Suggestions for second-pass anti-COVID-19 drugs based on the Artificial Intelligence measures of molecular similarity, shape and pharmacophore distribution.

Martyna Moskal[1+], Wiktor Beker[1+], Rafał Roszak[1], Ewa P. Gajewska[2], Agnieszka Wołos[1], Karol Molga[1], Sara Szymkuć[1], & Bartosz A. Grzybowski[1,2,3,4]*

[1] Allchemy, Inc., 2145 45th Street, Highland, IN 46322, USA

[2] Institute of Organic Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, Warsaw 02-224, Poland

[3] IBS Center for Soft and Living Matter and

[4] Department of Chemistry, UNIST, 50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan, 689-798, South Korea

[+]Authors contributed equally

*Correspondence to: nanogrzybowski@gmail.com

Abstract

Artificial Intelligence algorithms are used to identify "progeny" drugs that are similar to the "parents" already being tested against COVID-19. These algorithms assess similarity not only by the molecular make-up of the molecules, but also by the "context" in which specific functional groups are arranged and/or by three-dimensional distribution of pharmacophores. The parent-progeny relationships span same-indication drugs (mostly antivirals) as well as those in which the "progenies" have different and perhaps less intuitive primary indications (e.g., immunosuppressant or anti-cancer progenies from antiviral parents). The "progenies" are either already approved drugs or medications in advanced clinical trials – should the currently tested "parent" medicines fail in clinical trials, these "progenies" could be, therefore, re-purposed against the COVID-19 on the timescales relevant to the current pandemic.

Introduction

As the COVID-19 pandemic unfolds at an alarming rate, scientists and clinicians are desperately looking for effective countermeasures. With vaccines estimated to become available in no less than a year, and with the development of brand-new medications requiring even longer times, the focus of attention has been on the already approved drugs (or those in advanced clinical trials) that could be re-purposed against COVID-19 within much shorter time scales. In this respect, first WHO trials [1] are already ongoing and cover a range of potential therapies, from various antivirals, to antimalarial chloroquine/hydroxychloroquine, to interferon-beta, to antibody-rich plasma from COVID-19 recoverees. A literature search we conducted shows that, in total, 42 small-molecule medications are in various forms of clinical trials and some 29 other ones have been shown to exhibit some *in vitro* activity (for all literature sources, see Supplementary Information, **Table S1**). While much hope – and even popular hype – has been pinned at these studies, it should be emphasized that their outcome remains uncertain and, in the worst-case scenario, none of the currently tested drugs will prove effective. If this is so, the last-ditch effort would be to consider repurposing of other drugs – the question is, of course, which ones. Here, we use the methods of Artificial Intelligence, AI, to suggest such second-pass (or, "progeny") candidates based on their similarity to the ones already being considered (henceforth, "parent" drugs). We make such suggestions based on two measures of molecular similarity that are more advanced than the popular but inaccurate [2] metrics such as fingerprint-based Tanimoto coefficients: (1) the linguistics-inspired Mol2Vec embedding [3] and (2) the so-called Estimated Shape Representation (ESR) introduced here for rapid comparison of molecules based on their 3D shapes and spatial distribution of pharmacophoric features. These methods suggest multiple progeny drugs for several parent compounds, including cases in which one progeny is predicted (i) by both similarity methods for the same parent; or (ii) by one or both methods for multiple parents. While many progenies share the same primary indication as the parent (e.g., an antiviral progeny similar to an antiviral parent), many have different and perhaps less obvious indications (e.g., immunosuppressant progeny similar to an antiviral parent). We hope that at least some of these suggestions will prove useful and will merit additional *in silico* analyses (e.g., docking against COVID-specific targets [4,5]), *in vitro* screens, and careful scrutiny by clinicians.

Methods

Model choice and description. The crux of the AI approaches we use is to translate molecular structures into a high-dimensional vector space, in which similarity between compounds is reflected by the proximity of the corresponding points/vectors. We will refer to this data transformation as either "vectorization" or "embedding". In particular, we employ distance metrics based on two vectorization techniques: (1) the linguistics-inspired Mol2Vec embedding [3] and (2) Estimated Shape Representation (ESR) developed here as an extension to the model originally created by Skalic and co-workers [6].

In the Mol2Vec approach [3], the key concept is to view the molecule as a "sentence" comprised of "words" corresponding to substructures of predefined size (**Figure 1a**). Unlike in fingerprint approaches, these "words" are not assigned with arbitrary numbers but, instead, each is represented as a 300-dimensional vector reflecting colocations with other "words". Statistical information describing the so-called "corpus" [7] and reflecting colocations between large numbers of chemical "words" is derived from some comprehensive collection of organic molecules, here, ~19 million molecules from the Zinc database [8]. A molecule "sentence" is then a union of vectors describing its constituent "words". Effectively, Mol2Vec not only recognizes the counts of specific fragments in molecules, but also their mutual molecular "contexts".

Regarding the ESR approach, we began by training Skalic's Variational Autoencoder (VAE) [6] to generate SMILES strings resembling 3D shape and pharmacophoric features of a given seed molecule [9] (**Figure 1b**). In this method, a Convolutional Neural Network (CNN) takes as input a three dimensional structure of a molecule (optimized by one of the MM force-fields, in Skalic, MMFF) and encodes 3D distributions of selected properties (e.g., aromatic rings, H-bond acceptors and donors, spatial distribution of heavy atoms) into a 512-dimensional vector (so-called hidden representation). This vector is then perturbed with Gaussian noise to produce a set of similar vectors. Next, the Long-Short Term Memory (LSTM) module generates SMILES matching each of these perturbed representations, in effect generating additional yet similar molecules (in **Figure 1b**, examples shown on the *right*) having some shape and pharmacophoric similarity with the seed molecule. Importantly, Skalic showed that the similarity between these LSTM-output molecules and the input seed molecule is not as strongly dependent on seed's conformation as for the 512-dimensional hidden representation.

With the goal of minimizing any such conformational dependence, our idea was to design a function inverse to the SMILES generator – that is, one that would estimate encoded 3D information from 2D

molecular representation. To do so, we used a teacher-student training model originally devised for model compression [10,11]. This training scheme involved teaching a "student" model (here, a multiple layer perceptron, MLP, in **Figure 1c**) to mimic the behavior of a "teacher" (here, LSTM module from **Figure 1b**) based on the latter's input and output data. In this way, the relationship between 3D features and molecular topology learned by Skalic's model could be – to some extent – transferred to a simpler neural network, allowing this network to estimate molecular similarity in terms of 3D features, within milliseconds and without the more computationally-demanding and force-filed dependent generation of conformers and structure alignment (**Figure 1d**).

Preparation of the training set for the ESR model. First, we collected a random subset of 6,000 small molecules from the ZINC database [8] and divided it (randomly) into the training and test sets in 5:1 proportion. Next, we used each compound as a seed molecule in the Skalic's shape-captioning generator, thus obtaining, in total, 117,824 vectors corresponding to 68,734 unique SMILES strings (on average, 20 "similars" per one seed, with ca. 2% of "similars" discarded because of SMILES errors, as detected by RDKit [12]). We kept the entries duplicated with respect to SMILES, since in the generator setting of the original model, several slightly different vectors may lead to the same molecule. Since we wished our teacher-student training procedure to average over this redundancy, the dataset was balanced by introducing sample weights inversely proportional to the number of generated vectors per unique SMILES. For instance, if a given molecule appeared 5 times (each time with a different vector) in the set, each of its occurrences was assigned with a weight of 0.2. We note that 11 seed molecules from the test set generated SMILES overlapping with those present in the training set. This is likely a consequence of the similarity between these 11 seed compounds and the training set, and therefore we discarded them together with their descendants. Finally, we represented SMILES with ECFP4 fingerprints [13] kept as vectors of 2048 integers denoting substructure counts (instead of more commonly used binary values).

ESR model training. Before training, columns (features) with zero variance (computed over the training set) were removed from the data, thus reducing the input (fingerprint) dimensionality to 2044 and output (ESR vectors) from 512 to 196. The model hyperparameters, including L2 regularization and dropout factors, as well as numbers of layers and neurons, were optimized with hyperas [14] package for Bayesian optimization. In order to accelerate this step, we randomly selected 10% of both training and test sets, and then performed 50 optimization trials selecting models with the best score (mean squared error) over the test set. This procedure led to an architecture comprised of two layers with 512 and 196 neurons, respectively. Dropout mask with 0.1 dropout probability was applied to all connections and L2 regularization with coefficient equal to 10^{-5} applied to all weights.

Selection of "parent" drugs already considered for COVID-19 treatment. An extensive literature search for approved drugs that have shown therapeutic potential against COVID-19 resulted in 71 hits, of which 29 are currently being tested in clinical trials and 42 exhibited some activity in *in vitro* studies. This collection, along with pertinent literature references, is detailed in the Supplementary Information, **Table S1**. In the following, the selected drugs are referred to as "parents".

Selection of "progeny" drugs. We curated a collection combining (i) drugs and bioactive substances approved in major world jurisdictions and deposited in ZINC [8] and (ii) experimental, Phase 3 and 4 drug candidates from ChemBL [15]. These datasets were then cleared of duplicates, resulting in 1,634 ZINC drugs, 808 Phase 3 drugs, and 2014 Phase 4 ones.

Results and Discussion

One of our motivations of implementing Mol2Vec and ESR approaches is that traditional measures of molecular similarity gave largely unsatisfactory results. In particular, similarity evaluated by the popular metrics such as Morgan-fingerprint-based Tanimoto coefficient [16] (at the > 0.85 threshold [17]) made only rather trivial suggestions in which the parent and progeny shared a common scaffold with relatively small modifications in terms of functional groups (e.g., Cyclosporine parent and Volcosporin progeny differing in only vinyl vs. methyl groups; Ritonavir parent and its hydroxy-Ritonavir metabolite; Toremifene citrate and Tamoxifen differing in one chlorine atom). Unlike these traditional models – whose weaknesses are well documented [2] – we have hoped to capture not only the "make-up" of the molecules in terms of the fragments they contain, but also information about the mutual arrangement (Mol2Vec) or even spatial, 3D distributions of the groups sharing similar *properties* though not necessarily the same atoms (e.g., in ESR, similar H-bonding/accepting propensities of structurally different groups).

After vectorizing the SMILES of all parent and progeny drugs using Mol2Vec and ESR representations discussed above, we first visualized their distributions in multidimensional spaces constructed using t-distributed Stochastic Neighbor Embedding (t-SNE) [18] implemented in Sci-Kit learn module [19]. Two-dimensional projections of these spaces are shown in **Figure 2** and evidence different distributions, though sharing some similarities. For example, area densely occupied by compounds currently tested for COVID-19 (red circles) is also rich in drugs in Phase 3 clinical trials

(orange circles). Another observation is that analogs of Emtricabine, Azvudine and Ribavirin antivirals tend to form similar clusters in both projections.

Next, we calculated Euclidean distances between each parent and progeny drugs in both Mol2Vec and ESR representations. In order to provide a common scale for these molecules' similarity, we took the following steps: (i) in either of the representations, the distance matrix (i.e., the matrix of distances between parents and progenies) was divided by its largest element (the largest distance); (ii) such normalized distances, d_{ij}, between parent i and progeny j, were subjected to a non-linear transformation $s_{ij} = \left(1 + 100 \cdot d_{ij}\right)^{-1}$ where factor of 100 was introduced to increase the "contrast" between the closest neighbors and other compounds in the resulting plots. Importantly, after this transformation, similarity score s_{ij} of compounds being far apart (large d_{ij}) tends to 0, whereas for points laying within close proximity, it is close to 1.

For both vectorization methods, we then selected 150 most similar parent-progeny pairs and analyzed them further to exclude the most unlikely entries, such as metabolites that are not used as drugs (but are present in the ZINC's World Drugs collection), dietary supplements, contrast agents, and drugs whose properties make them extremely unlikely candidates for COVID-19 treatment (e.g., bone resorption drugs, topical agents, or drug transport media). After merging the results from Mol2Vec and ESR, we obtained 133 unique parent-progeny pairs in which there were 110 unique progenies. The similarities of these 110 progenies against approved and experimental parents considered for COVID-19 are summarized in s_{ij} heatmaps such as one shown in **Figure 3** (see also **Figures S1-S3** and **Table S2**). A more informative representation, however, is one in which the most similar parent-progeny pairs are connected by arrows (of length proportional to the s_{ij} metric calculated by either Mol2Vec or ESR).

Figures 4-7 show the top-scoring pairs in which the parents (in the "inner circle" of each figure) are drugs already tested against COVID-19 (either in clinical trials, **Figures 4**,**6**, or exhibiting *in vitro* activity, **Figures 5**,**7**) and their most similar progenies (in the "outer circle" of each figure) are suggestions for "similars" as predicted by either the ESR (**Figure 4,5**) or Mol2Vec (**Figures 6,7**) methods. The first observation is that the models capture, as should be expected, similarity between very close analogs sharing the same scaffolds (e.g., Emtricitabine antiviral parent vs. Lamivudine and other antiviral progenies in the *upper-left* "spider" in **Figure 4**; or Tenofovir vs. Adefovir antivirals in the *lower-right* "spider"). Less obviously, there are also pairs that do not look that similar in terms of 2D structures but do have similar 3D conformations. In fact, additional analyses by the so-called

ShaEP method [20] – quantifying similarity of conformations both in terms of shape and distribution of electrostatic potential – showed significant overlap even in such counterintuitive cases as Thalidomide parent and its Felbamate progeny (**Figure 4**): Although the latter's 2D structure appears "extended", its 3D conformation is more "cyclic" due to hydrogen-bond interactions (see superimposed conformations in **Figure S4**). Overall, the 3D similarity calculated by ShaEP and averaged over 133 parent-progeny pairs we consider is 0.75 (0.91 in terms of the shape-similarity contribution and 0.61 in terms of the distribution of electrostatic potential). In a broader context, these figures can be compared with the similarities between drugs that are known to act against the same (non-COVID-related) targets. For instance, within the family of 41 medications targeting Cox-2 (e.g. Aspirin, Ibuprofen, Naproxen, Diflunisal, Flurbiprofen), the similarity index is 0.57 (0.71 in terms of shape and 0.47 in terms of electrostatic potential), whereas for the family of 22 drugs targeting mitogen-activated protein (MAP) kinetase p38α, the corresponding numbers are 0.58 (0.74, 0.49). Such comparisons can serve as an independent validation that similarities captured by our AI methods are generally in line with traditional conformational-analysis approaches.

Each of the substances showed in **Figures 4-7** is accompanied by a colored marker specifying its primary therapeutic indication. These indications are mostly from the DrugBank (see **Tables S1** and **S2** for some additional references) and do not mean that a given substance does not have any other uses – for instance, primary indication of Arzoxifene is for breast and endometrial cancer treatments, but it is also in trials for post-menopausal osteoporosis treatment [21]. We recognize that some indications may be less relevant to COVID-19 than others but, at the same time, we note that parents coming from such seemingly unlikely classes have been considered in literature-reported COVID-19 studies (e.g., a CNS drug Thioridazine exhibited *in vitro* activity and is therefore included as a parent in **Figure 5**). This being said, our own – likely, subjective – focus has been on drugs that might mitigate the viral infection itself (antivirals), those that can be in some way related to immune response accompanying serious COVID-19 infections (immunosuppressant, immunomodulatory, antiasthmatics, antirheumatic and anti-inflammatory agents), or mucolytics helpful in airway clearance .

Some parent-progeny suggestions are obvious, e.g., the antiviral progenies very similar to the Emtricitabine parent in the upper-left "spider" in **Figure 4**. Focusing on less trivial pairs, we note an Azuvidine parent in **Figure 4** and Ribavirin parent found in both **Figures 5** and **7**. Among Azuvidine's progenies, Mizoribine immunosuppressant and Pidotimod immunomodulator stand out because of their relevant indications and because their 3D conformations show good overlap with Azuvidine (**Figures 8a,b**). Rivabirin also gives Mizoribine progeny and, additionally, Diphylline antiasthmatic,

Inosine antinflammatory, and Thioinosine immunosuppressant, all of which show significant 3D parent-progeny overlap (**Figures 8c-f**). The variety of these primary indications reflects different roles nucleotides and their derivatives play in processes ranging from signaling [22, 23]), to the inhibition of viral replication (e.g., Ribavirin [24, 25]), to immunosuppression (e.g., Mizoribine inhibits inosine and guanosine monophosphate synthetases [26]). Although detailed structural knowledge of SARS-CoV-2 proteins involved in viral replication is lacking, we do not find it inconceivable that some of the closely-shaped nucleotide-based immunosuppressants/immunomodulators might also show activity against viral replication.

Considering the family of Acetylcysteine mucolytic agent ("spiders" at the bottom of **Figure 4** and on the right of **Figure 6**), there are obvious, also mucolytic progenies (Mesna, Mecysteine and Carbocysteine) as well as three interesting progenies with different but, we think, promising indications: Bucillamine used in rheumatoid arthritis [27], Penicilamine used against Wilson's desease [28], and Tromethamine antiasthmatic [29, 30]. All of these progenies show significant conformational overlap with the Acetylcysteine parent (see examples in **Figures 9a-d**). Several of them contain a reactive thiol group, which may break disulfide bonds in some proteins or act by influencing the oxidative-reductive balance in the organism (in many cases, the actual mechanism remains unknown) [27]. Compounds of similar properties (but containing disulfide bridges rather than free SH groups) were shown to have potential therapeutic effect on COVID-19 [31].

Another interesting example is the Niclosamide-TCSA pair (**Figure 7**). Here, the parent drug, used typically as an anti-parasitic agent, was shown to exhibit antiviral activity [32] but also significant side-effects. We do not feel qualified enough to judge if the side-effects of the TCSA bacteriostat [33] are equally problematic. This being said, we note that this pair has a very good overlap and ShaEP score (**Figure 9e**), one of the best in our collection (and, interestingly, well above the average score in the family of COX-2 inhibitors mentioned earlier). Finally, connection between Tefonovir antiviral and Dyphilline antiasthmatic (**Figure 4**) strikes us as non-obvious but relevant given COVID's respiratory symptoms. As for other pairs, this one also shows good 3D overlap (**Figure 9f**). We note that Dyphilline is also a progeny of another antiviral, Ribavirin, we discussed earlier.

Hoping that Readers identify other pairs of potential clinical relevance, we provide **Figure 10** showing only those progenies that are predicted by one or both methods for different parents (*progenies with two or more incoming arrows*) and those predicted by both Mol2Vec and ESR methods (*red arrows*). Our suggestion for these pairs is that they correspond to the most conservative (also less "imaginative") choices from the full selection in **Figures 4-7**.

Conclusions

In summary, we performed AI analyses to suggest molecules that are similar – in terms of 3D shape and pharmacophore distributions – to the ones already being tested against COVID-19. The progeny compounds we identified might become useful should the currently-tested drugs fail to show desired effect. Conceptually, this work is a form of "reasoning by analogy/similarity," albeit with the use of AI methods more advanced and more accurate than traditional means of quantifying molecular similarity. Under ordinary circumstances, such analyses of drug re-purposing would likely be secondary to the efforts to develop brand new drugs. However, the timelines imposed by COVID-19 pandemics are not ordinary, and re-purposing appears to be the sole timely alternative.

Author contributions

W.B. and M.M. implemented most of the algorithms described in the paper. R.R. performed calculations in ShaEP. E.P.G, A.W., K.M. and S.S. helped with data collection and analysis of results. B.A.G. conceived and supervised the project. All authors contributed to the writing of the manuscript.

Acknowledgements

This work was supported by Allchemy's internal research funds. In addition, the authors are grateful to the U.S. DARPA supporting several modules within the Allchemy platform under contract B634874 (administered by the Lawrence Livermore National Laboratory). B.A.G. also gratefully acknowledges personal support from the Institute for Basic Science Korea, Project Code IBS-R020-D1.

Conflict of interest

The authors are owners and/or contractors of Allchemy, Inc. This being said, the authors declare no financial interest in the current work.

References

[1] Kupferschmidt, K. & Cohen, J. Race to find COVID-19 treatments accelerates. *Science* **367**, 1412–1413 (2020).

[2] Maggiora, G., Vogt, M., Stumpfe, D. & Bajorath, Molecular similarity in medicinal chemistry. *J. Med. Chem.* **57**, 3186–3204 (2014).

[3] Jaeger, S., Fulle, S. & Turk, S. Mol2vec: Unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* **58**, 27–35 (2018).

[4] Jeon, S. *et al.* Identification of antiviral drug candidates against SARS-CoV-2 from FDA-approved drugs. bioRxiv 2020.03.20.999730 (2020) doi:10.1101/2020.03.20.999730.

[5] Sekhar, T. Virtual Screening based prediction of potential drugs for COVID-19. Preprints (2020) doi:10.20944/preprints202002.0418.v2.

[6] Skalic, M., Jiménez, J., Sabbadin, D. & De Fabritiis, G. Shape-Based Generative Modeling for de Novo Drug Design. *J. Chem. Inf. Model.* **59**, 1205–1214 (2019).

[7] Cadeddu, A., Wylie, E. K., Jurczak, J., Wampler-Doty, M. & Grzybowski, B. A. Organic chemistry as a language and the implications of chemical linguistics for structural and retrosynthetic analyses. *Angew. Chemie Int. Ed.* **53**, 8108–8112 (2014).

[8] Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model*. **52**, 1757–1768 (2012).

[9] Gómez-Bombarelli, R. *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).

[10] Ding, H., Chen, K. & Huo, Q. Compressing CNN-DBLSTM models for OCR with teacher-student learning and Tucker decomposition. *Pattern Recognit.* **96**, 106957 (2019).

[11] Li, J., Zhao, R., Huang, J.-T. & Gong, Y. Learning small-size DNN with output-distribution-based criteria. *Proc. Interspeech* 1910–1914 (2014).

[12] RdKit: Open-source cheminformatics. www.rdkit.org.

[13] Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

[14] Keras + Hyperopt: A very simple wrapper for convenient hyperparameter optimization. http://maxpumperla.com/hyperas/

[15] Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2016).

[16] Willett, P., Barnard, J. M. & Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996 (1998).

[17] Sasmal, S., El Khoury, L. & Mobley, D. L. D3R Grand Challenge 4: ligand similarity and MM-GBSA-based pose prediction and affinity ranking for BACE-1 inhibitors. *J. Comput. Aided. Mol. Des.* **34**, 163–177 (2020).

[18] van der Maaten, L. J. P. & Hinton, G. E. Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

[19] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

[20] Vainio, M. J., Puranen, J. S. & Johnson, M. S. ShaEP: Molecular overlay based on shape and electrostatic potential. *J. Chem. Inf. Model.* **49**, 492–502 (2009).

[21] Arzoxifene https://www.drugbank.ca/drugs/DB06249#pharmacology

[22] Berg, J. M., L, T. J. & Stryer, L. *Biochemistry* (W H Freeman, 2002).

[23] Caffine. Drugbank https://www.drugbank.ca/drugs/DB00201#pharmacology.

[24] Ortega-Prieto, A. M. *et al.* Extinction of Hepatitis C virus by Ribavirin in hepatoma cells involves lethal mutagenesis. *PLoS One* **8**, e71039 (2013).

[25] Crotty, S., Cameron, C. & Andino, R. Ribavirin's antiviral mechanism of action: lethal mutagenesis? *J. Mol. Med.* **80**, 86–95 (2002).

[26] Yokota, S. Mizoribine: Mode of action and effects in clinical use. *Pediatr. Int.* **44**, 196–198 (2002).

[27] Amersi, F. *et al*. Bucillamine, a thiol antioxidant, prevents transplantation-associated reperfusion injury. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 8915–20 (2002).

[28] Peisach, J. & Blumberg, W. E. A mechanism for the action of penicillamine in the treatment of Wilson's disease. *Mol. Pharmacol.* **5**, 200–9 (1969).

[29] Kallet, R.H., Jasmer, R.M., Luce, J.M. *et al.* The treatment of acidosis in acute lung injury with tris-hydroxymethyl aminomethane (THAM). *Am. J. Resp. Crit. Care Med.* **161**, 1149–1153 (2000).

[30] Hoste, E.A., Colpaert, K., Vanholder, R.C., Lameire, N.H., De Waele, J.J., Blot, S.I., Colardyn, F.A. Sodium bicarbonate versus THAM in ICU patients with mild metabolic acidosis. *Journal of Nephrology*. **18**, 303–307 (2005).

[31] Lin, M.-H. *et al.* Disulfiram can inhibit MERS and SARS coronavirus papain-like proteases via different modes. *Antiviral Res.* **150**, 155–163 (2018).

[32] Xu, J., Shi, P.-Y., Li, H. & Zhou, J. Broad Spectrum Antiviral Agent Niclosamide and Its Therapeutic Potential. *ACS Infect. Dis.* (2020) doi:10.1021/acsinfecdis.0c00052.

[33]3,3’,4’,5-Tetrachlorosalicylanilide. PubChem https://pubchem.ncbi.nlm.nih.gov/compound/3_3_4_5-Tetrachlorosalicylanilide.

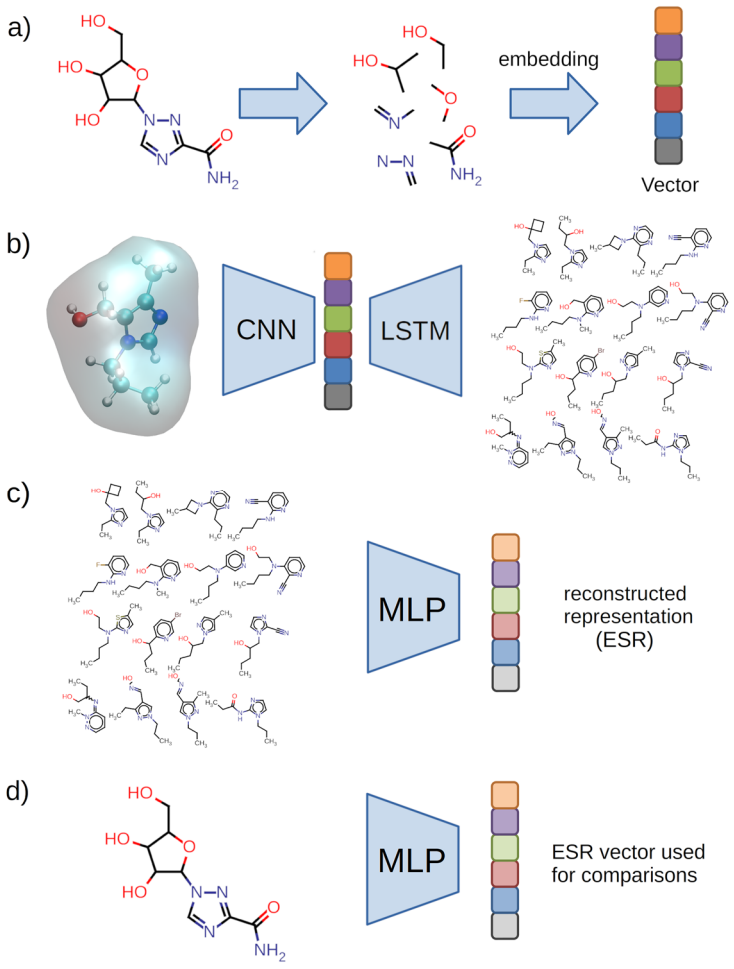

Figure 1. Schematic illustration of models used in this study. a) Mol2Vec [3] treats a molecule (here, Ribavirin antiviral drug shown on the *left*) as a "sentence", whose molecular substructures correspond to "words" (here, for the sake of clarity, the *middle* panel shows only a small subset of matching words). The "words" are each embedded in a 300-dimensional space and the entire molecule is a sum of all such vectors. **b)** The shape-captioning Variational Autoencoder (VAE) of Skalic and coworkers [6] – comprised of a Convolutional Neural Network (CNN) encoder and a Long-Short Term Memory (LSTM) decoder – takes a 3D structure of a seed molecule (here, a random molecule taken from ZINC, shown on the *left*) and generates novel molecules similar in terms of 3D shape and pharmacophoric similarity (examples shown on the *right*). **c)** In the current work, the input and output of the LSTM module "teaches" a Multi-Layered Perceptron (MLP) "student" the relationship between molecules' SMILES and their 3D features. **d)** The thus trained MLP is then used as an encoder, transforming SMILES representation into a 3D-aware ESR vector used in molecule-to-molecule comparisons in a manner similar to Mol2Vec.

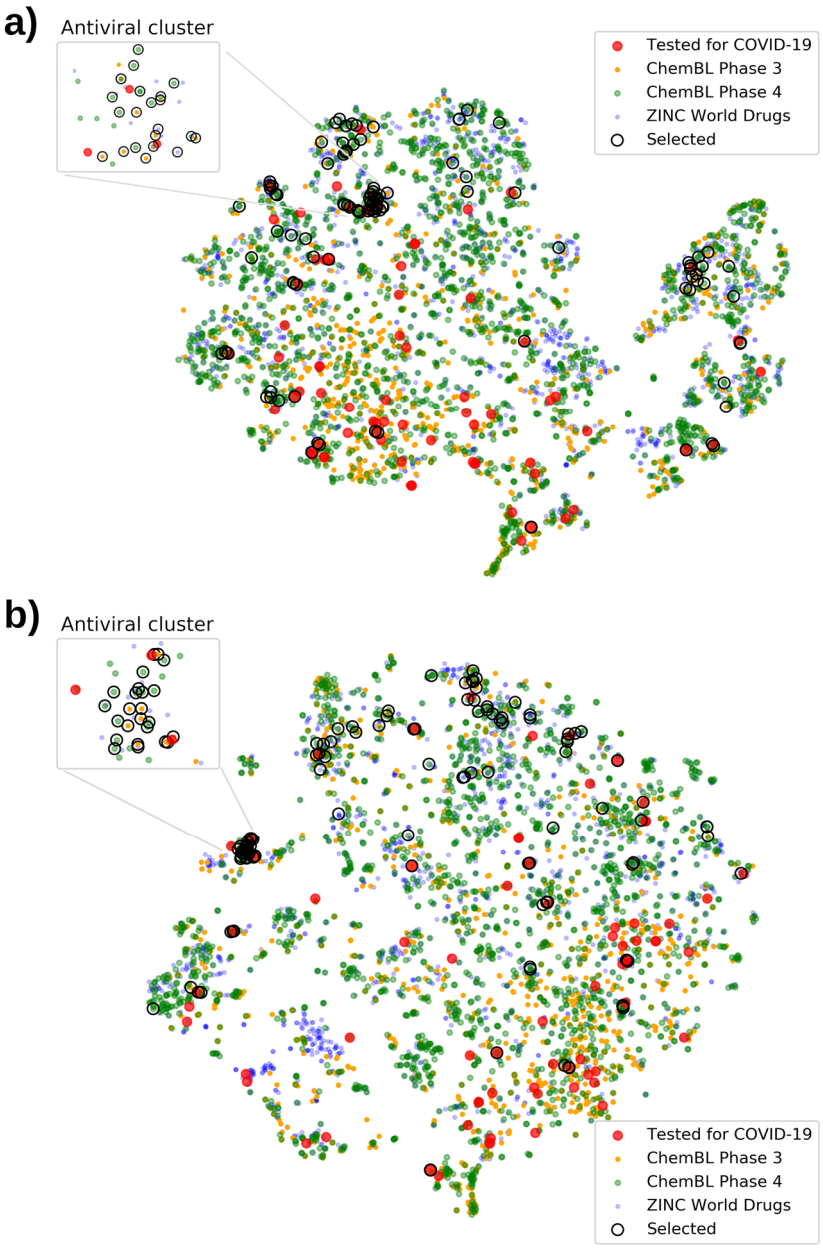

Figure 2. t-SNE projection of multidimensional drug space defined by a) ESR and b) Mol2Vec vectorizations. Red circles represent drugs already tested against COVID-19 (either in clinical trials or *in vitro* studies). Orange and green circles denote drug candidates in Phase 3 or 4 clinical trials and retrieved from ChemBL. Blue circles describe entries from ZINC World Drugs not present in the aforementioned sets. The regions corresponding to 110 progenies most similar to parent drugs are marked with additional black rings. Insets in the upper left corners of both a) and b) magnify an antiviral cluster around three drugs being tested for COVID-19: Emtricitabine (*top-most* red circle), Azvudine (*right-most* red circle) and Ribavirin (*left-most* red circle).

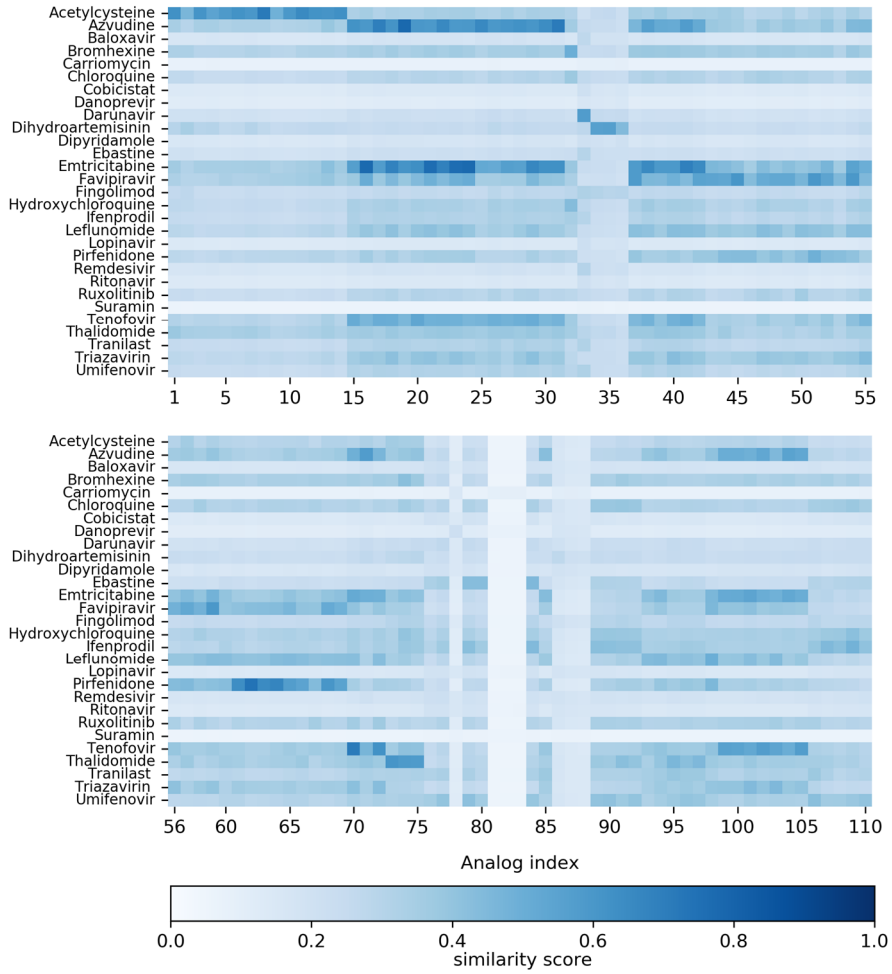

Figure 3. A "heatmap" quantifying similarity, s_{ij}, between drugs already being tested in clinical trials against COVID-19 and 110 unique, most-similar progenies found by the ESR method. Each row corresponds to one parent drug annotated with its common name. Progenies are denoted along the horizontal axis by numbers corresponding to those in **Table S2** detailing these progenies' names and primary therapeutic indications. Because of limited space, the matrix is divided into two parts shown one above the other. For other similarity maps (ESR-based for *in vitro* parents, Mol2Vec for clinical-trial parents, and Mol2Vec for *in vitro* parents) see SI, **Figures S1-S3**).

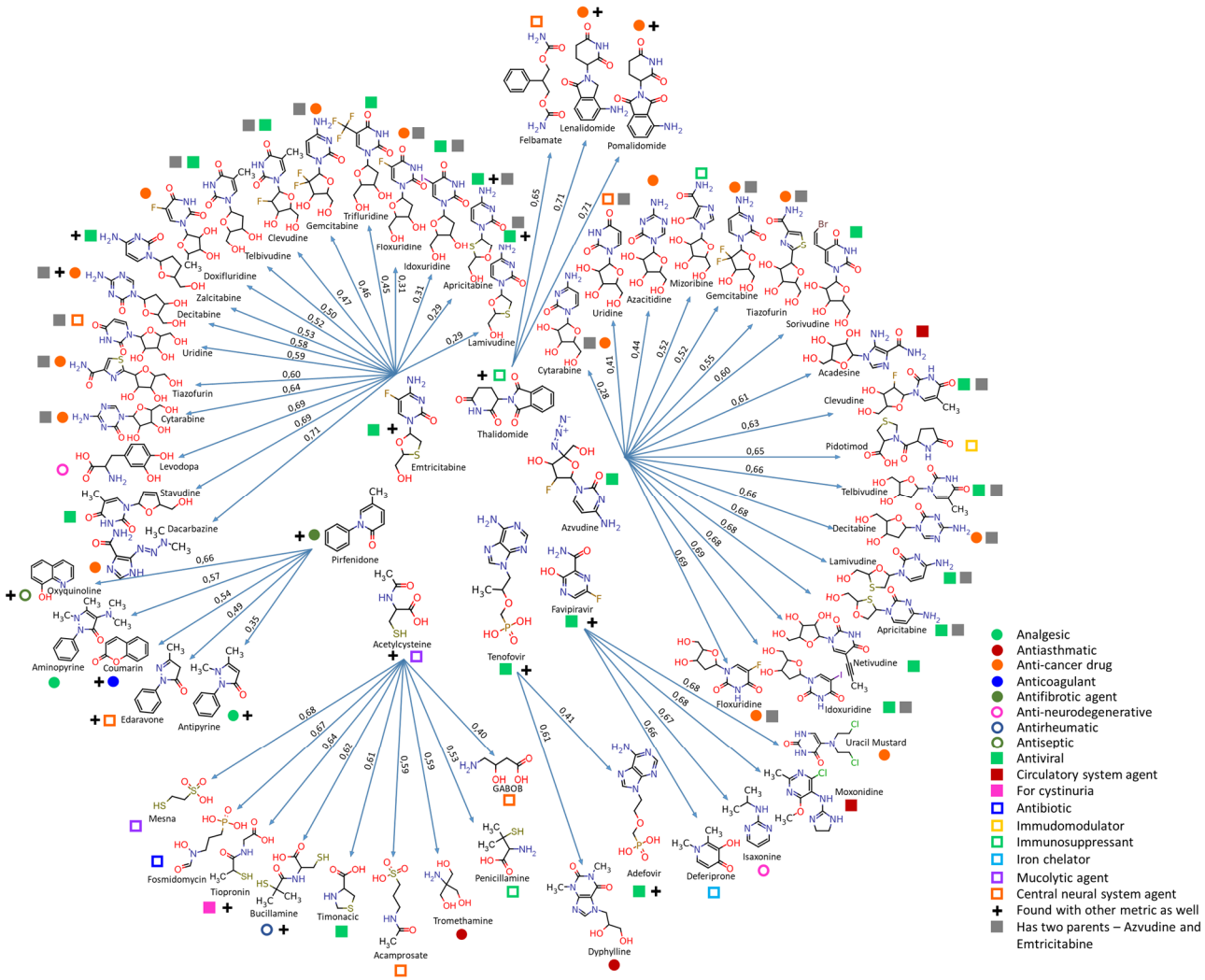

Figure 4. Parent drugs already tested in clinical trials for COVID-19 and their most relevant progenies found by the ESR similarity metric. The progenies were taken from the closest 150 neighbors of each parent compound, with exclusion of drug metabolites, dietary supplements, contrast agents, etc. Parent drugs for which no similar progenies were found are not shown. For each parent-progeny pair, the relative distance ($100*d_{ij}/d_{max}$, where d_{ij} is the distance in ESR space and d_{max} is the maximum distance in the set) is represented by an arrow of proportional length. In addition, therapeutic indications (e.g., antiviral, anti-inflammatory, etc.) are denoted by color markers explained in the legend on the lower right. Please note that these are *primary* indications (as provided in DrugBank and sometimes other sources, please see **Tables S1** and **S2**) and the drugs may have additional uses as well.

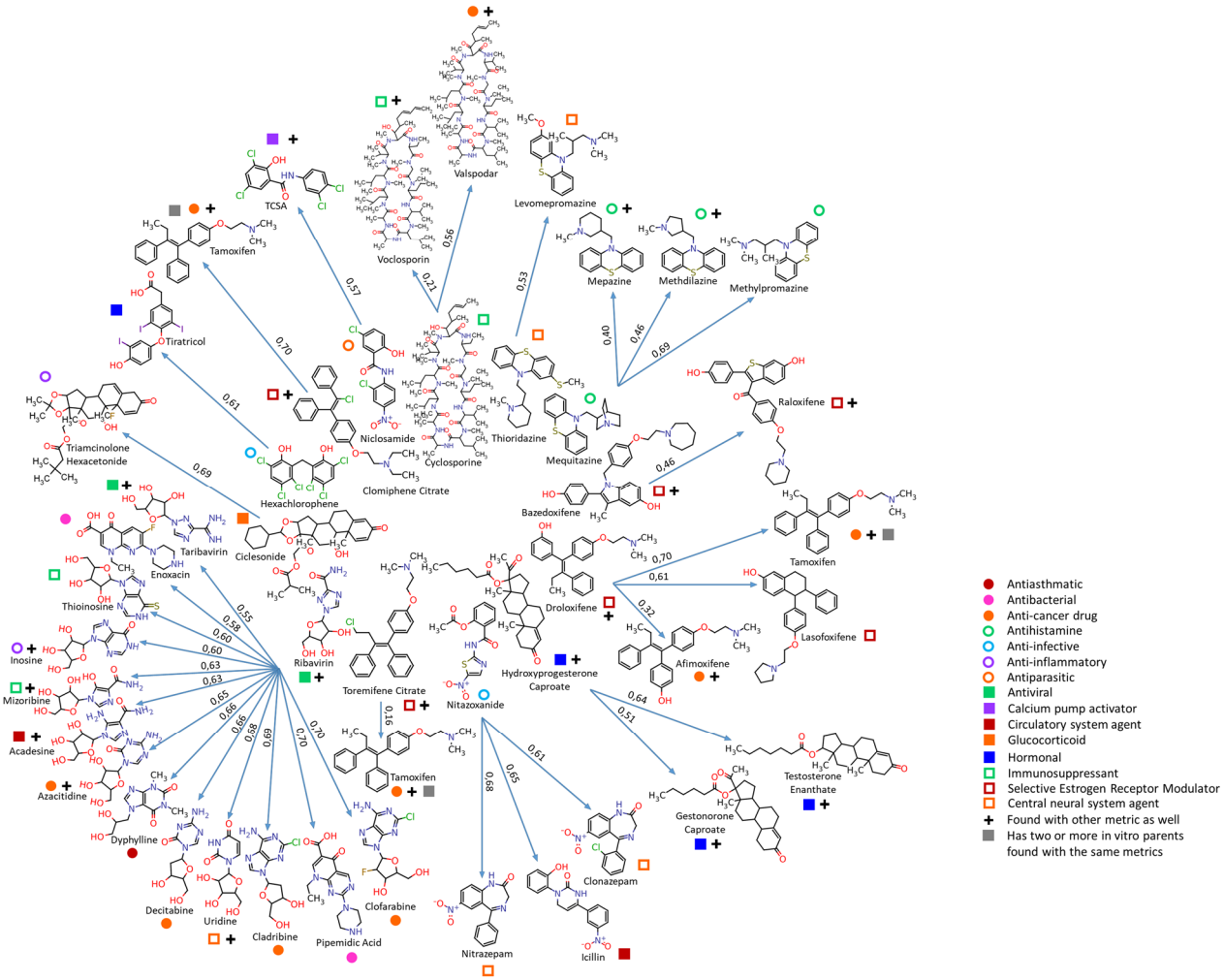

Figure 5. Parent drugs of verified *in vitro* activity against COVID-19 and their most relevant progenies found by the ESR similarity metric. The progenies were taken from the closest 150 neighbors of each parent compound, with exclusion of drug metabolites, dietary supplements, contrast agents, etc. Parent drugs for which no similar progenies were found are not shown. For each parent-progeny pair, the relative distance ($100*d_{ij}/d_{max}$, where d_{ij} is the distance in ESR space and d_{max} is the maximum distance in the set) is represented by an arrow of proportional length. In addition, therapeutic indications (e.g., antiviral, anti-inflammatory, etc.) are denoted by color markers explained in the legend on the lower right. Please note that these are *primary* indications (as provided in DrugBank and sometimes other sources, please see **Tables S1** and **S2**) and the drugs may have additional uses as well.

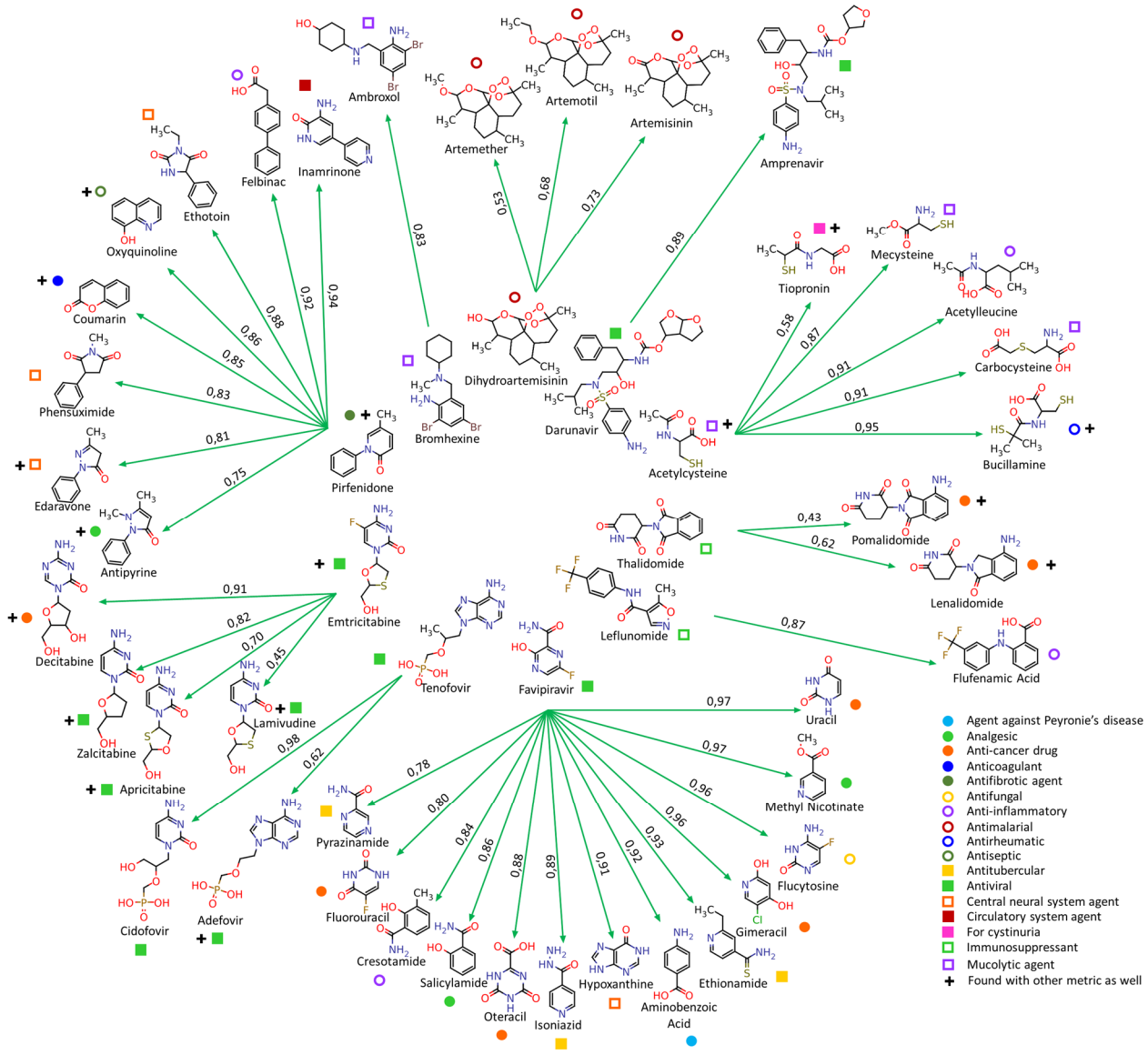

Figure 6. Parent drugs already tested in clinical trials for COVID-19 and their most relevant progenies found by the Mol2Vec similarity metric. The progenies were taken from the closest 150 neighbors of each parent compound, with exclusion of drug metabolites, dietary supplements, contrast agents, etc. Parent drugs for which no similar progenies were found are not shown. For each parent-progeny pair, the relative distance ($100*d_{ij}/d_{max}$, where d_{ij} is the distance in Mol2Vec space and d_{max} is the maximum distance in the set) is represented by an arrow of proportional length. In addition, therapeutic indications (e.g., antiviral, anti-inflammatory, etc.) are denoted by color markers explained in the legend on the lower right. Please note that these are *primary* indications (as provided in DrugBank and sometimes other sources, please see **Tables S1** and **S2**) and the drugs may have additional uses as well.

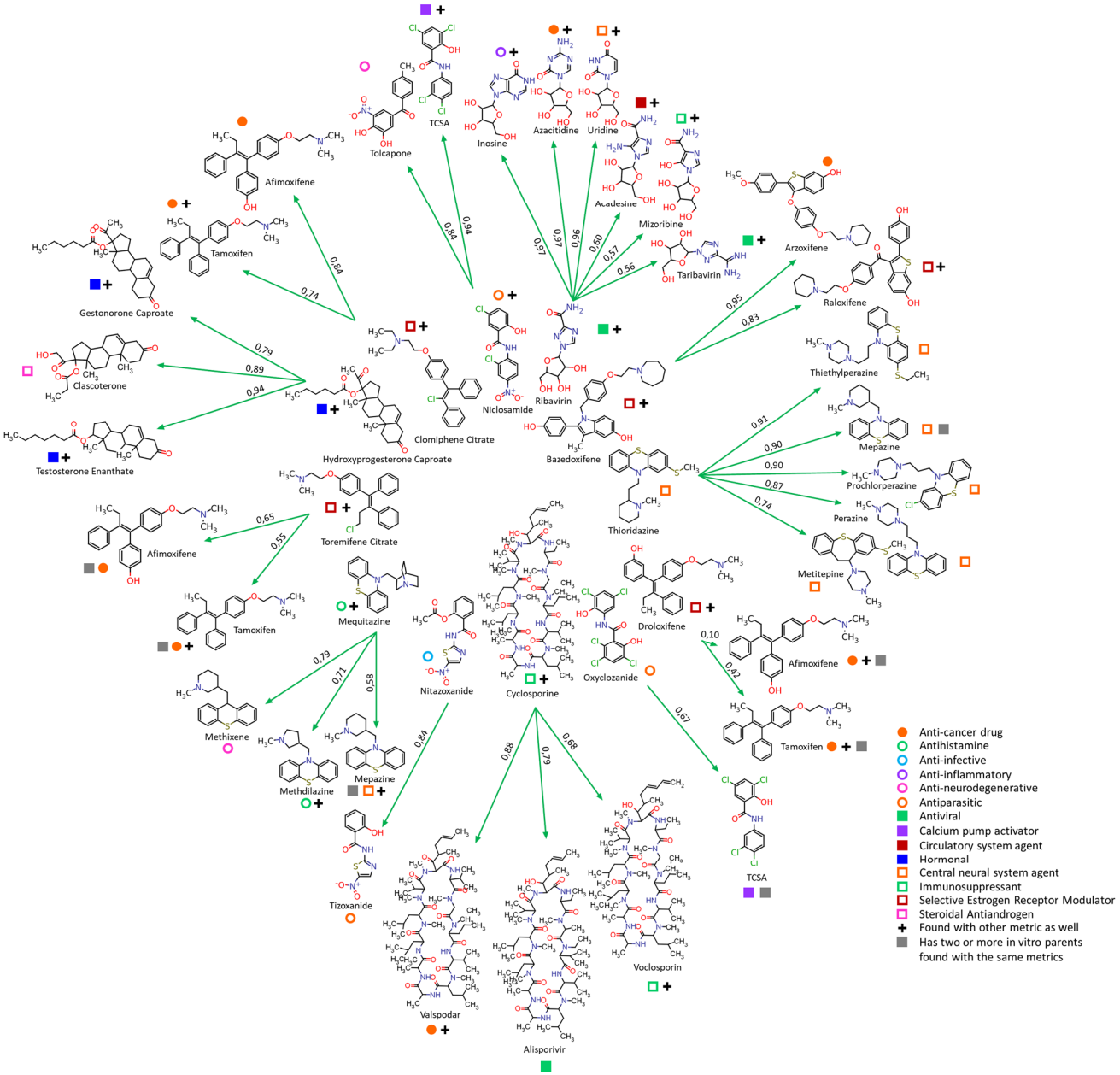

Figure 7. Parent drugs of verified *in vitro* activity against COVID-19 and their most relevant progenies found by the Mol2Vec similarity metric. The progenies were taken from the closest 150 neighbors of each parent compound, with exclusion of drug metabolites, dietary supplements, contrast agents, etc. Parent drugs for which no similar progenies were found are not shown. For each parent-progeny pair, the relative distance ($100*d_{ij}/d_{max}$, where d_{ij} is the distance in Mol2Vec space and d_{max} is the maximum distance in the set) is represented by an arrow of proportional length. In addition, therapeutic indications (e.g., antiviral, anti-inflammatory, etc.) are denoted by color markers explained in the legend on the lower right. Please note that these are *primary* indications (as provided in DrugBank and sometimes other sources, please see **Tables S1** and **S2**) and the drugs may have additional uses as well.

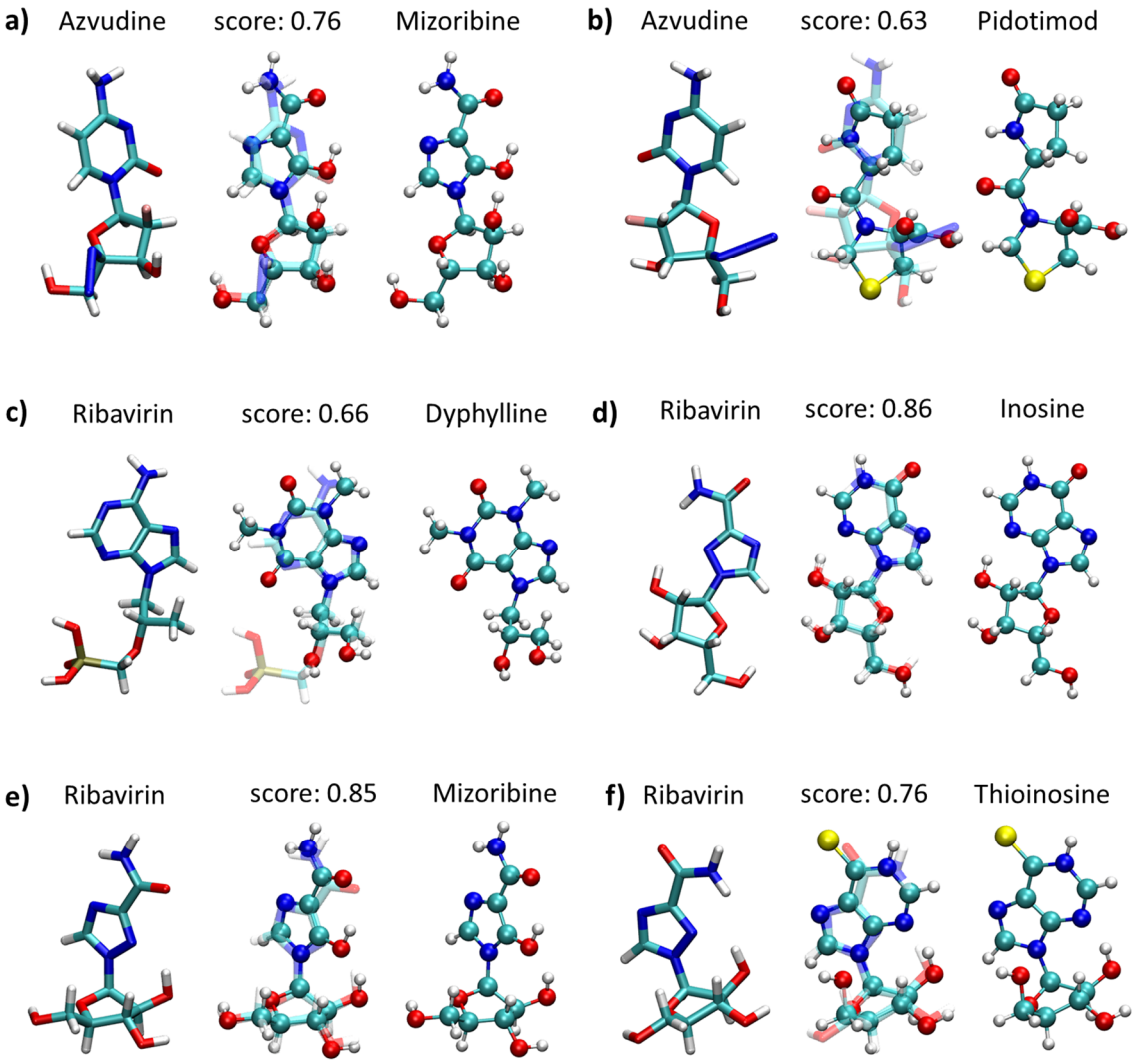

Figure 8. Best alignments of selected parent-progeny pairs obtained with ShaEP. In each panel, parent is shown on the *left* as a licorice model, progeny is drawn as a ball-and-stick model on the *right*, and the *middle* portion overlays the two structures. Numbers quantify alignment scores (including both shape and electrostatic factors). **a)** Parent Azvudine and progeny Mizoribine, score 0.76; **b)** Parent Azvudine and progeny Pidotimod, score 0.63; **c)** Parent Ribavirin and progeny Dyphylline; score 0.66, **d)** Parent Ribavirin and progeny Inosine; score 0.86; **e)** Parent Ribavirin and progeny Mizoribine, score 0.85, **f)** Parent Ribavirin and progeny Thioinosine, score 0.76.

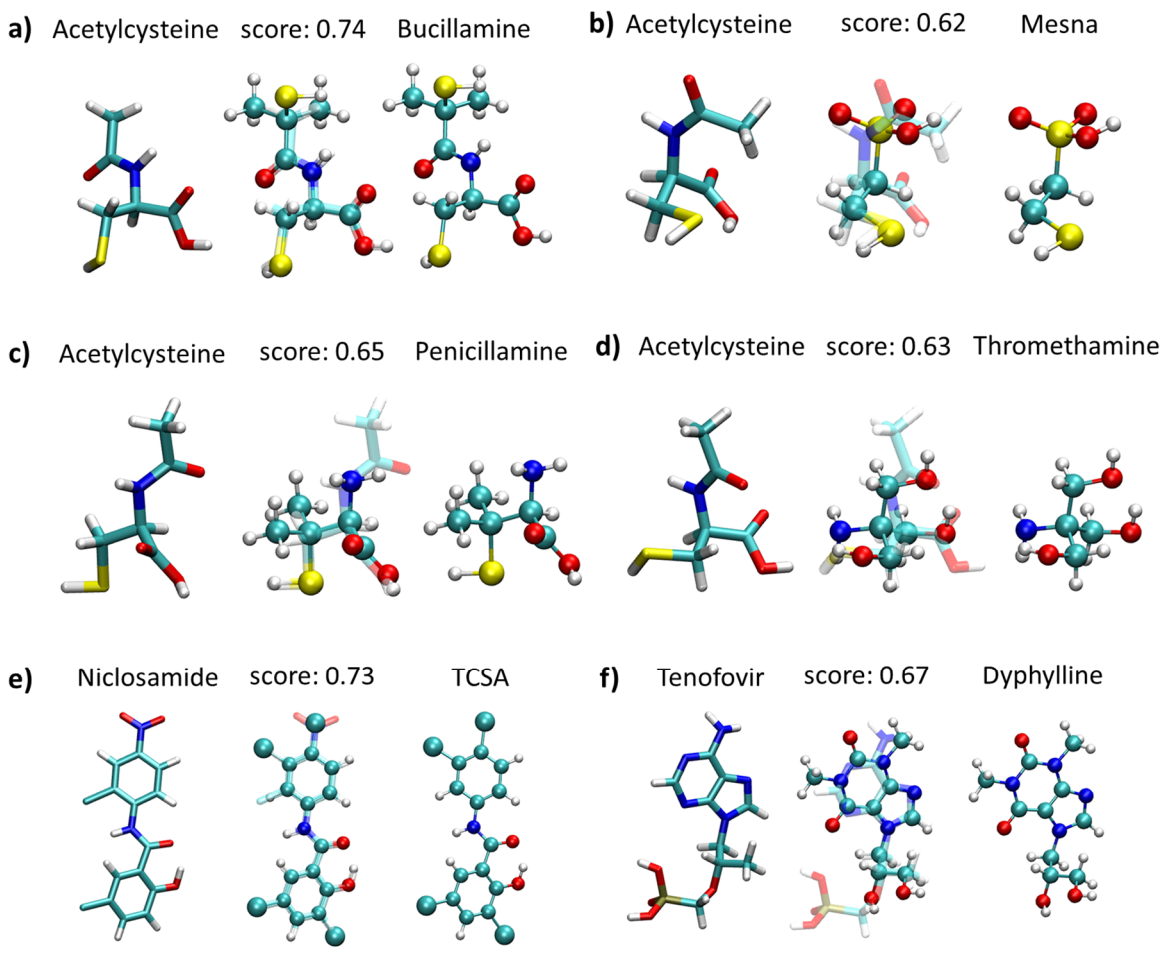

Figure 9. Best alignments of selected parent-progeny pairs obtained with ShaEP. In each panel, parent is shown on the *left* as a licorice model, progeny is drawn as a ball-and-stick model on the *right*, and the *middle* portion overlays the two structures. Numbers quantify alignment scores (including both shape and electrostatic factors). **a)** Parent Acetylcysteine and progeny Bucillamine, score 0.74; **b)** Parent Acetylcysteine and progeny Mesna, score 0.62; **c)** Parent Acetylcysteine and progeny Penicillamine, score 0.65; **d)** Parent Acetylcysteine and progeny Tromethamine, score 0.63; **e)** Parent Niclosamide and progeny TCSA (3,3',4',5-tetrachlorosalicylanilide), score 0.73; **f)** Parent Tenofovir and progeny Dyphylline, score 0.67.

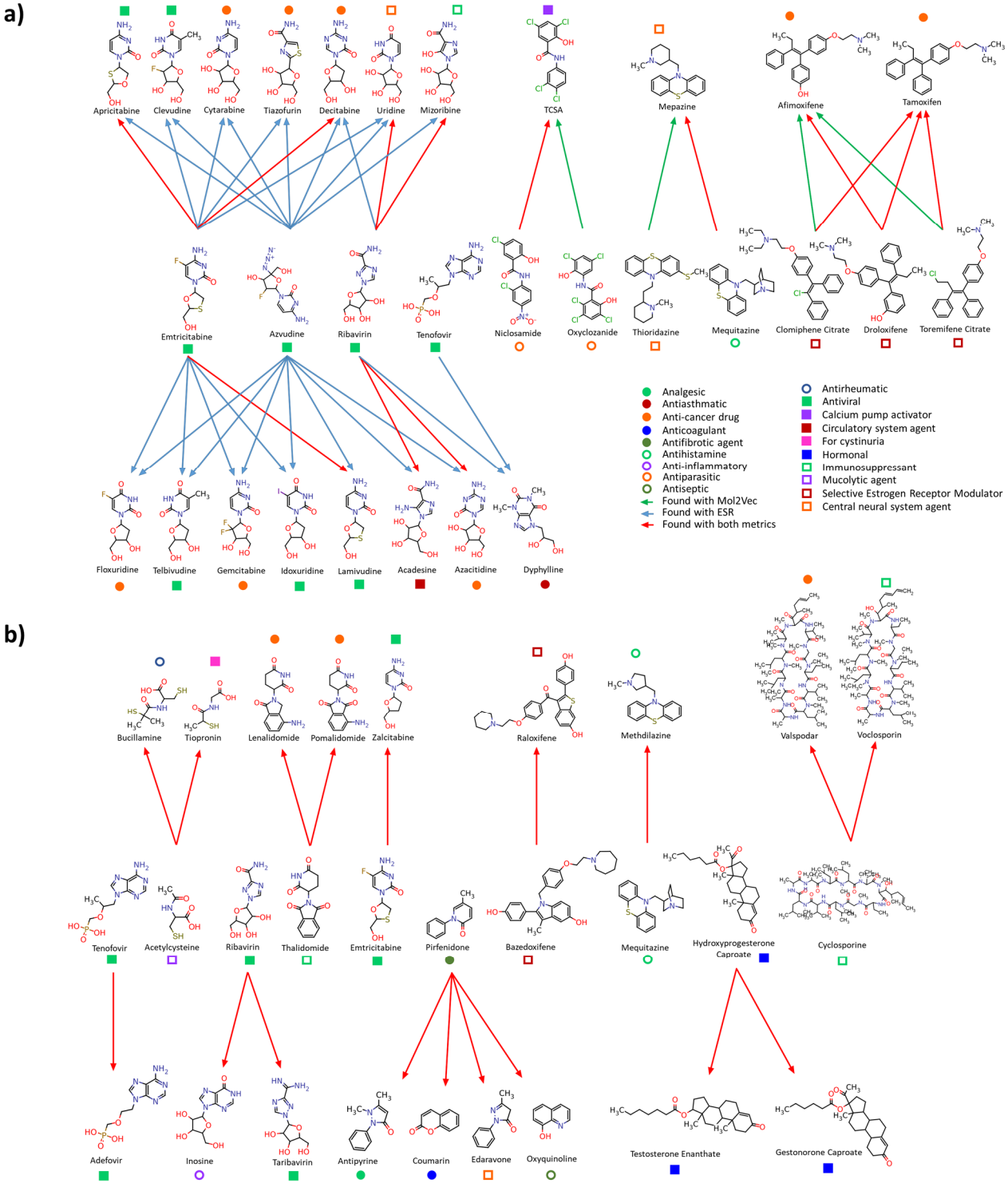

Figure 10. Parent-progeny pairs in which the progeny has several different parents (two or more arrows pointing towards one progeny drug) and/or is suggested by both ESR and Mol2Vec methods (red arrows). Panel **a)** shows progenies that have more than one parent. Panel **b)** shows progenies that have one parent but were found using both ESR and Mol2Vec methods.