Multi-reference Poly-conformational Computational Methods for De-novo Design, Optimization, and Repositioning of Pharmaceutical Compounds.

Vadim Alexandrov
Liquid Algo LLC
Alexander Kirpich
Georgia State University
Yuriy Gankin (yuriy.gankin@quantori.com)
Quantori LLC

Research Article

Keywords: COVID-19, conformers, multi-reference, poly-conformational, in silico, ligand-based, structure-based, SARS-COV-2, fingerprints, cheminformatics, similarity, virtual library, computational framework

Posted Date: February 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-153954/v1

Multi-reference poly-conformational computational methods for de-novo design, optimization, and repositioning of pharmaceutical compounds.

Vadim Alexandrov1, Alexander Kirpich2 and Yuriy Gankin*3

1 Liquid Algo LLC, 85 Thistle Lane, Hopewell Junction, NY 12533, United States of America

2 School of Public Health, Georgia State University, Atlanta, Georgia, United States, Atlanta, GA, United States of America

3 Quantori LLC, 625 Mass Ave, Cambridge, MA 02139, United States of America

*** Corresponding author**

E-mail: yuriy.gankin@quantori.com

ABSTRACT

The COVID-19 epidemic, SARS-CoV-2, that began in December of 2019 has drastically altered the aspects of daily life across the global society. Time-effective treatment of those infected has since become a major goal with multiple treatment strategies having been designed to prevent the progression of the disease into severe pneumonia. To date, no drug has been found to be 100% effective against SARS-COV-2, possibly because each candidate drug was targeting only one particular mechanism of action (MoA). Neither proposed up-to-date anti-SARS-COV-2 vaccine are 100% effective. To contribute to the process of finding a more robust small-molecule solution, utilizing several anti-SARS-COV-2 MoAs, a novel framework is presented; where the in silico generated set of virtual library compounds is compared to six known reference drugs: Chloroquine, Favipiravir, Remdesivir, JQ1, Apicidine, and Haloperidol which have been already used for SARS-CoV-2 treatment. The aims were: a) to present a universal search framework for potential candidate compounds based on the comparison of multiple similarities between compounds' conformers and b) to identify candidate compounds that are simultaneously "close" to each of the six known reference compounds that counteract SARS-CoV-2 via different mechanisms of action.

KEYWORDS

COVID-19, conformers, multi-reference, poly-conformational, in silico, ligand-based, structure-based, SARS-COV-2, fingerprints, cheminformatics, similarity, virtual library, computational framework

INTRODUCTION

The necessity of the transparent algorithm. The novel coronavirus SARS-CoV-2 was introduced into the human population in the Chinese city of Wuhan in the Province of Hubei in December of 2019 [1–4]. Since then the epidemic of SARS-CoV-2 has rapidly spread Worldwide. The World Health Organization (WHO) has officially declared the SARS-CoV-2 pandemic in March 2020 just three months after its emergence [5]. The novel coronavirus received an official name SARS-CoV-2 and the virus pandemic was called COVID-19 [6]. As of November 2020, there has been no virus-specific treatment for SARS-CoV-2 [7] but multiple drugs have been proposed, such as remdesivir [8], talampicillin, lurasidone, rubitecan, loprazolam [9,10], chloroquine, and hydroxychloroquine [9,10], atazanavir [11], azithromycin, quercetin, chloroquine, rapamycin, and doxycycline [12]. The formal evaluation and comparison of drugs can be performed by studying the compound properties by treating patients and performing clinical trials [7–10,12] or by studying the properties of the corresponding compounds in silico [9,11]. The computational methods are preliminary and exploratory, but are less invasive than clinical trials and allow the simultaneous study of large quantities of potential compounds. Since none of the proposed small molecules were found to be 100%-effective, we present a new computational method that combines the best features from the earlier individually imperfect small-molecule treatments.

Conformers as independent molecular entities. In real life, most compound molecules exist in multiple conformations (shapes) based on the surrounding environmental conditions. In particular, each 3D shape of a molecule dictates its biological activity and enables the molecule to fit into the binding pockets of proteins. Often, distinctly different chemical compounds that have similar shapes (and similar charge distributions along the molecular surface) can bind as long as the ligand's partial charges are positioned in the binding pocket the same way (i.e., form the same

hydrogen bonds). Therefore, it is beneficial to compare the shapes and surface distribution charges for target query and reference compounds on a conformer-by-conformer basis. If one of the conformers of the query molecule matches one of the conformers (especially bound-to-target) of the reference molecule, then there is a chance that the reference compound will also exhibit similar binding properties to the same target.

Alignment-free 3D-similarity scoring. OpenEye Scientific Software Inc. pioneered an algorithm and the corresponding tool ROCS [13] for comparing shapes of molecules by overlaying and measuring their molecular structures in silico and comparing differences between a query and target molecule. ROCS identifies potentially active compounds by comparing their shapes. Moreover, the ROCS tool is competitive and often superior to structure-based approaches in virtual screening [14,15] both in terms of overall performance and consistency [16]. As a result novel, molecular scaffolds have been identified by using ROCS against various targets which have been considered very difficult to address computationally [17].

Challenges with overlaying. The process of molecular shapes overlaying remains computationally intensive and often is a bottleneck in the search process for similar molecules. This remains despite the recent so-called PAPER implementation of ROCS on GPU [18] and the development of FastROCS [19] for large (>1B) compound libraries. Recently, alternative methods for overlaying have been introduced as a substitute for the ROCS approach. The alternative overlaying is performed by comparing shape-based descriptors (a.k.a conformer-level 3D fingerprints). An example of such an approach is ElectroShape implemented in the ODDT package [20] and is based on the algorithm that incorporates shape, chirality, and electrostatics [21,22], and represents each conformer via a fixed-length vector of real-valued numbers. Similarly the E3FP package [23] also utilizes an alignment-invariant 3D representation of molecular conformers as a

fixed-length binary vector for each conformer. These fingerprint-based approaches allow to calculate the similarity between two molecular shapes either as a Tanimoto distance (for binary fingerprints) or Euclidean distance (for real-valued fingerprints) computations. Such computations are orders of magnitude faster in comparison to alternative methods that require the actual alignment of the two compared conformers. Even though the calculation of a shape-based fingerprint for each conformer can be a rather computationally involved procedure, as soon as all conformers for the virtual library are fingerprinted and stored in a database, the similarity search for the query molecule in such a database is computationally quick. Therefore the computationally efficient method proposed here is expected to be very useful for finding candidate drugs for multi-target disease indications, ligand-based drug design, and drug repurposing applications.

METHODS

Conformer-by-conformer comparison. The proposed computational algorithm extends the currently available methods [20–23] and introduces additional search flexibility via the use of the compound conformers. The proposal is to compare multiple possible shapes, adopted via varying environmental conditions, of the same molecule (i.e., conformers) rather than just a single shape that was used before. In particular, the suggested approach is based on the matching of ligand-ligand fingerprints and goes beyond the known docking methods that utilize the simulated physical binding of a ligand to the target. The supporting theory behind the method is based on the decision to treat conformers, which might have different binding characteristics and properties, as independent entities. In such an approach each conformer has the corresponding independent alignment-free 3D-similarity scoring using the known multi-references. All conformers were generated using the ETKDG algorithm implemented in RDkit [24]. Benchmarking studies have

found ETKDG to be the best-performing freely available conformer generator up-to-date [25,26] providing diverse and chemically-meaningful conformers reproducing crystal conformations.

The authors have called the approach MultiRef3D to emphasize that it is a fast, alignment-free multi-objective optimization protocol that maximizes the 3D overlap of a query molecule's conformational ensemble with conformational ensembles of multiple reference ligands.

Efficiency and a conformer scoring. In the algorithm, each conformation is treated as an independent entity and is characterized by a vector of features (fingerprint) which describes its 3D shape along with the distribution of electrostatic charge across its molecular surface. The generated fingerprints reflect both 3D and surface charge (both denoted further as 3D+charge) for each query molecule conformer as well as for all conformers of the reference compound. In this case, each conformer is coded within the algorithm by a single fingerprint represented as a vector of numbers which ensures computational efficiency. Those fingerprints for each of the query molecule conformers are individually scored by Euclidean distance as a similarity measure concerning each conformer of the reference compound. The Euclidean distance can be viewed as an extension of the Tanimoto similarity measure for non-binary fingerprints. The fingerprinting of individual conformers for alignment-free comparisons became popular in the past couple of years [23,27–29] so the proposed method is built on those.

Objective Function Optimization. The sum of the conformer-to-conformer similarity scores between the query and a reference compound are compared via an objective similarity function W_c for each reference compound c. The goal is to maximize the sum of those individual objective similarity functions across all reference compounds of interest $c=1,2,...,C$ where c is a summation index for the desired set of reference compounds:

$$W_{All} = \sum_{c=1}^{C} W_c = \sum_{c=1}^{C} \sum_{q=1}^{Q} \sum_{r=1}^{R} S_{q,r.}^{(c)} \quad (1)$$

In formula (1) the summand $S_{q,r}^{(c)}$ is the similarity (overlap) of the query conformer q ($q=1,2,...,Q$) with the conformer r ($r=1,2,...,R$) for each reference compound c ($c=1,2,...,C$). For the real-valued fingerprints, the similarity summand between the pair of conformers of interest indexed by query index q and reference index r for compound c is calculated as:

$$S_{q,r}^{(c)} = 1 - (1/N)\sqrt{\sum_{n=1}^{N}(x_{q,n}^{(c)} - x_{r,n}^{(c)})^2} \quad (2)$$

where $x_{q,n}^{(c)}$ and $x_{r,n}^{(c)}$ are the corresponding normalized fingerprint vector coordinates for $n=1,2,...,N$. The length (the number of coordinates) of the fingerprint N is determined based on the problem-specific target-ligand interaction characteristics. Since the fingerprint coordinates $x_{q,n}^{(c)}$ and $x_{r,n}^{(c)}$ are normalized (i.e. have values between 0 and 1 for each coordinate n) the resulting overlap $S_{q,r}^{(c)}$ is maximized with the value equal to 1 when the fingerprints of both conformers are identical and can take the smallest value equal to 0 when all the fingerprint coordinates have a difference equal to 1 i.e. as different as possible at the normalized scale.

When the objective is to identify a novel compound for just a single active conformation ($r=1$) of one ($c=1$) reference compound (e.g. a reference ligand co-crystallized with one particular target) then all conformers for the query molecule are scored against only one active reference conformer. However, in the case when multiple reference compounds are bound to the same target (or sets of reference compounds bound to multiple targets), the total objective function comes into play. It is important to point out that the proposed method is not limited to the structure-based design situations: when several reference compounds are found to be active in a functional assay (and either the target(s) is unknown or the crystal structure of the target is not available) - the formula

works just as well (as long as the ligand structure is known). The method becomes especially handy, when there is a great diversity among active reference compounds, whether the target structural information is known or not – the objective function will extract and sum up the similarities for all of the relevant parts of the fingerprinted conformer representations responsible for the observed activity.

The query compound can be evaluated against multiple reference compounds on a conformer-by-conformer basis. In such a case, the corresponding similarity scores are summed and constitute the multi-reference conformer-level objective function to maximize. This can be readily used in a typical ligand-based design setting. However, instead of just searching for a shape analog of one of the conformers of a reference compound, in the case of multiple references, the algorithm performs a search for such a compound in the virtual library whose conformers have overlapped with conformers of each of those reference compounds. The latter will increase the chances that the selected virtual compound binds the same way to the corresponding targets of each of the references (i.e. the selected compound is capable of forming conformations that resemble active conformations responsible for the Mechanism-of-Action (MoA) of each of the references).

Selection of the reference compounds and conformers. In total six diverse compounds from ClinicalTrials.gov [30] that are currently undergoing clinical trials for SARS-CoV-2 treatment have been selected as reference compounds. These compounds are diverse in the sense that their chemical structures differ substantially (i.e. Tanimoto distance < 0.7) from each other, and they all have specificity for different targets, i.e. different Mechanism-of-Action (MoA). Three of these compounds (Chloroquine [31], Remdesivir [32], and Favipiravir [33]) have recently demonstrated significant efficacy against SARS-COV-2, whereas JQ1, Apicidin, and Haloperidol are already marketed compounds well-known for their efficacy against other disease indications. One hundred

conformers for each of the reference molecules were generated at the MMFF94 level of theory [34] and each conformer was ODDT-fingerprinted [20] and saved in the MongoDB database [35]. The ODDT implementation [20] of ElectroShape fingerprints [22] has been selected to demonstrate the proposed approach because these fingerprints are considered to be state-of-the-art in ligand-based virtual screening experiments [36,37], and they are not limited to binary values.

Sometimes (e.g. at the beginning of a drug discovery program) a good virtual library might not exist yet and fingerprinting the entire ZINC15 database [38] or Enamine REAL database [39] can be computationally expensive (it may take several months to ODDT-fingerprint either of these databases on a single CPU). Therefore in this work, two simplified approaches have been used to illustrate the method. The first one is a focused virtual library approach that screened "only" 100,000 compounds from focused Enamine libraries, such as "anti-viral-like". The second one is an optimization based on a focused reference compound which is already known to be effective for the target goals (e.g., start the optimization from a SARS-COV-2 reference compound, such as chloroquine).

Focused virtual library screening. For the first approach, the virtual library (query compounds) consisted of Enamine focused virtual sets (antiviral and other specialized libraries) from the Enamine REAL database [39]. Molecules from this virtual library were simultaneously evaluated against several antiviral reference drugs with different mechanisms of action (e.g. in the SARS-COV-2 case the three major currently pursued MoA-s are: ACE2 binding, Mpro and RdRP inhibition). A query molecule for which some of the conformers are similar in shape with conformers for all the reference drugs would receive a higher score. In this approach, multiple virtual compounds can be identified to have a good conformer overlap with the reference drugs.

The proposed in silico optimization algorithm from a focused reference compound (i.e. chloroquine) is as follows:

1. Start in-silico synthesis (transformation) of any active compound in the training set by applying chemical transformation rules from chemical reaction databases that ensure tractability ("synthesizability") of each new "in-silico" compound obtained using these rules (i.e each rule is a documented feasible transformation of a particular compound class or compound substructure).
2. After each transformation, fingerprint all conformers from the obtained novel in-silico compound and calculate the total overlap score with the reference conformers.
3. Based on the value of the total overlap score:
    a. reject novel compounds with low scores.
    b. keep transforming the top N (e.g. $N = 10$) highest scored compounds.
4. Stop the optimization procedure when the top score at the current iteration is no longer higher than the top score in the previous iteration.

RESULTS

Focused virtual library screening for SARS-CoV-2 compounds. For the first approach, scoring has been performed for the six compounds that are used for SARS-CoV-2 treatment i.e. Chloroquine, Remdesivir, Favipiravir, JQ1, Apicidine, and Haloperidol. The algorithm visual summary is displayed in Fig. 1 for the W_{All} objective function. Table 1 summarizes the direct application results of the focused antiviral Enamine virtual sub-library screening. The first two

220 columns of Table 1 contain query compounds from the Enamine REAL database [39] and their
221 computed overlap scores. The rows are sorted according to the total sum overlap score displayed
222 in the second column. The first two compared virtual compounds (Virt-cpd-001 thorough Virt-
223 cpd-004) have the maximum sum of scores without "gaps", i.e. none of the compounds have a
224 score equal to zero indicating no overlap. This is different from compound Virt-cpd-014 whose
225 conformers had no overlap with any of the conformers of Chloroquine, Remdesivir, and JQ1 but
226 had a "good" overlap (in terms of the score) with conformers of Favipiravir, Apicidin, and
227 Haloperidol.

228 **Table 1**. The query compounds from a virtual library (the first column) are sorted by their total
229 overlap score W_{All} (the second column). The values in the other columns correspond to the sums
230 of the overlap scores of the conformers for the corresponding reference compounds.

231

Virtual Compound	W_{All}	**Choloroquine**	**Remdesivir**	**Favipiravir**	**JQ1**	**Apicidin**	**Haloperidol**
Virt-cpd-001	354.37	58.25	58.30	59.14	59.38	59.34	59.94
Virt-cpd-002	283.28	48.73	24.64	36.07	50.53	61.31	62
Virt-cpd-003	282.95	48.1	25.8	37.26	49.85	60.61	61.34
Virt-cpd-004	269.63	37.87	23.46	46.09	39.12	61.06	62.03
Virt-cpd-005	241.03	29.27	0	59.51	29.87	60.58	61.78
Virt-cpd-006	209.92	34.42	34.61	34.93	34.94	35.35	35.67
Virt-cpd-007	182.34	21.04	22.94	38.17	21.41	38.08	40.7
Virt-cpd-008	173.60	45.31	20.82	0	46.62	21	39.85
Virt-cpd-009	172.77	32.23	32.1	0	45.28	29.81	33.36
Virt-cpd-010	168.55	32.29	11.72	16.28	33.49	31.1	43.67
Virt-cpd-011	165.05	0	18.2	31.89	20.03	42.59	52.34
Virt-cpd-012	155.42	39.85	0	11.76	34.63	34.39	34.8

Virt-cpd-013	152.10	39.57	0	0	39.44	26.97	46.12

When interpreting the results displayed in Table 1, the individual scores for each reference compound should be taken into account along with the total score. For example, while compounds Virt-cpd-012 and Virt-cpd-013 have very similar scores, conformers for Virt-cpd-014 have no overlap with conformers of Remdesivir or Favipiravir, whereas conformers of Virt-cpd-012 have no overlap only with conformers of Remdesivir. Since the objective is to find a compound whose conformers have overlap with conformers for all of the reference compounds that have been selected for the algorithm it is concluded that Virt-cpd-013 is substantially worse than Virt-cpd-012. The other illustrative example is a total score comparison of Virt-cpd-005 vs Virt-cpd-006. Even though the score for Virt-cpd-005 is higher than that for Virt-cpd-006, none of the conformers of Virt-cpd-005 have any overlap with conformers of Remdesivir. The total score of Virt-cpd-005 is higher because the overlap of its conformers with those of Apicidin and Haloperidol is a lot higher than the corresponding overlaps for Virt-cpd-006. Therefore based on Table 1. the compounds Virt-cpd-006, Virt-cpd-007, Virt-cpd-008, Virt-cpd-011, Virt-cpd-012, and Virt-cpd-013 are not recommended for the followup in-vivo tests since those compounds have no overlap with any of the conformers for one of the reference compounds. A filter can be imposed for those virtual compounds that have zero overlap with at least one of the references since the *multi-objective* goal is to identify those virtual compounds that have the potential to exhibit *all* aspects of the therapeutic intervention for each of the reference compounds. Therefore the compounds with at least one such zero overlap are not recommended for the follow-up *in-vivo* testing experiments.

For the visual illustration of the algorithm results two compounds with the highest scores from Table 1. have been presented in Fig. 1. It is with noting that these compounds form very flexible molecules due to their either sulphonyl or ether bridge around which the ring substructures can rotate, which ensures the ability of those molecules to accommodate different targets.

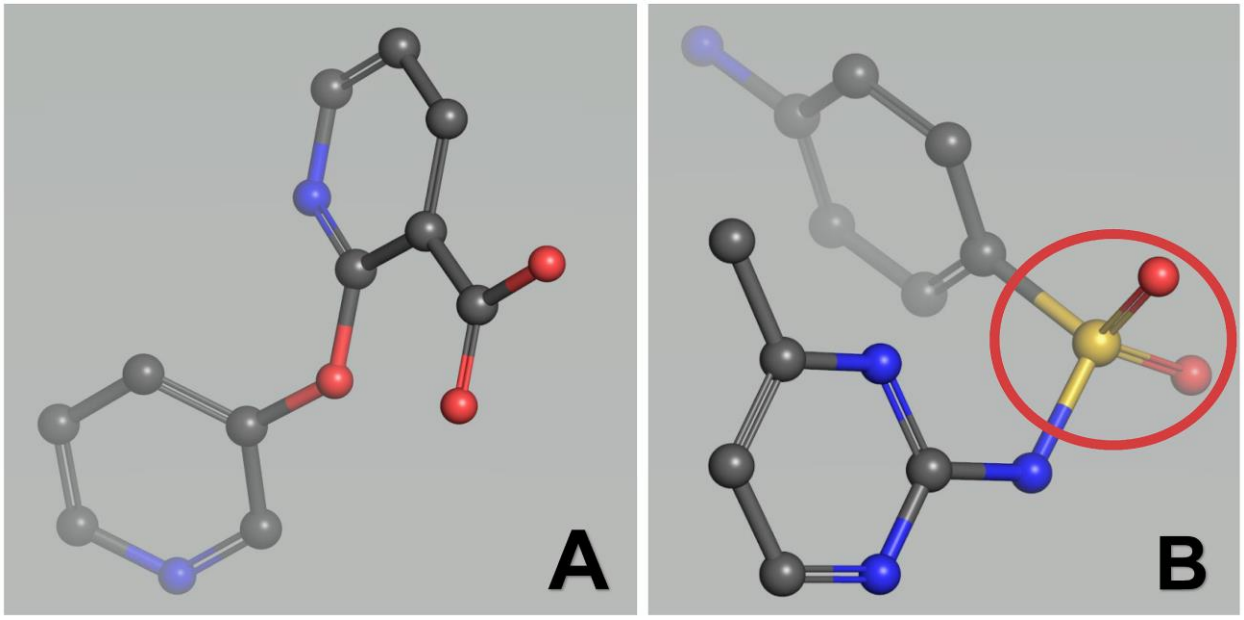

Fig 1. The compounds presented in panels **A** and **B** from the antiviral Enamine virtual sub-library collection that were found to maximize conformer overlap scores with the six reference compounds. In addition to that sulphonyl bridge in panel **B** (circled in red) is a signature of the classic antiviral compounds (e.g. well-known drug sulfapyridine), as well as the ether bond. The bridge allows for 3D flexibility for the molecule to change conformation and bind to multiple targets.

In-silico optimization of a specific drug (chloroquine). For the second approach chloroquine drug has been selected for the illustration of the optimization from a focused single reference compound. The selected chloroquine has been used as a starting point for the algorithm. For this starting point, it has been discovered that the in-silico modification/optimization procedure quickly

converges to a series of novel compounds with high scores exceeding the highest score in the original starting compound. The visual summary of the optimization results based on NOVA [40] in-silico synthesis and selection for higher overall overlap are presented in Fig 2.

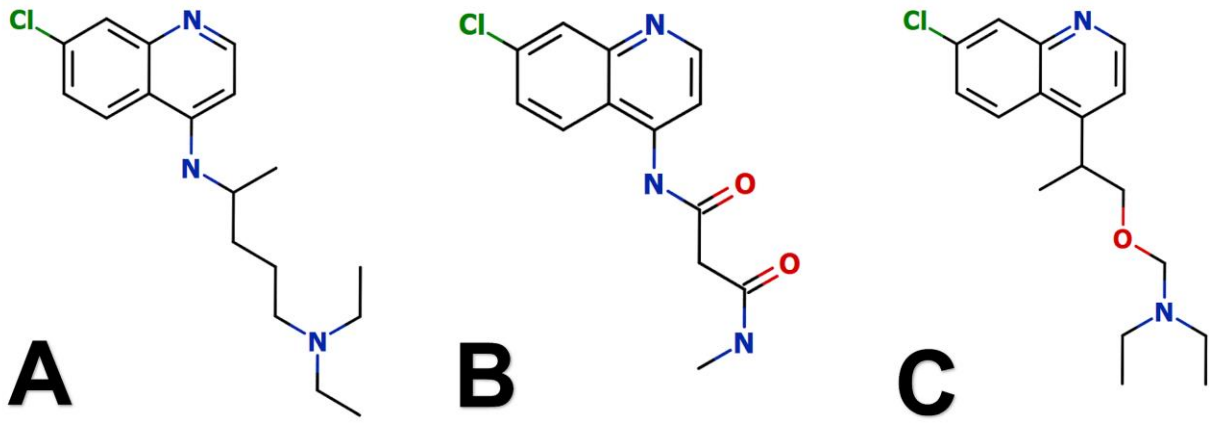

Fig 2. The original chloroquine is presented in panel **A**, while the chloroquine analogs optimized in the modify-score-select algorithm are presented in panels **B** and **C**.

In Fig. 2 the compounds in panels B and C can be visually compared to the original chloroquine structure in panel A. One can see that the two joined rings, the structural signature of chloroquine, remained in the final optimized structures untouched by the optimization process (modifications to that moiety would decrease the score and thus were rejected for the next iteration). This approach can be viewed as building and scoring virtual library on-the-fly where the seed is chosen to be one of the reference compounds. This can save a lot of compound generation scoring effort (in contrast to scoring a comprehensive 140-billion conformation) library but might not hop too far from the chemical neighborhood of the seed compound (the latter is not a drawback but rather a desired feature during lead optimization effort around e.g. a patented chemotype or well-researched chemical space in an already advanced lead optimization program).

DISCUSSION

Application for drug-repurposing. Depending on what is known about the indication or marketed drug of interest (targets, MoAs, other existing drugs for the same indication) the proposed methods (or a combination thereof) can be used to find other non-obvious molecules whose shape and the surface electrostatic charge is similar to that of the marketed drug. The methods can also be used to search for the cumulative similarity to conformers of the multiple drugs used to treat this disease indication.

Special note on the use of non-lowest energy conformations and SARS-CoV-2. Unlike what the majority of computational methods had assumed a couple of decades or so ago (e.g. in the CoMFA method [41]), recent research indicates that the bioactive conformation is not necessarily the lowest-energy conformation in the absence of the receptor [42–44]. In particular, as long as an increase in energy for less favorable conformation is compensated by its binding to the target, i.e. the total ligand-target energy is lower than the sum of the energies for the non-bound target and ligand, the bound state is favored. The proposed method emphasizes and relies on this ligand's ability to use its higher energy conformations depending on the target it attempts to bind. Moreover, in the proposed method multiple conformers of the query ligand have been compared with conformers from *multiple* reference compounds whose therapeutic effect of interest is achieved via different mechanisms of bindings to different targets, e.g. by inhibiting Main protease (Mpro) [45] and RNA-dependent RNA polymerase (RdRP) [46], while at the same time elevating pH in lysosomes to arrest the intracellular proliferation of SARS-CoV-2 [47,48]. An "ideal drug" would contain conformers that resemble (some) conformers of all of the reference drugs, thus by increasing chances that the drug inhibits SARS-CoV-2 via multi-MoA routes and is more effective than each individual reference drug.

Note on applications for structure-based designs. When the crystal structure of the target protein is known and the reference ligand is co-crystallized in its active conformation (structure-based design), we can use this information about the reference compound and evaluate the query molecules against only one, the active (co-crystallized) reference ligand conformation ($r = r_{active}$) in formulas (1) and (2). Confirmation by direct docking for the fingerprint-matched queries can be used to confirm the match.

Our methodology emphasizes pursuit of candidate compounds that achieve therapeutic effect (e.g. stops SARS-CoV-2 proliferation) by multiple MoA routes. A successful candidate compound would contain conformers targeting M-protease, RdRP and endosome virus trafficking MoAs all at the same time by increasing chances that the compound would protect against SARS-CoV-2 much more effectively. Naturally, all successful candidates would need to be further screened and filtered for proper ADME-Tox and other drug-likeness properties. Binding to anti-targets, e.g. hERG, can be explicitly incorporated to this methodology by adding the corresponding terms (similarities to known hERG-binding ligands) to the overlap sum with a negative sign. Even though many computational methods exist to evaluate hERG in particular as well as other common tox liabilities, when an anti-target is very specific and less commonly known as "pure tox target" (e.g. undesired binding to D2 receptor for many modern CNS drugs), the explicit inclusion of similarity score to such anti-target with a negative sign can greatly streamline the overall drug optimization process.

CONCLUSION

We have demonstrated the usefulness of the multi-reference optimization approach in various in-silico drug discovery settings and illustrated its application for de-novo ligand-based design, optimization, and repositioning of pharmaceutical compounds. The results allow the representation

of each molecule as an ensemble of flexible conformers that would choose the best possible conformation for each presented target-binding opportunity that can be applied in multiple settings. Application of this approach to SARS-CoV-2 produced several antiviral drug candidates that are designed to protect against SARS-CoV-2 by multiple mechanisms simultaneously.

LIST OF ABBREVIATIONS

ADME-Tox - Absorption, Distribution, Metabolism, Excretion and Toxicity

GPU - Graphics processing unit

CNS - Central nervous system

CoMFA - Comparative molecular field analysis

COVID-19 - Coronavirus Disease of 2019

CPU - Central processing unit

hERG - Human Ether-a-go-go-related Gene

MoA(s) - Mechanism of Action(s)

ODDT - Open Drug Discovery Toolkit

RNA - Ribonucleic acid

ROCS - Rapid overlay of chemical structures

SARS-CoV-2 - Severe acute respiratory syndrome coronavirus 2

WHO - World Health Organization

DECLARATIONS

Availability of data and materials

Code that has been used for analysis and for manuscript preparation can be found at Quantori public GitHub repository online [49]. Data (ligand structures) from REAL focused libraries can be downloaded from the Enamine Ltd. website [50].

Competing interests

The proposed method has been submitted for a patent. The patent application number is 63061790 at the United States Patent and Trademark Office and as of October 17, 2020, the patent is pending. The patent can be a source of financial income for authors Vadim Alexandrov (VA) and Yuriy Gankin (YG).

Funding

The author Vadim Alexandrov (VA), a founder of a consulting company Liquid Algo LLC in Hopewell Junction, New York, United States received no compensation for this work. The author Yuriy Gankin (YG) is employed by the commercial company Quantori in Cambridge, Massachusetts, United States. Alexander Kirpich (AK) is employed at the School of Public Health at a non-profit institution Georgia State University. Quantori provided support in the form of salary and relevant publication and patent fees for YG. YG received no compensation for this work. AK received no funding or any other financial support for this project.

Authors' contributions

Vadim Alexandrov (VA), Alexander Kirpich (AK), and Yuriy Gankin (YG) are the authors of the manuscript. VA and YG proposed the manuscript idea, obtained the data, implemented routine coding operations, and wrote the preliminary version of the manuscript. AK performed an additional literature review and wrote the final version of the manuscript. YG also provided the overall guidance for the project and participated in the manuscript preparation.

Acknowledgments

The authors would like to acknowledge Nika Tsutskiridze and Daviti Khatchilava for their assistance in extracting information from clinical trials and peer reviewed literature. The authors also want to acknowledge Alexander Proutsky and John Reynders for their suggestions and comments during the manuscript preparation.

Authors' information

VA holds a Ph.D. in Computational Chemistry from The University of Arizona and a Ph.D. in BioInformatics from Yale University. He has over 20 years of experience in computer-aided drug design and pioneered hit generation and lead optimization paradigm in high-throughput in-vivo phenotypic drug discovery.

AK is a biostatistician and an assistant professor in the Department of Population Health Sciences at the School of Public Health at Georgia State University. AK holds a Ph.D. in biostatistics from the University of Florida and has expertise in infectious disease statistical modeling (cholera, dengue, HIV, anthrax, and others), epidemiology, and bioinformatics. The interests and goals of AK are broad and target application of statistical methods to public health and biomedical research

391 questions and policies and to address statistical challenges, such as missing data, asymptomatic
392 infections, underreporting, and noise during data acquisition.

393 YG serves as Chief Scientific Officer at Quantori, a developer of intelligent IT and data science
394 solutions for life science, pharma, and healthcare organizations. Yuriy's expertise is in analytical
395 chemistry, data science, and chem- and bioinformatics. He holds a Ph.D. in Analytical Chemistry
396 from Tufts University and an MBA from the Massachusetts Institute of Technology Sloan School
397 of Management.

398
399
400
401
402
403
404
405
406
407
408
409
410
411
412

REFERENCES

1. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
2. Lu, H., Stratton, C. W. & Tang, Y.-W. Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. *J. Med. Virol.* **92**, 401–402 (2020).
3. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* vol. 395 497–506 (2020).
4. Chen, N. *et al.* Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet* vol. 395 507–513 (2020).
5. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-COVID-19---11-march-2020.
6. Naming the coronavirus disease (COVID-19) and the virus that causes it. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it.
7. Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C. & Di Napoli, R. Features, Evaluation, and Treatment of Coronavirus (COVID-19). in *StatPearls [Internet]* (StatPearls Publishing, 2020).
8. Sanders, J. M., Monogue, M. L., Jodlowski, T. Z. & Cutrell, J. B. Pharmacologic Treatments for Coronavirus Disease 2019 (COVID-19): A Review. *JAMA* (2020) doi:10.1001/jama.2020.6019.
9. Elmezayen, A. D., Al-Obaidi, A., Şahin, A. T. & Yelekçi, K. Drug repurposing for

coronavirus (COVID-19): screening of known drugs against coronavirus 3CL hydrolase and protease enzymes. *J. Biomol. Struct. Dyn.* 1–13 (2020).

10. Zhang, W. *et al.* The use of anti-inflammatory drugs in the treatment of people with severe coronavirus disease 2019 (COVID-19): The experience of clinical immunologists from China. *Clin. Immunol.* 108393 (2020).
11. Beck, B. R., Shin, B., Choi, Y., Park, S. & Kang, K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Comput. Struct. Biotechnol. J.* **18**, 784–790 (2020).
12. Sargiacomo, C., Sotgia, F. & Lisanti, M. P. COVID-19 and chronological aging: senolytics and other anti-aging drugs for the treatment or prevention of corona virus infection? *Aging* **12**, 6511–6517 (2020).
13. Rocs, O. E. OpenEye Scientific Software, Inc., Santa Fe, NM, USA. (2008).
14. Hawkins, P. C. D., Skillman, A. G. & Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **50**, 74–82 (2007).
15. Venhorst, J., Núñez, S., Terpstra, J. W. & Kruse, C. G. Assessment of scaffold hopping efficiency by use of molecular interaction fingerprints. *J. Med. Chem.* **51**, 3222–3229 (2008).
16. Sheridan, R. P., McGaughey, G. B. & Cornell, W. D. Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results. *J. Comput. Aided Mol. Des.* **22**, 257–265 (2008).
17. Rush, T. S., 3rd, Grant, J. A., Mosyak, L. & Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* **48**, 1489–1495 (2005).

18. Haque, I. S. & Pande, V. S. PAPER--accelerating parallel evaluations of ROCS. *J. Comput. Chem.* **31**, 117–132 (2010).
19. FastROCS. OpenEye Scientific Software, Inc., Santa Fe, NM, USA. (2011).
20. Wójcikowski, M., Kukiełka, M., Stepniewska-Dziubinska, M. M. & Siedlecki, P. Development of a protein-ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics* **35**, 1334–1341 (2019).
21. Ballester, P. J. & Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* **28**, 1711–1723 (2007).
22. Armstrong, M. S. *et al.* ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J. Comput. Aided Mol. Des.* **24**, 789–801 (2010).
23. Axen, S. D. *et al.* A Simple Representation of Three-Dimensional Molecular Structure. *J. Med. Chem.* **60**, 7393–7409 (2017).
24. Wang, S., Witek, J., Landrum, G. A. & Riniker, S. Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. *J. Chem. Inf. Model.* **60**, 2044–2058 (2020).
25. Friedrich, N.-O. *et al.* High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *J. Chem. Inf. Model.* **57**, 529–539 (2017).
26. Friedrich, N.-O. *et al.* Benchmarking Commercial Conformer Ensemble Generators. *J. Chem. Inf. Model.* **57**, 2719–2728 (2017).
27. Wójcikowski, M., Zielenkiewicz, P. & Siedlecki, P. Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. *J. Cheminform.* **7**, 26 (2015).
28. Gladysz, R. *et al.* Spectrophores as one-dimensional descriptors calculated from three-

dimensional atomic properties: applications ranging from scaffold hopping to multi-target virtual screening. *J. Cheminform.* **10**, 9 (2018).

29. Wang, Y. *et al.* TF3P: Three-Dimensional Force Fields Fingerprint Learned by Deep Capsular Network. *J. Chem. Inf. Model.* **60**, 2754–2765 (2020).
30. ClinicalTrials.gov. http://ClinicalTrials.gov.
31. Chen, X. & Geiger, J. D. Janus sword actions of chloroquine and hydroxychloroquine against COVID-19. *Cell. Signal.* **73**, 109706 (2020).
32. Frediansyah, A., Nainu, F., Dhama, K., Mudatsir, M. & Harapan, H. Remdesivir and its antiviral activity against COVID-19: A systematic review. *Clin Epidemiol Glob Health* (2020) doi:10.1016/j.cegh.2020.07.011.
33. Agrawal, U., Raju, R. & Udwadia, Z. F. Favipiravir: A new and emerging antiviral option in COVID-19. *Armed Forces Med. J. India* **76**, 370–376 (2020).
34. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* vol. 17 490–519 (1996).
35. MongoDB Atlas Database. https://www.mongodb.com.
36. Cortés-Cabrera, A., Morris, G. M., Finn, P. W., Morreale, A. & Gago, F. Comparison of ultra-fast 2D and 3D ligand and target descriptors for side effect prediction and network analysis in polypharmacology. *British Journal of Pharmacology* vol. 170 557–567 (2013).
37. Bonanno, E. & Ebejer, J.-P. Applying Machine Learning to Ultrafast Shape Recognition in Ligand-Based Virtual Screening. *Frontiers in Pharmacology* vol. 10 (2020).
38. http://zinc15.docking.org/. http://zinc15.docking.org/.
39. REAL Database - Enamine. https://enamine.net/library-synthesis/real-compounds/real-database.

40. Optibrium. Optibrium - StarDrop: Nova - A new generation of possibilities. https://www.optibrium.com/stardrop/stardrop-nova.php.
41. Gohda, K., Mori, I., Ohta, D. & Kikuchi, T. 10.1023/A:1008193217627. *Journal of Computer-Aided Molecular Design* vol. 14 265–275 (2000).
42. Mackerell, A. D., Jr. Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.* **25**, 1584–1604 (2004).
43. Hasegawa, K., Arakawa, M. & Funatsu, K. Rational choice of bioactive conformations through use of conformation analysis and 3-way partial least squares modeling. *Chemometrics Intellig. Lab. Syst.* **50**, 253–261 (2000).
44. Acharya, C., Coop, A., Polli, J. E. & Mackerell, A. D., Jr. Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. *Curr. Comput. Aided Drug Des.* **7**, 10–22 (2011).
45. Ullrich, S. & Nitsche, C. The SARS-CoV-2 main protease as drug target. *Bioorganic & Medicinal Chemistry Letters* vol. 30 127377 (2020).
46. Elfiky, A. A. SARS-CoV-2 RNA dependent RNA polymerase (RdRp) targeting: an perspective. *J. Biomol. Struct. Dyn.* 1–9 (2020).
47. Li, W. *et al.* Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* vol. 426 450–454 (2003).
48. Vincent, M. J. *et al.* Chloroquine is a potent inhibitor of SARS coronavirus infection and spread. *Virol. J.* **2**, 69 (2005).
49. Website. *github.com - Quantori repository* https://github.com/quantori/MultiRef3D.
50. Targeted Libraries. https://enamine.net/hit-finding/focused-libraries/.

Figures

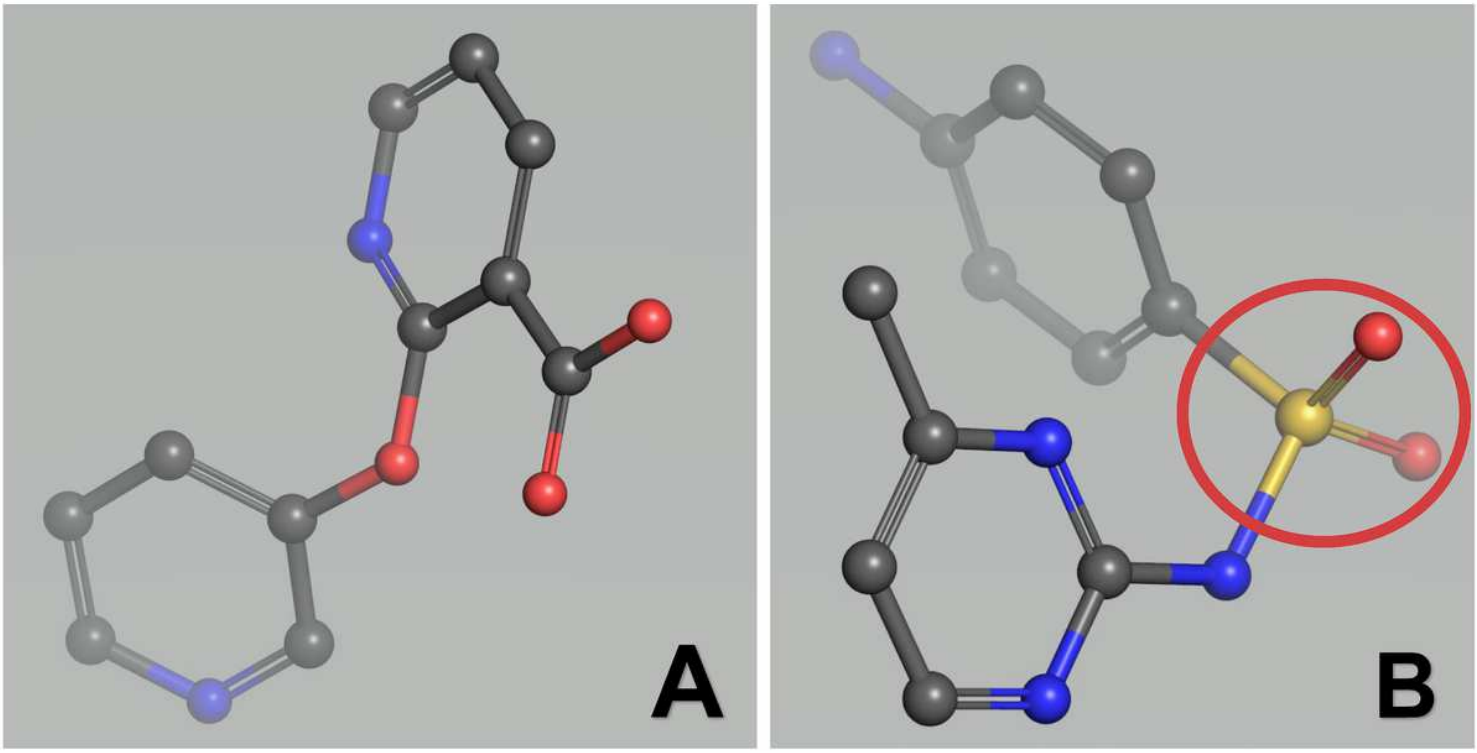

Figure 1

The compounds presented in panels A and B from the antiviral Enamine virtual sub-library collection that were found to maximize conformer overlap scores with the six reference compounds. In addition to that sulphonyl bridge in panel B (circled in red) is a signature of the classic antiviral compounds (e.g. well-known drug sulfapyridine), as well as the ether bond. The bridge allows for 3D flexibility for the molecule to change conformation and bind to multiple targets.

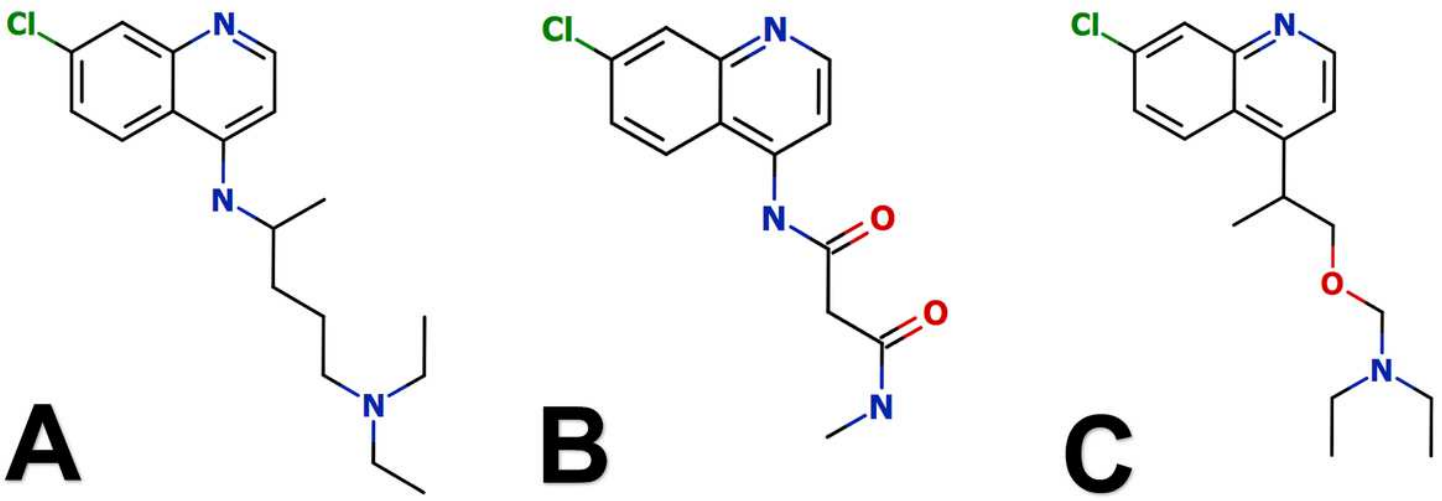

Figure 2

The original chloroquine is presented in panel A, while the chloroquine analogs optimized in the modify-score-select algorithm are presented in panels B and C.