Ligand-centered assessment of SARS-CoV-2 drug target models

A. Wlodawer[1], Z. Dauter[2], I. Shabalin[3,4], M. Gilski[5,6], D. Brzezinski[3,6,7], M. Kowiel[6], W. Minor[3,4], B. Rupp[8,9], **M. Jaskolski**[5,6]

[1]*Protein Structure Section, Macromolecular Crystallography Laboratory, NCI, Frederick, MD, USA*
[2]*Synchrotron Radiation Research Section, Macromolecular Crystallography Laboratory, NCI, Argonne National Laboratory, IL, USA*
[3]*Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, USA*
[4]*Center for Structural Genomics of Infectious Diseases (CSGID), Charlottesville, VA, USA*
[5]*Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University, Poznan, Poland*
[6]*Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland*
[7]*Institute of Computing Science, Poznan University of Technology, Poznan, Poland*
[8]*k.-k. Hofkristallamt, San Diego, CA, USA*
[9]*Institute of Genetic Epidemiology, Medical University Innsbruck, Austria*

Outline

1. Atomic structure determination and drug discovery
2. SARS-CoV-2 drug targets
3. Assessment protocol of SARS-CoV-2 structure models
4. Examples of detected problems
5. Future plans

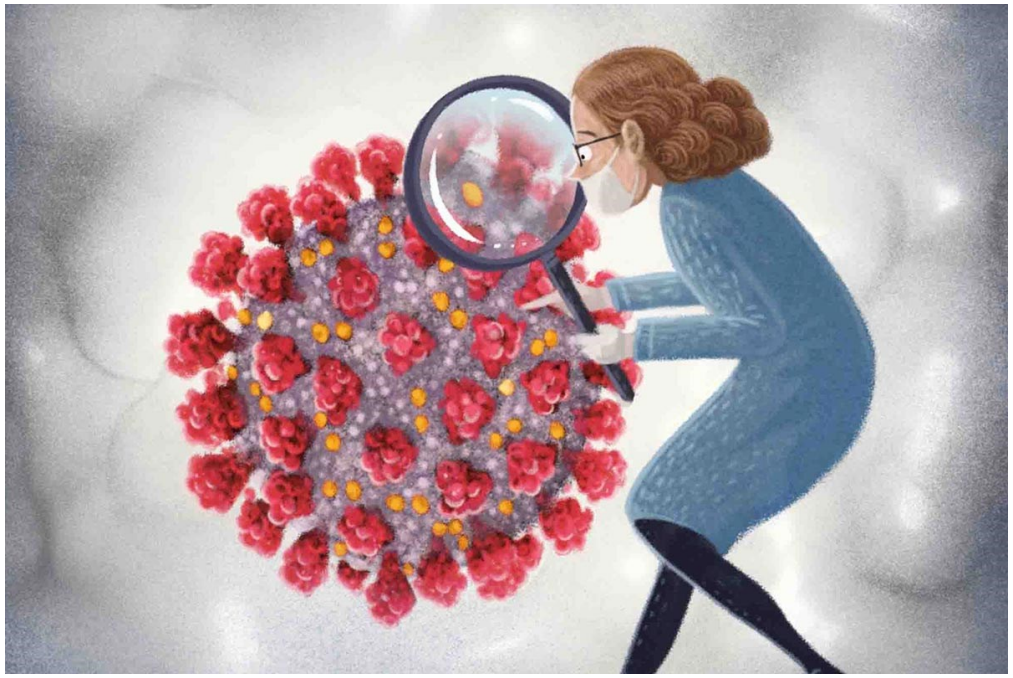
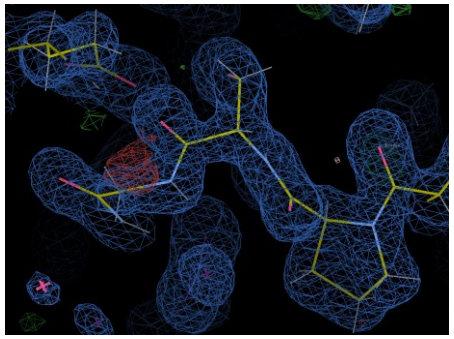


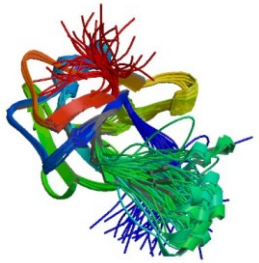
Illustration by Marcin Minor

Atomic structure determination

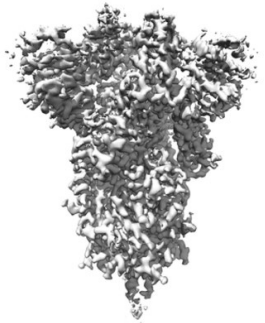
- The goal of structure determination is to experimentally reveal the 3D atomic architecture of a chemical compound (e.g. a protein)
- Main methods used for this purpose:



X-ray crystallography
(**6WNP**: SARS-CoV-2 Main Protease)



NMR spectroscopy
(**6YI3**: SARS-CoV-2 RBD)

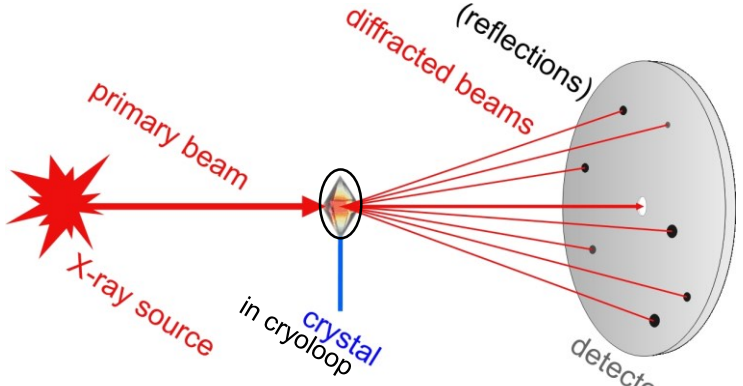


Cryo electron microscopy
(**6X29**: SARS-CoV-2 Spike)

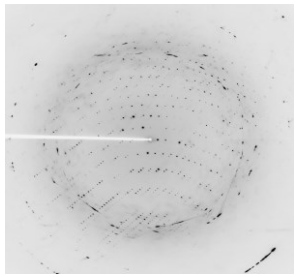
- The resulting 3D models are made publicly available through databases
- Protein structures are deposited in the Protein Data Bank (PDB)

X-ray crystallography

- Most popular structure determination method (89% PDB, 90% SARS-CoV-2)
- Offers highest resolution
- Best choice for drug design and fragment screening
- Like each structure determination method, requires a degree of human interpretation

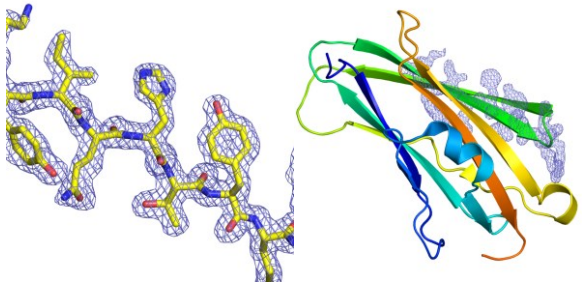


(a) X-ray diffraction experiment



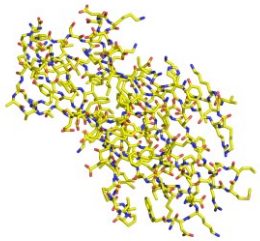
(b) Diffraction image

FT → phase problem

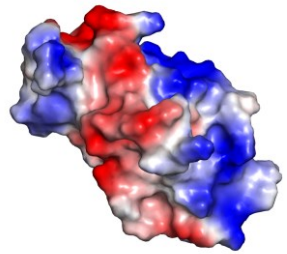


(c) Electron density

(d) Cartoon model



(e) All-atom model



(f) Surface

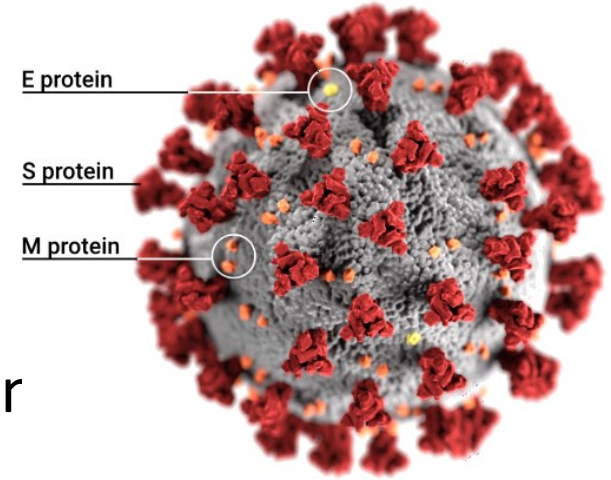
(g) deposit in PDB

WORLDWIDE PDB
PROTEIN DATA BANK

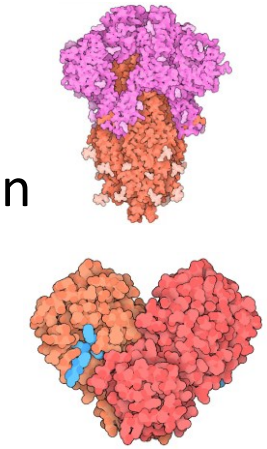
Structure-based drug design

- Knowledge of the atomic structure of biological macromolecules is necessary to understand the mechanisms of life processes
- In the case of viruses, such knowledge is the basis for the design of drugs (*bullets*) that target certain parts of the virus and block their function
- Usually this requires:
  - finding a suitable binding site (*pocket*) in one the virus's proteins
  - designing a small-molecule with tight & specific binding in that site
- With iteration cycles, this is the most rational way to develop efficient drugs targeting specific diseases
- HIV treatments have been designed this way

Drug targets for SARS-CoV-2



- SARS-CoV-2 consists of ~30 proteins and encapsulated RNA ger that codes those proteins
- The proteins can be classified as:
  - **Structural proteins**: M, E, S, N
  - Non-structural proteins (NSP): mainly **enzymes** (biocatalysts) and regulatory proteins
- The main proteins that can be used for drug design:
  - **Spike protein (S)**: structural protein that recognizes the ACE2 receptor on human cell; if this protein (or ACE2) is blocked by a drug, the virus will not be able to enter the host cell
  - **Main protease (Mpro)**: an enzyme whose function is to cut the viral polyproteins produced in the infected cell to their active form; if this enzyme is blocked by a drug, the virus will not be able to mature and will be non-infectious



Project goal

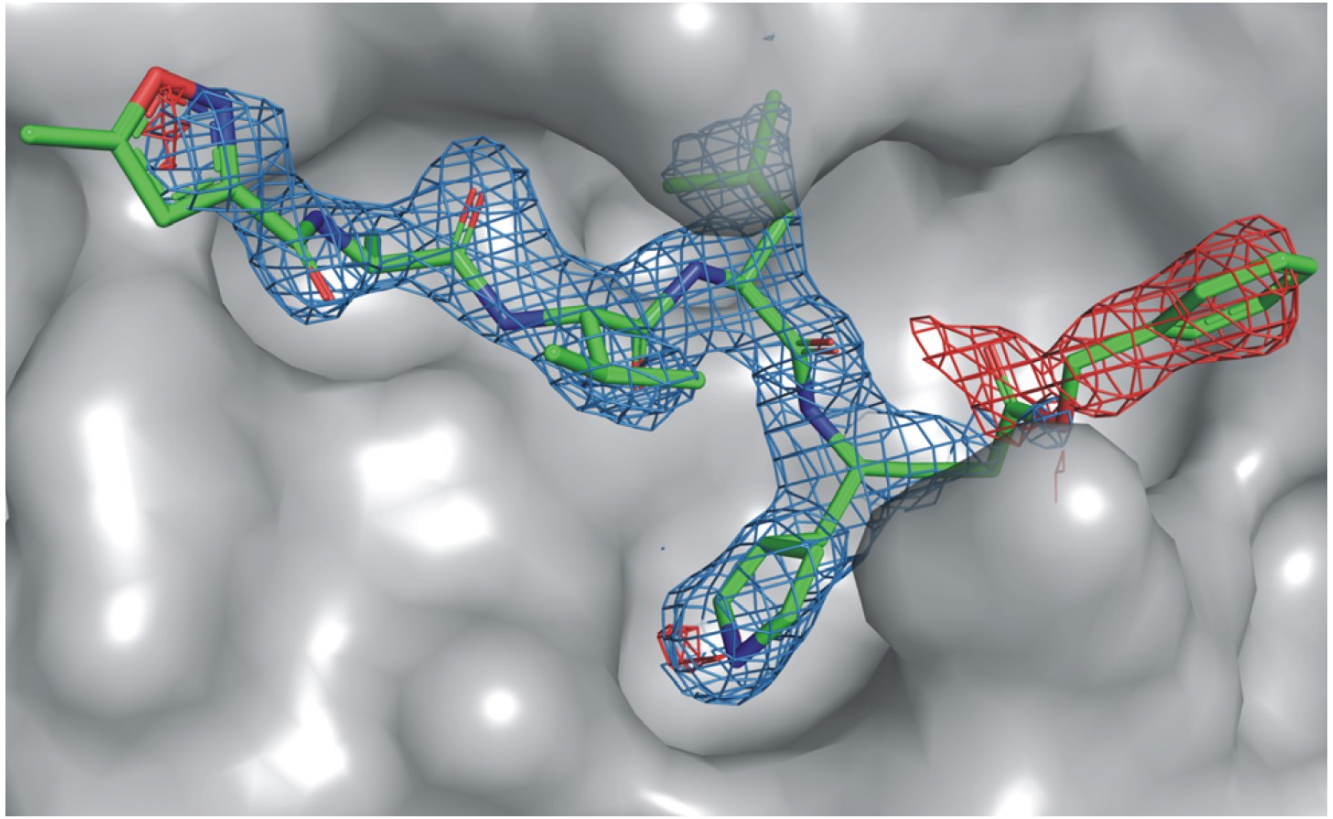
Critically evaluate the experimentally determined SARS-CoV-2 protein structures, with special focus on potential drug targets

Proposed assessment protocol

- Extract data from the PDB
- Look for raw diffraction data (IRRMC or Zenodo)
- Run validation tools:
  - MolProbity (geometry checking, assessment of the entire model)
  - Twilight (real space correlation coefficient, assessment of ligands)
- Pass data to expert structural biologists
- Determine protein type and ligand status
- If needed, re-refine the structure
  - Run ACHESYM (standardization of model placement in the unit cell)
  - [If interesting case] Prepare Molstack visualization for comparison

Example problems – incorrect ligand model

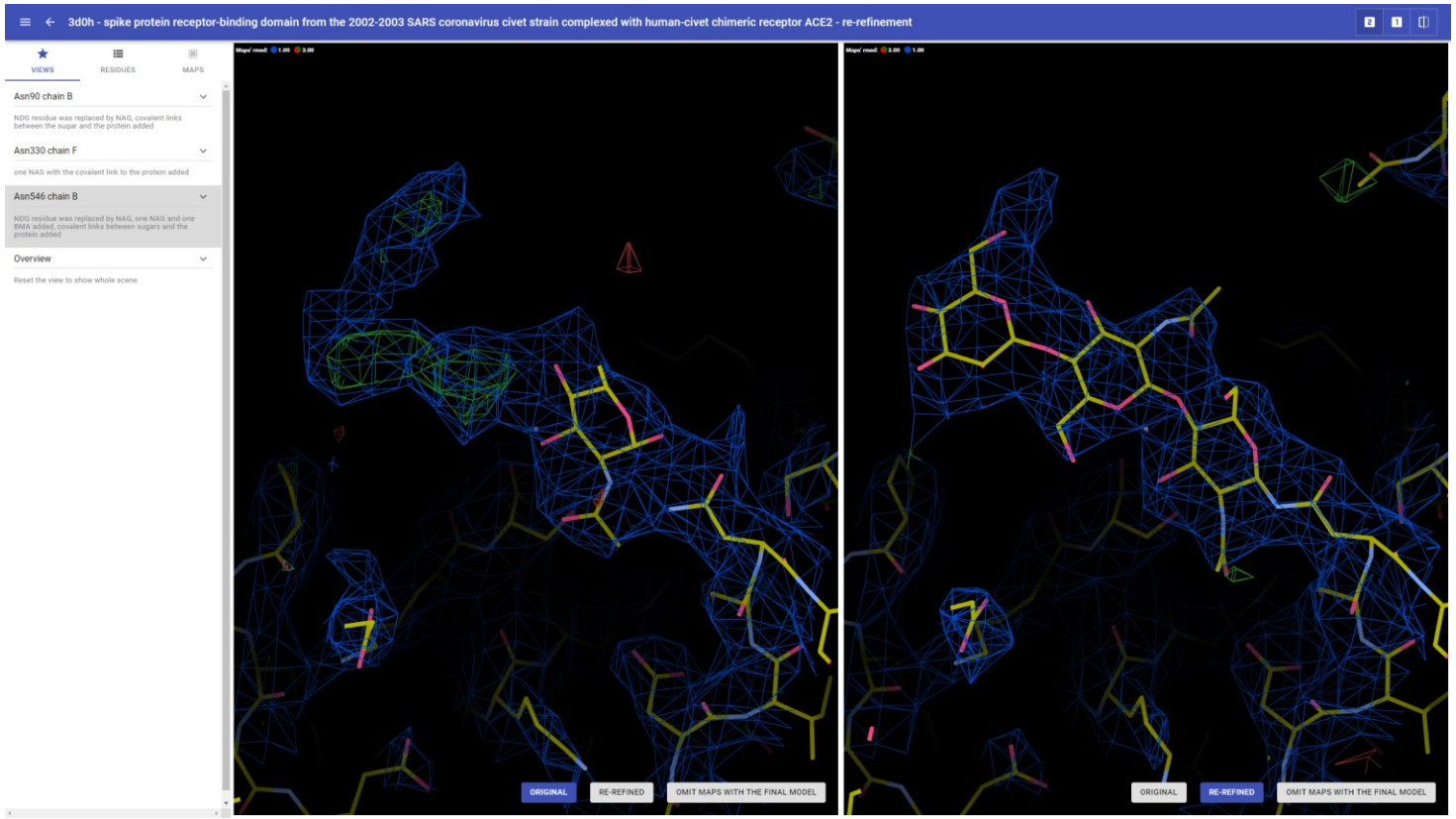
- Peptidic inhibitor in the substrate-binding site of the structure with PDB ID **6LU7**
- The presence of negative difference electron density (red contour) for the terminal benzyl group indicates that this group has been eliminated by hydrolysis and is not there



Incorrectly modeled inhibitor molecule in the protein binding site

Example problems – missing chain fragment

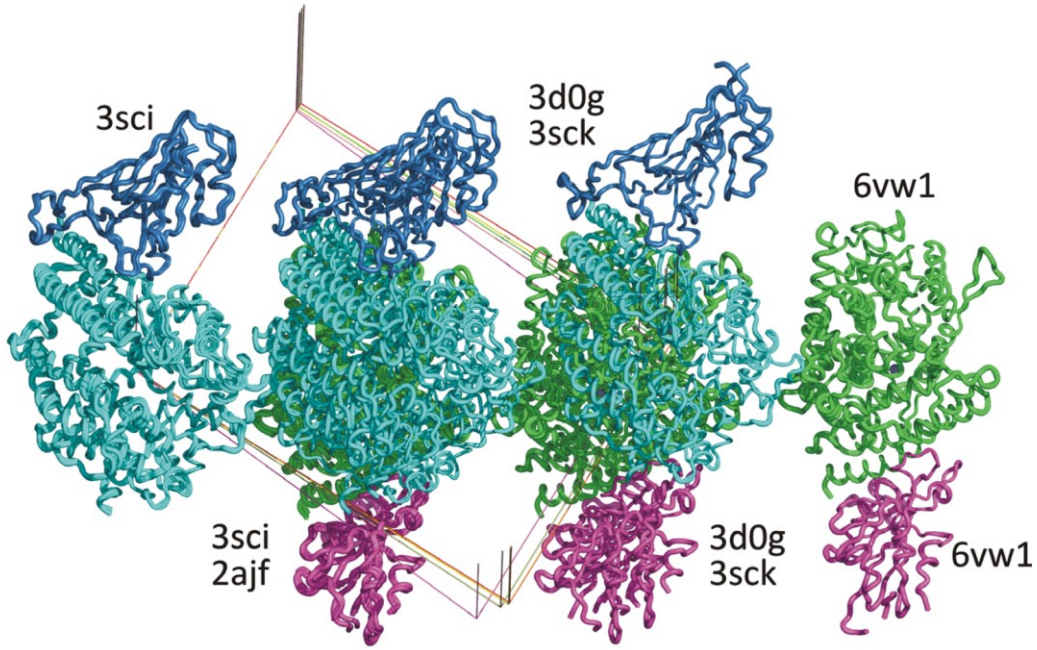
- Structure **3D0H**
- Three chemically linked carbohydrate molecules (NAG-NAG-BMA) should be connected to residue Asn546B
- Left panel shows the original (wrong) model
- Right panel after corrections



https://molstack.bioreproducibility.org/project/view/WrI2XslE978LiF95PQYo/

Example problems – unit cell placement

- Structures of the same protein although crystallized isomorphously are often presented inconsistently
- This means that different versions of the same protein are hard to compare
- To alleviate this issue we used our ACHESYM server in each re-refinement to unify model placement in the unit cell

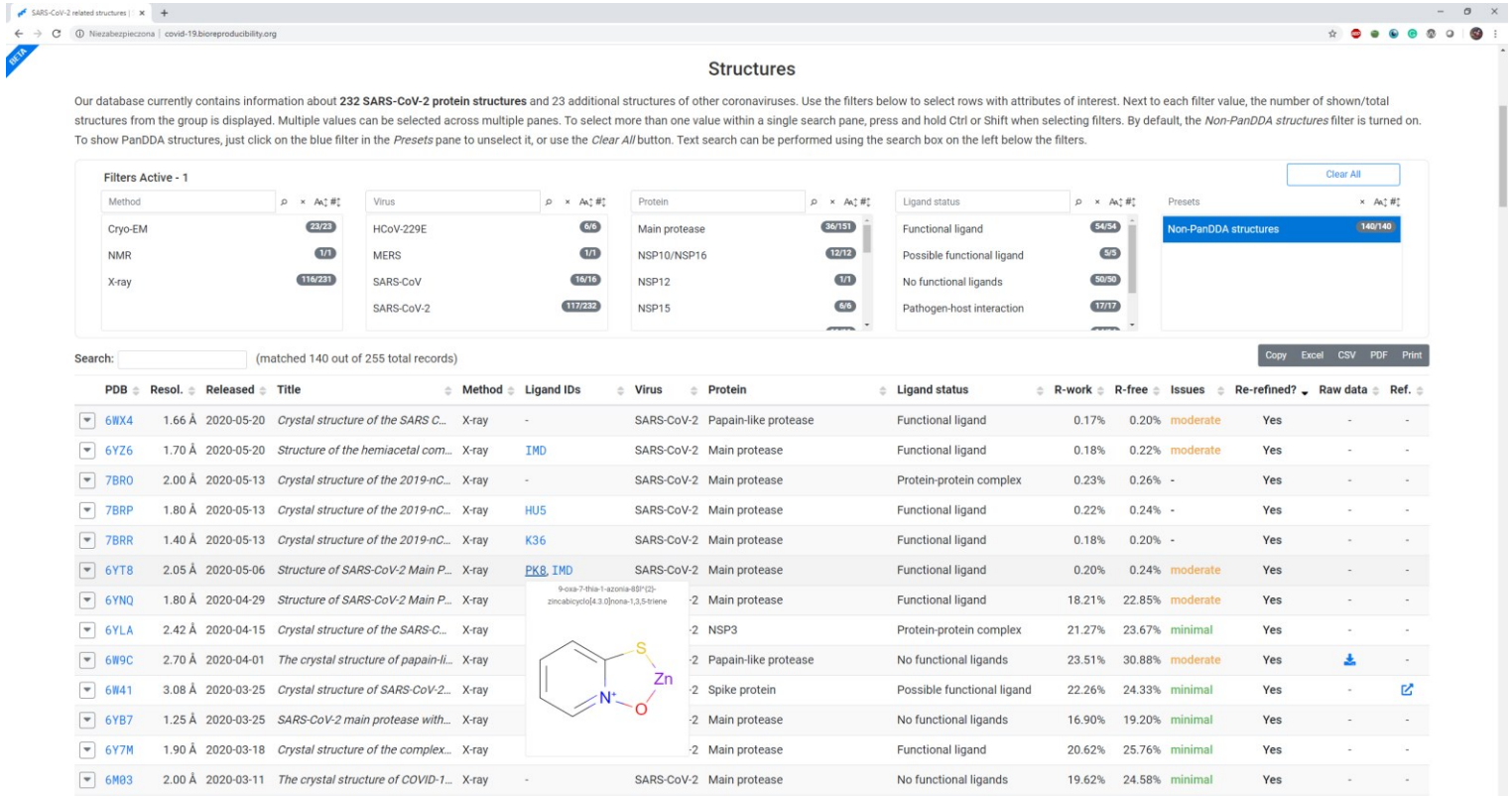


Protein structures after placing in isomorphous unit cells

Web resource

- Aggregates all the mined SARS-CoV-2 data
- Provides info about original model problems & links to re-refinements
- Classifies proteins according to:
  - experimental method
  - virus type
  - protein type
  - ligand status
- Allowing flexible and versatile selection of cases

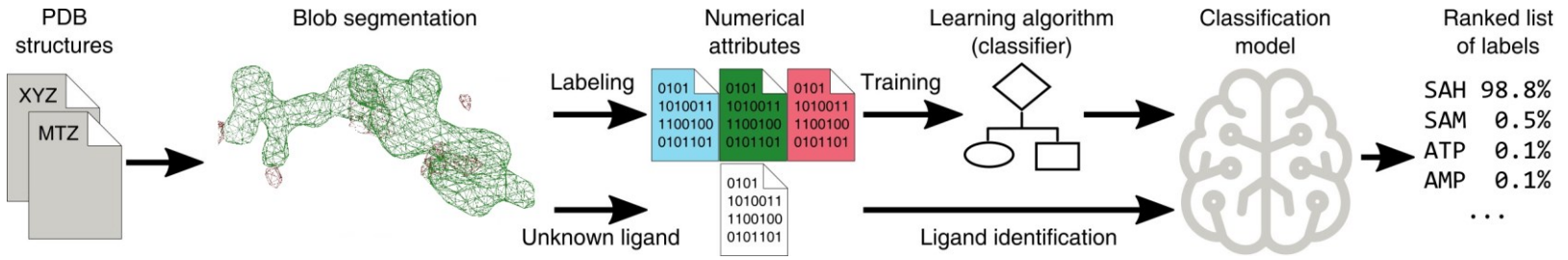
covid-19.bioreproducibility.org



https://covid-19.bioreproducibility.org

Future plans

- Use Machine Learning validation as an addition to correlation-based validation metrics (https://checkmyblob.bioreproducibility.org/)



- Work on combining genetic/structural visualizations with our quality assessment data (https://coronavirus3d.org/)
- Evaluate PanDDA fragment screening procedure to prevent flooding of the PDB with low-quality ligand complexes

Conclusions

- New structures of SARS-CoV-2 proteins with ligands appear every week
- Due to the accelerated pace of COVID-related science, these structures have to be double-checked for correctness as drug design targets
- We use bioinformatic tools and expert knowledge to review, validate & rectify these structures
- Through our covid-19.bioreproducibility.org server we want to pass our results on to the biomedical community
- We plan to expand it with new validation metrics and categorizations
- **Tools that combine knowledge and translate it to other fields are as important as tools that generate new knowledge within one field**