Summarizing known host-viral protein-protein interactions into an online atlas**:** We are strategically positioned to fulfill the need for an atlas of molecular interactions in SARS-CoV-2, with the unique structural and computational perspectives afforded by the collaboration between the Xiong and Gerstein laboratories. Our ongoing work to unravel HIV host-viral interactions will provide insights as HIV is the most well-studied virus and thus, is a valuable comparative resource for understanding SAR-SCoV-2 host-viral interactions 1. More specifically, several of the host proteins we have extensively studied31 were identified in a SARS-CoV-2 host-viral interactome 2. We are also very experienced in developing online resources for data-driven analyses, such as a database of molecular movements 3-25 (http://www.molmovdb.org/ — including GenoDock 26: http://genodock.molmovdb.org/), a pseudogene database 27 (http://pseudogene.org/), the Function-based prioritization of Sequence variants (FunSeq) and family of related tools 28.29 (http://funseq.gersteinlab.org/), and networks research 30-33 (http://networks.gersteinlab.org/ — including Publication Network Graph Utility [PubNet] 34: http://pubnet.gersteinlab.org/) resources. Our compendium will assist in combating the ongoing COVID-19 pandemic by creating key resources for the development of an effective SARS-CoV-2 vaccine and other therapeutics. The SHI atlas will be an educational hub, disseminating SARS-CoV-2 PPI information in a freely accessible and user-friendly format. The PIs’ interdisciplinary expertise in studying host-viral PPIs (Structural Interaction Network [SIN] 35, Dynamic SIN [DynaSIN] 23, tYNA 31, and TopicNet 36) ensures that the SHI atlas will be further expanded and kept up to date for the COVID-19 research community.

Applying RNA-seq clustering and machine learning to identify further interactions: We recently applied a simplified version of this approach to a similar type of RNA-seq dataset from asthmatic patients to identify the interactions between microbes and host genes. Now, we plan to extend this method and apply it to COVID-19. In our preliminary work, we analyzed the heterogeneous data of sputum samples collected from the airway of asthmatic patients. We applied our LDA-link tool and identified connections between genes and microbes reported elsewhere in the literature as well as novel observations. Notably, both fungi and bacteria showed these links, further highlighting the need to evaluate more than just bacteria when performing microbiome experiments in the airway. Several of the observed linkages showed strong literature precedence. For example, the gene lactotransferrin was linked to Aeromonas 37, the gene MUC6 was linked to Burkholderia 38 , the NFKB Inhibitor Zeta was linked to Haemophilus 39 , and IL1B was linked to Pasteurella 40 . We also identified several novel linkages, such as between IL1B and Haemophilus and between GCSAML and Candida. In addition to single gene-microbe pairs, we layered on pathway and cell deconvolution data to identify larger-scale effects of microbes.

Studying viral sequence variability and evolution to identify candidate key positions in the viral

genome and their prevalence across different lineages: In the context of New York State, we downloaded all 1,400 sequences available as of June 29th and aligned them in MAFFT. We obtained 50 reference sequences from GISAID on June 29th and added them to the existing New York sequence alignment (**Figure 3 a,b,c**). We analyzed the resulting alignment in BEAST (v1 and v2) using tip dates in the sequence header corresponding to the sample collection date. We used a strict molecular clock in all analyses along with a GTR substitution model and a gamma + invariant site heterogeneity model. We tested different tree models to find the best fit for the data. These included a coalescent logistic growth model and a coalescent extended Bayesian skyline plot. We analyzed the MCMC log files created by BEAST using Tracer to assess the validity of the tree based on estimated sample size values and the convergence of the posterior distribution (**Figure 3**). Using selected and representative reference sequences from the worldwide pandemic (**Figure 3b**), our preliminary results suggest that the large majority of New York viral sequences form a single outbreak and also suggest distinct lineages within that same local outbreak (**Figure 3c**).

****

**Figure 3**. (**a**) An unrooted phylogenetic tree based on a global subsampling of 3,746 SARS-CoV-2 genomes sampled between Dec. 2019 and July 2020, downloaded from nextstrain.org on 07/17/2020. According to nextstrain.org, worldwide sequences are divided into 5 main lineages (19A, 19B, 20A, 20B and 20C) denoting different outbreaks of the pandemic.  (**b**) Our selected 50 reference sequences spanning the nextstrain.org tree across the 5 main lineages. (**c**) Alignment of the 50 reference sequences from (**b**) with 1480 sequences from the New York outbreak, available as of 06/29/20 on GISAID. We inferred a phylogenetic tree using an extended Bayesian skyline coalescent model under a strict clock and constraints on tip dates. New York taxa labels are colored by date (dark blue for February and March, light blue for April, green for May). Reference sequences are colored according to their clade of origin. Our tree suggests that New York sequences largely form their respective outbreak event, clustering within group 20A.

Globally evaluating the constraints on host sequence variation: Our work on host protein variation will be based on our previous studies in this area. In particular, we have extensively analyzed patterns of genetic variation in non-coding regions, along with their coding targets. Our work employs metrics such as diversity and fraction of rare variants to characterize selection on various classes and subclasses of functional annotations. In recent studies, we developed a collection of tools that can identify sensitive and ultra-sensitive regions (i.e., those annotations under strong selective pressure, as determined using genomes from many individuals from diverse populations) 29,41,42. FunSeq links each noncoding mutation to target genes and prioritizes such variants based on scaled network connectivity 29,41. This tool identifies deleterious variants in many non-coding functional elements, including transcription factor (TF) binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitive sites. FunSeq also detects the disruptiveness of variants in TF binding sites (both LoF and gain-of function events). By integrating large-scale data from various resources (including ENCODE and the 1000 Genomes Project) with cancer genomics data, our method is able to prioritize the known TERT promoter driver mutations and score somatic recurrent mutations higher than those that are non-recurrent.

Mapping prioritized mutations onto the viral and host proteins and determining their impact on

structural and dynamic properties, using physically based and deep-learning models: Recently, we have developed a computational method, ThermoNet 43, for predicting how variants can impact the stability of a protein. ThermoNet is a deep 3D-convolutional neural network designed for structure-based prediction of the change in protein stability upon a point mutation. We propose to apply ThermoNet to the population variants in the complex structures of the key receptor proteins and PPIs to explore to what extent the variants could protect or render individuals more susceptible to SARS-CoV-2.

**References**

1. Fung, S. Y., Yuen, K. S., Ye, Z. W., Chan, C. P. & Jin, D. Y. A tug-of-war between severe acute respiratory syndrome coronavirus 2 and host antiviral defence: lessons from other pathogenic viruses. *Emerg Microbes Infect* **9**, 558-570, doi:10.1080/22221751.2020.1736644 (2020)
2. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459-468, doi:10.1038/s41586-020-2286-9 (2020).
3. Gerstein, M., Lesk, A. M. & Chothia, C. Structural mechanisms for domain movements in proteins. *Biochemistry* **33**, 6739-6749, doi:10.1021/bi00188a001 (1994).
4. Gerstein, M. & Krebs, W. A database of macromolecular motions. *Nucleic Acids Res* **26**, 4280-4290, doi:10.1093/nar/26.18.4280 (1998).
5. Gerstein, M. & Chothia, C. Perspectives: signal transduction. Proteins in motion. *Science* **285**, 1682-1683, doi:10.1126/science.285.5434.1682 (1999).
6. Krebs, W. G. & Gerstein, M. The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Res* **28**, 1665-1675, doi:10.1093/nar/28.8.1665 (2000).
7. Krebs, W. G. *et al.* Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. *Proteins* **48**, 682-695, doi:10.1002/prot.10168 (2002).
8. Echols, N., Milburn, D. & Gerstein, M. MolMovDB: analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Res* **31**, 478-482, doi:10.1093/nar/gkg104 (2003).
9. Krebs, W. G. *et al.* Tools and databases to analyze protein flexibility; approaches to mapping implied features onto sequences. *Methods Enzymol* **374**, 544-584, doi:10.1016/S0076-6879(03)74023-3 (2003).
10. Alexandrov, V. & Gerstein, M. Using 3D Hidden Markov Models that explicitly represent spatial coordinates to model and compare protein structures. *BMC Bioinformatics* **5**, 2, doi:10.1186/1471-2105-5-2 (2004).
11. Gerstein, M. & Echols, N. Exploring the range of protein flexibility, from a structural proteomics perspective. *Curr Opin Chem Biol* **8**, 14-19, doi:10.1016/j.cbpa.2003.12.006 (2004).
12. Goh, C. S., Milburn, D. & Gerstein, M. Conformational changes associated with protein-protein interactions. *Curr Opin Struct Biol* **14**, 104-109, doi:10.1016/j.sbi.2004.01.005 (2004).
13. Alexandrov, V. *et al.* Normal modes for predicting protein motions: a comprehensive database assessment and associated Web tool. *Protein Sci* **14**, 633-643, doi:10.1110/ps.04882105 (2005).
14. Flores, S. *et al.* The Database of Macromolecular Motions: new features added at the decade mark. *Nucleic Acids Res* **34**, D296-301, doi:10.1093/nar/gkj046 (2006).
15. Flores, S. C., Lu, L. J., Yang, J., Carriero, N. & Gerstein, M. B. Hinge Atlas: relating protein sequence to sites of structural flexibility. *BMC Bioinformatics* **8**, 167, doi:10.1186/1471-2105-8-167 (2007).
16. Flores, S. C. & Gerstein, M. B. FlexOracle: predicting flexible hinges by identification of stable domains. *BMC Bioinformatics* **8**, 215, doi:10.1186/1471-2105-8-215 (2007).
17. Flores, S. C. *et al.* HingeMaster: normal mode hinge prediction approach and integration of complementary predictors. *Proteins* **73**, 299-319, doi:10.1002/prot.22060 (2008).
18. Keating, K. S., Flores, S. C., Gerstein, M. B. & Kuhn, L. A. StoneHinge: hinge prediction by network analysis of individual protein structures. *Protein Sci* **18**, 359-371, doi:10.1002/pro.38 (2009).
19. Bhardwaj, N. & Gerstein, M. Relating protein conformational changes to packing efficiency and disorder. *Protein Sci* **18**, 1230-1240, doi:10.1002/pro.132 (2009).
20. Abyzov, A., Bjornson, R., Felipe, M. & Gerstein, M. RigidFinder: a fast and sensitive method to detect rigid blocks in large macromolecular complexes. *Proteins* **78**, 309-324, doi:10.1002/prot.22544 (2010).
21. Fytianos, K., Konstantinou, J., Kaidatzis, J. & Albanis, T. Multiresidue determination of insecticides and fungicides in fruits and vegetables by gas chromatography. *Bull Environ Contam Toxicol* **61**, 684-689, doi:10.1007/s001289900815 (1998).
22. Voss, N. R. & Gerstein, M. 3V: cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Res* **38**, W555-562, doi:10.1093/nar/gkq395 (2010).
23. Bhardwaj, N., Abyzov, A., Clarke, D., Shou, C. & Gerstein, M. B. Integration of protein motions with molecular networks reveals different mechanisms for permanent and transient interactions. *Protein Sci* **20**, 1745-1754, doi:10.1002/pro.710 (2011).
24. Flores, S. C. & Gerstein, M. B. Predicting protein ligand binding motions with the conformation explorer. *BMC Bioinformatics* **12**, 417, doi:10.1186/1471-2105-12-417 (2011).
25. Clarke, D. *et al.* Identifying Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation. *Structure* **24**, 826-837, doi:10.1016/j.str.2016.03.008 (2016).
26. Wang, B. *et al.* Building a Hybrid Physical-Statistical Classifier for Predicting the Effect of Variants Related to Protein-Drug Interactions. *Structure* **27**, 1469-1481.e1463, doi:10.1016/j.str.2019.06.001 (2019).
27. Karro, J. E. *et al.* Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res* **35**, D55-60, doi:10.1093/nar/gkl851 (2007).
28. Bahcall, O. FunSeq for cancer genomics. *Nature Genetics* **45**, 1273-1273, doi:10.1038/ng.2819 (2013).
29. Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**, 480, doi:10.1186/s13059-014-0480-5 (2014).
30. Yu, H., Zhu, X., Greenbaum, D., Karro, J. & Gerstein, M. TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res* **32**, 328-337, doi:10.1093/nar/gkh164 (2004).
31. 61 Yip, K. Y., Yu, H., Kim, P. M., Schultz, M. & Gerstein, M. The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. *Bioinformatics* **22**, 2968-2970, doi:10.1093/bioinformatics/btl488 (2006).
32. 62 Yu, H., Paccanaro, A., Trifonov, V. & Gerstein, M. Predicting interactions in protein networks by completing defective cliques. *Bioinformatics* **22**, 823-829, doi:10.1093/bioinformatics/btl014 (2006).
33. 63 Cheung, K. H. *et al.* YeastHub: a semantic web use case for integrating data in the life sciences domain. *Bioinformatics* **21 Suppl 1**, i85-96, doi:10.1093/bioinformatics/bti1026 (2005).
34. Douglas, S. M., Montelione, G. T. & Gerstein, M. PubNet: a flexible system for visualizing literature derived networks. *Genome Biol* **6**, R80, doi:10.1186/gb-2005-6-9-r80 (2005).
35. Kim, P. M., Lu, L. J., Xia, Y. & Gerstein, M. B. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314**, 1938-1941, doi:10.1126/science.1136174 (2006).
36. Lou, S. *et al.* TopicNet: a framework for measuring transcriptional regulatory network change. *Bioinformatics* **36**, i474-i481, doi:10.1093/bioinformatics/btaa403 (2020).
37. Ascencio, F., Ljungh, A. & Wadstrom, T. Characterization of lactoferrin binding by Aeromonas hydrophila. *Appl Environ Microbiol* **58**, 42-47, doi:10.1128/AEM.58.1.42-47.1992 (1992)
38. Sajjan, U., Keshavjee, S. & Forstner, J. Responses of well-differentiated airway epithelial cell cultures from healthy donors and patients with cystic fibrosis to Burkholderia cenocepacia infection. *Infect Immun* **72**, 4188-4199, doi:10.1128/IAI.72.7.4188-4199.2004 (2004).
39. Park, C. Y., Heo, J. N., Suk, K. & Lee, W. H. Sodium azide suppresses LPS-induced expression MCP-1 through regulating IκBζ and STAT1 activities in macrophages. *Cell Immunol* **315**, 64-70, doi:10.1016/j.cellimm.2017.02.007 (2017).
40. Hildebrand, D. *et al.* Granzyme A produces bioactive IL-1β through a nonapoptotic inflammasome-independent pathway. *Cell Rep* **9**, 910-917, doi:10.1016/j.celrep.2014.10.003 (2014).
41. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587, doi:10.1126/science.1235587 (2013).
42. Lou, S. *et al.* GRAM: A GeneRAlized Model to predict the molecular effect of a non-coding variant in a cell-type specific manner. *PLoS Genet* **15**, e1007860, doi:10.1371/journal.pgen.1007860 (2019).
43. Li, B., Yang, Y. T., Capra, J. A. & Gerstein, M. B. Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. *bioRxiv*, 2020.2002.2028.959874, doi:10.1101/2020.02.28.959874 (2020).