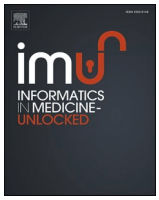
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Informatics in Medicine Unlocked 31 (2022) 100966

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu



Towards faster response against emerging epidemics and prediction of variants of concern

B. Robson [a,b,*]

[a] Ingine Inc., Cleveland, Ohio, USA
[b] The Dirac Foundation, Oxfordshire, UK

A B S T R A C T

The author, the journal, Computers in Biology and Medicine (CBM), and Elsevier Press more generally, played a helpful very early role in responding to COVID-19. Within a few days of the appearance of the "Wuhan Seafood isolate" genome on GenBank, a bioinformatics study was posted by the present author in ResearchGate in January 2020, "Preliminary Bioinformatics Studies on the Design of Synthetic Vaccines and Preventative Peptidomimetic Antagonists against the Wuhan Seafood Market Coronavirus. Possible Importance of the KRSFIEDLLFNKV Motif" DOI: 10.13140/RG.2.2.18275.09761. On February 2nd, 2020, a more thorough analysis was submitted to CBM, e-published on February 26, and formally published in April 2020, at about the same time as the virus named as 2019n-CoV was identified as essentially SARS and renames SARS-COV-2. This was followed by four further papers describing in more detail some previously unreported aspects of the early investigation. The speed of research and writing of the papers was made possible by knowledge-gathering tools. Based on this and earlier experiences with fast responses to emerging epidemics such as HIV and Mad Cow Disease, it is possible to envisage the nature of a speedier response to emerging epidemics and new variants of concern in established epidemics.

1. Introduction

1.1. Background

The rapid acquisition of knowledge about a newly emerging disease is crucial to the health of human, animal, and plant populations. The term *epidemic* from the Greek epi (on) and demos (people), possibly first used by Homer, is believed to have promoted in the medical setting as the title of the treatise attributed to Hippocrates. The interest of Hippocrates in medicine seems likely to have been greatly motivated by knowledge of a plague that killed a quarter of the population of Athens. *Epidemiology*, the study of epidemics, is typically described as the investigation of factors that determine the frequency and distribution of disease or other health-related conditions within a defined human population during a specified period. That description reflects the founding approach of John Snow (1813–1858) in large part because of his use of maps and statistics in tracing the source of a cholera outbreak in Soho, London, in 1854.

More simply expressed, epidemiology is the basic science of public health, but it remains that it is infectious disease and especially "new kinds" that are of concern to national and international authorities [1], that are of interest here. "New kinds" can mean various things. Not all epidemics in humans arise from well-studied pathogens (long known species of viruses, bacteria, etc.). Some arise as new species in the sense of a new identification and classification, but new strains of those that are very familiar, such as influenza A, tuberculosis, and potentially measles, can be serious enough. Coronaviruses illustrate both: all coronaviruses are relatively new to science, being discovered in the 1960s [2,3], but still fairly well studied prior to the rise of SARS in 2003 [4] and COVID-19 in 2019–2020 [5–9]. In contrast, HIV was reasonably described in terms such as "a totally new kind" of infectious disease in the sense of being essentially unknown to modern science when it was identified in the 1980s. It was a major factor that encouraged health journalist Laurie Garrett to worry that is would be a harbinger of other pandemics in her detailed 750-page book, "The coming Plague" [1]. Prior to that, in the 1950s and 1960s, there was great optimism in industrial nations as medical researchers declared "miracle breakthroughs" against infectious disease on what seemed like an almost weekly basis [1]. She reviews several important principles and practices of epidemiology and gives a good account of all actual and potential epidemics known to history up to that time, and potential future threats. However, the word "coronavirus" does not appear in Garrett's extensive index [1]. Because coronaviruses were discovered only around 1966 [2,3] and so a "relatively new kid on the block", especially as regards diseases serious to humans [4,5], the appearance of COVID-19 caught the world somewhat by surprise [6–11].

* Ingine Inc., Cleveland, Ohio, USA.
E-mail address: barryrobson@ingine.com.

https://doi.org/10.1016/j.imu.2022.100966
Received 23 March 2022; Received in revised form 5 May 2022; Accepted 11 May 2022
Available online 20 May 2022

1.2. Aims of this paper

This paper is of the "narrative review" and to some extent a "scope review" addressing the aspects of response to emerging infectious diseases that the author considers as potentially important, and which perhaps have previously been somewhat under-discussed, or that the author thinks should be approached in a somewhat different way. In comparison to some recent papers by the author, it is neither intended to be a research paper illustrating a rapid response to an emerging epidemic, nor even a review of the "rapid review" type often used to alert relevant authorities to some new pressing need. It is, however, about such rapid responses, and it does use as examples the three early papers by the author [7–9] on what became known as COVID-19, and subsequent papers, as well as referring to earlier efforts in outbreaks of AIDS and other infectious diseases, all of which illustrate, to vary extents, fast responses within the present author's experience.

The above is an important distinction to make, for reasons that highlight the problem. During the relatively leisurely, look-back writing of this present paper, which includes accounts of what is to be learned from the past and specifically what computational tools were found useful for the author's fast responses to COVID-19, there have been several more variants of the omicron variant (which *is* discussed), variants that are increasingly given informal names such as "deltacron" and "stealth omicron". Indeed, during a very few hours of delay in submission of this review, due to a purely technical issue, it became clearer that a new spike was due to a stealth variant BA.2 replacing BA.1 omicron which at the beginning of 2022 had grown to represent some 99% of cases, motivating some rewriting. Shortly prior to gally proof stage, BA.2.12.1, an offshoot of the BA.2 Omicron stealth variant, rose to represent 20%–30% of cases in the UK. But more recently still, there has been the news that companies like Moderna are developing muti-strain vaccines that can hopefully handle all of these. More recent surprises regarding quite different pathogens are mentioned in Section 5.

As well as describing some tools that worked well early in the appearance of COV-19, some further proposed algorithms are introduced to illustrate the kinds of developments that could facilitate a fast response to epidemics. Because the new variants can arise quickly and significantly change the nature of the disease, the new algorithms presented as examples should be considered as "templates" for future use.

Regarding these algorithms, some clarity may be provided as regards the overall purpose and exposition, because there would appear to be some mixing of qualitative and quantitative or semi-quantitative aspects, to an extent which is perhaps not common. This paper which emphasizes the importance of obtaining both kinds of knowledge to respond rapidly to emerging epidemics. But more importantly and unusually, the approach *integrates into a common canonical* form the knowledge from (a) mining structured data including such as data in CSV format (comma separated value format, essentially like a large spreadsheet), and DNA, RNA, and protein sequences on GenBank and (b) from mining unstructured data as authoritative natural language text, mainly obtained by "autosurfing" (automated surfing) of the Internet. One may also include as frequently used input (c) the knowledge extracted as above and retained for future use in a Knowledge Representation Store (KRS). The importance is that the integrated elements of knowledge in common format can be used in automated reasoning and prediction, as well as simply read directly by human eye, as informative to the user (but see below). Most often, the canonical form derived from either structured or textual data will be in a *semantic triple* format < *subject expression* | ***relationship expression*** | *object expression* >, analogous to subject-verb-object languages like English Here the expressions can contain biosequence data or other information, and the overall structure <…> is the commonest kind of basic element or "tag" of the Q-UEL language, discussed in Theory Section 2.1. It is analogous to an XML "tag" except that it may be associated with probabilistic values and used in the automated reasoning and prediction in a way that takes account of degrees of uncertainty or limited reliability in the source information. Q-UEL can be considered as an extension of XML for probabilistic semantics and Artificial Intelligence, and can be converted to XML, though the result is typically uglier and usually much harder for humans to read directly.

Some related comment may also be made on the nature of the informative content of this review. There are also more elaborate Q-UEL tag forms called *semantic multiples* corresponding to parsed structures of sentences that contain several relationships, verbal, prepositional, comparative, or logical. These tags are usually the initial form of the knowledge extracted from natural language text and are called XTRACT tags. They can either be read directly as sentences by the human eye, decomposed in semantic triples, or used directly by the computer in certain reasoning algorithms. However, while readable by the human eye, they can often appear stilted and somewhat robotic because the sentence (or sentences of subsentence) from which the XTRACT tag is derived is reorganized, i.e., reparsed, to facilitate computer use. Primarily, this is because the graph structure representing the parsed form of a sentence is always converted by natural language processing to a linear graph as much as possible, to facilitate decomposition into the component semantic triples. Consequently, for ease of readability, some such tags have been used as relevant information and re-expressed in the text of the present paper in a more readable English form. Although currently this is still largely a manual process, it is being progressively developed to allow metanalyses, systematic reviews, reports, and technical papers to be written automatically, at least as good initial drafts. At the same time, however, much of the present paper is what one would expect: simply a recollective review written by the author.

1.3. Epidemics and the roles of computers

The above arguably represents continuation of a natural trend in modern epidemiology. *Computational epidemiology* is a recognized field that uses techniques from mathematics, computer science, geographic information, and public health data to analyze the spread of diseases and the effectiveness of a public health intervention. Effective intervention for new pathogens and variants requires, for the most part, new diagnostics, vaccines, and therapeutic drugs. Developing effective diagnostics and vaccines, along with attempts at containment and other preventative measures, represent *primary prevention*. A response to emergent infectious disease also means developing effective therapeutic drugs to treat infected people (*secondary prevention*), and effective means of aiding recovery and diminishing the severity of after-effects (*tertiary prevention*) which is still somewhat imperfect in the case of so-called "long COVID". To achieve these "preventions". Garrett had noted in "The Coming Plague" that extensive data banks including genomics of pathogens would become important. This view was clearly correct, and today a rate-limiting step is the appearance of well-checked genomic details about new pathogens and strains in publicly accessible data banks, plus alerts to such submissions, as discussed and illustrated below.

For most researchers today, that essentially means when the genome (DNA or RNA sequence) of the identified causative agent is deposited in a data base and today that means primarily in *GenBank*, which is accessible at https://www.ncbi.nlm.nih.gov/nucleotide/. This is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations, produced and maintained by

the National Center for Biotechnology Information which is a member of the International Nucleotide Sequence Database Collaboration. Initially, one may then undertake bioinformatics analyses followed by computational chemistry studies of the proteins of the pathogen with a view to design or selection of diagnostics, vaccines, and potential drugs.

Except for some important preliminary knowledge-gathering by computer of the kind described later below, relatively little could be done by the present author regarding the disease subsequently called COVID-19 until near-finalized versions of the SARS-COV-2 virus genome became widely available via GenBank [10,11]. Computer and Internet are particularly essential to handle the molecular details and here computational protein science plays an important role. Knowledge of the RNA sequence representing the genome of SARS-COV-2 [10,11] was fundamental to the present author's growing COVID-19 project [7–9, 12–14] and detailed knowledge of that sequence is obviously required for what at the time seemed the surprising use of DNA and RNA vaccines for COVID-19 by the biopharmaceutical industry and major universities (e.g., Refs. [15,16]). The AI-style tools discussed below, developed by the present author and collaborators, were extremely helpful in facilitating a rapid response to COVID-19 without extensive resources, though aided and influenced by the author's earlier experience [17–19] in responding to the early stages of outbreaks of AIDS [18], Mad Cow Disease (Bovine Spongiform Encephalitis) [19], as well as flaviviruses, Ebola, and a variety of veterinary diseases [17]. In the case of COVID-19, it was arguably the first known extensive bioinformatics and diagnostic, vaccine, and drug design response, starting in January 2020 in the same few days in which the disease was characterized, i.e., the first steps toward an epidemiological *case definition* were taken (analyzed in some detail below).

The acquisition and application of knowledge remains important at all stages, however. One way to improve early detection is to monitor health-seeking behavior in the form of queries to online search engines such as Google, and more generally the Internet can play an important role. Its history and its layers are often considered as being in five or more stages; see Ref. [17] for discussion in an epidemiological context. The basic Internet connects computers and began in the 1960s when the US Department of Defense awarded contracts as early as the 1960s for packet network systems to connect computers, including the development of the ARPANET, which would become the first network to use the Internet Protocol. *World Wide Web 1*.0 connects web pages. Berners-Lee wrote a proposal in March 1989 for "a large hypertext database with typed links". *World Wide Web 2.0* connects people, using sites that use technology beyond the static pages of earlier Web sites. Essentially, it connects people by facilitating social networking. The term was coined in 1999 by Darcy DiNucci and was popularized by Tim O'Reilly in 2004. *World Wide Web 3.0* connects data and knowledge. iIs is normally considered as represented by *the Semantic Web*, a collaborative movement led by international standards body the World Wide Web Consortium (W3C). It aims at converting the current web, dominated by unstructured and semi-structured documents into a "web of data", particularly by using Resource Description Framework (RDF). *World Wide Web 4.0, The Thinking Web,* sometimes called 5.0 or higher according to classifications, will organize probabilistic knowledge and reason with it across multiple servers and help make decisions. Particularly when rendered capable of handling uncertainty and probability, it considered by the present author and collaborators as pressingly important for a variety of applications in medicine and the Q-UEL language [20–34] and, as an extension to that to epidemiological use cases, it was used in the early response to COVID-19 as described in this paper.

Although for any emerging epidemic the bioinformatics tools used locally and on the Internet are important for analysis of pathogens at the level of DNA and RNA sequences, the proteins for which they code quickly become a major part of the investigation. Up till relatively recently, there was some extent of a gap between the knowledge of the molecular details of a pathogen and the understanding of molecular sciences that would be desirable to make more finely tuned, sophisticated responses. Until the 1980s and even late 1990s, the molecular details used for subtypes, strains and variants were primarily about what antibodies could be raised by, and interact with, the surface proteins of the pathogen. This is especially famous publicly for the influenza A virus where the practice and the notation continues today: recall that the numbers in "Spanish flu" H1N1 (1918–1919), "Asian flu" H2N2 (1957–1958), "Hong Kong flu" H3N2, and "Bird flu" H5N1 (1997), H7N9 (2013), H5N6 (2014), and H5N8 (2016) all relate to the immunological typing of influenza A into subtypes using the immunological properties of external (spike-like) proteins hemagglutinin H and neuraminidase N, the number increasing with a new immunological sub-type (most of these still circulate extensively today). Although today the major step is determining the genome of a pathogen, attack by pathogens and defense against them, both naturally by the body and aided by science, is primarily a war between proteins, i.e., the proteins of the pathogen and the proteins of the host often aided by the proteins or peptides (in effect, small parts of proteins) contained in, or implied by, or used in making, a vaccine or diagnostic. The well-known success of DNA and RNA vaccines, used for the first time "outside the lab" in the COVID-19 pandemic, does not get round the fact that they use the cells of the vaccinated person to produce the required proteins, and that knowledge of the details of protein structure and function remains important. Not least, ongoing acquisition of knowledge about new pathogen variants and how they affect the function of pathogen proteins and the response of the host, is crucial, else vaccines (and perhaps also diagnostics) found effective in a first wave of an epidemic might be rendered useless in subsequent waves. Vaccines composed of pathogen proteins remain important or are emerging as potentially important tools in the modern armory, and new classes of vaccine based on chemosynthetic or cloned peptide copies of key parts of them continue to show promise as a new class of weapon in the laboratory, already proven in animal husbandry (as discussed below).

The pathogen genome is, of course, not the only genome of interest. Bioinformatics tools both locally and on the Internet are also important for a fast response because of the complexity of the human genome and human molecular biology required for understanding the host response. To add to that complexity, there is an aspect of personalized medicine (including simulations) in the interactions between the proteins of pathogen and host. This is because there are not only differences in the immunological state of individuals due to past exposure and the actions of some therapeutic agents on the immune system, but also all the important proteins of the human host response and host receptors to which the virus binds and enters the cell are at least to some degree *polymorphic*. That is, because of genomic differences, they vary from human individual to individual. For example, host cell receptors by which a virus can enter a cell can vary: notably in gene encoding CCR5, which acts as a co-receptor for HIV. Importantly, HLA (Human Leucocyte Antigen) proteins of the major histocompatibility complex (MHC), can vary, and variations in both types of proteins often lead to marked differences between individuals in susceptibility and severity to particular infectious diseases.

1.4. First responses: use of early pre-genomic knowledge in an emerging epidemic

As also described in this review, knowledge gathering tools are very valuable for gathering the appropriate data as input to bioinformatics and computational protein science, and to facilitate automatic use of those tools, but some comments should also be made on their importance in the very earliest stages of an emerging epidemic, even before the pathogen genome is available. This may be described as a more predominantly qualitative, preparative stage and certainly as a pre-bioinformatics stage, but it is important because the severity of the problem, as the prevalence and geographical distribution of a disease and comorbidities or fatalities from it, can evidently escalate very rapidly. Initially, there may be little knowledge of the disease, or even of the nature of the pathogen (and almost certainly not a detailed map of the pathogen genome), putting the science involved in a somewhat similar position to that available up to the 19th century. The first warning will be a sudden increase in the incidence and prevalence of a disease with characteristic symptoms and a significantly increased morbidity and/or fatality rate, whence it is likely to a previously unknown species or a new variant of a known one. There may, nonetheless, be some prior probabilities as degrees of belief. Medical anthropologist Edward Hudson stated that sexually transmitted syphilis is a disease of "advanced urbanization" whereas yaws (caused by a bacterium that enters skin abrasions and gives rise to small, crusted lesions which may develop into serious deep ulcers) was "a disease of village and the unsophisticated" [1].

It could still be asked as to why early knowledge is useful, beyond being, of course, a valuable step toward characterizing the pathogen and determine its genome. While overpopulation and modern travel means that an infectious disease today is no longer confined in space and time [1], the rise of communication technology, and especially the Internet, also means that knowledge is similarly no longer limited. One can imagine that many pandemics (global epidemics) in history could have been averted (primary prevention) or attenuated had the Internet been available in their time, even if vaccines, relevant effective pharmaceutical drugs, and even a significant knowledge of microbiology, were still absent. Although the first definitive appearance of the Black Death was in Crimea in 1347, and cholera was characterized in outbreaks in England and Italy, and Jessore India in the mid to late 19th century, they were known to ancient physicians at specific locations in Asia long before that. Once relevant knowledge is available, defensive action of a simple "low tech" nature is sometimes sufficient. Many lives may have been saved from cholera, that kills by dehydration, by the availability of large amounts of water and salt. When John Snow identified cholera as water-born, the only technology required was that for the removal of the handle of the pump to the infected supply. It is possible that in ancient India an early form of homeopathic medicine was used to expose people to low safe concentrations of cholera bacillus to confer immunity. As far as is known, vaccination *per se* originated in 1796, when Edward Jenner took fluid from a cowpox blister and scratched it into the skin of an eight-year-old boy. A single blister arose on the spot, but the boy soon recovered. Later, Jenner inoculated him with smallpox matter, and no disease developed. See Ref. [1] for historical discussions on these topics.

Even before obtaining molecular details, there is characteristically consideration of therapeutic chemical substances for using approaches to previous diseases with similar symptoms, recently by repurposing approved drugs and historically by using herbal remedies, in the hope that they may apply. Of course, use of herbal remedies persists strongly today, in all nations. Searching on Covid herbal (without quotes) got 344,000,000 Google hits at the time of writing, and many of the visible hits were clearly as the query intended. Contagion, the 2011 fictional movie about a deadly worldwide virus outbreak, captured the features of early stages of past epidemics by being partly focused on a herbal treatment *forsythia* that was scientifically ill-conceived and promoted by blogging within the plot of the story, but there is in fact an ancient Chinese herbal remedy believed to have microbial and anti-inflammatory activity based on that genus of plant of the same name. At the very least, previously used compounds have been judged safe for most patients by clinical trials, or by centuries of informal clinical trials (by abundant use) in the case of herbal remedies. Sometimes, as for COVID-19 and known herbal extracts such as emodin, ursanoic acids, and steroid-like compounds, similar in appearance to these in plants of genus *Forsythia*, make some scientific sense by modern criteria [8,9]. The chemical formulae of many such plant compounds, earn the nickname "dihydroxy-chicken-wire", a humorous behind-the scenes reference by pharmaceutical chemists to the flexible steel netting with hexagonal holes used to contain chickens and other small animals. However, they often have the above beneficial properties to varying extents, and while it could be said that herbal remedies have not *typically* proven as effective as novel therapeutic agents developed by science to combat infectious disease, it is also to be remembered that many of the therapeutic agents available today have been derived from or inspired by natural products, at very least as starting points for drug discovery and development.

1.5. New diseases versus new variants of concern

Modern researchers have a much larger toolbox for research and development of preventative measures, but as the discussion above implied, in responding to the earliest stages of epidemics, they are in a similar position to their historical predecessors. Early waves of epidemics may involve previously unknown species of pathogens, and subsequent new variants of the identified pathogen can be significantly different in ease and mode of infection, symptoms, and severity. Researchers are then shooting at targets that are unclear or have changed to become unclear respectively, and they must rely on clues from what appeared to work in apparently similar cases before. Hence, the question "How new is new?" is an important one. COVID-19 is currently the obvious example of both kinds of challenge. It is not of known historical concern and the word "coronavirus" did not appear in Garrett's extensive index [1]. Coronaviruses were first considered as potentially affecting human health around 1965–1966 [2–4] and were specifically described as "a new virus" found in the respiratory tracts of humans with common colds [2,3], and officially declared as of a new genus "coronaviridae" in the mid-1970s. They were subsequently known to be responsible for roughly 20%–30% of common colds, but they were only considered a source of potentially serious disease for adults and a global threat for the first time in the SARS outbreak in 2002 [4].

If, after an epidemic, a disease is established in a population (meaning that continues for a significant period or persists at some endemic level), then it provides a reservoir that can form the basis of a new epidemic still due to that same species of pathogen. Technically, an epidemic is an increase in the incidence of disease, etc. in a defined human population that is clearly in excess of that which was expected during a specified period, i.e., above the normal endemic level of disease in an area. It applies even if the endemic level is low or zero, as seemed the case for HIV. A common example of increase above a significant, though low, endemic level is that in Escambia County Florida, the average number of early syphilis cases reported per quarter increased from 15 in 1987 to 75 in 1990; it was of sufficient concern to the Florida Department of Health and Rehabilitative Services and CDC to

investigate used patient interview records to compare characteristics of patients with syphilis diagnosed before and during the increase of cases, and similar situations have occurred several times in many different locations [1]. Such an increase can be due to changes in population or behavioral changes in a human population but can of course be due to changes in the pathogens themselves.

In advanced pathogenic organisms, dangerous variations can appear due to sexual (or sex-like) recombination. This includes spreading of drug-resistance plasmids in bacteria, and in the case of viruses with segmented or modular domains in the genome such as influenza, new strains appear particularly rapidly by *reassortment* of the viral RNA between different variants in the same host cell. The latter has some of the character of *crossing over* in sexual reproduction, which facilitates efficient evolution by tending to preserve, as interchangeable building blocks, the genes or parts of genes responsible for proteins or their subdomains respectively (i.e., parts of larger proteins that were originally separate smaller proteins). Coronaviruses possess no extensive degree of such genomic segmentation to facilitate reassortment, though the appearance of "deltacron" COVID-19, a hybrid of the delta and omicron variants, indicates that reassortment like that in influenza can still occur, even it is with a lower probability. But even if evolution of a virus depends only on accepted mutations, that can be an efficient means of generating new variants if the pathogen has spread rapidly and represents a very large reservoir, as was the case with HIV. In a LinkedIn post on the April 13, 2020 that led to discussion on other sites, the present author calculated that based on global prevalence at time, and viral genome copies per host organism, that there could be at least 10^{21} SARS-COV-2 virus particles in the world, possibly 10^{26} bits or more of parallel viral computational power allocated to working by natural selection to produce variants that are better fitted to reproduce. Be that as it may, the number is inevitably astronomic.

Not all variations have serious consequences, but some do and thrive due to natural selective pressure. COVID-19 alpha was becoming dominant around the beginning of January 2021, delta around May 2021, and omicron in early January 2022. To consider the potential seriousness of accepted mutations in the pathogen proteins, the term "variant of concern" (VOC) is obviously a useful general concept. The term has primarily arisen in widespread use in connection with SARS-CoV-2. It is mainly used for variants of SARS-COV-2 where mutations in the spike protein receptor binding domain (RBD) substantially increase binding affinity (e.g., N501Y) in RBD-hACE2 complex (genetic data), while also being linked to rapid spread in human populations. Several national and international health organizations such as the Centers for Disease Control and Prevention (CDC) (US), Public Health England (PHE), the COVID-19 Genomics UK Consortium for the UK, and the Canadian COVID Genomics Network (CanCOGeN) use many or all of the following criteria to assess what is meant by "concern". These are, increased transmissibility, morbidity, or risk of "long COVID", ability to evade diagnostic tests, decreased susceptibility to neutralizing antibodies and/or antiviral drugs, ability to cause reinfection and/or infection of vaccinated persons, increased risk of serious conditions such as multisystem inflammatory syndrome or long-haul COVID, or increased affinity for particular groups (e.g. children, elderly, or immunocompromised patients). Variants that meet one or a few of these criteria may be labeled "variants of interest" or "variants under investigation" ('VUI'), pending further research. In the case of variants, once knowing the genome, there can be a level of prediction even before the new variant has impact, and in principle there is the opportunity to predict what the consequences might be of certain changes to the genome even before they happen. At present, very cases of the latter occur, insight comes as hindsight, and such predictions before-the-fact remain essentially "in principle". Some useful directions, however, are described here.

For VUIs, the concerns typically relate to changes in amino acid residues in host receptor binding sites and certain regions on the outside of surface proteins of the pathogen (the spike glycoprotein in the SARS-COV-2 case) that serve B-epitopes. These raise an antibody response and are the regions which bind to the antibodies so raised. Being at the surface and often in flexible loops, these change fairly readily, i.e., there is a higher probability of accepted mutations. In contrast, T-epitopes responsible for immune system memory in response to infection or vaccination can be buried inside the protein and exposed by proteolysis; because the amino acids have to fit in appropriately in the manner of a three-dimensional jigsaw, T-epitopes tend to change more slowly. Nonetheless, T-epitopes are not confined to such locations and, in the author's experience, they can often overlap with B-epitopes. That said, a sufficient number of T-epitopes will enable successful vaccination. Here, one functional distinction between B and T epitopes is seen in the omicron variant of COVID-19, in which the spike glycoprotein B-epitopes have changed so that vaccination no longer confers much resistance to infection, but the T-epitopes being often in proteins not at the surface, are largely unchanged by natural evolution, and the cellular response is still efficacious in reducing severity of the disease. In practice, predictions as B and T-epitopes may be boosted, or more correctly stated, ranked, by adding a score based on the appearance of certain amino acid residues in known B and T epitopes, but also based even more pragmatically based on past efficacy of such epitopes when synthesized, linked to a carrier protein, and tested in laboratory animals. For example, the author has often found the presence of histidine and tyrosine to be helpful in obtaining good response to raising antibodies. The size of a potential epitope, i.e., number of residues in it, can also be important for vaccine design, particularly for T-epitopes. The peptides are presented on the surface of an antigen-presenting cell, bound to major histocompatibility complex (MHC) molecules and certain cells in human hosts are specialized to present longer MHC class II peptides of 13–17 residues, while nucleated somatic cells mostly present shorter MHC class I peptides of 8–11 residues.

1.6. The example of the rise of COVID-19 in more detail

On December 31, 2019, the World Health Organization (WHO) was informed of a cluster of cases of pneumonia of unknown cause in Wuhan City, Hubei Province, China (e.g., Ref. [5]). The early official responses could neither reasonably be described as rapid nor as particularly well organized [6]. Because of an interest in emerging epidemics, the present author became aware of these cases very early in January but the news at that time was patchy and unclear; there were five critical days from December 30, 2019 to 3rd January in which the picture solidified [6], but the information was little more than that there was a potentially serious problem emerging, primarily and simply that this was not normal pneumonia. On January 4, 2020, the WHO reported on social media that there was a cluster of pneumonia cases, with no deaths, in Wuhan, Hubei province. On January 9, 2020, it was officially announced that a novel coronavirus had been identified in samples obtained from the Wuhan pneumonia cases, and around 11th January Chinese state media were reporting the first known death definitely caused by the virus. A diagnostic test was more-or-less publicly available by 13 January, on a very limited basis. Human-to-human transmission, a key step in the rise of a zoonotic disease (i.e., of animal origin) was only publicly confirmed by the 20th January [5,6].

All these were triggers that initiated interest in the present author, but preliminary bioinformatics studies of the genome could only begin around January 23, 2020, when Chinese researchers in association with the University of Sydney posted the updated the genome sequence considered as reasonably correct and complete as GenBank entry MN908947.3. That original entry stated in a comment that this sequence version replaced MN908947.2 on Jan 17, 2020, and the current entry at time of writing with minor revision is dated March 18, 2020 [7]. The GenBank entry at the time that it was used most extensively by the present author began as follows.

```
Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome
GenBank: MN908947.3
LOCUS       MN908947               29903 bp ss-RNA     linear   VRL 23-JAN-2020
DEFINITION  Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete
            genome.
ACCESSION   MN908947
VERSION     MN908947.3
KEYWORDS .
SOURCE      Wuhan seafood market pneumonia virus
ORGANISM  Wuhan seafood market pneumonia virus
            Viruses; Riboviria; Nidovirales; Cornidovirineae; Coronaviridae;
            Orthocoronavirinae; Betacoronavirus; unclassified Betacoronavirus.
REFERENCE   1  (bases 1 to 29903)
AUTHORS   Wu,F., Zhao,S., Yu,B., Chen,Y.-M., Wang,W., Hu,Y., Song,Z.-G.,
            Tao,Z.-W., Tian,J.-H., Pei,Y.-Y., Yuan,M.L., Zhang,Y.-L.,
            Dai,F.-H., Liu,Y., Wang,Q.-M., Zheng,J.-J., Xu,L., Holmes,E.C. and
            Zhang,Y.-Z.
TITLE     A novel coronavirus associated with a respiratory disease in Wuhan
            of Hubei province, China
JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 29903)
AUTHORS   Wu,F., Zhao,S., Yu,B., Chen,Y.-M., Wang,W., Hu,Y., Song,Z.-G.,
            Tao,Z.-W., Tian,J.-H., Pei,Y.-Y., Yuan,M.L., Zhang,Y.-L.,
            Dai,F.-H., Liu,Y., Wang,Q.-M., Zheng,J.-J., Xu,L., Holmes,E.C. and
            Zhang,Y.-Z.
TITLE     Direct Submission
JOURNAL   Submitted (05-JAN-2020) Shanghai Public Health Clinical Center &
            School of Public Health, Fudan University, Shanghai, China
COMMENT     On Jan 17, 2020 this sequence version replaced MN908947.2.
```

Of course, there were several fast responses by well-resourced laboratories such as Oxford University based on knowing the SARS-COV-2 genome, and these were soon directed at productive vaccines, as discussed below in Section 1.7. The present author responded primarily to the above GenBank entry with preliminary bioinformatics studies focusing on the spike glycoprotein, and a preprint was posted on ResearchGate on 30th January [7], emphasizing likely bat origins, important conserved regions, and diagnostic, vaccine and peptidomimetic design. Some aspects described, such as the essential SARS-like nature and immediate bat origins, could be considered controversial at the time, but less so later. On the same day, 30th January, the WHO declared a Public Health Emergency of International Concern. Recall that there are three phases in any fast response in such cases, being (1) awareness that there is a possible new infection disease that could become an epidemic, (2) isolating, identifying, and characterizing the pathogen responsible, and (3) obtaining access to a reasonably reliable genome sequence, at least the genes for important surface proteins (the spike glycoprotein in the case of the above virus). This is reflected in the title of the above preprint [7], which therefore referred to the Wuhan Seafood Market Coronavirus, and references were made in the text to the Wuhan seafood market isolate. In principle, the start of the author's project in its bioinformatics phase could have started several days before the 23rd: it is likely that earlier less well validated versions of the genome such as that submitted on the January 5, 2020 by the Shanghai Public Health Clinical Center & School of Public Health, Fudan University, Shanghai, China, could have been very valuable for bioinformatics analysis, but sequence errors can lead to false trails and wasted time, and totally incorrect sequence due sampling errors or contamination by other viruses or organisms are not unknown, leading to withdrawal from databases. Indeed, later news articles asserted that early sequences from early outbreaks in Wuhan were removed from a US government database by the scientists who deposited them, possibly due to similar concerns. As noted above, even when given the news that the virus was a coronavirus, relatively little more could be done by public researchers until the sequence of the viral genome was made widely accessible and verified. As also noted above, that in practice means, primarily, GenBank. Once accessible, the genomic sequence of a pathogen immediately makes possible bioinformatics studies that can relate the causative agent of a new epidemic to lessons learned from any already known related pathogens, lead on to protein science, then wet laboratory biotechnology, and then more rational design of diagnostics, vaccine, and pharmaceutical agents.

The preprint [7] was followed by two fuller reviewed papers by the present author in February and June 2020 [8,9]. All these papers highlighted the risk of emergent new strains and concentrated on the conserved segments of the spike glycoprotein that were at least partly exposed and that must be important to the infection by, and replication and survival of, the virus. Also, on 23rd January, a scientific preprint from the Wuhan institute of Virology was posted on Biorxiv, and e-published in Nature on the 3rd February [10], announcing that a bat virus with 96% similarity had been sequenced in a Yunnan cave in 2013. While earlier genome versions and even those removed are very likely have been important, it is also probably fair to say that the potential seriousness of the outbreak was not fully clear until around the time of the appearance of version MN908947.3. The Chinese researchers with the University of Sydney gave a fuller account of MN908947.3, describing the genes and the associations with other, in February 2020 [11]. It was not until much later, at least by the standards of these timescales, that on February 11, 2020, the WHO named the syndrome caused by this novel coronavirus COVID-19 (Coronavirus Disease 2019), and it was only formally classified as a pandemic as late as March 11, 2020. The severity is, of course, now very clear. On the day of first writing this sentence (January 2, 2022), there had been accumulatively some 300 million known cases of COVID-19 worldwide so far, and 5.5 million known deaths from it, with just under 4,000 reported for January 1st' 2022 alone. Rewriting on March 3, 2022, there were some 441 million cases and just under 6 million deaths. On close-to-final writing on March 18, 2022, there have been 467 million cases and just over 6 million deaths, and 6.28 million deaths at the time of final typesetting corrections. Probably all these numbers are gross underestimates, due to frequent mild or absent symptoms, misdiagnosis, and underreporting. Many journalists have speculated, perhaps not unreasonably, that such global numbers as quoted above could be as much as twice as high, or even three times as high. With the rise of the omicron variant, infection levels were accelerating globally but in many countries such as the US and UK, that peak has now passed. Either way, the situation in January 1st 2022 showed a hugely significant difference from the situation on January 1st 2020.

Early indication of a high fatality rate, essentially the probability of dying if one has the disease, is an important alert. In 1997, a few hundred people became infected with the avian A/H5N1 flu virus in Hong Kong and 18 people were hospitalized. Six of the hospitalized persons died. The rise of COVID-19 was somewhat more alarming. When the present author began the COVID-19 project, it was prior to the WHO announcement and there was just one death definitely due to the virus described in the news, and possibly two. But subsequently 17 deaths were reported as occurring by 23rd January, and in view of the small number of cases, this was sufficient to worry that there was a new

disease or variant with a high fatality rate. The state of knowledge at around the start of the study was that it was probably a form of severe acute respiratory syndrome, SARS, but not necessarily sufficiently close to be called SARS. The number of deaths as described at the beginning of this Section seemed consistent with the concern, nonetheless, because the fatality rate for the earlier SARS appeared to be around 9–10%, and the related Middle Eastern respiratory syndrome, MERS, was even higher at 34–35% (though this may be misleading as many mild cases may not have been reported). While the preliminary analyses indeed indicated that it was SARS [7,8] some authorities were then declaring it was not. Authorities may thus have possibly been making a fine distinction to alleviate public concern, although for researchers outside the innermost circles wishing to gather knowledge, that was possibly less than helpful. It is certainly now considered a SARS-like coronavirus sufficiently enough to justify the final name of SARS-COV-2. Recall that the virus in the earliest paper by the present author was referred to as the Wuhan Seafood Isolate coronavirus or just the Wuhan Seafood Isolate, and later 2019n-Cov: it was not until 11th February that it was named COVID-19 by the WHO, followed by the Coronavirus Study Group (CSG) of the International Committee on Taxonomy of Viruses who named the name of the causative agent as SARS-COV-2 (severe acute respiratory syndrome coronavirus 2) [5].

The author's early papers analyzing the above GenBank entry indicate the simplicity and power of bioinformatics, because as well as emphasizing that it was essentially SARS and that previous SARS studies were likely to be relevant, another interesting observation of the above papers were that "All the top matches are bat host species" [8]. That is probably the most popular choice of immediate host now, at the time of writing this review, although that remained controversial for some time. That is also to be seen in the light of understanding that the immediate host for SARS was known to be the civet, although a bat was subsequently determined to have been the source of civet infection. In July 2020, the present author also predicted that like influenza many coronaviruses and the spike glycoproteins contained hemagglutinin-like binding sites to bind host cell sialic acid, which was contrary to opinion at the time, but did not contain a neuraminidase (sialidase) or similar esterase to reverse the binding, suggesting an increased risk of hemagglutination between red cells, and between red cells and capillary wall, and hence increased risk of hemolytic anemia and multiple thrombosis and kidney damage. Two later papers in the series focused on conserved regions and variations in the coronavirus proteins, one highlighting highly conserved sequence motif in Nsp3 of SARS-CoV-2 as a potential therapeutic target, and one indicating how the highly conserved KRSFIEDLLFNKV motif, a target Achilles Heel of the virus associated with host cell entry predicted in the earlier papers [7–9], becomes more extensively exposed to antibody when antibodies bind elsewhere [14].

1.7. The biopharmaceutical response to COVID-19

The early responses by well-resourced organizations such as Oxford University was also rapid, but it might have been faster still with more immediate funding from government agencies. However, at the time the picture was unclear, the true extent of global danger was not obvious, and perhaps to many administrators it all seemed more academic. Another possible reason was that, traditionally, vaccines have been based on killed or attenuated viruses, without knowing the genome or any other molecular details, but times have changed and perhaps proposals sounded futuristic: responses based on knowing the genome represented almost all the academic and industry responses to COVID-19, and the use of RNA and DNA vaccines had never been tried before in a large-scale response to an epidemic. An awareness by research scientists of the true state of the art was important. In an interview in March 2020 with Bernarda Tundzhay, a health journalist, the present author (BR) was not as skeptical about the speed of development as other experts [15], but keeping in mind that one of the older but faster vaccines to develop, MMR, took 4 years to develop, that there is still no approved successful vaccine against HIV after more than 40 years, and that there is little long-term immunity to the common cold that is a coronavirus infection in 20%–30% of cases, so it was important not to raise false hopes. Quoting the present author the journalist stated that "Usually, it takes about one year to initially test vaccine or antiviral products before moving them into clinical trials… However, during the ongoing Covid-19 global pandemic, there is an obvious need for a quicker turnaround, but a rushed vaccine or antiviral of any kind could cause safety issues, such as where an autoimmune reaction is raised against a patient's own proteins, he added" [15]. The estimate of about a year was considered an optimistic fastest possible limit largely based on the time period for development of some animal vaccines (which typically undergo less stringent tests) but considering that the first steps of rollouts took place in mid-December 2020 and main rollouts in 2021, it was not a bad guess. The fast responses in vaccine development include the Oxford-AstraZenca DNA vaccine effort, said to have taken 11 months. The mRNA vaccine from Pfizer received FDA approval on December 11, 2020 and the Oxford-AstraZeneca vaccine shortly after. For many industrial nations, the one-year estimate was more-or-less "spot-on". AstraZeneca received a conditional marketing authorization valid throughout the European Union on January 29, 2021. Research and development on these vaccines and others such as the Moderna vaccine are considered by many experts and the press as unexpectedly rapid (e.g. Ref. [16]). Many attribute this to the RNA/DNA nature of the vaccine constructs, which were somewhat unexpected. Admittedly, the constructs were ready as a general method of combatting new pathogens, using a "cartridge" or "plug'n'play" approach. The Oxford construct used a common cold virus that infected chimpanzees, ChAdOx1 (Chimpanzee Adenovirus Oxford 1), programming the DNA to encode the spike protein. However, "ready" meant that successes had been largely confined to the laboratory and a few small trials. According to the Oxford group, prior to Covid, 330 people had been given ChAdOx1 vaccines for a variety of diseases ranging from flu to Zika virus, chikungunya, and prostate cancer [16].

There was one unstated prediction or presumption by the present author that was not correct. The use of DNA and RNA-based vaccines and particularly their rapid approval by the FDA etc. for human use were somewhat unexpected. The present author had focused on peptide-based vaccines [7,8,14]. They were still considered state-of-the art and new in the sense that use was still largely confined to veterinary medicine, as in Foot and Mouth Disease. The peptide approach still requires knowledge of the genome or the details of proteins generated from it, and the present author focused on pathogen protein analogues as prepared on a peptide synthesizer, in which he had most experience because they were considered for many years as the most promising new generation of "cartridge" or "plug n 'play" approaches to vaccines both in terms of fast response and relative safety compared with traditional vaccines made from killed or attenuated viruses [17]. By focusing on the parts protein sequences of the pathogen that appear to matter to a B-cell and T-cell response and ignoring only those details concerning unnecessary biological features that might lead to adverse effects, the peptide approach remains attractive. But also, using the DNA or RNA approach in no way diminishes the huge benefits of the computational and knowledge-based approach, the use of bioinformatics, the appreciation of the fundamental features required for diagnostics and vaccine, and the identification of, and response to, variants of concern, as follows.

2. Theory by example

2.1. Extracting knowledge

The mathematical theory of knowledge underlying the Q-UEL language and related inference and prediction methods, as used by the author in relation to the rise of COVID-19, may appear unfamiliar. Consequently, it should be emphasized that it is presented here

primarily as a way of pulling together formal discussion of the *kind* of information and computational tools required. The fight against emerging disease has priority over any personal opinions regarding mathematical elegance, so it is fortunate that there are many beneficial features of the approach used that could be reimplemented in somewhat different, and perhaps more familiar, ways. The focus is on what the tools need to do, of which these are but examples. In the author's opinion, however, the mathematical basis has considerable advantages, and it is arguably the natural and conservative solution, at least in the sense that it builds on a highly successful standard in physics that goes back to the 1930s. Descriptions of the Q-UEL language and uses of it are provided in many published papers: see especially Refs [20–37]. Examples of literature sources from which relevant *qualitative* knowledge (discussed later below) can be extracted are Refs [38–41]. The approach in the case of making quantitative, probabilistic predictions from knowledge associated with probabilities is particularly emphasized in Refs. [24,29–31], and the basic theory of the inference method that underlies such predictions, called the Hyperbolic Dirac Net (HDN), is give in Refs. [42–45]. One way to introduce the ideas more briefly in the context of emerging diseases that are relatively new to science is as follows.

Possibly the most general theoretical statement that can be made about knowledge gathering methods is that if X is something new and hitherto unknown, and Y is something with similar features that is well known and known to have properties Z, then it is certainly worth considering, subject to further scientific investigation, that X has properties Z. Recall comments relevant to the idea of "similar features" such as symptoms for drug repurposing and herbal remedies to treat emerging epidemics including COVID-19 (Section 1.4). This natural and obvious approach has been well-known in a more structured form to pharmaceutical chemists for drug discovery purposes and is useful for similar pharmaceutical reasons even when X and Y are large structures such as pathogens or pathogen proteins. An important step in the present author's COVID-19 project was the observation that the amino acid residue sequence of the spike protein of the pathogen in the Wuhan seafood market isolate was closely related to that of the coronavirus responsible for earlier SARS. In the present cases of interest, knowledge k regarding X, Y, and Z, say as k(X), k(Y), k(Z) and important joint knowledge k(X; Y) and k(Y; Z) from which k(X; Z) is deduced on the above assumption, can take diverse initial forms. Recall (Section 1.2) that the main classes are (a) structured data including of spreadsheets, tables, and last data types such as DNA, RNA, and protein sequences, (b) unstructured data such as authoritative natural language text on webpages, and (c) knowledge repository stores (KRS) containing elements of knowledge extracted from both of the above sources into canonical form that both computers and humans can easily read and use to draw inference, such as in the case of the Q-UEL language discussed below. One might say that there is some functional model for inference *f* such that one can write *f*(k(X; Y), k(Y; Z)) → k(X; Z). In the case of structured sources it is perhaps particularly useful to see the k in k(X; Y) and k(Y; Z) as association constants (with a natural logarithm as Fano mutual information), from which k(X; Z) is deduced using certain interdependency and independency assumptions.

Such associations are, for the above and many other purposes, quantified wherever possible as K(A; B) = P(A|B)/P(A) = P(B|A)/P(B) and conditional probabilities P(A|B) and P(B|A). While K(A; B) is symmetrical, i.e. K(A; B) = K(B; A), the *probability dual* {P(A|B), P(B|A)} discussed below is, in a sense, a kind of dualized or directionalized K(A; B). In general, P(A|B) is not equal to P(B|A): they are mutually related by Bayes' Rule that can be expressed as P(A|B)P(B) = P(B|A)P(A). The above dual is a prominent a feature of the Q-UEL language [20–33] which is based on the Dirac notation. The dual is particularly important in the construction of inference nets that, unlike Bayes nets, and not confined to a Directed Acyclic Graph (DAG) and are thus free of the severe independency assumptions that this can imply (see e.g. Ref. [30]). Recall (Section 1.2) that Q-UEL can be considered as an extension to XML for probabilistic semantics and Artificial Intelligence. Note that one cannot deduce K(A; B) precisely from P(A|B) and P(B|A), nor *vice versa* (though there are constraints) but from just these three measures K(A; B), P(A|B), and P(B|A), many other probabilistic measures can be calculated, notably prior probabilities P(A) including prevalence, joint probabilities P(A,B), negative forms P(not A, B), and so on, and hence many basic measures of epidemiology and evidence based medicine such as positive and negative predictive value, predictive odds, likelihood ratios including risk factors, odds ratios and so on. Unlike XML, in Q-UEL, all tags can take on algebraic and arithmetic force and can be used directly as building blocks of inference networks for automated reasoning. Also recall (Section 1.2) that an example of a typical basic Q-UEL tag being used in programming mode for automated reasoning by an inference net is

$$< \textit{subject expression} \mid \textit{relationship expression} \mid \textit{object expression} > = \{\textit{pfwd}, \textit{pbwd}\}$$

Here entries that are replaced by specific values are in italics. *pfwd* (probability forward) refers to conditional probability such as of form P(A|B), say 0.93, and *pbwd* is its adjoint form such as P(B|A), say 0.27, these being important when tags are used in construction of a Hyperbolic Dirac Net (HDN) as an inference net as described later below, in Section 4.1. Conditional probabilities P(A|B) and P(B|A) are usually sufficient when we can interpret Dirac's basic brake <A|B> (see below) as <A | **if** | B> or <B | **are** | A>, or (with caution) < A | **'is caused by'** | B > and < B | **causes** | A >. All these have an analogous interpretation in quantum mechanics, in terms of vectors <A| and |B>, and Hermitian operator and matrices [21–30]. When the value of an association constant, say 5.64, is also to be assigned, its value can be included in various ways, e.g. by assigning the values {*pfwd*, *pbwd*}, *assoc* to the tag. When generated tags are generated by structured data mining, stored for long term use in the Knowledge Representation Store (KRS), exchange on the Internet, the format is as follows.

< *subject-expression* Pfwd:=*pfwd* | *relationship-expression* assoc:=*assoc* | *object-expression* Pbwd:=*pbwd* >

For implementation of the method in automated reasoning including by inference nets, which unlike Bayes Nets can be bidirectional general graph and evolve under rules of categorical logical, grammar, and definitions, it is important that the value of the above tag, say in general < A | **R** | B> when analogous to <A|B> = <A | **if** | B>, has the following hyperbolic complex value.

$$<\text{A}|\,\mathbf{R}\,|\,\text{B}> = \tfrac{1}{2}\,[\text{Pfwd} + \text{Pbwd}] + \tfrac{1}{2}\,\boldsymbol{h}\,[\text{Pfwd} - \text{Pbwd}] = \{\text{P(A|B)}, \text{P(B|A)}\} \quad (1)$$

Here $\boldsymbol{h}$ is the hyperbolic or split-complex imaginary number such that $\boldsymbol{hh} = +1$, rediscovered in various guises by Dirac (e.g., as linear operators and γ-matrices). The above also reveals that the *probability dual* {P(A|B), P(B|A)} is one way of writing that value, i.e., of writing quantum mechanical, but purely $\boldsymbol{h}$-complex, *probability amplitudes*. Q-UEL stands for "Quantum Universal Exchange Language" because it builds on the Dirac notation and Dirac's associated algebra for quantum mechanics. See Refs. [21–33] for explanation and discussion; the important point for present purposes is that these tags representing elements of knowledge can be brought together in various ways for probabilistic inference and probabilistic semantic reasoning (Section 1.2). In Sections below, the emphasis is on the Q-UEL tags generated by the various methods, and they are sufficiently readable by the human eye that their use in reasoning, and their use in carrying knowledge between algorithms, can be appreciated intuitively (see in particular Refs [24–31]).

While the Semantic Web is not inherently probabilistic and there is lack of agreement on the best probabilistic approach (see discussion in Ref. [21]), the Q-UEL language is compatible with it because probabilistic inference includes handling the case of certainty, i.e., with P = 1 as a limiting case. However, there is a twist. Because statements and data on the World Wide Web are not always true or certain, or represent an

error or do not apply to all possible cases (e.g., do not have universal scope), or change with new information (as was and is commonly the case for web pages about COVID-19), Q-UEL tags from data mining unstructured data as natural language text and bioinformatics sources are often annotated as to provenance and time-stamped. Importantly, in any computation using them, they are treated as *assertions*. They take the value 1 as in e.g. P(A|B) = P(B|A) = 1, and odds take the value 1, by default until there is evidence to the contrary. Perhaps at first counter-intuitively, 1 indicates uncertainty or lack of impacting knowledge. In such cases, Q-UEL tags also assume that by default K(A; B) = 1 and collectively satisfy the requirement that the mutual information content is I(A; B) = ln(K(A; B)) = 0, importantly have no effect on a purely multiplicative inference net, and are in accord with Karl Popper's theory of scientific knowledge discussed in e.g. Refs. [21,24,25,27].

2.2. Some preliminary Q-UEL tag examples from the COVID-19 studies

For example, during the first few days of the SARS-COV-2 studies methods of interacting with the Internet and source data developed for human genomics [32,33] (see below) were refined to generate a tag containing the spike glycoprotein sequence [8] that could be stored in a Knowledge Representation Store (KRS) for future use as input and as a record of the data at the time [9].

male, that can stand alone. The metadata operator ':=' has the value, more specific instance, or example to the right. One advantage of this is that different ontological structures for the same basic kind of information can be expressed and combined, and a main purpose of Q-UEL is to enable interoperability by using this as a kind of "tourist phrase book". Another difference from XML is that several attributes form a logical expression and the default logical operator between attributes, if not shown, is AND (arguably, if the organization of attributes in XML implies a logical expression at all, the basic formalism confines it to AND logic). The above is an example of a well-annotated tag of the type that is usually for permanent storage in the Knowledge Representation store KRS. Simpler tags derived from them can be used as working tags, sometimes temporary, for semi-interactive explorative studies.

In contrast to the above tag that has implied default Pfwd, Pbwd, and assoc attributes of value 1, the following is a tag derived by analysis of many such sequences and the associated three-dimensional structures where known, to derive statistical relationships between amino acid residue sequence patter and *secondary structure* as α-helix (H), β-peated sheet or extended chain (E), and coil or loop (C). Two example applications of this are to study the effect of sequence changes in new variants, and to predict surface coil or loop as a potential B-epitope that can initiate antibody response with antibodies capable of binding to that region.

```
<Q-UEL-ORF-PROTEIN:=(application:='Perl version v5.16.3':= 'GenBank query',
tagtime(gmt):= 'Sun Feb 2 15:41:23 2020' source:=('GenBank
entry':=https://www.ncbi.nlm.nih.gov/nuccore/MN908947.3,
process:='GenBank query':= https://www.ncbi.nlm.nih.gov/genbank/,
definition:='Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome.
', accession:=MN908947, version:=MN908947.3))
ORF:='orf1ab'
product:= 'orf1ab polyprotein'
'protein id':= 'QHD43415.1'
sequence:= 'IUPAC 1 letter aa code':=
'MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQHLKDGTCGLVEVE
:
[omitted for brevity]
:
GQINDMILSLLSKGRLIIRENNRVVISSDVLVNN'
'size class':='protein':=full:=7097
|'has a well conserved subsequence as':='transformed by':=(converter:=BLASTPXTRACTION-
exptal4, 'using':=BLASTp:=
'https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LO
C=blasthome:=('Database':='Non-redundant protein sequences (nr)', 'Organism':=
'Viruses (taxid:10239)','exclude:= '+', top:=100)) |
'conserved sequence':='IUPAC 1 letter aa code':=VVVNAANVYLKHGGGVAGALNK'
Q-UEL-ORF-PROTEIN>
```

```
<Q-UEL-QGOR  conformation(i):=H  Pfwd:=0.32  | if:=assoc:=2.53 |  residue(i-
9):=L residue(i-4):=E  residue(i-2):=L  Pbwd:=0.10  Q-UEL-QGOR>
```

Note that, consistent with the opening remarks in Section 2.1, and because of the implied defaults in Q-UEL concerning probabilistic values, one can in this kind of case make use of similar ideas without reference to the underlying mathematics. The tag is quantitative but only in the sense that it carries biosequence information: there are no probabilities mentioned. Indeed, the above could obviously be automatically expressed in XML, although the result is typically more complicated and less readable and would still need a Q-UEL-like system to interpret and use the information carried. For example, unlike XML attributes, those in Q-UEL can have a rich formal ontological structure, e.g. defined as the Attribute Metadata Language (AML), e.g. a form A:=B, or A:=(B,C) or A:=(B:=C, D:=(E:=F,G)), and so on. Minimally it has the form *metadata:=value*, e.g. *gender:=male* although some exceptions are optionally allowed for authorized nominal categorical data such as

Note that tag value attributes Pfwd, Pbwd and assoc are no longer absent and so no longer imply value 1 by default. In essence, the above is an example of a quantitative, probabilistic element from machine learning and many such tags describing many protein sequences are used to predict the conformation of *each* amino acid residue in a given sequence as being in an H, E, or C state. In earlier parlance used in the field of protein science, it represents a GOR *parameter* for secondary structure prediction and similar kinds of prediction, but now in Q-UEL format. The input for the machine learning process is typically the data in proteins of known sequence and conformation derived directly or indirectly from the Protein Data Bank https://www.rcsb.org/. For example, the following is the sequence of the receptor binding domain in 6M0J, the crystal structure of SARS-CoV-2 spike receptor-binding domain bound with ACE2 which is of particular interest in Results

Section 4. It also further illustrates some common features of Q-UEL use. It is another importance source of amino acid residue sequence information. The methods for generating tags including known secondary structure descriptions and the above Q-UEL-QGOR tag as a result of machine learning from them are described in Refs. [29,33,43,45].

determined the three receptor-binding C-terminal domain 1 (CTD1s) of the S1 subunits in symmetric "down" positions. The binding of the "down" CTD1s to the SARS-CoV-1 binding sites to receptor ACE2 was not possible due to steric clashes, suggesting that the conformation 1 represents a receptor-binding inactive state. Conformations 2–4 also examined were found to be symmetric showing that the ACE2 binding

```
<Q-UEL-PDB-SEQUENCE
'protein id':='PDB entry':= 'FASTA header':='6M0J_2;Chain B[auth E];Spike protein
S1;Severe acute respiratory syndrome coronavirus 2 (2697049)'
|'has sequence'|
sequence:='IUPAC 1 letter aa code':=
RVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
DEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYF
PLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNF
comment:='ACE2 Receptor Binding Domain HHHHHH C-terminal production tail peptide
removed'
reference:=https://pubmed.ncbi.nlm.nih.gov/32225176/:=(authors:='Lan, J., Ge,
J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang,
L., Wang, X.', title:=Structure of the SARS-CoV-2 spike receptor-binding
domain bound to the ACE2 receptor', journal:=Nature, volume:=581,pages:=215-
220, year:=2020)'
Q-UEL-PDB-SEQUENCE>
```

For knowledge that is primarily from structured bio-sequence data combined with unstructured, textual data, points relating to more general principles are most profitably illustrated by examples in overview. Recall that when the author of this review began his studies on the Wuhan Seafood Market isolate entry MN908947.3, it had only been known for a few days that it was a coronavirus, and knowledge gathering tools were required because he had relatively little knowledge of coronaviruses. Further, authorities were maintaining, perhaps to alleviate public concerns, that it was not SARS. The immediate impression gained by use of tools to access standard bioinformatics tools and database on the Internet was that the important spike glycoprotein was essentially the SARS spike glycoprotein. In the first few days, the closest homology (sequence match) of the spike protein with proteins on the protein entries of GenBank by BLASTP were bat coronavirus spike glycoproteins with 77%–81% sequence identity. Just after the work of first study [8] was completed, but in for the second paper [9] submitted on 2nd February, a bat coronavirus spike glycoprotein with 97.41% identity of that of the Wuhan Seafood Market isolate appeared on GenBank, on 29th (submitted 27th) January 2020. This was the controversial posting by researchers at the CAS Key Laboratory of Special Pathogens, Wuhan Institute of Virology, Center for Biosafety Mega-Science, Wuhan. However, even the first matches unlocked a wealth of relevant information that was known for SARS-COV-1 and likely to apply to SARS-COV-2.

Importantly, the SARS coronavirus spike glycoprotein of the 2002–2003 infection, now called SARS-COV-1, was known to bind the ACE2 receptor initially and to have two cleavage sites associated with entry to the host cell. The three-dimensional structures SARS-COV-1 spike glycoproteins had been determined, notably entry 5XLR that had been obtained in 2017 by cryo-electron microscopy to 3.8 Å and refined by conformational calculations, and it was possible to overlay the corresponding important sites such as cleavage sites [8,9]. That study region rotates away from the "down" position by different angles to an "up" position, while the "up" CTD1 exposed the receptor-binding site for ACE2 engagement. It was also known that the above conformational change is also required for the binding of SARS-CoV-1 neutralizing antibodies targeting CTD1. This description could be extended to other betacoronaviruses using CTD1 of the S1 subunit for receptor binding. The beta coronaviruses include OC43 and HKU1 (which can cause the common cold) of lineage A, SARS-COV-1 and SARS-COV-2 both of same lineage B, and MERS-CoV of lineage C. Consequently, it was reasonable to suppose that the SARS-COV-2 spike glycoprotein had similar structure and behavior to that of SARS-COV-1 [8,9], as turned out to be the case when SARS-COV-2 spike glycoprotein structure PDB entry 6VYB became available for comparison on 11th March [10].

Q-UEL tags in the earlier genomic studies, generated some 9 months prior to COVID-19, were still retained in the KRS, and played a role in further studies investigating the role of mitochondrial signaling. Signaling by peptides encoded on small open reading frames in the mitochondrial DNA mitochondria is known to be involved in response to cell stress. One of these earlier example tags is as follows.

```
<Q-UEL-ORF-PROTEIN:=(application:='Perl version v5.16.3':=FASTAtoTags.txt,
tagtime(gmt):='Thu Mar  8 13:46:53 2018' source:=(patient:=75229,
process:=interpretation:='ORFfinder+BLAST':='https://www.ncbi.nlm.nih.gov/orffinder/')
'for patient':=75229 'putative mitochondrial protein':='best match':='ORF2:472:567
small proline-rich protein 2G [Mus musculus]' | is | sequence:='IUPAC 1 letter aa
code':='MLLISSTQPPPILPSTHTPLLTPYPEPTKPQ' 'size class':='peptide-miniprotein (21-35
aa)':=31 Q-UEL-ORF-PROTEIN>
```

Nonetheless, subsequent studies of knowledge gathering from the literature showed that in some respects mitochondria seek to "carry on regardless" to maintain basic housekeeping functions in the cell [32]. In contrast, though, the knowledge captured on tags in the earlier studies usefully contained galectin-3, which also expressed in mitochondria as well as nucleus, cytoplasm, cell surface, and extracellular space, is involved in hyperinflammation and fibrosis in severe covid-19 patients. As further auto-surfing revealed, Galectin-3 regulates mitochondrial stability and antiapoptotic function, mitochondrion, cell surface, and extracellular space. It appeared to be a molecule worth considering both as a target for inhibitors and perhaps as a diagnostic biomarker for extreme COVID-19 response including acute respiratory failure. This

turns out to be the case and these studies and the literature will be reported elsewhere.

2.3. Such approaches are important irrespective of the vaccine development strategy

Since the above examples relate to protein sequence and structure, while the vaccines produced by industry against COVID-19 were (for the first time to any extensive degree) DNA and RNA vaccines, it is useful to comment on why the kind of Theory and implementation discussed above (and Methods described below) remain important. In principle, in considering the model *f*(k(X; Y), k(Y; Z)) → k(X; Z), one might argue that it is best written as *f* (k(X; Y |c), k(Y; Z |c)) → k(X; Z |c), that ensure that it relates to some specific context or conditions c. Asking that we can replace k(X; Z |c) to k(X; Z |c') might well cause *f* to fail. This is reflected in the discussion in Section 2.1 regarding assertions and particularly considerations of *scope* of a statement. Notably, it is not necessarily obvious that some features of the present author's early studies on design of vaccines against SARS-COV-2 were necessarily relevant to the kinds of vaccines that were the first and most successful in combatting the COVID-19 pandemic. The initial papers focused largely on *peptide-based* vaccines, diagnostics, and peptidomimetic drugs [7–9,12–14]. However, the computational aspects, including prediction of the epitopic sites in the pathogen proteins that form the basis of hoped-for peptide vaccines, remain no less valid and no less general. That is because these sites appear in the biotechnologically produced proteins, the spike glycoprotein in the COVID-19 vaccine case, in most cases encoded in the RNA or DNA of the COVID-19 vaccine constructs as discussed above in Introduction Section 1.5. Most notably, these sites can change with different variants, and have done so for variants of concern.

For example, the first tag in Section 2.1 containing the spike glycoprotein sequence could then be used to access other information such as three dimensional structures determined experimentally for SARS-COV-1 and SARS-COV-2 proteins and their interaction with antibodies and the ACE2 receptor, and initiate initial bioinformatics and protein structure analysis studies such as changes in exposure of sidechains considered as the basis for peptide-based diagnostics, vaccines and peptidomimetic drugs [14]. Here # indicates ACE2 binding, and @ antibody binding. A conformationally disordered loop is ~. Extent that sidechain is buried is given by scale 0,1,2,3,4,5,6,7,8,9,X. A smoothed score over neighbors is shown in every case below it, as indicated by 'sm.' At the end of the description to the right, and in the smoothed score the residue obscured by glycosylation is indicated by % [15]. For more details, see Ref. [15].

```
<Q-UEL-VALIDATED-PROTEIN:=(application:='Perl version v5.16.3':=FASTAtoTags.txt,
tagtime(gmt):='Thu Mar  8 13:46:53 2018',
source:=ValidatedMitoProteins.txt:='http://lifeserv.bgu.ac.il/wb/jeichler/MPA/')
'validated protein':='Galectin-3 MPAPH0003' | 'is associated with' |
mitochondrion:=human Q-UEL-VALIDATED-PROTEIN>
```

```
<Q-UEL-PEPTIDOMIMETIC/PEPVACCINE/DIAGNOSTIC
'Interaction block':=447-506:=(
'############################################################ ACE2 binding loops',
'@@@@@@@@@@                          @@@@@@@@@@@@@@          BD23-B binding loops',
'       @@@@@                 @@@@@@@@@@@@@@@@@@@@     CC12.1+CR3022 binding loops',
)
| 'is associated with' |

Pathogen:=virus:=SARS-COV-2
Genome:=Segment:='spike protein''3D structures':=Selected:='Ingine exposure analysis
output'='Standard SARS-COV-2 Ref block':=447-506
'Sidechain exposure and covalent glycosylation block':=447-506:=(
'GNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQ  506',
'~~X95565~~~~~63854653788~~~~~~~~~~~~~~~~~~~85555~~~~~~~~~~~~ 6ZP5-B closed',
'~~676666~~~~~65%56%66666~~~~~~~~~~~~~~~~~~~65554~~~~~~~~~~~~ 6ZP5-B closed sm.',
'~~X95465~~~~~63854537888~~~~~~~~~~~~~~~~~~~85555~~~~~~~~~~~~ 6ZP7-B open',
'~~666666~~~~~65%55%66666~~~~~~~~~~~~~~~~~~~55544~~~~~~~~~~~~ 6ZP7-B open sm.',
'~985553~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~X856672479~~8573 7BYR-A BD23 Fab',
'~332222~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~2223334445555555 7BYR-A BD23 Fab sm.',
'203723314337571643135456624334888385723661321112221146258463 7BYR-B BD23 Fab',
'3443333443%%%%%44%4%%4%455555655665554433222%12222334444443 7BYR-B BD23 Fab sm.',
'~~595653~~~~~838654847999~~~~~~~~~~~~~~~~9X9866476779~~~~9893 7BYR-C BD23 Fab',
'~~999999998777%8899XX9988~~~~~~~~~~~~~~~XX9887778899~~~~9876 7BYR-C BD23 Fab sm.',
'225514111248761854555767924524878487867632251133120145027442 6MOJ RBD+ACE2',
'444333443444555655665555556565666776655444332222223333333332 6MOJ RBD+ACE2 sm.',
'~5825212347741X64665767932512798498887773236125514145813842 3 6XC3 RBD+CC12.1+CR3022',
'~766544434445556556666555555556667776666554443334344444343 32 6XC3 RBD+CC12.1+CR3022
sm.',
)
Q-UEL-PEPTIDOMIMETIC/PEPVACCINE/DIAGNOSTIC>
```

Inevitably every COVID-19 vaccine tried, tested, and found promising inevitably features the spike glycoprotein subsequence targets believed to be first discussed and proposed publicly by the present author, albeit because they generate or contain the entire SARS-COV-2 spike protein or most of it. They include the content of the above tag which is important as the site of binding of many antibodies raised against the spike protein. However, it was noted that region of the sequence is variable amongst coronaviruses, and attention shifted to the above highly conserved KRSFIEDLLFNKV motif [7–9]. The earlier work [7–9] including tags like that above, nonetheless also remains relevant in its details. The whole spike protein does have advantages of multiple sites and polyvalency, and of taking care of glycosylation in a natural way that complicate (but by no means disqualify) use of peptide vaccines, but the possible advantage of focusing on individual segments rather than just presenting the whole spike protein is that it is likely to focus the immune response or conserved regions. For example, the current notorious omicron strain of SARS-COV-2 has many mutations in

the regions reviewed prior to omicron in Ref. [14] as most readily binding antibodies and thus responds to the current vaccines rather weakly as far as the B-cell antibody response is concerned. Consequently, vaccinated individuals can be readily infected even though the cellular T-cell response remains so that the disease is less severe. There is the argument that this situation of a high incidence rate and low fatality rate is an advantage in terms of building up heard immunity, a seeming satisfactory approach if the natural infection were to approach the effect of large-scale vaccination by an attenuated virus preparation. It would seem a very risky strategy, however, for many reasons, not least because there is no guarantee that a new highly infectious strain will be "less kind" regarding the fatality rate.

3. Methods by example: practical support from automatic knowledge gathering

By "Methods" here is meant an account of how Internet information and particularly the Q-UEL tags derived from it are used in the context of strategy for responding to an emerging epidemic, especially as regards the workflow. It is also arguably appropriate to consider here those less obvious knowledge-gathering tools that support the above studies of Section 2 in a practical context, and which speedily provide an initial orientation as to how to go about the kinds of bioinformatics studies touched upon as examples in Theory Section 2. Although it is somewhat of a simplification, the software tools used in the COVID-19 project could be broadly classified into three types that may be arbitrarily called A, B, and C. This also reflects, to a large extent, their order of use in a workflow. Tools A comprise those involved in knowledge gathering from the internet, including generation of alerts based on news items regarding possible new infectious diseases or variants of concern. The results of this may initiate and shape a subsequent project, so they are of crucial importance in the sense that tools B and C will not be invoked to address a potential epidemic without them. Tools B are those that interact with web pages to obtain data and make use of standard bioinformatics tools. Notably, DNA, RNA, and protein sequences are extracted from GenBank https://www.ncbi.nlm.nih.gov/genbank/, with annotation when desired, SIXPACK, e.g. at https://www.ebi.ac.uk/Tools/st/emboss_sixpack/is used to convert DNA or RNA sequence to protein amino acid residue sequences (primary structure), BLASTP e.g. at https://blast.ncbi.nlm.nih.gov/Blast.cgi, is employed to find similar sequences, Clustal Omega, e.g. at https://www.ebi.ac.uk/Tools/msa/clustalo/, is used to align many similar sequences and construct evolutionary trees of relationships, and the PROSITE data base at https://prosite.expasy.org/is used to annotate protein sequences. There were three justifications for this use of standard tools. The first is that there is little point in "reinventing the wheel": much effort by others has gone into the development and refinement of the algorithms. The second is that the methods are indeed standards in effect, with standard default options, and researchers can expect them to behave in a familiar way and for the results to have a particular meaning. The third is that, for the present author and collaborators, this use of the World Wide Web is entirely consistent with the original intent for Q-UEL to be a Web-centered interoperability language [21–24]. The use of a Q-UEL approach to facilitate tool use is that, where appropriate, the public tools can be accessed "behind the scenes" and integrated together with the rest of the Q-UEL system in the manner of a workbench. That was particularly valuable because the integrated Biology Workbench developed at the University of Champagne-Urbana and later implemented at the San Diego supercomputer center has been unavailable for some time due to funding difficulties (http://workbench.sdsc.edu/). Tools C include those algorithms which are not available as standards on the Web, or which do not produce exactly the kind of information required, or in the required form. Those of importance in the COVID-9 project included (i) improved methods of secondary structure prediction, including prediction of surface loops as potential epitopes for vaccine and diagnostic development that can achieve 90%–99% three state (α-helix, β-sheet, loop) accuracy by making make maximal use of large numbers of known protein three dimensional structures without alignment, (ii) prediction of binding sites on pathogen proteins that bind to host sialic acids and hence e.g. mucins in the respiratory and alimentary tracts, and (iii) measurements of sidechain exposure along the protein sequence where the three dimensional structure is known to assist in design of diagnostics, vaccines and peptidomimetics, with particular emphasis on discovering and reporting how exposure increases or decreases with antibody and receptor binding locally and at remote sites when data is available for such complexes. More recently (and not previously described) there are two further kinds of tools that are being developed to help combat COVID-19. These are discussed in Sections 4.3 and 4.4 later below.

All these processes can be linked in a workflow by information exchange and control via the Q-UEL language. Protein modeling and three-dimensional simulations such as the use of molecular dynamics and molecular mechanics are a natural further step to assess conformational and binding free energies, and the means of using the Q-UEL language to drive these will be described elsewhere, but they played a relatively small role in the COVID-19 project with the important exception of ligand binding studies for drug discovery purposes, i.e., determining which potential chemical compounds had appropriate least free energy of binding and hence appropriate binding strength. Otherwise, the emphasis was on use of empirical experimental data when available, including data from the experimentally determined three-dimensional protein structures and protein complexes, not least because of the need for considerable computer power needed to calculate accurately important entropy contributions, as discussed later below.

The impression should not be given that, for a new and unexpected crisis such as an emerging new kind of epidemic, all tools will be available and well-honed though practical experience. In the above application of Q-UEL, a concept of methodological importance well-known to programmers is *extreme programming* (XP). See below. This is not a well-defined formal approach, however. The relevance here is that the essence of a useful response method to combat emerging epidemics is that it is speedy, accurate, and successful, or at least plausible and logically founded, given all the knowledge that is principle available at the time, but not all requirements can be predicted in advance. For example, the molecular biology of pathogens of many epidemics such as AIDS and Mad Cow Disease had several novel aspects. Overall, the approach taken for COVID-19 could be described as using Q-UEL as an architectural principle, but importantly also as a means of facilitating extreme programing. Extreme programming is a software development philosophy very suited to unexpected features of emerging epidemics because it intended to allow rapid response to changing requirements while ensuring a reasonable degree of software quality. In practice, such an approach was necessitated in the COVID-19 project by the following considerations. Despite the present author's interest in bioinformatics and rapid response to emerging infectious diseases, neither the author nor the Q-UEL system were well-prepared with knowledge and expertise concerning coronaviruses. Indeed, Q-UEL *per se* had been primarily developed for use cases in clinical decision support, e.g., diagnosis, selection of best therapy, and prognosis and determination of risk, and for detection of medical claims anomalies and fraud. The diseases of particular interest had been such as congestive heart failure, renal failure, and cancers, i.e., not primarily infectious diseases nor necessarily

associated with them. An extensive bioinformatics component had only begun to be introduced to facilitate the study of genetic factors. Epidemiological interest, although represented, had been primarily concerned with toxicological aspects of public health such as air quality and its impact on clinical decision support. Monitoring of infectious disease in populations had only just begun to extend that in a natural way. Consequently, in the early days of the COVID-19 project, Q-UEL and associated tools were repurposed and adapted "on the fly" and with frequent manual intervention because the pressing need and focus was naturally regarding a rapid response to the rise of COVID-19, not on further commercial tool development.

Fortunately, the specification of Q-UEL lends itself well to such rapid application and adaptation, i.e., as extreme programming. That includes capture of expertise by progressive conversion of manual intervention to an automatic protocol, usually expressed in Q-UEL itself. Typically, this involves taking a template program in which a Q-UEL tag is represented with parts that are variables and modifying preceding regular expressions (match-and-edit instructions) to extract text and numeric information from webpages or other incoming information, to assign values to those variables. Such short program scripts (including data capture, regular expressions, and tag template) are known as *converters*, because they convert a variety of source information in diverse formats to the canonical Q-UEL form. In many instances, the incoming information can already be in Q-UEL form, and often this involves combining several tags, containing knowledge from different sources, into one tag.

As indicated earlier above, the important initial steps in workflow, from which all else follows, involved the use of Q-UEL to auto-surf the Internet and text sources to extract knowledge from natural language text [25–27], given one or more simple initiating queries such as "SARS" or (later) "COVID-19". Several queries like multiple choice answers in a multiple-choice medical licensing exam, and a body of text analogous to the examination question, can be used along with tests of valid authoritative biomedical text by lists of appropriate Latin and Greek roots, medical terms, along with a dictionary of words and phrases more characteristic of non-authoritative medical sites and non-medical sites generally, help ensure relevance and focus [26]. Prior to that, automatic monitoring of the Internet can alert that a new disease or variant can be arising [17].

At present, there is inevitably some screening by humans of the information being obtained to identify cases that are more likely to be of genuine concern, though clearly developments in AI will help filter the wealth of information that can be generated. In most cases this involves manual or semi-automated curation of Q-UEL XTRACT tags that carry extensive annotation about the source and context and are time-stamped. There are two reasons why the present author personally suspected and investigated an emerging pandemic from the Wuhan Seafood Market isolate with some promising features from that study as described above. The first is simply because of a continuing interest in identifying and responding to emerging epidemics. The present author had experience for several years as an epidemiologist in the Caribbean studying emerging viruses such as Zika [17] when also a Professor of Epidemiology, Biostatistics, and Evidence Based Medicine, as well as with earlier pandemics. Earlier, he also had earlier experience leading the teams that responded first to HIV [18] followed by several HIV diagnostics patents, and on Mad Cow Disease (BSE), inventing the vaccine marketed worldwide by Abbott Laboratories, e.g. Ref. [19], as well as several animal vaccines, diagnostics, or immunotherapeutic agents. However, he was almost completely ignorant regarding coronaviruses, and the availability of Q-UEL knowledge-gathering tools was hugely beneficial. It was an approach which can also be applied to structured and semi-structured data including genomics and bioinformatics data [32,33].

While experience of any kind of problem can certainly help, a main priority should be to capture that expertise so that any researcher can use it. Moreover, it should be possible, where desirable, to make research study recoverable and *reproducible* when used by the same researcher, or anyone else. It is primarily these considerations that necessitate automation, but it also means as discussed in Section 2 that knowledge captured should include information as to circumstances and *provenance*. For example, the autosurfing of the Internet for latest updates regarding COVID-19 encountered the following elements of information from Wikipedia, cited and dated as shown, which is important as Wikipedia content can be updated, and this was especially so for the COVID-19 pandemic. Some further examples of the Q-UEL approach, and of Q-UEL tags relevant here, are as follows. For example, an original extract (an XTRACT tag) obtained by autorsurfing and knowledge

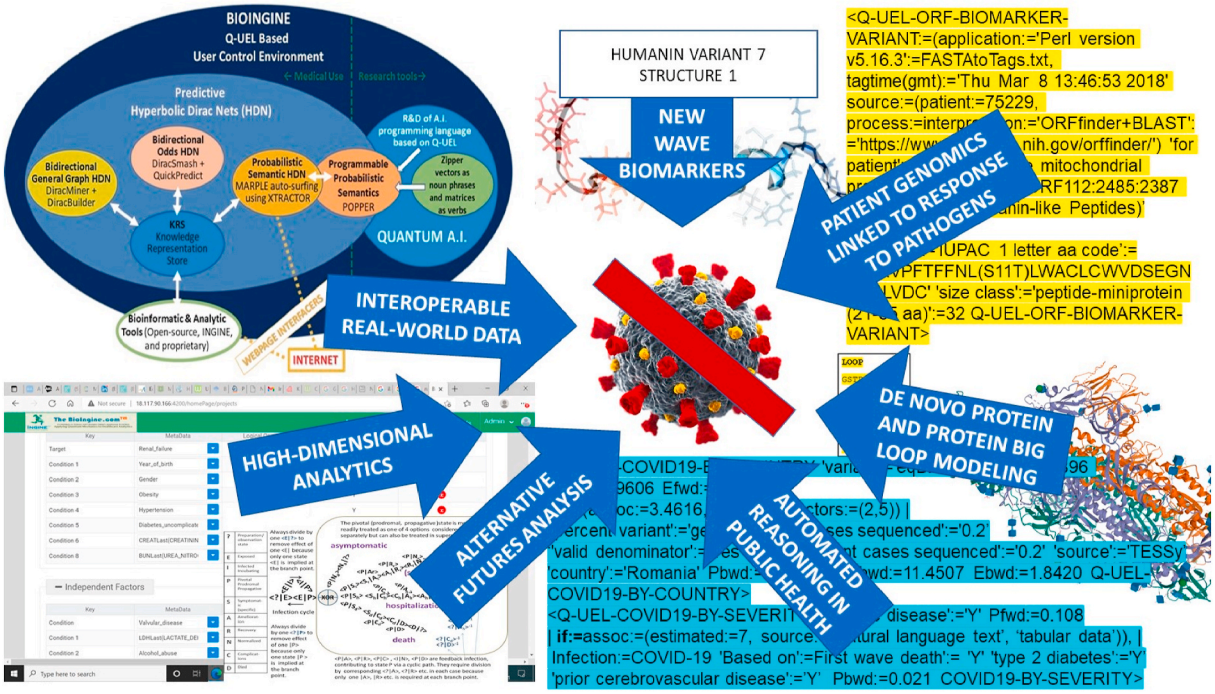


Fig. 1. Seven key technologies that are argued to be important for early response to emerging epidemics.

capture was as follows.

```
<Q-UEL-XTRACT-Marple41W. "Severe acute respiratory syndrome |as| SARS |^is|
`a viral respiratory _disease
[0https://en.wikipedia.org/wiki/Respiratory_disease] |of| zoonotic
[0https://en.wikipedia.org/wiki/Zoonotic] origin |^caused by| `the SARS
coronavirus [0https://en.wikipedia.org/wiki/SARS_coronavirus] |as| SARS-CoV
|Between| November 2002 {AND} July 2003 'an (^outbreak) |of|  SARS |in|
southern China [0https://en.wikipedia.org/wiki/Southern_China] |^caused| `an
eventual 8098 cases |with _value ^resulting in| 774 _deaths reported |with
_value in| 17 countries
[1https://www.sciencedirect.com/science/article/abs/pii/S0277953606004060?via
%3Dihub]  |with| `the majority |of| cases |in ^mainland| China {AND} Hong
Kong [2https://www.who.int/csr/sars/country/table2004_04_21/en/]  |with
_value as| 96/ fatality [0https://en.wikipedia.org/wiki/Case_fatality] _rate
|according to| `the World Health _Organization
[0https://en.wikipedia.org/wiki/World_Health_Organization] |as| (WHO)
[2https://www.who.int/csr/sars/country/table2004_04_21/en/]  |as| `No cases
|of| SARS |^have ^been ^reported| worldwide |with _value since|
[3https://www.nhs.uk/conditions/sars/;  |In| late 2017 Chinese scientists
|^traced| `the virus |through| `the intermediary |of| civets
[0https://en.wikipedia.org/wiki/Civets] |to | cave-dwelling| horseshoe bats|
[0https://en.wikipedia.org/wiki/Horseshoe_bat]" | 'was extracted from' |
source:='https://en.wikipedia.org/w/index.php?title=SARS&amp;redirect=no'
time:='Fri Jan 31 15:20:26 2020' extract:=69  Q-UEL-XTRACT-Marple41W>
```

As mentioned in Introduction Section 1.2, the sometimes-stilted form when a Q-UEL tag is read directly by eye (which it need not be) is because sentences, subsentences or integrated sentences are reparsed into as linear sentence structure as possible, so that if required they can be easily decomposed into semantic triples, i.e. $<A \mid \textbf{\textit{relationship}} \mid B>$, $<$ **B** $\mid$ ***relationship*** $\mid C>$, and so on, and used in the most common forms of automated inference. They are still natural-language-like, which facilitates development, debugging, and maintenance. They are also intended to be responsible for robustness of a medical system if part of the IT and communications infrastructure is lost in a disaster, since they can still be understood by eye with relatively very little effort.

By August 2020 it was possible to obtain more detail regarding understanding of the pandemic and its symptoms, for example, as follows. The unusual link law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode = L0050039] in the second of these tags is till active at the time of writing this paper, and relates to as study article "Special Act for Prevention, Relief and Revitalization Measures for Severe Pneumonia with Novel Pathogens" from the Chinese Ministry of Health and Welfare, as a study item for Chinese law students.

Following general knowledge capture of the above kind, integration of knowledge with the kinds of bioinformatics studies of Section 2 becomes possible. Several other useful observations could quickly be made, including many regarding potential "in-a-pill" therapeutics and much closer to the X, Y, Z model (Section 2.1) as applied by pharmaceutical chemists. Notably, it was known that the small plant compound *emodin* was an inhibitor of SARS-COV-1 infection and cell entry and of the human inflammatory response enzyme 11β-hydroxysteroid dehydrogenase type 1. Though perhaps more familiar as a laxative, it has many known beneficial effects including anticancer, anti-inflammatory, antiviral, antibacterial, anti-allergic, anti-osteoporotic, anti-diabetic, immunosuppressive, neuroprotective and hepatoprotective properties. It was thus possible for the present author to continue early computational studies on the experimentally known potent inhibitors of that enzyme, now extended to computational studies on emodin and similar compounds (e.g. Refs. [8,9]).

In later exploring a broader panel of potential drugs and protein targets, Remdesivir as a broad-spectrum antiviral now in use against SARS-COV-2 was not in the published list of which the present author explored binding affinities, but the closely related natural ADP-ribose-1″-phophate compound and drugs with similar binding affinities and some similar features such as cancer and lymphoma drugs, and notably Favipiravir which at high doses has potent antiviral activity in SARS-CoV-2−infected test animals, were investigated and reported [13]. First,

```
<Q-UEL-marple41Wuhan.txt "`the group |of| _diseases |For| `the group |^see|
Coronavirus _disease [0https://en.wikipedia.org/wiki/Coronavirus_disease]" |
'was extracted from'|  source:='https://en.wikipedia.org/wiki/Coronavirus
disease 2019' time:='Sat Aug  8 11:23:47 2020' extract:=59 Q-UEL-
marple41Wuhan.txt>

<Q-UEL-marple41Wuhan.txt "_disease |^called(?)| COVID &and 2019-nCoV acute
respiratory _disease &and Sars-Cov &and Novel coronavirus pneumonia &and
Severe pneumonia |with| novel pathogens
[0https://law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode=L0050039]" | 'was
extracted from'| source:='https://en.wikipedia.org/wiki/Coronavirus disease
2019' time:='Sat Aug  8 11:23:47 2020' extract:=68 Q-UEL-marple41Wuhan.txt>

<Q-UEL-marple41Wuhan.txt
"[0https://en.wikipedia.org/wiki/Specialty_(medicine)] Infectious _disease
[0https://en.wikipedia.org/wiki/Infectious_disease_(medical_specialty)]Sympto
msFever |(as context for)| _cough |of| fatigue shortness |of| _breath _loss
|of| _smell"  | 'was extracted from'|
source:='https://en.wikipedia.org/wiki/Coronavirus disease 2019' time:='Sat
Aug  8 11:23:47 2020' extract:=70  Q-UEL-marple41Wuhan.txt>
```

for the latter study, a highly conserved sequence motif in Nsp3 of SARS-CoV-2 was investigated as a therapeutic target to capture knowledge about the functions of similar sequences, in examples of automated surfing of the Internet to gather related in other virus and prokaryotic and eukaryotic organisms in semi-structured data sites.

```
<Q-UEL-BIOINFORMATICS:=( application:='Perl version v5.16.3':=BLASTtestScript.txt,
time:= ' Thu May 27 14:32:49 2020')
tool:=BLASTp:= https://blast.ncbi.nlm.nih.gov/Blast.cgi,
'source header check':= '!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
', html:=http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd,
Query:= 'VVVNAANVYLKHGGGVAGALNK'
'input parameters':=(standard, 'job title':= 'VVVNAAN domain core', database:= 'Non
redundant protein sequences (nr) ', organism:=exclude:= 'viruses (taxid:10239) ', 'max
target sequences':=100)
| generated |
'commonly recurrent strings':=( 'macro':=63, 'domain':=63, 'domain-containing':=58,
'TPA':=37, 'ribose':=13,  'ADP':=13, 'ATP':=12, 'AAA family':=12)
'query hits':=( 'macro':=63, 'AMP':=0, 'phosphat':=0, 'polymeras':=0)
Q-UEL-BIOINFORMATICS>
```

By more general auto-surfing, it was quickly found that the SARS-COV-2 protein of interest contained a universal nucleotide binding domain called the macro domain.

```
<Q-UEL-marple39 "`the Macro domain &or A1pp domain |^is| `a module |In| molecular
biology &or, |as| `a module |with _value of about| 180 amino acids
[0https://en.wikipedia.org/wiki/Amino_acid] |^can ^bind| ADP-ribose
[0https://en.wikipedia.org/wiki/ADP-ribose], `an NAD
[0https://en.wikipedia.org/wiki/Nicotinamide_adenine_dinucleotide] metabolite
[0https://en.wikipedia.org/wiki/Metabolite] |^related| ligands
[0https://en.wikipedia.org/wiki/Ligands]"
(source:='https://en.wikipedia.org/wiki/macro domain' time:='Tue May 26 10:43:36 2020'
extract:=62) Q-UEL-marple39>
```

Not all the tools found of value in the COVID-19 project can be discussed here, but in briefest possible review the following is of interest: see Fig. 1 as a guide in the context of COVID-19 and future possible epidemics. In the opinion of author of the present review, there are seven key technologies that will be important for early response to emerging epidemics involving new pathogens or strains. As Fig. 1 states, the Q-UEL language [21–38] has been used to help study them and is progressively incorporating them. Proceeding from the top clockwise, these are (i) the use of new generations of peptide biomarkers [32,33], (ii) analysis of patient genomics (including proteomics) regarding response to pathogens [28,29,32], (iii) improved *de novo* modeling of proteins such that large loops on polymorphic patient proteins, and not least of those on the pathogen proteins and their interactions with receptors and antibodies, can be simulated (including with better entropy calculations) [34], (iv) automated reasoning in public health [36,37], (v) alternative futures analysis (discussed below, using the example of different paths in development of COVID infections), (vi) high-dimensional analytics [24,27,29,30], and (vii) management of Real-World Data including interoperability [20–23,27,31]. These are believed to meet at least many of the challenges raised in Ref. [1].

4. Example results and discussion

4.1. General observations

In most of the previous Q-UEL papers the new Q-UEL tags and algorithms that the papers introduced have been considered as new results to be placed in the Results Section, but the emphasis in this paper is review is to a large extent on established Q-UEL tags and algorithms and on their benefits for a fast response to an emerging epidemic, e.g., for rapid production of diagnostics and then vaccines, and ideally therapeutic drugs. It therefore seems unfortunate that in no case could it be said that a diagnostic, vaccine, or new drug against the COVID-19 virus SARS-COV-2 was developed directly by the author, even in collaboration. However, it was almost certainly possible in principle, at least on a laboratory scale and if given sufficient laboratory resources, because diagnostics and sometimes vaccines based on raising antibodies in test animals were soon constructed and tested by collaborators in previous emerging epidemics based on earlier computational methods such as those described in Refs. [17–19]. It is relevant to note, though, that the major difference in the case of the earlier efforts was the much greater span of time between isolation of the pathogen responsible and proposals for diagnostics and vaccines. In the case of AIDS, the virus was isolated in May 1983 but the proposals by the authors and collaborators were made in January 1987, although it took some time for researchers to establish sequences for genes for at least two variants of a key surface protein, comparison of which was important to establish subsequences as vaccine targets [18].

More to the point, for COVID-19 the speed of response and publication and in particular the high level of citations of the first two peer-reviewed publications and increasing daily indicates that the work has been helpful to COV-19 researchers. Several aspects discovered by the bioinformatics and knowledge-gathering approaches described above have been followed up particularly well by other researchers, notably the subsequence KRSFIEDLLFNKV which has been commercially available as a peptide product by several commercial organizations for research purposes, as discussed in Section 4. 2 below. Although the ACE2 receptor binding region and antibody binding region was well studied in the above studies [7–9,14], as well as prediction of the binding of the spike glycoprotein head to host cell sialic acids [12] and the core of the highly conserved Nsp3 domain as potential pharmaceutical targets [13], probably the best-known observation from the above studies is that KRSFIEDLLFNKV is a likely Achilles heel for SARS-COV-2 [7–9]. This area of research is discussed as an example in section 4.2. Some useful new tools may be considered as results, though in many cases they were modifications of algorithms in previous papers. That included bioinformatics tools and tag types that were considered in papers prior to those reporting the COVID-19 study, e.g., the report on

the study the mitochondrial genome [32] published just before the COVID-19 investigations (on January 12, 2020). Other tools can reasonably be considered as technological results of the COVID-19 project. These tools included an algorithm for predicting sialic acid non-covalent binding sites on proteins [12], and a novel algorithm for sidechain exposure and study of the effects of, e.g., viral surface proteins interacting with antibodies and host cell receptors [14], which is described in more detail in Section 4.5 below. There was use of a new protein secondary structure prediction algorithm capable of three-state with 90–99% accuracy when using contemporary large data bases protein structure [45]. This was mainly split away from the main line of COVID-19 papers, so its value within the COVID-19 project, and first application to SARS-COV-2 proteins is emphasized here. For present purposes, the main importance of this is that loops are of particular interest as verifying putative B-epitopes which raise an antibody response and as regions which bind to the antibodies so raised.

4.2. Importance of invariant regions: example of the KRSFIEDLLFNKV motif

An important consequence of the project overall was the importance placed on invariant regions of the proteins of the group of viruses to which SARS-COV-2 belongs. As was pointed out from the very first paper, a highly conserved region across at least a group of the coronaviruses, and especially one at the protein surface and exposed to the environment, means at least that (a) it serves some important function or functions for members of that group, and (b) also represents a site that is not likely to change easily so that established diagnostics, vaccines and drugs based on the original causative agents of the epidemic might become useless. In other words, they represent an Achilles heel of the virus. Many things about SARS-COV-2 seem obvious in hindsight but were not so at the time. Once the first papers [7–9] had confirmed that the Wuhan Seafood Market isolate was essentially SARS, and once detailed alignments of spike glycoproteins from many spike proteins were performed with particular attention to the KRSFIEDLLFNKV region, its likely importance quickly became clear. It was notable by virtue of being functionally important to the SARS coronavirus and less susceptible to accepted mutations. While the ACE2 region has also now been well studied as a target by other researchers, it was soon seen to be variable amongst coronavirus and at risk from escape mutations, a notion supported by the emergence of the omicron variant (see below). In contrast, KRSFIEDLLFNKV is a well conserved motif across all the coronaviruses, and arguably recognizable beyond them into the nidoviruses. The S2′ region including it (see below), residues 800–839, is quite well conserved in the coronaviruses in the sense that amino-acid substitutions by accepted mutations that would be considered conservative substitutions, i.e., they have similar physicochemical and conformational properties. This is especially so around the C-terminal (right hand side) arginine R constituting the S2' cleavage point in bold and underlined, though substitutions of the phenylalanine F by other hydrophobic residues in common. Notably, RSAIEDLLFDKV is characteristic of common cold coronavirus, also found in the coronaviruses of dogs, cats, rodents, pigs, rabbits, camels, ferret badgers, raccoon dogs, etc. [9]. So far this has, as predicted, been conserved in SARS-COV-2 variants, and the accepted mutations in the omicron variant are not found in this region (mutations D796Y and N856K are the closest along the sequence).

The subsequence KRSFIEDLLFNKV is functionally important to the virus because it includes the S2′ cleavage site at the arginine R, involved in the key stage of virus entry into the host cell. Presumably to protect itself from antibodies and untimely enzymic cleavage, the spike glycoprotein exposes its functional sites in a series of steps. KRSFIEDLLFNKV is partially exposed even in the closed state and more fully exposed after binding of the spike to ACE2, as well as after antibody binding at and near the ACE2 binding site [9]. This S2′ cleavage appears to be absolutely required for reconfiguration of the spike to attain significant levels of fusion between virus and host cell membrane following initial binding at ACE2. The S2′ site resides at some distance from the ACE2 binding region, at the stem of "bundle-of-flowers" of the trimeric S proteins, so the details of the activation mechanism prior to SARS were not obvious. The conformation 1 of SARS coronavirus earlier determined experimentally was three-fold symmetric and has all the three receptor-binding C-terminal domain 1 (CTD1s) of the S1 subunits in "down" positions. It was primarily pre-COVID-19 studies on the cleavage process of S protein of MERS-CoV that had shown how S1/S2 cleavage occurs first and increases the exposure S2' site for enzymic cleavage, and because of spike glycoprotein homologies with SARS and the Wuhan Seafood Market isolate, a similar cell entry mechanism seemed likely, as follows. The "down CTD1" of S1 protein locates immediately above the S2 subunit, and the opening of CTD1, especially by binding the receptor, would remove steric restraints and trigger rearrangement of the spike protein trimer with the release of the S1 subunits and extension of pre-fusion S2 helixes to form a post-fusion S2 long helix bundle. Seeking to blocking this crucial cell entry process for SARS-COV-2 has become a recognized strategy [38–40]. For example, in a subsequent paper [38] by a Chinese research group it was shown that EK1, a pan-coronavirus fusion inhibitor, targeted the HR1 domain (894–966) of S2 protein and could inhibit infection by many human coronaviruses.

The segment KRSFIEDLLFNKV has now been studied by many other researchers. Genetic Engineering and Biotechnology News of December 14, 2021 [39], referring a research paper on Keeping SARS-CoV-2 Reinfections at Bay [40] that cites the present author, notes that "*To counteract the loss of valuable financial resources and specialized professional facilities produced by escape mutations during the development of vaccine, a convergent effort toward discovery of highly immunogenic conserved sequence of viral proteins is dire need of the day. It has been reported that the amino acid sequence motif KRSFIEDLLFNKV, found in spike protein, is one of the conserved regions in coronaviridae family. The motif is partially associated with cellular entry of virus into host cell. Researchers consider this sequence as one of the most vulnerable yet conserved sequence in coronaviruses. Exploration regarding the role of this spike protein sequence is necessary to assess and confirm the degree of attachment of virus to host cell. It can be a valuable target for long-term immunity …*". The peptide KRSFIEDLLFNKV has been advertised by many peptide synthesis companies and studies on and similar sequences as synthetic peptides have revealed interesting physicochemical properties. Notably, the conformation and aggregation of the peptide corresponding to the variant RSAIEDLLFDKV mentioned above has been studied [41].

4.3. Preliminary studies for a tool for predicting the severity of SARS-COV-2 variants

Appropriate amongst what may be considered as results of the project are some new methods that emerge from that project. The following is a recent preliminary result in the author's COVID-19 project that has not previously been described and was developed while writing this paper. Normally, the natural and obvious choice of many researchers for *predicting* "variant of concern" (VOC) in advance of significant epidemiological data concerning its effects, but when the original and current genomes are known, is to estimate by computer simulations the free energy of binding of ACE2 receptors and/or

antibodies [46]. Earlier simulations can be extended to include the affects of the accepted mutations in variants, and so assess whether they are VOCs [47]. A VOC score is thus envisaged by many researchers as ultimately based on computed free energy differences, between the bound and unbound state. The computational chemistry simulation approach is valuable and was used by the author to study the binding of potential drugs against COVID-19 [8,9,13] in cases where the binding site on the protein was relatively rigid. The problem is that such calculations of this kind are not wholly reliable not least because of difficulties in converging the entropy that can dominate determination of stable protein structures and protein-protein interactions [34]. This situation is worsened for computing a VOC score for SARS-COV-2 because the free energy of importance is the difference between unbound and bound states involving large conformationally disordered loops, mainly in the unbound reference state. Solving the proper behavior of a large conformationally flexible loop is akin to solving the *protein folding problem* [34]. But to worsen matters still further, the bound state is unstable in the sense that the whole spike glycoprotein is a flexible machine for cell entry. The functionally important conformational changes of the spike glycoprotein that occur on ACE2 and/or antibody binding and can be seen in experimental three-dimensional structure determinations of the complexes in atomic detail [14]. For simulations, this adds a further layer of much greater complexity. More reliable entropy calculations can only be attained by a great deal of computing power, and IBM's Blue Gene supercomputer, originally motivated by the desire to solve the *de novo* protein folding problem still failed to solve it despite many other useful success in protein science [34]. Even if we consider such computations as potentially highly accurate, it seems clear that for large proteins that can exists in several conformational states, with large flexible loops, that computation of the overall free energy (from enthalpy and entropy) can take a very long time even on a powerful computer.

The argument is that quick "red flag" VOC score is needed that will at least put focus on what variants need to be examined first by more sophisticated computationally intensive methods. Building on the kind of tag shown in Section 2.2, the essential content of the Q-UEL tag attribute as displayed for alignment new variants would be exemplified as follows. A new feature is the use Greek characters **χ**, **φ**, **Π**, etc. that represents the nature of the change as difference between the physicochemical properties of the amino acid residue at that locus in the original Wuhan sequence and those of the accepted mutation, as given along with the basis of the algorithm in Table 1. The table also contains parameters, rounded to the nearest integer, that add up to a score for the new sequence based on the kinds of output described soon below.

The empirical approach was as follows. The scoring will be such that the score as a Variant of Concern is 0 for the original Wuhan strain (the concern is as to it being an even more worrying variant, not that the original Wuhan strain was not worrying). At this stage this scoring method is intended to be crude in the sense of bundling together several transmissibility, morbidity, "long COVID", ability to evade diagnostic tests, neutralizing antibodies or drugs, ability to cause reinfection and severity of associated symptoms in different population groups. Initially, parameters were semi-automatically assigned by reference to SARS-COV-2 and its variants of concern but also with SARS-COV-1 variants, aligning multiple sequences and noting what amino acid residue features differed in variants that appeared to have had more serious consequences. Middle East respiratory syndrome-related coronavirus (MERS-CoV), e.g., GenBank entry QGV13484.1, was also initially included because as SARS-like outbreak starting in 2012 in the middle East has some similarities to SARS and COVID-19, but while SARS-COV-1 and SARS-COV-2 belongs to betacoronaviruses lineage B, MERS belongs to betacoronviruses lineage C with substantial differences, with a recognizable if weak sequence similarity only beginning at the segment RVQPTESIVRFPNITNLCP of Wuhan SARS-COV-2 and EAKPSGSVVEQAEGVECD of QGV13484.1. This is outside the SARS-COV-2 RDB sequence, and so MERS was excluded from the investigation. Including SARS-COV-1 variants and comparison with SARS-COV-2 Wuhan reference sequence and variants must, of course, also be done cautiously, but it is possible to identify analogous residues and consider them as variants of the reference Wuhan strain, to some extent. For example, using the standard Clustal Omega tool at https://www.ebi.ac.uk/Tools/msa/clustalo/, the sections of sequence associated with the ACE2 receptor binding domain in Wuhan reference strain GenBank MN908947.3 is aligned with Protein Data Bank entry 6NB6 which uses the SARS-COV-1 sequence in the structural determination deposited in 2018 of SARS-CoV complex with human neutralizing S230 antibody Fab fragment. Note that SARS-COV-1 represented by 6NB6 has significant changes in the ACE2 and antibody binding region involving the SARS-COV-1 6NB6 segment NVPFSPDGKPCTP-PALN (see also later below), not least the deletion '-', but polar, small non-polar (hydrophobic) or large non-polar character tends to be conserved which is in the accord with the principles of conservative substitution and do allow the customary groupings of residues with similar physicochemical properties as a starting point, and do earmark locations which are highly conserved across the SARS-COV-1 and SARS-COV2 betacaronavirus lineage B group. Amongst the changes that are more remarkable is that the asparagine N at the start of

Table 1
Preliminary example parameterization to predict variants of concern.

Change from Wuhan reference	Symbol	Both antibody @ and ACE2 # binding	Antibody binding residue @ in at least one structure	Antibody binding residue @ in at least one structure and flexible loop ~ or buried index <2 (unsmoothed) in at least one structure	ACE2 binding residue # in at least one structure	ACE2 binding residue @ in at least one structure and flexible loop ~ or buried index <2 (unsmoothed) in at least one structure	Neither antibody nor ACE2 binding in any aligned structure
Positively charged from neutral	+	10	9	8	8	7	1
Negatively charged from neutral	-	8	7	6	6	5	1
Switches charge	±	9	8	7	4	3	1
Polar neutral from charged	**π**	4	3	2	2	2	1
Polar neutral or G from hydrophobic	**Π**	5	4	3	3	2	1
Hydrophobic from polar neutral or G	**φ**	7	6	8	5	4	1
Hydrophobic from charged	**Φ**	8	7	9	6	5	1
Changes but stays in positive, negative, polar neutral or G, or hydrophobic class	**χ**	2	2	2	2	2	1

the above segment has a small polar sidechain while a glutamate E carrying a negatively charged sidechain occurs at the same locus in SARS-COV-1. This might be a feature that give rise to the medical consequences of SARS-COV-2 compared with SARS-COV-1, perhaps the greater transmissibility of the former.

physicochemical properties of amino acid residues in SARS-COV-2 variants of concern using **Π** to indicate a change from a highly polar (charred) sidechain to a hydrophobic one, **χ** to indicate a change but to a sidechain of similar properties, + to indicate a positive charge from neutral, - to indicate a negative charge from neutral, ± a change of charge, and **φ** a hydrophobic from polar neutral or glycine (G). These

```
Wuhan       RVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFK    60
6NB6_A      RVVPSGDVVRFPNITNLCPFGEVFNATKFPSVYAWERKKISNCVADYSVLYNSTFFSTFK    60
            ** *: .:********************:* *****:**:**************: *****
Wuhan       CYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNS    120
6NB6_A      CYGVSATKLNDLCFSNVYADSFVVKGDDVRQIAPGQTGVIADYNYKLPDDFMGCVLAWNT    120
            ***** ********:*********::**:********** ************ ***:***:
Wuhan       NNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQ    180
6NB6_A      RNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCTP-PALNCYWPLNDYGFY    179
            .*:*:.  ***** ** :*:.:*:*******.  :. ...**.   .:***:**:.***
Wuhan       PTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKK    240
6NB6_A      TTTGIGYQPYRVVVLSFELLNAPATVCGPKLSTDLIKNQCVNFNFNGLTGTGVLTPSSKR    239
             *.*:***************:********* **:*:**:**************** *.*:
Wuhan       FLPFQQFG     248
6NB6_A      FQPFQQFG     247
            * ******
```

Importantly, also considered were the relationships between the different changes and the changes in conformation of regions when antibody or ACE2 bound, as seen in three-dimensional structure determination deposited in the Protein Data Bank www.rcsb.org, and analyzed as to degrees of sidechain exposure and chain conformational flexibility [14]. For example, the following is a condensed form of one of the blocks of the RBD (described more extensively and in more detail later below). Blocks were as in the original paper [14] chosen to conveniently capture, and appropriately partition, important features, e.g., that regions around as well as in the receptor binding domain that are potentially affecting antibody binding are included in the blocks.

and the rest shown in Table 1 emerged as the significant changes from this study, and it is important to keep in mind that these do not relate to probability of accepted mutations over many proteins, and hence evolution's notion of physicochemical similarity [48], but to the consequences of selective pressure on SARS-COV-2 in favor of continued replication of the virus.

Although the scoring parameters in Table 1 are empirical, they appear to fit some rationales. They do *not* reflect the general trend in protein evolution to conserve residues with similar sidechain properties [48]. They reflect, in contrast, selective pressure to reject binding by antibodies formed against previous variants (encountered by infection or vaccination) while at the same seeking to strengthen ACE2 binding,

```
(1) GNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQ 506 (Wuhan strain)
(2) GNYNYLYRLFRKSNLKPFERDISTEIYQAGNKPCNGVAGFNCYFPLRSYSFRPTYGVGHQ omicron(21K)
(3) GNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCTPPA-LNCYWPLNDYGFYTTTGIGYQ 6ACC-B (SARS-COV-1)
(4) ############################################################# ACE2 binding loops
(5) @@@@@@@@@@                          @@@@@@@@@@@@@@@@          BD23-B binding loops
(6) ~~X95565~~~~~63854653788~~~~~~~~~~~~~~~~~~~~85555~~~~~~~~~~~~ 6ZP5-B closed
(7) 203723314337571643135456624334888385723661321112221146258463 7BYR-B BD23 Fab
(8) 225514111248761854555767924524878487867632251133120145027442 6M0J RBD+ACE2
(9)       Π                       χ+     ±        +  Π+  ϕ   +
```

The scoring is based on a classification that includes the conformational effects of antibody and ACE2 binding but is based on the classes of physicochemical change that became apparent in the sequence comparisons in the most preliminary studies. The two kinds of change are combined in the following ways. Above, rows (1)–(3) show the Wuhan reference sequence, omicron variant, and a SARS-COV-1 subsequence in the main part of the receptor binding domain RDB. (4) shows ACE2 receptor binding contacts, (5) shows antibody binding contacts. Rows (6)–(8) show numbers 0–9 and X (for 10) show the extent of being buried aways from solvent in the indicated experimental structures, with ~ indicating conformation disordered loop in the indicated three-dimensional structures. Row (9) indicates changes in the

but presumably only to limited degree since the coronavirus have had substantial past opportunities to refine ACE2 binding. An accepted mutation that involves a radical change in physical properties and occurs in binding loop binding antibody or ACE2 is likely to have more serious implications, since it suggests a strong selective pressure, irrespective of whether the loop is disordered prior to binding, noting that the conformation of the loop on binding is not in general similar in the antibody and ACE2 binding cases. If the mutation occurs just in an antibody binding loop or in a separate ACE2 binding loop, and the loop is conformationally disordered prior to binding, it appears that a radical mutation is more likely to be accommodated by significant conformational adjustments of the binding loop, so that the specific nature of the change in physicochemical properties of the sidechain is somewhat less important.

A fuller analysis of the whole receptor binding domain and neighboring sequence including the other strains is as follows.

```
BLOCK 10.
S1 RECEPTOR BINDING DOMAIN (RBD) 333-526,331-527
      (...
VRFPNItNLCPFGEVFnAtRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTK 386
VRFPNItNLCPFGEVFnAtRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTK alpha(20I)
VRFPNItNLCPFGEVFnAtRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTK beta(20H)
VRFPNItNLCPFGEVFnAtRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTK gamma(20J)
VRFPNItNLCPFGEVFnAtRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTK delta(21AIJ)
VRFPNItNLCPFDEVFnAtRFASVYAWNRKRISNCVADYSVLYNLAPFFTFKCYGVSPTK omicron(21K)
VRFPNItNLCPFGEVFnAtRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTK lamda(21G)
VRFPNItNLCPFGEVFnAtKFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTK mu(21H)
             -      χ                              ϕ ϕ ϕ
32135488837456436388444142154753482225446357576355361752136 5 6ZP5-B closed
4444%%555556%%55%%5444444444444444444444555%5%%554444444443 6ZP5-B closed sm.
32135488737446336388434132154753482225446357576355362752136 5 6ZP7-B open
4444%%555555%%5%%%4444444434444444444444555%5%%555444444443 6ZP7-B open sm.
53~~~~~~648534336698463175557775364636447458858846452374536 5 7BYR-A BD23 Fab
78~~~~~~987%%%5%%555555555555555555555555%%66666%555544544455 7BYR-A BD23 Fab sm.
31146776937446335386242033062673481225256356477256362452145 4 7BYR-B BD23 Fab
4444%555655%%%5%%4444333%33344444444444%%55%55%554444443343 7BYR-B BD23 Fab sm.
42349887736545446588442235255743593335466546688456563363367 5 7BYR-C BD23 Fab
55555666666%%%5%%554544444445555555555555%66666%6665555555555 7BYR-C BD23 Fab sm.
......X98254473364981220321646723825264464665873553735766468 6M0J RBD+ACE2
......99876%%%5%%5444344333344445555545555%%55%565555555655566 6M0J RBD+ACE2
.......X836459336497133043175782393526455445698254230323334 6 6XC3 RBD+CC12.1+CR3022
.......X987%%%5%%5444444444445555555454554%%55%%%4444333333444 6XC3 sm.
binding loop                              @@@@@@@@@@@@@@@@@@@@ 6XC3 CC12.1+CR3022

BLOCK 11.
LNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVG  446
LNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVG alpha(200I)
LNDLCFTNVYADSFVIRGDEVRQIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVG beta(20H)
LNDLCFTNVYADSFVIRGDEVRQIAPGQTGTIADYNYKLPDDFTGCVIAWNSNNLDSKVG gamma(20J)
LNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVG delta(21AIJ)
LNDLCFTNVYADSFVIRGDEVRQIAPGQTGNIADYNYKLPDDFTGCVIAWNSNKLDSKVS omicron(21K)
LNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVG lamda(21G)
LNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVG mu(21H)
                                 π                          +     χ
22522311031021438664363345567472146224237753112114847766~~~~ 6ZP5-B closed
33332222122233334444554545544444443444443333334334%456667~~~~ 6ZP5-B closed sm.
22522312031021438664363335567472146123237753212114847766~~~~ 6ZP7-B open
33332222122233334444554545544444443334443333334334%456667~~~~ 6ZP7-B open sm.
23675313141121121274474345767663238245577773211112502676 5~~~ 7BYR-A BD23 Fab
2333456789999999X98777777766654332223444444455555554444433~~~ 7BYR-A BD23 Fab sm.
~~~~226321040200100130623613335575610221232168152121023137 72 7BYR-B BD23 Fab
~~~~33332222111111112223333444433333332222333333332223333444 7BYR-B BD23 Fab sm.
25743313142121143266574545678765237232346772133325416866 4~~~ 7BYR-C BD23 Fab
44444333222222333444555666655555544444444444443344445678 9~~~ 7BYR-C BD23 Fab sm.
23766413140011014062383344568652032013228926212213302762168 7 6M0J RBD+ACE2
55544333321112222333344555544443332333434433332333344445554 6M0J RBD+ACE2 sm.
                                                    ######### ACE2 binding loops
35766323140101011052482333655242000123228925112113302872 17X~ 6XC3 RBD+CC12.1+CR3022
54444433221111112223334444433332221223333333333233333456778~ 6XC3 sm.
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@ binding loop   6XC3 CC12.1+CR3022binding

BLOCK 13-14 (includes interesting SARS-COV-1 features).
GNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQ 506
GNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTYGVGYQ alpha(20I.some E484K
GNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQ beta(20H)
GNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQ gamma(20J)
GNYNYRYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQ delta (20AIJ)
GNYNYLYRLFRKSNLKPFERDISTEIYQAGNKPCNGVAGFNCYFPLRSYSFRPTYGVGHQ omicron(21K)
```

The scoring proceeds for the example of omicron as follows, noting that one reason that the method is preliminary is that it depends on what constitutes an ACE2 or antibody binding loop and a given degree of sidechain exposure depends on the author's method of classification of these features. Nonetheless, it is precisely defined with respect to the classifications in the above blocks, and the rationale and methods are given in Ref. [14]. Another reason for the preliminary nature is that it is obvious that any scheme is likely to score it highly relative to the Wuhan strain simply because of the large number of accepted mutations, but while at time of writing it is suspected that omicron should rate lower because of a decrease in severity of symptoms, it is early days for assessing that for all groups such as diabetics, and in particular including

```
GNYNYQYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGSNCYFPLQSYGFQPTNGVGYQ lamda(21G)
GNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQ mu(21H)
     Π                          χ+     ±        +  Π +  ϕ   +
~~X95565~~~~~63854653788~~~~~~~~~~~~~~~~~~85555~~~~~~~~~~~~~ 6ZP5-B closed
~~676666~~~~~65%56%66666~~~~~~~~~~~~~~~~~~~~65554~~~~~~~~~~~~ 6ZP5-B closed sm.
~~X95465~~~~~63854537888~~~~~~~~~~~~~~~~~~~~85555~~~~~~~~~~~~ 6ZP7-B open
~~666666~~~~~65%55%66666~~~~~~~~~~~~~~~~~~~~55544~~~~~~~~~~~~ 6ZP7-B open sm.
~985553~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~X856672479~~8573 7BYR-A BD23 Fab
~332222~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~2223334445555555 7BYR-A BD23 Fab sm.
203723314337571643135456624334888385723661321111221146258463 7BYR-B BD23 Fab
3443333443%%%%%%44%4%%4%4555555655665554433222%12222334444443 7BYR-B BD23 Fab sm.
~~595653~~~~~838654847999~~~~~~~~~~~~~~~~9X9866476779~~~~9893 7BYR-C BD23 Fab
~~999999998777%8899XX9988~~~~~~~~~~~~~~~~XX9887778899~~~~9876 7BYR-C BD23 Fab sm.
@@@@@@@@@@                          @@@@@@@@@@@@@@@@           BD23-B binding loops
225514111248761854555767924524878487867632251133120145027442 6M0J RBD+ACE2
444333344344455565566555555656566677665544433222222333333332 6M0J RBD+ACE2 sm.
############################################################## ACE2 binding loops
~5825212347741X64665767932512798498887773236125514458138423 6XC3 RBD+CC12.1+CR3022
~766544434445556556666555555556667776666554443334344444434332 6XC3 sm.
       @@@@@                    @@@@@@@@@@@@@@@@@@@@      CC12.1+CR3022 binding loops
996824312549771954565767934546978497877962662276272748468473 6ZP5-A (SARS-COV-2)
877765555455556666666656666666667777776666555444555555555443 6ZP5-A (SARS-COV-2)
GNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCTP-PALNCYWPLNDYGFYTTTGIGYQ 6ACC-B (SARS-COV-1)
448755335446782753244567774369689494869948725652426695694830
555555555555555555555556667676677877766666655555555556555544


BLOCK 15.
Spike domain 1 C-terminal residual domain 527-588, 528-588
                    (...
PYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG 566
PYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG alpha(20I)
PYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG beta(20H)
PYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG gamma(20J)
PYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG delta(21AIJ)
PYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLKGTGVLTESNKKFLPFQQFG omicron(21K)
PYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG lamda(21G)
PYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG mu(21H)
                                        +
~~~211223457547330104695585567630422133554602137295615721440 6ZP5-B closed
~~~433333444444444444445566555444333333333434444444%44433444 6ZP5-B closed sm.
~~~211213457548330104695585567530422133544502137294615721450 6ZP7-B open
~~~333333444444444444445566555444333333333334444444%44444444 6ZP5-B open sm.
134112223269764520335896685547731525257555612247396615731650 7BYR-A BD23 Fab
433222334444444555555556666555544444444444444454455%55544444 7BYR-A BD23 Fab sm.
122011212345445330115575586567730313223434512137386814720130 7BYR-B BD23 Fab
332212222333333333334445566555544333333333334444%%%44433333 7BYR-B BD23 Fab sm.
022111223477X7964155..................................... 6M0J RBD+ACE2
22222234455656677888..................................... 6M0J RBD+ACE2 sm.
032111313658X8X641525X................................... 6XC3 RBD+CC12.1+CR3022
2222223445666666677889................................... 6XC3 sm.
          @@@@@   @@@@
```

. (*continued*).

any “long COVID” consequences. Nonetheless, this is likely to apply in the early days of any new variant arising for any pathogen, perhaps especially a virus. Working through the sequence, the first two mutations in the omicron strain are neither in ACE2 binding loops # nor in antibody loops @, and so score 1 each by Table 1. The possibility of some effect on binding ACE2 and/or antibodies cannot be disregarded. The nest three mutations in antibody binding loops @ are designated **φ** which means that they are changes from polar neutral or glycine residues to residues with hydrophobic (non-polar) sidechains, and score 9 each by the Table. A mutation designated **π** signifies a charged residue replaced by a polar neutral residue, and occurs in an antobody binding loop of the spike protein, and so scores 3. A mutation designated + neutral residue changing to a positive charged residue in an ACE loop, but not a conformationally disorganized loop, scores 8. A conservative mutation indicated by **χ** scores 2. The next four mutations **χ**, +, ±, and + all occur both in ACE2 and antibody binding loops of the spike protein and score highly at 10 each even though in conformationally flexible loops. The next three mutations +, **φ**, and + are in ACE2 binding loops that are disordered in the absence of binding, and score 7, 4, and 7 respectively. The last mutation is neither in an antibody nor ACE2 receptor binding loop so scores 1. Note that the last and 16th mutation lies outside the receptor binding domain and so is not one of the 15 mutations normally considered for omicron. However, its inclusion, as defined by the author’s block convention, and could be considered as having a potential effect. The total score for omicron is 102 ignoring this last residue and 103 including it, relative to the original Wuhan strain that scores 0 by definition.

4.4. Modeling the clinical effects of SARS-COV-2 variants

Also appropriate amongst what may be considered as results of the author’s COVID-19 project is a consideration of methods for predicting

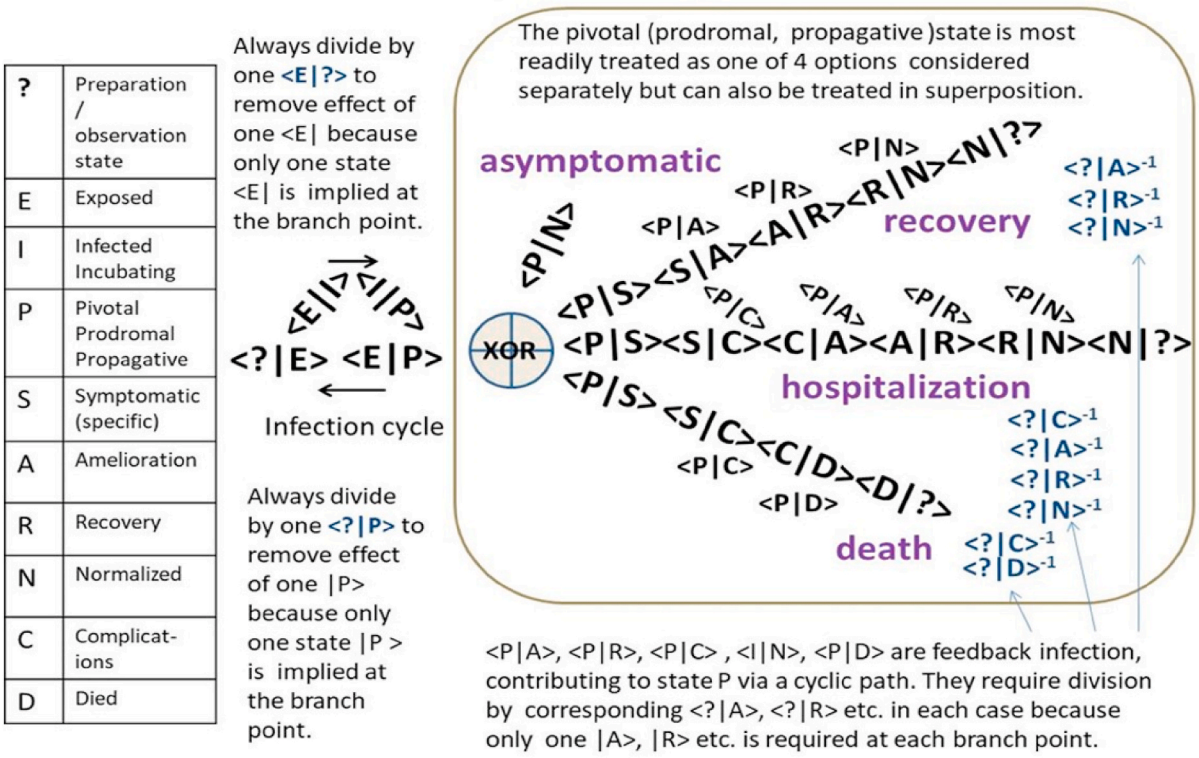


Fig. 2. A basic hyperbolic Dirac net for alternative futures analysis in the case of a patient exposed to a disease such as COVID-19.

the effects of variants of concern on the course of disease, as seen both from the perspective of individual patients and in terms of impact on the healthcare system as a whole. The regions that vary significantly in evolution of SARS-COV-2 are of interest to public health for three main reasons, (i) possible escape from vaccines and appropriate antiviral drugs, (ii) because of the effect that it has epidemiologically on the way the virus spreads, and (iii) because of the effect that the variation has clinically, regarding the development of the disease in infected persons. Such considerations require a further layer of modeling. To that end, the other result of the study associated with the present paper and not previously described, was the use of Q-UEL and the Hyperbolic Dirac Net (HDN) approach [30,36,42–44] approach to study the stages of the epidemic for a new variant, and the course of the disease in patients. Such graphs can be used in the manner of an epidemiologist's *chain rule*, i.e., the mortality rate for a population given prevalence and probability of exposure, probability of infection given exposure, and so on including conditional probabilities of symptoms, complications, and death.

See Fig. 2, which reflects the probabilities and hence Q-UEL tags that are required as described in Ref. [30], to make maximum use of information by minimizing independency assumptions, and to avoid counting the same information more than once. Such an HDN of the simplest type is essentially a Bayes Net making use of Q-UEL algebra to construct a probabilistic knowledge graph that is a Bidirectional General Graph (BGG), i.e., bidirectional, and optionally including cyclic paths that can be solved without iteration [30]. A Bayes Net is, in contrast, a more restricted Directed Acyclic Graph (DAG) by definition. Such approaches, including Bayes Nets, imply use of logical AND between probabilities and the assumption of certain independencies, although in this study the method was extended to include logical OR to enable "Alternative Futures Analysis", see Fig. 2, that can be applied both to "what if' studies in responses to epidemics and the different possible outcomes for a patient exposed and then infected by COVID-19. This kind of inference net approach does depend on having epidemiological data for a new variant or to emerging diseases other than COVID diseases because such data is required to parametrize the required probabilities, although the above scoring measures for VOCs can be used to predict probabilities that can be tested in "what if" computer experiments. Note that decision trees, clinical pathways, and most epidemiological graphs usually indicate a flow from left to right, and so are flipped around compared with Bayes Nets and HDNs that follow the conditional probability notation P(A|B) = P("A←B") [30]. The convention suitable for the former is used in Fig. 2. Fortunately, the use of dual probabilities [30] makes this change in convention simple, though the one used should always be clearly stated. Mathematically, it is a matter of a sign convention: the choice here is equivelent to saying that we follow Eqn. (1) but take the complex conjugate * of the braket <A|B> = <B|A>*, i.e. change the sign of the imaginary part, but omit all the *, taking them as understood, for brevity.

Such HDNs were initially calibrated for COVID-17 B.1.1.7 and related early variants and then adjusted for omicron discussed in the previous Section 4.3 though it did not include the new variants related to the omicron that arose while writing this paper, and for which details as to clinical effect were in some cases relatively sparse. For example, in response to the queries, this following tag was one of many tags for a variety of European counties that had at some stage 50% or more of B.1.1.7 amongst tested people, and also had association constants with the conditions of 3 or more factors.

< Q-UEL-COVID19-BY-COUNTRY 'variant':='eqB.1.1.7' Pfwd:=0.8896 Ofwd:=3.9606 Efwd:=0.6371 | **if**:=(assoc:=3.4616, count:=7, factors:=(2,5)) | 'percent variant':='ge50' 'percent cases sequenced':='0.2' 'valid denominator':='Yes' **with** 'percent cases sequenced':='0.2' 'source':='TESSy' 'country':='Romania' Pbwd:=0.006057 Obwd:=11.4507 Ebwd:=1.8420 Q-UEL-COVID19-BY-COUNTRY>

Probabilities used in this study are purely examples at this stage and represented an amalgam of information from various sources. Since shortly before submission BA.2 extensively replaced BA.1 omicron such that the probabilities appropriate to BA.1 are likely to be obsolete and misleading at the time of reading this, and the probability values discussed later below relate to data for the second peak of September 24, 2020 to March 28, 2021 for which more data is available. One reason for caution is that the Q-UEL knowledge-gathering techniques were valuable but also illustrated that it is dangerous to extract statistics of a previous wave to predict the statistics at the very start of a wave due to a new variant. Recall that COVID-19 alpha was becoming dominant around the beginning of January 2021, delta around May 2021, and omicron in early January 2022. Typically, pathogens are assumed to become milder in time due to natural selective pressure to survive better, and due to the deaths of less susceptible hosts over several generations. Also, growing experience in dealing with the disease shifts the spectrum of severity in the direction of less severe outcomes. Omicron patients had a 53% reduced risk of hospitalization and a 91% reduced risk of death compared with patients who had the delta variant (though of course the population *mortality rate* will increase if the incidence and prevalence of a wave is much higher than a previous wave). It was tempting that the early statistics for in-hospital fatalities from severe symptoms of COVID-19 in the first wave of early 2020 be considered as indicative of "serious covid" in the second, primarily alpha, wave. So, for example, knowledge capture indicated that in March 1 and May 11, 2020, the probability of dying of serious (in-hospital) COVID-19 if aged 50–59 was about 0.055 (but 0.135 if the patient had type 1 diabetes) in the first, which became the probability of serious symptoms for the second wave, and 0.25 for patients aged 70–79 (but 0.27 if the patient had type 1 diabetes), which similarly became the probabilities of serious COVID-19 in the following wave. The following is from semi-structured data in tabular form captured from a web page in the form of an XTRACT tag but simplified to a CTRACT tag, implying a degree of automated curation.

<Q-UEL-CTRACT 'Severe disease':='Y' Pfwd:=0.108 | **if** | Infection:=COVID-19 'Based on':=First wave death':= 'Y'
'type 2 diabetes':='Y' 'prior cerebrovascular disease':='Y' Pbwd:=0.021 Q-UEL-CTRACT>

However, in the early stages of the alpha B.1.1.7 variant the collected knowledge suggested it significantly *increased* the risk of hospitalization and the fatality-rate for patients aged 70 or less, but decreased the fatality rate for older patients. Consequently, for variants arising at this time and in the future, Fig. 2 is to be considered as a *template* irrespective of specific probabilities that should to be introduced as required.

Some methods and examples for the use of this template are as follows. The principles of *coherence* (mutual consistency of probabilities) that should be considered are discussed in Ref. [30], Should this template need to be varied, Ref [30] also gives a step-by-step account of manual construction of small inference nets, though semi-automated [24] and automated [29] methods are usually used, except when Q-UEL is used in a programming language mode [36]. That is to say, except for an Expert System approach in which the human expert user enters the probabilities [36], but all still first require datamining of epidemiological sources to generate the tags with the required probabilities (and association constants). In Fig. 2, the state '?' is a state of observation or preparation of a probability, such that $P(?) = 1$, and a tag like $<A|?>$ is analogous to a self or prior probability in a Bayes' Net. Coherence means establishing other reasonable values under the constraint that the probabilities used must be such that Bayes' Rule and normalization and marginal summation are satisfied, and that the sum of all paths from origin to terminal nodes to the right, is probability 1.0. By such means it is possible to fill many gaps in probability assignements and fully quantify such a graph, but it varies from country to country, variant to variant, and not least from patient type to patient according the conditioning factors such as ethnic and socioeconomic group, and also not least by genomic "molecular ethnicity", in ways that are as yet to be fully understood. Useful sources of data hit upon by the knowledge gathering methods include the European Surveillance System (TESSy) and GISAID database of the WHO Global Influenza Surveillance and Responses System (GISRS). The available data is in this case seen for the most part as clean by being available in several well-defined tabular formats, and in having no unknowns or ambiguities.

For example, in the first calibration of the template, knowledge captured by the Q-UEL system indicated that among US counties with populations greater than 500,000 people, during the week ending June 13, 2020, the median estimate of the county level probability of a confirmed infection was 1 infection in 40,500 person contacts. Using the knowledge gathering techniques, it was found that if each person interacts with 50 people a day face to face contact, plus e.g., supermarket exposure. It assumed 2 weeks to show infection, the probability is of the order (1/40,500) x 50 $\times$ 14 = 0.0346. For COVID-19, data to that date suggest that 80% of infections are mild or asymptomatic, 15% are severe infection, requiring oxygen and 5% are critical infections, requiring ventilation. It was also found that a probability of 0.021 was reported for serious, hospitalized, discharged out of population. Probability 0.00071 was reported as prevalence of complications, 0.00048 was reported as cause specific mortality rate, and 0.062 was reported as case specific fatality rate.

5. Conclusions

The above paper sought to illustrate ways in which computers and the Internet can help combat emerging disease and described as "Results" some preliminary methods indicating directions in which computational tools might be further developed to meet that challenge. Such studies are still incomplete, and efforts by many workers will doubtless continue to be developed for several years, improved and fine-honed by experience of their use in meeting hitherto unnoticed species or variants of pathogens. The approach by the present author is ultimately rooted in some mathematics that is not widely known, but there is as yet nothing to say that it is not a valid candidate and insightful example for providing the general kind of tools required. As stated in Theory Section 2.1, the focus is on what such tools need to do. The approach described above is to be seen only as an example of a way to achieve that.

In developing a design approach based on bioinformatics, the appropriateness for practical development of diagnostics, vaccines, and peptidomimetics is constantly to be kept in mind. In this paper, there has been some large degree of emphasis on techniques most relevant to making use of peptide synthetic chemistry and laboratory immunology. This is simply because the focus of the initial papers was to some extent with peptide-based vaccines in mind, or somewhat similarly, epitopes inserted as loops into cloned proteins. It was the peptide approach that seemed the most modern, and that had been successful in other cases in the hands of the author and collaborators, as well as in the laboratories of many other workers and in veterinary medicine. However, these peptide-centric considerations are by no means outmoded, even in the light of RNA and DNA vaccines. Such peptide-based tools still have the benefit of focusing on and using only the parts that matter for the effect desired and run less risk of side-affects due to the presence of unnecessary material, such as hemagglutination or autoimmune responses in

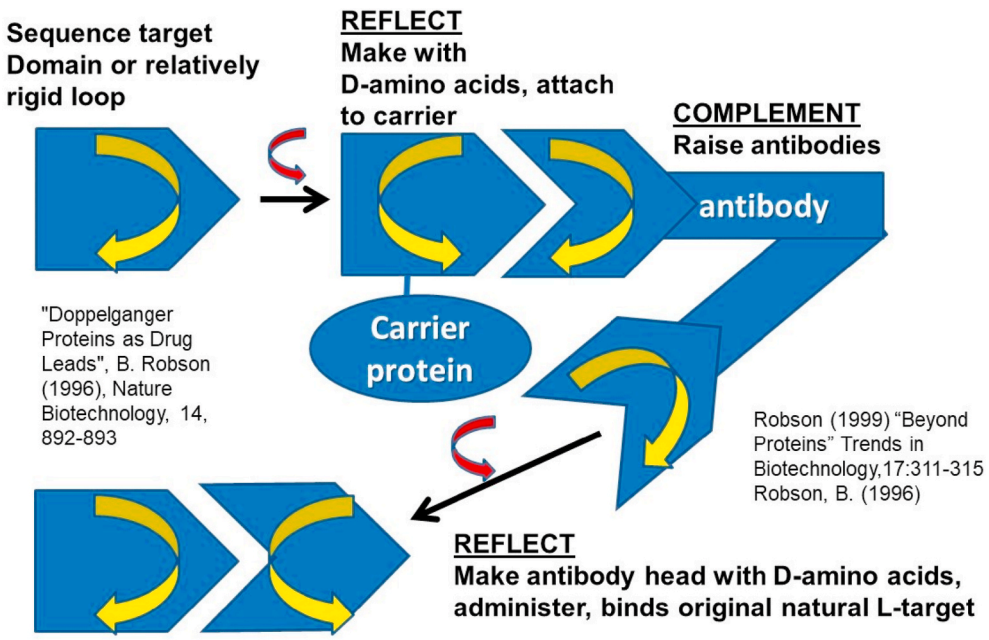


Fig. 3. The reflect-complement-reflect method.

some patients. There is a considerable body of emerging literature on the risks of RNA and DNA which will be analyzed elsewhere: much seems alarmist, and some have been retracted. In view of the obvious life-saving success of RNA and DNA vaccines, it might seem ungracious to consider these concerns. However, one cannot easily put aside that in scale, much larger than peptide methods and commensurate with killed or attenuated viruses as vaccines, the RNA and DNA vaccines are still big constructs, that act on complex systems inside human cells, and contain features that may not always be fully understood.

There is even room for a more fundamental development of the peptide-based approach. One that could bring the response of bio-nanotechnology to the development of novel protein-like compounds interacting with target proteins [49] is shown in Fig. 3, reproduced from the sister journal [14], and using the technology described by the present author and colleagues in Refs. [50–54]. In essence, this *reflect--complement-reflect method* requires synthesizing a viral or host protein or protein domain target using D-amino acids (the first reflect step), attaching that to a L-amino acid protein carrier to raise antibodies (the complement or *wet-lab-fit-to-binding-site* step), and synthesizing nanobodies (here meaning antibody heads) out of D-amino acids (the final reflect step) to interact with the original protein target. Peptides and proteins made entirely from D-amino acid residues fold up in space and function in mirror image to their L-amino acid counterparts, but the resulting complex structure cannot be considered biological, and is not subject to rapid proteolysis in the patient (though ultimately degraded intracellularly).

This present paper also touched upon, and illustrated, the more general kinds of development by the author and collaborators in the areas on knowledge management and AI which are in progress with the author and collaborators, illustrated by describing their application in the COVID-19 pandemic. In closing summary, the major finding has been, not unexpectedly, that access to fullest possible knowledge of emerging and previous epidemics, gathered from different places and different times, is extremely useful in rapid defense against emerging disease. It is also important to update this dynamically, as real-world data in real time. However, integration of knowledge from diverse sources has not been a strong feature of response to emerging epidemics in the past, and efforts like these described here are required. Some would argue that the world's response could have been faster in the case of COVID-19 [6], and if so, we should learn from it. Coronaviruses may not be the pathogens involved in next pandemic. Concerns arise constantly. For example, in preparing an early draft version of the present paper, a warning sign for avian influenza in Barkby, Leicestershire, UK, was photographed on Sunday, December 12, 2021, e.g. Ref [55], apparently the result of transmission from chickens to a single farmer, that could involve a new strain. The patient was isolated, the WHO informed, and doubtless the viral RNA has already been sequenced. By late February 2022 there has been no further news but should this event have emerged as the seed for a new epidemic, an official, clear, well-advertised reference to finding the sequence quickly on GenBank and gathering all relevant knowledge would have been the important first steps to realize and facilitate the kind of approaches described here. At the time of submitting this final version, there is also an emergence of what appears to be a new form of viral hepatitis, reversing the COVID-19 story by focusing on children as the most affected part of the population. And then yet again, at the time of doing the galley proofs of this paper, there is the rise of "monkey pox" in the human population. There seems to be a relentless progression of potential new pandemics that makes the previous decades almost seem like a lull. But throughout, the argument remains the same: gathering and bringing together new knowledge from all sources to tackle new emergent diseases remains imperative.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Barry Robson reports a relationship with Ingine Inc. that includes: equity or stocks. The author is an Assistant Editor of this journal - Barry Robson.

Acknowledgement

The author acknowledges the support of Ingine Inc. and collaborators in the cited papers who have contributed to some of the ideas expressed here. This paper is a voluntary work without pay in the interests of battling the COV-19 pandemic, and is not related to any grant, hire, or contractual arrangement.

References

[1] Garrett L. The coming plague. Newly emerging diseases in a world out of balance. Farrar, Straus and Giroux; 1994.

[2] Tyrrell DA, Bynoe ML. Cultivation of viruses from a high proportion of patients with colds. Lancet 1966;1:76–7. https://doi.org/10.1016/s0140-6736(66)92364-6.

[3] Hamre D, Procknow JJ. A new virus isolated from the human respiratory tract. Proc Soc Exp Biol Med 1966;121:190–3. https://doi.org/10.3181/00379727-121-30734.

[4] Masters PS. The molecular biology of coronaviruses. Adv Virus Res 2006;66:193–292. https://doi.org/10.1016/S0065-3527(06)66005-3.

[5] https://www.gov.uk/government/publications/wuhan-novel-coronavirus-background-information/wuhan-novel-coronavirus-epidemiology-virology-and-clinical-features. [Accessed 29 December 2021].

[6] https://www.bbc.co.uk/news/world-55756452. [Accessed 1 January 2022]. Accessed.

[7] Robson B. Preliminary bioinformatics studies on the design of synthetic vaccines and preventative peptidomimetic Antagonists against the Wuhan seafood market coronavirus. Possible importance of the KRSFIEDLLFNKV motif, Epub 30th January on ResearchGate 2020; https://doi.org/10.13140/RG.2.2.18275.09761.

[8] Robson B. Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus. Comput Biol Med 2020:103670. https://doi.org/10.1016/j.compbiomed.2020.103670. Epub 2020 Feb 26.

[9] Robson B. COVID-19 coronavirus spike protein analysis for synthetic vaccines, a peptidomimetic antagonist, and therapeutic drugs, and analysis of a proposed Achilles' heel conserved region to minimize probability of escape mutations and drug resistance. Comput Biol Med 2020;121:103749. https://doi.org/10.1016/j.compbiomed.2020.103749. Epub 2020 Apr 11.

[10] Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579(7798):270–3. https://doi.org/10.1038/s41586-020-2012-7. Epub 2020 Feb 3.

[11] Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan J, Xie Z, Ma J, Liu WJ, Wang D, Xu W, Holmes EC, Gao GF, Wu G, Chen W, Shi W, Tan W. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Feb 22 Lancet 2020;395(10224):565–74. https://doi.org/10.1016/S0140-6736(20)30251-8. Epub 2020 Jan 30.

[12] Robson B. Bioinformatics studies on a function of the SARS-CoV-2 spike glycoprotein as the binding of host sialic acid glycans. Comput Biol Med 2020;122:103849. https://doi.org/10.1016/j.compbiomed.2020.103849. Epub 2020 Jun 8.

[13] Robson B. The use of knowledge management tools in viroinformatics. Example study of a highly conserved sequence motif in Nsp3 of SARS-CoV-2 as a therapeutic target. Comput Biol Med 2020:125. https://doi.org/10.1016/j.compbiomed.2020.103963. Epub 2020 Aug.

[14] Robson B. Techniques assisting peptide vaccine and peptidomimetic design. Sidechain exposure in the SARS-CoV-2 spike glycoprotein. Comput Biol Med 2021;128:104124. https://doi.org/10.1016/j.compbiomed.2020.104124. Epub 2020 Nov 21.

[15] https://www.pharmaceutical-technology.com/comment/covid-19-vaccine-availability/. [Accessed 2 January 2022].

[16] Gallagher J. BBC News, 23 November, Oxford vaccine: how did they make it so quickly. https://www.bbc.co.uk/news/health-55041371; 2020.

[17] Robson B. From Zika to flu and back again. CAVIRC (Caribbean anti-virus informatics research center). https://doi.org/10.13140/RG.2.1.5000.6808; 2016.

[18] Robson B, Fishleigh RV, Morrison CA. Prediction of HIV vaccine. Nature 1987;4:325–95. https://doi.org/10.1038/325395a0.

[19] Fishleigh RV, Robson B, Mee RP. Fragments of prion proteins, US Patent 5,773,572,.

[20] Robson B, Caruso TP, Balis UGJ. Considerations , for a universal Exchange Language for healthcare. In: Proceedings of 2011 IEEE 13th international conference on e-health networking, applications and Services. Healthcom); 2011. p. 173–6.

[21] Robson B, Caruso TP, Balis UGJ. Suggestions for a web based universal exchange and inference language for medicine. 1 Comput Biol Med 2013;43(12):2297–310. https://doi.org/10.1016/j.compbiomed.2013.09.010. Epub 2013 Sep. 20.

[22] Robson B, Caruso TP. A universal exchange language for healthcare. In: Lehmann CU, Ammenwerth E, Nohr C, editors. Stud health technol inform. vol. 192. IOS Press; 2013. p. 949.

[23] Robson B, Caruso TP, Balis UGJ. Suggestions for a web based universal exchange and inference language for medicine. Continuity of patient care with PCAST disaggregation. Comput Biol Med 2014;56:51–66.

[24] Robson B, Boray S. Implementation of a web based universal exchange and inference language for medicine. Sparse data, probabilities, and inference in data mining of clinical data repositories. Comput Biol Med 2015;66:82–102. https://doi.org/10.1016/j.compbiomed.2014.10.022. Epub 2014 Nov 4.

[25] Robson, B, Boray S. Interesting things for computer systems to do: keeping and data mining millions of patient records, guiding patients and physicians, and passing medical licensing exams, Bioinformatics and Biomedicine (BIBM), Proceedings 2015 IEEE International Conference. 21015l; 1397-1404.

[26] Robson B. Boray, data-mining to build a knowledge representation store for clinical decision support. Studies on curation and validation based on machine performance in multiple choice medical licensing examinations. Comput Biol Med 2015;73:71–93. https://doi.org/10.1016/j.compbiomed.2016.02.010. Epub 2016 Feb 26.

[27] Robson B. Studies in using a universal exchange and inference language for evidence based medicine. Semi-automated learning and reasoning for PICO methodology, systematic review, and environmental epidemiology. Comput Biol Med 2016;79:299–323. https://doi.org/10.1016/j.compbiomed.2016.10.009. Epub 2016 Oct 17.

[28] Robson B, Boray S. Studies of the role of a smart web for precision medicine supported by biobanking. Per Med 2016;13(4):361–80. https://doi.org/10.2217/pme-2015-0012. Epub 2016 Jul 5.

[29] Robson B, Boray S. Studies in the extensively automatic construction of large odds-based inference networks from structured data. Examples from medical, bioinformatics, and health insurance claims data. Comput Biol Med 2018;95:147–66. https://doi.org/10.1016/j.compbiomed.2018.02.013. Epub 2018 Mar 21.

[30] Robson B. Bidirectional General Graphs for inference. Principles and implications for medicine. Comput Biol Med 2019;10:382–99. https://doi.org/10.1016/j.compbiomed.2019.04.005. Epub 2019 Apr 13.

[31] Robson B, Boray S, Weisman J. Mining real-world high dimensional structured data in medicine and its use in decision support. Some different perspectives on unknowns, interdependency, and distinguishability. Comput Biol Med 2022;141:105118. https://doi.org/10.1016/j.compbiomed.2021.105118.

[32] Robson B. Extension of the Quantum Universal Exchange Language to precision medicine and drug lead discovery. Preliminary example studies using the mitochondrial genome. Comput Biol Med 2020 Feb;117:103621. https://doi.org/10.1016/j.compbiomed.2020.103621. Epub 2020 Jan 20.

[33] Robson B. Computers and preventative diagnosis. A survey with bioinformatics examples of mitochondrial small open reading frame peptides as portents of a new generation of powerful biomarkers. Comput Biol Med 2021;140:105116. https://doi.org/10.1016/j.compbiomed.2021.105116.

[34] Robson B. De novo protein folding on computers. Benefits and challenges. Comput Biol Med 2022;143:105292. https://doi.org/10.1016/j.compbiomed.2022.105292.

[35] Robson B. Towards new tools for pharmacoepidemiology. Adv Pharmacoepidemiol Drug Saf 2013;1:6. https://doi.org/10.4172/2167-1052.100012.

[36] Robson B. POPPER, a simple programming language for probabilistic semantic inference in medicine. Comput Biol Med 2015;56:107–23. https://doi.org/10.1016/j.compbiomed.2014.10.011. Epub 2014 Nov.

[37] Robson B, Boray S. Studies in the use of data mining, prediction algorithms, and a universal exchange and inference language in the analysis of socioeconomic health data. Comput Biol Med 2019;112:103369. https://doi.org/10.1016/j.compbiomed.2019.103369. Epub 2019 Jul 25.

[38] Xia S, Liu M, Wang C, Xu W, Lan Q, Feng S, Qi F, Bao L, Du L, Liu S, Qin C, Sun F, Shi Z, Zhu Y, Jiang S, Lu L. Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. Cell Res 2020;30(4):343–55. https://doi.org/10.1038/s41422-020-0305-x. Epub 2020 Mar 30.

[39] Ahsan H, Sonn JK, Lee YS, Islam SU, Khalil SK. Keeping SARS-CoV-2 reinfections at Bay, genetic engineering and biotechnology news. https://www.genengnews.com/immunology/keeping-sars-cov-2-reinfections-at-bay/. [Accessed 14 December 2021].

[40] Ahsan H, Sonn JK, Lee YS, Islam SU, Khalil SK. An overview about the role of adaptive immunity in keeping SARS-CoV-2 reinfections at Bay. Viral Immunol 2021;34(9):588–96. https://doi.org/10.1089/vim.2021.0017. Epub 2021 Jun 8.

[41] Castelletto V, Hamley IW. Amyloid and hydrogel formation of a peptide sequence from a coronavirus spike protein. American Chemical Society Nano; 2022. https://pubs.acs.org/doi/10.1021/acsnano.1c10658.

[42] Robson B. The new physician as unwitting quantum mechanic: is adapting Dirac's inference system best practice for personalized medicine, genomics, and proteomics? J Proteome Res 2007;6(8):3114–26. https://doi.org/10.1021/pr070098h. Epub 2007 Jul 3.

[43] Robson B. Hyperbolic Dirac Nets for medical decision support. Theory, methods, and comparison with Bayes Nets. Comput Biol Med 2014;51:183–97. https://doi.org/10.1016/j.compbiomed.2014.03.014.

[44] Deckelman S, Robson B. Split-complex numbers and Dirac bra-kets. Commun Inf Syst 2015;14(3):135–49. https://www.academia.edu/33231629/Split-complex_numbers_and_Dirac_i_bra-kets_i_.

[45] Robson B. Testing machine learning techniques for general application by using protein secondary structure prediction. A brief survey with studies of pitfalls and benefits using a simple progressive learning approach. Comput Biol Med 2021;138:104883. https://doi.org/10.1016/j.compbiomed.2021.104883. Epub 2021 Sep. 23.

[46] SARS-CoV-2 and ACE2 receptor: combination of molecular dynamics simulation and density functional calculation. J Chem Inf Model 2021;61(9):4425–41. https://doi.org/10.1021/acs.jcim.1c00560. Epub 2021 Aug 24.

[47] Chen C, Boorla VS, Banerjee D, Chowdhury R, Cavener VS, Nissly RH, Gontu A, Boyle NR, Vandegrift K, Nair MS, Kuchipudi SV, Maranas CD. Computational prediction of the effect of amino acid changes on the binding affinity between SARS-CoV-2 spike RBD and human ACE2. Proc Natl Acad Sci U S A 2021;118(42):e2106480118. https://doi.org/10.1073/pnas.2106480118.

[48] French S, Robson B. What is a conservative substitution? J Mol Evol 1983;19:171–5. https://doi.org/10.1007/BF02300754.

[49] Robson B. Doppelgänger proteins as drug leads. Nat Biotechnol 1996;14(7):892–3. https://doi.org/10.1038/nbt0796-892.

[50] Canne LE, Figliozzi GM, Robson B, Siani MA, Simon RJ. N-Alkoxy amid backbone protection in BOC chemistry : improved synthesis of a 'difficult sequence. Protein Sci 1996;5(1):72. 1996.
[51] Siani MA, Canne LE, Simon RJ, Figliozzi GM, Robson B, Thompson DA, Simon RJ. Chemical synthesis and activity of D-superoxide dismutase. Protein Sci 1996;5 (Suppl.1):72.
[52] Siani MA, Canne LE, Figliozzi GM, Robson B. Total chemical synthesis of proteins including chemokines and their analogues. Chemokines: International Business Communications; 1997.
[53] Robson B. Pseudoproteins: non-protein protein-like machines. http://www.foresight.org/Conferences/MNT6/Abstracts/Robson/index.html; 1999.
[54] Robson B. Beyond proteins. Trends Biotechnol 1999;17(8):311–5. https://doi.org/10.1016/s0167-7799(99)01339-6.
[55] https://www.aol.co.uk/news/human-case-bird-flu-detected-141440541.html.