frontiers | Frontiers in Bioinformatics

TYPE Original Research
PUBLISHED 24 May 2024
DOI 10.3389/fbinf.2024.1397968

Human cytokine and coronavirus nucleocapsid protein interactivity using large-scale virtual screens

Phillip J. Tomezsko[1*†], Colby T. Ford[2,3,4*†], Avery E. Meyer[1], Adam M. Michaleas[1] and Rafael Jaimes III[1*]

[1]MIT Lincoln Laboratory, Lexington, MA, United States, [2]Tuple LLC, Charlotte, NC, United States, [3]University of North Carolina at Charlotte, Department of Bioinformatics and Genomics, Charlotte, NC, United States, [4]University of North Carolina at Charlotte, Center for Computational Intelligence to Predict Health and Environmental Risks (CIPHER), Charlotte, NC, United States

Check for updates

OPEN ACCESS

EDITED BY
Wen Wei,
Arizona State University, United States

REVIEWED BY
Devlina Chakravarty,
National Library of Medicine (NIH), United States
Pasqualina D'Ursi,
National Research Council (CNR), Italy
Bin Hu,
Los Alamos National Laboratory (DOE),
United States

*CORRESPONDENCE
Phillip J. Tomezsko,
✉ phillip.tomezsko@ll.mit.edu
Colby T. Ford,
✉ colby@tuple.xyz
Rafael Jaimes III,
✉ rafael.jaimes@ll.mit.edu

†These authors have contributed equally to this work

RECEIVED 08 March 2024
ACCEPTED 26 April 2024
PUBLISHED 24 May 2024

CITATION
Tomezsko PJ, Ford CT, Meyer AE, Michaleas AM, Jaimes R III (2024), Human cytokine and coronavirus nucleocapsid protein interactivity using large-scale virtual screens.
Front. Bioinform. 4:1397968.
doi: 10.3389/fbinf.2024.1397968

Understanding the interactions between SARS-CoV-2 and the human immune system is paramount to the characterization of novel variants as the virus co-evolves with the human host. In this study, we employed state-of-the-art molecular docking tools to conduct large-scale virtual screens, predicting the binding affinities between 64 human cytokines against 17 nucleocapsid proteins from six betacoronaviruses. Our comprehensive *in silico* analyses reveal specific changes in cytokine-nucleocapsid protein interactions, shedding light on potential modulators of the host immune response during infection. These findings offer valuable insights into the molecular mechanisms underlying viral pathogenesis and may guide the future development of targeted interventions. This manuscript serves as insight into the comparison of deep learning based AlphaFold2-Multimer and the semi-physicochemical based HADDOCK for protein-protein docking. We show the two methods are complementary in their predictive capabilities. We also introduce a novel algorithm for rapidly assessing the binding interface of protein-protein docks using graph edit distance: graph-based interface residue assessment function (GIRAF). The high-performance computational framework presented here will not only aid in accelerating the discovery of effective interventions against emerging viral threats, but extend to other applications of high throughput protein-protein screens.

KEYWORDS

nucleocapsid (N) protein, SARS-CoV-2, coronavirus, MERS- and SARS-CoV, OC43-CoV, AlphaFold2, protein-protein interaction (PPI), docking

Introduction

SARS-CoV-2 evolution over the course of the pandemic has led to sustained, continued waves of infections. The variants of concern have shown a high degree of mutation relative to the prevailing strain at the time of their emergence. Most research has focused on the impact of mutations on the spike proteins' ability to enter cells and evade antibodies, whether the antibodies are therapeutics or induced from vaccination or induced from prior infection (Wall et al., 2021; Willett et al., 2022). The main reason for the focus on spike is driven by the fact that the most selective pressure on the virus is applied to the spike and the mutations have consequences for the therapeutics that are available. Outside of spike, the nucleocapsid (N) gene of SARS-CoV-2 has had lineage defining mutations in each of the variants of concern (Emma, 2021). A robust platform

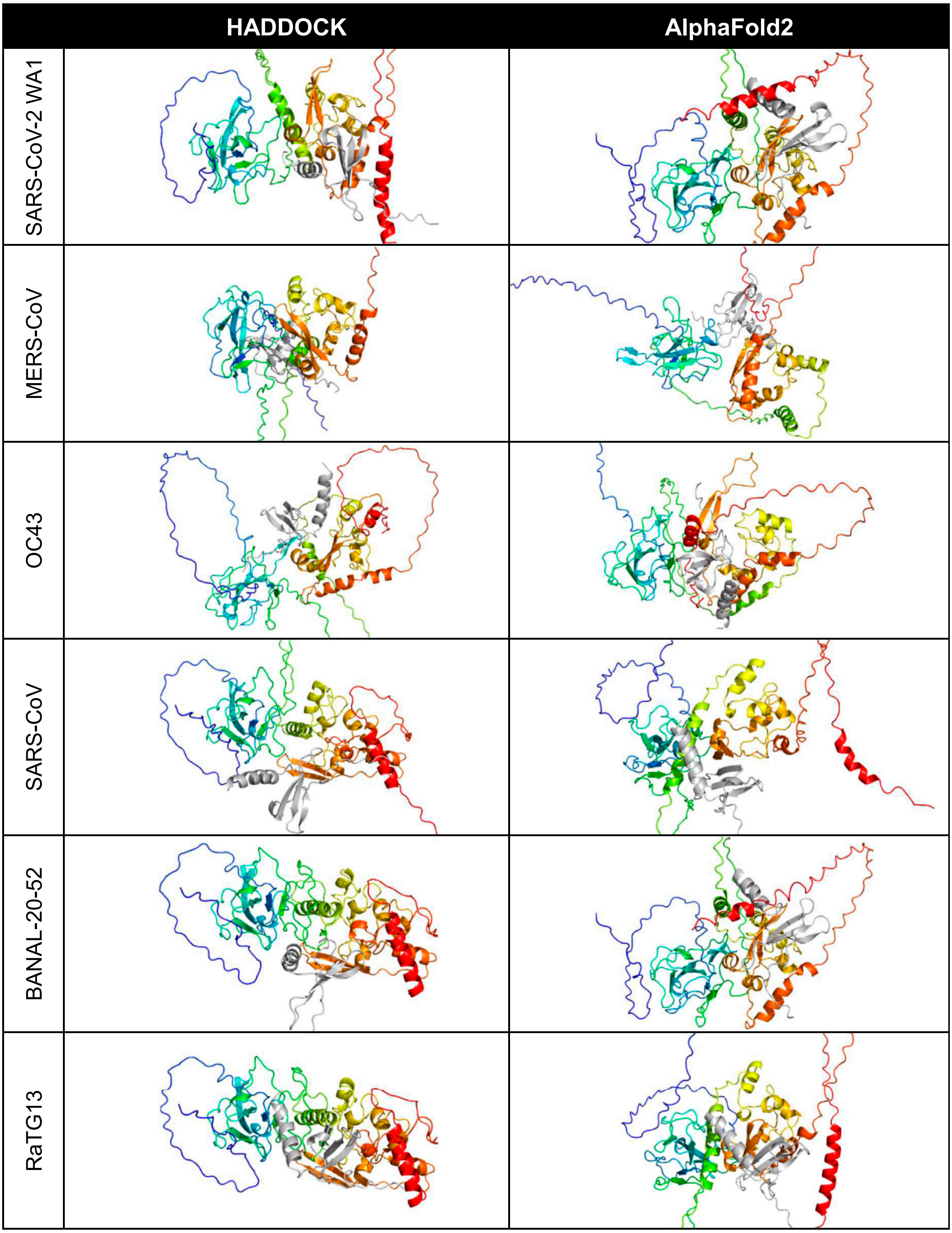

FIGURE 1
Predicted binding site of CXCL12β across selected coronaviruses. The proteins are shown in cartoon style with the N protein colored as rainbow (blue is the amino/N-terminus on the left and red is the carboxyl/C-terminus on the right in each cell) and the CXCL12β cytokine is colored in gray.

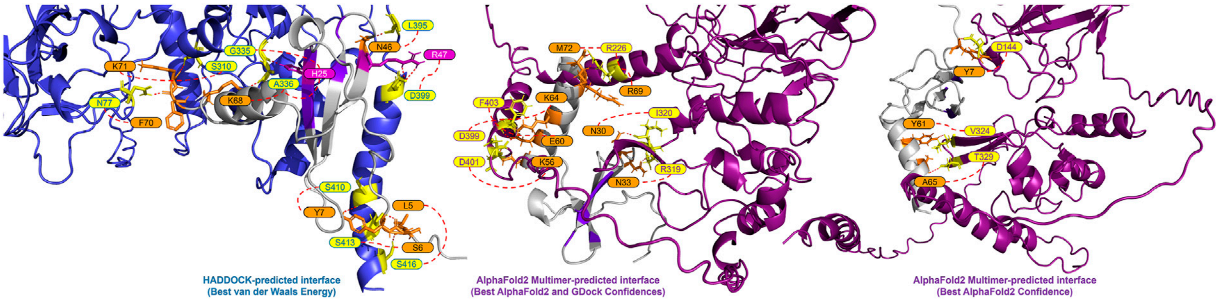
L395
N46
R47
G335
S310
K71
H25
A336
D399
N77
K68
F70
S410
L5
Y7
S413
S6
S416
HADDOCK-predicted interface
(Best van der Waals Energy)
M72
R226
K64
R69
F403
I320
N30
D399
E60
D401
K56
R319
N33
AlphaFold2 Multimer-predicted interface
(Best AlphaFold2 and GDock Confidences)
D144
Y7
Y61
V324
T329
A65
AlphaFold2 Multimer-predicted interface
(Best AlphaFold2 Confidence)

FIGURE 2
Predicted interface between the SARS-CoV-2 WA-1 N protein and CXCL12β. The proteins are shown in cartoon style with the N protein colored as periwinkle in the HADDOCK-predicted complex and dark magenta in the AlphaFold2 Multimer-predicted complexes (with polar interacting residues shown as yellow sticks). The CXCL12β protein is shown in cartoon style in gray (with polar interacting residues shown as orange sticks). Residues in the GAG binding site on CXCL12β are shown in purple (with polar interacting residues shown as magenta sticks).

TABLE 1 Correlations between AlphaFold2-Multimer and HADDOCK docking predictions across various binding metrics. Significant correlations are shown in bold.

p-values/R^2		HADDOCK			AlphaFold2 multimer		
		van der Waals energy	PRODIGY ΔG	FoldX ΔG	van der Waals energy	PRODIGY ΔG	FoldX ΔG
HADDOCK	van der Waals Energy		**0.9867**	0.0830	0.0037	0.0122	0.0384
	PRODIGY ΔG	**<0.0001**		**0.6095**	**0.8560**	**0.8935**	0.0366
	FoldX ΔG	**0.0210**	**<0.0001**		0.0515	0.0419	0.0582
AlphaFold2 Multimer	van der Waals Energy	0.6349	**<0.0001**	0.0713		**0.5741**	0.1605
	PRODIGY ΔG	0.3852	**<0.0001**	0.1045	**<0.0001**		0.0161
	FoldX ΔG	0.1206	0.1267	0.0548	**0.0010**	0.3178	

Bold indicates significant p-values (alpha < 0.05) and their corresponding correlations.

for studying the effects of nucleocapsid mutations on viral RNA delivery and expression has identified mutations in the nucleocapsid gene of Alpha and Delta variants that may contribute to higher transmissibility (Syed et al., 2021).

Another critical function of SARS-CoV-2 N is to bind and sequester a subset of 11 cytokines in order to disrupt immune signaling (Domingo López-Muñoz et al., 2022). Elevated cytokine levels are part of the pathogenesis of severe COVID-19 (Fajgenbaum, 2020). The elevated cytokine levels, a hallmark of cytokine storm, are driven by an inability to resolve infection rapidly and to moderate a proinflammatory immune response (Fajgenbaum, 2020; Merad et al., 2022). The set of cytokines that SARS-CoV-2 binds is shared with another distantly related human betacoronavirus, HCoV-OC43, though HCoV-OC43 also binds to an additional six cytokines with high affinity (Domingo López-Muñoz et al., 2022; Domingo López-Muñoz et al., 2023). Binding to at least one cytokine, CXCL12β, was determined experimentally for human betacoronaviruses SARS-CoV and MERS-CoV in addition to SARS-CoV-2 and HCoV-OC43. Binding of CXCL12β was shown to inhibit chemotaxis of macrophages in a transwell culture study. Other viruses also employ extracellular proteins that bind to specific subsets of cytokines with high affinity, most notably the SECRET-domain containing proteins of the poxviruses (Alejo et al., 2006; Xue et al., 2011). The presence or absence of certain SECRET-domain containing proteins has important consequences on the pathogenicity of the poxvirus.

An *in silico* docking screen of the panel of 11 cytokines tested experimentally against SARS-CoV-2 original reference strain and key variants of concern was developed in the current study. The *in silico* screen was set up on a high-performance computing system. AlphaFold2-Multimer (Evans et al., 2022) and HADDOCK (van Zundert et al., 2016) were utilized in parallel to predict cytokine binding sites on N, as GDockScore (McFee and Kim, 2023) and PRODIGY (Vangone and MJJ Bonvin, 2015) were used to further assess the properties of the predicted cytokine binding. The goals were to identify the cytokine binding sites of the experimental cytokine hits and determine if their binding has been impacted by continuing evolution in the human host over the course of the pandemic. N proteins from human betacoronaviruses and close relatives of SARS-CoV-2 were included in the *in silico* screen as well to test the conservation of cytokine binding in the broader betacoronavirus family. Finally, we wanted to test the ability

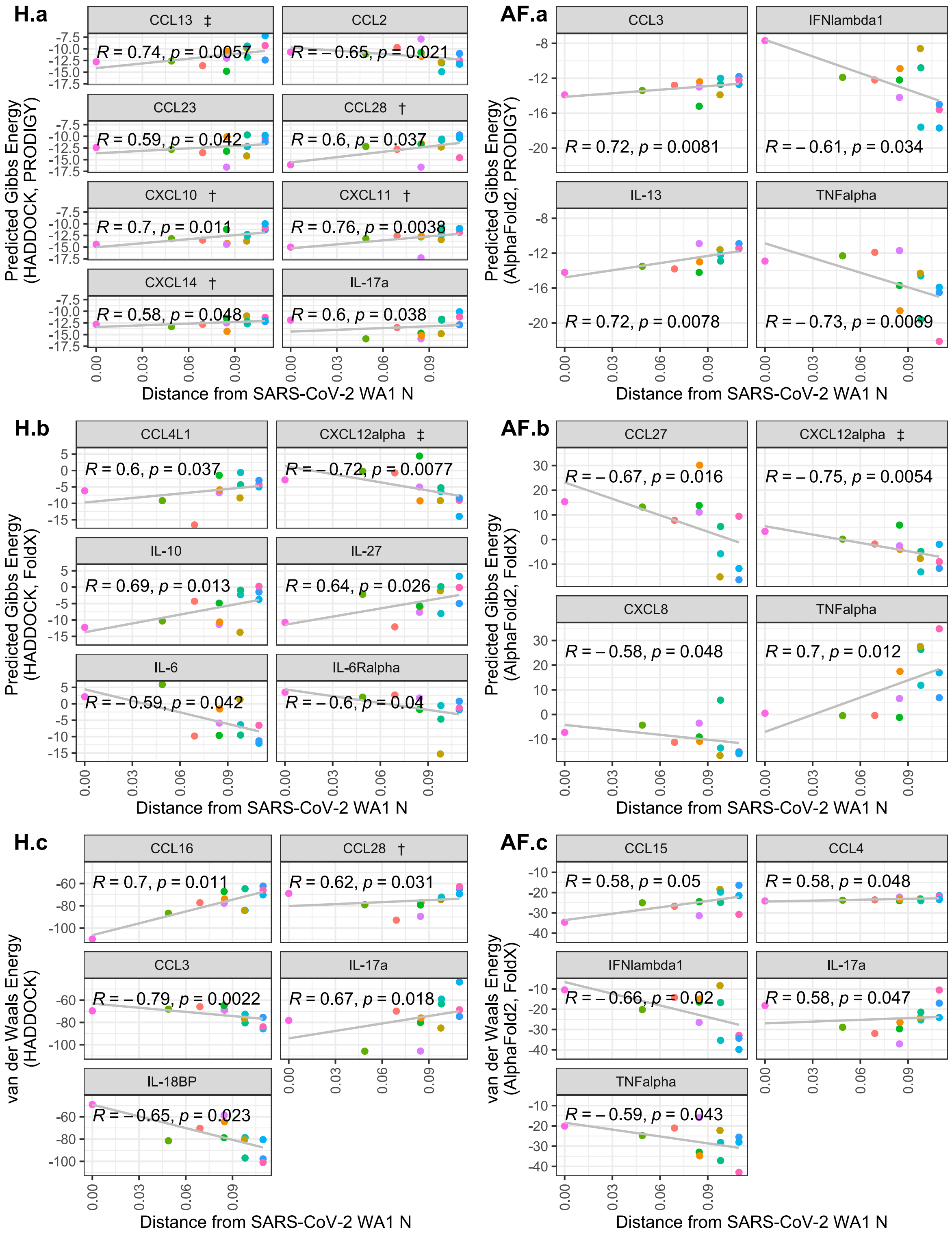

FIGURE 3
Scatterplots of cytokine binding affinity for all cytokines with a statistically-significant correlation to genetic distance. Shown is the relationship between the Fitch distance from the SARS-CoV-2 WA1 N variant vs. HADDOCK and AlphaFold2 predicted binding affinity metrics. Each point represents a SARS-CoV-2 variant, with the most distant points representing Omicron subvariants. Cytokines referenced in López-Muñoz et al. (2022) are denoted by † and ‡ for WA-1/OC43 and OC43 only hits, respectively.

of the *in silico* tools to identify interactions with cytokines using the full 64 cytokine panel utilized experimentally. The full cytokine panel docking screen could be utilized in order to track how evolution of N impacts this function continuing forward in the pandemic as well test other viral proteins for the ability to sequester cytokines.

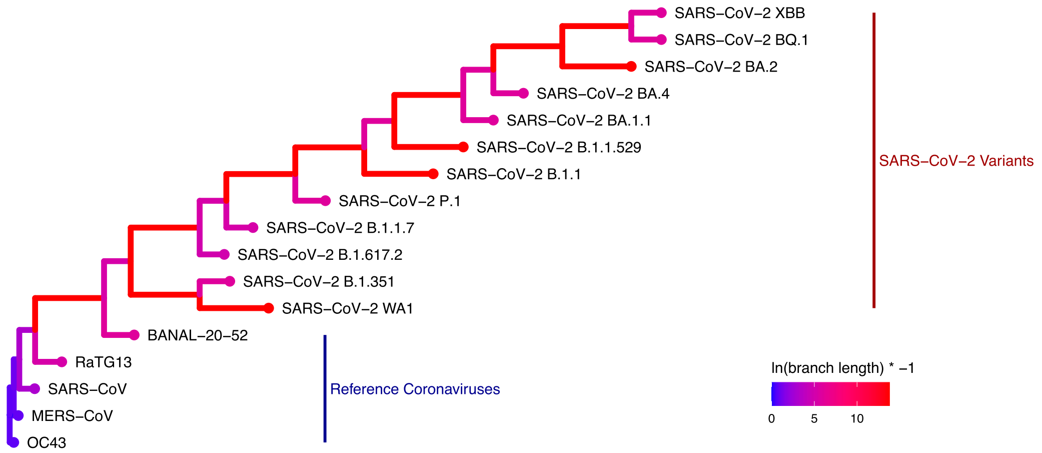

FIGURE 4
Phylogenetic tree generated with RAxML depicting the genetic distance of the 17 coronavirus N proteins used in this study (Colors represent the log-transformed branch length: ln (branch length) × − 1.).

Results

We predicted 17 N protein structures across six betacoronaviruses (SARS-CoV, SARS-CoV-2, MERS-CoV, HCoV-OC43, RaTG13 and BANAL-20-52), all of which generally matched the expected topology of the N protein (Peng et al., 2020). For example, the predicted structure of the SARS-CoV-2 N WA1 strain had two structured domains that recapitulated two crystallized domains (PDB 6M3M for the N-terminal RNA binding domain, total RMSD 0.851Å and 6WZO for the dimerization domain, total RMSD 1.178Å) as well as three flexible domains with low AlphaFold2 confidence (pTM) that corresponded with the three intrinsically disordered domains. AlphaFold2 was also used to generate the structure of the 64 cytokines. The generated N and cytokine structures were used as inputs for HADDOCK, located here: https://github.com/tuplexyz/SARS-CoV-2_N-Cytokine_Docking/tree/main/haddock/inputs.

Computing benchmarks

Running on the MIT SuperCloud high-performance computing (HPC) environment (Reuther et al., 2018), significant throughput was achieved for the numerous computational docking tasks.

AlphaFold2 performance

AlphaFold2-Multimer jobs took 493 ± 21 min of walltime per multimer run on nodes containing an Nvidia Volta V100 GPU, which contain 5,120 CUDA cores and 640 tensor cores. To complete the 1,088 AlphaFold2-Multimer experiments, 36 nodes were used, completing the entire set of experiments in approximately 12 days. A later subset was run on Ampere A100 GPUs to investigate if the newer generation GPU would increase speedups; the mean runtime decreased to 378min ±80 min (noting that the variability of duration increased). The A100 GPUs have more CUDA cores but fewer tensor cores, 6,912 and 432 respectively.

HADDOCK Performance

HADDOCK jobs took 707 ± 40 min of walltime on Intel Xeon Phi 7,210 cluster nodes, which contain 64 physical cores. On Intel Xeon Platinum 8,260 nodes, which contain 48 physical cores, HADDOCK took 133 ± 92 min of walltime. The speedup on the Xeon Platinum nodes was likely due to the much newer CPU architecture and higher maximum clockspeed of the Xeon Platinum (3.9 GHz) compared to the Xeon Phi (1.5 GHz). While HADDOCK performed as quickly as 74 min on the Xeon Platinum nodes, it had a wide range of runtimes, depending on the cytokine and number of surface residues selected against which to dock. Cytokines with multiple chains (IL-12p70, IL-23, IL-27, and IL-35) took approximately three times as long. The 1,088 HADDOCK docking experiments were run on 64 of the Xeon Platinum 8,260 nodes, completing the entire set of experiments in approximately 36 h.

Nucleocapsid cytokine binding sites

HADDOCK and AlphaFold2-Multimer were used to identify potential binding sites of the cytokines with the nucleocapsid proteins (Figure 1). CXCL12β was chosen for initial representative structure modeling because it has been experimentally shown to bind to HCoV-OC43 and SARS-CoV-2 WA1 nucleocapsid in bio-layer interferometry assays (Domingo López-Muñoz et al., 2022; Domingo López-Muñoz et al., 2023). SARS-CoV, MERS-CoV, HCoV-OC43 and SARS-CoV-2 nucleocapsid were additionally shown to inhibit CXCL12β dependent migration of macrophages in a transwell assay (Domingo López-Muñoz et al., 2022; Domingo López-Muñoz et al., 2023). For the six betacoronaviruses (using SARS-CoV-2 WA1 as the representative strain of SARS-CoV-2), HADDOCK docked CXCL12β to locations around the dimerization interface, with contributing contacts from other domains of the protein. AlphaFold2-Multimer complexed the cytokine consistently at the dimerization interface, including interacting residues in the beta sheets of both the nucleocapsid dimerization interface and CXCL12β.

A closer look at the HADDOCK and AlphaFold2-Multimer docking of CXCL12β showed that HADDOCK did not include any beta sheet residues from the dimerization interface, but

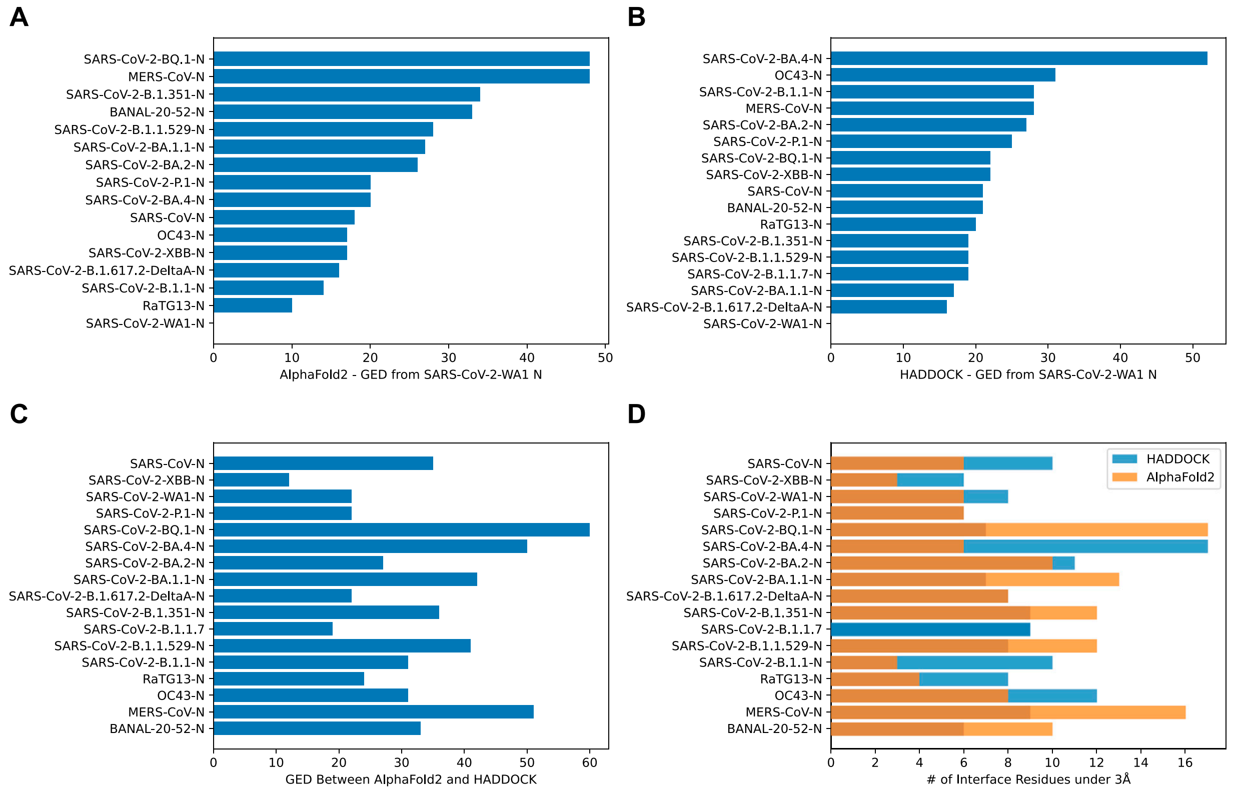

FIGURE 5
Bar charts depicting the graph edit distance for CXCL12β using GIRAF. **(A)** AlphaFold2 GED from SARS-CoV-2 WA1 N. Note, there were no interface residues less than 3Å for SARS-CoV-2-B.1.1.7. **(B)** HADDOCK GED from SARS-CoV-2 WA1 N. **(C)** GED between AlphaFold2 and HADDOCK SARS-CoV-2 WA1 N. **(D)** Number of interface residues under 3Å.

included polar contacts with residues on either side of the interface loop (S310, G335 and A336) (Figure 2). Other interacting residues were located in the structured N-terminal RNA binding domain and the flexible C-terminal region. The AlphaFold2-Multimer complex included two polar contacts of residues directly involved in dimerization (R319 and I320), as well as resides in the flexible C-terminal region. Two residues on CXCL12β involved in glycosaminoglycan (GAG) binding (H25 and R47) were involved in the HADDOCK predicted polar contacts, whereas the GAG-binding residues showed hydrogen bonding but no polar contacts in the AlphaFold2-Multimer predicted complex (Panitz et al., 2016; Crijns et al., 2020). The SARS-CoV-2 WA1 nucleocapsid CXCL12β interaction was previously shown to be competitive with heparin sulfate and chondroitin sulfate (Domingo López-Muñoz et al., 2022).

Cytokine strength

Across the 64 cytokines that were tested, binding affinities were highly correlated between the AlphaFold2-Multimer predictions and HADDOCK. However, the ranges of the scores may vary significantly. For example, the range of the HADDOCK van der Waals energies is [-109.82, −38.45] compared to [-56.76, −10.46] for the AlphaFold2-Multimer predictions (measured by FoldX). See Table 1.

Genetic distance

Correlating the cytokine binding affinity metrics with Fitch distance of each SARS-CoV-2 variants' N protein from the SARS-CoV-2 WA1 N protein, there were significant positive correlations for some of the 64 cytokines, including some which were hits in experimental SARS-CoV-2 and/or HCoV-OC43 screens (Domingo López-Muñoz et al., 2022; Domingo López-Muñoz et al., 2023). Namely, CCL28, CXCL10, CXCL11, CXCL13 and CXCL14 from the HADDOCK predictions show that the binding Gibbs energy predicted by PRODIGY weakened as the genetic distance of the N protein increases from SARS-CoV-2 WA1 (Figure 3 H. a). No significant trends were observed for the predicted AlphaFold2 Gibbs energy by PRODIGY of the cytokines from the experimental screens (Figure 3 AF. a). When correlating the Fitch distance to predicted Gibbs energy by FoldX, only CXCL12α showed a significant correlation of the experimental hits. For this cytokine, both HADDOCK and AlphaFold2 predictions yield a negative correlation with genetic distance, indicating a stronger association with CXCL12α over the course of pandemic (Figure 3 H.b and AF. b). The correlation between Fitch distance van der Waals energy was significant for only CCL28 of the experimental hits, again a positive correlation predicted by HADDOCK indicating less favorable binding (Figure 3 H.C and AF. c). Several cytokines not identified in either experimental screen had significant positive correlations with

Fitch distance. However, cytokines that were non-binders in the screen and had increased Gibbs energy or van der Waals energy would still be expected to be non-binders. Nine cytokines that were negative in both screens had a significant negative correlation in one of the three metrics of binding strength: CXCL8, CCL27, IFNλ1, TNFα, CCL2, CCL3, IL-6, IL6Rα and IL-18BP. Experimental validation could verify if any of these cytokines are inhibited variants of SARS-CoV-2 N even though SARS-CoV-2 WA1 N did not inhibit them.

A phylogenetic tree using only N protein sequences was generated using RAxML (Figure 4). The expected phylogenetic topology was observed of a tree generated using both reference coronaviruses like HCoV-OC43, SARS-CoV, and MERS-CoV along with SARS-CoV-2 variants. SARS-CoV-2 variants were grouped in a monophyletic clade for this analysis.

Breadth of CXCL12β binding

Graph Edit Distance (GED) provided a view into how the N-protein receptor (i.e., the binding pocket) to CXCL12β changed between different N-proteins. A novel algorithm was developed in order to determine GED of the bind site, named graph-based interface residue assessment function (GIRAF). Both AlphaFold2 and HADDOCK structures concurred with relatively low GED of SARS-CoV-2-B.1.1 and SARS-CoV-2-B.1.617.2-DeltaA (Figure 5A, B, respectively). AlphaFold2 had high GED for MERS-CoV and SARS-CoV-2-BQ.1, suggesting far evolutionary distance in the binding pocket compared to the SARS-COV2-WA1 baseline (Figure 5A).

There was marked disagreement between AlphaFold2 and HADDOCK for SARS-CoV-2-BQ.1 (Figure 5C). This was partly due to the discrepancy in the number of interfacing residues on the AlphaFold2 structure compared to the HADDOCK structure (Figure 5D). AlphaFold2 and HADDOCK reported a similar number of interfacing residues on other N proteins; both agreed the number of interfacing residues on SARS-CoV-2-XBB and SARS-CoV-2-P.1 as relatively low. AlphaFold2 reported six residues on SARS-CoV-2-WA1 compared to HADDOCK's eight residues.

TABLE 2 Coronavirus strains and NCBI accession numbers.

Virus	Strain	Accession
BatCoV	BANAL-20-52	MZ937000.1
MERS-CoV	England1	NC_038294.1
HcoV	OC43	NC_006213.1
BatCoV	RaTG13	MN996532.2
SARS-CoV	BJ01	AY278488.2
SARS-CoV-2	B.1.1	MT233522.1
SARS-CoV-2	B.1.1.529	OL919777.1
SARS-CoV-2	B.1.1.7	OP879258.1
SARS-CoV-2	B.1.351	MW571126.1
SARS-CoV-2	B.1.617.2	MW931310.1
SARS-CoV-2	BA.1.1	OP969456
SARS-CoV-2	BA.2	OP968961.1
SARS-CoV-2	BA.4	ON324341.1
SARS-CoV-2	BQ.1	OP839317.1
SARS-CoV-2	P.1	MW520923.1
SARS-CoV-2	XBB	OP847716.1

Discussion

In silico structural prediction and molecular docking tools were utilized in order to interrogate the potential evolutionary change of the interaction between SARS-CoV-2 nucleocapsid proteins and 64 human cytokines. Other betacoronaviruses were also included in the analysis, as several human betacoronaviruses had previously been shown to have a similar cytokine inhibitory phenotype. Additionally, HCoV-OC43 was shown to bind 17 human cytokines, including all 11 cytokines that SARS-CoV-2 bound despite being distantly related (Domingo López-Muñoz et al., 2022; Domingo López-Muñoz et al., 2023). We identified the nucleocapsid dimerization domain as an important site of multiple cytokine interactions, with AlphaFold2 consistently identifying the dimerization loop specifically as the interaction site. We also identified five cytokines from the experimental screens (CCL28, CXCL10, CXCL11, CXCL13, CXCL14) that had significantly reduced binding to N as SARS-CoV-2 evolved by at least one *in silico* metric, and only one cytokine from the experimental screens that had increased binding (CXCL12α). CXCL12α was identified in the HCoV-OC43 screen but not the SARS-CoV-2 screen; all of the hits from the SARS-CoV-2 screen with a significant difference in predicted binding energy were reduced. Finally, we identified nine cytokines that were negative in both experimental screens but had increased interactions with N variants. These cytokines could be potential new targets of inhibition by N.

We expected that interactions with N and cytokines would generally get stronger as SARS-CoV-2 co-evolved with humans, similar to mutations in spike increasing affinity for ACE2 (Markov et al., 2023). Contrary to our hypothesis, we found that the many statistically significant changes were weakening of the cytokine interaction as mutations accumulated. Multiple variants of the Omicron group were included in the analysis and were the most distant from the reference WA1 strain. The decreasing capacity for inhibition of cytokine signaling could be at least partially responsible for the decreased severity of disease observed with Omicron. Of the four cytokine that HADDOCK predicted would bind with decreased affinity, CXCL10 stands out as it is a strong predictor of disease severity (Chen et al., 2020; Yang et al., 2020; Wang et al., 2021). Decreased disregulation of CXCL10 by Omicron N could contribute to any related phenotypic change towards lower severity.

In addition to interrogating how the binding might change over the course of the pandemic for the cytokines that were experimentally shown to bind to SARS-CoV-2 WA1 N, we also identified a series of cytokines that may be inhibited by Omicron that were not inhibited by the original strain of SARS-CoV-2. These include predicted increased binding of CXCL8, CCL27, IFNλ1, and TNFα by AlphaFold2 and CCL2, CCL3, IL-6, IL6Rα and IL-18BP by HADDOCK. Though the predicted affinity and/or van der Waals associated is lower for the variants compared to the reference strain, it is not certain that these

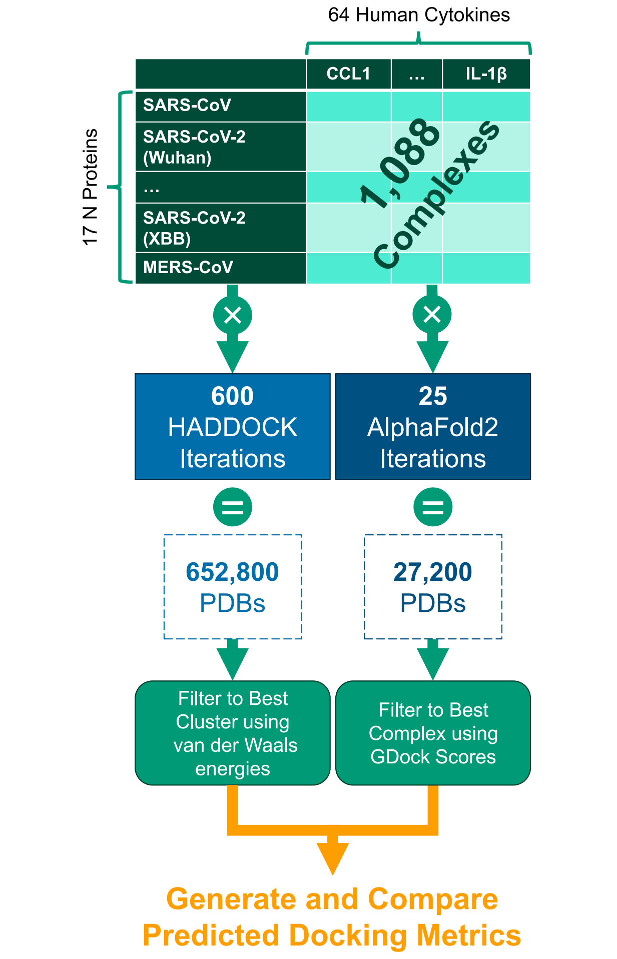

FIGURE 6
Experiment generation workflow for the 1,088 cytokine-N complexes.

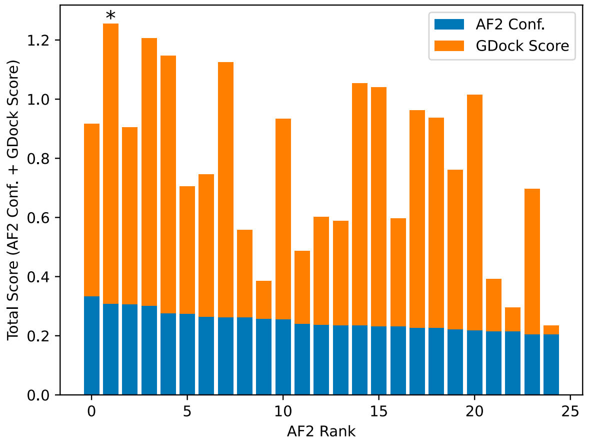

FIGURE 7
Example total score calculation for SARS-CoV-2-WA1 bound to CXCL12β. AlphaFold2 confidence was calculated by ipTM + pTM (see text for details). The AlphaFold2 rank 0 had a slightly higher confidence than the rank 1, however the rank one structure had a significantly higher GDock score which led to a higher total score. In this case, the rank one structure was selected as the best structure.

cytokines would be inhibited because it is unclear if the lower affinities translate to physiological binding and the predicted binding sites may not be conducive to inhibition of cytokine function.

The predicted SARS-CoV-2 N binding to cytokine shares several features with the better characterized SECRET-cytokine binding. The binding to N typically involves a beta sheet to beta sheet interaction. One of the two main beta sheets involved in cytokine binding is the dimerization interface of SARS-CoV-2 N. The beta sheet interaction frequently involves the oligomerization interface of the cytokine. The binding both the GAG-binding domain oligomerization interface stabilized by a flexible arm describe the interaction motif of the SECRET proteins, at least one interaction of which has been crystallized (PDB: 3ONA) (Xue et al., 2011).

Computational approaches to modeling binding are an attractive solution for variant tracking and biosurveillance in order to deal with the huge influx of sequences. Utilization of structural and docking predictions could be used to test the functional interactions as new variant sequences are identified. Previous computational studies have provided accurate early insights into viral mutations and their affects on human health (Piplani et al., 2021; Ford et al., 2022; VanBlargan et al., 2022; Ford et al., 2023; Sugano et al., 2023).

In this work, we have not only showcased the utility of *in silico*-based protein modeling, but also the importance of scalability through high-performance computing. HPC-enabled frameworks, as used here, allow for tremendous throughput improvement, reducing costly and slow lab-based workloads.

This approach will be important for proactively determining when a pathogen has acquired new phenotypes, such as increased transmissibility, pathogenicity, or zoonotic potential. In the long term, robust computational assays for several functions would need to be developed in order to track those higher-order viral features such as transmissibility.

Methods

Overall approach

The docking experiment set in this study contained 64 human cytokines (Supplementary Table S1) and the nucleocapsid (N) protein from 17 different coronaviruses, 12 of which were from SARS-CoV-2 variants (Table 2). The experiment generation consisted of all possible combinations of the cytokines and N proteins for a total of 1,088 cytokine-N protein complexes.

Each experiment was submitted through both AlphaFold2-Multimer and HADDOCK docking systems. These tools generated numerous PDB files of predicted protein complex conformations. From these outputs, the best representative complex PDB structure was selected through various filtering techniques (described below) and then compared across the full experiment set. See Figure 6.

AlphaFold2 protein complex prediction

AlphaFold2 (version 2.3.2) was compiled into a Singularity container and queried against a reference database constructed on 2023-04-12 according to the author's instructions (Jumper et al., 2021). The database was compressed into a SquashFS file and bound directly to the

Singularity container. The specification for the AlphaFold2 Singularity container can be found at: https://github.com/mit-ll/AlphaFold.

AlphaFold2 was run in multimer mode and used default settings to construct five randomly seeded models with five predictions each, for a total of 25 models (PDBs) per complex. Amber relaxation was used on all resulting 25 models. AlphaFold2 was run on both standalone and high performance computing platforms.

For the standalone system, we used an NVIDIA DGX A100 80GB server. The system was equipped with dual AMD Rome 7,742 CPUs, 2TB RAM, and 8 NVIDIA A100-SXM 80GB GPUs.

The MIT Lincoln Laboratory TX-GAIA and MIT SuperCloud (Reuther et al., 2018) systems were the environments used for the HPC-enabled AlphaFold2 pipeline prototyping and performance benchmarking. The compute nodes each consisted of a Intel Xeon Gold 6248 2.5 GHz CPU with 40 cores, 377GB RAM, and Intel Omni-Path with 2 NVIDIA Tesla V100 GPUs.

Signal peptides were identified and removed from cytokine sequences using SignalP6.0 web portal[1] before structure prediction with AlphaFold2 (Teufel et al., 2022).

AlphaFold2 confidence

AlphaFold2 confidence was given as the interface predicted template model score plus predicted template model score (ipTM + pTM). The ipTM was weighted by 80% and the pTM was weighted by 20%, as described by the authors (Evans et al., 2022). The total AlphaFold2 confidence (ipTM + pTM) ranged from [0, 1], where one was the highest confidence.

GDock score

GDockScore, a graph-based deep learning model to assess the docking of two proteins, was included to evaluate AlphaFold2 models (McFee and Kim, 2023). The model was pre-trained by the original authors on docking outputs generated from Protein Data Bank, RosettaDock, HADDOCK decoys, and the ZDOCK Protein Docking Benchmark-to include a wide variety of protein complexes and ensure generalization. *GDockScore* achieved state-of-the-art on the CAPRI Score Set, a challenging dataset for developing docking scoring functions (Lensink and Wodak, 2014). *GDockScore* ranged from 0 to 1, where a score of 0 coincided to an unfeasible complex and a score of one coincided with high probability that the protein complex is comparable to a high quality CAPRI complex. The authors showed a *GDockScore* of under 0.2 approximates to a complex that was unlikely to exist (i.e., bad protein-protein dock) and a *GDockScore* range from 0.2 to one coincided with an increasing predicted quality.

Best AlphaFold2 structure selection

AlphaFold2 ranked the 25 structures by confidence (AF2 conf.). We added the raw AF2 conf. to the GDock score (i.e., total score) for all 25 structures and selected the highest score for each multimer experiment. The highest total score was not necessarily the best ranked from AF2 conf. alone. The best 1,088 structures were used for all downstream analysis. See Figure 7.

HADDOCK docking

For each of the 1,088 experiments generated in this study, HADDOCK v2.4 (van Zundert et al., 2016) was used to dock the N protein and cytokine in each experiment.

HADDOCK, High Ambiguity Driven protein-protein DOCKing, is a biomolecular modeling software that provides docking predictions for provided structures using an information-driven flexible docking approach.

For this study, we utilized a Docker containerized version of HADDOCK, which contains all of the software dependencies to allow HADDOCK to run more readily in an HPC environment. HADDOCK was run on both 64 physical cores (Intel Xeon Phi 7,210) and 48 physical cores (Intel Xeon Platinum 8,260). HADDOCK needed to be compiled against the number of physical cores; more information can be found at: https://github.com/colbyford/HADDOCKer, or on DockerHub at: https://hub.docker.com/r/cford38/haddock.

HADDOCK experiment setup

Run parameters for each experiment were generated programmatically, defining the N protein as the static object around which the cytokine protein is positioned and measured. Other experiment files were also copied or created programmatically that include the scripts to run the docking process, define restraints, and specify input PDB files.

HADDOCK required the definition of active/inactive residue restraints (AIRs) to help guide the protein docking process. To avoid biasing the docking placement of the cytokine, all surface residues were selected as "active" and were then included in the AIR file on which to dock.

The logic for this programmatic generation of HADDOCK experiment files is available in the supplementary GitHub repository.

HADDOCK outputs

The HADDOCK docking process consisted of three steps:

it0: Randomization of orientations and rigid-body minimization.
it1: Semi-flexible simulated annealing through molecular dynamics simulations in torsion angle space.
itw: Refinement by energy minimization in Cartesian space with explicit solvent (i.e., in water).

Each of the above steps generated 200 PDB files of the proteins in a complex for a total of 600 PDB files. From the final, water-refined (*itw*) set of outputs, the "best" PDB file was retrieved from the cluster of predictions with the lowest van der Waals energy. This representative PDB of each N-cytokine complex was then used in subsequent analyses and comparisons.

[1] SingalP6.0 web portal: https://services.healthtech.dtu.dk/services/SignalP-6.0/

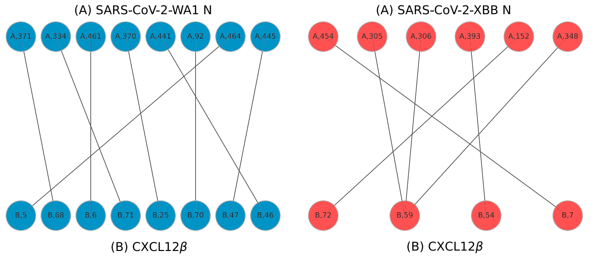

FIGURE 8
Example bigraphs for SARS-CoV-2-WA1 **(A** chain, left) and SARS-CoV-2-XBB **(A** chain, right) bound to CXCL12*β* **(B** chains) on models predicted by HADDOCK.

HADDOCK metrics

The HADDOCK system outputted multiple metrics for the predicted binding affinities and an output set of PDB files containing the N protein docked against the cytokine protein. Some main metrics included:

- van der Waals intermolecular energy (*Evdw*)
- Electrostatic intermolecular energy (*Eelec*)
- Desolvation energy (*Edesol*)
- Restraints violation energy (*Eair*)
- Buried surface area (*BSA*)
- HADDOCK score:
  $1.0Evdw + 0.2Eelec + 1.0Edesol + 0.1Eair$

Prodigy ΔG prediction

PRODIGY, a tool to predict the binding affinity of protein-protein complexes from structural data, was used on each complex in this study (Vangone and MJJ Bonvin, 2015). The predicted binding affinities were reported as Gibbs energy, shown as ΔG (in Kcal/mol units).

The predicted ΔG values were calculated by counting the number of various interfacial contacts (ICs) between the chains of the input complex (N protein and cytokine) along with some properties of the non-interacting surfaces (NIS) using the following equation:

$$\begin{aligned}\Delta G_{predicted} = & -0.09459 \cdot ICs_{charged/charged}\\ & -0.10007 \cdot ICs_{charged/apolar}\\ & +0.19577 \cdot ICs_{polar/polar}\\ & -0.22671 \cdot ICs_{polar/apolar}\\ & +0.18681 \cdot NIS\%_{apolar}\\ & +0.38100 \cdot NIS\%_{charged}\\ & -15.9433\end{aligned}$$

FoldX ΔG prediction

FoldX (Schymkowitz et al., 2005), a tool that evaluates protein-protein complex interactions and their stability, was also used to predict ΔG for all of the complexes. In the command line, the AF2 and HADDOCK structures were first repaired using the PDB repairing function, RepairPDB, which minimized the complex energy by rotating identified side chain residues.

FoldX calculated ΔG for the repaired structures using a linear combination of empirically derived energy terms including van der Waals, solvation energy for apolar and polar groups, and electrostatic contribution of charged groups.

Graph-based interaction residue assessment function (GIRAF)

The interfaces of all complexes were first processed with *INTERCAAT* (Grudman et al., 2021) to produce a full list of all interacting residue pairs between the N-protein and bound cytokine. A bigraph was created using Python NetworkX v3.1 (Hagberg et al., 2008) where the nodes were defined as each residue and the edges (ie. links) were defined as the predicted distance between the residues (in angstroms) by *INTERCAAT*. Distances longer than 3 angstroms were discarded. A lookup table was constructed for the 17 N-proteins using a consensus sequence from multiple sequence alignment (Geneious, Dotmatics, Inc) so that each residue number was aligned with the consensus (Supplementary Data S1). Baseline graphs were generated for SARS-CoV-2-WA1 with CXCL12*β* from both AlphaFold2 and HADDOCK predicted complexes. Graphs were then generated for the other N-proteins with CXCL12*β*. Graph Edit Distance (GED) was used as a metric to determine how the binding pocket of two N-proteins differed. A greater GED was indicative that more residues needed to be swapped in or out to match the binding pocket between a pair of two N-proteins. In other words, a low GED meant that the binding pockets were similar. GED was computed between each of the other N-proteins with SARS-CoV-2-WA1 as a baseline. GED was also computed between each N:CXCL12*β* complex from AlphaFold2 and HADDOCK to evaluate differences in the predicted binding pockets of both methods.

The number of interfacing residues on N-protein from both AlphaFold2 and HADDOCK were enumerated.

GED between a pair of graphs was defined as follows:

$$GED(g_1, g_2) = \min_{(e_1,\ldots,e_k)\in\mathcal{P}(g_1,g_2)} \Sigma_{i=1}^{k} c(e_i)$$

where:

g_1: Computed graph of a baseline complex (e.g., SARS-CoV-2-WA1-N bound to CXCL12*β*.

g_2: Computed graph of a query complex (e.g., SARS-CoV-2-XBB-N bound to CXCL12*β*.

e: Graph edit operation (residue substitution, insertion, or deletion).

k: Step.

P: Set of all possible edit paths to match g_1 to g_2.

c: Cost of the edits to match g_1 to g_2, which is minimized.

GED minimization was capped at 10 s as we saw that further significant cost minimization over the set of edit paths did not occur beyond this point. GED costs were equally weighted a value of one for substitutions, insertions, or deletions.

Example bigraphs are shown in Figure 8 where interfacing residues between the N protein and a cytokine, shown as labeled dots, are connected to one another.

Protein structure and results visualization

Protein complexes were visualized using using PyMOL (Schrödinger, 2015). PyMol's analysis capabilities were employed to detect and show interfacing residues (polar interactions within 3.0Å) between the N protein and cytokine in each complex.

Graphs were rendered using the ggplot2 v3.4.2 R package (Hadley, 2016) or Matplotlib v3.8.1 (Hunter, 2007). The phylogenetic tree figure was rendered using the ggtree v3.6.2 R package (Yu et al., 2017).

N protein alignment, genetic distance, and tree generation

The sequences for the N protein of each variant were aligned using Muscle v3.8.425 (Edgar, 2004) and was written in FASTA format. The multiple sequence alignment is available in the (Supplementary Table S2). From the aligned FASTA file, the pairwise distances were computed using the Fitch matrix (Fitch, 1966) from the seqinr v4.2.30 R package (Charif et al., 2007). For Figure 4, the distance from the SARS-CoV-2 WA1 isolate was used as the reference point.

A maximum likelihood phylogenetic tree of the N protein alignment was generated with RAxML v8.2.12 (Stamatakis, 2014). This tree was generated using the "rapid bootstrap" mode with 100 replicates and rooted on the OC43 taxon.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/tuplexyz/SARSCoV-2_N-Cytokine_Docking.

Author contributions

PT: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing. CF: Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing–original draft, Writing–review and editing. AMe: Investigation, Visualization, Writing–original draft, Writing–review and editing. AMi: Writing–original draft, Writing–review and editing, Methodology, Software. RJ: Methodology, Software, Writing–original draft, Writing–review and editing, Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Project administration, Resources, Supervision, Validation, Visualization.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Department of the Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of the Air Force.

Acknowledgments

The authors would like to thank Alberto D. López-Muñoz and Jonathan Yewdell at the National Institutes of Health for guidance on understanding coronavirus nucleocapsid and cytokine interactions. The authors also want to acknowledge Darrell Ricke (MIT LL) for initial help with debugging AlphaFold2, Keegan Quiqley (MIT LL) and Menghua Rachel Wu (MIT CSAIL) for helpful suggestions, the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing resources and consultation, the MIT BeaverWorks facilities, and Jon Halter from the UNC Charlotte University Research Computing group.

Conflict of interest

Author CF was employed by Tuple LLC.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2024.1397968/full#supplementary-material

References

Alejo, A., Ruiz-Argüello, M. B., Ho, Y., Smith, V. P., Saraiva, M., and Alcami, A. (2006). A chemokine-binding domain in the tumor necrosis factor receptor from variola (smallpox) virus. *Proc. Natl. Acad. Sci.* 103 (15), 5995–6000. doi:10.1073/pnas.0510462103

Charif, D., and Lobry, J. R. (2007). "SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis," in Structural approaches to sequence evolution: molecules, networks, populations, *biological and medical physics, biomedical engineering*. Editors U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo (Verlag, New York: Springer), 207–232.

Chen, Y., Wang, J., Liu, C., Su, L., Zhang, D., Fan, J., et al. (2020). IP-10 and MCP-1 as biomarkers associated with disease severity of COVID-19. *Mol. Med.* 29 (26), 97–10. doi:10.1186/s10020-020-00230-x

Crijns, H., Vincent, V., and Paul, P. (2020). Targeting chemokine—glycosaminoglycan interactions to inhibit inflammation. *Front. Immunol.* 11, 483. doi:10.3389/fimmu.2020.00483

Domingo López-Muñoz, A., Kosik, I., Holly, J., and Yewdell, J. W. (2022). Cell surface SARS-CoV-2 nucleocapsid protein modulates innate and adaptive immunity. *Sci. Adv.* 8 (31), eabp9770. doi:10.1126/sciadv.abp9770

Domingo López-Muñoz, A., Santos, J. J. S., and Yewdell, J. W. (2023). Cell surface nucleocapsid protein expression: a betacoronavirus immunomodulatory strategy. *Proc. Natl. Acad. Sci.* 120 (28), e2304087120. doi:10.1073/pnas.2304087120

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinforma.* 5 (1), 113. doi:10.1186/1471-2105-5-113

Emma, B. (2021) *Hodcroft. CoVariants: SARS-CoV-2 mutations and variants of interest*.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., et al. (2022). Protein complex prediction with AlphaFold-Multimer. *bioRxiv*. doi:10.1101/2021.10.04.463034

Fajgenbaum, D. J. (2020). Cytokine storm. *N. Engl. J. Med.* 383 (23), 2255–2273. doi:10.1056/NEJMra2026131

Fitch, W. M. (1966). An improved method of testing for evolutionary homology. *J. Mol. Biol.* 16 (1), 9–16. ISSN 0022-2836. doi:10.1016/S0022-2836(66)80258-9

Ford, C. T., Machado, D. J., and Janies, D. A. (2022). Predictions of the SARS-CoV-2 Omicron variant (B.1.1.529) spike protein receptor-binding domain structure and neutralizing antibody interactions. *Front. Virology* 2. doi:10.3389/fviro.2022.830202

Ford, C. T., Yasa, S., Jacob Machado, D., White, R. A., and Janies, D. A. (2023). Predicting changes in neutralizing antibody activity for SARS-CoV-2 XBB.1.5 using *in silico* protein modeling. *Front. Virology* 3. ISSN 2673-818X. doi:10.3389/fviro.2023.1172027

Grudman, S., Fajardo, J. E., and Fiser, A. (2021). INTERCAAT: identifying interface residues between macromolecules. *Bioinformatics* 38 (2), 554–555. ISSN 1367-4803. doi:10.1093/bioinformatics/btab596

Hadley, W. (2016) *ggplot2: elegant graphics for data analysis*. Berlin, Germany: Springer International Publishing. doi:10.1007/978-3-319-24277-4

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). "Exploring network structure, dynamics, and function using NetworkX," in *Proceedings of the 7th Python in science conference*. Editors G. Varoquaux, T. Vaught, and J. Millman (Pasadena, CA USA: Scipy), 11–15.

Hunter, J. D. (2007). Matplotlib: a 2d graphics environment. *Comput. Sci. Eng.* 9 (3), 90–95. doi:10.1109/MCSE.2007.55

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2

Lensink, M. F., and Wodak, S. J. (2014). Score_set: a CAPRI benchmark for scoring protein complexes. *Proteins Struct. Funct. Bioinforma.* 82 (11), 3163–3169. doi:10.1002/prot.24678

Markov, P. V., Ghafari, M., Lythgoe, K., Simmonds, P., Stilianakis, N. I., Katzourakis, A., et al. (2023). The evolution of sars-cov-2. *Nat. Rev. Microbiol.* 21, 361–379. doi:10.1038/s41579-023-00878-2

McFee, M., and Kim, P. M. (2023). GDockScore: a graph-based protein–protein docking scoring function. *Bioinforma. Adv.*, 3(1):vbad072. doi:10.1093/bioadv/vbad072

Merad, M., Blish, C. A., Sallusto, F., and Iwasaki, A. (2022). The immunology and immunopathology of covid-19. *Science* 375 (6585), 1122–1127. doi:10.1126/science.abm8108

Panitz, N., Theisgen, S., Samsonov, S. A., Gehrcke, J.-P., Baumann, L., Bellmann-Sickert, K., et al. (2016). The structural investigation of glycosaminoglycan binding to CXCL12 displays distinct interaction sites. *Glycobiology* 26 (11), 1209–1221. ISSN 0959-6658. doi:10.1093/glycob/cww059

Peng, Y., Du, N., Lei, Y., Dorje, S., Qi, J., Luo, T., et al. (2020). Structures of the SARS-CoV-2 nucleocapsid and their perspectives for drug design. *EMBO J.* 39 (20), 1059388–e106010. doi:10.15252/embj.2020105938

Piplani, S., Singh, P. K., Winkler, D. A., and Petrovsky, N. (2021). *In silico* comparison of SARS-CoV-2 spike protein-ACE2 binding affinities across species and implications for virus origin. *Sci. Rep.* 11 (1), 13063. ISSN 2045-2322. doi:10.1038/s41598-021-92388-5

Reuther, A., Kepner, J., Byun, C., Samsi, S., Arcand, W., Bestor, D., et al. (2018). "Interactive supercomputing on 40,000 cores for machine learning and data analysis," in 2018 IEEE High Performance extreme Computing Conference (HPEC), Boston, MA, USA, September 25-29, 2023, 1–6.

Schrödinger, L. L. C. (2015) *The PyMOL molecular graphics system*. version 1.8.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Res.* 33 (2), W382–W388. ISSN 0305-1048. doi:10.1093/nar/gki387

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30 (9), 1312–1313. ISSN 1367-4803. doi:10.1093/bioinformatics/btu033

Sugano, A., Murakami, J., Kataguchi, H., Ohta, M., Someya, Y., Kimura, S., et al. (2023). *In silico* binding affinity of the spike protein with ACE2 and the relative evolutionary distance of S gene may be potential factors rapidly obtained for the initial risk of SARS-CoV-2. *Microb. Risk Anal.* 25, 100278. ISSN 2352-3522. doi:10.1016/j.mran.2023.100278

Syed, A. M., Taha, T. Y., Tabata, T., Chen, I. P., Ciling, A., Khalid, M. M., et al. (2021). Rapid assessment of SARS-CoV-2–evolved variants using virus-like particles. *Science* 374 (6575), 1626–1632. doi:10.1126/science.abl6184

Teufel, F., Armenteros, J. J. A., Johansen, A. R., Gislason, M. H., Pihl, S. I., Tsirgos, K. D., et al. (2022). SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* 40, 1023–1025. doi:10.1038/s41587-021-01156-3

VanBlargan, L. A., Errico, J. M., Halfmann, P. J., Zost, S. J., Crowe, J. E., Purcell, L. A., et al. (2022). An infectious SARS-CoV-2 B.1.1.529 Omicron virus escapes neutralization by therapeutic monoclonal antibodies. *Nat. Med.* 28 (3), 490–495. ISSN 1546-170X. doi:10.1038/s41591-021-01678-y

Vangone, A., and Mjj Bonvin, A. (2015). Contacts-based prediction of binding affinity in protein–protein complexes. *eLife* 4, e07454. ISSN 2050-084X. doi:10.7554/eLife.07454

van Zundert, G. C. P., Rodrigues, J. P. G. L. M., Trellet, M., Schmitz, C., Kastritis, P. L., Karaca, E., et al. (2016). The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.* 428 (4), 720–725. doi:10.1016/j.jmb.2015.09.014

Wall, E. C., Wu, M., Harvey, R., Kelly, G., Warchal, S., Sawyer, C., et al. (2021). Neutralising antibody activity against SARS-CoV-2 VOCs B.1.617.2 and B.1.351 by BNT162b2 vaccination. *Lancet* 397 (10292), 2331–2333. ISSN 0140-6736. doi:10.1016/S0140-6736(21)01290-3

Wang, G.-L., Gao, H.-X., Wang, Y.-L., Xiao, W., Liu, Y.-Z., Lu, J.-H., et al. (2021). Serum IP-10 and IL-7 levels are associated with disease severity of coronavirus disease 2019. *Cytokine* 142, 155500. ISSN 1043-4666. doi:10.1016/j.cyto.2021.155500

Willett, B. J., Grove, J., MacLean, O. A., Wilkie, C., De Lorenzo, G., Furnon, W., et al. (2022). SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway. *Nat. Microbiol.* 7 (8), 1161–1179. doi:10.1038/s41564-022-01143-7

Xue, X., Lu, Q., Wei, H., Wang, D., Chen, D., He, G., et al. (2011). Structural basis of chemokine sequestration by CrmD, a poxvirus-encoded tumor necrosis factor receptor. *PLOS Pathog.* 7 (7), 10021622–e1002213. doi:10.1371/journal.ppat.1002162

Yang, Y., Shen, C., Li, J., Yuan, J., Wei, J., Huang, F., et al. (2020). Plasma IP-10 and MCP-3 levels are highly associated with disease severity and predict the progression of COVID-19. *J. Allergy Clin. Immunol.* 146 (1), 119–127.e4. doi:10.1016/j.jaci.2020.04.027

Yu, G., Smith, D., Zhu, H., Guan, Yi, and Lam, T.T.-Y. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36. doi:10.1111/2041-210X.12628