Review of: "A machine learning platform to estimate anti-SARS-CoV-2 activities"

Christina Eckhardt[1]

1 Columbia University

Potential competing interests: The author(s) declared that no potential competing interests exist.

Govinda et al. leveraged data from the National Center for Advancing Translational Sciences (NCATS) Covid-19 portal to develop machine learning models that estimate anti-SARS-COV-2 activity of molecular compounds. The models evaluate the degree to which different molecular compounds target viral infectivity, viral entry, viral replication, and human cell toxicity. The platform provides a robust and rapid approach to screening compounds for anti-SARS-COV-2 activity, and may streamline identification and development of novel pharmacologic interventions. I applaud the authors on developing a novel platform that has potential to accelerate identification of new Covid-19 therapies. My concerns relate primarily to the structure of the manuscript and potential undescribed limitations of the models. Below are some minor comments:

Minor Comments

1. The first paragraph of the discussion reads like a continuation of the Results section. It is also somewhat atypical to introduce a new figure in the Discussion. I suggest moving the enumeration of Figure 4 to the Results section.
2. Consider opening the discussion with a summary of how the computational models were developed and how REDIAL-2020 will help identify new therapeutic candidates for Covid-19.
3. Please label each panel in Figure 4 so it is easier to visually distinguish the information being presented in each panel
4. Please clarify in the legend for Figure 4 that each point in each panel represents an individual compound.
5. Please introduce Table 2 in the Results section rather than the Discussion.
6. The authors acknowledge the possibility that intra- and inter-experiment variability may impact the applicability of their models to external compounds. Please expand upon how this impacts the applicability of the models. Please also compose a discrete paragraph in the Discussion that delineates the limitations of the proposed models.
7. The second-to-last sentence in the discussion state “For example by using the TMPRSS2 inhibition assay[38] data from the NCATS COVID-19 portal or additional NCATS data as they become available in the public domain.” I suspect the authors are suggesting they may include additional machine learning

models in future versions of REDIAL-2020, but this is not entirely clear. Please revise this sentence to clarify the meaning.

8. Do similar models designed to identify new therapeutic compounds for Covid-19 exist in the literature? If so, please provide a comparison of REDIAL-2020 to existing machine learning models.