descriptors included 40 well-known physicochemical parameters like molecular weight (MW), the logarithm of the *n*-octanol/water partition coefficient (log P), the number of Lipinski violations (Lip viol), number of atoms (#atom), synthetic accessibility (SA), the energies of the lowest unoccupied molecular orbitals (LUMO) and of the highest occupied molecular orbitals (HOMO), the number of rotatable bonds, the water solubility, the formal charge (Charge), Oprea lead-likeness score (Oprea Lead), the number of chiral centres (#chiral), the number of basic (#basic) and acidic atoms (#acid), the molar refractivity (mr), the total polar surface area (TPSA), the molecular volume (vol), the dipole moments, the polarizabilities, the number of H-bond donors and acceptors, etc. The dimensionality reduction of the computed descriptors was conducted by principal component analysis (PCA) using MOE [48]. Scaffold analysis was preceded by the Retrosynthetic Combinatorial Analysis Procedure (RECAP) [49] implemented in MOE [48]. This consists in fragmenting each molecule by breaking the bonds that are estimated to be those that can be formed when synthesising each molecule from its constituent building blocks by common synthetic reactions. Thus, a unique extended SMILES string and the fragment's name, which retains the chemical context of the broken bond, was assigned to each resulting fragment, as described by Weininger [50]. This was applied to both the NPs and SDs datasets to determine the most frequent chemical scaffolds and the statistics on the frequency of the individual fragments were generated, while retaining only scaffolds with at least 10 atoms.

TABLE 1 | Distribution of compounds according to SARS-CoV-2 drug targets reported in the literature.

Compound type	No. cpds	M^{pro}[a]	Spike/ACE2[b]	RdRp[c]	PL^{pro}[d]	Dual targets	Multiple targets	Other targets[e]	Unknown target
NPs	620	223	18	2	38	152	12	17	158
SDs	618	295	39	46	5	20	2	57	154

[a] Main protease;
[b] Viral spike protein in complex with the human angiotensin-converting enzyme 2;
[c] RNA-dependent-RNA polymerase;
[d] Papain-like protease;
[e] These include the viral methyltransferase, the S-protein, the N-protein, and those preventing the human cathepsin L and serine protease TMPRSS2 from recognizing and binding with the viral spike protein.

FIGURE 1 | **Overview of the manually curated database of anti-SARS-COV-2 compounds. A**. Methodology for the preparation of NP and SD datasets. **B**. Proportion of unique synthetic derivatives (SD, 622) and natural products (NP, 618) available in the database. **C**. Natural product likeness distribution of SD vs NP (bin size 15). **D**. Distribution of the Synthetic Bayesian Accessibility (SYBA) score of SD vs NP (bin size 15). **E**. Selection of scaffolds over-represented in the SD and NP subsets. **F**. T-SNE representation of the chemical space of the dataset when compounds are described by UniMol and WHALES, respectively. Green dots indicate NP and Purple dots indicate SD. **G**. Top properties contributing to the PCA components 1 and 2 in the whole dataset, calculated with MOE descriptors (HBD: number of H-bond donors, SA: synthetic accessibility, Acid: number of acidic atoms, Basic: number of basic atoms, Lip viol: number of Lipinski violations, Opr lead: Oprea′s lead-likeness score, AM1 LUMO: lowest unoccupied molecular orbital energy computed by the AMI semiempirical method, Charge: total formal charge). **H**. Distribution of several ADMET properties in NP (green) vs SD (purple). Data is represented as percentile of Drugbank, calculated by ADMET-AI. Distributions were obtained with kernel density estimation.

2.3 | Chemical Properties Calculation and ADMET Rediction

To visualize chemical spaces, t-SNE plots were generated using Uni-Mol [51] and WHALES [52] descriptors. Both were calculated using their implementation in the Ersilia Model Hub (https://ersilia.io/model-hub) [53], references eos39co and eos24ur, respectively. The natural product-like score [54–56] was calculated using the RDKit package via its implementation in the Ersilia Model Hub (reference eos8ioa). The synthetic accessibility has been calculated using the SYBA package [57] (reference eos7pw8). The SARS-CoV-2 predicted activities have been calculated using the Ersilia implementation of REDIAL-2020 [58] (reference eos8fth), and ADMET properties have been calculated using the ADMET-AI package [59] (reference eos7d58). We used the openTSNE implementation (Python) with Euclidean distance, perplexity 30 and 500 iterations.

2.4 | Exploration of the Chemical Space of the Manually Curated Dataset by Ultra-Large Library Screening

We used the freely available CHEESE API (https://cheese.deep-medchem.com) to search against the ZINC15 and Enamine REAL databases. For each query compound, we used four similarity search modes, namely "2D fingerprint", "3D shape", "3D electrostatic" and "consensus"; 100 nearest-neighbours using Euclidean distance were retrieved for each search mode with the "high accuracy" option. All molecules were indexed with their InChIKeys and optionally flattened (i.e. stereochemistry removed) to obtain a de-duplicated list. The data aggregation pipeline resulting from this ultra-large-scale search is fully reproducible from the code repository specified in the Code availability section.

2.5 | Compound Prioritisation

To prioritise the compounds obtained from the ultra-large scale similarity search, we developed two criteria. On one hand, we summed the number of occurrences of each retrieved compound across the four search methods (namely Morgan, 3D shape, 3D electrostatics, and consensus), multiplied by the Tanimoto coefficient (T_c) with respect to the query compounds. On the other hand, we developed an ensemble of binary classifiers aimed at scoring the probability of a given compound belonging to the anti-SARS-CoV-2 chemical space. As a reference chemical space, we used our manually annotated compounds, and as "negative" (null) sets we used DrugBank compounds and three subsamples of the ChEMBL database (v33) with a maximum positive-negative imbalance of 1:10. We also trained a classifier using the ChemDiv Coronavirus Library as positive, and a 100 k-scale diversity library from the same vendor as negatives. All classifiers were trained using Ersilia's LazyQSAR [37] framework based on Morgan counts fingerprints (radius 3, 2048 dimensions) and the autoML framework FLAML (random forests and LGBM) [60] with a time budget of 60 seconds. Based on five 80:20 stratified train-test splits, all classifiers satisfactorily performed within the range of 0.75–0.85 AUROC. Finally, since the similarity and the classifier ranks are two genuinely different ranking approaches, we merged them into a consensus rank using the rank averages.

3 | Results and Discussion

3.1 | Literature Review Provides a Comprehensive Curated Anti-SARS-CoV-2 Library

The procedure for gathering literature evidence for the biological activities of the SARS-CoV-2 compounds has been summarised in Figure 1A. It was found that the synthetic compounds belong to quite diverse classes like indoles and peptidomimetics, well-known for their antiviral activities [61–65], as well as antimalarials like chloroquine and its analogues. The naturally occurring compound library was rich in terpenoids, flavonoids, and alkaloids, including the recently discovered hits like salvinorin A and deacetylgedunin which block SARS-CoV-2 viral cell entry by inhibiting the transmembrane protease, serine 2, an enzyme that in humans is encoded by the TMPRSS2 gene [66–68]. Our analysis rendered a final dataset of 618 unique NPs and 620 unique SDs (Figure 1B). After data collection, we sought to understand the characteristics of our dataset. As expected, NPs present a higher natural product-likeness score and, conversely, a lower synthetic accessibility when compared to SDs (Figure 1C and D). Interestingly, a retrosynthetic analysis of the NP library provided 421 scaffolds, revealing that oxygen-containing rings like sugars and polyphenol moieties are the most abundant chemical building blocks in their biosynthesis (Figure 1E). On the other hand, 793 chemical scaffolds resulted from the retrosynthetic analysis of the SD library, revealing a higher diversity in terms of ring types and constituent atoms, with many halogen-, O-, N- and S-containing chains and rings. A comparison of the top-ten most abundant scaffolds in each dataset and not abundant (freq < 3) in DrugBank scaffolds, revealed that the NP fragments contain sugar moieties, polyphenolic rings, and non-oxygenated aliphatics. In contrast, the SD fragments contain heterocyclic rings, aromatic rings, and aliphatic systems, with multiple N-atoms, fewer O-atoms than in the NP fragments, and some S-atoms and halogens (supplementary Figure S1).

To visualise the chemical space of our dataset, we chose two different molecular descriptor techniques. On one hand, Uni-Mol [69] (a deep-learning embedding technique pre-trained on over 209 million molecular conformations) has chemical information in 3D space. On the other hand, WHALES descriptors are a small set of physicochemical parameters that capture both molecular 3D shapes and partial charges, making them suited for scaffold hopping exercises. Figure 1F shows how NPs and SDs cluster together much better when represented with WHALES descriptors, indicating that, despite having dissimilar 3D structures, they may retain similar charge patterns, an essential characteristic to bind to the pockets of their targets in SARS-CoV-2. To further inspect the chemical space of our manually-curated compounds, we used MOE descriptors to build a 2D PCA and analysed the top contributors to defining components 1 and 2 (Figure 1G), which highlighted the descriptors corresponding to Tudor Oprea's test

for lead-likeness (Oprea Lead) [70] and synthetic accessibility (SA) [71], along with the number of basic atoms and the number of H-bond donors, which are all empirical rules that generally characterise drugs and lead compounds. The cumulative variances recovered with the two principal components were 46.72% and 58.33%, respectively. The weights of the descriptors used in PCA analysis have been included in the updated Supplementary Data (Data S1). This means that, within our literature-curated collection, there is wide variability in terms of drug- and lead-likeness. According to the second principal component, the number of acidic and basic atoms, as well as the LUMO features (often associated with chemical reactivities) contribute to the diversity of the dataset.

Finally, we aimed to compare our hand-curated dataset with the chemical space of approved drugs available from DrugBank. To that end, we leveraged a recently published AI/ML model, ADMET-AI [72], which has been trained on reference datasets from the Therapeutics Data Commons [73]. In Figure 1H, we show results from the ADMET-AI predictions as percentiles with respect to approved drugs. Thus, a percentile of 50 means that a given value corresponds to the median value of those observed in the drug space, while extremely high (~100) and low (~0) percentiles indicate deviations from the properties observed in approved drugs. When comparing SD and NP compounds, there was no apparent clear distinction between the two datasets for MW, log P, solubility, inhibition of the cytochrome CYP2 C9, NR-PPAR-γ, and SR-ARE. However, for the descriptors BBB, NR-AR-LBD, and skin toxicity, the NP dataset seems to have a higher proportion of compounds above the 50th percentile, whereas this was the contrary for computed descriptors related to drug absorption (e.g. intestinal absorption and bioavailability), distribution, e.g. ability to cross the blood-brain barrier (BBB), metabolism, e.g. the ability to interact with CYP3 A4 enzymes and toxicity e.g. drug-induced liver injury (DILI), carcinogenesis, and inhibition of the human ether-a-go-go-related gene (hERG). It must be mentioned that the dysfunction of hERG often causes cardiac arrhythmia and sudden death, implying that compounds that block hERG channels are considered toxic. Collectively, and as expected, this indicates that natural product compounds tend to present more liabilities, which is why they are often considered as starting points that require further optimization from an ADMET perspective. Both NPs and SDs are skewed towards relatively high MW and low solubility with respect to approved drugs, and, as expected in compounds not yet progressed to the clinics, there is an enrichment of potential CYP liabilities and toxicity pathways, reinforcing the notion that this set of compounds should be used as a starting collection to identify a larger set of optimised compounds.

3.2 | Distribution of Compounds by Drug Target Based on Literature Information

In addition, we carefully annotated our curated collection with target information, when possible. A summary of the various targets identified in the literature from *in vitro* assays and putative targets predicted by molecular simulations is given in Table 1. It was observed that the main protease (M^{pro}) is the most represented target in the two datasets (36% and 48% for NPs and SDs, respectively). Besides, several compounds have more than one target, including dual protease inhibitors like those that inhibit both M^{pro} and the papain-like protease (PL^{pro}), as well as those that inhibit both M^{pro} and the RNA-dependent-RNA polymerase (RdRp), and those that inhibit both M^{pro} and the viral spike in complex with the human angiotensin-converting enzyme 2 (spike/ACE2) and other protein targets. In both the NP and SD datasets, a small number of the compounds inhibit more than two targets and are classified as multi-target compounds, while a significant number have no known target. This last category corresponds to 25% of both NPs and SDs (supplementary Figure S2).

3.3 | Ultra-Large Library Screening Around the Anti-SARS-CoV-2 Chemical Space

Having defined and characterised the chemical space of manually-curated compounds, we carried out a systematic similarity search against two of the most widely used compound libraries for virtual screening, namely ZINC15 [41] and Enamine REAL [39,40]. ZINC is a compendium of commercially available molecules, and Enamine REAL offers an enumerated billion-scale library of make-on-demand molecules based on a large collection of building blocks. Even the most basic chemoinformatics operations such as similarity search can become prohibitive at such scales, more so in resource-limited settings where computing capacity is low. Thus, we used the online server CHEESE which leverages an embedding-based method to index compounds and speed up the similarity search. The approach capitalises on recent advances in AI embedding techniques initially developed for image and text data, which require fast queries over extremely large databases. In particular, it uses "semantic similarity" search techniques over small molecule embedding vectors, returning the k-neighbors of the seed compound. An advantage of the CHEESE methodology is that it allows performing 3D-based searches, which can be advantageous when the query molecule is IP-protected or difficult to synthesise, as is the case of NP compounds.

We successfully carried out a search of 1231 compounds and obtained, in total, a set of unique 225,774 hits, of which 152,901 remained after flattening out stereochemistry information to remove redundancy. The results of the search correspond to four queries (namely, Morgan (2D) similarity, 3D-shape and 3D-electrostatics, and a consensus measure) against both ZINC15 and Enamine REAL. We retrieved 100 nearest neighbours per search request, obtaining a relatively balanced set of structurally similar compounds, with Tanimoto similarity ($T_c > 0.7$) and more distant ones (Figure 2A). The rankings from the classifier and the similarity search were significantly different and, therefore, we argue that they can be combined in a blended measure that captures both magnitudes. Generally, amongst the top-100 list, ZINC compounds were more abundant than those from Enamine REAL, albeit with more redundancy when the stereochemistry was removed. Enamine REAL is a make-on-demand library based on a predefined set of building blocks and, by definition, it enumerates easily synthesizable compounds. Thus, as expected, natural products were generally less similar to Enamine REAL

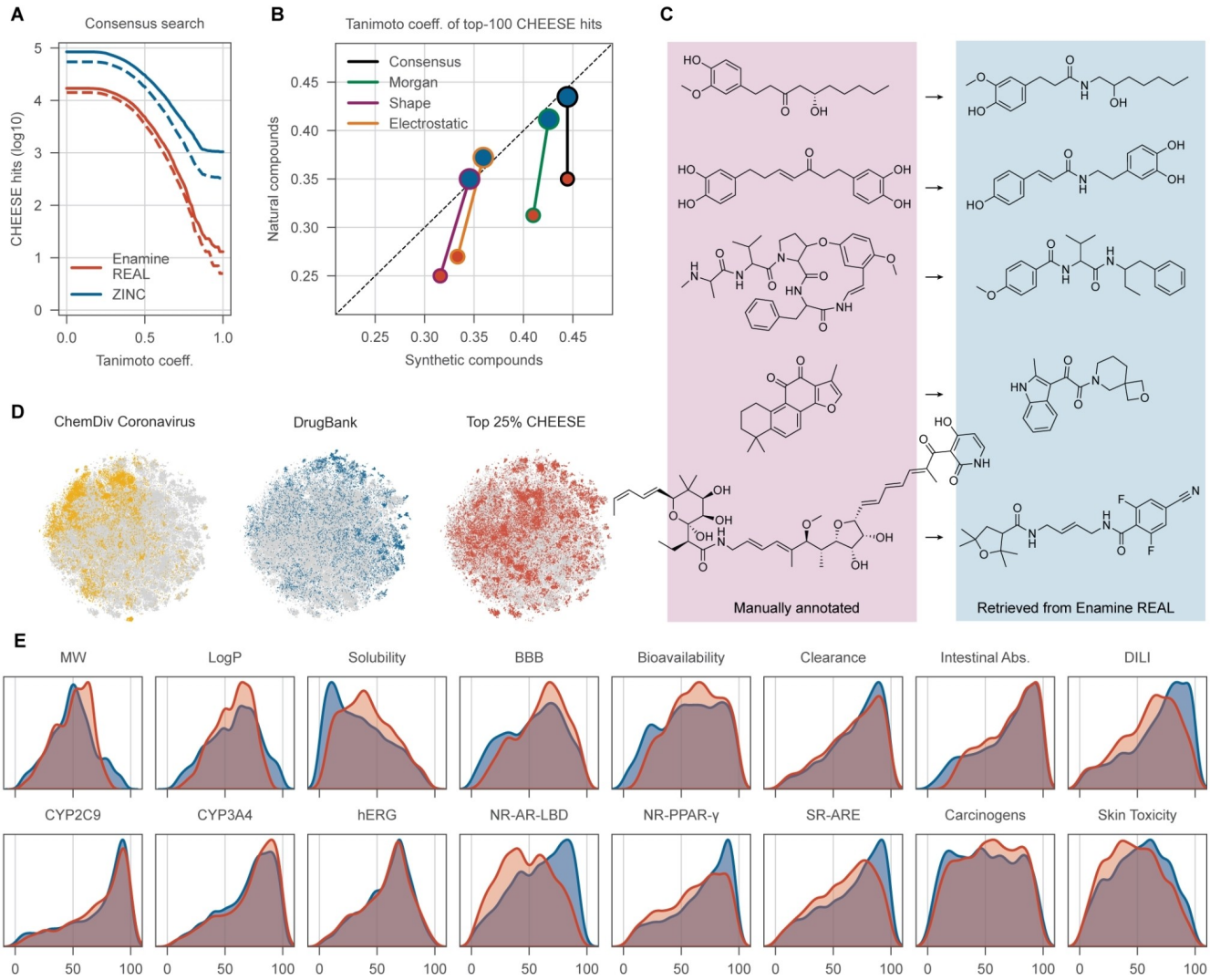

FIGURE 2 | **CHEESE Search of the chemical space around anti SARS-COV-2 compounds. A**. Number of CHEESE hits in the top 100 (nearest neighbours) using the 'consensus' search. In red, we show the number of hits from ZINC and, in blue, the hits from Enamine REAL. The line indicates the number of hits found at a given similarity score and above. Solid lines indicate the raw results from CHEESE, and dashed lines correspond to the flattened results (i.e. stereochemistry removed). **B**. Average Tanimoto similarity of search hits from SD manually-annotated compounds (*x*-axis) and NP compounds (*y*-axis). Size of the dot is proportional to the number of molecules obtained from ZINC (blue) and Enamine REAL (red). Coloured lines denote the type of similarity search. **C**. Example results from the search. On the left, manually annotated NPs are shown. On the right, retrieved compounds Enamine REAL are shown. **D**. 2D projections (t-SNE) plots based on Uni-Mol embeddings, highlighting the ChemDiv Coronavirus Library (yellow), DrugBank molecules (blue) and the top 25% compounds from the virtual library based on the average rank between the similarity score and the classifier score. **E**. Distribution of several ADMET properties in ZINC (blue) vs Enamine REAL (red) compounds. Data is represented as percentile of Drugbank, calculated by ADMET-AI.

compounds than ZINC compounds from a 2D structure perspective (Figure 2B), and hits from 3D-electrostatics and 3D-shape searches tended to give more distal compounds, which may be helpful for scaffold hopping. Generally, the consensus CHEESE score captures structural similarity while providing a slightly better balance between Enamine REAL and ZINC hits than a mere Morgan fingerprints search (Figure 2B).

We then scored the list of compounds based on (a) their similarity to query compounds and (b) their probability of being associated with the SARS-CoV-2 chemical space. These are two simple and indicative measures that can be used to navigate the relatively large collection (>150 k) when screening capacity is limited. To assign a score to the latter, we built an ensemble of binary classifiers capable of discriminating between compounds in our manually curated dataset from randomly sampled compounds in the medicinal chemistry space, as well as between compounds from ChemDiv's Coronavirus Library and a diverse, agnostic collection from the same vendor. As expected, ZINC compounds were ranked higher in the similarity score (Mann-Whitney statistic $2{\cdot}10^9$, P-value~0), while we could still find 6,066 compounds from the Enamine-REAL database that were dissimilar ($T_c < 0.5$) to any compound in the query list but still ranked in the top 20% of the classifier list. In Figure 2C, a few examples are shown

where starting from a natural product compound with high natural product-likeness (>2), it was possible to find make-on-demand hits from Enamine REAL (some of them only retrievable via a 3D search in the CHEESE embedding space) that appear to have a high probability of being interpolated in the chemical space associated with SARS-CoV-2. All the scores are annotated in an easy-to-navigate table as specified in the Data Availability section. When we inspected the ADMET properties of the expanded collection (Figure 2E), we observed that especially for Enamine REAL compounds, properties like MW, logP, solubility, BBB penetration and bioavailability were quite centred or well distributed with respect to approved drugs, and certainly much better than those of NPs (Figure 1H). While, generally, some ADMET liabilities remained (e.g. CYPs), in some cases such as the toxicity pathway NR-AR-LBD the profile was much improved with respect to hand-curated compounds.

Interestingly, when we mapped the abovementioned ChemDiv Coronavirus Library along with DrugBank compounds and our set of >150 k molecules, we observed that the ChemDiv set was focused on a relatively well-defined region (Figure 2E) with respect to our set of compounds and the DrugBank collection. This suggests that our expanded library can be a good starting point for screening purposes against SARS-CoV-2 generally. Since this set has been generated with a ligand-centred approach using a diverse set of mechanisms of action, and including both natural and synthetic compounds, the library is expected to have broad applicability within this field of research.

3.4 | Proposed Libraries Retain SARS-CoV-2 Predicted Activity

The goal of this study is not to provide a short list of anti-SARS-CoV-2 molecules with strong confidence. Rather, we wanted to offer a virtual screening library that can be used as a go-to option in this disease area, ensuring that compounds are purchasable and inspired by compounds with reported evidence in the literature. As an exploratory assessment of the potential of our collection across a broad range of virtual screening tasks related to COVID-19, we chose to use REDIAL-2020, a compendium of open-source machine learning (ML) models containing QSAR predictors for *in vitro* endpoints of viral load reduction. In Figure 3 we can see that, compared to DrugBank compounds (mimicking an unbiased drug repurposing exercise), both our manual collection and the ChemDiv Coronavirus Library tend to perform better in several tasks, most notably in the AlphaLISA screen testing the spike/ACE2 interaction. Differences were observed between NPs and SDs, with NPs having, for example, higher scores in the 3CL and AlphaLISA predictions, and lower in the TruHit counterscreen. Our expanded library was also enriched in high AlphaLISA scores, although, as in the case of the ChemDiv Coronavirus Library and the SD hand-curated compounds, it would be advisable to control for TruHit counterscreen hits. Other enriched predictions are ACE2 blocking and pseudotyped particle entry (PPE) both for SARS-CoV and MERS, suggesting a broad applicability of our collection. We did not obtain particularly high scores in the 3CL predictions, which means that the library is probably not particularly enriched in this class of compounds. However, note that at a classification score above 0.6 (approximately the median for the manually annotated molecules), we still have 26,440 candidates for this activity.

4 | Conclusions

In an attempt to understand the chemical space of potential lead compounds for drug discovery against COVID-19, we have characterised the chemical space of naturally occurring and synthetically derived small molecules that inhibit the growth of the SARS-CoV-2 virus. We have compared the two datasets of compounds hand-curated from the literature by descriptor

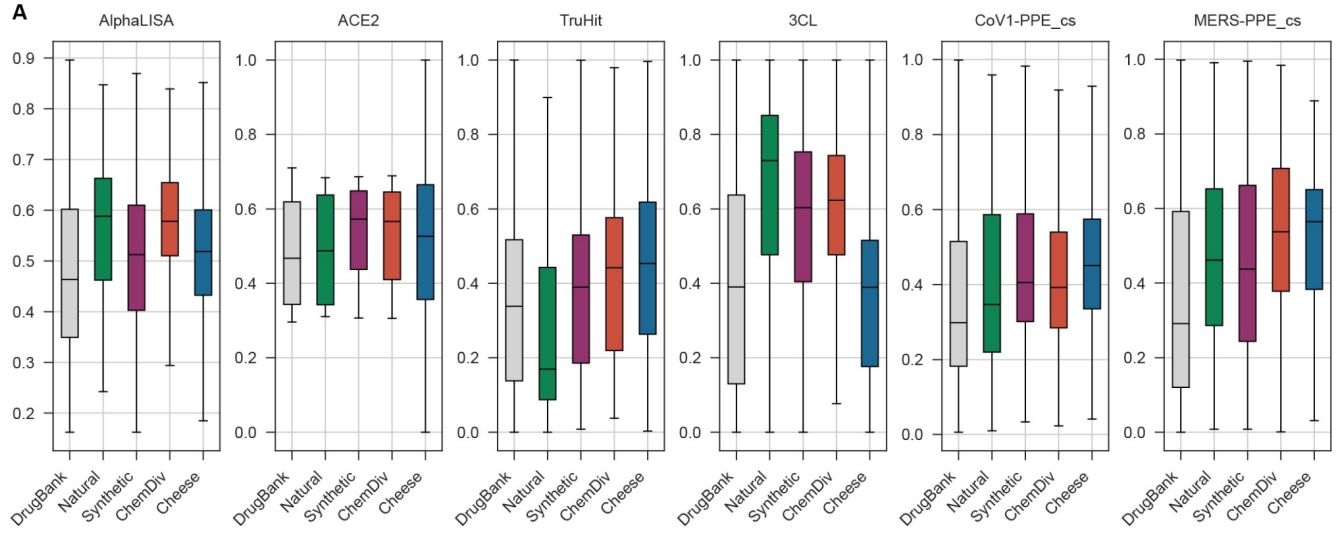

FIGURE 3 | **Predicted activity against SARS-COV-2. A**. Prediction of the activity of the compounds in DrugBank, ChemDiv, CHEESE and the manually curated Natural and Synthetic libraries against several endpoints related to SARS-COV-2 Activity, calculated with REDIAL-2020. Results are presented as probability of 1 (Active).

calculation, principal component analysis and scaffold analysis. It was observed that most of the compounds act by inhibiting the main protease, while several compounds could also be dual and multiple inhibitors. We then derived an expanded chemical space of over 150 k purchasable compounds with either 2D or 3D relatedness to the manually-curated collection. It is planned that, in follow-up studies, these compounds will be virtually screened through pharmacophore modelling and protein-ligand interactions with the view of identifying a small subset of ligands that could putatively bind to the targets reported in the literature. These will then be screened *in vitro* to identify novel antivirals which were not originally reported in the literature.

With make-on-demand libraries growing at an exponential rate, it is important to devise ways to efficiently exploit these libraries and use them to develop custom and smaller virtual collections [74] like our African Natural Products Database (ANPDB) [75,76]. Searching across billion-scale libraries is still computationally intensive and becomes prohibitive in resource-limited settings such as laboratories in Africa, as is our case. Here, we have demonstrated how a well-defined methodology for literature curation, coupled with a simple and fast methodology to search ultra-large chemical spaces, can yield a manageable number of molecules to be used in subsequent virtual screening tasks. We have proved the concept for SARS-CoV-2, a pathogen for which we have invested efforts in our group, but the approach is disease-agnostic and could be applied to any other area for which some compounds are annotated in the literature. Given the infrastructural limitations of chemistry laboratories in our setting, we chose to use purchasable compounds from either ZINC or Enamine REAL databases. The size of the current library (150 k) is amenable for low-resource computing and we expect it to be useful to other researchers pursuing COVID-19 treatments based on small molecules and in a cost-effective manner. Indeed, our overarching plan is to apply this pipeline to other disease areas and targets of interest to our team, including neglected tropical diseases that disproportionately affect people living in the global South.

Code Availability

All code used in the study is available for download at https://github.com/ersilia-os/sars-cov-2-chemspace.

Author Contributions

Conceptualization: Fidele Ntie-Kang and Miquel Duran-Frigola; Data curation: Jude Y. Betow, Clovis S. Metuge, Simeon Akame, Vanessa A. Shu, and Oyere T. Ebob; Formal analysis: Gemma Turon, Fidele Ntie-Kang and Miquel Duran-Frigola; Funding acquisition: Fidele Ntie-Kang; Investigation: Jude Y. Betow, Gemma Turon, Miquel Duran-Frigola and Fidele Ntie-Kang; Methodology: Jude Y. Betow, Gemma Turon, Miquel Duran-Frigola and Fidele Ntie-Kang; Project administration: Jude Y. Betow, Gemma Turon and Fidele Ntie-Kang; Software: Gemma Turon, Miquel Duran-Frigola and Fidele Ntie-Kang; Resources: Gemma Turon, Miquel Duran-Frigola and Fidele Ntie-Kang; Supervision: Fidele Ntie-Kang and Miquel Duran-Frigola; Validation: Fidele Ntie-Kang, Gemma Turon, and Miquel Duran-Frigola; Writing – original draft: Jude Y. Betow, Gemma Turon, Miquel Duran-Frigola and Fidele Ntie-Kang; Writing – review & editing: everyone.

Acknowledgments

We acknowledge financial support from the Bill & Melinda Gates Foundation through the Calestous Juma Science Leadership Fellowship awarded to FNK (grant award number: INV-036848 through the University of Buea). FNK also acknowledges joint funding from the Bill & Melinda Gates Foundation (award number: INV-055897) and LifeArc (Grant ID: 10646) under the African Drug Discovery Accelerator program. FNK acknowledges further funding from the Alexander von Humboldt Foundation for a Research Group Linkage project (Ref [3].4-1156361-CMR-IP). We acknowledge the technical support of Dr. Conrad V. Simoben.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All data used in the study is available for download at https://github.com/ersilia-os/sars-cov-2-chemspace. The library of compounds is referenced in the README file of this repository.

References

1. C. W. Coley, "Defining and Exploring Chemical Spaces", *Trends Chemistry* 3 (2020): 133–145, https://doi.org/10.1016/j.trechm.2020.11.004.

2. J. Wang, J. Mao, M. Wang, X. Le, Y. Wang, "Explore Drug-Like Space with Deep Generative Models", *Methods* 210 (2023): 52–59, https://doi.org/10.1016/j.ymeth.2023.01.004.

3. C. A. Lipinski, "Lead- and Drug-Like Compounds: the Rule-of-Five Revolution", *Drug Discovery Today* 1 (2004): 337–341, https://doi.org/10.1016/j.ddtec.2004.11.007.

4. J. L. Medina-Franco, A. L. Chávez-Hernández, E. López-López, F. I. Saldívar-González, *Molecular Informatics* 41 (2022): e2200116, https://doi.org/10.1002/minf.202200116.

5. P. S. Gromski, A. B. Henson, J. M. Granda, L. Cronin, "How to Explore Chemical Space Using Algorithms and Automation", *Nature Reviews Chemistry* 3 (2019): 119–128, https://doi.org/10.1038/s41570-018-0066-y.

6. C. Dobson, "Chemical Space and Biology", *Nature* 432 (2004): 824–828.

7. C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, "Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings 1PII of Original Article: S0169-409X(96)00423-1", *Advanced Drug Delivery Reviews* 46 (2001): 3–26, https://doi.org/10.1016/S0169-409X(00)00129-0.

8. A. K. Ghose, V. N. Viswanadhan, J. J. A. Wendoloski, "A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases", *Journal of Combinatorial Chemistry* 1 (1999): 55–68, https://doi.org/10.1021/cc9800071.

9. D. F. Veber, S. R. Johnson, H.-Y. Cheng, B. R. Smith, K. W. Ward, F. D. Kopple, "Molecular Properties That Influence the Oral Bioavailability of Drug Candidates", *Journal of Medicinal Chemistry* 45 (2002): 2615–2623, https://doi.org/10.1021/jm020017n.

10. W. J. Egan, K. M. Merz, J. J. Baldwin, "Prediction of Drug Absorption Using Multivariate Statistics", *Journal of Medicinal Chemistry* 43 (2000): 3867–3877, https://doi.org/10.1021/jm000292e.

11. G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, A. L. Hopkins, "Quantifying the Chemical Beauty of Drugs", *Nature Chemistry* 4 (2012): 90–98, https://doi.org/10.1038/nchem.1243.

12. B. Li, Z. Wang, Z. Liu, "DrugMetric: Quantitative Drug-Likeness Scoring Based on Chemical Space Distance", *Briefings in Bioinformatics* 25 (2024): bbae321, https://doi.org/10.1093/bib/bbae321.

13. P. G. Polishchuk, T. I. Madzhidov, A. Varnek, "Estimation of the Size of Drug-Like Chemical Space Based on GDB-17 Data", *Journal of Computer-Aided Molecular Design* 27 (2013): 675–679, https://doi.org/10.1007/s10822-013-9672-4.

14. K. L. Drew, H. Baiman, P. Khwaounjoo, B. Yu, J. Reynisson, "Size Estimation of Chemical Space: How Big is it?" *Journal of Pharmacy and Pharmacology* 64 (2012): 490–495, https://doi.org/10.1111/j.2042-7158.2011.01424.x.

15. G. Maggiora, In: K. Martinez-Mayorga, J. L. Medina-Franco (Eds) *Foodinformatics* (Springer: Cham, 2014) 1–81.

16. T. I. Oprea and J. Gottfries, "Chemography: The Art of Navigating in Chemical Space", *Journal of Combinatorial Chemistry* 3 (2001): 157–166, https://doi.org/10.1021/cc0000388.

17. X. Lucas, B. A. Grüning, S. Bleher, S. Günther, "The Purchasable Chemical Space: A Detailed Picture", *Journal of Chemical Information and Modeling* 55 (2015): 915–924, https://doi.org/10.1021/acs.jcim.5b00116.

18. J.-L. Reymond, R. van Deursen, L. C. Blum, L. Ruddigkeit, "Chemical Space as a Source for New Drugs", *Medicinal Chemistry Communications* 1 (2010): 30–38, https://doi.org/10.1039/C0MD00020E.

19. H. L. Barazorda-Ccahuana, K. E. Juárez-Mercado, J. L. Medina-Franco, M. A. Chavez-Fumagalli, M. A., *Journal of Visualized Experiments* 211 (2024): https://doi.org/10.3791/66349.

20. Y. Liu, C. Mathis, M. D. Bajczyk, S. M. Marshall, L. Wilbraham, L. Cronin, "Exploring and Mapping Chemical Space with Molecular Assembly Trees", *Science Advances* 7 (2021): eabj2465, https://doi.org/10.1126/sciadv.abj2465.

21. C. W. Coley, "Defining and Exploring Chemical Spaces", *Trends Chemistry* 3 (2021): 133–145, https://doi.org/10.1016/j.trechm.2020.11.004.

22. S. Lemonick, "Exploring chemical space: can AI take us where no human has gone before?" *Chemical Engineering News* 98 (2020): 30–35.

23. O. A. von Lilienfeld, K. R. Müller, A. Tkatchenko, "Exploring Chemical Compound Space with Quantum-Based Machine Learning", *Nature Reviews Chemistry* 4 (2020): 347–358, https://doi.org/10.1038/s41570-020-0189-9.

24. R. Pikalyova, Y. Zabolotna, D. Horvath, G. Marcou, A. Varnek, "Chemical Library Space: Definition and DNA-Encoded Library Comparison Study Case", *Journal of Chemical Information and Modeling* 63 (2023): 4042–4055, https://doi.org/10.1021/acs.jcim.

25. P. J. Tonge, "Drug–Target Kinetics in Drug Discovery", *ACS Chemical Neuroscience* 9 (2018): 29–39, https://doi.org/10.1021/acschemneuro.7b00185.

26. S. Stocker, G. Csányi, K. Reuter, J. T. Margraf, "Machine Learning in Chemical Reaction Space", *Nature Communications* 11 (2020): 5505, https://doi.org/10.1038/s41467-020-19267-x.

27. M. Y. McGrady, S. M. Colby, J. R. Nuñez, R. S. Renslow, T. O. Metz, "AI for Chemical Space Gap Filling and Novel Compound Generation", *arXiv* (2024): arXiv:2201.12398v1, https://doi.org/10.48550/arXiv.2201.12398

28. Y. Khalak, G. Tresadern, D. F. Hahn, B. L. de Groot, V. Gapsys, "Chemical Space Exploration with Active Learning and Alchemical Free Energies", *Journal of Chemical Theory and Computation* 18 (2022): 6259–6270, https://doi.org/10.1021/acs.jctc.

29. Y. Yang, K. Yao, M. P. Repasky, "Efficient Exploration of Chemical Space with Docking and Deep Learning", *Journal of Chemical Theory and Computation* 17 (2021): 7106–7119, https://doi.org/10.1021/acs.jctc.1c00810.

30. C. Lu, S. Liu, W. Shi, "Systemic Evolutionary Chemical Space Exploration for Drug Discovery", *Journal of Cheminformatics* 14 (2022): 19, https://doi.org/10.1186/s13321-022-00598-4.

31. M. Šícho, S. Luukkonen, H. W. van den Maagdenberg, L. Schoenmaker, O. J. M. Béquignon, G. J. P. van Westen, "DrugEx: Deep Learning Models and Tools for Exploration of Drug-Like Chemical Space", *Journal of Chemical Information and Modeling* 63 (2023): 3629–3636, https://doi.org/10.1021/acs.jcim.3c00434.

32. K. Edfeldt, A. M. Edwards, O. Engkvist, "A Data Science Roadmap for Open Science Organizations Engaged in Early-Stage Drug Discovery", *Nature Communications* 15 (2024): 5640, https://doi.org/10.1038/s41467-024-49777-x.

33. F. Grisoni, B. J. H. Huisman, A. L. Button, "Combining Generative Artificial Intelligence and On-Chip Synthesis for de Novo Drug Design", *Science Advances* 7 (2021): eabg3338, https://doi.org/10.1126/sciadv.abg3338.

34. A. von Delft, M. D. Hall, A. D. Kwong, "Accelerating Antiviral Drug Discovery: Lessons from COVID-19", *Nature Reviews Drug Discovery* 22 (2023): 585–603, https://doi.org/10.1038/s41573-023-00692-8.

35. L. Riva, S. Yuan, X. Yin, "Discovery of SARS-CoV-2 Antiviral Drugs through Large-Scale Compound Repurposing", *Nature* 586 (2020): 113–119, https://doi.org/10.1038/s41586-020-2577-1.

36. C. T. Namba-Nzanguim, G. Turon, C. V. Simoben, "Artificial Intelligence for Antiviral Drug Discovery in Low Resourced Settings: A Perspective", *Frontiers in Drug Discovery* 2 (2022): 1013285, https://doi.org/10.3389/fddsv.2022.1013285.

37. G. Turon, J. Hlozek, J. G. Woodland, A. Kumar, K. Chibale, M. Duran-Frigola, "First Fully-Automated AI/ML Virtual Screening Cascade Implemented at a Drug Discovery Centre in Africa", *Nature Communications* 14 (2023): 5736, https://doi.org/10.1038/s41467-023-41512-2.

38. G. Turon, M. Njoroge, M. Mulubwa, K. Chibale, M. Duran-Frigola, "AI can Help to Tailor Drugs for Africa — but Africans should Lead the Way", *Nature* 628 (2024): 265–267, https://doi.org/10.1038/d41586-024-01001-y.

39. O. O. Grygorenko, D. S. Radchenko, I. Dziuba, A. Chuprina, K. E. Gubina, Y. S. Moroz, "Generating Multibillion Chemical Space of Readily Accessible Screening Compounds", *iScience* 23 (2020): 101681, https://doi.org/10.1016/j.isci.2020.101681.

40. A. N. Shivanyuk, S. V. Ryabukhin, A. V. Bogolyubsky, D. M. Mykytenko, A. A. Chupryna, W. Heilman, A. N. Kostyuk, A. A. Tolmachev, "Enamine real database: making chemical diversity real", *Chemistry Today* 25 (2007): 58–59.

41. T. Sterling and J. J. Irwin, "ZINC 15 – Ligand Discovery for Everyone", *Journal of Chemical Information and Modeling* 55 (2015): 2324–2337, https://doi.org/10.1021/acs.jcim.5b00559.

42. ChemDiv CORONAVIRUS Library (https://www.chemdiv.com/catalog/focused-and-targeted-libraries/coronavirus-library/).

43. D. S. Wishart, Y. D. Feunang, A. C. Guo, "DrugBank 5.0: A Major Update to the DrugBank Database for 2018", *Nucleic Acids Research* 46 (2018): D1074–D1082, https://doi.org/10.1093/nar/gkx1037.

44. O. T. Ebob, S. B. Babiaka, F. Ntie-Kang, "Natural Products as Potential Lead Compounds for Drug Discovery Against SARS-CoV-2", *Natural Products and Bioprospecting* 11 (2021): 611–628, https://doi.org/10.1007/s13659-021-00317-w.

45. S. Kim, P. A. Thiessen, E. E. Bolton, "PubChem Substance and Compound Databases", *Nucleic Acids Research* 44 (2016): D1202–D1213, https://doi.org/10.1093/nar/gkv951.

46. COVID-19 HELP, https://covid19-help.org/home/.

47. MedChemExpress, https://www.medchemexpress.com.

48. Chemical Computing Group, Molecular, Operating Environment (MOE), version 2016.08 (2016).

49. X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann, "RECAPRetrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry", *Journal of Chemical Information and Computer Sciences* 38 (1998): 511–522, https://doi.org/10.1021/ci970429i.

50. D. Weininger, "SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules", *Journal of Chemical Information and Computer* 28 (1988): 31–36, https://doi.org/10.1021/ci00057a005.

51. G. Zhou, Z. Gao, Q. Ding, H. Zheng, H. Xu, Z. Wei, L. Zhang, G. Ke, "Uni-Mol: A Universal 3D Molecular Representation Learning Framework", *ChemRxiv*. V1 (2023): https://doi.org/10.26434/chemrxiv-2022-jjm0j-v4.

52. F. Grisoni, G. Schneider, "Molecular Scaffold Hopping via Holistic Molecular Representation", *Methods in Molecular Biology* 2266 (2021): 11–35, https://doi.org/10.1007/978-1-0716-1209-5.

53. G. Turon, M. Duran-Frigola, "Ersilia Model Hub: a repository of AI/ML for neglected tropical diseases", Zonodo Software (2021). Published December 15, 2021.

54. P. Ertl, S. Roggo, A. Schuffenhauer, "Natural Product-Likeness Score and Its Application for Prioritization of Compound Libraries", *Journal of Chemical Information and Modeling* 48 (2008): 68–74, https://doi.org/10.1021/ci700286x.

55. K. V. Jayaseelan, P. Moreno, A. Truszkowski, P. Ertl, C. Steinbeck, "Natural Product-Likeness Score Revisited: An Open-Source, Open-Data Implementation", *BMC Bioinformatics* 13 (2012): 106, https://doi.org/10.1186/1471-2105-13-106.

56. RDkit NP-scorer: https://github.com/rdkit/rdkit/blob/master/Contrib/NP Score/npscorer.py.

57. M. Voršilák, M. Kolář, I. Čmelo, D. Svozil, "SYBA: Bayesian Estimation of Synthetic Accessibility of Organic Compounds", *Journal of Cheminformatics* 12 (2020): 35, https://doi.org/10.1186/s13321-020-00439-2.

58. B. K. C. Govinda, G. Bocci, S. Verma, "A Machine Learning Platform to Estimate Anti-SARS-CoV-2 Activities", *Nature Machine Intelligence* 3 (2021): 527–535, https://doi.org/10.1038/s42256-021-00335-w.

59. K. Swanson, P. Walther, J. Leitz, "ADMET-AI: A Machine Learning ADMET Platform for Evaluation of Large-Scale Chemical Libraries", *Bioinformatics* 40 (2024): btae416, https://doi.org/10.1093/bioinformatics/btae416.

60. C. Wang, Q. Wu, M. Weimer, E. Zhu, "FLAML: A Fast and Lightweight AutoML Library," *arXiv* v1 (2021): arXiv:1911.04706v3, https://doi.org/10.48550/arXiv.1911.04706.

61. M. Z. Zhang, Q. Chen, G. F. Yang, "A Review on Recent Developments of Indole-Containing Antiviral Agents", *European Journal of Medicinal Chemistry* 89 (2015): 421–441, https://doi.org/10.1016/j.ejmech.2014.10.065.

62. U. L. Urmi, S. Attard, A. K. Vijay, "Antiviral Activity of Anthranilamide Peptidomimetics against Herpes Simplex Virus 1 and a Coronavirus", *Antibiotics* 12 (2023): 1436, https://doi.org/10.3390/antibiotics12091436.

63. D. Ding, S. Xu, E. F. da Silva-Júnior, X. Liu, P. Zhan, "Medicinal Chemistry Insights into Antiviral Peptidomimetics", *Drug Discovery Today* 28 (2023): 103468, https://doi.org/10.1016/j.drudis.2022.103468.

64. M. S. Mousavi Maleki, M. Rostamian, H. Madanchi, "Antimicrobial Peptides and other Peptide-Like Therapeutics as Promising Candidates to Combat SARS-CoV-2", *Expert Review of Anti Infective Therapy* 19 (2021): 1205–1217, https://doi.org/10.1080/14787210.2021.1912593.

65. Y.-S. H. Mahmoud, Y. A. M. M. Elshaier, N. M. A. Shama, "Antiviral Activities of Plant-Derived Indole and β-Carboline Alkaloids against Human and Avian Influenza Viruses", *Scientific Reports* 13 (2023): 1612, https://doi.org/10.1038/s41598-023-27954-0.

66. A. Khursheed, V. Jain, A. Rasool, M. A. Rather, N. A. Malik, A. H. Shalla, "Molecular Scaffolds from Mother Nature as Possible Lead Compounds in Drug Design and Discovery against Coronaviruses: A Landscape Analysis of Published Literature and Molecular Docking Studies", *Microbial Pathogenesis* 157 (2021): 104933, https://doi.org/10.1016/j.micpath.2021.104933.

67. M. Omrani, M. Keshavarz, S. Nejad Ebrahimi, "Potential Natural Products Against Respiratory Viruses: A Perspective to Develop Anti-COVID-19 Medicines", *Frontiers in Pharmacology* 11 (2021): 586993, https://doi.org/10.3389/fphar.2020.586993.

68. J. Shawon, Z. Akter, M. M. Hossen, "Current Landscape of Natural Products against Coronaviruses: Perspectives in COVID-19 Treatment and Anti-viral Mechanism", *Current Pharmaceutical Design* 26 (2020): 5241–5260, https://doi.org/10.2174/1381612826666201106093912.

69. G. Zhou, Z. Gao, Q. Ding, "Uni-Mol: A Universal 3D Molecular Representation Learning Framework," https://github.com/deepmodeling/Uni-Mol.

70. T. I. Oprea, *Journal of Computer-Aided Molecular Design* 16 (2002): 325–334, https://doi.org/10.1023/A:1020877402759.

71. P. Ertl and A. Schuffenhauer, "Estimation of Synthetic Accessibility Score of Drug-Like Molecules Based on Molecular Complexity and Fragment Contributions", *Journal of Cheminformatics* 1 (2009): 8, https://doi.org/10.1186/1758-2946-1-8.

72. K. Swanson, P. Walther, J. Leitz, S. Mukherjee, J. C. Wu, R. V. Shivnaraine, J. Zou, "ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries." *Bioinformatics* 40 (2024)), btae416, https://doi.org/10.1093/bioinformatics/btae416.

73. K. Huang, T. Fu, W. Gao, "Artificial Intelligence Foundation for Therapeutic Science", *Nature Chemical Biology* 18 (2022): 1033–1036, https://doi.org/10.1038/s41589-022-01131-2.

74. M. Duran-Frigola, M. Berton, R. Blanco, et al, "Bioactivity Profile Similarities to Expand the Repertoire of COVID-19 Drugs", *Journal of Chemical Information and Modeling* 60 (2020): 5730–5734, https://doi.org/10.1021/acs.jcim.0c00420.

75. F. Ntie-Kang, K. K. Telukunta, K. Döring, "NANPDB: A Resource for Natural Products from Northern African Sources", *Journal of Natural Products* 80 (2017): 2067–2076, https://doi.org/10.1021/acs.jnatprod.7b00283.

76. C. V. Simoben, A. Qaseem, A. F. A. Moumbock, "Pharmacoinformatic Investigation of Medicinal Plants from East Africa", *Molecular Informatics* 39 (2020): e2000163, https://doi.org/10.1002/minf.202000163.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.