Virulence network of interacting domains of influenza a and mouse proteins

Teng Ann Ng, Shamima Rashid and Chee Keong Kwoh*

School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

There exist several databases that provide virus-host protein interactions. While most provide curated records of interacting virus-host protein pairs, information on the strain-specific virulence factors or protein domains involved, is lacking. Some databases offer incomplete coverage of influenza strains because of the need to sift through vast amounts of literature (including those of major viruses including HIV and Dengue, besides others). None have offered complete, strain specific protein-protein interaction records for the influenza A group of viruses. In this paper, we present a comprehensive network of predicted domain-domain interaction(s) (DDI) between influenza A virus (IAV) and mouse host proteins, that will allow the systematic study of disease factors by taking the virulence information (lethal dose) into account. From a previously published dataset of lethal dose studies of IAV infection in mice, we constructed an interacting domain network of mouse and viral protein domains as nodes with weighted edges. The edges were scored with the Domain Interaction Statistical Potential (DISPOT) to indicate putative DDI. The virulence network can be easily navigated *via* a web browser, with the associated virulence information (LD_{50} values) prominently displayed. The network will aid influenza A disease modeling by providing strain-specific virulence levels with interacting protein domains. It can possibly contribute to computational methods for uncovering influenza infection mechanisms mediated through protein domain interactions between viral and host proteins. It is available at https://iav-ppi.onrender.com/home.

KEYWORDS

influenza a, lethal dose 50, protein, virulence, mouse model, domain-domain interaction

1 Introduction

Influenza A virus (IAV) is a single stranded, positive ribonucleic acid (RNA) virus that is a respiratory pathogen across many species such as humans, swine, and wild waterfowl. It consists of eight genomic segments which encode at least 11 proteins. The structure and organization of the virus particle is shown in Figure 1, which is reproduced here from the work of Jung and Lee (Jung and Lee, 2020).

Hemagglutinin (HA) and neuraminidase (NA) on the viral particle surface, are the proteins responsible for mediating entry into and cleavage from the host cell, respectively. Matrix protein 1 (M1) is a component of the viral envelop while matrix protein 2 (M2) is found below the lipid bilayer of the viral membrane, strengthening it. Together with the nucleoprotein (NP), they form the ribonucleoprotein complex (indicated as vRNP in Figure 1). The final three proteins are the polymerase basic 1 frame 2 (PB1-F2), and non-structural proteins 1 and 2 (NS1 and NS2), respectively.

FIGURE 1
IAV particle and its fully assembled constituent proteins. Genomic RNA segments are shown in green, wrapped around nucleoproteins. The hetero-trimeric RNA-dependent RNA polymerase complex comprising of PB1, PB2, and PA is shown in orange, light and dark blue circles. This figure is by Jung and Lee (Jung and Lee, 2020).

FIGURE 2
Implementation procedure.
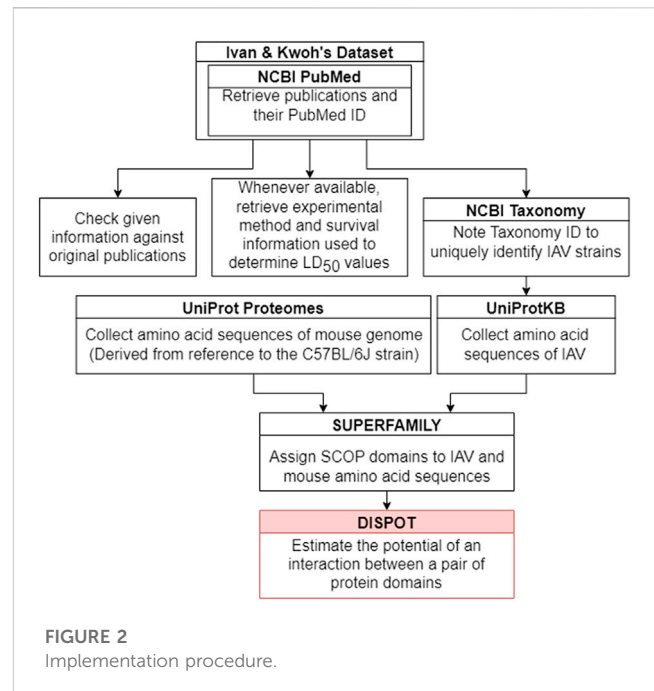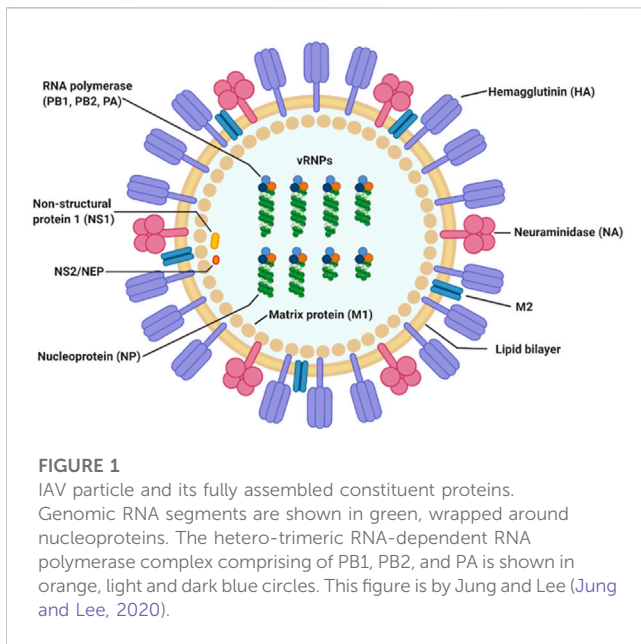
IAV can be highly pathogenic in humans and several highly virulent strains have already caused millions of deaths worldwide in multiple pandemic events. Estimated death tolls for the 1918 (H1N1), 1957 (H2N2) and 1968 (H3N2) pandemics are 50 million (Johnson and Mueller, 2002), 1.1 million (Viboud et al., 2016) and 1-4 million (Rogers, 1968), respectively. Further, IAV triggers various respiratory illnesses seasonally, making it endemic in human populations. The yearly number of deaths due to influenza associated respiratory illness from seasonal influenza has been estimated to vary between nearly 300,000 to 646,000 (Iuliano et al., 2018). Since it is endemic in populations, hosts harboring seasonal influenza strains can act as a reservoir for reassortment events, leading to cross-infection with other circulating pathogens such as SARS-CoV-2 to form potentially harmful recombinant strains (Swets et al., 2022). These attributes highlight the complexity of disease factors of respiratory pathogens and indicate the need of wide-scale influenza studies. They also make the continual monitoring of public health outcomes necessary.

Infectious studies of mouse models can help to elucidate host factors responsible for virulence, since they are cost effective, reproducible and allow mechanistic analyses that may not be directly conducted on humans due to ethical reasons (Sarkar and Heise, 2019). One way to measure the pathogenicity of IAV is by obtaining the lethal dose at which 50% of the inoculated animal test population is infected or perishes (abbreviated here as LD_{50}) (Eugene, 2001). By comparing outcomes of influenza infections in different strains of mice, differences due to allelic variations in mice strains could be possibly be established (Lu et al., 1999).

On one hand, databases such as HPIDB (Ammari et al., 2016) and STRING Viruses (Cook et al., 2018) besides several others have already covered the interactions between influenza A and human proteins in an extensive manner. In comparison, the interactions in mouse hosts are lacking. There exist very few database records of IAV-mouse interactions (for both experimental and computational methods).

On the other hand, it is challenging to directly study the effect of influenza A virulence in human hosts owing to ethical considerations. Mice have been used to infer disease pathology of IAV in humans (Lu et al., 1999). While mice contain significant differences in body size and distribution that affect tissue tropism in pathogenesis (Masemann et al., 2020; Perlman, 2016), at present they are widely accepted pre-clinical models for linking virulence levels with IAV-host interactions (Masemann et al., 2020). Collecting IAV-mouse protein interactions provides a practical approach to identifying virulence factors. As an example, after identifying influential mouse host factors from a network of predicted interactions with IAV proteins, the corresponding set of human homologues (target proteins) can be determined from a combination of homology mapping, associated virulence levels and literature evidence. *In-vitro* interactions found to be occurring amongst target proteins and IAV (*via* biochemistry assays or cell cultures) could assist in designing knock-out factors or drug targets that will allow *in-vivo* validation of the interaction in mouse models.

Data records of mouse model infectious studies had been previously collected in an earlier work by F.X. Ivan and C.K. Kwoh (Ivan and Kwoh, 2019). Their study highlighted the role of protein sites of PB2 in influenza virulence by a systematic meta-analysis using rule-based models to predict the virulence. Therefore, a link between macroscopic virulence labels (such as LD_{50} categories) and protein-protein interactions could prove beneficial in understanding the factors contributing to IAV virulence. Domain-domain interaction(s) DDI can be particularly useful because a protein domain is often a discrete functional unit that is modular, and protein-protein interactions rely on combinations of DDI (Itzhaki et al., 2006; Alborzi et al., 2021). Hence here, the network was constructed with domains representing nodes. While 'domain-domain' interactions are by definition a subset of 'protein-protein' interactions, here the quoted terms are used interchangeably, unless specified otherwise.

TABLE 1 Summary statistics of data collected.

# journal publications	57
# IAV subtypes	14
# IAV strains	109
# mouse proteins	2419
# DDIs	1936

TABLE 2 Three-class virulence classification.

Low	Intermediate	High
$LD_{50} > 10^{6.0}$	$LD_{50} \leq 10^{6.0}$ and $LD_{50} > 10^{3.0}$	$LD_{50} \leq 10^{3.0}$

To systematically identify potential interactions between IAV and mouse host-proteins, a protein network consisting of putative DDI between IAV and mouse proteins, scored by the Domain Interaction Statistical Potential (DISPOT) (Narykov et al., 2019) was developed in this work.

The protein network is presented in a clear graphical user interface (GUI) that easily shows the LD_{50} values and interacting protein domains from the C57BL/6J mouse strain as identified by DISPOT. The virulence network of interacting protein domains will assist studies of IAV disease modeling by providing data of putative interacting protein domains that are associated with their LD_{50} values.

The rest of this manuscript is organized as follows. Section 2 details the contents and design and describes the data presented in this database and the data curation procedure. Section 3 details the web server implementation, describing the tools used, graphical user interface layout and functionality. Section 4 details discussion of this data. Section 5 outlines the proposed future work. Section 6 summarizes and concludes this paper.

2 Contents and design

Figure 2 outlines the steps taken to implement the IAV-Mouse protein-protein interaction (PPI) database. The IAV-Mouse PPI web server can be accessed at: https://iav-ppi.onrender.com/home.

Table 1 provides an overview of the data collected. Five out of the eight RNA segments of IAV genome, namely PB1, HA, NA, M1, and NS1 were found to contain the interacting pathogen protein domains. In summary, 31 unique pairs of DDIs were found between seven IAV protein domains and 29 mouse protein domains.

This work built on initial data records of mouse model infectious studies collected in the previous work by F.X. Ivan and C.K. Kwoh (Ivan and Kwoh, 2019). Their same process of assigning virulence levels was followed here. LD_{50} value was the key information needed to identify the virulence class of a specific IAV strain. Virulence was classified as two-class (avirulent/virulent) and three-class (low/intermediate/high) (*shown in* Figure 3). Essentially, the total infection records classified as "virulent" under the two-class problem is the sum of records classified as "intermediate" and "high" under the three-class problem. Likewise, the total infection records classified as "avirulent" is equivalent to the number of records classified as "low". For the three-class virulence classification, LD_{50} thresholds of $10^{3.0}$ and $10^{6.0}$ were applied (*shown in* Table 2), referencing thresholds that are used by World Health Organization (WHO), for classification of influenza virulence in mice (EID_{50} infection unit) (WHO, 2003). LD_{50} infection units include Plaque-forming Unit (PFU), Focus-forming Unit (FFU), 50% Egg Infective Dose (EID_{50}), 50% Tissue Culture Infectious Dose ($TCID_{50}$) and 50% Cell Culture Infectious Dose ($CCID_{50}$), where the equality across all units was assumed.

SUPERFAMILY 2.0 sequence search (https://supfam.org/sequence/search) (Pandurangan et al., 2019) was used to map regions of an amino acid sequence to at least one Structural Classification of Proteins (SCOP) superfamily using the SUPERFAMILY hidden Markov models. SCOP is a
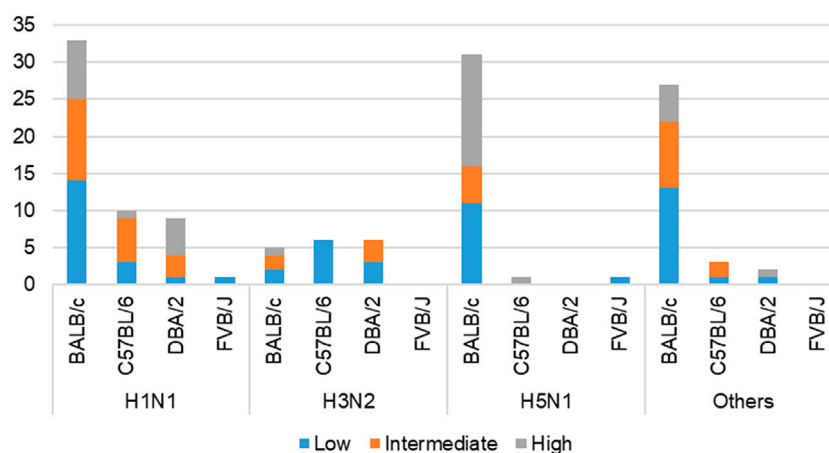

FIGURE 3
Cross-tabulation of IAV subtypes and mouse strains, colored according to three-class virulence classification problem. 'Others' refers to the aggregation of infection records from IAV subtypes—H1N2, H3N8, H5N2, H5N5, H5N6, H5N8, H7N2, H7N3, H7N7, and H7N9.

representation of structure-based hierarchical classification of relationships between protein domains, with "family" being the first level and "superfamily" being the second level. Protein domains from the same SCOP family are strongly related and frequently share the same function (Andreeva et al., 2004).

DISPOT (http://dispot.korkinlab.org/home/pairs) (Narykov et al., 2019) served as the web tool to determine presence of DDIs between pairs of IAV and mouse SCOP superfamily domains. DISPOT uses exclusively DDIs from DOMMINO (Kuang et al., 2012), an in-depth database of structurally resolved macromolecular interactions, where data about DDIs is the amplest, as its source of data. For a given domain pair, DISPOT returns a statistical potential, denoted as the probability P_{ij}. Statistical potentials take values across the entire scale of real numbers. Negative and positive values can be respectively interpreted as having more or less than average number of DDIs in the DOMMINO database. Neutral values are corresponding to the number of DDIs close to the average number. "No information" will be returned instead of a numeric value if the DOMMINO database does not have an entry for the particular domain pair.

The DISPOT calculation of statistical potential formula is given in the equation as follows (Narykov et al., 2019):

$$P_{ij} = \frac{1}{Z_2} \ln \frac{M_{pij}}{M_{mean}}$$

$$\text{where } Z_2 = \sum \sum \ln \frac{M_{pkl}}{M_{mean}}$$

Z_2 is the natural logarithm of observed frequencies of interactions between domains in the DOMMINO database. M_{mean} is the average number of interactions for a pair of domain families, calculated from the non-redundant DOMMINO dataset. Non-redundant refers to two corresponding pairs of domains that do not share 95% or more sequence identity (Narykov et al., 2019).

2.1 Dataset

All 57 journal publications reviewed in this work were retrieved from National Centre for Biotechnology (NCBI) PubMed (Lindberg, 2000), where LD_{50} values were explicitly stated in them. 55 publications referenced the supplementary information given in F.X. Ivan and C.K. Kwoh's publication ("Additional file 5: Supplementary Table S1") (Ivan and Kwoh, 2019), where LD_{50} values were stated as "values given". LD_{50} values reflected in their dataset were checked against the original publications and some missing records were added. Additionally, seven new records from two other papers (Shi et al., 2017) and (Shi et al., 2018) were documented.

2.1.1 Data cleaning

The preliminary dataset presented in this work (https://github.com/tengann/IAV-Host-PPI-Database/blob/main/RawData_2022.xlsm) holds 488 infection records involving wild-type, laboratory, mouse-adapted, recombinant or mutant IAV strains. IAV genomes of wild-type strains are in their natural and non-mutated form while laboratory strains were prepared by means of reverse genetics. Mouse-adapted strains were derived from serial lung-to-lung passages of virus in mice. Genetic amino acid sequences of

mutant virus were changed through point mutations *via* single amino acid substitutions. Recombinant strains were formed by the combination of protein segments from at least two different IAV strains.

The initial dataset was manually curated to only include records involving wild-type or laboratory IAV strains, thereby reducing the number of infections to 190 (Supplementary Figure S1). Subsequently, infection records comprising wild-type or laboratory strains where their Taxonomy identification (ID) number (otherwise known as accession number) (Schoch et al., 2020) could not be found were dropped, further reducing the records to 166 (*shown in* Figure 4). In these cases, it was not possible to retrieve the complete protein sequences of IAV gene segments for SCOP domain assignment *via* SUPERFAMILY 2.0.

Lastly, multiple records concerning the same combination of IAV strain and mouse genome were condensed into a single record, adopting the approach from F.X. Ivan and C.K. Kwoh's publication (Ivan and Kwoh, 2019). From this process, the tally of infection records was reduced to 139 (*shown in* Figure 4). Whenever possible, the majority class of the three-class virulence assignment scheme was selected. Otherwise, the class that is more or most virulent was considered. Next, if only the lower bound of the LD_{50} value was presented, the record with the highest lower bound was selected. For cases where the lowest exact or upper bound of LD_{50} value was provided, the record was selected. The final cleaned dataset containing 109 unique IAV strains was used to derive the network of interacting protein domains.

2.2 Data annotation

Firstly, to distinguish between all journal publications referenced, the NCBI PubMed (https://pubmed.ncbi.nlm.nih.gov/) (Lindberg, 2000) ID number uniquely assigned to each publication record was noted. Information collected in F.X. Ivan and C.K. Kwoh's dataset consists of IAV strain, mouse strain, LD_{50} value and infection unit. Then, in this paper, to provide a deeper insight into how the LD_{50} value was determined in each separate experiment, additional evidence, namely, the experimental method, weight loss and/or survival remarks and LD_{50} calculation method were documented. Also, for each IAV strain, the Taxonomy ID number, a unique ten-digit code that designates classification and specialization was retrieved from NCBI Taxonomy database (https://www.ncbi.nlm.nih.gov/ taxonomy) (Schoch et al., 2020).

Amino acid sequences of both IAV and mouse proteins were retrieved from the UniProt (release 2021_03) protein knowledgebase (UniProtKB) (https://www.uniprot.org/) (Bairoch et al., 2005). IAV protein sequences were retrieved using strain names and/or matching Taxonomy ID number where available. Mouse protein sequences were retrieved using the Proteome ID number, UP000000589. This reference proteome was derived from the genome sequence of mouse strain C57BL/6J, with Taxonomy ID number, 10090. For this work, among 55, 315 protein records that were available, 17, 120 Swiss-Prot gold star reviewed entries (https://www.uniprot.org/uniprotkb?query= UP000000589) were retrieved. Swiss-Prot reviewed refers to records with information fully and manually extracted from literature or curator-evaluated computational analysis (Bairoch and Apweiler,
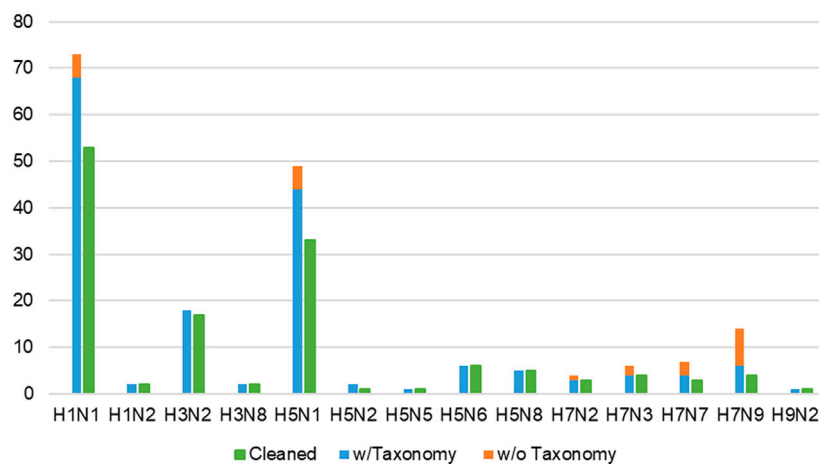
FIGURE 4
Proportion of all wild-type/laboratory IAV strains, separated into with and without taxonomy ID numbers, against remaining records (green bars) in the cleaned dataset. Infection records involving wild-type IAV where taxonomy ID number could not be found (orange bars) were omitted, reducing the number of infection records to 166 (blue bars). "Cleaned" refers to the final dataset of 139 infection records (green bars) available in IAV-Mouse PPI database, where each record corresponds to a unique combination of IAV and mouse strain.

TABLE 3 IAV Domains identified by SCOP Superfamily. Red indicates domains identified as interacting with mouse proteins, while '-' indicates no identified domains.

IAV segment	SCOP superfamily name/Accession number
PB2	PB2 C-terminal domain-like/160453
PB1	DNA/RNA polymerases/56672
PB1-F2	-
HA	Viral protein domain/49818
	Influenza hemagglutinin (stalk)/58064
NP	Flu NP-like/161003
NA	Sialidases/50939
M1	Influenza virus matrix protein M1/48145
	Alpha-catenin/vinculin-like/47220
	Methyl-accepting chemotaxis protein (MCP) signaling domain/58104
M2	-
NS1	NS1 effector domain-like/143021
	S15/NS1 RNA-binding domain/47060
NS2	Nonstructural protein NS2, NEP, M1-binding domain/101156
	Spectrin repeat/46966

1997). It strives to provide high-quality annotations with a minimal level of redundancy and high level of integration with other databases.

As an average protein consists of two or more domains, domain start and end residue numbers, corresponding to regions of protein sequences matching each assigned SCOP domain (by SUPERFAMILY 2.0) were independently noted for every protein sequence retrieved from UniProtKB. Since each domain has its

distinct structure and biological function, only a subset of domains constituting each protein are involved in the interaction between a pair of proteins. Thus, this enhances the complexity of host-pathogen protein interaction analysis (Narykov et al., 2019).

Overall, IAV strains in the dataset were found to comprise proteins domains belonging to 13 SCOP superfamilies (Table 3). Then, these domains were paired up individually with 1102 unique
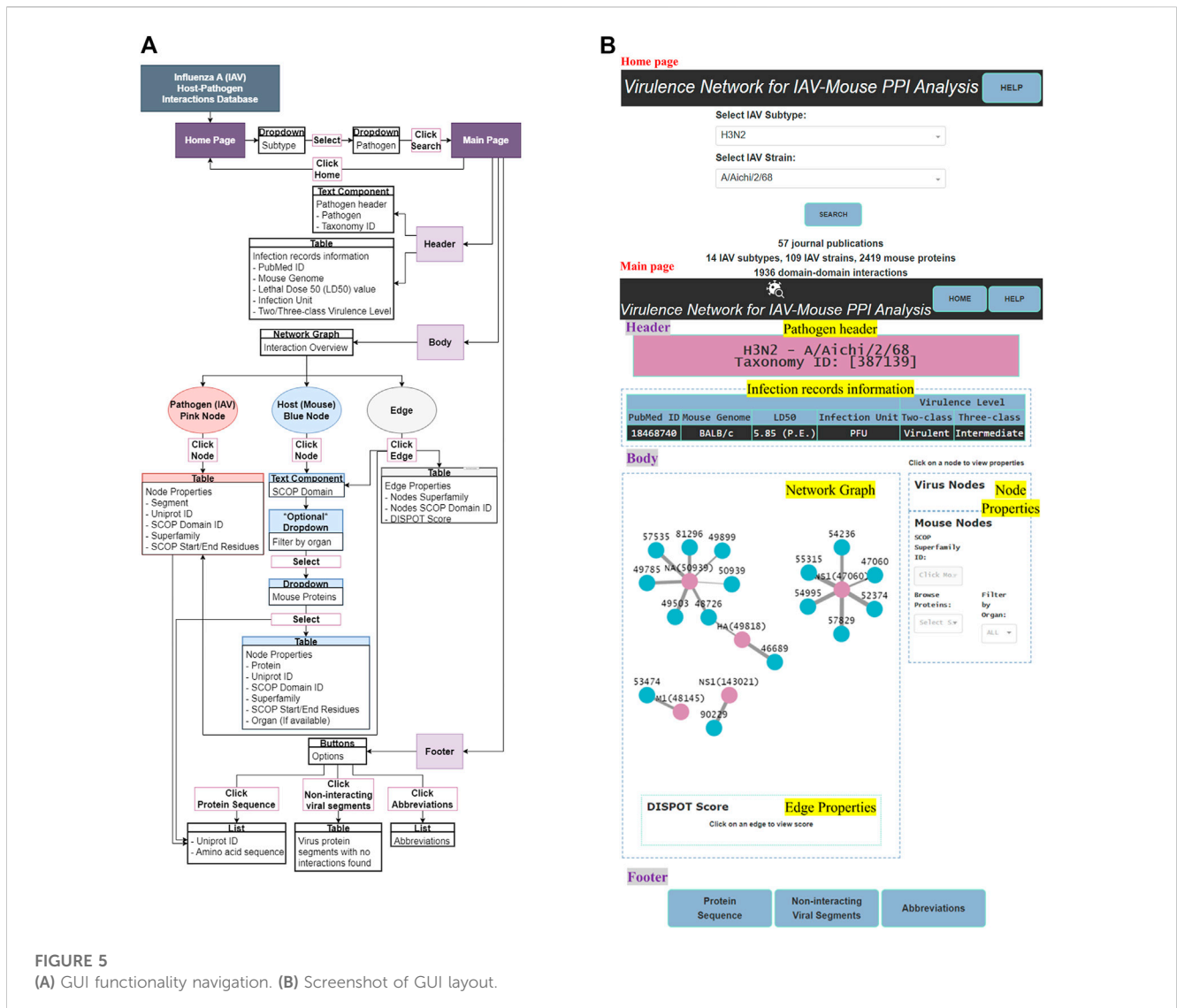
FIGURE 5
(A) GUI functionality navigation. **(B)** Screenshot of GUI layout.

domains found among the mouse protein sequences, forming a total of 14, 326 IAV-mouse protein domain pairs. Subsequently, domain pairs were fed as input into DISPOT, for calculation of statistical potential. Finally, seven domains in IAV proteins (indicated in red in Table 3) and 29 domains in mouse proteins were found to be involved in the host-pathogen PPI. Out of the 17, 120 mouse protein sequences retrieved, 2419 unique proteins were found to contain to at least one of the 29 interacting SCOP domains. In addition, the mouse protein localization in vital organs (lungs, brain, liver, kidney, spleen and heart) or blood was noted, whenever available. (available in Supplementary Figure S2).

3 Web server implementation

IAV-Mouse PPI web server GUI has a comprehensible interface, made up of two pages, with various features, including browsing *via* subtype and strain to view information collected from literature searches, an interactive network graph with accompanying information on node and edge attributes as well as amino acid

sequences extracted from UniProt. Figure 5 illustrates navigation and layout of the web server's GUI.

3.1 Tools

Firstly, Microsoft Excel 2016 was used to store and organize data collected from literature. Secondly, the web interface was developed on the code editor Visual Studio Code V1.71.2, with Python V3.7.4 as the programming language. Python libraries used were Pandas V1.3.5, for transforming comma-separated values (csv) from Excel files to dataframes. Dash V2.3.1 was the framework for designing the application's functionalities, described as follows: Dash bootstrap components V0.3.0 for building the application's layout, graph visualization component Dash Cytoscape V0.3.0 for constructing the interactive network graph. Beautiful Soup V4.11.1 was the HTML parser used for pulling protein sequences from the UniProt database. Lastly, cloud hosting application, Render (https://dashboard. render.com/) web service with Python web

server gateway interface HTTP server, Gunicorn V20.1.0 was utilized to build and run the web interface entirely in the cloud.

3.2 Graphical user interface (GUI)

3.2.1 Home page

The home page features a dependent dropdown component to firstly allow the user to search for a specific IAV subtype and subsequently, browse and select the pathogen(s) belonging to the selected subtype group.

3.2.2 Main page

The main page features three main sections—header, body and footer (shown in Figure 5B).

3.2.2.1 Header

The header section consists of two components—pathogen header text component and infection records information table.

3.2.2.1.1 Pathogen header. This text component displays the IAV subtype, pathogen name and Taxonomy ID, based on selections made by the user in the home page.

3.2.2.1.2 Infection records information. This section presents infection records information collected directly from literature and additional information collected from web tools, NCBI Taxonomy and UniProt databases in the form of a table. Information is filtered to present only those relevant to the user's selections in the home page.

3.2.2.2 Body

The body section consists of two panels, where the left panel displays the network graph and edge properties. The right panel is divided into two subsections and displays the virus and mouse node properties, respectively.

3.2.2.2.1 Network graph. The network graph was designed such that the user can clearly differentiate between IAV and mouse nodes by colors, where pink was assigned to IAV nodes and blue to mouse nodes. User can differentiate interaction statistical potentials by edge weights, where a thicker edge line represents a higher possibility of interaction. Also, the edge color will change to blue upon clicking, to highlight the edge selection.

3.2.2.2.2 Node and edge properties. Node properties include either the IAV protein segment or name of mouse protein, UniProt ID, SCOP superfamily ID, superfamily name and SCOP start/end residue(s). Edge properties comprise IAV and mouse SCOP superfamily ID and name with the matching DISPOT statistical potential score. All IAV node properties will be populated upon clicking of any pink IAV node while only the mouse SCOP superfamily ID field will be populated upon clicking of any blue mouse node in the network graph. Similarly, the former, together with its respective edge property will be presented upon click on any edge. To display all mouse node properties, a mouse protein first needs to be selected from the "Browse Proteins" dropdown under the mouse node properties subsection.

3.2.2.3 Footer

The footer provides the user with the following supplementary information—protein sequence, non-interacting viral segments and abbreviations.

3.2.2.3.1 Protein sequences. Web scraping was applied to extract amino acid sequences of IAV and mouse proteins from the UniProt database.

3.2.2.3.2 Non-interacting viral segments. Non-interacting IAV protein segments consist of the following cases: 1) The UniProt ID could not be found. Therefore, the protein sequence for input to SUPERFAMILY 2.0 is unknown and no domain information could be retrieved. In this case, the UniProt ID field was indicated with "Not Found" and remaining information was labelled as "N/A". 2) Some protein sequences retrieved were not mapped to any SCOP superfamily based on the SUPERFAMILY 2.0 database. As such, there was no protein domain information for input to DISPOT. 3) IAV domain information was available but the DOMMINO database did not have any entry between the IAV domain and all of 17, 120 retrieved mouse proteins, hence no interaction information was returned by DISPOT.

3.2.2.3.2 Abbreviations. This section conveys extra information to the user; specifically, the definitions and expansions of abbreviations used as well as notes targeted to help the user better comprehend the network graph.

4 Discussion

This section lists domain pairs identified with high scoring interaction potentials. In section 4.1 the interacting domain pairs are verified against actual protein-protein interactions identified in biochemistry or proteomics literature. It is organized as a series of discussions on the functional role of interacting domains pairs, per paragraph.

4.1 Interacting protein domains

According to DISPOT scores obtained, the top 10 domain pairs with strongest statistical interaction potentials are as listed in Table 4.

Of 109 unique IAV strains presented in this database, the PB1 gene segment of three strains (Table 5) were assigned the DNA/RNA polymerases domain. However, despite the strong interactions, it was not possible to ascertain if the presence of the DNA/RNA polymerases domain has an impact on the pathogenicity, due to variation in virulence levels across the different IAV strains.

The M1 gene segment of eight IAV strains were assigned a Methyl-accepting chemotaxis protein (MCP) signaling domain instead of alpha-catenin/vinculin-like. No interaction was found between the MCP signaling domain with all domains present in mouse proteins. Comparing results from experiments carried out in H3N2 IAV strains (Tables 6, 7), especially on mice strains BALB/c and DBA/2, it is evident that presence of the alpha-catenin/vinculin-like domain is a virulence factor responsible for IAV infection.

TABLE 4 Top 10 Interactions according to statistical potentials returned by DISPOT. A more negative DISPOT score indicates a higher possibility of interaction.

IAV segment	IAV SCOP domain	Mouse SCOP domain	DISPOT score
NS1	S15/NS1 RNA-binding	Nucleotidylyl transferase	-4.301855801
		L30e-like	
PB1	DNA/RNA polymerases	DNA clamp	
		5′ to 3′ exonuclease, C-terminal subdomain	
		PIN domain-like	
		N-acetylmuramoyl-L-alanine amidase-like	
NS1	S15/NS1 RNA-binding	Zn-binding ribosomal proteins	−3.608708621
		Ribosomal protein S6	
M1	Alpha-catenin/vinculin-like	PH domain-like	
		I/LWEQ domain	

TABLE 5 IAV strains with DNA/RNA polymerases domain assigned to PB1 gene segment.

IAV strain	Subtype	Three-class virulence level
A/PuertoRico/8/1934	H1N1	High (BALB/c, DBA/2)
		Intermediate (C57BL/6J)
rA/X-31	H3N2	Low (BALB/c)
A/HongKong/97/98	H5N1	Low (BALB/c)

Alpha-catenins are members of the vinculin family of proteins. Vinculin is an actin-binding protein. Protease treatment revealed that actin present in the interior of influenza virions presumably participates in moving viral components to the assembly site and cytoskeletal reorganization that occurs during bud formation (Peng et al., 2012). Actin is a family of globular multi-functional proteins that form microfilaments in the cytoskeleton. The host cytoskeletal network takes part in transport of viral components in the cell, predominantly during the stages of virus entry and exit (Shaw et al., 2008).

Pleckstrin homology (PH) domain-like is a short peptide module often found in cytoskeletal proteins (Yao et al., 1999). Although cytoskeletal elements are known to be associated with M1, the underlying mechanisms are not clear (Zhao et al., 2017). Ezrin, [EZRI_MOUSE (UniProt accession number: P26040)] has a PH-like domain. Based on meta-analysis of IAV interactome studies on the M1 gene segment conducted by (Chua et al., 2022), Ezrin was discovered to be a common interactor and is a positive regulator of virus replication.

Information from UniProtKB indicates that majority of proteins that contain a nucleotidylyl transferase domain possess the tRNA ligase enzyme, otherwise known as aminoacyl-tRNA synthetase (ARSs). ARSs play a crucial role in protein synthesis by attaching amino acids to their cognate transfer RNAs (tRNAs) (Nie et al., 2019). Specifically, Cysteine-tRNA ligase [SYCC_MOUSE (UniProt accession number: Q9ER72)], an interactor of NS1, catalyzes the ATP-dependent ligation of cysteine to tRNA (Cys) and plays a role in translation (de Chassey et al., 2013). Furthermore, ARSs plays a vital role in the development of immune cells because of their involvement in maturation, transcription, activation, and recruitment of immune cells. More significantly, ARSs regulate various biological processes and act as signaling molecules in infectious disease (Nie et al., 2019), which supports the high DISPOT score (≈−4.302) predicted for the S15/NS1 RNA-binding domain in NS1 segments.

Ubiquitin-40S ribosomal protein S27a, [RS27A_MOUSE (UniProt accession number: P62983)], is a protein with the Zn-binding ribosomal protein domain. It is an NS1-interacting host protein node, classified as belonging to the apoptosis pathway (Thulasi Raman and Zhou, 2016). Although not required for ribosome function, it plays an important role in the life cycle of IAV through regulating viral nucleic acid replication and gene transcription. When interrupted in host cells, the replication and infectivity of IAV is stopped (Li, 2019).

CCCH zinc finger present in mouse proteins is the sole domain that interacts with NS1 effector domain-like instead of the S15/NS1 RNA-binding domain in the NS1 segment of IAV. This domain pair has a DISPOT statistical potential score of -2.916 (rounded to 3 d.p.). Mouse protein, cleavage and polyadenylation specificity factor subunit 4 (CPSF4) [CPSF4_MOUSE (UniProt accession number: Q8BQZ5)] contains the CCCH zinc finger domain. The interaction between NS1 and CPSF4 controls the alternative splicing of tumor protein p53 (TP53) transcripts, and alters the expression of TP53 isoforms in parallel. As a result, cellular innate response, particularly via type I interferon secretion is regulated, leading to efficient viral replication (Dubois et al., 2019).

The Immunoglobulin (Ig) domain, otherwise known as antibodies is the sole SCOP protein domain that interacts with three IAV domains, namely DNA/RNA polymerases, Viral protein domain and Sialidases in the PB1, HA, and NA segments respectively. Immunoglobulin is the most abundant domain found among the 2419 unique mouse proteins containing interacting SCOP domains (Supplementary Figure S2). During natural infection with IAVs, immune response against both HA and NA will be evoked (Creytens et al., 2021). IgM response is dominant in primary infection, while IgG response is dominant in

TABLE 6 IAV strains with a Methyl-accepting chemotaxis protein (MCP) signaling domain assigned to M1 gene segment instead of alpha-catenin/vinculin-like.

IAV strain	Subtype	Three-class virulence level
A/Aichi/2/68	H3N2	Intermediate (BALB/c)
A/Brisbane/10/2007		Low (C75BL/6, DBA/2)
A/Memphis/8/1988		Low (BALB/c)
A/Panama/2007/1999		Low (C57BL/6, DBA/2)
A/Wisconsin/67/2005		Low (C57BL/6, DBA/2)
rA/X-31		Low (BALB/c)
A/duck/Guangxi/53/2002	H5N1	Low (BALB/c)
A/chicken/Shandong/lx1023/2007	H9N2	Low (BALB/c)

TABLE 7 H3N2 strains with alpha-catenin/vinculin-like domain assigned to M1 gene segment.

IAV strain	Three-class virulence level
A/Hong Kong/1/1968	Low (C57BL/6, DBA/2)
A/Philippines/2/1982	High (BALB/c)
A/swine/Spain/54008/2004	Low (C57BL/6)
	Intermediate (DBA/2)
A/swine/Texas/4199-2/1998	Low (C57BL/6)
	Intermediate (DBA/2)
A/Victoria/3/1975	Intermediate (BALB/c)

secondary infection, for Ig secretion. IgA present in nasal secretions can neutralize HA and NA of IAVs (Chen et al., 2018).

4.2 Non-interacting protein domains

The non-interacting influenza hemagglutinin (stalk) SCOP domain present in the HA is an example that not all domains constituting a protein are involved in interaction between a pair of proteins. The stalk evolves slower than the receptor binding head and it is suggested that it has to remain structurally conserved owing to its role in membrane fusion (Kirkpatrick et al., 2018). Studies have also suggested that the stalk domain is not under immune pressure (Wu and Wilson, 2020; Petrova and Russell, 2018). Additionally, mutations in the stalk domain do not drastically impact virus binding or aid in avoiding neutralizing antibody responses from the host (Kirkpatrick et al., 2018). Therefore a potential "true negative" interaction is also identified in the protein domain network, in line with experimental findings.

4.3 Limitations

Generally, DISPOT works as a tool to streamline the PPI prediction problem through providing insight on the possibility of specific DDIs in a given physical PPI. However, it is not a classification method and statistical potentials returned are useful for ranking DDIs but do not directly translate to the probability score. DISPOT which solely uses information about interactions between protein domain should not be used as a standalone PPI prediction tool to identify virulence factors responsible for IAV infections (Narykov et al., 2019). Based on results of this work, IAV genomes across different strains comprise highly similar domains due to their similar structure (i.e., eight segments, encoding at least 11 proteins) and biological function. Furthermore, interactions that involve protein structures are facilitated not only by the protein domains, but also by various non-structured regions, such as interdomain linkers, N and C terminal structures or sequences, protein peptides (Kuang et al., 2012). Therefore, utilizing DISPOT exclusively may produce high number of false negative or false positive PPI predictions.

Mitochondria play an imperative role in antiviral innate immune response through the mitochondiral antiviral-signaling protein (MAVS) [MAVS_MOUSE (UniProt accession number: Q8VCF0)] protein, a component of the retinoic acid-inducible gene I (RIG-I) antiviral pathway. This pathway along with multiple others, is essential for combating and resolving viral infection, repair of damaged tissues, and generating adaptive immune response. It has been revealed that PB1-F2 inhibits antiviral cytokines and enhances expression of inflammatory cytokines through direct interaction with MAVS and other components of the RIG-I/MAVS system (Kamal et al., 2017). However, as protein sequences of both PB1-F2 and MAVS were not assigned any SCOP domain by SUPERFAMILY 2.0, it was not possible to verify this interaction via DISPOT.

A homeodomain-like domain was identified by DISPOT to be interacting with the viral protein domain present in the HA gene segment of IAV, with a statistical potential score of −3.203 (rounded to 3 d.p.). However a study conducted by (Farooq et al., 2020), which integrates both IAV-Mouse PPIs detected using either small-scale or large-scale researches carried out experimentally or computationally found no evidence for an interaction with HA. In (Farooq et al., 2020), homeobox protein MOX-2 (MEOX2) [MEOX2_MOUSE, (UniProt accession number: P32443)], containing the homeodomain-like

domain, was identified to be interacting with IAV gene segments PB1, PA, NA, and M2 but not HA. By comparison, for DISPOT no interaction was detected between the homeodomain-like domain and segments PB1 and NA. Likewise, as protein sequences of segments PA and M2 were not assigned any SCOP domain by SUPERFAMILY 2.0, DISPOT could not be used to ascertain these interactions. Further, the cellular localisation for proteins with homeodomain-like domains was found to be in the nucleus according to UniProt, which indicates its interaction with HA would be unlikely, given that HA, mediates cell-surface recognition and viral entry.

Additionally, the reason for virulence levels to differ across mouse strains infected with the same IAV strain has not been uncovered as protein sequences of mouse strains retrieved from UniProt were derived from referencing the C57BL/6J mouse strain only. This limitation is because currently whole proteome sequences of other mouse strains (i.e., BALB/c, DBA/2 and FVB/J) are not available in any public database. Translation of strain specific genomic sequences to whole proteomes is a challenging task needing extensive experimental effort. In this work, obtaining these proteomes by means of experimental protein sequencing was not possible as the necessary materials and labor were not available.

5 Future work

To bridge the gaps in this work, sequence-based PPI prediction methods can be employed to substantiate DDIs identified by DISPOT. An example is the Human-Virus Protein-Protein Interactions (HVPPI) web server, developed by X.Yang and colleagues (Yang et al., 2020a). HVPPI applied an unsupervised sequence embedding technique (*doc2vec*) to represent protein sequences as low-dimensional rich feature vectors. Then, a random forest classifier was trained using a training dataset that covers known PPIs between human and all viruses to predict human-virus PPIs. Lastly, the HVPPI web server automatically calculates the interaction probability of a query protein pair. The data to be used as input to HVPPI can be constructed as follows: Firstly, human protein sequences can be obtained using mouse and human homologs. Next, all protein sequences with SCOP domain(s) assigned to them can be trimmed, following the collected start and end residue numbers. Subsequently, trimmed protein sequences can be paired corresponding to DDIs recognized by DISPOT. Interaction probabilities provided as predicted outputs by HVPPI for each IAV-human protein pair can then be used to detect false positives. For protein sequences not assigned to any SCOP domain, complete protein sequences can be used, which will in turn aid with the detection of false negatives.

As an extension of this work from the raw dataset, instead of ignoring non-standard strains, the protein sequences of recombinant or mutant IAV strains can be reproduced *via* manually changing the protein sequences of wild-type or laboratory IAV strains that are available in UniProt. This enriches the dataset further.

The DISPOT statistical potentials, HVPPI interaction probabilities and LD_{50} values can be incorporated to represent the PPI network as a weighted undirected graph. Later, graph embedding methods can be applied to this weighted graph to learn low-dimensional node representations (Yue et al., 2020). Structural information of PPI, such as the degree, position and neighbouring nodes in a graph has been recognized to be helpful in PPI prediction (Yang et al., 2020b).

6 Conclusion

As IAV is a significant danger to global human health and life, it is critical to have deeper, accurate as well as reliable insights and knowledge on the virulence factors responsible for IAV infections to counteract potential outbreaks (Ivan and Kwoh, 2019). This work built upon a previously curated dataset of lethal dose studies of IAV infection in mice. Thereafter, superfamily domains involved in DDIs between IAV and mice were discovered, and ranked according to statistical interaction potentials calculated by DISPOT. A one-stop web server integrating information collated from literature and various databases, namely, NCBI Taxonomy, UniProt and SUPERFAMILY 2.0 with the DDI network was constructed. Furthermore, the web server is scalable and can seamlessly accommodate addition of new functions and data when future research is carried out.

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://iav-ppi.onrender.com/home https://github.com/tengann/IAV-Host-PPI-Database.

Author contributions

TN and SR conceived and designed the method. TN implemented the method. TN and SR wrote the manuscript. CK supervised throughout the conception, design, implementation and writing stages. All authors reviewed and approved the final manuscript.

Funding

Acknowledgments

Conflict of interest

The 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.2023.1123993/full#supplementary-material

References

Alborzi, S. Z., Nacer, A. A., Najjar, H., Ritchie, D. W., and Ppidomainminer, M. D. D. (2021). Inferring domain-domain interactions from multiple sources of protein-protein interactions. *PLOS Comput. Biol.* 17 (8), e1008844. doi:10.1371/journal.pcbi.1008844

Ammari, M. G., Gresham, C. R., McCarthy, F. M., and Nanduri, B. (2016). Hpidb 2.0: A curated database for host–pathogen interactions. *Database* 2016, baw103. doi:10.1093/database/baw103

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2004). Scop database in 2004: Refinements integrate structure and sequence family data. *Nucleic Acids Res.* 32, 226D–229D. doi:10.1093/nar/gkh039

Bairoch, A., and Apweiler, R. (1997). The swiss-prot protein sequence data bank and its supplement trembl. *Nucleic Acids Res.* 25 (1), 31–36. doi:10.1093/nar/25.1.31

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2005). The universal protein resource (uniprot). *Nucleic Acids Res.* 33, D154–D159. doi:10.1093/nar/gki070

Chen, X., Liu, S., Goraya, M. U., Maarouf, M., Huang, S., and Chen, J. L. (2018). Host immune response to influenza a virus infection. *Front. Immunol.* 9, 320. doi:10.3389/fimmu.2018.00320

Chua, S. C. J. H., Cui, J., Engelberg, D., and Lim, L. H. K. (2022). A review and meta-analysis of influenza interactome studies. *Front. Microbiol.* 13, 869406. doi:10.3389/fmicb.2022.869406

Cook, H. V., Doncheva, N. T., Szklarczyk, D., von Mering, C., and Jensen, L. J. (2018). Viruses.STRING: A virus-host protein-protein interaction database. *Viruses* 10 (10), 519. doi:10.3390/v10100519

Creytens, S., Pascha, M. N., Ballegeer, M., Saelens, X., and de Haan, C. A. M. (2021). Influenza neuraminidase characteristics and potential as a vaccine target. *Front. Immunol.* 12, 786617. doi:10.3389/fimmu.2021.786617

de Chassey, B., Aublin-Gex, A., Ruggieri, A., Meyniel-Schicklin, L., Pradezynski, F., Davoust, N., et al. (2013). The interactomes of influenza virus ns1 and ns2 proteins identify new host factors and provide insights for adar1 playing a supportive role in virus replication. *PLoS Pathog.* 9 (7), e1003440. doi:10.1371/journal.ppat.1003440

Dubois, J., Traversier, A., Julien, T., Padey, B., Lina, B., Bourdon, J. C., et al. (2019). The nonstructural ns1 protein of influenza viruses modulates tp53 splicing through host factor cpsf4. *J. Virol.* 93 (7), 021688–e2218. doi:10.1128/jvi.02168-18

Eugene, W. N. (2001). *Microbiology: A human perspective*. 3rd edition. Dubuque, Iowa: McGraw-Hill.

Farooq, Q. U. A., Shaukat, Z., Aiman, S., Zhou, T., and Li, C. (2020). A systems biology-driven approach to construct a comprehensive protein interaction network of influenza a virus with its host. *BMC Infect. Dis.* 20 (1), 480. doi:10.1186/s12879-020-05214-0

Itzhaki, Z., Akiva, E., Altuvia, Y., and Margalit, H. (2006). Evolutionary conservation of domain-domain interactions. *Genome Biol.* 7 (12), R125. doi:10.1186/gb-2006-7-12-r125

Iuliano, A. D., Roguski, K. M., Chang, H. H., Muscatello, D. J., Palekar, R., Tempia, S., et al. (2018). Estimates of global seasonal influenza-associated respiratory mortality: A modelling study. *Lancet* 391, 101271285–101271300. doi:10.1016/s0140-6736(17)33293-2

Ivan, F. X., and Kwoh, C. K. (2019). Rule-based meta-analysis reveals the major role of pb2 in influencing influenza a virus virulence in mice. *BMC Genomics* 20 (9), 973. doi:10.1186/s12864-019-6295-8

Johnson, N. P. A. S., and Mueller, J. (2002). Updating the accounts: Global mortality of the 1918-1920 "Spanish&quot; influenza pandemic. *Bull. Hist. Med.* 76 (1), 105–115. doi:10.1353/bhm.2002.0022

Jung, H. E., and Lee, H. K. (2020). Host protective immune responses against influenza a virus infection. *Viruses* 12 (5), 504. doi:10.3390/v12050504

Kamal, R. P., Alymova, I. V., and York, I. A. (2017). Evolution and virulence of influenza a virus protein pb1-f2. *Int. J. Mol. Sci.* 19 (1), 96. doi:10.3390/ijms19010096

Kirkpatrick, E., Qiu, X., Wilson, P. C., Bahl, J., and Krammer, F. (2018). The influenza virus hemagglutinin head evolves faster than the stalk domain. *Sci. Rep.* 8 (1), 10432. doi:10.1038/s41598-018-28706-1

Kuang, X., Han, J. G., Zhao, N., Pang, B., Shyu, C. R., and Korkin, D. (2012). Dommino: A database of macromolecular interactions. *Nucleic Acids Res.* 40, D501–D506. doi:10.1093/nar/gkr1128

Li, S. (2019). Regulation of ribosomal proteins on viral infection. *Cells* 8 (5), 508. doi:10.3390/cells8050508

Lindberg, D. A. (2000). Internet access to the national library of medicine. *Eff. Clin. Pract.* 3 (5), 256–260.

Lu, X., Tumpey, T. M., Morken, T., Zaki, S. R., Cox, N. J., and Katz, J. M. (1999). A mouse model for the evaluation of pathogenesis and immunity to influenza a (h5n1) viruses isolated from humans. *J. Virol.* 73 (7), 5903–5911. doi:10.1128/jvi.73.7.5903-5911.1999

Masemann, D., Ludwig, S., and Boergeling, Y. (2020). Advances in transgenic mouse models to study infections by human pathogenic viruses. *Int. J. Mol. Sci.* 21 (23), 9289. doi:10.3390/ijms21239289

Narykov, O., Bogatov, D., and Korkin, D. (2019). Dispot: A simple knowledge-based protein domain interaction statistical potential. *Bioinformatics* 35 (24), 5374–5378. doi:10.1093/bioinformatics/btz587

Nie, A., Sun, B., Fu, Z., and Yu, D. (2019). Roles of aminoacyl-trna synthetases in immune regulation and immune diseases. *Cell. Death Dis.* 10 (12), 901. doi:10.1038/s41419-019-2145-5

Pandurangan, A. P., Stahlhacke, J., Oates, M. E., Smithers, B., and Gough, J. (2019). The superfamily 2.0 database: A significant proteome update and a new webserver. *Nucleic Acids Res.* 47 (D1), D490–D494. doi:10.1093/nar/gky1130

Peng, X., Maiers, J. L., Choudhury, D., Craig, S. W., and DeMali, K. A. (2012). α-catenin uses a novel mechanism to activate vinculin. *J. Biol. Chem.* 287 (10), 7728–7737. doi:10.1074/jbc.m111.297481

Perlman, R. L. (2016). Mouse models of human disease: An evolutionary perspective. *Evol. Med. Public Health* 2016 (1), 170–176. doi:10.1093/emph/eow014

Petrova, V. N., and Russell, C. A. (2018). The evolution of seasonal influenza viruses. *Nat. Rev. Microbiol.* 16 (1), 47–60. doi:10.1038/nrmicro.2017.118

Rogers, K. (1968). "Flu pandemic," in *Encyclopedia britannica*. 15th edition (Chicago, Illinois, United States: Encyclopaedia Britannica, Inc.). Avaliable At: https://www.britannica.com/event/1968-flu-pandemic.

Sarkar, S., and Heise, M. T. (2019). Mouse models as resources for studying infectious diseases. *Clin. Ther.* 41 (10), 1912–1922. doi:10.1016/j.clinthera.2019.08.010

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., et al. (2020). Ncbi taxonomy: A comprehensive update on curation, resources and tools. *Database (Oxford)* 2020, baaa062. doi:10.1093/database/baaa062

Shaw, M. L., Stone, K. L., Colangelo, C. M., Gulcicek, E. E., and Palese, P. (2008). Cellular proteins in influenza virus particles. *PLOS Pathog.* 4 (6), e1000085. doi:10.1371/journal.ppat.1000085

Shi, J., Deng, G., Kong, H., Gu, C., Ma, S., Yin, X., et al. (2017). H7n9 virulent mutants detected in chickens in China pose an increased threat to humans. *Cell. Res.* 27 (12), 1409–1421. doi:10.1038/cr.2017.129

Shi, J., Deng, G., Ma, S., Zeng, X., Yin, X., Li, M., et al. (2018). Rapid evolution of h7n9 highly pathogenic viruses that emerged in China in 2017. *Cell. Host Microbe* 24 (4), 558–568. doi:10.1016/j.chom.2018.08.006

Swets, M. C., Russell, C. D., Harrison, E. M., Docherty, A. B., Lone, N., Girvan, M., et al. (2022). Sars-cov-2 co-infection with influenza viruses, respiratory syncytial virus, or adenoviruses. *Lancet* 399, 103341463–103341464. doi:10.1016/s0140-6736(22)00383-x

Thulasi Raman, S. N., and Zhou, Y. (2016). Networks of host factors that interact with ns1 protein of influenza a virus. *Front. Microbiol.* 7, 654. doi:10.3389/fmicb.2016.00654

Viboud, C., Simonsen, L., Fuentes, R., Flores, J., Miller, M. A., and Chowell, G. (2016). Global mortality impact of the 1957-1959 influenza pandemic. *J. Infect. Dis.* 213 (5), 738–745. doi:10.1093/infdis/jiv534

WHO (2003). Production of pilot lots of inactivated influenza vaccine in response to a pandemic threat: An interim biosafety risk assessment. *Wkly. Epidemiol. Rec.* 78 (47), 405–408.

Wu, N. C., and Wilson, I. A. (2020). Structural biology of influenza hemagglutinin: An amaranthine adventure. *Viruses* 12 (9), 1053. doi:10.3390/v12091053

Yang, F., Fan, K., Song, D., and Lin, H. (2020a). Graph-based prediction of protein-protein interactions with attributed signed graph embedding. *BMC Bioinforma.* 21 (1), 323. doi:10.1186/s12859-020-03646-8

Yang, X., Yang, S., Li, Q., Wuchty, S., and Zhang, Z. (2020b). Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. *Comput. Struct. Biotechnol. J.* 18, 153–161. doi:10.1016/j.csbj.2019.12.005

Yao, L., Janmey, P., Frigeri, L. G., Han, W., Fujita, J., Kawakami, Y., et al. (1999). Pleckstrin homology domains interact with filamentous actin. *J. Biol. Chem.* 274 (28), 19752–19761. doi:10.1074/jbc.274.28.19752

Yue, X., Wang, Z., Huang, J., Parthasarathy, S., Moosavinasab, S., Huang, Y., et al. (2020). Graph embedding on biomedical networks: Methods, applications and evaluations. *Bioinformatics* 36 (4), 1241–1251. doi:10.1093/bioinformatics/btz718

Zhao, M., Wang, L., and Li, S. (2017). Influenza a virus-host protein interactions control viral pathogenesis. *Int. J. Mol. Sci.* 18 (8), 1673. doi:10.3390/ijms18081673