# Amenders Assemble!

Iris D. Young, James S. Fraser
Department of Bioengineering and Therapeutic Sciences, UCSF
jfraser@fraserlab.com

"Until such time as the world ends, we will act as though it intends to spin on." The efforts of scientific researchers responding to the COVID-19 pandemic have been nothing short of heroic. The breakneck pace of this research has depended not only on individual researchers' incredible work ethics but also on strategic collaborations, flexible working environments, and streamlined dissemination of information. There are also international efforts designed to take full advantage of a distributed workforce, such as the COVID Moonshot (https://www.biorxiv.org/content/10.1101/2020.10.29.339317v1) or the Billion Molecules against COVID-19 JEDI grand challenge (https://www.covid19.jedi.group/), and even some that leverage crowdsourcing, such as Folding@Home (https://foldingathome.org/diseases/infectious-diseases/covid-19/). X-ray crystallography and cryoEM facilities are granting priority access to proposals related to SARS-CoV-2 and a number of journals are making COVID-19-related publications available to the public for free. Many researchers are also accelerating research by posting their COVID-related manuscripts directly to preprint servers such as bioRxiv or medRxiv. The promise of these preprint servers is that peer review can take place in real time and in full public view.

In a recent issue of Biophysical Journal (https://pubmed.ncbi.nlm.nih.gov/33460600/), Croll and coworkers describe such a real time “peer review” process (in a manuscript we peer reviewed!). Hundreds of COVID-related structures have been deposited to the Protein Data Bank (PDB). Although automated validation steps built into the deposition pipeline catch a variety of map and model pathologies before structures are accepted, good validation metrics can miss important details. Worse, once errors are discovered, it may be too late to stop the propagation of structural errors to derivative structures. But we don't judge people (or structures) by their worst mistakes. The PDB has recently introduced a versioning system that enables deposition authors to update the entry while maintaining the existing PDB ID code.

Several collaborative efforts have sprung up to address the proliferation of SARS-CoV-2 structures. In addition to collections curated by the PDB itself, the CoV3D database (https://cov3d.ibbr.umd.edu/) provides a landing place for structural biologists to find all available SARS-CoV-2, SARS-CoV-1 and MERS-CoV structures in a sensible organization. The Coronavirus Structural Task Force (https://insidecorona.net/) goes a step further: the members of the task force examine and re-refine deposited SARS-CoV-2 and SARS-CoV-1 structures, post these results, contact the authors of the original structures with their findings, and encourage them to reversion their depositions. In Croll et al, the idea was to bring together a group of remarkable people to perform the re-refinement process, ensuring that the records in the PDB are maintained as up-to-date as possible.

Their first step in examining a newly deposited structure is running automated comprehensive validation using traditional metrics: Ramachandran plots, clashscores, bond and angle rmsds, etc. In the resolution range of most SARS-CoV-2 structures, CaBLAM scores had the best performance in the latest CryoEM Model Challenge (Croll ref 13), so the authors prioritize residues flagged by CaBLAM for remediation. They also pay close attention to biologically important regions of the

structures, such as active sites or domains known to undergo conformational changes, and features especially prone to mismodeling, such as metal ions and disulfide bridges. Finally, painstakingly, they perform a final once-over the entire structure residue by residue using the hybrid molecular dynamics-molecular graphics program ISOLDE (https://journals.iucr.org/d/issues/2018/06/00/ic5101/). This process consistently produces improved structures by nearly all metrics. The lone exception is bond lengths, which are optimized to systematically different values by the AMBER force field than the values they are compared against during validation; this is remedied by a final round of refinement in Phenix.

The authors describe how in several cases they were able to collaborate with original deposition authors to re-version the records in the PDB. For example, The RNA-binding nucleocapsid phosphoprotein structure (6vyo) featured what the authors describe as a "structural equivalent of a typo" at a metal center: two zinc ions modeled where there should be one and its ligands. This type of error has a serious impact on downstream uses of the structure if not corrected. In another case, structure 6w9c of the papain-like protease contained poorly modeled zinc finger domains with incorrect coordination numbers. The authors' corrections to this model were later validated by a higher resolution structure of the same macromolecule. The same happened for 6w41, a spike receptor binding domain, in which the authors corrected conformations of disulfide staples and hydrogen bonding of several peptides. Among the more egregious errors they identified was a register shift in structure 7btf of the RNA-dependent RNA polymerase. They also re-refined another structure of the polymerase in complex with RNA and remdesivir, which came with the unique challenge of having remdesivir bound at about half occupancy, complicating analyses near the active site.

While some of the modeling errors they encountered could have been discovered with automated validation tools, there are hard limits. For example, real space correlation coefficients can identify atoms for which there is not sufficient density present, but not density in which atoms should be modeled. It is also possible to strike the wrong balance between fit to data and fit to priors, which necessarily varies with resolution. The authors explored one such case in depth: at low resolution, there is insufficient justification to model unusual rotamers, but the authors of 7bv2 enforced trans conformations of *all* peptides, including two prolines that are better modeled as cis. In the extreme case, the statistically unlikely total absence of outliers is itself a cause for concern (http://dx.doi.org/10.1016/j.str.2020.08.005). Unfortunately, adding new metrics to address overuses of existing metrics will never be able to identify all possible problems with a structure; on the contrary, each new metric lends itself automatically to a new mode of overfitting. The authors emphasize that the "gold standard" for validation of a structure should always involve the examination of every residue by human eyes at least once.

The accelerated pace of publishing in particular has the potential to be a double-edged sword: readers get early access to research — after all, no amount of money ever bought a second of time — but without the security that traditional peer review has already occurred. Now, the scientific community has to decide if it's going to step up or not. To enable the heroic teams of amenders to assemble, structural biologists should make available their structures as early as possible (https://asapbio.org/asappdb).