Star-shaped molecules cross-bind SARS-CoV2 spike α-helices *in silico*

short title: star-shaped stabilization of coronavirus
Coll, J*.
Department of Biotechnology. Instituto Nacional de Investigacion y Tecnología Agraria y Alimentaria, INIA. Madrid, Spain.

* Corresponding author
Email: juliocollm@gmail.com (JC)
Julio Coll, orcid: 0000-0001-8496-3493

Abstract

Despite docking to the isolated α-helix residues 960-1010 ("spring-loaded switch-folding", SLSF) of wild-type S spike trimers of Severe Acute Respiratory Syndrome coronavirus (SARS)-CoV2, the star-shaped hydrophobic Tinosorb failed to dock to SLSF inside the S (S-SLSF) and to inhibit viral-host cell membrane fusion[1]. This work discovered computational star-shaped-similar molecules exhibiting lower binding-scores (higher-affinities) to S-SLSF with probable cross-binding properties of their targeted α-helices but with lower hydrophobicities and smaller molecular sizes. Most star-shaped-similar leads contained Trihydroxyl-Triphenyls arms branching from each of the three carbons of a central Triazine core (TTT). Deconstruction of TTTs by core-replacement (X), fragment extension (F), and 2D deep-screening among millions of molecular possibilities, found additional leads that by combining structural features (F+TTX) reduced their binding-scores to S-SLSF. Such leads maximize their possibilities to stabilize wild-type S-SLSF α-helices, with the aim to reduce host-coronavirus membrane fusion using drug-like ligands rather than mutations.

Keywords: S spike; star-shaped molecules; triazine core; prefusion; coronavirus; deep-learning; SARS CoV2; spring-loaded switch-folding

Introduction

Tinosorb was previously identified as the computational lead for binding to the spring-loaded switch-folding (SLSF) of wild-type S spikes of SARS-CoV2[1]. SLSF was previously defined as the S spike sequence that expands from the 960 to the 1010 amino acid residues forming 3x3 α-helices in the trimer wild-type prefusion states (Figure 1, Bottom). Computational screening of ~130000 natural compounds predicted lower leads to wild-type SLSF trimers rather than to monomers and to other PP mutated conformers. Preferential binding to the SLSF trimers in the low nM range, its star-shaped 3-fold symmetric molecular structure[2] and its fitting to the inner part of the 3x3 SLSF α-helices, together with the existence of side and top-bottom surface-accessible cavities in the prefusion S wild-type trimer, suggested that Tinosorb may show binding and binding-dependent biological activity. Hypothetically, Tinosorb's cross-binding of the inner space of the S-SLSF 3x3 α-helices could stabilize its prefusion states to inhibit fusion, similarly to PP mutations [3-5]. However, Tinosorb's binding to S-SLSF (SLSF inside the whole S trimer) could not be demonstrated and *in vitro* assays using S pseudotyped VSV-infectivity, did not inhibit fusion[1]. Tinosorb's relatively large size compared to the narrow surface-accessibility to S-SLSF together with its low solubility in aqueous media, respectively, may have contributed to such failures. This work computationally explores, Tinosorb-like star-shaped molecules among less hydrophobic molecular alternatives for improved binding to S-SLSF. Because whole S trimers may be the best target to stabilize prefusion virions or to interact with the earliest steps of viral infection, the S-SLSF wild-type trimer model (6xr8 ID in the RCSB protein data bank) was selected for this work.

The previously defined amino acid sequence of SLSF (residues 960-1010) contained part of the HR1 heptad-repeat (910 to 988) and part of the CH central helix (986-1033) of the S2 subunit of the S spike of SARS-CoV2. In the S wild-type prefusion conformation, the S-SLSF α-helices are central to the inner space of the S trimers but could be surface-accessible to small molecules through cavities of ~ 7-20 Å of diameter at the side and top-bottom axis[1].

The SLSF amino acid sequence monomer contains two amino-terminal, small and aligned α-helices separated by an small loop, and a larger α-helix separated by a 975-987 loop in an elbow-like folded spring-loaded 3x3 α-helices conformation (Figure 1, Bottom). Before viral-cellular membrane fusion, the folded spring-loaded mechanism unfolds and the 3x3 α-helixes elongate to 3x1 α-helices in the trimers. After elongation, one coiled-coil bundle involving the newly formed 3 HR1-CH and other 3 HR2 helices, originates the linearly rigid trimer conformation typical of active fusion and postfusion states. Similar spring-loaded mechanisms are common to many other enveloped viruses [6, 7]. In coronaviruses and other viruses, mutations to prolines in the SLSF folded loop stabilize the S trimers at their prefusion states resulting in inactivation of viral-host membrane fusion and inhibition of infectivity [3-5].

Most previous and abundant experimental and computational search for prediction of anti-coronavirus activity has been focused on approved drugs (drug repurposing)[8, 9] to protein targets on either cell hosts or coronavirus. For instance, most recent experimental screening for intracellular inhibitors of coronavirus-infected drug-treated cells identified 90 compounds with EC_{50} < 96000 nM [10] among ~ 12000 drugs from the Repurposing, Focused Rescue and Accelerated MedChem (ReFRAME) bank (https://reframedb.org). Additive interactions enhancing the inhibition of viral RNA-dependent RNA polymerase (RdRp) by the nucleoside analogues remdesivir (EC_{50} of 123 nM) or amilimod (11 nM) were experimentally demonstrated for several drugs. On the other hand, computational work on coronavirus targets have been mainly focused on the HR1-HR2 helices bundle on the S spike [12, 13, 14,15, 16] , the S1 spike surface interphase with the ACE2 human receptor [11] , and/or the active sites of RdRp and of viral proteases [9, 17, 18]. Although targeting the HR1-HR2 helices bundle with synthetic peptides, only reported modest inhibition of infectivity[12, 13, 14,15, 16], searching for more potent small ligands among star-shaped molecules targeting a possible stabilization of their S-SLSF 3x3 α-helices rather than by mutations may be justified by both experimental and computational screens.

Here, several highly imaginative strategies previously developed by others[19] have been used for computational screens, such as similar searches, *de novo* generation of compounds with drug-like properties [20-23], available fragment-libraries to chemically fine-tune identified leads [24-27], machine learning involving convolutional neural networks (CNN) using molecular 2D images as inputs[28], and filtering for synthetic feasibility or for presence in catalog/approved (purchasable/repositioning) drugs [25, 29, 30]. To explore star-shaped molecule alternatives for the possible opportunities raised by Tinosorb, a variety of such methods have been combined by following an step-by-step strategy.

In the work to be described here, Tinosorb-similars binding to SLSF, identified a common Triazine core branched in its carbons by hydroxyl-phenyls (Trihydroxyl-Triphenyl-Triazine, TTT). Deconstruction of TTT unrevealed a few core alternatives (TTX) with similar and sometimes lower binding-scores to S-SLSF. Further explorations by phenyl fragment extension (F+TTX), discover more hydrophilic leads that combined small fragments, additional cores and different phenyl positions to maintain and/or to improve binding to S-SLSF at the low nM range. All the above mentioned methods together with our own previous data involving 3D docking[1] allowed the training of high-accuracy deep-learning models using 2D images of molecules as inputs [28]. The derived convolutional neural networks (CNN) were then optimized to screen a few larger libraries of compounds such as those recently designed for maximal chemotype diversity among the purchasable chemical space[23], and the last release available of the CHEMBL large chemical data collection. The 3D-docking screening for the deep-learning proposed candidates, added a few more leads to the list.

Taken together the results predicted ~ 50 leads with greater drug-like characteristics and possibilities to stabilize S-SLSF than the initial Tinosorb molecule. Although more computational candidates may still be found by exploring more, larger chemical spaces or newly designed libraries focused in star-shaped molecules, whether the leads already identified would experimentally bind to S-SLSFs and inhibit viral fusion remains yet to be demonstrated.

Materials and Methods

Molecular characteristics of Tinosorb

Tinosorb (PubMed ID 135487856, CHEMBL 2104956, CAS 187393-00-6), also called bemotrizinol (2,2'-[6-(4-methoxyphenyl)-1,3,5-Triazine-2,4-diyl]bis{5-[(2-ethylhexyl)oxy]phenol}) or bis-ethylhexythexyloxyphenol-methoxyphenyl-Triazine, has a molecular weight of 627.8 Dalton (see Figure S1). Tinosorb adsorbs ultraviolet UV-A and B from 280 to 400 nm preventing the formation of free radicals induced in the skin by sun exposure reducing tissue oxidation. Due to its low water solubility (0.33 µg/ml, logP of 10.4), Tinosorb is used in topical creams in oiled mixtures. Tinosorb has no estrogenic, nor androgenic effects. Its oral/dermal toxicities are estimated to LD_{50} >2g/Kg.

Libraries of possible ligands used in this work

Tinosorb-similar 1746 molecules were downloaded from PubChem https://pubchem.ncbi.nlm.nih.gov/#query =*smiles*=similarity). Trihydroxyl-Triphenyl-Triazine (135616181 ID), and Triazine (9262 ID)-similars downloaded from PubChem resulted in 599 and 279689 molecules, respectively. The number of Triazine-similars were downsized to 4346 molecules <700 Dalton and < 6 logP.

Core-replacement screening was performed on each of the seeSAR's 1 Gb zipped fragment libraries (pdb and zinc data bases containing ~20 million of fragments each). Fragment extension screened the seeSAR's library of 100 small fragments enriched with home-designed 10 fragments of 5-7 non-hydrogen atoms.

The SuperNatural II SNII library (http://bioinf-applied.charite.de /supernatural_new/index.php) was splitted in nineteen *.sdf files each containing different molecular weight ranges from 16 to 380 Daltons as described before [1]. The splitted sdf files were randomly sampled to supply 10 high-binding-score inactive or negative ligands per file to contribute to the design of one training-set to train 2D deep-learning models. A 0.5 million library of maximized chemotype diversity among the purchasable space[23] and the ~ 2 million Chembl28 latest release library, filtered between 250-500 Daltons to ~1.5 million compound (http://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/ releases/chembl_28/), were screened by the DEEPscreen[28] T13 developed model.

BioSolveIt infiniSee extremely large libraries (CHEMriya_11bn_2021-05 $1.1x10^{10}$ compounds , GalaXi_2.1bn_2020-11 $2.1x10^{9}$, KnpwledgeSpace_290tr_2019-05 $2.9x10^{14}$, and REALSpace_19bn_2021-04 $1.9x10^{10}$) were screened by the infiniSee program for 1000 or 10000 compounds with 0.9-1.0 target similarity and 80-90% minimum similarity thresholds to TTT.

Compounds available in commercial catalogs were searched through the ZINC data base by supplying their smiles (http://zinc15.docking.org/) and/or searched in building-block catalogs (Sigma, BLDpharm). Duplicates of drugs retrieved from several sources were eliminated by OpenBabel (https://sourceforge.net/projects/openbabel/postdownload)[1].

3D trimeric wild-type 6xr8 S spike model

To explore virtual binding of ligands to the prefusion state of wild-type SARS-COV-2, the S spring-loaded switch-folding (SLSF) expanding S amino acid residues from 960 to 1100 in each monomer, was extracted from the trimer model and used as previously defined[1]. The whole trimeric S 6xr8 molecule model (**R**esearch **C**ollaboratory for **S**tructural **B**ioinformatics, RCSB, **P**rotein **D**ata **B**ank PDB ID) of the wild-type closed, all-down S conformer was used to target S-SLSF with the aid of an i9 computer with 48 CPUs[1].

3D-docking screening by two algorithms

Two different and complementary algorithms (AutoDockVina and seeSAR), were employed for 3D-docking. The programs differed on both, **i)** generation of ligand conformations (also known as poses) with high probability of binding to the target binding-pocket (set at 10 per ligand) and **ii)** quantification by binding-score estimations of each conformation or pose. The two programs set the target protein binding-pocket as rigid (maintaining constant covalent lengths and angles) and the ligands as flexible (using their rotatable bonds to generate different conformations) [1, 31].

The AutoDockVina included in the PyRx 0.9.8. package [32](https://pyrx.sourceforge.io/) uses multithreading on multi-core e7 / i9 computers [33] to speed up docking. To generate possible bound conformations (poses), AutoDockVina uses Lamarckian genetic algorithms [34]. The corresponding conformation-dependent Gibbs free-energies (ΔG) are calculated using semi-empirical data [35]. To perform the docking, the *.sdf files were first ffu energy minimized and charges added to obtain *.pdbqt files (PyRx-Bable program). Grids including only the SLSF inner 3x3 α-helices were used. Only the pose with the lowest binding-score (*.out.pdbqt) was retained for analysis. To compare with seeSAR values, the output ΔG energies in kcal/mol were converted to constant inhibition (Ki) in nM concentrations, as described before [1, 31].

To predict possible seeSAR poses, the vs.10 package (https://www.biosolveit.de/SeeSAR/) uses the FLEXx incremental fragment construction method based on software developed for computer vision and pattern recognition [36].The corresponding conformation-dependent Gibbs free-energies (ΔG) are calculated by the HYDE scoring function computing HYdration and DEsolvation values (as calibrated with octanol/water partition data, logP) [37, 38]. To reduce false positives, the HYDE calculations include not only favorable but also unfavorable interactions [39]. To perform dockings, the unique binding-pocket internal to the SLSF α-helices or all the 36 binding-pockets (average of 17 amino acids per pocket) predicted by the seeSAR in the whole S trimer (Figure 3,C) were selected. Only the pose with the lowest binding-score was selected for analysis. The corresponding binding-scores were expressed in the mean predicted value calculated from the HYDE lower-higher nM estimates (100-fold range) delivered by the program.

To facilitate interpretation, the order by binding-score profiles were first analysed graphically using the Origin program (OriginPro 2015, 64 bit Sr1 b9.2.257, Northampton, MA, USA). Binding-score estimations differed <10 % when repeated in different tests (n=3-5). The predicted ligand-protein complexes were visualized in PyRx, seeSAR and/or PyMOL (https://www.pymol.org/).

Core-replacement

The Triazine core or each of the Trihydroxyl-Triphenyl groups of TTTs were selected (Figure S2, AB) to carry out replacements using the seeSAR inspirator module to screen tenths of million (~1 Gb zipped) each of the provided pdb and zinc fragment libraries. Each iteration of the core-replacement feature selects for the best-similar 10 cores that maintain the rest of the molecule intact while docking to SLSF. Up to 90 best-fitting new cores were retrieved from each library and their corresponding 3D-binding-scores to SLSF estimated by seeSAR docking. The resulting SLSF leads were finally docked to S-SLSF.

Fragment extension

The seeSAR fragment extension feature (Figure S2, C) was applied using the sdf file provided by seeSAR containing 100 small molecular weight fragments. Additionally, 10 home-made fragments between 1 to 7 non-hydrogen atoms designed in MarvinSketch 17.1.30.0 (Chemaxon, Oracle Co) were also included. To explore each of the carbons at the hydroxyl-phenyl groups, cores were fixed while F+TTT compounds having fragment extensions (Figure S2) at each of the hydroxyl-phenyl **C1**s (**RED**, **GREEN**, **BLUE**) were generated by the program by docking to SLSF. To explore **C1** bound fragments, additional F+TTT compounds containing one, two or three SLSF best-binding fragments in different non-symmetric **C1** positions for each fragment were manually generated in MolSoft (Molbrowser vs3.9-1bWin64bit). No fragment combinations with different fragments at each **C1** position were generated for docking because of their high number of possibilities. All the resulting SLSF leads were 3D-docked to S-SLSF.

Preparation of training sets for machine learning

Our previous library of thousands of 3D-docked compounds separated in 19 different molecular weight files[1, 31] were randomly sampled for 10 binding-score inactives or negatives >100 nM for each molecular weight file (under sampling the majoritarian class)[40]. Other 30 TTT-similars > 100 nM were added. The 48 F+TTX of < 0.2 nM binding-score were used as actives. To balance the final training-set (1 for actives and 0 for inactives), the randomized inactives were pooled in a 4:1 ratio with the 48 actives. To provide a common identifier to best compare previous and present docking results with large target libraries, *inchikeys* were calculated and added to the training-set as molecular_names using the DataWarrior program (Osiris, vs 5.5.0. Idorsia Pharmaceuticals Ltd). Possible duplicates were eliminated by OpenBabel. For training the model, the final sdf file was randomly splitted in 60 % for training, 20 % for validation and 20 % for test.

Learning models using chemical fingerprints as inputs

All chemical fingerprint types were obtained from the PADEL descriptor tool (http://padel.nus.edu.sg/software/ padeldescriptor). The resulting PADEL files were splitted into molecular descriptors and algorithm-specific fingerprint bins for comparative tests. Regression prediction models such as AdaBoost, Tree, NeuralNetwork, and others provided by the Orange vs3.27.1 package (http://www.ailab.si/orange) resulted in accuracies ~ 85 %. However, their predictions on ligands which were never-seen-before were < 20 % (not shown).

Convolutional neural networks (CNN) deep-learning models using 2D-molecular images as inputs

The recently released DEEPScreen software [28], using input molecular 2D-images rather than chemical fingerprints for higher prediction success, was adapted to the present purpose. The final strategy was applied in three steps: **i)** 3D-docking to obtain enough negative and positive binding-scores to train a CNN T13 model, **ii)** 2D-deep-learning applying the CNN T13 model to downsize large libraries to a lower number of docking-candidates, and **iii)** final 3D-docking of identified candidates for validation. The final CNN T13 model optimizing the parameters provided by DEEPScreen and their predictions of never-see-before ligands were performed using python and pytorch home-made scripts. The accuracy of the T13 CNN model was 97.4 % as estimated using the whole training-set as input.

Results

Docking of Tinosorb-similars to SLSF

Thousands of Tinosorb-similars were downloaded from PubMed and docked by seeSAR or AutoDockVina to isolated SLSF sequences (S residues 960 to 1010 extracted from the 6xr8 wild-type trimer conformer). Results predicted 10^2-10^5-fold lower binding-scores when using seeSAR compared to AutoDockVina (Figure 1 top, red hexagons *versus* blue hexagons).

Visual comparative observations of the best-conformational poses suggested that those obtained by seeSAR were interdigitated within the 3x3 α-helices (Figure 1 Down **C**,**D**). In contrast, preferential interactions with the sides of two adjacent α-helices rather than with three, were obtained by AutoDockVina (Figure 1 Down **A**,**B**). These observations may offer some explanation for the differences in binding-scores between the 2 programs. Together, these data suggested that to find out new candidates targeting the inner space or binding-site of the SLSF 3x3 α-helices, seeSAR was the best option. Therefore, seeSAR was chosen for the rest of the work.

The majority of the Tinosorb-similar leads contained chemotypes having a central Triazine core (N at 1,3 and 5 positions), with 3 hydroxyl-phenyl groups linked to the core carbons (such structures have been called star-shaped molecules) [2]. Additional structural variations in the leads consisted in different fragments linked to the **C1** of each phenyl group forming star-shaped molecules > 620 Daltons (F+TTT, Fragment-Trihydroxyl-Triphenyl-Triazine). However, most of these newly identified leads were of higher hydrophobicities and/or molecular weights than Tinosorb (see some selected examples in Table 1). Thus, although a total of 57 F+TTT Tinosorb-similars (~ 3 % of the initial Tinosorb-similars) predicted lower binding-scores than Tinosorb (Figure 1 up, **red large hexagon**), only one had low, drug-like logP solubility (Table1).

To further understand the binding requirements of the identified star-shaped molecules, several deconstructions of the F+TTT were then explored targeting wild-type S-SLSF whole S trimers rather than isolated SLSF trimers.

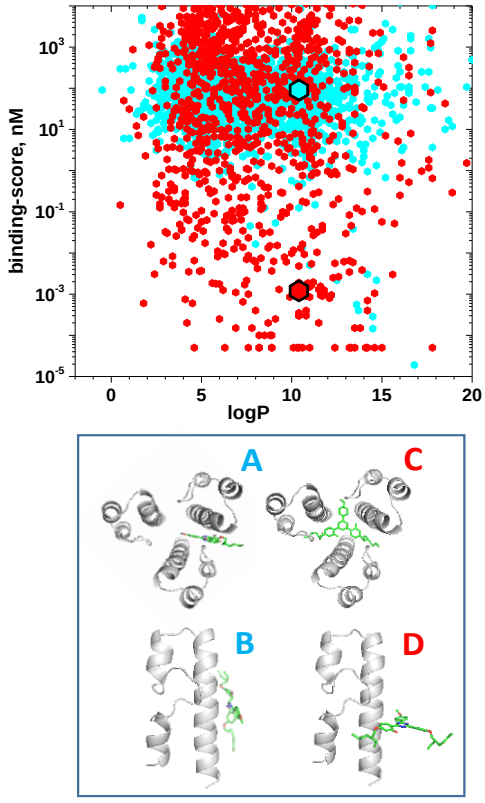

Figure 1
TOP) Binding-scores of Tinosorb-similars *versus* their logP estimations. BOTTOM) best conformational poses of Tinosorb-bound to SLSF by AutoDockVina (A,B) and seeSAR (C,D)

TOP

Binding-scores and logP of 1746 PubMed-downloaded Tinosorb-similars docked to isolated SLSF 6xr8.
Blue hexagons, AutoDockVina. **Red hexagons**, seeSAR.
Black-edged blue hexagon, Tinosorb by AutoDockVina.
Black-edged red hexagon, Tinosorb by seeSAR.

BOTTOM

Green, Tinosorb.
Gray, SLSF 960-1010 amino acid sequence ribbons
A, Top-view of trimeric SLSF 3x3 α-helices complexed to Tinosorb by AutoDockVina.
B, Side-view of one SLSF complexed to Tinosorb by AutoDockVina. Two monomers were removed for clarity.
C, Top-view of trimeric SLSF 3x3 α-helices complexed to Tinosorb by seeSAR.
D, Side-view of one SLSF complexed to Tinosorb by seeSAR. Two monomers were removed for clarity.

Table 1
Leads from Tinosorb-similars docked to SLSF showing different fragments linked to C1 phenyls

PubChem, ID	Binding-score,nM	logP	MW	Smiles	Trihydroxyl-Triphenyl – Triazine (TTT)
136025237	0.00005	4.4	621.7	CCC(CO-C1	
155024065	0.00005	6.2	621.7	CCCOCOC1=CC(=C(C=C1)	
135783913	0.00005	7.9	609.7	C1CCC(-C1)	
135611720	0.00005	8.2	615.8	CCC(C)CO-C1	
136417298	0.00005	8.2	615.8	CC(C)CCO-C1	
136044044	0.00005	8.3	623.7	C1CCC(C-C1)	
135740105	0.00005	8.7	747.9	CCCCOCCCCO-C1	
136383973	0.00005	8.7	773.9	CC(C)CCCCCOC(=O)C(C)O-C1	
136058049	0.00005	9.0	773.9	CCCCCCCCOC(=O)C(C)O-C1	
149408938	0.00005	9.1	643.8	CCCCC(CC)CO-C1	
148346816	0.00005	9.5	755.9	CCCCC(CC)C(=O)O-C1	
136030929	0.00005	9.6	627.8	CCC-C1	
136467720	0.00005	11.0	699.9	CCCCCCCO-C1	
137127598	0.00005	11.6	712.0	CCCCC(C)CO-C1	
142723568	0.00005	12.0	.790.0	CCCCCCCCOO-C1	
.....	.....	.....	.....	.....	
135487856	0.00100	10.4	627.8	*CCCCC(CC)CO-C1	

Examples from the 57 leads from Tinosorb-similars docked to SLSF ordered by their logP. **135487856**, Tinosorb (58th of the leads). The same fragments were attached to the **C1** carbons at each of the Trihydroxyl-Triphenyls bound to the Triazine core (F+TTT, star-shaped molecules with 3-fold symmetries). *, Tinosorb's one of the fragments is chemically different (Figure S1).
Binding-score, mean of the seeSAR's estimations in nM.
LogP, PubChem estimation of hydrophobicity (molecular partition ratio between water and octanol).
MW, molecular weight in Daltons.
Smiles, fragment formula expressed in Smiles (Simplified Molecular Input Line Entry Systems).

Docking of TTT-similars to SLSF and S-SLSF

The binding-scores of the PubMed similars to the TTT chemotype (without any fragments) were compared using both SLSF and S-SLSF targets. The docking results confirmed some lead requirements for the TTT chemotype and revealed the existence of a relative high number of F+TTT leads with lower binding-scores to S-SLSF than to SLSF (Figure 2), suggesting that additional interactions with other residues within the S trimer may contribute to reduce their binding-scores. Most of the TTT leads showed fragments attached to the **C1** carbons symmetrically located in front of the C4-C bond between phenyls and their Triazine-cores. Other were asymmetrical.

Although there were not many new ligands identified by applying this search method, in this case these experiments discover that the lowest binding-score was predicted to the TTT chemotype without any fragments. Although TTT predicted a higher binding-score than Tinosorb-SLSF, it showed lower molecular size, lower hydrophobicity and symmetry, improving its drug-like properties (Figure 2). There were two other leads which suggested there may exist more variations such as having only 2 hydroxyls in the Triphenyl groups and/or two N in the core.

Visual inspection of the TTT leads bound to S-SLSF confirmed that most of them docked inside the SLSF 3x3 α-helices (Figure 3), in contrast to Tinosorb [1]. Further exploration of the TTT+S-SLSF complex showed that their interactions implicated the T^{998} and Q^{1002} residues from the 3x3 α-helices by forming hydrogen bridges with the Trihydroxyls. Most of the rest of the TTT atoms including those of the Triphenyls and their attached fragments contributed to the final binding-score only by favoring desolvation (according to seeSAR structural analysis confirmed in CCP4) (not shown).

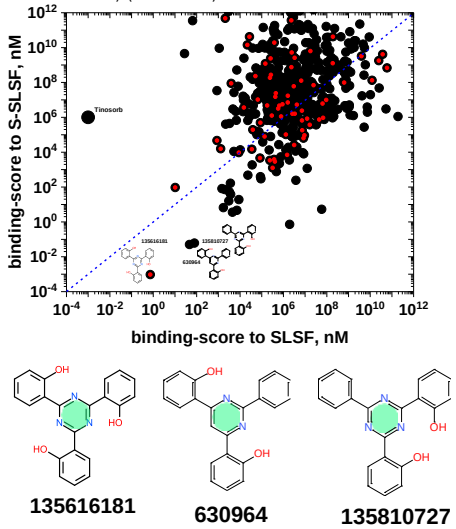

Figure 2
Binding-scores of TTT-similars to SLSF and S-SLSF

Downloaded PubMed 599 TTT-similars were docked to SLSF and S-SLSF trimers.
Black circles, asymmetrical molecules.
Red circles, 3-fold symmetrical molecules.
Black numbers, PubChem IDs.
Blue hatched line, equal SLSF and S-SLSF binding-scores. The binding-score of Tinosorb was included here for comparison but its binding to S-SLSF was outside the 3x3 α-helices[1]
Down, 2D structures of some leads.

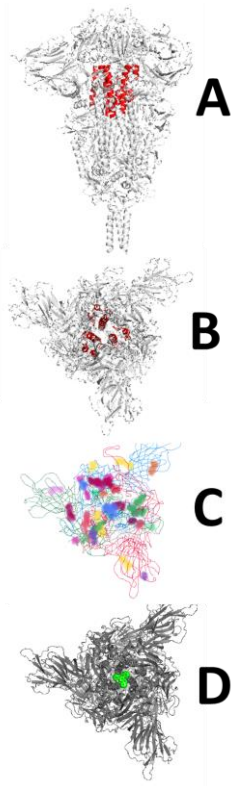

Figure 3
Mapping of the S-SLSF sequences at the S trimer (A,B), S binding-pockets of the whole trimers predicted by seeSAR (C) and mapping of S-SLSF+ TTT complexes (D)

A, side view of wild-type S 6xr8 trimer and location of the 3x3 α-helices of the S-SLSF sequences (**red**).
Gray lines, rest of amino acid residues of the S trimer

B, top view of wild-type S 6xr8 trimer and central location of its 3x3 α-helices of the S-SLSF sequences showing the central cavity of 7-20 Å wide (**red**).
Gray lines, rest of the amino acid residues of the S trimer

C, top view of the binding-pockets predicted by seeSAR in the exposed surface of the wild-type S 6xr8 trimer in several background colors. **RED, GREEN , BLUE lines**, rest of amino acid residues. Each S monomer was colored by seeSAR in a different color.

D, Mapping of the TTT docked to S-SLSF. **Green**, TTT docked to the wild-type S 6xr8 trimer in competition with the 36 binding-pockets schematized in C.
Gray lines, amino acids residues of the S trimer

Triazine-core replacement

Since the Triazine core with Nitrogen at positions 1,3,5, was one of the features of most leads, attempts were made to explore binding to SLSF and/or S-SLSF trimers of any similars. Searching PubMed for Triazine-similars yielded an excessive number of 279689 molecules. Therefore, those were downsized to 4346 molecules applying drug-like criteria. However, when docked to SLSF, none of them predicted any binding-scores lower or similar to those of TTT (not shown).

An alternative way to search for Triazine-core alternatives, was the seeSAR core-replacement feature which maintains the rest of the Trihydroxyl-Triphenyl groups while substituting cores for binding to SLSF (Figure S2). Docking results from millions of possibilities identified 4 new cores (TTX chemotypes) with similar binding-scores to SLSF than TTT. Binding-scores were reduced ~50 fold when docked to S-SLSF (~10^{-2} to 10^{-4} nM ranges, respectively), suggesting again the participation of other residues outside the SLSF sequences. These new X cores have only one or two N rather than the 3 N present in TTT. Two of them (N1 and N35) were identified in both libraries used for the core-replacement (Figure 4).

These results suggested new molecules containing other than Triazine-cores (TTX) and new fragments attached to them which may result in lower binding-scores to S-SLSF.

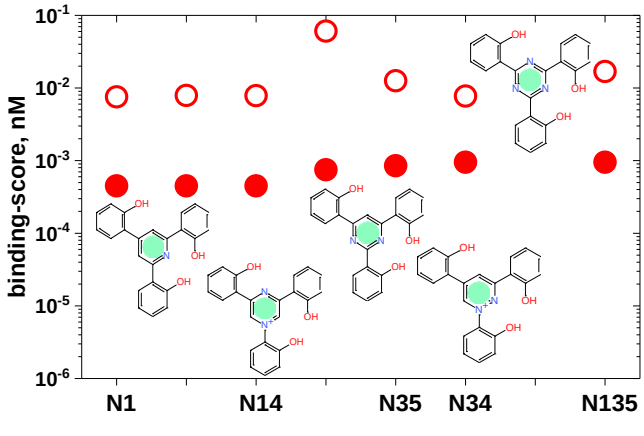

Figure 4
Triazine-core replacement (TTX) binding-scores to SLSF and S-SLSF
The N135 was replaced by X-cores using the core-replacement feature of seeSAR which screens the pdb / zinc fragment-libraries of tenths of millions each (Figure S2). **Y-axis**, binding-scores of best pose of the newly generated TTX core-replaced molecules. **X-axis,** leads with cores labeled by N followed by their position in the 6 atom cores. The numbers correspond to the pbd and zinc fragment IDs were as follows N1 (3qx502P1H5I, zinc263631 fragments), N14 (zinc1593398 fragments), N35 (3bhh5CP1B600F, zinc8300484 fragments), N34 (zinc1564326 fragment) and N135 (PubChem 135616181, Figure 2).
Open red circles, TTX binding-scores to SLSF.
Solid red circles, TTX binding-scores to S-SLSF.

Fragment search by similars and extension

Because the use of the "build evolutionary library" option of DataWarrior could not generate any lower binding-score alternatives for the 5 new cores, alternative methods were explored.

Trihydroxyl-Triphenyl groups of Tinosorb and of TTTs were tilted to each other around their rotable C4-C Triazine bonds when docked to SLSF (Figure S1). Therefore there were 5 C positions for each phenyl group to which fragments could be added. To study whether or not there were any differences among the Triphenyls, fragment-extension to each of them (labeled as **RED**, **GREEN** and **BLUE**, according to their default colors in seeSAR), were independently studied (Figure S2). Only the C4-C Triazine bonds were maintained intact while the rest of the C1-C6 positions including the C3 hydroxyl position, were targeted for possible fragment extension. Binding of the resulting F+TTT molecules were then evaluated by docking to SLSF. Docking results showed that the fragments attached to **C1** yielded the lowest binding-scores (Figure 5), confirming previous observations made on Tinosorb, Tinosorb-similars and TTT-similars. Therefore, to further explore binding of F+TTXs to S-SLSF, only their **C1** positions were targeted.

A first search for any possible **C1**-fragments among PubMed only found similars for N1, N35 and N135 cores but none of those displayed drug-like characteristics (i.e., high hydrophobicities).

As an alternative, the fragment extension feature of seeSAR was then applied to the 5 new cores testing each of their Trihydroxyl-Triphenyls for 110 fragment extensions (a library of fragments provided by seeSAR and enriched by 10 home-made designs). One-by-one fragment extensions were made at each of the **C1s** (**RED**, **GREEN**, **BLUE**) of the TTX Trihydroxyl-Triphenyl bonds. The resulting F+TTX molecules were then docked to SLSF (Table S2). Docking results showed that most, but not all, lead fragments bound to the three RGB **C1** positions with similar binding-scores (Figure S3, red bars), and that 1, 2 or 3 fragments resulting in any symmetric or asymmetric structures were among the possible leads. Those fragments/positions which showed binding-scores to SLSF < 0.2 nM (Table S1) were selected to construct the corresponding F+TTX complete molecules for docking to S-SLSF. In this case, the S-SLSF dockings together with the previous results, identified dozens of leads (see Figure 5). Since in contrast to Tinosorb, those new lead logP solubilities were between 3-6, steric inhibition rather than hydrophobicity could partially explain extreme differences in binding-scores when changing targets from SLSF to S-SLSF. For instance, some of the the larger or charged fragments that generated leads to SLSF, resulted in too high binding-scores when docked to S-SLSF (>10^6 nM), similarly to what occurred with Tinosorb[1].

Together, the new F+TTXs leads obtained from fragment extension showed a group which docked to S-SLSF on the ~ 0.001 nM range while a second group gathered around the ~0.1 nM range (Figure 6). Up to 43.7 % of the identified leads were molecules found in PubMed while the rest corresponded to newly described molecules, mainly identified by fragment extension. One of the lowest binding-scores, consisted in a new chemotype substituting the Trihydroxyls by Trimethoxyl groups, resulting in a Trimethoxyl-Triphenyl-Pyrimidine molecule. However, further search did not identified any lower binding-score among its 510 PubMed similars, nor fragment extension attempts corresponding to its N35 core identified any lower or similar binding-scores to S-SLSF (not shown).

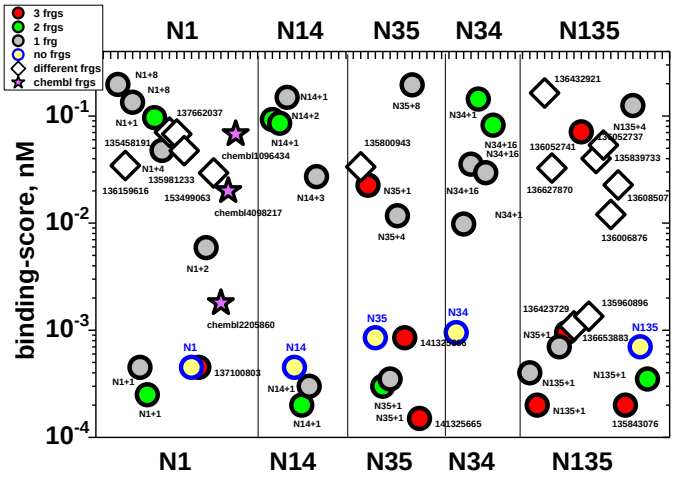

Figure 5
Leads to S-SLSF
The leads were grouped by the core (Figure 4)+ the ID number of the fragments (Table S3).
Red circles, same fragments in the 3 **C1s**.
Green circles, same fragments in 2 **C1s**.
Gray circles, fragment in 1 **C1**.
Blue-edged yellow circles, original N1, N14, N35, N34, N135 cores (Figure 4).
Open diamonds, heterogeneous fragment combinations (PubMed IDs-similars).
Purple stars, heterogeneous fragment combinations (CHEMBL IDs predicted by the CNN T13 model).

Screening large libraries by a CNN T13 deep-learning classifying model

To select for the corresponding leads with higher probabilities to bind to S-SLSF, a newly developed deep-learning CNN model called T13 was trained using 2D molecular images (learning rate followed in Figure S5).The T13 predicted candidates were then 3D-docked to S-SLSF.

A library of computationally synthetic 500000 compounds designed to cover a maximum of purchasable chemotypes was screened by T13. Results of the model predicted 105 possible candidates. However, their lowest binding-scores when 3D-docked to S-SLSF were in the high nM ranges. Therefore no new leads were obtained from this library.

A library of ~ 2 million compounds downloaded from the last Chembl28 release and downsized to ~1.5 million drug-like ligands, was screened by the CNN T13 deep-learning model using 2D molecular images. Results of the model predicted 8751 possible candidates. To downsize that large number, 3D-docking was first made to SLSF. The resulting 34 candidates predicting < 1 nM binding-scores to SLSF were finally 3D-docked to S-SLSF and 4 new leads identified. Since one of the leads was previously identified, the other 3 leads were added to the final proposed lead list (Figure 5, **purple stars**).

Screening larger libraries by infiniSee

To explore wider chemical spaces, the BioSolveIt's infiniSee program was used to screen 4 libraries containing 10^9-10^4 compounds each for TTT-similars. However such screening attempts, did not identified any TTT-similars with 3D binding-scores to S-SLSF <0.2 nM among the best 1000 similar compounds in any of the target similarities or thresholds studied (binding-scores >3-3.8 nM). Additional attempts targeting 10000 compounds were also not successful to identify any lower S-SLSF leads (not shown).

Drug-like properties of the leads to S-SLSF

The corresponding *in silico* pharmacokinetic parameters, physicochemical and toxicity ADME predicted to leads (Figure 5), showed that many of them were moderately soluble, complied with Lipinski rules and have gastrointestinal good permeability predictions (Table S2).

Binding leads to computationally PP-mutated S trimers

Preliminary results indicated that some of the leads predicted large differences when docked to other SLSF or S-SLSF representative conformers selected from our previous study[1] (data not shown).

Because any amino acid sequence differences among those conformers and wild-type 6xr8 were only due to their PP mutations, the same mutations were computationally introduced into the amino acid sequence of wild-type 6xr8 S-SLSF. The corresponding 3D trimer model derived from the 6xr8 PP mutated sequence was then docked to the leads. Restricting the analysis to lead poses predicting binding-scores < 0.2 nM, the results showed that an estimated 30.9 % of the 6xr8 lead poses altered their binding-scores by the introduction of the PP mutations into its sequence (Figure 6). In some leads those differences were of several orders of magnitude corresponding to poses that did not cross-bind the 3x3 α-helices of S-SLSF (not shown).

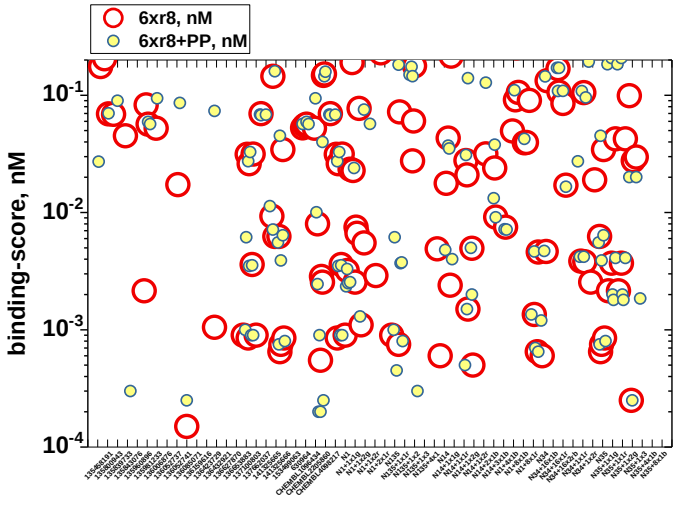

Figure 6
Comparison of binding-scores between leads from wild-type and PP mutated 6xr8 S-SLSF trimers
The wild-type sequence of 6xr8 S was computationally mutated to the previously reported $P^{986}P^{987}$ amino acid positions stabilizing the trimer at the prefusion states and inhibiting the coronavirus infection[3-5]. The corresponding 3D models were build with the Swiss model server and wild-type and PP mutant 6xr8 S-SLSF docked to the F+TTX leads of Figure 5. Ten poses per lead were obtained and those with binding-scores < 0.2 nM represented.
X-axis, F+TTX leads labeled as in Figure 5.
Red circles, binding-scores of wild type 6xr8 S-SLSF.
Yellow circles, binding-scores of PP mutant 6xr8 S-SLSF.

Discussion

Binding-scores of natural compounds to isolated S residues 960-1010 (SLSF) were lower to trimers than to monomers and to wild-type rather than to PP-mutated conformers[1]. The predicted lead was Tinosorb with binding-scores at the very low nM range (0.003 nM)[1]. Tinosorb is a highly hydrophobic molecule with a 3-fold star-shaped architecture containing two phenyls and one methoxyl groups linked to the carbons of one Triazine central core. Docked Tinosorb fitted the inner empty space inside the SLSF 3x3 α-helices. However, Tinosorb failed to bind to S-SLSF and to inhibit S-pseudotyped VSV-infection, raising the question of whether or not other star-shaped molecules could be found for those purposes. A first computationally exploration to such question was attempted here.

Docking of thousands of Tinosorb-similars to SLSF resulted in the identification of ~ 50 leads with lower binding-scores. However, all those leads except one, had higher hydrophobicity and larger sizes than Tinosorb and therefore they may have reduced possibilities to bind to S-SLSF. Nevertheless, many of them showed common architectures including identical fragments linked to the all **C1** carbons at their Trihydroxyl-Triphenyl groups (F+TTT). The unique SLSF lead identified with the lowest hydrophobicity, contained a fragment of only 4 carbons and one oxygen. The above commented results suggested further searches for more leads to SLSF and S-SLSF among TTT-similars (without any fragments). The new docking results confirmed that the simplest TTT architecture also predicted lower hydrophobicities, size and binding-scores to both SLSF and S-SLSF, prompting us for additional explorations searching for new leads using core-replacement and/or fragment extension. The application of both computational techniques resulted in the identification of dozens of star-shaped alternatives that including short fragments (F+TTX) predicted leads with more hydrophilic lower binding-scores to SLSF and reducing 2-8-fold those to S-SLSF. Most probably, the lower binding-scores to S-SLSF could be explained by their fragment interactions with other residues outside the S-SLSF sequences, such as Y^{756} and F^{759} , in addition to those between the Trihydroxyls and S-SLSF T^{998} and Q^{1002} (identified by visual seeSAR inspection of the bound complexes). Among all the identified leads there were structures with a few new cores cotaining N and 1 to 3 small fragments bound to the **C1s** of their Trihydroxyl-Triphenyl groups.

In the previously reported docking of hundred of thousands of natural compounds of <380 Daltons to the trimer or monomer SLSFs from 9 conformers, nor Tinosorb, nor TTT-similar leads were detected[1]. Because most SLSF or S-SLSF leads described here had higher molecular weights, it is probable that their lower size could explain such failures. Perhaps, further docking screenings including higher molecular size candidates may have found other star-shaped molecules among the natural compounds. However, that exploration would require much more intensive computation since their relative abundances in several chemical banks was very low. Nevertheless, the leads or active (positives) and not-binding or inactive (negatives) ligands identified in this and previous work[1], respectively, were successfully used to design training-sets to optimize predictive deep-learning CNN models. The optimal T13 deep-learning model feeding on 2D molecular images and trained by the compounds identified as above, successfully identified new leads in a short time, been capable of detecting one of our previously identified leads among millions of never-seen-before compounds. However, only 4 star-shaped new structures could be detected by T13 among ~ 2.5 million compounds. Also, the screening with infiniSee of much larger libraries (10^9-10^{14} compounds) although detected many TTT-similars, did not contributed any other leads to S-SLSF. Further explorations of wider chemical spaces using the above mentioned T13 model may be tried in the future. However, although more leads may be reasonably expected, perhaps with some new chemotypes and higher hydrophilicity, it is possible that the abundance of star-shaped compounds in actual chemical banks may be limited. Perhaps to fully explore all possibilities, new computationally libraries of "synthetic" compounds potentially targeting S-SLSF may be required. On the other hand, the star-shaped molecules which virtually bound to S-SLSF were best detected when using work-intensive core-replacement or fragment-extension methods. Most probably, further automatizing those search methods by new algorithm designs to target new chemical data bases, would contribute to discover new star-shaped molecules targeting the 3x3 α-helices of S-SLSF. Since α-helices are present in many coiled-coils that participate in important biological interactions, such methods may be of interest to other fields.

Identifying the star-shaped lead molecules would have been more difficult using AutoDockVina or similar algorithms. In this particular case, only the seeSAR algorithm detected the star-shaped structures fitting the inner part of the 3x3 SLSF α-helices. Opposite examples in which AutoDockVina performed better than seeSAR were also found, such as to detect anti-VKORC1 molecules to identify possible rodenticides [41] or to predict graphene interactions with detoxifying enzymes [42]. It seems likely that the best fitting algorithms may depend not only of using the most adequate data bases but also of each particular docking problem.

One of the challenges for successful experimental prediction of leads targeting S-SLSF is how accessible is the SLSF 3x3 α-helices in the wild-type coronavirus particle. The partial accessibility to S-SLSF predicted by modeling the wild-type closed all-down S trimer, suggests that fixed S-SLSF may be reached

even at that highly-compacted conformation. Theoretically, experimental accession should be possible for any leads with the lowest binding-scores and smaller molecular sizes, provided they could be efficiently water-solubilized. At this respect, the conformer-dependent binding-score lead variations despite differing only in their PP mutations but with similar 3D structures (low RMSDs), was again remarkable. Apparently, most of the observed differences were highly dependent on small 3D variations on their inner 3x3 α-helices. Thus superposition of 6xr8 and 6xr8+PP SLSF trimers predicted that an small total widening (~ 1 Å) of the internal space between one or several of the 3x3 α-helices in the PP mutant compared to the wild-type conformer. That could be enough to explain the increasing binding-scores of 3-4 orders of magnitude. Seldom addressed, because of the wide acceptance of the PP mutations to stabilize the S protein to develop vaccines, these observations agreed with those made in crystal structures in other coronaviruses, which predicted small but significant changes in the inner trimer 3x3 α-helix S-SLSF in their PP mutants [7]. However, there have been few studies comparing wild-type and PP-mutated spike conformations taking into account the dinamic nature of the corresponding conformations[3].

Although the more relaxed open (RBD-up) S structures together with the flexible nature of the protein conformations, suggested an experimental increase of accessibility to S-SLSF, preliminary docking results to PP-mutated but fixed conformers, suggested that the inner empty space between their S-SLSF 3x3 α-helices were widened enough in the PP-mutants, which will result in an apparent increase of the binding-scores of some wild-type leads. To our knowledge, the absence of 3D structures of wild-type S open conformations, do not allow yet to make any accurate docking predictions with fixed target 3D protein targets. A possible alternative will be to introduce molecular dynamic procedures, which will best mimic possible conformational variations on the S-SLSF during the docking proccesses, however computational costs for studying the whole S-SLSF trimer would be very high. Another possible alternative to explore, would be to computationally revert the PP mutations to the ^{986}KV wild-type sequence. Before interpreting those possible results, experimental binding and fusion inhibition, would need to be investigated.

Visualization of the docked complexes to S-SLSF predicted that interactions with each of the 3x3 α-helices may slightly differ among leads. Thus, for most leads, the main contributions to the final binding-scores, implicated hydrogen bridges between the Trihydroxyls of the Triphenyl groups and the three T^{998} and Q^{1002} residues of the S trimer. The total binding-score estimations by seeSAR implicated lead-dependent combinations of the above mentioned binding residues of the S-SLSF α-helices and of others located inside or outside of S-SLSF. There were none or few alternative fragments detected which could replace the contribution of the Trihydroxyl groups attached to the C5 carbon of the phenyl groups, confirming their importance in lead binding. The small distance-requirement for such hydrogen bridges may also explain why any small displacement of the α-helixes, like those previously described for crystalized structures of PP mutants [3-5], could increase binding-scores. On the other hand, the other atoms which either belong to the Triphenyls or to different fragment structures, only slightly contributed to the final binding-score calculations. Their small per atom contributions were due to favor desolvation (displacement of water molecules). Although weak, these individual-atom interactions, taken together made important contributions to the final computation of each lead-dependent binding-score. Most fragment alternatives containing either positive or negative charges generated unfavorable interactions, increasing their corresponding estimations of binding-scores, and suggesting that charges were not involved in start-shaped lead bindings to S-SLSF. Taken all the above ideas together, we may conclude that these star-shaped F+TTX leads may be highly specific and unique for the wild-type 6xr8 conformers. It is probable that any of the 1, 2 or 3 receptor-binding domains (RBD)-up conformations of the prefusion S trimers would require other leads to virtually cross-bind their 3x3 S-SLSF α-helixes.

The identified leads were more hydrophilic, of smaller size and maintained their binding-scores low compared to Tinosorb. However, some of these leads may still be difficult to dissolve in water to be tested *in vitro* and there are some that may be inhibitors of important detoxifying cytochromes, which it is usually interpreted as physiologically problematic for *in vivo* drug-like purposes. Therefore, these practical aspects should be also considered when evaluating any possible validation of these compounds. Assays such as testing by experimental binding of leads to isolated recombinant S, inhibition of S pseudotyped VSV fusion (as described before[1]) and/or possible blocking of coronavirus cellular infection, could be employed to validate *in vitro* some of the proposed lead predictions. Considering that those new F+TTX leads with 3-fold symmetry will be most favorable for chemical synthesis, those may be preferred for initial experimental tests. Additionally, both from a physiological point of view and because computational data indicated that the smaller the size, the lower the binding-score to S-SLSF, the lead candidates or other still to be found with simpler chemical structures may be preferable for *in vitro* and *in vivo* testing.

Supporting information

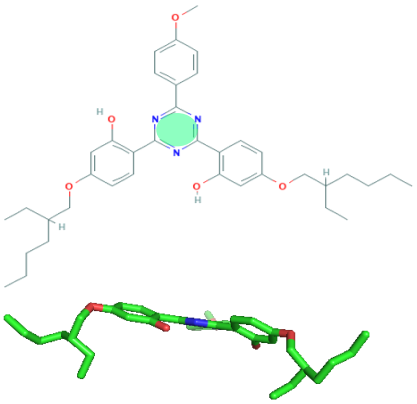

Figure S1
Molecular 2D and 3D molecular structures of Tinosorb
A. 2D representation of Tinosorb (bis-ethylhexythexyloxyphenol-methoxyphenyl-Triazine)
B. 3D representation of Tinosorb best conformational pose when docked to SLSF

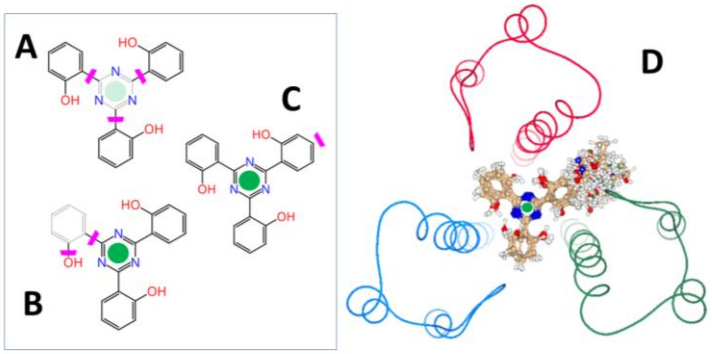

Figure S2
TTT core- / phenyl-replacements and hydroxyl-phenyl C1 fragment extensions
Core- and phenyl-replacements (A,B) were screened among ~40 millions of fragments of seeSAR's supplied libraries. Hydroxyl-phenyl **C1** fragment extensions (C) were screened among the 100 low molecular weight fragment library provided by seeSAR and enriched with 10 home-made fragments. Each of the resulting new F-TTT molecules were then docked to SLSF.
A, core-replacements in **gray** and the rest of the maintained structures in **black**, separated by the **pink** insertion locations.
B, phenyl-replacements in **gray** and the rest of the maintained structures in **black**, separated by the **pink** insertion locations.
C, hydroxyl-phenyl **C1** extension separated by the **pink** insertion location.
D, example of merged fragments extended in position **C1** at the **RED** hydroxyl-phenyl.

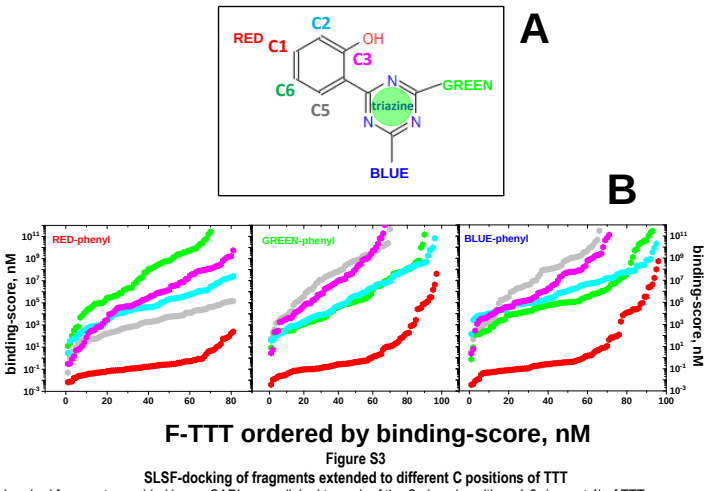

Figure S3
SLSF-docking of fragments extended to different C positions of TTT
One hundred fragments provided by seeSAR's were linked to each of the C phenyl positions 1-6 (except 4) of TTT. **RED**, **GREEN** and **BLUE** hydroxyl-phenyls were independently extended. Binding of the resulting new F+TTTs were then estimated by docking to SLSF.
A, scheme of the **TTT** molecular structure, with only the **RED** hydroxyl-phenyl drawn (**GREEN** and **BLUE** hydroxyl-phenyls were omitted in the TTT structure for clarity).
B, Binding-score profiles to SLSF of the resulting F+TTT complexes.
Red hexagons, **C1**.
Blue hexagons, C2.
Purple hexagons, C3.
Gray hexagons, C5.
Green hexagons, C6.

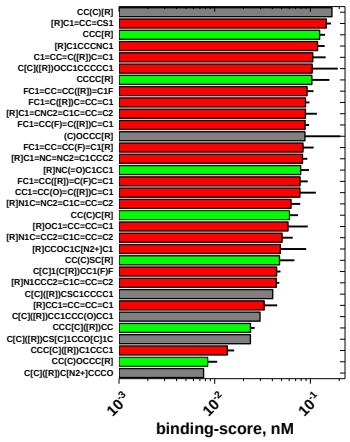

Figure S4

Mean RGB binding-scores to SLSF of the C1 fragment extension leads at F+TTTs

The fragments represented in their smiles formula were those selected among the seeSAR's fragment extension leads having binding-scores < 0.2 nM. Mean ± sd were calculated from the binding-scores of the 3 RGB hydroxyl-phenyl groups at each of the F+TTT complexes.

Y-axis, fragment smiles formula where [R] indicates the covalent bond to **C1**.

Gray bars, bound only to one of the RGB. **Red bars**, fragments with > 7 non-hydrogen atoms.

Green bars, fragments with < 7 non-hydrogen atoms.

Table S1

SLSF binding-scores of cores + fragment extensions

			phenyl RED	GREEN	BLUE
core		C1-fragment	nM	nM	nM
N1	N1+	[R]			
	N1+1	C[R]			
	N1+2	CC[R]	0.14	0.11	0.11
	N1+3	CCC[R]	0.05	0.03	0.05
	N1+4	CCCC[R]	0.07	0.03	0.05
	N1+5	CC(C)SC[R]	0.02	0.01	0.01
	N1+6	CC(C)C[R]	0.02	0.03	0.01
	N1+7	CC(C)OCCC[R]	0.01	0.02	0.07
	N1+8	(C)OCCC[R]	0.00	0.23	0.01
	N1+9	CS[R]	0.09	0.12	0.08
	N1+10	CC(O)[R]	0.13	**1.30**	0.09
	N1+11	COCN[R]	0.03	0.09	**7.63**
	N1+12	CCC[C]([R])CC	0.11	0.01	0.01
N14	N14+	[R]			
	N14+1	C[R]			
	N14+2	CC[R]	0.11	0.07	0.10
	N14+3	CCC[R]	0.06	0.03	0.02
	N14+4	CCCC[R]	0.09	0.02	0.02
	N14+5	CC(C)SC[R]	0.05	0.01	0.04
	N14+6	CC(C)C[R]	0.14	0.01	0.03
	N14+9	CS[R]	0.12	0.07	0.10
	N14+13	CC(C)[R]	**260.74**	0.04	0.07
	N14+14	COC([R])=O	0.12	0.08	0.08
	N14+15	C1COCCC1[R]	0.91	0.03	0.09
N35	N35+0	[R]			
	N35+1	C[R]			
	N35+3	CCC[R]	0.41	0.07	0.08
	N35+4	CCCC[R]	0.12	0.06	0.03
	N35+5	CC(C)SC[R]	0.01	0.01	0.01
	N35+6	CC(C)C[R]	0.05	0.05	0.05
	N35+7	CC(C)OCCC[R]	0.00	0.00	0.01
	N35+8	(C)OCCC[R]	0.01	0.01	0.05
	N35+12	CCC[C]([R])CC	0.01	0.01	0.01
N34	N34+	[R]			
	N34+1	C[R]			
	N34+13	CC(C)[R]	**33.70**	0.12	0.06
	N34+16	CO[R]	1.34	1.34	0.17
N135	N135+	[R]			
	N135+1	C[R]	0.38	0.40	0.45
	N135+3	CCC[R]	0.12	0.14	0.10
	N135+4	CCCC[R]	0.16	0.09	0.05
	N135+5	CC(C)SC[R]	0.06	0.02	0.04
	N135+6	CC(C)C[R]	0.04	0.05	0.07
	N135+7	CC(C)OCCC[R]	0.01	0.01	0.01
	N135+8	(C)OCCC[R]	0.01	0.22	0.02
	N135+12	CCC[C]([R])CC	0.02	0.02	0.02

F+TTX structures were generated by the seeSAR fragment extension of hydroxyl-phenyls **C1s** in 5 different cores. The resulting structures were labeled by an N followed by their N positions in the core as in Figure 4. The selected fragments were arbitrarily numbered from +1 to +16 (i.e., N34+13). Those fragments represented by individual [R] corresponded to the initial core-replacement molecules of Figure 4. The represented fragment binding leads to SLSF were defined as those F+TTX structures with predicting binding-scores < 0.2 nM. The tabulated fragments were then computationally drawn as been bound to their corresponding **C1** positions on the final F+TTX molecules to be tested by docking to S-SLSF (final results in Figure 5). Some of the minimal fragments C[R] were rejected by the fragment extension program, but included in the final F+TTX reconstructed molecules for docking to S-SLSF. **Numbers in red**, fragments with a high nM binding-score having low probability of fitting the SLSF and/or the S-SLSF inner 3x3 α-helices.

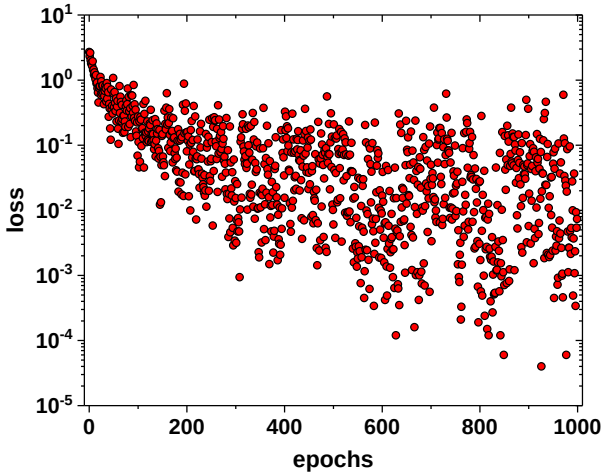

Figure S5

Learning curve of the DEEPScreen CNN T13 model developed for high-throughput screening of binding candidates among large libraries

The training-set contained 48 F+TTX as positives (1) (Figure 5) and randomized/size-selected 30 TTT-similars + 162 SNII as negatives (0). To train the T13 model, the resulting training-set of 240 classified compounds was randomized and splitted in 60 % for training, 20 % for validation and 20 % for test. The DEEPScreen convolutional neural network (CNN) T13 model prediction performance were true for 80.0 % of the 48 positives and for 97.4 % of all the 240 classified compounds of the training-set.

Loss, mean differences between the T13 model predictions and their classifications.

Epochs, number of forward/backward iterations through all training-set in our DEEPScreen CNN T13.

Table S2

Lead drug-like characteristics predicted by the SwissADME web server

ID or Name	nM	MW	Log P	sol	GIA	1A2	2C19	2C9	2D6	3A4	LipK	PAINS	SynAcc
141325665	0.0001	398.5	4.6										3.1
135843076	0.0002	399.4	4.2										2.7
N14+1x2	0.0002	385.4	3.0										3.2
N135+1x3	0.0002	399.4	4.2										2.7
N1+1x2	0.0002	383.4	4.6										3.0
N14+1x1	0.0003	371.4	2.6										3.1
N35+1x2	0.0003	384.4	4.2										3.0
N135+1x2	0.0003	385.4	3.9										2.6
N35+1x1	0.0003	370.4	3.9										2.9
N135+1x1	0.0004	371.4	3.5										2.5
137100803	0.0004	355.4	4.0										2.8
N14	0.0004	357.4	2.3										3.0
N1	0.0004	355.4	4.0										2.8
N1+1x1	0.0004	369.4	4.4										2.9
N135	0.0007	357.4	3.2										2.4
N35+1x1	0.0007	370.4	3.9										2.9
141325666	0.0008	356.4	3.7										2.8
N35	0.0008	356.4	3.7										2.8
136653883	0.0009	357.4	3.2										2.4
N34	0.0009	357.4	2.7										3.0
136423729	0.0011	415.4	3.5										2.9
135960896	0.0013	413.5	4.6										2.9
chembl2205860	0.0018	444.4	5.5										2.9
N1+2x1	0.0058	383.4	4.7										3.0
N34+1x1	0.0098	371.4	3.1										3.2
N35+4x1	0.0117	412.5	4.9										3.2
136006876	0.0120	383.4	4.6										2.7
chembl4098217	0.0205	373.8	5.0										2.8
135839733	0.0227	355.4	4.0										2.5
N14+3x1	0.0271	399.5	3.3										3.3
153499063	0.0294	323.4	4.8										2.6
N34+16x1	0.0295	387.4	2.8										3.2
N35+1x3	0.0325	398.5	4.6										3.1
136627870	0.0327	427.5	4.9										3.1
135800943	0.0335	340.4	3.9										2.7
136159616	0.0344	339.4	4.4										2.8
N34+16x1	0.0354	387.4	2.7										3.2
136085071	0.0400	355.4	4.0										2.5
135981233	0.0474	339.4	4.4										2.7
N1+4x1	0.0476	411.5	5.4										3.2
630964	0.0505	340.4	4.0										2.7
136052741	0.0535	401.5	4.4										3.9
137662037	0.0678	373.8	5.0										2.8
chembl1096434	0.0681	339.4	4.4										2.7
135458191	0.0706	353.4	4.7										2.8
136052737	0.0710	401.5	4.1										3.9
N34+16x2	0.0820	417.4	2.7										3.4
N14+1x2	0.0855	385.4	3.0										3.2
N14+2x1	0.0926	385.4	3.0										3.1
N1+1x2	0.0965	383.4	4.7										3.0
N135+4x1	0.1250	413.5	4.6										2.9
N1+8x1	0.1343	427.5	4.7										3.2
N34+1x2	0.1437	385.4	3.4										3.3
N14+1x1	0.1491	371.4	2.7										3.1
136432921	0.1641	423.4	5.0										2.8
N35+8x1	0.1948	428.5	4.2										3.2
N1+8x1	0.1969	427.5	4.7										3.2

Lead numbers, PubMed IDs

Lead chembl number, CHEMBL IDs

Lead N numbers, core (Figure 4) + fragment number (Table S1) x 1,2 or 3 fragments .

LIPK, number of violations of Lipinski rules that would make the ligand less likely to be an orally administrable drug if >5.

LogP, consensus value of multiple predictions of lipophilicity.

1A2, 2C19, 2C9, 2D6, 3A4, in green the leads predicted to inhibit the main detoxyfying cytochromes P450.

GIA, in green predictions of high gastro-intestinal adsorption.

PAIN, Pan Assay Interference Structures (PAINS), alerting of the number of chemical fragments that return false positives in virtual binding.

Green, favorable

Light green, moderately favorable.

Empty in white backgrounds, unfavorable.

SynAcc, synthetic difficulties calculated by fragmentation of the leads and ranged from 1-10 (the highest the most difficult to synthesize)

Funding
The work was carried out without any external financial contribution

Competing interests
The author declares no competing interests

Authors' contributions
JC designed, performed and analyzed the dockings and deep-screen, and drafted the manuscript.

Acknowledgements
Thanks are specially due to Dr. Judd Dunkan of Awridian Ltd at United Kingdom by kindly providing the 500k synthetic chemotype-maximized purchasable library, to Dr. Tunca Doğan of the University of Ankara at Turkey for his help to understand his simple and powerful DEEPScreen program, to Dr. Markus Gastreich of BioSolveIT GmbH at Germany for his zoom-help with the inspirator features of seeSAR. Thanks are also due to Dr. Alberto Villena from the University of Leon (Spain), Luis Maestre (telecommunication engineer) from Madrid and to Dr. Ignacio Garcia from the Hospital Gomez Ulla of Madrid (Spain) for their help with the bibliography, to Dra.Maria M.Lorenzo for her preliminary tests and to Dr. Rafael Blasco at INIA (Madrid, Spain) for his original ideas and discussions.

References

[1]Coll, J. Would it be possible to stabilize prefusion SARS-COV-2 spikes with ligands? *ChemRxiv.* 2020, https://doi.org/10.26434/chemrxiv.13453919.v2

[2]Diab, H.M., A.M. Abdelmoniem, M.R. Shaaban, I.A. Abdelhamid and A.H.M. Elwahy. An overview on synthetic strategies for the construction of star-shaped molecules. *Royal Society Chemistry Advances.* 2019, 9: 16606-16682. http://dx.doi.org/10.1039/c9ra02749a.

[3]Henderson, R., R.J. Edwards, K. Mansouri, K. Janowska, V. Stalls, S. Gobeil, . . . P. Acharya. Controlling the SARS-CoV-2 Spike Glycoprotein Conformation. *bioRxiv.* 2020: http://dx.doi.org/10.1101/2020.05.18.102087.

[4]Hsieh, C.L., J.A. Goldsmith, J.M. Schaub, A.M. DiVenere, H.C. Kuo, K. Javanmardi, . . . J.S. McLellan. Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science.* 2020: science.abd0826 [pii], http://dx.doi.org/10.1126/science.abd0826.

[5]Xiong, X., K. Qu, K.A. Ciazynska, M. Hosmillo, A.P. Carter, S. Ebrahimi, . . . J.A.G. Briggs. A thermostable, closed SARS-CoV-2 spike protein trimer. *Nat Struct Mol Biol.* 2020: http://dx.doi.org/10.1038/s41594-020-0478-5, 10.1038/s41594-020-0478-5 [pii].

[6]Carr, C.M. and P.S. Kim. A spring-loaded mechanism for the conformational change of influenza hemagglutinin. *Cell.* 1993, 73: 823-832. https://doi.org/10.1016/0092-8674(93)90260-W.

[7]Pallesen, J., N. Wang, K.S. Corbett, D. Wrapp, R.N. Kirchdoerfer, H.L. Turner, . . . J.S. McLellan. Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. *Proc Natl Acad Sci U S A.* 2017, 114: E7348-E7357. 1707304114 [pii], http://dx.doi.org/10.1073/pnas.1707304114.

[8]Kandeel, M., A.H.M. Abdelrahman, K. Oh-Hashi, A. Ibrahim, K.N. Venugopala, M.A. Morsy and M.A.A. Ibrahim. Repurposing of FDA-approved antivirals, antibiotics, anthelmintics, antioxidants, and cell protectives against SARS-CoV-2 papain-like protease. *J Biomol Struct Dyn.* 2020: 1-8. http://dx.doi.org/10.1080/07391102.2020.1784291

[9]Kandeel, M. and M. Al-Nazawi. Virtual screening and repurposing of FDA approved drugs against COVID-19 main protease. *Life Sci.* 2020, 251: 117627. S0024-3205(20)30375-1 [pii], http://dx.doi.org/10.1016/j.lfs.2020.117627.

[10]Bakowski, M.A., N. Beutler, K.C. Wolff, M.G. Kirkpatrick, E. Chen, T.-T.H. Nguyen, . . . T.F. Rogers. Drug repurposing screens identify chemical entities for the development of COVID-19 interventions. *Nature Communications.* 2021, 12: 3309-3323. https://doi.org/10.1038/s41467-021-23328-0.

[11]Wu, C., Y. Liu, Y. Yang, P. Zhang, W. Zhong, Y. Wang, . . . H. Li. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm Sin B.* 2020: http://dx.doi.org/10.1016/j.apsb.2020.02.008, S2211-3835(20)30299-9 [pii].

[12]Xia, S., L. Yan, W. Xu, A.S. Agrawal, A. Algaissi, C.K. Tseng, . . . L. Lu. A pan-coronavirus fusion inhibitor targeting the HR1 domain of human coronavirus spike. *Sci Adv.* 2019, 5: eaav4580. http://dx.doi.org/10.1126/sciadv.aav4580, aav4580 [pii].

[13]Xia, S., Y. Zhu, M. Liu, Q. Lan, W. Xu, Y. Wu, . . . L. Lu. Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein. *Cell Mol Immunol.* 2020, 17: 765-767. http://dx.doi.org/10.1038/s41423-020-0374-2, 10.1038/s41423-020-0374-2 [pii].

[14]Wang, C., S. Xia, P. Zhang, T. Zhang, W. Wang, Y. Tian, . . . K. Liu. Discovery of Hydrocarbon-Stapled Short alpha-Helical Peptides as Promising Middle East Respiratory Syndrome Coronavirus (MERS-CoV) Fusion Inhibitors. *J Med Chem.* 2018, 61: 2018-2026. http://dx.doi.org/10.1021/acs.jmedchem.7b01732.

[15]Cannalire, R., I. Stefanelli, C. Cerchia, A.R. Beccari, S. Pelliccia and V. Summa. SARS-CoV-2 Entry Inhibitors: Small Molecules and Peptides Targeting Virus or Host Cells. *Int J Mol Sci.* 2020, 21: ijms21165707 [pii], http://dx.doi.org/10.3390/ijms21165707.

[16]Tang, T., M. Bidon, J.A. Jaimes, G.R. Whittaker and S. Daniel. Coronavirus membrane fusion mechanism offers a potential target for antiviral development. *Antiviral Res.* 2020, 178: 104792. S0166-3542(20)30206-0 [pii], http://dx.doi.org/10.1016/j.antiviral.2020.104792.

[17]Ruan, Z., C. Liu, Y. Guo, Z. He, X. Huang, X. Jia and T. Yang. SARS-CoV-2 and SARS-CoV: Virtual Screening of Potential inhibitors targeting RNA-dependent RNA polymerase activity (NSP12). *J Med Virol.* 2020: http://dx.doi.org/10.1002/jmv.26222.

[18]Tsuji, M. Potential anti-SARS-CoV-2 drug candidates identified through virtual screening of the ChEMBL database for compounds that target the main coronavirus protease. *FEBS Open Bio.* 2020, 10: 995-1004. http://dx.doi.org/10.1002/2211-5463.12875.

[19]de Souza Neto, L.R., J.T. Moreira-Filho, B.J. Neves, R. Maidana, A.C.R. Guimaraes, N. Furnham, . . . F.P. Silva, Jr. In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery. *Front Chem.* 2020, 8: 93. http://dx.doi.org/10.3389/fchem.2020.00093.

[20]Yang, X., J. Zhang, K. Yoshizoe, K. Terayama and K. Tsuda. ChemTS: an efficient python library for de novo molecular generation. *Sci Technol Adv Mater.* 2017, 18: 972-976. http://dx.doi.org/10.1080/14686996.2017.1401424, 1401424 [pii].

[21]Spiegel, J.O. and J.D. Durrant. AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization. *J Cheminform.* 2020, 12: 25. http://dx.doi.org/10.1186/s13321-020-00429-4.

[22]Chevillard, F., S. Stotani, A. Karawajczyk, S. Hristeva, E. Pardon, J. Steyaert, . . . P. Kolb. Interrogating dense ligand chemical space with a forward-synthetic library. *Proc Natl Acad Sci U S A.* 2019, 116: 11496-11501. 1818718116 [pii], http://dx.doi.org/10.1073/pnas.1818718116.

[23]Volochnyuk, D.M., S.V. Ryabukhin, Y.S. Moroz, O. Savych, A. Chuprina, D. Horvath, . . . D.B. Judd. Evolution of commercially available compounds for HTS. *Drug Discov Today.* 2019, 24: 390-402. S1359-6446(18)30242-3 [pii], http://dx.doi.org/10.1016/j.drudis.2018.10.016.

[24]Liu, T., M. Naderi, C. Alvin, S. Mukhopadhyay and M. Brylinski. Break Down in Order To Build Up: Decomposing Small Molecules for Fragment-Based Drug Design with eMolFrag. *J Chem Inf Model.* 2017, 57: 627-631. http://dx.doi.org/10.1021/acs.jcim.6b00596.

[25]Patel, H., W.D. Ihlenfeldt, P.N. Judson, Y.S. Moroz, Y. Pevzner, M.L. Peach, . . . M.C. Nicklaus. SAVI, in silico generation of billions of easily synthesizable compounds through expert-system type rules. *Sci Data.* 2020, 7: 384. http://dx.doi.org/10.1038/s41597-020-00727-4.

[26]Polishchuk, P. CReM: chemically reasonable mutations framework for structure generation. *J Cheminform.* 2020, 12: 28. http://dx.doi.org/10.1186/s13321-020-00431-w.

[27]Jahnke, W., D.A. Erlanson, I.J.P. de Esch, C.N. Johnson, P.N. Mortenson, Y. Ochi and T. Urushima. Fragment-to-Lead Medicinal Chemistry Publications in 2019. *J Med Chem.* 2020, 63: 15494-15507. http://dx.doi.org/10.1021/acs.jmedchem.0c01608.

[28]Rifaioglu, A.S., E. Nalbat, V. Atalay, M.J. Martin, R. Cetin-Atalay and T. Dogan. DEEPScreen: high performance drug-target interaction prediction with convolutional neural networks using 2-D structural compound representations. *Chem Sci.* 2020, 11: 2531-2557. http:/dx.doi.org/10.1039/c9sc03414e.

[29]Polishchuk, P. Control of Synthetic Feasibility of Compounds Generated with CReM. *J Chem Inf Model.* 2020, 60: 6074-6080. http://dx.doi.org/10.1021/acs.jcim.0c00792.

[30]Yang, T., Z. Li, Y. Chen, D. Feng, G. Wang, Z. Fu, . . . M. Zheng. DrugSpaceX: a large screenable and synthetically tractable database extending drug space. *Nucleic Acids Res.* 2021, 49: D1170-D1178. 5940503 [pii], http://dx.doi.org/10.1093/nar/gkaa920.

[31]Blasco, R. and J.M. Coll. In silico screening for natural ligands to non-structural nsp7 conformers of SARS coronaviruses. *ChemRxiv.* 2020, https://doi.org/10.26434/chemrxiv.12952115.v2: https://doi.org/10.26434/chemrxiv.12952115.v2.

[32]Dallakyan, S. and A.J. Olson. Small-molecule library screening by docking with PyRx. *Methods Mol Biol.* 2015, 1263: 243-50. http://dx.doi.org/10.1007/978-1-4939-2269-7_19.

[33]Trott, O. and A.J. Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem.* 2010, 31: 455-61. http://dx.doi.org/10.1002/jcc.21334.

[34]Morris, G.M., R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell and A.J. Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem.* 2009, 30: 2785-91. http://dx.doi.org/10.1002/jcc.21256.

[35]Huey, R., G.M. Morris, A.J. Olson and D.S. Goodsell. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem.* 2007, 28: 1145-52. http://dx.doi.org/10.1002/jcc.20634.

[36]Rarey, M., B. Kramer, T. Lengauer and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol.* 1996, 261: 470-89. S0022-2836(96)90477-5 [pii], http://dx.doi.org/10.1006/jmbi.1996.0477.

[37]Schneider, N., S. Hindle, G. Lange, R. Klein, J. Albrecht, H. Briem, . . . M. Rarey. Substantial improvements in large-scale redocking and screening using the novel HYDE scoring function. *J Comput Aided Mol Des.* 2012, 26: 701-23. http://dx.doi.org/10.1007/s10822-011-9531-0.

[38]Schneider, N., G. Lange, S. Hindle, R. Klein and M. Rarey. A consistent description of HYdrogen bond and DEhydration energies in protein-ligand complexes: methods behind the HYDE scoring function. *J Comput Aided Mol Des.* 2013, 27: 15-29. http://dx.doi.org/10.1007/s10822-012-9626-2.

[39]Reau, M., F. Langenfeld, J.F. Zagury and M. Montes. Predicting the affinity of Farnesoid X Receptor ligands through a hierarchical ranking protocol: a D3R Grand Challenge 2 case study. *J Comput Aided Mol Des.* 2018, 32: 231-238. 10.1007/s10822-017-0063-0 [pii] http://dx.doi.org/10.1007/s10822-017-0063-0.

[40]Garcia, S. and F. Herrera. Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy. *Evol Comput.* 2009, 17: 275-306. http://dx.doi.org/10.1162/evco.2009.17.3.275.

[41]Bermejo-Nogales, A., J.M. Navas and J.M. Coll. Computational ligands to VKORC1s and CYPs. Could they predict new anticoagulant rodenticides? *BioRxiv.* 2021, https://biorxiv.org/cgi/content/short/2021.01.22.426921v1

[42]Connolly, M.C., J.M. Navas and J. Coll. Prediction of nanographene binding-scores to trout cellular receptors and cytochromes. *BioRxiv.* 2021: https://doi.org/10.1101/2021.02.20.432107.