Coronavirus canSAR – a data-driven, AI-enabled, drug discovery resource for the research community

Costas Mitsopoulos[1,2*], Albert A. Antolin[1*], Eloy Villasclaras Fernandez[1*], Patrizio Di Micco[1*], Ioan L. Mica[1*], Joseph E Tym*[1], Ka Hing Che[2], James Campbell[1], Domenico Sanfelice[1], Ian Collins[2], Paul Workman[2] and Bissan Al-Lazikani[1,2]

[1.] The Department of Data Science, The Institute of Cancer Research, London, SM2 5NG, UK
[2.] The Cancer Research UK Cancer Therapeutics Unit, The Institute of Cancer Research, London, SM2 5NG, UK

*Authors contributed equally to this work

Correspondence:
Bissan.al-lazikani@icr.ac.uk
Paul.Workman@icr.ac.uk

Abstract:

We describe an AI-enabled, integrated Coronavirus drug discovery knowledgebase, free for the research community. Its goal is to make accessible up to date information relevant to drug discovery for SARS-CoV-2 and other coronaviruses. It builds on great knowledge from across therapeutic areas and provides unbiased, systematic, objective information to empower the international effort.

The search for effective therapies and for novel drugs to tackle Covid-19 is an unparalleled international effort, incorporating a spectrum of activity from rapidly established clinical trials of existing drugs and laboratory-based drug screens, all the way to using artificial intelligence to predict likely active compounds. This heroic effort is unfortunately taking place within a chaotic information landscape[1] where data are often dispersed and rarely ever connected in meaningful ways. This means that the same paths are often repeatedly trodden, while truly novel opportunities may be missed. Moreover, the Covid-19 endeavour needs to learn from previous knowledge – not only from the field of viral drug discovery, including efforts that accompanied the SARS and MERS outbreaks, but also from other disease areas.

For over 10 years, our canSAR resource[2] (canSAR.icr.ac.uk) has provided an extensive, interconnected knowledgebase for cancer drug discovery and translational research that is publicly available to the research community worldwide. It integrates billions of experimental measurements from across genomics, chemistry, pharmacology, structural and systems biology, clinical trials and more. The data are fully interlinked to allow researchers to seamlessly and rapidly explore inter-domain connections. canSAR also employs a powerful suite of artificial intelligence (AI) tools[3,4] that have been helping translational researchers uncover novel druggable targets and generate new hypotheses[5-7]. canSAR is being used by >150,000 unique researchers from >350 countries, in both academia and industry.

When we built canSAR, a founding principle was that oncology must learn from other fields. Hence, we made canSAR as comprehensive as possible. While genomic data in canSAR are focused on oncology, much of the remaining data are disease-agnostic. For example, canSAR contains >2million bioactive small molecules assayed against proteins and models from across all biology – not just cancer and not only human. It also contains all 3-dimensional protein structures from the Protein Data Bank. Moreover, canSAR is automatically updated weekly to ensure where possible that it maintains the most up-to-date relevant data. As a result, canSAR is already being used by researchers from fields outside oncology.

To support the international community in the fight against Covid-19, we have adapted canSAR to coronavirus research and made it freely available on corona.cansar.icr.ac.uk. In addition to existing functionality, we also generated special live reports on key aspects of coronavirus drug discovery research. These reports summarise and curate information about key areas of interest – such as 3D structures and clinical trials – and these summaries are automatically updated as new data become available.

In this first release, we have focused on four major areas: 3D structural analysis and 'ligandability' assessment for drug discovery; drugs in clinical trials; virus-host molecular interactions; and chemical probes for mechanistic studies.

3D protein structural analysis

At the time of writing, Coronavirus canSAR contained >500,000 individual protein structural snapshots for >50,000 proteins from 13,857 organisms. For these, we applied our AI methods[4] to assess >4 million cavities for ligandability (the suitability to bind a small-molecule). In brief, our method examines all 3D structures and identifies all candidate cavities on the protein surface. We calculate 35 geometric and physicochemical properties for each of these cavities which are then classified by our AI algorithm for their suitability for small-molecule binding. Using the same method, we also analysed the 3D structural interfaces of >117,000 biological complexes and 640,988 interface cavities. All results for all structural analyses and AI assessments are provided through a user-friendly interface on the canSAR website.

In particular, at the time of writing, we analysed 229 SARS-CoV-2 (Covid-19), 443 SARS-CoV (SARS) and 181 MERS-CoV (MERS) protein structural snapshots. Some 661 structures are from X-ray, 157 from cryo-EM and 35 from NMR. We found 80 structures of eight human proteins, belonging to seven different families, that bind viral proteins. We analysed >8,100 cavities on >850 protein structural snapshots. Through this analysis, we identified 284 ligandable pockets and 339 ligandable cavities at the interfaces of protein complexes.

Importantly, we find novel ligandable cavities at the interface between the coronavirus Spike protein and their receptor ACE2, and at the complex's interface with the Sodium-dependent neutral amino acid transporter B(0)AT1 (Figure 1). These live analyses that are updated weekly as new structural data emerge provide drug discoverers with clear, data-driven hypotheses for exploring drug intervention approaches.

Drugs in clinical trials

Coronavirus canSAR maintains a daily update of clinical trials for Covid-19, SARS and MERS from clinicatrials.gov. At the time of writing, Coronavirus canSAR contained >1180 clinical trials, of which 520 are interventional trials. Although the trial descriptions are not structured and are composed of human-written text, we use canSAR's framework to identify and map drug names to these trials, and we also map the trials to active substance classifications from the Anatomical Therapeutic Chemical (ATC) Classification System. This helps classify trials based on the therapeutic strategies adopted (direct antivirals, immunostimulants, immunosuppressives, etc).

At the time of writing there were 150 drugs under investigation in open trials for Covid-19. Of these 10 are direct acting antivirals and 16 are immunosuppressants. Despite continuing lack of clear evidence of the efficacy of the chloroquine class of antimalarial drugs in Covid-19, they remain the most studied drug class with 133 of 520 trials investigating them alone or in combination with other therapeutics. In contrast, so far only 65 trials are utilising directly acting antivirals or angiotensin blockers. Similarly, while many antiviral trials are examining the efficacy of nucleoside analogues that are RNA polymerase inhibitors, for some other inhibitors the mechanisms are not so clear.

The Coronavirus interactome

Building on our curated and pharmacologically annotated human protein-protein interaction database which contains >900,000 interactions from over 60,000 peer reviewed publications, we augmented this with a further 500 experimentally validated interactions between coronavirus and human proteins from high-quality experimental data (e.g. ref[8]). We curated interactive cellular communications maps to allow researchers to explore coronavirus molecular interactions with human proteins, and identify existing drug targets and novel druggable proteins and protein interactions. We constructed three molecular maps which cover: 1) SARS-CoV-2 proteins and human protein partners that meet strict experimental validation criteria; 2) SARS-CoV-2 proteins and human protein partners that meet less strict criteria but still require experimental evidence; and 3) as per map (1) but additionally including known interactions with SARS-CoV and MERS-CoV proteins.

Our analysis of the strictly defined SARS-CoV-2-human cellular interactome shown in Figure 1 contains nine targets of FDA approved drugs, nine SARS-CoV-2 and 137 human proteins that we have assessed to be druggable using our AI structural ligandability algorithm. In addition, we identify 17 crystallographically-resolved protein binding interfaces, eight of which contain ligandable pockets, offering further intervention possibilities.

Chemical probes for investigating Coronavirus

High quality small-molecule chemical probes can be powerful tools to explore biology and validate drug targets, especially when used alongside genetic perturbation[9]. However, the use of low quality or flawed chemical probes can produce misleading results that affect the robustness and reproducibility of biomedical research[10]. There is considerable lack of clarity and quality around chemical probes for Coronavirus targets, for example with a series of promiscuous frequent hitters being referred to erroneously as ‘drug leads with clinical potential’ for the Covid-19 main protease Mpro[11]. There is an urgent need to support the scientific community and empower them to select high quality chemical probes.

Informed by existing resources, The Chemical Probes Portal[12] (chemicalprobes.org) and Probe Miner[13] (probeminer.icr.ac.uk), we have embarked on curation and assessment of small-molecule tools for use in coronavirus mechanistic and drug discovery research. At the time of writing, Coronavirus canSAR contains 18 curated, selective chemical probes that act on 11 targets that are either viral or human proteins hijacked by Covid-19. These can be used relatively safely in experiments with minimal risk of significant off-target action or chemical reactivity. We also list 12 compounds that have been proposed in the literature for use in coronavirus mechanistic studies, but that are problematic and may produce misleading results due to poor selectivity or other factors that make them unsuitable as chemical probes. We will continue updating our chemical probe and caution list curation as Coronavirus canSAR expands.

Integrated knowledge

A key enabler of Coronavirus canSAR is that all data are fully interconnected. Drugs in clinical trials are interlinked with their molecular targets, other clinical trials outside Covid-19 in which they are being investigated, 3D protein structural complexes, and published in-vitro or in-vivo activities from key sources. Likewise, viral proteins or host proteins that are utilised by the virus are fully annotated and interlinked with 3D structures, published inhibitors, protein-protein interactions and more. This enables the researchers to identify key biological interactions and novel drug target opportunities.

In the next months, we will expand Coronavirus canSAR to include more data and also additional types of data including gene-expression profiling and large-scale drug screens. We will maintain Coronavirus canSAR and keep it freely available for the community, and expand it to ensure it remains an objective, data-rich resource. We will also respond to feedback from users to ensure this resource facilitates a more informed, data-driven, international response against Covid-19.

References

1 Mullard, A. Coordinating the COVID-19 pipeline. *Nature reviews. Drug discovery*, doi:10.1038/d41573-020-00068-2 (2020).

2 Coker, E. A. *et al.* canSAR: update to the cancer translational research and drug discovery knowledgebase. *Nucleic acids research* **47**, D917-D922, doi:10.1093/nar/gky1129 (2019).

3 Mitsopoulos, C., Schierz, A. C., Workman, P. & Al-Lazikani, B. Distinctive Behaviors of Druggable Proteins in Cellular Networks. *PLoS computational biology* **11**, e1004597, doi:10.1371/journal.pcbi.1004597 (2015).

4 Bulusu, K. C., Tym, J. E., Coker, E. A., Schierz, A. C. & Al-Lazikani, B. canSAR: updated cancer research and drug discovery knowledgebase. *Nucleic acids research* **42**, D1040-1047, doi:10.1093/nar/gkt1182 (2014).

5 Wedge, D. C. *et al.* Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. *Nature genetics* **50**, 682-692, doi:10.1038/s41588-018-0086-z (2018).

6 Chau, C. H., O'Keefe, B. R. & Figg, W. D. The canSAR data hub for drug discovery. *Lancet Oncol* **17**, 286, doi:10.1016/S1470-2045(16)00095-4 (2016).

7 Patel, M. N., Halling-Brown, M. D., Tym, J. E., Workman, P. & Al-Lazikani, B. Objective assessment of cancer genes for drug discovery. *Nature reviews. Drug discovery* **12**, 35-50, doi:10.1038/nrd3913 (2013).

8 Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, doi:10.1038/s41586-020-2286-9 (2020).

9 Workman, P. & Collins, I. Probing the probes: fitness factors for small molecule tools. *Chem Biol* **17**, 561-577, doi:10.1016/j.chembiol.2010.05.013 (2010).

10 Blagg, J. & Workman, P. Choose and Use Your Chemical Probe Wisely to Explore Cancer Biology. *Cancer cell* **32**, 9-25, doi:10.1016/j.ccell.2017.06.005 (2017).

11 Jin, Z. *et al.* Structure of M(pro) from COVID-19 virus and discovery of its inhibitors. *Nature*, doi:10.1038/s41586-020-2223-y (2020).

12 Arrowsmith, C. H. *et al.* The promise and peril of chemical probes. *Nat Chem Biol* **11**, 536-541, doi:10.1038/nchembio.1867 (2015).

13 Antolin, A. A. *et al.* Objective, Quantitative, Data-Driven Assessment of Chemical Probes. *Cell chemical biology* **25**, 194-205 e195, doi:10.1016/j.chembiol.2017.11.004 (2018).

Competing Interests

CM, AAA, EVF, PdM, ILM, JET, KHC, JC, DS, IC, PW and BAL are employees of the Institute of Cancer Research (ICR) which has a commercial interest in a range of drug targets. The ICR operates a Rewards to Inventors scheme through which employees of the ICR may receive financial benefit following commercial licensing. PW is a former employee of AstraZeneca and received research funding from Vernalis, Astex, AstraZeneca, BACIT and Sixth Element Capital/CRT Pioneer Fund. PW is a consultant/scientific advisory board member for NextechInvest, Storm Therapeutics, Astex Pharmaceuticals and CV6, and holds stock in Chroma Therapeutics, NextInvest and Storm Therapeutics. He is also a Non-Executive Director of Storm Therapeutics and the Royal Marsden NHS Trust and a Board Director of the non-profit Chemical Probes Portal. BAL is currently or has been a consultant to Astex Pharmaceuticals, GSK, Astelas Pharma and Difiniens AG (member of Astra Zeneca group). BAL is a former employee of Inpharmatica Ltd. IC is currently or has been a consultant to Epidarex LLP, AdoRx Therapeutics and Enterprise Therapeutics, has received research funding from Astex, Merck KGaA, Janssen Biopharma, Monte Rosa Therapeutics and Sixth Element Capital/CRT Pioneer Fund, and holds stock in Monte Rosa Therapeutics AG. IC is a former employee of Merck Sharp & Dohme.

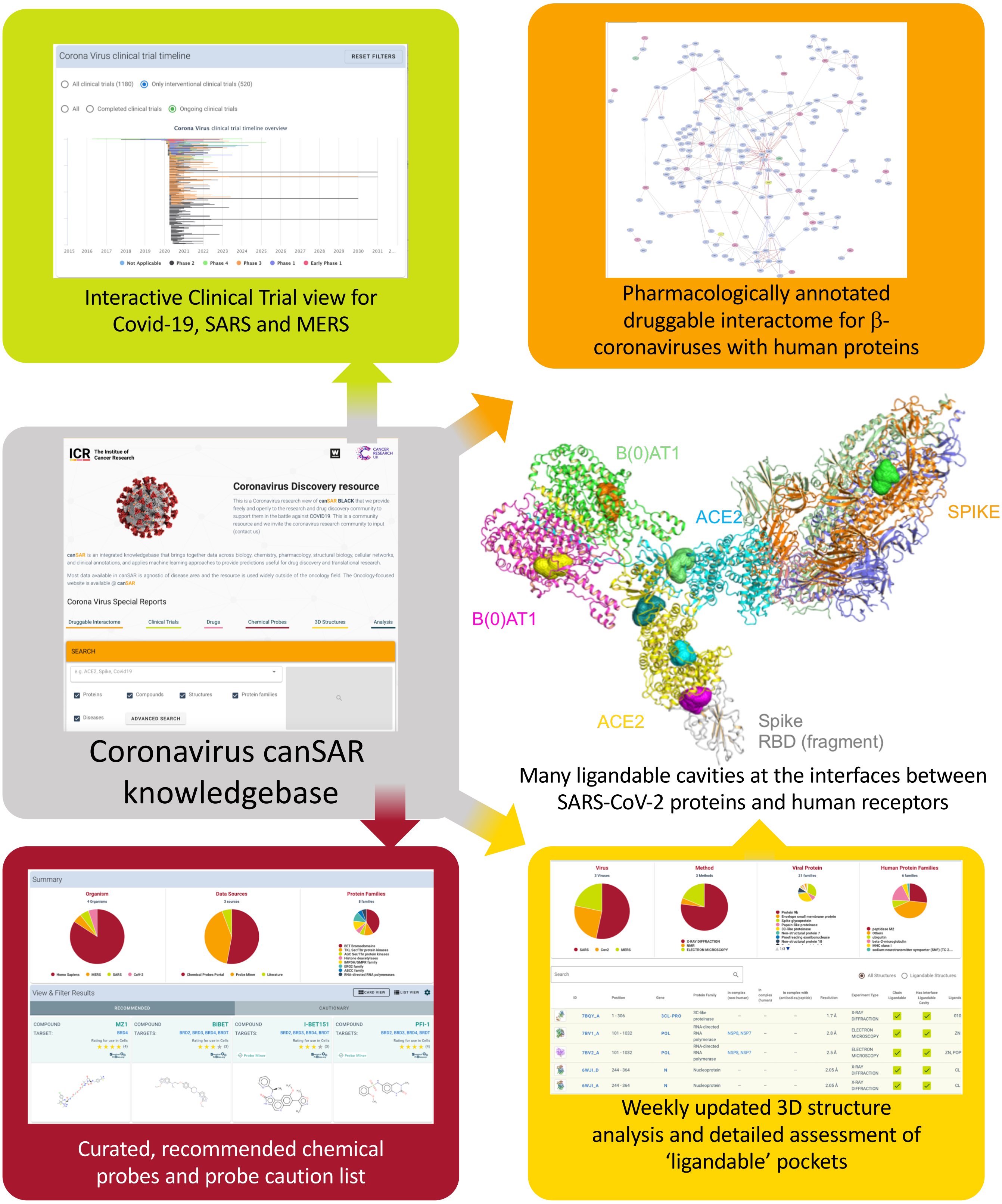
Acknowledgements

canSAR is funded by the Cancer Research UK Drug Discovery Committee strategic award to BAL and PW ‘canSAR: enhancing the drug discovery knowledgebase’ (C35696/A23187). PW, BAL and IC are also funded by a programme grant from CRUK Cancer Therapeutics Unit (C309/A25144). PW and BAL are funded by a Wellcome Trust Biomedical Resource Award to operate and develop the Chemical Probes Portal (WT212969/Z/18/Z). AAA is a Sir Henry Wellcome Postdoctoral Fellow (WT204735/Z/16/Z). We thank the Heather Beckwith Charitable Settlement and The John L Beckwith Charitable Trusts for their generous support of our High Performance Computing facility.

We thank John P. Overington for providing curation of approved antiviral drugs and for useful discussion and advice. We also thank Veronica Garcia-Perez and Karen J. Cross for their advice and support in the completion of this work.

Figure 1

The Coronavirus canSAR knowledgebase serves as a portal to enable the analysis and investigation of interlinked data for use in coronavirus drug discovery and translational research.


Corona Virus clinical trial timeline
RESET FILTERS
○ All clinical trials (1180) ● Only interventional clinical trials (520)
○ All ○ Completed clinical trials ● Ongoing clinical trials
Corona Virus clinical trial timeline overview
2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026 2027 2028 2029 2030 2031
● Not Applicable ● Phase 2 ● Phase 4 ● Phase 3 ● Phase 1 ● Early Phase 1
ICR The Institute of Cancer Research
CANCER RESEARCH UK
Coronavirus Discovery resource
This is a Coronavirus research view of canSAR BLACK that we provide freely and openly to the research and drug discovery community to support them in the battle against COVID19. This is a community resource and we invite the coronavirus research community to input (contact us)
canSAR is an integrated knowledgebase that brings together data across biology, chemistry, pharmacology, structural biology, cellular networks, and clinical annotations, and applies machine learning approaches to provide predictions useful for drug discovery and translational research.
Most data available in canSAR is agnostic of disease area and the resource is used widely outside of the oncology field. The Oncology-focused website is available @ canSAR
Corona Virus Special Reports
Druggable Interactome
Clinical Trials
Drugs
Chemical Probes
3D Structures
Analysis
SEARCH
e.g. ACE2, Spike, Covid19
☑ Proteins ☑ Compounds ☑ Structures ☑ Protein families
☑ Diseases
ADVANCED SEARCH
B(0)AT1
ACE2
SPIKE
B(0)AT1
ACE2
Spike
RBD (fragment)
Summary
Organism
4 Organisms
● Homo Sapiens ● MERS ● SARS ● CoV-2
Data Sources
3 sources
● Chemical Probes Portal ● Probe Miner ● Literature
Protein Families
8 families
BET Bromodomains
TKL Ser/Thr protein kinases
AGC Ser/Thr protein kinases
Histone deacetylases
IMPDH/GMPR family
ERG2 family
ABCC family
RNA-directed RNA polymerases
View & Filter Results
CARD VIEW
LIST VIEW
RECOMMENDED
CAUTIONARY
COMPOUND MZ1
TARGET: BRD4
COMPOUND BiBET
TARGETS: BRD2, BRD3, BRD4, BRDT
COMPOUND I-BET151
TARGETS: BRD2, BRD3, BRD4, BRDT
COMPOUND PFI-1
TARGETS: BRD2, BRD3, BRD4, BRDT
Rating for use in Cells
Probe Miner
Virus
3 Viruses
● SARS ● Cov2 ● MERS
Method
3 Methods
● X-RAY DIFFRACTION ● NMR ● ELECTRON MICROSCOPY
Viral Protein
21 families
Protein 9b
Envelope small membrane protein
Spike glycoprotein
Papain-like proteinase
3C-like proteinase
Non-structural protein 7
Proofreading exoribonuclease
Non-structural protein 10
Human Protein Families
6 families
peptidase M2
Others
ubiquitin
beta-2-microglobulin
MHC class I
sodium:neurotransmitter symporter (SNF) (TC 2...
Search
● All Structures ○ Ligandable Structures
ID
Position
Gene
Protein Family
Experiment Type
Resolution
Chain Ligandable
Has Interface Ligandable Cavity
Ligands
7BQY_A
7BV1_A
7BV2_A
6WJI_D
6WJI_A

Interactive Clinical Trial view for Covid-19, SARS and MERS

Pharmacologically annotated druggable interactome for β-coronaviruses with human proteins

Coronavirus canSAR knowledgebase

Many ligandable cavities at the interfaces between SARS-CoV-2 proteins and human receptors

Curated, recommended chemical probes and probe caution list

Weekly updated 3D structure analysis and detailed assessment of ‘ligandable' pockets