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.

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

B. Robson

PII: S0010-4825(20)30128-1

DOI: https://doi.org/10.1016/j.compbiomed.2020.103749

Reference: CBM 103749

To appear in: *Computers in Biology and Medicine*

Received Date: 15 March 2020

Revised Date: 3 April 2020

Accepted Date: 3 April 2020

Please cite this article as: B. Robson, 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, *Computers in Biology and Medicine* (2020), doi: https://doi.org/10.1016/j.compbiomed.2020.103749.

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.

B. Robson
Ingine Inc. Cleveland Ohio USA,
and The Dirac Foundation, Oxfordshire UK

This paper continues a recent study of the spike protein sequence of the COVID-19 virus (SARS-CoV-2). It is also in part an introductory review to relevant computational techniques for tackling viral threats, using COVID-19 as an example. Q-UEL tools for facilitating access to knowledge and bioinformatics tools were again used for efficiency, but the focus in this paper is even more on the virus. Subsequence KRSFIEDLLFNKV of the S2' spike glycoprotein proteolytic cleavage site continues to appear important. Here it is shown to be recognizable in the common cold coronaviruses, avian coronaviruses and possibly as traces in the nidoviruses of reptiles and fish. Its function or functions thus seem important to the coronaviruses. It might represent SARS-CoV-2 Achilles' Heel, less likely to acquire resistance by mutation, as has happened in some early SARS vaccine studies discussed in the previous paper. Preliminary conformational analysis of the receptor (ACE2) binding site of the spike protein is carried suggesting that while it is somewhat conserved, it appears to be more variable than KRSFIEDLLFNKV. However compounds like emodin that inhibit SARS entry, apparently by binding ACE2, might also have functions at several different human protein binding studies. The enzyme 11β-hydroxysteroid dehydrogenase type 1 is again argued to be a convenient model pharmacophore perhaps representing an ensemble of targets, and it is noted that it occurs both in lung and alimentary tract. Perhaps it benefits the virus to block an inflammatory response by inhibiting the dehydrogenase, but a fairly complex web involves several possible targets.

Keywords: coronavirus; 2019-nCoV; SARS-CoV-2; COVID-19, Wuhan Seafood Market Coronavirus; Bionformatics; Synthetic Vaccine; Peptidomimetic; Retroinverso; Q-UEL language;

1. Introduction and Review.

1.1. Background.

Coronaviruses have been known to medicine for some time [1], but it is of course only very recently that the coronavirus SARS-CoV-2, the COVID-19 virus new and dangerous to humans, was identified. It is believed to be related to an initial cluster of pneumonia cases associated with a seafood and fresh meat market in Wuhan, China, [2]. Current case rates at the time of writing are close to one million with close to 60,000 deaths. The genomic relationships to other coronaviruses were quickly examined by Lu et al. to shed light on the origins, epidemiology, and receptor binding of the virus [2]. On January 17th, 2019, the Wuhan isolate Genbank entry MN908947.3 replaced MN908947.2, and MN908947.3 probably represents an adequate stable description of the sequence for research into that strain isolate, and was immediately investigated by the present author [3, 4]. Originally, it was seen by authorities as a coronavirus outbreak but not as SARS (Severe Acute Respiratory Syndrome). However, its genomic relationships examined in refs [3, 4] also showed many fairly close correlations with the genomes of SARS-CoV in the previous human (but not pandemic) outbreaks and in pigs, bats and civets, and the emphasis was on finding subsequences that are well conserved across coronavirus strains and species. The earliest patients suffering from what is now called COVID-19 had overall 99·98% genome sequence identity to the above Wuhan isolate, so that one may reasonably say that it is the origin of COVID-19, and its virus SARS-Cov-2 [2]. The earlier Wuhan isolates also related (with 88% identity) to two bat-derived severe acute respiratory syndrome (SARS)-like coronaviruses collected in 2018 in Zhoushan, China, but differed more from SARS-CoV (at about 79%) and MERS-CoV (at about 50%) [2]. The Wuhan and related isolates revealed a coronavirus that resides in the subgenus Sarbecovirus of the genus Betacoronavirus [2], and although genetically distinct from its predecessor SARS-CoV it appeared to have similar external binding proteins, meaning here the spike glycoprotein discussed extensively in the present paper. See Section 1.3 below for introduction to this protein, which also discusses some further early identified genomic correlations. In addition, the rest of this present paper discusses many other genomic relationships relevant to the design of synthetic vaccines and therapeutic antagonists againstCOVID-19.

One problem is that COVID-19 is a new pathogen posing a global threat and so presents new challenges both in primary prevention, where a vaccine is required, and in secondary prevention, where a therapeutic compound (ideally, “in a pill”) is required to treat patients who are infected. It might also present challenges for tertiary prevention, which seeks to remedy a persistent level of infection, or to prevent recurrence even to essentially the same strain, as discussed in Section 1.2. A main problem of concern, and a point of the present paper, is the likely appearance of new strains with resistance to vaccines and therapeutic agents. At the time of writing, confirmed cases double globally every 6 days, and undetected cases are expected to be much higher (the current plateauing of reported cases in China offers a glimpse that this this should

attenuate soon, but estimates of how and when are varied). With a significant portion of humanity already infected, there is enhanced probability of successful “escape mutations” in the genome of the virus. Development of vaccines and perhaps particularly therapeutics that could, but do not, take account of this by targeting less variable protein regions could be a huge waste of resource and a dangerous delay.

COVID-19 is, of course, by far the most serious, but not the first SARS outbreak of concern to humans, and coronaviruses have for decades been of veterinary concern [1]. However, it still remains true that zoonotic coronaviruses have only rather recently seriously impacted humans, as far as is known. They include SARS-CoV (2002, *Betacoronavirus*, subgenus *Sarbecovirus*), and MERS-CoV (2012, *Betacoronavirus*, subgenus *Merbecovirus*). Although the idea that SARS-CoV-2 was distinct from SARS-CoV was originally discouraged, distinction is here a matter of degree. By usual criteria they are fairly closely related, genetically clustering within *Betacoronavirus* subgenus *Sarbecovirus.* Until very recently, SARS-CoV, effectively SARS-CoV-1, was the primary reference point and model regarding molecular and functional details, and it remains important.

Shortly after the appearance in GenBank of the apparently final version of the Wuhan seafood market isolate MN908947.3 [2], the present author compared a variety of coronavirus genomes [3, 4]. The KRSFIEDLLFNKV protein subsequence was seen as a potential Achilles’ heel because it is exposed or potentially exposable, being required for proteolytic activation cleavage, and importantly is also a well-conserved feature on the surface of the virus [3,4]. Being well conserved suggests that mutations are much less easily “accepted”, meaning that the virus is less likely to survive more than one or two generations. As discussed below, the conservation is in a region of protein on the virus surface concerned with at least one step of lung cell entry, interesting because coronaviruses seem to be able readily adjust to alternative means of entry, possibly hinting at additional roles for the subsequence. Whether or not that is the case, the above motif seems a likely primary target for synthetic vaccines and a basis for drug discovery, and was proposed as such [3, 4]. It is a motif that was found to be quite well conserved even in more distantly related coronaviruses [3, 4], and the present paper also explores how far that seems to extend. It includes the common cold coronaviruses. Another potential subsequence of interest popular with researchers is also examined (the ACE2 binding domain discussed below), but the above remains popular with the present author because of its relatively high degree of conservation.

1.2. Implication of the Common Cold?

At first glance, of the three kinds of prevention, tertiary prevention, i.e. including trying to insure that the disease does not recur, seems the least worrying. In the

present authors' opinion, however, it relates to a specter that recently haunted COVID-19 vaccine research, and which might still cause some concerns. This is the question of why there is no significant immunity acquired by the body to prevent recurrence of common cold, of which up to roughly 30% of cases are believed to be due to coronaviruses. Fortunately, at time of final writing of this paper, news reports indicate that neutralizing antibodies can be found in patients who have had COVID-19. However, with the risks of escape mutations of the virus in mind, it remains worthwhile considering whether the subsequence KRSFIEDLLFNKV, again, found to be well conserved [3, 4] across many coronaviruses [3, 4], is still present in common cold coronavirus. This is in order to force better immune response by targeting using synthetic or cloned vaccines with this epitope. Most common cold strains fall into one of two coronavirus serotypes: OC43-like and 229E-like, which are the main examples discussed below. While the common cold is generally considered as mainly an upper respiratory tract infection and a mundane inconvenience, common human coronaviruses *Betacoronavirus* HCoV-OC43 and HCoV-HKU1, as well as *Alphacoronavirus* HCoV-229E, also cause severe lower respiratory tract infections in children and the elderly. Some discussion is also given to HCoV-HKU1 in this paper.

1.3. The Spike Glycoprotein.

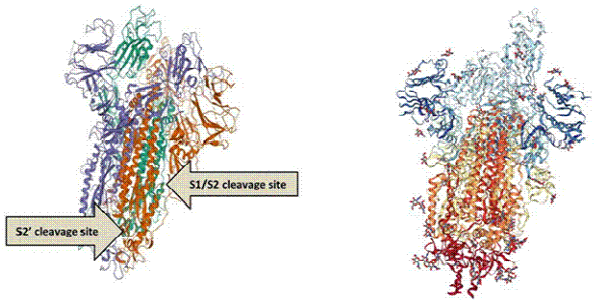
The above motif KRSFIEDLLFNKV occurs in the spike glycoprotein [4] responsible for initial binding of previous SARS coronaviruses to lung cells and their activation of the spike protein by a proteolytic cleavage [5-7]. The spike glycoprotein (or just "spike protein") is the familiar spike that studs the surface of the coronavirus, giving it the appearance of a crown to electron microscopy, hence "corona" (Latin: crown). After the completion of the first version of the previous paper [3], a bat virus with 97.41% identity of the amino acid sequence of the spike protein discussed extensively in the present paper, was entered into Genbank as entry QHR63300.1. As of the time of final writing this on April 2nd 2020, there is 100% match of this protein with entry YP_009724390.1 that appears to be a same or similar to the above Wuhan isolate. The top hundred non-redundant matching entries found using MN908947.3 by BLASTp at https://blast.ncbi.nlm.nih.gov/Blast.cgi (see below) for MN908947.3 spike protein used here vary from the above 100% match down to 75.80%, such as AAU04646.1, which is a civet isolate.

In viruses, proteins of a similar protruding nature, e.g. the hemagglutinin of influenza A, are primary targets for vaccine development, and important targets for development of therapeutic drugs that seek to block the virus from infecting host cells. At the time of the current project, only the three dimensional structures of the SARS-CoV spike proteins of the earlier SARS outbreak was known (e.g. ref [8]), which has only 75%-81% sequence match to SARS-CoV-2 [3]. Note that it is customary to write

SARS-CoV rather than SARS-CoV-1. RNA viruses mutate with high frequency but so far the differences in spike proteins in emergent SARS-CoV-2 variants are much less. At the time of the study in late February and early March 2020, the amino acid residue sequences of the spike proteins of COVID-19 isolates from different states and countries, such as California, Brazil, Taiwan, and India, remain identical or almost so. For example, with respect to the original Wuhan isolate [2], phenylalanine (F) is replaced by cysteine (C) as residue 797 in a Swedish isolate, and alanine (A) is replaced by valine (V) as residue 990 in an Indian isolate. As of 21st March 2020, largest variants in the SARS-CoV-2 genome as a whole show 99.9% nucleotide sequence match, which for a genome of 29,858 RNA bases, suggests approximately 30 base changes, and of the order of 5 in the spike glycoprotein gene of 3821 nucleotides. That then suggests roughly 1 to 3 amino acid differences in the spike protein of current (March 2020) SARS-CoV-2 variants, consistent with the above more specific observations of isolates from California etc. A single amino acid change can, of course, sometimes have significant effect, e.g. on the aggressive character of a coronavirus, and so be considered as creating a new strain. Some new strains are being reported at the time of writing, but to the author's knowledge none of them are spike protein variations, and more specifically none are as yet in the KRSFIEDLLFNKV subsequence.

Fig. 1.

The Spike Glycoprotein of SARS-CoV (left) and SARS-CoV-2 (right), Showing the two Proteolytic Cleavage Sites well Established in SARS-CoV. The Arginine (R) in the conserved motif KRSFIEDLLFNKV is the cleavage point in S2'.



approximate positions of the cleavage points in the SARS-Cov S1 spike protein superimposed on the PDB entry 5XLR.

PDB entry 6VYB SARS-CoV-2 spike ectodomain structure (open state)

The left hand side of Fig. 1 shows the SARS-CoV (previous SARS) S1 spike glycoprotein within the trimer that makes up the spike. The right hand side shows SARS-CoV-2, the SARS of current concern. All human SARS coronaviruses (and indeed the spike proteins of many other related coronaviruses) appear similar in overall

conformation, and the variations seen in experimental structures are probably more to do with crystallization or other preparation methods, particularly regarding solvent details and ligands. SARS-CoV, on the left, has been well studied and still serves as the reference model. In order to fuse with and infect cells, the spike protein needs to be in an open state; presumably the closed state makes it less vulnerable to antibodies. On the left, Fig. 1 also shows the approximate positions of the cleavage points superimposed on the Protein Data Bank (PDB) entry 5XLR for SARS-CoV. Reading from the N-terminus of S1, the important functional elements of SARS coronaviruses deduced from SARS-CoV studies [5, 6] and applicable to SARS-CoV-2 are the S1 N-terminal domain (S1-NTD), the S1 C-terminal domain (S1-CTD), the S1/S2 site as the first protease cleavage site as a loop between a pleated sheet and a-helix, the fusion peptide (FP) associated with a highly disordered loop between two a-helices which contains the second cleavage site S2', and a heptad repeat (HR). The Arginine (R) in the conserved motif KRSFIEDLLFNKV that was of interest in the previous study [2] is the cleavage point in S2'.

Recall that the KRSFIEDLLFNKV subsequence associated with S2' is potentially important, not least because it must be exposed or exposable (because it permits proteolytic cleavage) and therefore the site cannot be well shielded. The experimental three dimensional structures of coronavirus spike proteins do not for the most part reveal the large amount of glycosylation that protects most of the spike protein surface. Possibly the major problem, however, is not so much in the selection of accessible surface regions as a basis for design entry inhibitor and vaccine design [8, 9] but that the coronavirus readily escapes from such agents by mutation, including in the spike protein [10, 11]. This is the further importance of being a highly conserved motif, i.e. a subsequence that does not readily change from strain to strain except for a conservative sidechain replacement in more remote strains. Of course, as one carries the study forward to more distantly related viruses, one expects the motif to differ at some stage, and this is investigated later below. In contrast, nonetheless, the PIGAG motif "associate with the S1/S2 cleavage site disappears in coronaviruses that are not too distantly related [3,4]. As noted above, a high degree of conservation of KRSFIEDLLFNKV in the face of genetic indicates that it is in some way important to the virus, presumably for the proteolytic activation cleavage, and/or initial binding to lung cells, but there could be other interactions with other proteins, i.e. to reduce an inflammatory response, as discussed later below.

1.4. Review of Strategies for Design of Synthetic Vaccines and Pharmaceutical Agents.

Modern computer-driven strategies, and the kind of chemical products that they help produce, differ substantially from the earlier and more familiar approaches in which the computer played little if any role. In large part this is due to the invention of

automatable peptide synthesis by Merrifield in 1963, who used solid phase peptide synthesis based on crosslinked polystyrene beads [12]. Traditional vaccines are purely biological, being composed of dead or attenuated strains of pathogen (meaning mainly, viruses and bacteria). In contrast, a synthetic vaccine is a vaccine consisting mainly of synthetic peptides but also sometimes carbohydrates, often linked to a carrier protein to render it immunogenic. Such vaccines produced *via* chemical synthesis are safer because they do not involve cell-derived material or biological processes for production. Their purity can be controlled as in the case of classical drugs. The world's first synthetic vaccine was created in 1982 from diphtheria toxin by Louis Chedid (scientist) from the Pasteur Institute and Michael Sela from the Weizmann Institute. In 1986, Manuel Elkin Patarroyo created SPf66, the first version of a synthetic vaccine for Malaria. Primarily applications so far have been veterinary. Many early vaccines used dead samples of foot and mouth disease virus to inoculate animals, but they caused real outbreaks. Scientists discovered that a vaccine could be made using only a single key protein from the virus, and later also found that loops from the surface proteins could be cloned or used in cloned or synthetic constructs. Novartis Vaccine and Diagnostics, among other companies, developed a synthetic approach that very rapidly generates vaccine viruses from amino acid sequence data in order to be able to administer vaccinations early in a pandemic outbreak.

Traditional vaccines have so far remained the popular choice, but during the H1N1 outbreak in 2009, they only became available in large quantities after the peak of human infections. This was a learning experience for vaccine companies. Creating vaccines synthetically would be currently more expensive but has the ability to increase the speed of production and to retune and fine tune the solution to combat new variations in pathogens. This is all especially important in the event of a pandemic. Synthetic vaccines are also considered to be safer by researchers than vaccines grown from e.g. eggs or from bacterial cultures (in the latter case there may even be other viruses present). Cloned proteins can however reflect the same desirable principles; regions of pathogen amino acid sequence acting as epitopes (see below) can be presented as loops at the surface of a cloned protein. The general idea is that synthetic vaccines are freer of contaminants and focus on the essential features of the required immune response. They can also be developed in a more logical step by step approach. For example, sometimes diagnostics are considered as a useful early step on the way to a vaccine, since they are only required to raise antibodies in animals such as sheep for diagnostic kit production, not to be safe in humans and also raise immune system memory and a cellular as well as antibody response.

Synthetic vaccines also have the advantage that they can be seen as cartridge vaccines, meaning that they contain bits and pieces that can readily be replaced by others to update the vaccine in order to combat new strains of pathogen. A synthetic

vaccine thus has several functional components, looking somewhat like a Swiss Army Knife under the electron microscope. The key component reproduces the essential features of a pathogen protein that the immune system sees. It is an epitope that typically means a patch of some 5 to 20 amino acid residues. Reproduced as a short peptide, epitopes can be considered as haptens. Haptens are substances with a low molecular weight such as peptides, small proteins and drug molecules that are generally not immunogenic and require the aid of a carrier protein to stimulate a response from the immune system in the form of antibody production. There are two main types of epitope, B and T, discussed in Theory Section 2. A synthetic vaccine consists of T-epitopes as haptens (for cell response and immune system memory), molecular adjuvant (e.g. muramyl dipeptide), and possibly excitatory or anti-inhibitory peptides. B-epitopes are good for raising antibodies in e.g. sheep to use in diagnostics/biosensors, all attached to, or cloned into, a carrier protein. The latter must be safe but at the same time sufficiently different from any human protein to avoid autoimmune disease. Used extensively as a carrier protein in the production of antibodies for research, biotechnology and therapeutic applications, keyhole limpet hemocyanin (KLH) is the most widely employed carrier protein, and least for studies using laboratory animals. For humans the food and drug authorities may have other preferences for carrier protein, but KLH illustrates the desired features. Its large size and numerous epitopes generate a substantial immune response, and abundance of lysine residues for coupling haptens allows a high hapten:carrier protein ratio, increasing the likelihood of generating hapten-specific antibodies. Because KLH is derived from the limpet, a gastropod, it is phylogenetically distant from mammalian proteins, thus reducing false positives in immunologically-based research techniques in mammalian model organisms, and clinically avoiding autoimmune effects. So far, the food and drug authorities do not seem to have favored synthetic vaccines for human use, but this may be more to do with the peptides themselves than the carrier proteins available. The earlier methods of peptide synthesis did not achieve high levels of purity. However, this has changed and quite elaborate peptides as well as proteins can be made, facilitated by making peptide synthesizers run fast to avoid the slower side reactions, and by methods that join shorter synthetic peptides into longer chains [43-46].

One of the original motivations for the present study was to capture experience and design strategies from vaccine, diagnostic and antagonist design [12-21]. Methods by the author and colleagues ranged from the Expert System Approach to automated bioinformatics and protein modeling [26-28] and automated drug design (e.g. refs [29-32]). See also ref [33]. More recently there has been an automated approach based on the proposed Q-UEL language [34-37]. The more fine-grained principles for the design of synthetic peptide vaccines, and antagonist peptides made of D-amino acids, are discussed in some detail in the previous paper [3]. A variety of bioinformatics techniques are available to help in development of these solutions (e.g. ref [38-42]), as

well as computational (e.g. refs [33]) and synthetic techniques (e.g. ref [43]). The Q-UEL language [33-37] used in the preceding work [3] is also a means of gathering relevant information from the World Wide Web efficiently when encountering a new problem such as an epidemic caused by a new virus, or at least a problem new to the researcher [3, 4, 38]. It also enables more automated interaction with websites for publically available bioinformatics tools. The motivation for this was all the stronger because the popular highly integrated approach to bioinformatics' called the Biology Workbench at the University of San Diego Supercomputer Center has been no longer available for some time [39]. However, the standard bioinformatics tools (e.g. refs [40-42]) used in the present study can of course be used readily by researchers reasonably experienced in bioinformatics.

Although peptidomimetics (containing amino acids that would not occur in normal ribosome-based biosynthesis) have been considered by authors as a basis for haptens in synthetic vaccines, they are in the author's opinion probably best considered as potential therapeutic antagonists. In the present study, the specific aims include design of molecules to impede binding and activation of the spike glycoprotein at the surface of lung cells [5-7]. Synthetic peptides copied from subsequences in the spike protein could be used directly for such clinical purposes, but then an important design step would be to render them resistant to biodegradation by human proteases. This is typically by inclusion of D-amino acid residues [44-47].

Previously, in the author's personal opinion, peptides and peptidiomimetics have been currently best considered as first steps in the research and development of small organic "in a pill molecules" of the traditional kind favored the by the pharmaceutical industry. Their role there nonetheless is a powerful one, linking amino acid sequences seen in nature, conveniently already "designed" by millions of years of evolution, to (typically) smaller novel organic molecules designed to have van-der-Waal's and electrostatic features in the binding site. However, the author's reticence has been largely based on cost, including cost in changing traditional production strategy, and in the reservations of food and drug authorities, but fairly recently all that appears to be changing. THPdb (http://crdd.osdd.net/raghava/thpdb/) includes an example of a manually curated repository of peptides and related molecules approved by the US Food and Drug Administration (FDA). Over some 70 peptide drugs are approved in the US and other major markets, and those in pipelines and in or approaching clinical trials may now be exceeding 200 entries. As natural compounds, peptide drugs are typically less toxic than more traditional chemical candidates. Although D-amino acids are not natural features of ribosomal production of peptides and proteins, human metabolism can handle them. They occur in gut microbes and ingested material and in human proteins they form spontaneously in a kind of aging process from some amino acids *in situ* in protein sequences (e.g. L to D-aspartate). Diverse D-amino acids such as D-

serine, D-aspartate, D-alanine, and D-cysteine are found as free amino acids and small peptides as well as in some proteins, and quite commonly in mammals. They are often found having playing important roles in the nervous system. For example, N-Methyl-D-aspartate (NMDA) receptors are associated with learning and memory and D-Serine, D-aspartate, and D-alanine bind to those receptors. Hydrogen sulfide generated from D-cysteine reduces disulfide bonds in receptors and potentiates their activity. Peptides made of D-amino acids resist not only normal proteolytic degradation but also resist an immune response (unless attached to a carry protein) [44, 46]. They persist some 4-6 days in the body, which is an ideal time period for pharmaceutical action, and are ultimately degraded to safe products (probably mainly in the peroxisomes and by enzymes in the kidneys) [44, 46]. The negative aspect is that they do have higher entropy to overcome than many drugs of more traditional form, but in practice this appears to be more a barrier to computer simulation of binding than to the real molecule, as extensively discussed below.

Studying the binding of synthetic peptides or small organic molecules to human proteins benefits from computer simulations of the solute-solvent system, and it was early found that these should ideally include water molecules in a detailed way because there are hydrogen bonding options between water molecules and amino acid residues which are not particularly intuitive [48, 49]. In most cases, the spatial locations of hydrogen atoms are deduced rather than seen in experimental protein and peptide three dimensional structures. This is likely to impact considerations of docking ligands to protein targets. In the present author's opinion, this provides a beneficial possibility for retroinverso designs [3] made by reversing the sequence and using D-amino acids that has the unfortunate or complicating effect of interchanging the C=O and N-H groups in the backbone of the synthetic peptide [3]. The beneficial possibility is that, for example, a repulsive C=O…O=C electrostatic interaction between a synthetic peptide and the spike protein could be ameliorated in the manner C=O…H…O=C where the H is a water, serine or threonine hydrogen atom, or by C=O…H-O-H…O=C, albeit that in practice the water molecule likely lies more to the side of the O..O interaction vector. Somewhat similar considerations apply to any N-H…H-N interactions that can ameliorated by the lone pair orbitals of an oxygen atom. Both could also involve tautomerization and/or rearrangement of internal hydrogen bonding networks (e.g. in the manner …O-H…O-H…. to …H-O..H—O…). Today, to take care of such matters, researchers consider docking of ligand to protein and high grade molecular dynamics simulations of the overall solute-solvent system by molecular dynamics, at least as the final refinement step [50], but even the awareness that the above compensations and others can take place can make it worthwhile to synthesize and test a proposal. Somewhat similarly, design of peptide synthetic vaccines and diagnostics can make direct use of peptides duplicating sequence motifs in the pathogen protein found by bioinformatics and relatively simple computational tools. After that, researchers often go

straight to synthesis and experimental immunological testing of the constructs rather than using complex simulations [51-53]. *Epitope predictions* for SARS-CoV-2 (simply meaning the choice of amino acid residue subsequences to synthesize for synthetic vaccines, but also for peptidomimetic antagonists) have been made by several authors (e.g. ref [54]). They have typically made use of extensive historical experimental data about the amino acid residue sequences of epitopes such as the Epitope Database and Analysis Resource (IEDB) and the Virus Pathogen Resource (ViPR) (e.g. ref [54]).

1.5. Human Protein Targets for Design of Therapeutics against COVID-19.

The immune system by its nature can make its own adjustments to recognize pathogens and vaccines, but designing some kind of therapeutic antagonist against virus binding to the lung cells requires rather more consideration about what human protein the spike protein is binding. Bioinformatics as the study of biosequences is a powerful tool, but it is well known that having the detailed three dimensional structure of the human protein target for a potential new pharmaceutical agent, or to which a virus attaches, is a great benefit to rational computer-aided design. Studies specifically investigating human protein binding and activation of previously known SARS viruses have for some years been carried out by several groups (e.g. [54-57]). It seems reasonably well agreed that angiotensin converting enzyme type 2 (ACE2) is responsible for binding the SARS associated with the 2002 outbreak, combined with a proteolytic cleavage to activate the spike protein, for which type II transmembrane serine protease (TMPRSS2) is the current popular candidate [3]. Several three dimensional structures are known for ACE2 complexed with SARS spike protein e.g. protein data bank (PDB) entry (6ACG) and of variants of the latter (e.g. TMPRSS2 protein data bank entry 2OQ5).

However, the full story involving human cell surface proteins (with which SARS-CoV-2 interacts in order to infect and replicate) is possibly not quite as firmly established at the time of this present study as some summaries would suggest. The origin of the general problem for a more detailed conformational chemistry approach is that diversity of genome and means of infecting cells are readily generated in nature in the case of different virus hosts, virus strains, and species jumps, and it is long established that the binding shows variation in the receptors used that correspond to viral groups. There have been alternative proposed candidates for initial binding receptors, e.g. carcinoembryonic antigen-related cell adhesion molecule 1 (CEACAM1), and various dipeptidyl peptidases. Highly virulent coronaviruses that form syncytia between cells can even spread in a receptor-independent fashion. Even when an initial binding receptor such as ACE2 is identified for a coronavirus, initial uncertainty or enduring complexity for the rest of the entry process may be the norm. Many other human proteases present in the lung seem capable of cleaving various sites on the spike protein and which could cause its activation. For example, a variety of proteases

such as trypsin, tryptase Clara, mini-plasmin, human airway trypsin-like protease (HAT), and TMPRSS2 (transmembrane protease, serine 2) are known to cleave the glycoprotein hemagglutinin (HA) of influenza A viruses as prerequisite for the fusion between viral and host cell membranes and viral cell entry. Human airway trypsin like protease (HAT), TMPRSS3, TMPRSS4, TMPRSS6 have also all been considered by SARS researchers at various stages. Other human proteins that might have similar involvement to the above in the SARS-CoV-2 case, and that are also affected by the same antagonists against the SARS-Cov-2 targets in the preceding paragraph, have also attracted the attention of researchers. The trypsin-like serine protease hepsin which has a fairly broad action and which is significantly inhibited by a diverse set of ligands, a particular example of one such binding is represented by protein data bank entry 5ce1. Even intracellular proteases could be released on cell damage resulting from the first wave of lung infection or from other disease or tissue trauma. Some variants and strains may use other, as yet unknown, proteins, or sugars, to assist entry. It is also plausible the spike protein might be activated by other proteases on *exit* from the epithelial lung cells, so allowing it efficiently to infect other cells. The spike glycoprotein of SARS-CoV-2 also has the so-called furin cleavage sequence (PRRARS or PRRARS), which is an extension to the so-called PIGAG motif of ref [3]. Consistent with the present author's preferred choice of KRSFIEDLLFNKV motif, coronaviruses with high sequence homology (such as that isolated from a bat in Yunnan in 2013), lack the furin cleavage sequence. Nonetheless, because furin proteases are abundant in the respiratory tract, SARS-CoV-2 spike glycoprotein might be cleaved on exit from cells.

Even if the means of binding, activation and entry is well established for a viral strain, recall that a single RNA base difference resulting in a single amino acid residue difference could alter all that, and there also appear to be several other possibilities that the virus can exploit *in parallel*. Indeed, somewhat similarly, potential inhibitors of SARS entry and/or activation proposed by researchers (e.g. refs [55-61]) may work by several routes in parallel, and significantly at least three mechanisms were reported in one relevant study [61].

1.6. The Pharmacophore Approach.

Once a target protein and its relevant binding site are clearly understood, methods are available for screening available ligands (binding molecules) to bind to those sites as potential antagonists, or even for "growing" or evolving antagonist molecules in those sites, whether smaller organic molecules [29-31] or peptides [32]. Pharmaceutical chemists have long used evidence and hunches to deduce a *pharmacophore*, i.e. an abstract description of recurrent molecular features that are necessary for molecular recognition of the ligand by the protein [3]. A pharmacophore ultimately implies at least a schematic model of the interfacial surface between ligand and protein, but in practice, a pharmacophore tends to be either considered from the

perception of the ligand (one compares similar inhibitors etc.) or from the perception of the binding site (one compares positions of key residues in the binding site). The choice depends on the quality of each kind of data, but could involve both. Historically, drug design was frequently based only on indirect deduction of binding site features using the chemical features of the ligands which successfully inhibit (or in a few instances excite) a response. This is essentially the use of *Quantitative Structure Activity Relationships* (QSAR). In effect the perception of the binding site was indirect and typically based on the chemist's expertise and hunches, and so often extremely "fuzzy". Subsequent elucidation of many protein structures with clear pictures of their binding sites led to a crisper physical perspective, exemplified by a study [50] that included many ligand molecules in the present investigation, and so faced some similar issues. In the approach which may now be considered traditional, docking calculations are fast, using grid maps that consist of a three dimensional lattice of regularly spaced points, centered on some region of interest of the protein target under study.

As discussed above, ACE2 and TMPRSS2 are very likely correct targets, but again they are not necessarily the only targets even for cell entry of current SARS-CoV-2, and the mechanisms used by each new coronavirus strain can differ, as the result of even a single amino acid residue change. In such circumstances, the conservation of the KRSFIEDLLFNKV motif might be considered suspicious. The activation cleavage is at the arginine (R) and workers tend to conclude that this site is more essential for action than S1/S2, and mutation of the arginine (R) specifically inhibits trypsin-dependent fusion in both cell–cell fusion and laboratory assays. But also, with the arginine retained, many other proteases can active the spike protein as above, and others can do so in laboratory conditions. Because of the conservation, one might therefore hold the seemingly reasonable hypothesis that this site is not also susceptible to cleavage and activation by other extracellular proteolytic enzymes, but also doing something else. Whether or not this is so, all this complexity makes detailed interaction models of spike protein binding and activation difficult, and while the "best bet" for ranking the choices of target protein may currently *seem* obvious, making a reasonable, currently conventional, choice which is actually an incorrect assumption can delay productive research into therapeutic agents. In the case of the hunt for prevention and cure of virus diseases, and particularly COVID-19, there seems to be increased justification for a "fuzzier" set-theoretic picture of a pharmacophore as an *ensemble* of different binding sites, or of ligands in a ligand-oriented perspective, as follows. Many of these, and perhaps all, suggest that even if one is using an incorrect picture of the mechanisms of entry and replication, even using the "wrong" or less important protein target, one might achieve some success. In brief summary, the justifications for the ensemble pharmacophore in the coronavirus case, i.e. the contributions to "fuzziness", include *parsimony*, that proteins and parts of proteins sometimes have more than one function [12] encouraged by limited numbers of accessible sites (due to e.g.

glycosylation) and exemplified by parallel alternative mechanisms of cell entry, multiple methods of drug action, escape from scientific defense measures by virus mutation, polymorphism of human proteins involved, different expression levels of human proteins involved, and the potential problem of the "specter of vaccine development" (concerns about missing the appropriate region of the virus that allows common cold viruses to escape the appropriate immune response). To the above may be of course added the fact that even if an experimental researcher is convinced of the value a specific protein as appropriate target, the picture for the computational chemist is a fuzzy one. The system itself, real and simulated, is to be seen as a *statistical mechanical ensemble* of multiple states, sampled over the population of molecules and across their conformational behavior in time. Not least, protein binding sites are often partially disordered before binding, and in any case there may be several binding modes. Picking the right one can be difficult because there is a fine balance between solvent and conformational entropy, and entropy is notoriously hard to compute [12].

Given this argued uncertainty as to the nature of the target protein and its binding site, a broader initial net as an *ensemble pharmacophore* can help. Docking approaches are continually being improved by researchers, and recently include ways of combining features that could ultimately relate to different protein binding sites. While many authors of these studies include the word "ensemble" in their discussion of pharmacophores, they appeared to be significantly different to the particular means of combining multiple pharmacophores that was explored here. However, the present author has had his attention drawn to some that are rather similar and the approach of Kumar [33] appears to particularly akin, especially in regard to distributions of expected values and use of weighting. Kumar's description [33] thus suffices, and briefly stated, it explores the ability of an ensemble of selected protein-ligand complexes to populate pharmacophore space in the ligand binding site, assesses the importance of pharmacophore features using Poisson statistical and information-theoretic entropy calculations, and generates the pharmacophore models with high probabilities. A scoring function then combines all the resultant high-scoring pharmacophore models. There is one significant operational difference between Kumar's approach and that used here. Recall that in the more traditional docking approach, it is the ligand as candidate drug that is typically seen as the variable and constantly changing and in many studies "evolving" the ligand chemistry with the pharmacophore is the basis of drug design [29-32]. Ref [50], related to the present study, has aspects of that applied in a different way. Kumar's approach can, however, combine the perspective of both pharmacophore and ligand as conceptual variables. Despite that, the present author's approach, as used in the present overall project, considers one candidate ligand at a time. This seems less efficient from the point of view of designing candidates, and not even as smart as the older single, non-probabilistic pharmacophore approaches [29-32]. Nonetheless, a single, simpler one-ligand-at-a-time strategy is both adequate and appropriate in the

present case. This is because there is already a data collection of candidate antagonists to build on [50], as discussed in Section 1.7.

Approaches of the ensemble pharmacophore kind are currently highly recommended for investigation of SARS-CoV-2 and for the spike protein in particular, again because of some uncertainties and the likely multiple functions of some spike protein features. However, it has not as yet had significant impact in the present study. The approach actually taken remains consistent in the sense that inclusion of one particular source for a pharmacophore, an enzyme considered by the author, was evidently going to dominate the ensemble because of certain similarities in the antagonists of SARS virus entry and inhibitors of the enzyme [3], given the knowledge available at this time. That choice is not, however, obvious, as follows.

1.7. A More Controversial Selection of a Representative Protein as Pharmacophore.

What may be more controversial is the case when there is a representative choice and it is a protein that is not obviously relevant to the target protein, or simply not “on the radar” of coronavirus researchers. What makes it a candidate is not necessarily that it is relevant to viral infection and not necessarily that it has an evolutionary relationship to proteins that are considered relevant, although this is a question addressed briefly in this paper. Rather, it may simply be based on the pragmatic notion that there may be ligands, potential binding molecules as antagonists, which are common to both more popular choices human target proteins and a less obvious candidate. The small organic molecule emodin has been found to inhibit SARS coronavirus entry [59, 60], as also so have other compounds some of which have emodin-like features [3, 61]. Similar molecules, and importantly emodin itself, are also inhibitors of 11β-hydroxysteroid dehydrogenase type 1, an example of a steroid binding enzyme [62]. It is normally anchored within the endoplasmic reticulum through an N-terminal transmembrane domain. Its involvement as a protein target is here based on a chemical justification. A biological justification might be that this enzyme is involved in the inflammation response which a coronavirus might also benefit by inhibiting. If so, the goal is not, of course, to help the virus by inhibiting at the same target which it would also gain by inhibiting, but rather to inhibit protein targets more crucial to it, i.e. for cell entry and possibly replication which are even more crucial to the virus. Some inhibition of 11β-hydroxysteroid dehydrogenase type 1 might even be a desirable thing because excessive or prolonged inflammation (including in response to pathogens) is well known to be potentially damaging to the host. An excessive inappropriate immune response may also include the basis of allergic reactions and even of autoimmune diseases.

A pragmatic reason for this choice of protein as pharmacophore is that was also one of those protein-ligand interaction systems that have been well studied by the

present author and collaborators [50]. Such studies pursued the idea of using a more rigid molecular framework, including the steroid framework and fragments of it, as a more rigid scaffold for active drug groups [50]. Importantly, that study and subsequent work has already established data base of compounds that bind to 11β-hydroxysteroid dehydrogenase type 1, and it includes many molecules including some discussed in this paper that again have some of the features of emodin. It also includes many weak binders that could also be much stronger binders at what turns out to be a more obviously relevant protein target. These issues can be addressed quickly in the laboratory and certainly seem worthy of investigation before addressing the more popular targets.

1.8. Do the Peptidomimetics and the Smaller Organic Antagonists Act at the Same Site?

There was a further implication in the previous paper [3], though not a requirement for its main arguments, that the peptides designed on the basis of the KRSFIEDLLFNKV motif bind the same KRSFIEDLLFNKV site as do emodin-like molecules. That seems currently to be an even less reliable assumption than the assumption that the above steroid dehydrogenase enzyme is relevant to coronavirus biology, and it is not of course an assumption that even matters if either a peptidomimetic and/or small organic molecule is found effective. However, again keeping in mind that there are a limited number of accessible, conserved sites in the spike protein, and that these may be involved in multiple mechanisms as discussed above, common targets for action of both peptides and smaller organics like emodin seems plausible. Partially the problem is extensive glycosylation. It is well known that glycosylation plays an important role in receptor-ligand recognition but also have structural influence in receptor-ligand recognition because of its bulky shape caused by branched side chains. For that and other reasons it may be that the KRSFIEDLLFNKV site is, with just a little variation, almost the only site on the spike protein that is persistently recognizable in coronavirus strains, and so presumably carrying out an important function and accessible, as also discussed in this paper. Angiotensin converting enzyme 2 (ACE2) binding is however also considered in this paper.

2. Theory.

2.1. Theory behind the General Strategy.

A number of ideas and principles, borrowed in established and recent design of synthetic vaccines and petidomimetics, were used (see ref [3] for discussion and e.g. refs [63-69]), as well as some of the ideas that lie behind the popular ZINC data base [70]. As discussed in refs [3, 4], the present investigation started as a use case for the Hyperbolic Dirac Net (HDN) and particularly the associated Q-UEL language for

automated inference [34-37]. The theory has been discussed elsewhere, e.g. in refs [34-37], which relate more to the practical and general uses of Q-UEL. These considerations are less important here because present studies can be reproduced by standard bioinformatics and molecular modeling means. Nonetheless, it is doubtful that the research for refs [3,4] could have been done and written up so rapidly without the aid of Q-UEL to interact with websites of the World Wide Web, gather knowledge, and facilitate use of the publically available bioinformatics tools [3].

2.2. Basic Principles of Epitope Prediction for Design of Synthetic Vaccines.

The challenge is ultimately one of molecular recognition but in practice many key principles for hapten design relate to distinguishing types of naturally occurring epitope. By the term “epitope” in this paper is meant “continuous epitope”, though several smaller epitopes may be joined to represent a discontinuous epitope in which conformation and relative position in space can sometimes be important. While a synthetic construct implies the use of synthetic chemistry typically combined with a judicious carrier protein to which the peptide is linked chemically, constructs can also be obtained by cloning, using protein engineering principles [12]. The terms B-epitope and T-epitope relate to the traditional picture of a bone marrow B or thymus T response. B cell epitopes occur at the surface of the protein against which an immune system response is required. They are recognized by B cell receptors or antibodies in their native structure, and are concerned with the bone marrow response and antibody production. T epitopes may be buried inside protein structures and released by proteolysis, and are traditionally considered as concerned with a cellular response and immune system memory, i.e. active immunity. Continuous B cell epitope prediction is very similar to T cell epitope prediction. The focus is on B-epitopes here, though a B-epitope can also be (or overlap with) a T-epitope especially if it has a significant content of hydrophobic residues. Prediction of these has traditionally been based has mainly been based on the amino acid properties such as hydrophilicity, charge, exposed surface area and secondary structure. There are many predictive algorithms available, but the present author prefers a more “expert system” kind of approach that incudes experimental data, though the above biophysical considerations certainly still play a strong role (see below).

2.3. Some Theoretical Issues Related to Design of Antagonists of COVID-19 Infection.

The previous paper [3] focused primarily on design of synthetic peptides as infection antagonists. However, partly for the reason of greater conformational flexibility discussed below, smaller less flexible organic molecules (i.e. with fewer rotatable bonds) are the traditional province of the synthetic chemist rather than use of an automated peptide synthesizer, are preferred for pharmaceutical application.

Consideration of peptides is more often considered as merely a useful intermediate step in more traditional pharmaceutical compound design. Biodegradability *per se* of peptides is not the main concern, since including D-amino acids in the design prevents proteolysis. In preliminary docking and simulation studies, the peptides do bind to 11β-hydroxysteroid dehydrogenase type 1, but less strongly and with several binding modes [3]. This weaker binding is not in itself a contraindication of the idea that these peptides bind at the same site as the more rigid non-peptide molecules, because it is an expected consequence of the much greater flexibility of peptides compared with molecules with, for example, multiple aromatic ring scaffolds. Conventional wisdom (e.g. ref [12]) frequently uses the rule-of-thumb that the total change in intramolecular (bond rotational) entropy of a peptide ligand is roughly $T\Delta S_{Total}$ = 1.5 kcal·mol^{-1} per residue at 300 K, corresponding approximately to a 12-fold reduction in conformational freedom per residue on binding. Because van der Wall's and hydrogen bonding tend to be very roughly equivalent for peptides in water and in well bound forms, the water entropy effects known as hydrophobic effects (along with electrostatic forces) play an important role in determining the balance of energies and final outcome. KRSFIEDLLFNKV would thus cost about +19.5 kcal/mole entropic contribution to bind rigidly, primarily compensated by hydrophobic contacts at up to about -1.7 kcal/mole in going from an aqueous to a non-polar environment, i.e. -22.1 kcal/mole for a 13 residue peptide or analogue of KRSFIEDLLFNKV. That example would not favor binding, but the proper calculation is in the details which of should show balance that favors good binding if that is found to be the case experimentally. Despite the above comments, the flexibility of peptides does provide more opportunities to fit a specific binding site, i.e. they can show some accommodation and they are more tolerant to imperfections in the design process. However, this is also an argument for their importance as an intermediate step in the design of more conventional pharmaceutical agents.

3. Methods.

3.1. Computational Methods.

The main methods are essentially standard bioinformatics approaches as used in refs [3, 4]. Some methods, e.g. rules for epitope prediction, are best discussed in context in Results Section 4. The Q-UEL methods specifically for bioinformatics are discussed in [38], and those for computational chemistry and docking of compounds are those using KRUNCH as described in ref [3] and the appendix to ref [50]. They are somewhat unorthodox by focusing on heuristics to handle the multiple energy minimum problem, but the end effect is probably similar to that of long runs using high grade molecular dynamics calculations, given opportunities for calibration [50]. Epitope predictions lie in more traditional "one dimensional" bioinformatics, and in this paper and the previous paper depended on predictions using a GOR4 secondary structure

prediction of α-helix (h), extended chain or β-sheet (e), and coil or loop (c). The reason for this and the particular use of GOR4 is discussed in ref [3], but briefly, it is in part because sections predicted by runs of c tend to be immunogenic even if they are incorrect as structure predictions [3]. However, charged residues in α-helices and β-sheets are believed to be occasionally B-epitopes, and short sections extended chain can effectively imply loops. The core and initial rules for B-epitope prediction used in the present study consider

(i) surface exposure when a three dimension structure is known, but allowing for conformational adjustment to expose residue when in a likely disordered or flexible loop, scores +2,
(ii) known exposure based on other kind experiment, which also recognizes the possibility that a partially buried site by the above criteria can be brought to the surface on binding, notably for proteolytic cleavage [3],
(iii) runs of amino acid from the set [STNQY] score +1, from the set [DEKHR] they score +2, and from the set [LIVFCM] they score -1,
(iv) runs of secondary structure prediction as coil or loop c, though runs of three or less e and the first and last three of helix h can be considered as c for this purpose, score +1, and
(v) the motif NX(S/T)X of asparagine (N) serine (S) or threonine (T), where X means "not a proline" (P) scores +2. However, this will not permit a corresponding peptidomimetic or vaccine without considering glycopeptide synthesis technology. See discussion below, which would justify a negative score, depending on the technology available.

In addition, these may be combined with predictions based on significant homology with proven epitopes in data bases, which has already been done by several groups for SARS-CoV-19 (e.g. ref [54]).

3.2. Data Sources.

For sources of data concerning COVID-19 virus spike proteins, GenBank and the Protein Data Bank were the main sources. There was some use of in-house collections of data, e.g. of typical B-epitopes and T-epitopes, although publically available collections would probably serve the same function. There was also use also of a data base of non-peptide ligand molecules of potential interest already generated during and since the work described in ref [50] that was used where appropriate. Many of these molecules (including emodin) are also found on the public ZINC data base [70] as indicated in Results Section 4 below, but several, including derivatives of carbenoxelone, are not, and these derivatives are of interest as potential coronavirus antagonists. To look up an entry on the ZINC data base by the codes used in this and other papers, one can go to http://zinc15.docking.org/ substances/ and enter

ZINC00011032. In an automated approach such as that favored by Q-UEL, a variable (such as a Perl variable $mol) to ZINC00011032 and is set an the Q-UEL application goes to http://zinc15.docking.org/substances/search/?q=$mol. Any references to experimental binding results concern data from cited papers, and see for example ref [69] for typical methods used for natural herbal compounds. As discussed in ref [3], Q-UEL helped gather these in the form of Q-UEL knowledge representation tags, so they become part of the growing knowledge Representation store.

3.3. Notation.

In regard to peptides and proteins, Table 1 used in ref [3] shows the standard IUPAC one-letter codes used for amino acid residues in sequences throughout this paper.

Table 1.
One Letter Amino Acid Codes Used in the Text.

One letter code	Amino acid	Conservative replacements
A	alanine	A, E, S, T
C	cysteine/cystine	S, T, V
D	aspartic acid	E
E	glutamic acid	A, D
F	phenylalanine	M, W, Y
G	glycine	N, P
H	histidine	K, R
I	isoleucine	L, V
K	Lysine	H, R
L	leucine	I, V
M	methionine	F, W, Y
N	asparagine	G, D, Q
P	proline	G
Q	glutamine	N, E
R	arginine	H, K
S	serine	A, T
T	threonine	A, I, S
V	valine	A, I, L
W	tryptophan	F, M, Y
Y	tyrosine	F. M, W

Conservative replacements are those common substitutions from a peptide design perspective, but for example phenylalanine (F), isoleucine (I), and alanine (A) are seen as natural substitutions that appear in discussion of spike protein sequence motifs later below. These amino acid residues have hydrophobic sidechains but they are not conservative replacements but rather substantially different size. A reasonable explanation is of course that sidechain size conservation matters less when the sidechains are at exposed at the surface of the protein. Similar notions underlie the idea that what can readily replace what is not always an equal probability in each direction. In that respect, Table 1 tends to reflect the changes that are used in the present project for design, when starting from epitopes.

4. Results.

4.1. Epitope Prediction.

The previous paper [3] should not give the impression that the specific motifs discussed (and particularly KRSFIEDLLFNKV) are the only sections likely of the SARS-CoV-2 spike protein to be of interest in the above respect. The preference for one choice was based on (a) conservation across many strains, suggesting that the site has an important function and is likely at the spike surface, and (b) avoiding the shielding of the spike protein by extensive glycosylation. The dramatic effect of relaxing these restrictions is a major point of this Section, in which a large number of candidates are found. Over-prediction is not necessarily a bad thing, because once a laboratory has a peptide synthesizer and other tools for constructing and testing designs, it is relatively easy and cheap to test and reject ideas, and more problematic to miss opportunities. The intention here is also to cover most possibilities, to enable index numbers to be assigned to them according to their order in the sequence (putative epitope 1 etc). Consequently, in future one may then readily refer to the index number, or speak of a new proposal or experimental epitope extending, overlapping, or even lying between two of these epitopes. They are primarily to be seen as B-epitope predictions, though they are favored if some T-epitopic character is also expected.

An initial step is based on adding up weights as described in Methods Section 3.2. In practice, there was also some judicious use of expertise and an epitope data base in an attempt to refine assignments. Recall that the trimeric SARS coronavirus (SARS-CoV) spike glycoprotein consists of three S1-S2 heterodimers. Some of these will be shielded by that configuration during most of the life cycle of the virus, but not necessarily in every S protein monomer, and also shielded by glycosylation. The higher scoring predicted epitopes in the sequence below are underlined and in **bold**, and are primarily to be considered as B-epitopes but with some extension to include T-epitope character where possible. Also included in these predictions are those using the Immune Epitope Database and Analysis Resource (IEDB) and the Virus Pathogen Resource (ViPR) which have already been made [54] (see later below). These are shown in underlined, **bold**, and in *italics* in the following, and since some are contiguous sections that look like a single long representation in the following, they are also stated separately below. It is apparent that while focus was on just KRSFIEDLLFNKV, if strain variation and glycosylation are ignored then much of the spike protein sequence contains epitope candidates.

```
        10        20        30        40        50        60        70
         |         |         |         |         |         |         |
MFVFLVLLPLV SSQCVNLTTRT QLPPA YTNSFTR GVYY PDKVFRSSVLHSTQD LFLPFF SNVT WF HAIH V
ccceeeeeccc ccccccccccc ccccc cccccc ccccc ccccceeeccccccccc ccccc ccccc cee eeeee
SGTNGTKRFDN PVLPF NDGVYFASTEKSN IIRGWIFG TTLDSKTQS LLIV NNATN VVI KVCE FQFCNDPF
```

Many of the epitopes predicted in the present study overlap with prediction made using the Immune Epitope Database and Analysis Resource (IEDB) and the Virus Pathogen Resource (ViPR) [54], and these comprised the following.

(1) DAVDCALDPLSETKCTLKS FTVEKGIYQTSN
(2) VCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDP
(3) QTLEILDITPCSFGGVSVIGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGS
(4) FSQILPDPSKPSKRSFIE
(5) FGAGAALQIPFAMQMAYRFNGI

Recall that one of the reasons for the original single preferred candidate KRSFIEDLLFNKV was that many of predicted epitopes contain evidence of glycosylation, reflecting the last of the “rules” (v) in Methods Section 3.1 above. That rule has, however, a special status, and the present author has tended to consider them undesirable for synthetic vaccine or diagnostic development. It indicates likely glycosylation of the protein. The bulky oligosaccharides so attached can be

immunogenic, but they are rather difficult to work with synthetically, traditionally expected to make bulk production expensive, and may be variable in structure which cannot typically be seen in detail in experimental three dimensional protein structures (typically as obtained by X-ray crystallography or high grade electron microscopy). Antibodies that are raised against the glycosylated surface patch of the protein or corresponding synthetic glycopeptides may be specific for their carbohydrate units. These can be recognized irrespective of the peptides, or in the context of the adjacent amino acid residues. Conformation and exposure of B-peptide epitopes of glycoproteins may be modulated by glycosylation because of intramolecular carbohydrate-protein interactions. The beneficial versus undesirable effects of glycosylation in synthetic vaccines is also a complex matter. Glycosylation may be essential for reactivity with the antibody, but conversely it may in effect inactivate the capabilities of a section of amino acid sequence to function as a B-epitope, which seems to be a very good reason for giving the glycosylation motif a strong negative rather than positive score. Unfortunately this will depend on the structure of the antigenic site and antibody fine specificity, and the recognition mechanisms involved are not fully clear. There is a (usually) positive aspect, however, in the current view that similar effects of glycosylation apply to T-cell-dependent cellular immune and IgG antibody responses, and that glycosylated peptides can elicit glycopeptide-specific T cell clones after being bound and presented by MHC class I or II molecules. It is of course only a positive aspect if the intended effect is obtained by the synthetic construct.

4.2. Persistence of the KRSFIEDLLFNKV Motif with Minor Variations in Common Cold Coronaviruses.

The overall spike glycoprotein protein sequence shown above changes across the coronaviruses, but the KRSFIEDLLFNKV subsequence is most notable amongst the exceptions. It extends to the common cold coronaviruses with minor variation, and may imply a better targeted approach to stimulate immunity. For common colds caused by the rhinovirus, recent research suggests misdirection of antibody responses against a non-protective epitope as a mechanism how the virus escapes immunity and so permits recurrent infections [71]. A clearer understanding of conserved subsequences in coronaviruses may also help tune the action of Toll-like receptors to initiate the appropriate response. These are a class of proteins that play a key role in the innate immune system. They are single-pass membrane-spanning receptors usually expressed on sentinel cells (e.g. macrophages and dendritic cells) that recognize structurally conserved molecular features of pathogens [72].

Despite concerns about two or more strains of COVID-19 virus appearing, these are not big changes for present purposes. It is sufficient to consider the sequence of the original Wuhan isolate as reference in comparisons for present purposes, i.e. for

comparing the spike protein sequences of other coronaviruses. Recall that as discussed in Introduction Section 1.3, at the time of the study in late February and early March 2020, the sequences of the spike proteins of COVID-19 isolates from different states and countries, such as California, Brazil, Taiwan, and India, remain identical or almost so. For example, with respect to the original Wuhan isolate [2], phenylalanine (F) is replaced by cysteine (C) as residue 797 in a Swedish isolate, and alanine (A) is replaced by valine (V) as residue 990 in an Indian isolate. Neither of these relate to the sequence motif KRSFIEDLLFNKV of particular interest here.

In the initial studies [3,4], the genome of the common cold coronavirus, and particularly the sequence of the spike protein, was considered sufficiently far from that of the COVID-19 virus so as to be less relevant to that problem. While looking at differing sequences is essential for detection of conserved motifs, very different and less relevant pathogens are unlikely to preserve them, except perhaps as pattern matches involving quite complex substitution rules. However, the appearance of the COVID-19 KRSFIEDLLFNKV motif does appear in common cold coronaviruses and with typically at most two relatively conservative substitutions. That is in the sense of preserving hydrophobic sidechain as discussed above in Methods Section 3. The conservative aspartate (D) and asparagine (N) replacement is also fairly common in the motif in the sequences examined. An example shown below is a Clustal Omega alignment of the COVID-19 virus spike protein original Wuhan Seafood Market isolate (GenBank entry MN908947.3) with spike proteins representatives members of the two major common cold coronaviruses strains 229E and OC43 (GenBank Entries NP_073551.1 and AIV41987.1). Despite radical sequence differences for the spike protein sequences overall (only 12.8% identity, well within the range for a random match), the underlined sequence motif KRSFIEDLLFNKV of COVID-19 virus is essentially retained as that sequence, except that alanine (A) replaces phenylalanine (F) in the common cold coronavirus (which is moderately conservative at the surface of a protein) and a conservative leucine for valine substation in one case. In the sequence (not shown) of HCoV-HKU1 which is often associated with more serious cases of cold-like diseases the above motif is still noticeable as RSFFEDLLFDKV in which the isoleucine (I) is replaced by phenylalanine (F). The “A for F” modified motif RSAIEDLLFDKV is also found in the coronaviruses of dogs, cats, rodents, pigs, rabbits, camels, ferret badgers, raccoon dogs, amongst others. All of these might be eaten by humans in certain countries and notably they are, for the most part, species that live in close proximity to humans.

```
NP_073551.1      MFVLLVAYALL-------HIAGCQTTNGLNTSYSVCN-GCVGYSE---------NVFAVE         43
MN908947.3       MFVFL-VLLPLVSSQCVNLTTRTQLPPAYTNSFTR-------------------GVYY-P         39
AIV41987.1       MFLILLISLPTAF--AVIGDLNCPLDPRLKGSFNNRDTGPPSISTDTVDVTNGLGTYYVL         58
                 **::*                         . *:.                    ..:

NP_073551.1      SGGYIPSDFAFNNWFLL---TNTSSVVDGVVRSFQPLLLNCLWSVSGL------------         88
MN908947.3       DKV-----------------FRSSVLHSTQDLFLPFFSNVTWFHAIHVSGT-NGTKRFD         80
```

```
AIV41987.1      DRVYLNTTLFLNGYYPTSGSTYRNMALKGTDKL------STLWFKPPFLSDFINGIFAKV  112
                .                     . .:...        .  *

NP_073551.1     ----RFTTGFVYFNGTGRGDCKGFSSDVLSDVIRYNLNFEENLRRGTILFKTSYGVVVFY  144
MN908947.3      NPVLPFNDGVYFASTEKSNIIRGWI---FG-----T--TLDSKTQSLLIVNNATNVVIKV  130
AIV41987.1      KNTKVFKDGVMYSEFPAITIGSTFVNTSYSVVVQPR--TINSTQDGV--NKLQGLLEVSV  168
                     *. *. : .        :     .          :.   .    :    : :

NP_073551.1     CTNNT----------------------LVSGDAHIPFGTVLGNFYCFVNTTIGNETTSA  181
MN908947.3      CEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGN-----  185
AIV41987.1      CQYNMCEYPHTICHP--KLGNHFKELWHLDTGVVSCLYK---RNFTYDVNATY-------  216
                *  :                        : :.     :      *   :: .

NP_073551.1     FVGALPKTVREFVISRTGHFYINGYRYFTLGNVEAV-----NF-------------NVT-  222
MN908947.3      -----FKNLREFVFKNIDGYFK---IYSKHTPINLVRDLPQGFSALEPLVDLPIGINITR  237
AIV41987.1      ---------LYFHFYQEGGTFY---AYFTDT----------GFV-TKFLFNVYLGMALS-  252
                          * : . .  :     * .           .*              ::

NP_073551.1     ----------------------TAETTDFCTVALASYADVLVNVSQTSIANIIYC-NSVI  259
MN908947.3      FQTLLALHRSYLTPGDSSSGWTA-GAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPL  296
AIV41987.1      --------HYYVMPLTCISRLDIGFTLEYWVTPLTPRQYLLAFNQDGIIFNAVDCMSDFM  304
                                       :  : .  *     :*   .:  * : : *  . :

NP_073551.1     NRLRCDQLSFDVPDGFYSTSPIQSVELPVS-------------------IVSLPVYHKH  299
MN908947.3      SETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRK  356
AIV41987.1      SEIKCKTQSIAPPTGVYELNGYTVQPIADVYRRKPDLPNCNIEAWLNDKSVPSPLNWERK  364
                .. :*   *:    *.*. .                              . *   :.::

NP_073551.1     TFIVLYVDFKP--QSGGGKCFNCYP---AGVN------ITLANFN---ETKGPLCVDT--  343
MN908947.3      RISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTG  416
AIV41987.1      TFSNCNFNMSSLMSFIQADSFTCNNIDAAKIYGMCFSSITIDKFAIPNRRKVDLQLGNLG  424
                 :     : .   .    . *.*     : :    :   .*        :  .

NP_073551.1     ----SH------FTT-----KYVAVYANVGRWSAS------------------------  363
MN908947.3      KIADYNYKLPDDFTGCVIAWNSNNLDSKVGG----NYNYLYRLFRKSNLKPFERDIST--  470
AIV41987.1      YLQSSNYRIDTTATSCQLYYNLPAANVSVSRFNPSTWNKRFGFIEDSVFVPQPTGVFTNH  484
                     :      *     :      .*.

NP_073551.1     ----------INTGNCPFS--FGK-----------------------------------  375
MN908947.3      ------EIYQA--GSTPCNGVEGFN-----------------------------------  487
AIV41987.1      SVVYAQHCFKAPKNFCPCSSCPGKNNGIGTCPAGTNYLTCDNLCTLDPITFKAPDTYKCP  544
                              . *.   *

NP_073551.1     -VNNFVKFGSVCFSLKDIPGGCAMPIVANWAYSKYYTIGSLYVSWSDGDGITGV------  428
MN908947.3      ----------------CYF--------PLQSYGFQPTNGVGYQPYRV------VVLS  514
AIV41987.1      QTKSLVGIGEHCSGLAVKSDYC--------GNNSCTCQPQAFLGWSADSCLQGDKCNIFA  596
                                          ..       :.:.

NP_073551.1     -----------PQPVEGVSSFMNVTLDKCTKYNIYDVSGVGVIRVSNDTFLNGITY--TS  475
MN908947.3      FELLHA--PATVCG---PKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLP-FQQFGRD  568
AIV41987.1      NFILHDVNNGLTCSTDLQKANTEIELGVCVNYDLYGISGQGIFVEVNATYYNSWQNLLYD  656
                              .   ::  . *.:::: .::* *::   * .:         .

NP_073551.1     TSGNLLGFKDVTKGTIYSITPCNPPDQLVVYQQAVVGAM---LSENF-TSYGFSNVV---  528
MN908947.3      IADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQ  628
AIV41987.1      SNGNLYGFRDYITNRTFMIHSCYSGRVSAAYH--ANSSEPALLFRNIKCNYVFNNSLTRQ  714
                  ..  ..:*  .     *  *      .       .     * .:.  .

NP_073551.1     ELPKFFY--------------------ASNGTYNCTDAVLTYSSFGVCADGSIIAVQPR  567
MN908947.3      LTPTWRVYSTGSNVFQTRAGCLIGAEH--VNNSYECDIP----IGAGICASYQTQTNSPR  682
AIV41987.1      LQPI-------NYSFDSYLGCVVNAYNSTAISVQTCDLT----VGSGYCVDYSKNRRSRR  763
                  *                              .   *        . * *.. .    . *

NP_073551.1     NVS-----------------YDSVSAIVTANLSIPSNWTTSVQVEYLQITSTPIVVDCST  610
MN908947.3      RARSVAS-QSIIAYTMS-LGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTM  740
AIV41987.1      AITTGYRFTNFEPFTVNSVNDSLEPVGGLYEIQIPSEFTIGNMEEFIQTSSPKVTIDCAA  823
                                .   .    .: **:::* .   * :  :     :**:

NP_073551.1     YVCNGNVRCVELLKQYTSACKTIEDALRNSARLESADVSEMLTFD-KKAFTLANV-----  664
MN908947.3      YICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQV-KQIYKTPPIKDF--  797
AIV41987.1      FVCGDYAACKLQLVEYGSFCDNINAILTEVNELLDTTQLQVANSLMNGVTLSTKLKDGVN  883
```

```
                ::*.. . *   * :* * *  ::  *       .    ::     :       :

NP_073551.1     SSFGDYNLSSVIPSLPTSGSRVAGRSAIEDILFSKLVTSGLGTVDADYKKCTKGLSIADL          724
MN908947.3      ---GGFNFSQI----LPDPSKPSKRSFIEDLLFNKVTLADAGFIK-QYGDCLGDIAARDL          849
AIV41987.1      FNVDDINFSPVLGCLGSECSKASSRSAIEDLLFDKVKLSDVGFVE-AYNNCTGGAEIRDL          942
                   .. *:* :      . *: : ** ***:**.*:  :. * :.  * .*  .    **

NP_073551.1     ACAQYYNGIMVLPGVADAERMAMYTGSLIGGIALGGLTSA----VSIPFSLAIQARLNYV          780
MN908947.3      ICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGI          909
AIV41987.1      ICVQSYKGIKVLPPLLSENQISGYTLAATSASLFPPWTAAAG----VPFYLNVQYRINGL          998
                 *.* ::*: *** :   : :: ** :  ..      * .      :** : :  *:* :

NP_073551.1     ALQTDVLQENQKILAASFNKAMTNIVDAFTGVNDAITQTSQALQTVATALNKIQDVVNQQ          840
MN908947.3      GVTQNVLYENQKLIANQFNSAIGKIQDSLS--------------STASALGKLQDVVNQN          955
AIV41987.1      GVTMDVLSQNQKLIASAFNNALHAIQQGFD--------------ATNSALVKIQAVVNAN         1044
                .:  :** :***::*  **.*:  * :.:               :. :** *:* *** :

NP_073551.1     GNSLNHLTSQLRQNFQAISSSIQAIYDRLDTIQADQQVDRLITGRLAALNVFVSHTLTKY          900
MN908947.3      AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRA         1015
AIV41987.1      AEALNNLLQQLSNRFGAISASLQEILSRLDALEAEAQIDRLINGRLTALNAYVSQQLSDS         1104
                .::** * .** ..* ***: :: * .*** ::*: *:****.*** :*:.:*:: *

NP_073551.1     TEVRASRQLAQQKVNECVKSQSKRYGFCGNGTHIFSIVNAAPEGLVFLHTVLLPTQYKDV          960
MN908947.3      AEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNF         1075
AIV41987.1      TLVKFSAAQAMEKVNECVKSQSSRINFCGNGNHIISLVQNAPYGLYFIHFNYVPTKYVTA         1164
                : :: *   *  *:.*** .**.* .***:* *::*: : ** *: *:*   :*::

NP_073551.1     EAWSGLCVDGTNGYVLRQPNLALYKEGNYYRITSRIMFEPRIPTMADFVQIENCNVTFVN         1020
MN908947.3      TTAPAICHDGKA-HFPREGV--FVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGI         1132
AIV41987.1      KVSPGLCIAGNRGIAPKSGY--FVNVNNTWMYTGSGYYYPEPITENNVVVMSTCAVNYTK         1222
                 .  .:*  *.     :.    : . .. :  *    : *.  *  : .   .* *

NP_073551.1     ISRSELQTIVPEYIDVNKTLQELSYKLPNYTVPDLVVEQYNQTILNLTSEISTLENKSAE         1080
MN908947.3      VNNTVYDPLQPELDSFKEELDKY---FKNHTSPDVDL-----------GDISGINASVVN         1178
AIV41987.1      APYVMLNTSIPNLPDFKEELDQW---FKNQTSVAPDL-----------S-LDYINVTFLD         1267
                      :   *:  ..:: *::    : * *     :           . :. :: .  :

NP_073551.1     LNYTVQKLQTLIDNINSTLVDLKWLNRVETYIKWPWWVWLCISVVLIFVVSMLLLCCCST         1140
MN908947.3      IQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMT         1238
AIV41987.1      LQVEMNRLQEAIKVLNHSYINLKDIGTYEYYVKWPWYVWLLICLAGVAMLVLLFFICCCT         1327
                ::  :::*:   . :* : ::*: :.  * *:****::** :    : :: : :: ** *

NP_073551.1     GCCGFFSCFASSIRGCCESTKLPYYD-VEKIHIQ--    1173
MN908947.3      SCC-SCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT    1273
AIV41987.1      GCGTSCFKK---CGGCCDDYTGYQELVIKTSHDD--    1358
                .*          .**.        ::  : .
```

The “PIGAG” motif does not show up in the above alignment, as is also the case in many other distantly related coronaviruses [3, 4]. However there is a subsequence PIGTNYRSCESTT in the HCoV-HKU1 spike protein that appears to relate to PIGAGICASYQTQ in the COVID-19 virus (recall that HCoV-HKU1 is a common cold virus, albeit usually associated with more severe, lower respiratory tract cases). In contrast, not only does the KRSFIEDLLFNKV motif stand out as potentially important to the COVID-9 virus by virtue of such comparisons, but also a match with that motif is almost the only continuous stretch of amino acid residues in most alignments like that above. The subsequence KWPWYIWL is an exception that is of interest and a characteristic feature of many SARS coronaviruses. It is not, however, considered further in the present paper, except to note that it does not appear to be associated with a COVID-19 virus spike protein proteolytic cleavage site. These sites are most prominently

Trypsin: S1/S2 HTVSLLRSTSQKSIVAYTMSL, S2' LPDPLKPTKRSFIEDLLFNKV; cathepsin: S1/S2 HTVSLLRSTSQKSIVAYTMSL;
Elastase: S2' LPDPLKPTKRSFIEDLLFNKV,
Plasmin: S1/S2 HTVSLLRSTSQKSIVAYTMSL, S2' LPDPLKPTKRSFIEDLLFNKV,
TMPRSS1: S1/S2 HTVSLLRSTSQKSIVAYTMSL;
TMPRSS2: Multiple sites;
TMPRSS11a: S1/S2 HTVSLLRSTSQKSIVAYTMSL, S2' LPDPLKPTKRSFIEDLLFNKV.

4.3. Variations in the KRSFIEDLLFNKV Motif Across a Broader Range of Coronaviruses.

As one looks out to more distant relatives, there are a number of variations in the KRSFIEDLLFNKV motif which, despite large variations in spike protein sequence as a whole, are still recognizable in the spike proteins of coronaviruses of diverse various host species, as shown for some examples in Table 2. The most noticeable variation is the occasional substitution of the cleavage point arginine (R) by a G. Rather than disrupt the possibility of cleavage, however, it is seemingly displacing that role to a arginine (R) or lysine (K) that lies to the N-terminal (left) side of the motif. It is interesting that this commonly retains firmly the IEDLLF core of the motif.

Table 2
Some Modifications of the RSFIEDLLFNKV Motif in Mammalian Hosts

Motif	Example	Description
RSFIEDLLFNKV	MN908947.3	SARS-CoV-2 and related coronaviruses, especially bat, civet, pig
RSIIEDLLFNKV	AJD09591.1	Porcine epidemic diarrhea virus
RSFFEDLLFDKL	ADX59495.1	Chaerephon bat coronavirus/Kenya/KY22/2006
RSFVEDLLFDKV	APD51483.1	NL63-related bat coronavirus
RSFIEDLLFDKI	YP_009336484.1	Lucheng Rn rat coronavirus
RSVLEDLLFDKI	ASF90465.1	Wencheng Sm shrew coronavirus
RSAIEDLLFNKV	AAP72150.1	Canine Coronavirus
RSAVEDLLFNKV	ADC35472.1	Feline coronavirus
RSAVEDLLFDKV	ABI14448.1	Feline coronavirus
RSAIEDLLFDKV	AIV41987.1	Common cold, also found in the coronaviruses of dogs, cats, rodents, pigs, rabbits, camels, ferret badgers, raccoon dogs, etc.
RSAIEDILFSKL	NP_073551.1	Common cold
RSAIEDLLFSKV	ASV64340.1	Porcine coronavirus (transmissible gastroenteritis of pigs, TGEV).
RSAIEDLLFAKV	ABG89301.1	Porcine TGEV Miller M6
RSAIEDILFSKV	ALK28767.1	229E-related bat coronavirus
RSFFEDLLFDKV	NC_006577.2	Human HCoV-HKU1 "Flu-ish" cold
RKYRSAIEDLLFDKV	ADU17734.1	Canine coronavirus
RKYRSAIEDLLFDKV	BAN67909.1	Feline coronavirus
RKYRSTIEDLLFDKV	BAP19067.1	Feline coronavirus
RKY**G**SAIEDLLFDKV	AAY32596.1	Feline coronavirus
ENK**G**SFIEDLLFDKV	AZF86124.1	Bat-CoV/P.kuhlii/Italy/3398-19/2015

EGK**G**SFIEDLLFDKV	YP_009201730.1	Bat - [BtNv-AlphaCoV/SC2013]
DNR**G**SFIEDLLFDKV	QGX41957.1	Western Australian microbat
VQK**G**SFIEDLLFNKV	AHA61268.1	Porcine epidemic diarrhea virus
VQKRSFIEDLLFNKV	QGA88709.1	Porcine epidemic diarrhea virus

The notion that the KRSFIEDLLFNKV motif overall plays an important role, and presumably a common or similar function across at least a very large number of known coronaviruses, still seems a reasonable one. Most important of course is that it is at least the case for the SARS-CoV-19 virus and its near relatives. At this time, no match with a coronavirus in GeneBank has been detected by the author by BLAST-p using queries with no phenylalanine (F), e.g. RSAIEDLLLDKV, RSAIEDLLIDKV, RSAIEDLLADKV, RSAIEDLLMDKV, RSAIEDLLWDKV, and RSAIEDLLYDKV as queries, but the search has not been exhaustive because it would not be too contradictory to any of the current hypotheses if some were found. In the group with the inserted glycine (G) replacement of initial argine (R) by the similar positively charged lysine (K) is common. However, as long as the motif is significantly recognizable, no histidine (H) as opposed to initial arginine (R) has been found.

4.4. Variations in the KRSFIEDLLFNKV Motif in Avian Coronaviruses

The motif cannot extend to other strains indefinitely as recognizable because at some point the evolutionary tree will bring up virus and hosts subject to quite different selective pressures, and the motif is not the definition of coronaviruses. However, it still persists as recognizable in birds such as duck (e.g GenBank KX266757, KC119407 white-eye bird CoV HKU (NC016991), magpie-robin (shama) CoV HKU18 (NC016993)) strains, a selction which spans a large range of coronavirus genome sizes. See the alignments below compared with the Wuhan seafood market isolate Genbank MN908947.3, showing the motif underlined and in bold.

```
CLUSTAL O(1.2.4) multiple sequence alignment


MN908947.3      -----MFVFLVLLPLVSSQ----------------------------------------    14
KX266757        MLVKSLFLVTLLFALSSASLYD--------------------------------------    22
KC119407.1      MLGKSLLIVTVLFALCSATLYT--------------------------------------    22
KM454473        MLGKSLLIVTVLFALCSATLYT--------------------------------------    22
NC016991        -MQRIILISTILYCARALTLADKMLDLLTFPGAHHYFR---GDLQTLHSRISAESYSVN-    55
NC016993        -MRGAILTLILVTSVKASPLADSVLDFLTFPGAHSYLHPRRGDLGALGNRMRANIRNSQT    59
                     ::   ::    :

MN908947.3      --CVN-------------------------------------------------------    17
KX266757        ----------------------------------------------------------    22
KC119407.1      ----------------------------------------------------------    22
KM454473        ----------------------------------------------------------    22
NC016991        ---------------PYDQYNYQTDSDYYINKSVHLIAPLTNLTLPISGLHRSMQPLRV    99
NC016993        DVCTTIQQGGFIPSTFTFPQWYVLTNGSTFLQGE----------------YTLSQPLLA    102
```

```
MN908947.3      ---LTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTN    74
KX266757        ------------------------NDTY-----------------VYYYQ--------    31
KC119407.1      ------------------------H-DY-----------------VYYYQ--------    30
KM454473        ------------------------H-DY-----------------VYYYQ--------    30
NC016991        GCIFGASNKIDQGFT---ISGMTYPLAYCV-----------------PPFYQ--------   131
NC016993        NAHFCPRKNSDGYWRYSFNNSCLFPDHRCQ-----------------DHWYD--------   137
                                                                 ::.

MN908947.3      GTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNA-------TNVV    127
KX266757        --------------------SAFRPPNGWHLNGGAYAVVNVSSQTNNAGIAPECTVGII    70
KC119407.1      --------------------SAYRPPNGWHLQGGAYAVVNSTNKFNNAGAASECSVGVL    69
KM454473        --------------------SAYRPPNGWHLQGGAYAVVNSTNKFNNAGAASECSVGVL    69
NC016991        --------------------VTNVTYDA--------------------------MRLL    143
NC016993        --------------------SQNPICLGWNNTFGLSDN---------------IRININ    161
                                    .                                 :

MN908947.3      IKVCEFQF-------------------CNDPFLGVYYHKNNKSWMESEFRVYSSANNCT    167
KX266757        SGDTV-----------------FNASSIAMTAPVGQG-----MQW--------SKSQFCT    100
KC119407.1      FNYTN----GNDVGYN------NSASSVAMTAPL-PG-----MSW--------SKTQFCT    105
KM454473        FNYTN----GNDVGYN------NSASSVAMTAPL-PG-----MSW--------SKTQFCT    105
NC016991        FAFADLNSTGDFLRINTKTMGMLNVSCSASPTPLGHQDADR---TF----YGYNKQLYCY    196
NC016993        ISHDEYQSHGGYVSLTLESGSVVNITCTNNSDPSTVTLATSLLPWA----RAIDQPMYCF    217
                                                                 ..   *

MN908947.3      FEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLV    227
KX266757        AHC--------------NFSDITVFVTHCYA------SGAGKCPLTGLIPKGHIRISAMR    140
KC119407.1      AHC--------------NFSDFTVFVTHCFA---------NSCPLTGRIEENHIRVSAMR    142
KM454473        AHC--------------NFSDFTVFVTHCFA---------NSCPLTGRIEENHIRVSAMR    142
NC016991        LDT--------------P--------------------AGMQYMGPLPANLTEITLFR    220
NC016993        ANL--------------T-------------T------GTASQLDFMGMLPPLVSELAFDR    245
                 .                                          :     :

MN908947.3      DL----PIGINITRFQTL-------------LALHRSYLTPGDSSS-----GW-TAGAAA    264
KX266757        NHTLFYNLTVSVSKYPTFKSLQCVDNFTAVYLNGDLVFTSNQTTDVISAGVYFKSGGPIT    200
KC119407.1      NGSLFYNLTVSVSKYPKFKSLQCVNNFTSVYLNGDLVFTSNKTTDVIGAGVYFKAGGPIT    202
KM454473        NGSLFYNLTVSVSKYPKFKSLQCVNNFTSVYLNGDLVFTSNKTTDVIGAGVYFKAGGPIT    202
NC016991        TG--------------------------QIYTNGFHLGTIPSELTYVY----LDKLAFQN    250
NC016993        TG--------------------------GIYINGYRYYLTSALRDVDF---KLKRNDTAE    276


MN908947.3      YYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQP-T    323
KX266757        Y--KVMK-EFKVLAYFVNGTAQDVILCDDTPRGLLACQYNTGNFSDGFYPFTNSSLV--K    255
KC119407.1      Y--KIMK-EFKVLAYFVNGTVQDVILCDNSPRGLLACQYNTGNFSDGFYPFTNFSLV--K    257
KM454473        Y--KIMK-EFKVLAYFVNGTVQDVILCDNSPRGLLACQYNTGNFSDGFYPFTNFSLV--K    257
NC016991        KTVCMMANLTDTLITLNHTVIQQVTYCEKDAVQALACQQSTHQLQDGFYSDPAPAVNNLP    310
NC016993        YFAVTWANYTDVHLSVDAGAIEKIKYCNT-PLDRLACDMNVFNLSDGVYSYTSLEKASVP    335
                              .  .   *         *  .   ...*.*

MN908947.3      ESIVRFPNITNLC----------PFGEVFNATRFAS-VYAWNRKRISNCVADYSV-----    367
KX266757        QRFVVY---RENSVNTTLTLTNYTFHNETNAQPNSGGVYTI-STYQTKTAQSGYYNFNLS    311
KC119407.1      DRFIVY---RESSTNTTLELTNFTFTNVSNASPNSGGVDTF-QLYQTHTAQDGYYNFNLS    313
KM454473        DRFIVY---RESSTNTTLELTNFTFTNVSNASPNSGGVDTF-QLYQTHTAQDGYYNFNLS    313
NC016991        KTLVTLPKIAESSTLQINVSATYSYGSASGSI----KLSYNGSSNNSHCVQTPYFKLEQN    366
NC016993        ETFVTLPVYSNHTYVTINTS--YTVGSCVNCPPISSTIDIMHARNDTLCVNSRQFTVRLN    393
                . ::      :              .  ..      :       :  .

MN908947.3      -----LYNSASF----STFKCYGVSPTKL-----NDLCFTNVYADSFVIRGDEVRQIAPG    413
KX266757        FLSSFVYKESNYMYGSYHPRCSFRPETINNGLWFNSLAVSLAYGP---------------    356
KC119407.1      FLSSFVYKPSDFMYGSYHPNCNFRPENINNGLWFNSLSVSLTYGP---------------    358
KM454473        FLSSFVYKPSDFMYGSYHPNCNFRPENINNGLWFNSLSVSLTYGP---------------    358
NC016991        LVC-----SGGCSVRIETLTCPFDLNAVSNGMSFQQFCVSTVSG----------------    405
NC016993        THHHAQY-PQYFSTAFVAGTCPFTLPNINNYLTFGSVCFSTVNN----------------    436
                              *             ....: .
```

```
MN908947.3      QTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIY    473
KX266757        --------------LQGGCKQSVFQG----RATCCYAYSYN------------------    379
KC119407.1      --------------IQGGCKQSVFSN----KATCCYAYSYR------------------    381
KM454473        --------------IQGGCKQSVFSN----KATCCYAYSYR------------------    381
NC016991        ----------------QCSMQAIVN----TGQ-PWGYV--------------------    422
NC016993        ----------------GGCTIHV--------QK-VWNHQY--------------------    451
                                 *              : :

MN908947.3      QAGSTPCNGVEG------FNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAP-ATVCG    526
KX266757        --GPRMCKGVYSGQLLQDFECGLL-----------VYVTKS---DGSRIQTATKPPVIT    422
KC119407.1      --GPTRCRGVYRGELMQYFECGLL-----------VYVTKS---DGSRIQTRSEPLVLT    424
KM454473        --GPTRCRGVYRGELMQYFECGLL-----------VYVTKS---DGSRIQTRSEPLVLT    424
NC016991        --------------------TST-----------LYVTYV---EGQSFTGT--S-SDQ    443
NC016993        ------------------HTFGT-----------IYVAYQ---DGNYITALPQP-STG    476
                                               *       .  :

MN908947.3      PKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG---RDIADTTD-----AVRD    578
KX266757        QHNYNNITLNTCVDYNIYGRVGQGFITNVTDSAASYNYLADAGLAILDTSGAIDIFVVQG    482
KC119407.1      QYNYNNITLNKCVEYNIYGRVGQGFITNVTEATANYSYLADGGLAILDTSGAIDIFVVRG    484
KM454473        QYNYNNITLNKCVEYNIYGRVGQGFITNVTEATANYSYLADGGLAILDTSGAIDIFVVRG    484
NC016991        IEDLTVLHLDQCTSYTIYGVSGTGVITLSDLQLP-------HGITFRAANGELS--AFKN    494
NC016993        VADISTVHLDVCTKYSIYGKTGTGVIRETNQSYT-------AGLYYTSSSGDLL--AFKN    527
                  . . :  : *.:.: *  * *.:                    :..     ..:.

MN908947.3      PQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYS-    637
KX266757        EYGLNYYKVNPCEDVNQQFVVSGGKL---VGILTSRNETGSQ--PLE-----NQFYIKLT    532
KC119407.1      AYGPNYYKVNPCEDVNQQFVVSGGNL---VGILTSHNETDSE--FIE-----NQFYIKLT    534
KM454473        AYGPNYYKVNPCEDVNQQFVVSGGNL---VGILTSHNETDSE--FIE-----NQFYIKLT    534
NC016991        TTTGDVYTIQPCSLPAQLA-IIDSTI---VGAITSTNE--SY--GFSNTIVTPTFYY---    543
NC016993        VTTQKVYSVTPCTLASQVA-VYNNSI---LAAFTSTANLTAI--DFNYTIATPTFYY---    578
                   .   : **       .  .    :. : .  :       :        :

MN908947.3      TGSNVFQTRAGCLIGAEHVNNSYECDIPI----GAGICASYQTQTNSPRRARSVASQSII    693
KX266757        NGSRR----------LRRSISSNVTICPYVSYGRYCIEPDGSLKQIVPQELQH-------    575
KC119407.1      NGTRR----------SRRSVTENVTNCPYVSYGKFCIKPDGSLSIIVPQELKQ-------    577
KM454473        NGTRR----------SRRSVTENVTNCPYVSYGKFCIKPDGSLSIIVPQELKQ-------    577
NC016991        -----------------STNATSNCTAPKISYGELGVCADGSIGAVSQLQDSK-------    579
NC016993        -----------------HSIGNETCEQPVITYGSIGLCPGGGLRLAHPTEDAA-------    614
                                   *       :  .         .

MN908947.3      AYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLL    753
KX266757        -------FVAPLLNVTEHVLIPNSFNLTVTDEYIQTRMDKVQINCLQYVCGNSIECRKLF    628
KC119407.1      -------FVSPLLNVTEHVLIPNSFNLTVTDEYIQTRMDKVQINCLQYVCGNSLNCRKLF    630
KM454473        -------FVSPLLNVTEHVLIPNSFNLTVTDEYIQTRMDKVQINCLQYVCGNSLNCRKLF    630
NC016991        ------PSIVP--LYTGEIEIPASFKLSVQTEYLQVQTEQVVIDCPKYVCNGNPRCLQLL    631
NC016993        ------PILVP--ISTSNISIPKNFTVSIQTEYIQIEQQPVVVDCRQYVCNGNPRCLQLL    666
                         .  : ** .*.:::  * :      . ::*  *:*... .* :*:

MN908947.3      LQYGSFCTQLNRALTGIAVEQDKNTQEVFA-QVKQIYKTPPIKDFG--GFNFSQILP-DP    809
KX266757        RQYGPVCDNILSVVNSVGQKEDMELLNFYSSTKPKGFDTPVLSNVSTGAFNISLLLT-PP    687
KC119407.1      QQYGPVCENILSIVNSVGQKEDMELLSFYSSTKPAGYNAPVFSNISTGDFNISLLLT-PP    689
KM454473        QQYGPVCENILSIVNSVGQKEDMELLSFYSSTKPAGYNAPVFSNISTGDFNISLLLT-PP    689
NC016991        AQYTSACSNIESALHSSAQLDSREITMMFQ-TSSQSVELANITNFQG---DYNFSMILPT    687
NC016993        QQYTSACSTIEQALSLNARLEASSIQDLLT-YSPETLVLANISNFDSGDLNYNLSSLLPK    725
                 **   *  :   :   .  :  .   .            :.:.     : .

MN908947.3      SKPSK<u>**RSFIEDLLFNKV**</u>TLADAGF-IKQYGDCLGDI--AARDLICAQKFNGLTVLPPLLT    866
KX266757        SSPSG<u>**RSFIEDLLFTSV**</u>ETVGLPT-DAEYKKCTAGPLGTLKDLICAREYNGLLVLPPIIT    746
KC119407.1      SSPRG<u>**RSFIEDLLFTSV**</u>ETVGLPT-DAEYKKCTAGPLGTLKDLICAREYNGLLVLPPIIT    748
KM454473        SSPRG<u>**RSFIEDLLFTSV**</u>ETVGLPT-DAEYKKCTAGPLGTLKDLICAREYNGLLVLPPIIT    748
NC016991        LPGKD<u>**RSAIEDLLFDKV**</u>VTNGLGTVDQDYKSCSKGI--AVADLVCAQYYNGIMVLPGVVD    745
NC016993        E-LYG<u>**KSAIEDLLFNKV**</u>VTNGLGTVDQDYKACTNGM--SIADLVCAQYYNGIMVLPGVAG    782
```

```
                    :* ****** .*   .      :*  *  .   :  **:**: :**: *** :

MN908947.3      DEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQ    926
KX266757        ADMQTMYTASLVGAMAFGG----ITSAAAIPFATQIQARINHLGITQSLLMKNQEKIAAS    802
KC119407.1      ADMQTMYTASLVGSMAFGG----ITAAGAIPFATQIQARINHLGITQSLLLKNQEKIAAS    804
KM454473        ADMQTMYTASLVGSMAFGG----ITAAGAIPFATQIQARINHLGITQSLLLKNQEKIAAS    804
NC016991        AEKMAMYTGSLTGAMVFGG----LTAAAAIPFSTAVQARLNYVALQTNVLQENQKILAES    801
NC016993        PEKMAQYTASLTGAMVFGG----ITAASAIPFSLAVQSRLNYVALQTDVLQQNQQLLADS    838
                 :  : **.:* ..   .*      :*  ***:  :  *:* :.:  .:* :**: :* .

MN908947.3      FNSAIGKIQDSLSSTASA--------------LGKLQDVVNQNAQALNTLVKQLSSNFGA    972
KX266757        FNKAIGHMQEGFRSTSLA--------------LQQVQDVVNKQSAILMETMNSLNKNFGA    848
KC119407.1      FNKAIGHMQEGFRSTSLA--------------LQQVQDVVNKQSAILTETMNSLNKNFGA    850
KM454473        FNKAIGHMQEGFRSTSLA--------------LQQVQDVVNKQSAILTETMNSLNKNFGA    850
NC016991        FNQAVGNISLALSNVNTAIQQTSEALLTVSNAINKIQTVVNQQGEALAHLTAQLSQNFQA    861
NC016993        FNNAIGNITLAFKEVSEGLSQVSGAVATVANALTKVQTVVNEQGHALATLTQQLANNFQA    898
                **.*:*::  .: ..  .             : ::* ***::.  *   .* .** *

MN908947.3      ISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSEC    1032
KX266757        ISSVIQDIYAQLDAIQADAQVDRLITGRLSSLSVLASAKQSEYLRVSQQRELATQKINEC    908
KC119407.1      ISSVIQDIYAQLDVIQADAQVDRLITGRLSSLSVLASAKQSEYIRVSQQRELATQKINEC    910
KM454473        ISSVIQDIYAQLDVIQADAQVDRLITGRLSSLSVLASAKQSEYIRVSQQRELATQKINEC    910
NC016991        ISTSIQDIYNRLDQIQADQQVDRLITGRLAALNAYVTQLLNKLSQVRQSRILAEQKINEC    921
NC016993        ISASISDIYNRLNQLEADAQVDRLITGRLASLNAFVTQTLSKLAEVRQQRQLATDKVNEC    958
                **: :.**  :*: ::*: *:******** :*..::   .  .:  .  **  *:.**

MN908947.3      VLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHF---    1089
KX266757        VKSQSTRYGFCGSGRHVLSIPQNAPNGIVFIHFTYTPESFVNVTAIVGFCVNPPNASQYA    968
KC119407.1      VKSQSNRYGFCGSGRHVLSIPQNAPNGIVFIHFSYTPESFVNVTAIVGFCVQPANASQYA    970
KM454473        VKSQSNRYGFCGSGRHVLSIPQNAPNGIVFIHFSYTPESFVNVTAIVGFCVQPANASQYA    970
NC016991        VKSQSSRYGFCGNGTHLFSLTQAAPNGIFFMHAVLVPQTFQPVVAYAGICVDGYGYS---    978
NC016993        VKSQSPRYGFCGNGTHLFSIVNAAPQGLLFFHTVLLPTQYAYVQAFSGICYNGIALA---    1015
                * .** * .***.* *::*: .**:*:.*:*    *     . :  .:* :

MN908947.3      ----PREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPE-    1144
KX266757        IVPVNDRGVFIQVNGTYYITSRDMYMPRDITAGDIVTLTSCQANYVSVNKTVITTFVDND    1028
KC119407.1      IVPVNSRGIFIQVNGSYYITARDMYMPRDITAGDIVTLTSCQANYVNVNKTVITTFVEDD    1030
KM454473        IVPVNSRGIFIQVNGSYYITARDMYMPRDITAGDIVTLTSCQANYVNVNKTVITTFVEDD    1030
NC016991        --L-QPQLVLYNLNDSYRITPRNMFEPRTPTQSVFIPLTTCSVDFVNVTANNVSIIIPD-    1034
NC016993        --LNDPTLALFKNGDKYLVSPRNMYQPRVPAQADFVYIETCTITYLNLTDLTIDVVIPD-    1072
                      : . .  : :: *::: *:  :    .   .*      :.      .  :

MN908947.3      LDSFKEELDKYFKN-------HTSPDVDL-----------GDISGINASVVNIQKEIDRL    1186
KX266757        DFDFYDELSKWWNDTKHELPDF-----DEFNYTIPVLNISN--------------EIDRI    1069
KC119407.1      DFDFDDELSKWWNDTKHELPDF-----DDFNYTVPILNISG--------------EIDRI    1071
KM454473        DFDFDDELSKWWNDTKHELPDF-----DDFNYTVPILNISG--------------EIDRI    1071
NC016991        YVD----VNKTVSDIINGLPNYSYPELSLDRFNHTILNLSQEIEDLQIRSQNLSATAELL    1090
NC016993        YVD----VNQTVNDILSKLPNSTGPSLTIDQYNNTILNLTTEIADLNNRTQNLSDVVQNL    1128
                  .    :.:  .:                                          : :

MN908947.3      NEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKG    1246
KX266757        QEVIQGLNDSLIDLETLSILKTYIKWPWYVWLAIAFAVIIFILILGWVFFMTGCCG----    1125
KC119407.1      QGVIQGLNDSIINLEELSIIKTYIKWPWYVWLAIGFAIIIFILILGWVFFMTGCCG----    1127
KM454473        QGVIQGLNDSIINLEELSIIKTYIKWPWYVWLAIGFAIIIFILILGWVFFMTGCCG----    1127
NC016991        QQYIDNLNNTLVDLEWLNRVETYLKWPWYIWLLIFLAIAAFATILVTIFLCTGCCGGCFG    1150
NC016993        EEYIHKLNATLVDLDWLNRVETYIKWPWWVWLLITLAIVAFVVILVTIFLCTGCCGGCFG    1188
                :   . ** ::::*: *.  : *:****::** :  .:  :  :   :   *.**.

MN908947.3      CC-SCGSCCKFDEDDSE-PV-LKGVK-----LHYT-----    1273
KX266757        CCCGCFGIIPLMSKCGKKSSYYTTFDNDVVTEQYRPKKSV    1165
KC119407.1      CCCGCFGIIPLMSKCGKKSSYYTTFDNDVVTEQYRPKKSV    1167
KM454473        CCCGCFGIIPLMSKCGKKSSYYTTFDNDVVTEQYRPKKSV    1167
NC016991        CCGGCFGLFSKKRRLSSEPT-PVSFK---------LKEW-    1179
```

```
NC016993        CCGGCFGLFSHNKRNTESIP-ITSFK---------LKEW-    1217
                ** .* .          .       ..
```

4.5. Traces of the KRSFIEDLLFNKV Motif in Nidoviruses of Reptiles and Fish.

Some indication of the limit of the survival of the motif RSFIEDLLFNKV as the researcher departs from SARS-CoV-19 might be given by the nidoviruses other than coronaviruses. Somewhat coronavirus-like nidoviruses are common as e.g. reptile viruses. The order Nidovirales contains enveloped, positive-strand RNA viruses with the largest known RNA genomes. Nidoviruses have been identified in snakes. They appear to be most closely related to coronavirus subfamily Torovirinae, and might be best represented as a genus in this subfamily. Sequences suggestive of RSFIEDLLFNKV, e.g. KNFIDLLLAGF do occur in genomes such as the ball python genome, but these really lie beyond the limit of serious detection. For example, Clustal Omega gives 18% exact match between the Wuhan isolate and spike protein nidovirus 1 of the reptile shingleback, but the motif is barely recognizable.

```
MN908947.3     NRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDL--       821
YP_009666261.1 VQS-----------------IAQILETEPLPST-KLDFRTEENNVT-KITLSFTQEVAS       560
                ::                 : ** :* *: .  ::*     : :   . ** :::
```

Including fish nidovirus of the Pacific salmon (GenBank QEG08239.1) is notable here because it supports the above alignment because it is preserved, but GTLYWLDY of the salmon nidovirus is far from KRSFIEDL and the nearest preceding plausible cleavage point is an arginine (R) 10 residues in the N-terminal direction (to the left). However, a similar occurs in some mammalian coronaviruses and so that residue may still play a similar role as an activation cleavage.

```
QEG08239.1          VLPQTYATAMLTRFIPPPVSISGTLYWLDYPDVFV--YSGNVAFDQPT-----------  817
MN908947.3          ILPD------------PSKPSKRSFIEDL-------LFNKVTLAD--AGFIKQYGDCLG  842
YP_009666261.1      EENN------------VT-KITLSFTQEVASTLTQRTINSKQLATPKLNQLKAWYQMTK  588
                       :                 :  :         ..  :
```

4.6. Tentative Matches of the KRSFIEDLLFNKV Motif with Human Proteins.

Looking for similar motifs in human proteins has a somewhat different motive. It makes sense in that, if there is significant match with subsequences, they might represent features of proteins to which both the spike protein and other human proteins may bind, irrespective of any other justification for commonality. Even if coincidental, as epitopes similar to those in a proposed synthetic vaccine they are always of possible interest in assessing the risk of cross-reaction and inducing autoimmunity in synthetic vaccine designs, and on certain occasion with peptidomimetics that induce an immune response, perhaps by a binding strongly to a human protein that the designer did not intend. As discussed in ref [3], there is a motif match at 56% identity with 77%

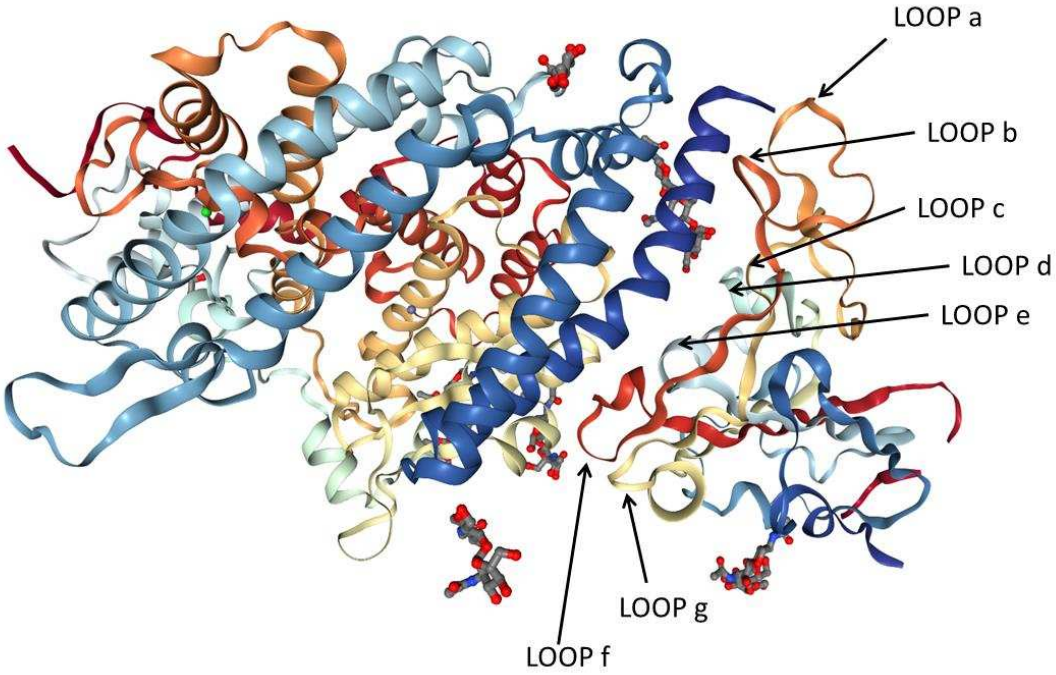
coverage is with tumor protein D55 isoform 2 [Homo sapiens], ID: NP_001001874.2, and similarly with Tumor protein D52-like 3 [Homo sapiens] ID: AAH33792.1. Next match is in regard to neprilsyn entries at only 56% match and 55% coverage. None of these are sufficient close of concern regarding induction of an autoimmune response. Some fairly close matches of KRSFIEDLLFNKV and of the "A for F" modified motif RSAIEDLLFDKV have come to light that might plausibly have a biological significance if supported by biological relevance, but are more likely to be random matches. Selecting only for human proteins, hits vary from 100% cover with 50% identity to 62% cover with 92% identity. These hits cannot be considered significant for peptides of this length in isolation from other evidence. However, a few seem worth recording for future reference in regard a potential biological function for the virus. As already noted [3], RRSFIDELAFGRG a section of a human semaphorin (GenBank NP_001243276.1) produced in response to lung disorders. Running RSAIEDLLFDKV itself in BLASTP generates 100 coronavirus hits. RNAREELLFD is found in human MHC class II antigen, GenBank AXN55588.1. RNAREELLFD is found in human immunoglobulin heavy chain junction region GenBank MCG49633.1. DLLFEKV is found in human tubulin, gamma complex associated protein 6, isoform CRA_d GenBank EAW73510.1. E3 is of interest with 75% identity 87% matches for SFLEELLFin KHKSFLEELLF in ubiquitin. The cellular E3 ubiquitin ligase ring-finger and CHY zinc-finger domain-containing 1 (RCHY1) have been identified as interacting partners of the viral SARS-unique domain (SUD) and papain-like protease (PL^{pro}), with the involvement of cellular p53 as antagonist of coronaviral replication. Down-regulation of p53 is a major player in antiviral innate immunity [72]. Again, however, these matches remain tenuous. GenBank has of the order of 0.2 billion nucleic acid sequences but a 13 residue peptide can have 81,920,000 billion sequences.

4.7. Are Spike Glycoprotein ACE2 Binding Region Features Well Conserved?

While the KRSFIEDLLFNKV motif remains favored by the author as a target at this time, identifying the amino acid residues in ACE2 and the spike protein is important. It may for example involve conserved residues that are not together in a continuous sequence. While a conserved run of amino acid residues is sufficient to be on the list of candidates for an important site, important sites are not necessarily conserved runs of amino acid residues. Here is shown that there is some conservation, but significant variation compared with RSFIEDLLFNKV. Subsequences RSFIEDLLFNKV and PIGAGICASY…R discussed in ref [3] as motifs associated with activation cleavage sites do not lie in the receptor (ACE2) binding domain of the SARS-Cov-2 Spike glycoprotein. The relationships between the whole spike protein and the receptor binding domains in PDB entries 6M17 and PDB 6VW1 are shown in the alignment below. Note that the above receptor binding domain precedes the above motifs in the sequence. A three dimensional perspective is required for an appreciation of the

important sequence features. In Fig. 2, the PDB 6VW1 binding domain is on the right, bound to ACE2 on the left.

Fig. 2.
Structure of ACE2 Interacting with Spike Glycoprotein Receptor Domain (Protein Data Bank Entry 6VW1)



Of course, not all the receptor binding domain is interacting intimately with ACE2. The sections of the receptor binding domain that do interact with ACE2 are also shown (underlined). To facilitate deeper analysis, the loops on the spike protein receptor binding domain were initially classified as loops a,b,c,d,e, and f in order of visual perspective, then joined into three subsequences 1, 2, 3 that contain these loops. The part of the spike glycoprotein sequence that represents the receptor (ACE2) binding domain can be shown by considering the proteins used in the two structural determinations 6M17 and 6VW1 in the Protein Data Bank, shown below in an alignment made using Clustal Omega alignment. Note that the above receptor binding domain precedes the above motifs in the sequence.

```
CLUSTAL O(1.2.4) multiple sequence alignment
WuhanSeafood      MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFS  60
PDB6M17           ------------------------------------------------------------  0
PDB6VW1           ------------------------------------------------------------  0

WuhanSeafood      NVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIV  120
PDB6M17           ------------------------------------------------------------  0
PDB6VW1           ------------------------------------------------------------  0
```

```
WuhanSeafood      NNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLE  180
PDB6M17           ------------------------------------------------------------  0
PDB6VW1           ------------------------------------------------------------  0

WuhanSeafood      GKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQT  240
PDB6M17           ------------------------------------------------------------  0
PDB6VW1           ------------------------------------------------------------  0

WuhanSeafood      LLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETK  300
PDB6M17           ------------------------------------------------------------  0
PDB6VW1           ------------------------------------------------------------  0

WuhanSeafood      CTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISN  360
PDB6M17           ------------------RVVPSGDVVRFPNITNLCPFGEVFNATKFPSVYAWERKKISN  42
PDB6VW1           ------------------RVVPSGDVVRFPNITNLCPFGEVFNATKFPSVYAWERKKISN  42
                                    ** *: .:*******************:* *****:**:***
                                                                LOOP 1e     LOOP 1d
WuhanSeafood  361 CVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIAD  420
PDB6M17           CVADYSVLYNSTFFSTFKCYGVSATKLNDLCFSNVYADSFVVKGDDVRQIAPGQTGVIAD  102
PDB6VW1           CVADYSVLYNSTFFSTFKCYGVSATKLNDLCFSNVYADSFVVKGDDVRQIAPGQTGVIAD  102
                  **********: ********** ********:********.:**:********** ***
                   LOOP 1c             LOOP 2g                        LOOP 3a
WuhanSeafood  421 YNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPC  480
PDB6M17           YNYKLPDDFMGCVLAWNTRNIDATSTGN-------------KPFERDISTEIYQAGSTPC  149
PDB6VW1           YNYKLPDDFMGCVLAWNTRNIDATSTGNYNYKYRLFRKSNLKPFERDISTEIYQAGSTPC  162
                  ********* ***:***:.*:*:.  **             ******************
                       LOOP 3b  LOOP 3f
WuhanSeafood  481 NGVEGFNCYFPLQSYGFQPTKGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVN  540
PDB6M17           NGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLP-------------KLSTDLIK-----  191
PDB6VW1           NGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLNAPATVCGPKLSTDLIK-----  217
                  ******************:***********             * **:*:*

WuhanSeafood      FNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP  600
PDB6M17           ------------------------------------------------------------  191
PDB6VW1           ------------------------------------------------------------  217

WuhanSeafood      GTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY  660
PDB6M17           ------------------------------------------------------------  191
PDB6VW1           ------------------------------------------------------------  217

WuhanSeafood      ECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI  720
PDB6M17           ------------------------------------------------------------  191
PDB6VW1           ------------------------------------------------------------  217

WuhanSeafood      SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE  780
PDB6M17           ------------------------------------------------------------  191
PDB6VW1           ------------------------------------------------------------  217

WuhanSeafood      VFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC  840
PDB6M17           ------------------------------------------------------------  191
PDB6VW1           ------------------------------------------------------------  217

WuhanSeafood      LGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM  900
PDB6M17           ------------------------------------------------------------  191
PDB6VW1           ------------------------------------------------------------  217

WuhanSeafood      QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALN  960
PDB6M17           ------------------------------------------------------------  191
PDB6VW1           ------------------------------------------------------------  217

WuhanSeafood      TLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA  1020
PDB6M17           ------------------------------------------------------------  191
PDB6VW1           ------------------------------------------------------------  217
```

```
WuhanSeafood      SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA  1080
PDB6M17           ------------------------------------------------------------  191
PDB6VW1           ------------------------------------------------------------  217
```

The amino acids residues in bold and underlined font are the subsequences of ACE2 that interact with the above spike protein ACE2 binding domain loops, which are indicated above each subsequence. These include some longer range electrostatic interactions and potential solvent effects. Those also in italics ***DKFNHEAEDLFY***, ***DKFNHEAEDLFY*** and ***KGDFR*** have particularly strong interactions.

```
LOOP 3a 3b 1c 1d 1e 3f                                  LOOP 3f 2g LOOP 1d
STIEEQAKTFLDKFNHEAEDLFYQSSLASWNYNTNITEENVQNMNNAGDKWSAFLKEQSTLAQMYPLQEIQNLTVKLQLQ
ALQQNGSSVLSEDKSKRLNTILNTMSTIYSTGKVCNPDNPQECLLLEPGLNEIMANSLDYNERLWAWESWRSEVGKQLRP
LYEEYVVLKNEMARANHYEDYGDYWRGDYEVNGVDGYDYSRGQLIEDVEHTFEEIKPLYEHLHAYVRAKLMNAYPSYISP
                                                        LOOP 2f 3G
IGCLPAHLLGDMWGRFWTNLYSLTVPFGQKPNIDVTDAMVDQAWDAQRIFKEAEKFFVSVGLPNMTQGFWENSMLTDPGN
          LOOP 3f
VQKAVCHPTAWDLGKGDFRILMCTKVTMDDFLTAHHEMGHIQYDMAYAAQPFLLRNGANEGFHEAVGEIMSLSAATPKHL
KSIGLLSPDFQEDNETEINFLLKQALTIVGTLPFTYMLEKWRWMVFKGEIPKDQWMKKWWEMKREIVGVVEPVPHDETYC
DPASLFHVSNDYSFIRYYTRTLYQFQFQEALCQAAKHEGPLHKCDISNSTEAGQKLFNMLRLGKSEPWTLALENVVGAKN
MNVRPLLNYFEPLFTWLKDQNKNSFVGWSTDWSPYAD
```

As a reference perspective, the *full* sequence for ACE2 as angiotensin-converting enzyme 2 isoform X1 [Homo sapiens] GenBank entry XP_011543851.1, is as follows. The part in the three dimensional structure above is in bold underlined font.

```
MSSSSWLLLSLVAVTAAQSTIEEQAKTFLDKFNHEAEDLFYQSSLASWNYNTNITEENVQNMNNAGDKWSAFLKEQSTLAQMYPLQEIQNLTV
KLQLQALQQNGSSVLSEDKSKRLNTILNTMSTIYSTGKVCNPDNPQECLLLEPGLNEIMANSLDYNERLWAWESWRSEVGKQLRPLYEEYVVLK
NEMARANHYEDYGDYWRGDYEVNGVDGYDYSRGQLIEDVEHTFEEIKPLYEHLHAYVRAKLMNAYPSYISPIGCLPAHLLGDMWGRFWTNLY
SLTVPFGQKPNIDVTDAMVDQAWDAQRIFKEAEKFFVSVGLPNMTQGFWENSMLTDPGNVQKAVCHPTAWDLGKGDFRILMCTKVTMDD
FLTAHHEMGHIQYDMAYAAQPFLLRNGANEGFHEAVGEIMSLSAATPKHLKSIGLLSPDFQEDNETEINFLLKQALTIVGTLPFTYMLEKWRWM
VFKGEIPKDQWMKKWWEMKREIVGVVEPVPHDETYCDPASLFHVSNDYSFIRYYTRTLYQFQFQEALCQAAKHEGPLHKCDISNSTEAGQKLF
NMLRLGKSEPWTLALENVVGAKNMNVRPLLNYFEPLFTWLKDQNKNSFVGWSTDWSPYADQSIKVRISLKSALGDKAYEWNDNEMYLFRSSV
AYAMRQYFLKVKNQMILFGEEDVRVANLKPRISFNFFVTAPKNVSDIIPRTEVEKAIRMSRSRINDAFRLNDNSLEFLGIQPTLGPPNQPPVSIWLIVF
GVVMGVIVVGIVILIFTGIRDRKKPTPLLGKSWLTAILKD
```

Note that at least in these particular experimental structures there is an involvement of “glycosylation-like” molecules. For example, in 6VW1 there are well localized N-acetyl-D-glucosamine, β-D-mannose, and 1,2-ethanediol molecules that make significant interactions in a glue-like manner, and essentially “glue around the edges”. However there no obvious indication of involvement of glycosylation in the main interior interaction face of the complex. The intimate interactions are protein-protein, i.e. between amino acid residues.

Primarily, but not solely, ACE2 and spike glycoprotein association involves interaction between the bent α-helix residues 19-54 (STIEE…NYNTN) of ACE2 and an extended chain configuration, effectively a stretched loop, that runs from residue 485-500 (GFNCY….YGQPT) of the spike glycoprotein and involves or ends in loops 3a, 3b, and 3f. In the case of ACE2 interacting with the ACE2 binding domain of the spike glycoprotein, one could in principle imagine blocking the ACE2 as receptor with a mimic

of the spike protein surface, or blocking the receptor binding site of the spike glycoprotein protein with a mimic of the ACE2 receptor surface. In other kinds of infection the former is usually considered more plausible, but the latter would not interfere with normal function of ACE2 and it is of course essentially the way in which immune system, and notably antibodies, work. Possibly the main argument against this second choice it is that it is essentially equivalent to using antibodies raised against the spike protein, i.e. in effect, passive immunization.

S**T**IE**EQ**A**KTF**L***DKF***N***HE***A***EDL***F***YQS***S**LA**S**WN**
G F N C Y F P L Q S Y G F Q P T

At the time of final writing, various news articles are drawing attention to potential use of the upper sequence or parts of it, which is that of the α-helix of ACE2 (for example https://scitechdaily.com/mit-chemists-have-developed-a-peptide-that-could-block-covid-19/). In the above "alignment", the helix contains 35 residues and the extended chain below contains 16. They have similar length as is as expected for such structures. An α-helix has a rise of 1.5 Å per residue along its axis and there are in typical protein helices with turn variations that imply up to 2.0 Å. The β-strand or a similar extended chain in the spike glycoprotein that interacts with it has a rise of circa 3.5 Å per residue. This general geometry naturally puts the two sequences above in roughly the one-to-one spatial correspondence shown. Note that this is not intended to be a sequence match representation; these chains have to interact. In that respect, there is a lack of charged residues (acidic and basic sidechains) in the extended chain of the spike glycoprotein in structure 6VW1, although an aspartic acid (D) replaces the serine (S) in some strains, arginine (R) replaces asparagine (N) in others, and so on (e.g. see BLASTp alignments later below). A detailed backbone view is confusingly cluttered, but one may identify residues that interact at the boundary between ACE2 and spike protein. All sidechains in the above spike protein subsequence GFNCYFPLQSYGFQPT either make close contact or are likely to have some influence at the interface. It is perhaps useful to have the initial mental picture that, very roughly speaking, the planes of the peptide groups are tangential to the above α-helix surface, rather than constituting an extended chain that makes an edge-on approach. As even Fig. 2 suffice to makes clear, however, the extended chain, like any so-called extended chain in proteins in practice, is essentially a visible helix of larger pitch, resembling a very stretched-out α-helix, and is itself slightly supercoiled to wrap around the ACE2 α-helix. In this case, this tends to follow the elbow or bend in the α-helix, staying roughly parallel to the local axis of the α-helix, so as to make intimate contact overall.

4.8. Consideration of the ACE2 Binding Region for Synthetic Vaccine and Antagonist Design.

If GFNCYFPLQSYGFQPT is to be use as an epitope analogue, the cysteine (C) may be tested as a convenient linker to a carrier protein, otherwise replaced by serine (S) as a close analogue. As far as peptide antagonists are concerned, the difficulty with using the above sequences STIEE…. and GFNCY…. is that they are readily degraded by host proteases. This would not occur if the peptide is made entirely of D-amino-acid restudies. A retro-inverso peptide [3, 46] is made up of D-amino acids in a reversed sequence to the subsequence which is seeks to mimic, and in the extended conformation assumes a side chain topology similar to that of the original native peptide, but with backbone N-H and carbonyl C=O groups interchanged.

(NH_3^+)-dextro-[NWSALSSQYFLDEAEHNFKDLFTKAQEEITS] -(COO^-)

(NH_3^+)-dextro-[TPQFGYSQLPFYSNFG]-(COO^-)

These are peptidomimetics of the subsequences sequences STIEE…. In ACE2 and GFNCY….. in the spike glycoprotein respectively. The cysteine (C) in ….CNFG in the second molecule may be a convenient linker for an epitope for a vaccine but should be replaced by serine (S) in an antagonist. Recall that the problem of having the backbone amide N-H and carbonyl C=O groups interchanged is that if, in the original section of backbone being mimicked, any N-H and C=O groups form a hydrogen bond with recipient and donor groups in the protein, those hydrogen bonds are now disrupted in the intended competitive antagonsist, e.g. they would be unstable N-H…H-N or C=O…O=C interactions. It would thus seem a significant advantage in using the ACE2 mimic, because that is essentially an α-helix which uses up its backbone amide and carbonyl groups. However, retro-inverso α-helices are not typically found in the areas that have shown some degree of success [47], such as antigenic mimicry. It would nonetheless seem to be of value to test both of the retro-inverso peptides in laboratory studies.

As to developing the above further both as the basis of a synthetic vaccine, or as a peptidomimetic, and as to the worth of extending the studies to small organic drug molecules, everything in the above depends on the extent to which GFNCYFPLQSYGFQPT can produce escape mutations which might soon rend such solutions useless. As in the previous paper, we can relate this to variations of the above sequence bother in closer and much more distant relatives. As the following shows, using BLASTp we do not have to go very far from SARS-CoV-2 to find matches with only part of this sequence (coverage) and differences within that area of partial match:

In the original Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, GenBank ID MN908947.3, the subsequence in this regions is FNCYFPLQSYGF, and the following are examples of coverage as found by BLASTp.

NCYFPLRGYGF – Wuhan seafood market isolate Genbank ID: MN908947.3
NCYWPLRGYGF - [SARS coronavirus C028] Civet GenBank ID: AAV98001.1
NCYWPLKGYGF - [SARS coronavirus PC4-137] palm civet GenBank ID:AAV49720.1
NCYWPLNGYGF - [SARS coronavirus CS21] Civet GenBank ID: ABF68958.1
NCYWPLNDYGF - [Bat SARS-like coronavirus] Bat GenBank ID: ATO98218.1
NCTWP----GF - [Feline coronavirus] Feline GenBank ID: AMD11161.1
NCYP---AGVN - [Human common cold coronavirus 229E] GenBank ID: NP_073551.1
TCNNIDAAKIY – [Human common cold coronavirus OC43] GenBank ID: AIV41987.1

The last of the above BLASTp at https://blast.ncbi.nlm.nih.gov/Blast.cgi match results differs in total alignment by Clustal Omega at https://www.ebi.ac.uk/Tools/msa/clustalo/, as follows, but of course this illustrates the high degree of variation that occurs as one proceeds on to coronaviruses less related to the Wuhan seafood market isolate that is believed to be associated with the origin of COVID-19.

```
MN908947.3       RDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVC   525
NP_073551.1      -------------PIVANWAYSKYYTIGSLYVSWSDGD-----------GITGVPQPVE   433
                             *  .  .:. *: : *  .. ::*            :  .*  *
```

For completeness, note that the alignment obtained in the region that BLASTp indicated is as follows.

```
MN908947.3       YNSASFSTFKCYGVSP------------TKLNDLCFTNVYADSFVIRGDEVRQIAPG--   413
NP_073551.1      PQSGGGKCFNCYPAGVNITLANFNETKGPLCVDTSHFTTKYVAVYANVGRWSASINTGNC   369
                  :*.. . *:** ..               ::   **. *.  :.  *    .*  *
```

Phenylalanine (F) commonly immediately precedes many of these matching subsequences NCYFP… NCYWP… etc., and the conservative substitution tryptophan (W) substitution for the second phenylalanine (F) is also very common, so it may be worth noting that FNCTWP is a subsequence in the mammalian vomeronasal type-2 receptor 1 on sensory cells within the main nasal chamber that detects heavy moisture-borne odor particles, and FNCTWP is also found in in dynein. Many viruses require the minus-end–directed dynein motor complex transport on microtubules from cell surface toward the nucleus, and dynein in addition to kinesins for the transport toward the plasma membrane. However a direct connection to viral infection, while tempting, is far from obvious as to any mechanistic or evolutionary explanation. Also, dynein nuclear shuttle transport may be less relevant to the coronavirus (an RNA virus), but certainly RNA viruses can rely on the dynein system (e.g. hanta virus uses it for endoplasmic reticulum-Golgi intermediate compartment). At very least, the above illustrates the kinds of further, perhaps immediately less obvious, functions that the above ACE2 binding domain of the spike glycoprotein, and the above motif, might have.

Within the coronaviruses, there is some degree of conservation that suggests that NCYWPLNDYGF is a segment for the virus to conserve, and a hint that FNCTWPGF is the key part, but there are soon very clearly significant variations across

coronaviruses of different hosts as we depart from the Wuhan seafood market isolate compared with the RSFIEDLLFNKV motif in the S2' cleavage regions [3]. Small organic drugs design to mimic this section, or simply designed to designed to antagonize ACE2 binding, are thus potentially susceptible to escape mutations, i.e. rapid appearance of drug resistance.

4.9. Binding Studies with 11β-hydroxysteroid dehydrogenase type 1 as Model Pharmacaphore.

11β-hydroxysteroid dehydrogenase type 1, which is inhibited by emodin, was an interim model pharmacophore of choice [3]. At this point in the development of the argument for optimal targets for vaccines and therapeutic antagonists, the above target fits in as follows. While the above regarding ACE2 binding must be kept in mind for antagonist development, as noted above the motif is not well conserved, and so could be prone to development of escape mutations, i.e. acquired resistance to vaccines and therapeutic antagonists. Because of the dominant theme of an ACE2 α-helix interacting with an extended chain loop of the spike glycoprotein, the structure of the interaction region is fairly easy to deduce for various SARS strains, and there was as yet no obvious strongly recurrent theme of significant conserved residues that are discontinuous (i.e. not together in the same subsequence) that could be interacting closely with ACE2. At the same time, while emodin appears to act at the ACE2 binding site [59], it remains of interest because there are complexities [60, 61] as discussed in Introduction Section 1.6. Notably, the ACE2 binding domain of the spike protein and the binding sequence discussed above might bind other human proteins and might have other functions that emodin and related compounds, related in the sense that they are at least consistent with pharmacophore features, might inhibit. A priori, the binding properties of emodin and the choice of 11β-hydroxysteroid dehydrogenase type 1 as model pharmacophore could equally relate to the RSFIEDLLFNKV site, or some other site, or a mix of several. The case for interaction vomeronasal type-2 receptor 1 and dynein discussed above was at best marginal, but these examples illustrated the diversity other kinds of functions, important to the virus, that might apply. In any event, any relations between emodin and similar and potentially related molecules remains of interest to impeding SARS-CoV-2 entry and the worse casualty would be the continuity of the story developed above, which is intended to illustrate a flow of reasoning in using the standard tools of bioinformatics.

11β-hydroxysteroid dehydrogenase type 1 is interesting as accommodating a great variety of ligands at the steroid binding site, but not without a degree of specificity as to general features of the ligands, and so far these resemble those of potential SARS-CoV-2 therapeutics. Keeping in mind the *refutation prin*ciple [3] that a pharmacophore (or contribution to a pharmacophore ensemble) the dehydrogenase is worthy of use

until a new ligand or other information proves otherwise. So far *pharmacophore validation* here, i.e. a demonstration that it is a suitable pharmacophore model until proven otherwise, has been based circumstantially on emodin and compounds looking chemically similar to it, that are known in practice or argued theoretically to interact with SARS virus entry in some way and bind at least weakly, experimentally or computationally, to 11β-hydroxysteroid dehydrogenase type 1 [3]. A review of compounds that are known experimentally to inhibit the dehydrogenase and known experimentally inhibit coronavirus entry, replication and maturation is being prepared. However, validation is also extensively based on a weaker but larger body of preliminary binding studies involving a variety of antagonists of coronavirus infection and very often other kinds of virus infection, that also bind *at least* very weakly to the dehydrogenase (see discussion on "very weakly" below). Most of these, emodin-like and otherwise, were first found by Q-UEL knowledge gathering tools as used in the initial coronavirus study [3] combined with "very early candidate selection rules" based on estimates of the mean binding strength of groups when binding well. Note that a hydrogen bond worth about -4 kcal/mole is nonetheless effectively zero when binding well, because it is relative to binding to water. In contrast aromatic and large aliphatic and are worth circa -3 kcal/mole due to hydrophobic interactions which depend on being considered relative to water. There are more complex electrostatic and intramolecular entropic considerations beyond present scope, noting that at least preliminary study of the interaction with 11β-hydroxysteroid dehydrogenase type 1 is the arbitrator. Weak and very weak candidates are also considered because there may be multiple binding modes that will take a great deal of computer time to explore but which could yield lower binding fre energies.

This produces a fairly "mixed bag" of compounds, based on the argument that viruses and coronavirus in particular may use each of its limited number of exposed or exposable sites for several purposes, and the coronavirus seems to be able to readily adjust to new mechanisms under the selective pressure of drugs and vaccines. The details of these molecules and studies are the subject of a further paper that will also discuss some interesting unifying themes. Briefly, they include many names as hoped-for drugs against the coronavirus that appear in the news and Internet discussion. It is convenient to see them as dividing into three classes

(i) Quinone-like. A "quinone" is any of a class of aromatic compounds having two carbonyl or ketone C=O functional groups in the same six-membered ring, though in "quinone-like" the author includes include many compounds resembling steroid fragments that may have many or just one carbonyl groups and several rings. This group includes 9,10-anthraquinone and derivatives that relate to many important drugs some with suggestive laxative and antiinflammatory functions, collectively called anthracenediones. They include ubiquinone as coenzyme Q, and various shorter aliphatic chain forms hydroxyl-decyl-ubiquinone and shorter aliphatic chain forms, laxatives such as dantron, emodin, and aloe

emodin, and some of the senna glycosides, antimalarials such as rufigallol, antineoplastics used in the treatment of cancer, such as mitoxantrone, pixantrone, and the anthracyclines. Caution is reuired in reading this list as a list of potential therapeutics, because anthraquinone derivatives rhein, aloe emodin or anthrone that lacks the methyl group, parietin (physcion), to some extent emodin itself, and chrysophanol extracted from *Cassia occidentalis* are toxic and known to cause hepatomyoencephalopathy in children. It is a medical term effectively defined to cover lethargy, jaundice, and altered senses of children in India after consumption of Cassia seeds.

(ii) Steroid-like. This group includes some plant steroid-like compounds such as carbenoxolone itself from liquorish (licorice) and others found in soy and sprouts.17β-estradiol (the endogenous ligand responsible for the growth and development of many tissues) diethylstilbestrol (a synthetic estrogen); 7-methyl-benz[*a*]anthracene-3,9-diol (a possible natural product from a common polyaromatic hydrocarbon) is also of interest. This group resembles group (i), but the concern for this group (ii) is that molecules like emodin that are known to antagonize viral or other infections are generally smaller, so it possible that a more relevant pharmacophore would sterically exclude a large steroid-like ring.

(iii) Quinine-like. These should not be confused with "quinone-like". Quinine is an alkaloid derived from cinchona bark, used to treat malaria and as an ingredient of tonic water. A common feature is, nonetheless the abundance of aromatic and other rings that in the quinine-like case include nitrogen, so variously resembling pyrimidines, purines, histidine and tryptophan. This group is of current considerable interest as potential thereputics for COVID-19. Of particular interest are Chloroquine, Theophylline, Tavipiravir, Baloxavir marboxil. Some ACE and ACE2 inhibitors can be classified in this group. They are weak but not very weak binders as discussed later below. Camostat, a serine protease inhibitor that has been considered as a potential therapeutic for COVI-19 is convenient to place in this class because of its analogues but it does not itself include a nitrogen atom within a ring.

There are several possible intriguing biological connections that will be discussed elsewhere. One might be briefly mentioned when considering combined therapeutic use of a member of each set. Ubiquinone-like compounds can inhibit ubiquinone sites that work in concert with NADH and NADPH cofactor sites. The latter in turn are often inhibited by the quinine-like members.

Many other above compounds generally bind "very weakly", though steroid-like compounds are strong binders and many quinine-like compounds are medium binders: these are discussed below. Binding strength is of course a matter of degree. RT (where R is the gas constant and T the absolute temperature) is 0.593 at 298^{o}K, i.e. circa 0.6 at biological temperatures, so 1 kcal/mole is not significant above thermal noise. Free energies of 2, 3, 4, and 5 correspond to binding association constants of 5, 148, 786,

and 4160. The above free energies are usually expressed as negative, for the perspective from the associated system. Considering that absolute values are much less reliable than relative values in this field, one might conservatively consider a binding energy of -3.5 kcal/mole as worthy justification for keeping a compound on a list, if one does not wish to reject prematurely, and this seems reasonable if one still has in mind the refutation principle. This includes the mental picture that a model pharmacophore such as 11β-hydroxysteroid dehydrogenase type 1 has a fairly large cavity which does not provide strong steric inhibition to the candidate ligands, but new evidence might show that a large ligand such as a steroid might be too big to fit the real target which the experimental data is describing. In other words, deficiencies in the pharmacophore model will start to show up when considering larger potential drugs.

There is also the benefit of using 11β-hydroxysteroid dehydrogenase type 1 as model pharmacophore that the present author has a data base of experimental and computational studies on compounds that bind to it. It should be stated, nonetheless, that any case for any common *evolutionary relationship* between this dehydrogenase and the spike protein binding receptor ACE2 would be, at best, marginal. 11β-hydroxysteroid dehydrogenase type 1 has 292 residues and ACE2 has 613. There is a 24% identity match of amino acid residues in the region of best possible match of the dehydrogenase. There is also further 19% conservative substitution (CLUSTAL ':', i.e. conservation between groups of strongly similar properties with a score greater than .5 on the PAM 250 matrix). If taken alone, this would provide some basis for further exploring a relationship. Admittedly, the conventional rule of thumb is that any two sequences are considered homologous if they are more than 30% exact amino acid residue matches, and strictly speaking this should apply over their entire lengths (this is discussed in Chapter 8 of ref [12] and a brief review of standard tools is given in refs [2] and [38]). Nonetheless, caution is required because the 30% exact match criterion is well known to miss many easily detected homologs and 15-20% is sometime found supported by evidence of evolutionary and functional relationship. For example, alignments between common cold and SARS-CoV-2 spike proteins already discussed above are in this range, but there is every good reason to believe a common ancestry, there is an overall conformational similarity, and essential features of some sequence motifs are preserved. There is some sense of comparable fold motifs with ACE2 comprising two 11β-hydroxysteroid dehydrogenase type 1 folds. The dehydrogenase is a bundle of some 12 well defined, roughly parallel and antiparallel α-helices of up to about 30 residues, interspersed by 7 short β-pleated sheet strands. ACE2 has some 20 well defined, predominantly and very roughly parallel and antiparallel α-helices of up to about 30 residues, interspersed by 6 short β-pleated sheet strands. If there is a common evolutionary origin of 11β-hydroxysteroid dehydrogenase type 1 and ACE2

domains, it is distant, but it remains marginally possible and more extensive conformational analysis is underway.

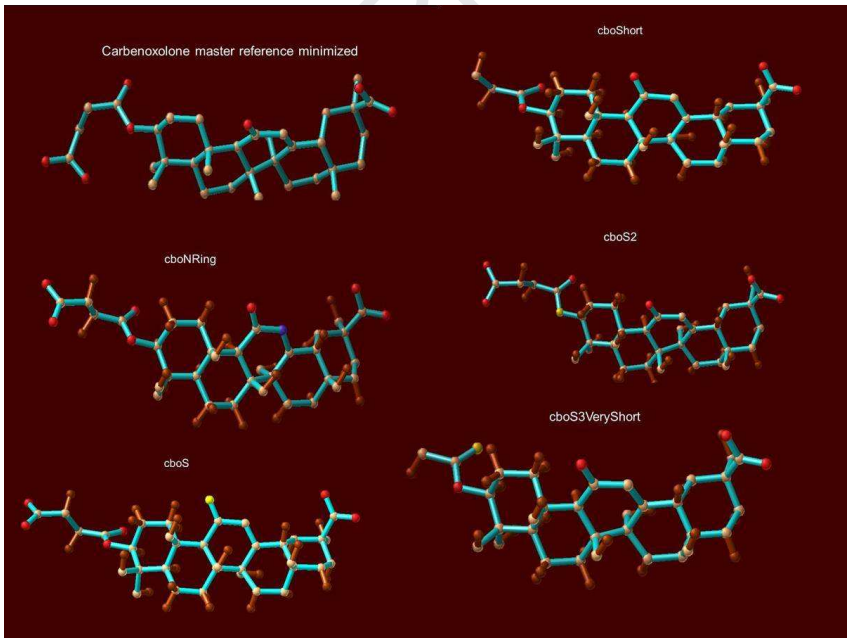
There is even less evidence of homology between 11β-hydroxysteroid dehydrogenase type 1 and TMPRSS2, although a serine residue is highly conserved in the catalytic site in both cases, which arguably makes it worthy of some initial exploration. TMPRSS2 comprises distinct cystine rich scavenger domain (residues 150-242) and a serine protease domain (residues 255-484). CLUSTAL O(1.2.4) multiple sequence alignment gives an exact match of amino acid of only 17.5%. There are some grounds for further investigation in the future. There is also further 17.5% conservative substitution (CLUSTAL ':', see above). For TMPRSS2 there are some suggestive short section matches in same order of appearance, e.g. AQYYYS with AYYYYS, VVSHC with VVSHC, LYHSD with LFHDD, and GILRQS with GALRQE, which by some arguments slightly increase statistical significance. No significant conformational homology is apparent, so it is even more likely to be a chance match, and any argument for similarity between the proteins would be on the basis of some kind of convergent evolution based on certain common ligands, recalling again that the coronavirus might benefit from inhibiting an inflammatory response.

Preliminary studies on the panel of ligands discussed below suggest some degree of binding (-4.5 kcal/mole and better, i.e. more negative) to both the above and 11β-hydroxysteroid dehydrogenase type 1, but these studies are still not fully complete and low energies may yet be obtained. The most substantial data base of results that can reasonably be considered final is in large part from the original studies [50]. There carbenoxolone was automatically evolved (by automatic editing of its chemical structure) under the combined selective pressure of improve binding to 11β-hydroxysteroid dehydrogenase type 1 while avoiding significant match with compounds covered by all US patents [50], and subsequent docking and high grade molecular dynamic simulations were carried out on IBM's Blue Gene [50]. Many subsequent studies have, however been carried out on using KRUNCH on a personal computer, because in the initial study it predicted well the Blue Gene results providing that the KRUNCH binding energies obtained were corrected (or refined) to fit the Blue Gene results by a linear regression formula [50].

Recall again that it is on the basis of similarities between some compounds that antagonize SARS virus entry and bind the steroid dehydrogenase, plus a notable commonality in the case of emodin (i.e. it binds both), that this model pharmacophore was chosen. Since emodin and many other compounds of interest contain two or three or more aromatic rings, it is reasonable, at least as an initial tactic, that one may regard them as pieces of the steroid ring system and start them in the steroid binding cavity in the same "plane" as the steroid ring. In such a case involving minor variations as

sidechains on the original steroid core, the way to make initial fit to using carbenoxolone as guide is obvious. However, the flat view of a steroid is misleading. The steroid ring system can “buckle” in various cis-trans combinations of bonds in the rings, and the longer sidechain conformations preferred on the basis of intramolecular energy are perhaps not obvious. Although the rotation barriers for most of the transitions are clearly above the thermal energy (kT) energy conformations (0.6 kcal/mole), the associated energy demands for buckling of parts of the steroid ring system of variously and roughly 2.5 to 5.0 kcal/mole are less that the ligand-receptor binding the associated energy demands are below the gain in energy from ligand-receptor binding to the protein target. This is shown in the high grade quantum mechanical Hartree-Fock GAMESS calculations on Blue Gene in the original study [50] but which have not been described in the literature. Minimized energy conformers of steroid-like compounds considered are shown in Fig. 3.

Fig. 3.
Preferred Conformers of Example Steroid-Like Analogues by Hartree-Fock Calculations.

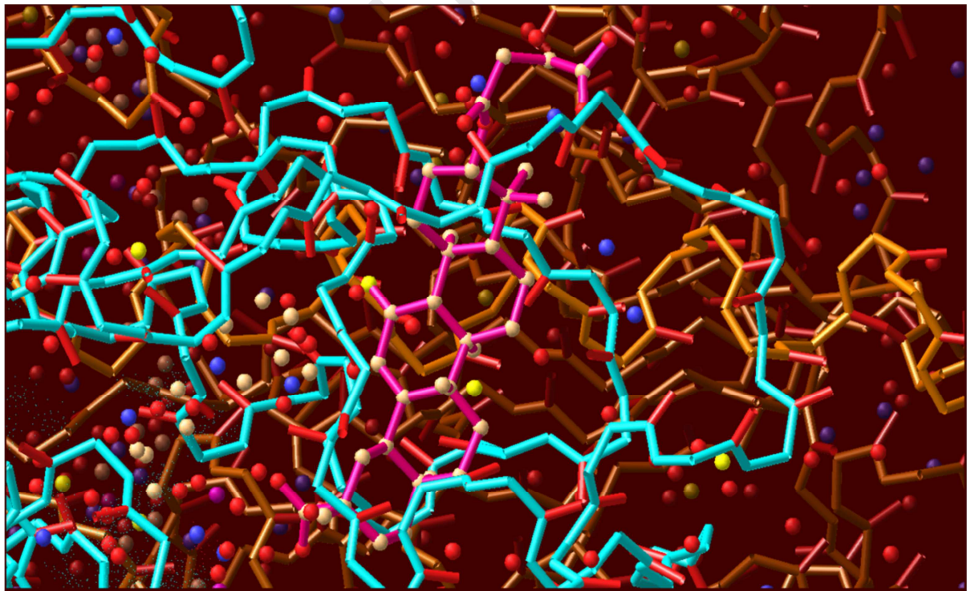


Such *calculations in vacuo* are less reliable for the charged species, but one may obtain a qualitative assessment from relative values and comparative uncharged species. These compounds are also shown more clearly from the chemist’s perspective in two dimensional formula format, later below. Fig. 4 shows one of early analogues of carbenoxolone (the thioketone derivative cboS1 discussed later below) in the 11β-hydroxysteroid dehydrogenase type 1 steroid binding site. The particular interest in this

compound is as follows. Since in the original study [50] KRUNCH judged this as the strongest binder at -16.8 kcal/mole, this compound was frequently used as a starting template for initial docking configurations when using KRUNCH. This is even though (a) it is probably an unlikely choice for a chemist to use in practice because of likely oligimerization of the thioketone groups, and even though (b) Corphos (also known as Cortisol 21-phosphate, Cortisol, phosphate, Hydrocortisone-21-phosphate or 21-Hydrocortisonephosphoric acid) was the strongest at -16.8 kcal mole when using instead the AMBER force field for molecular dynamics on IBM's Blue Gene [50]. The thioketone still retained a reasonable binding energy of -16.3 kcal/mole in the latter study, however, i.e. effectively the same binding strength within the state of the art. Fig. 4 does not of itself give details of any ligand-protein interactions (but see discussion in refs [3, 49]), although it does illustrate the tightest of fit. That is, except to the lower right of the thioketone ring of the ligand, which does appear to relate to genuine opportunities for additional groups to be added to carbenoxolone at that position.

Fig. 4.

A Carbenoxolone Analogue *In Situ* in the 11β-hydroxysteroid dehydrogenase type 1 Steroid Binding Cavity. It is the strongest binder in the KRUNCH Modeling System.

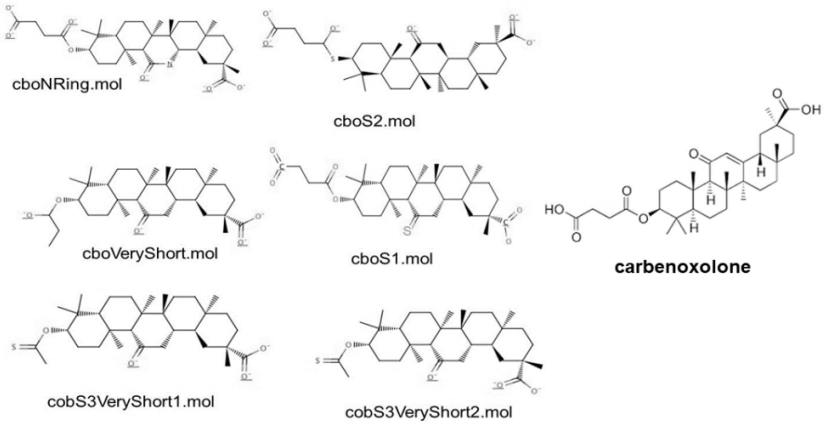


Carbenoxolone and initial closely related derivatives derived in that study [50] are shown in Fig. 4, binding in the range, -17 to -14 kcal/mole. Accuracy and limited realism of such methods does not really justify more precise statements on binding energy, and the classification of binding below is as strong, medium, and weak, but see ref [50] for more detail on some of the compounds. Authors variously consider binding energies -5

to -9 as a safe requirement for significant binding, but again this is subject to considerations of accuracy and almost all agree that it is only the relative values that are significant. Note that while they are often interpreteted as estimates of binding free energy, the entropy component, particularly of the aqueous solvent and solute-solvent intercations, is difficult to estimate. Experimental binding values of ligands in general in biological systems typically range from -4 to -16 kcal/mole, though over 95% lie in the range -7 to -13 kcal/mole. The thiioketone derivatives are more of theoretical interest in binding studies because in practice they may cause oligmerization

Fig. 5

Organic Compounds Binding the Pharmacophore. Strong Binders.
From the original study [50]. The estimated binding energy is in the range -17 to -14 kcal/mole. These were deigned from carbenoxolone with the intent to have a stronger or comparable strong binding (-16 kcal/mole). Corphos, cboNRing, and cboS2 bind at -17 kcal/mole.



Recall that the two peptide analogues of features of the spike protein of interest [3] are as follows.

Original L-Mimetic. (NH_3^+)-GPSKRSFIEDLLFNKVTLAC-(COO^-)
retroinverso mimetic (NH_3^+)-dextro-[GNFLLDEIFSRKSRKSPC]-(COO^-)

So far, simulations only show these to be binding relatively weakly at -10 and -8 kcal/mole respectively, but these compounds are highly flexible with a theoretical internal *in vacuo* conformational entropy corresponding to about -19.5 kcal/mole as discussed in Theory Section 2, show multiple binding modes and conformers on binding, and may not yet be complete. A high performance computer like IBM's Blue Gene used in the earlier drug design study [50] would certainly help.

In Fig. 6 is shown a set of compounds from the ZINC data base [69], and most were identified from the original 11β-hydroxysteroid dehydrogenase type 1 study [50] and subsequent studies. These bind significantly by usual criteria, but more weakly. They are in the range found for the synthetic peptides of interest, but have much less conformational freedom. A small few not shown here did appear in the original higher grade studies, but the reasonable binding energies could not be reproduced for reasons that are not as yet clear.

Fig. 6.

Organic Compounds Binding the Pharmacophore. Medium Binders.
From the ZINC data base. The estimated binding energy is in the range -9 to -11 Kcal/Mole.

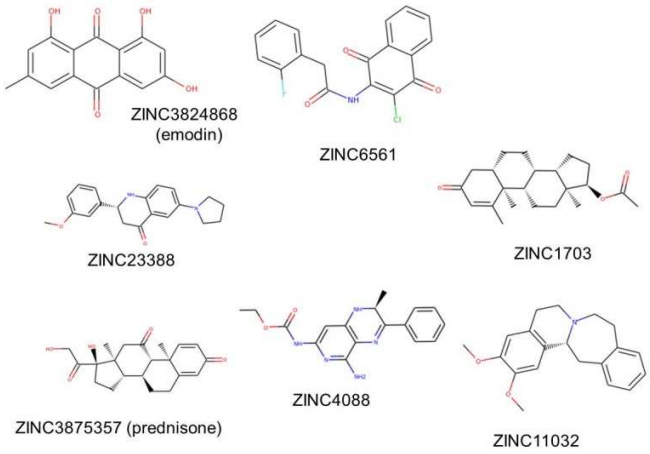
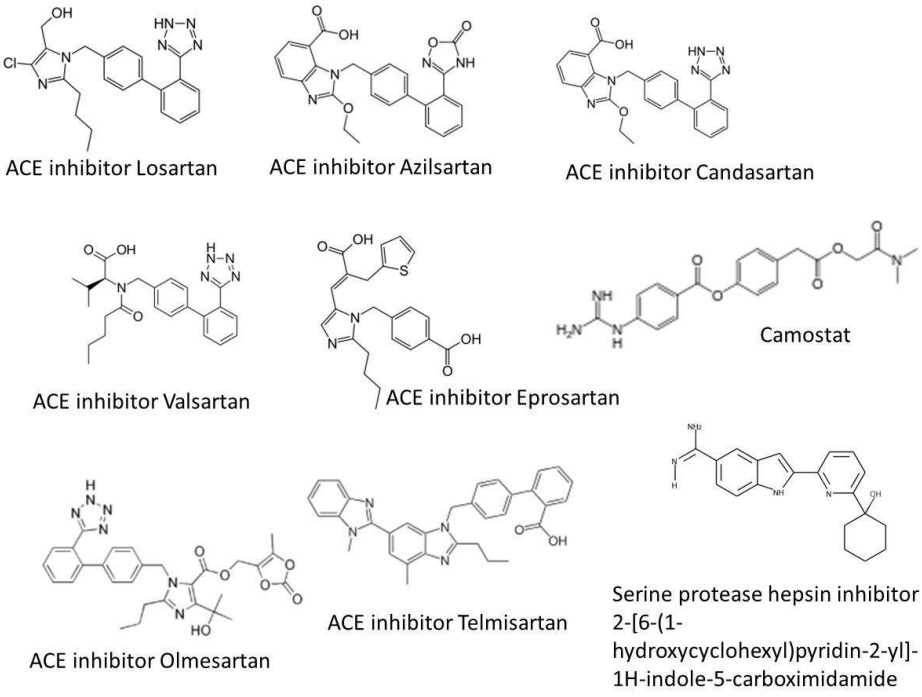


Fig. 7 shows some weaker binding results using Krunch [50]. *Great caution is required in drawing conclusions from the compounds in Fig. 3.* Camostat is definitely of interest as an inhibitor of the ACE2 protein to which the spike protein initially binds for cell entry and does seem effective in blocking entry [74], and similarly hepsin results are of related interest, e.g. as it is a potential alternative entry point. The ACE inhibitors also looked initially interesting by virtue of certain similarities to the other potential ligands, and of course because of their binding in this theoretical study, but most of the traditional ACE inhibitors are commonly viewed as not inhibiting ACE 2. Taking a drug such as Valsartan that acts on ACE might up-regulate ACE2, so facilitating virus entry, [75] but emerging information is revealing a complicated picture: see Discussion and Conclusions. There are possible explanations that would still allow for competing with spike protein binding, but these seem somewhat unlikely. Most probably, the binding is sufficiently weak that the normal substrate, and also the spike protein, displace it. Aromatic group interactions may be important here [76].

Fig. 7.

Organic Compounds Binding the Pharmacophore. Weak Binders.
Selected known drugs or proposals (caution: use of ACE inhibitors might be counterproductive, see text). The estimated binding energy is in the range -5 to -7 kcal/mole.



Some consideration has been given to prediction of ligand binding site motifs, but so far these have proven essentially negative as regards interesting results that might shed any further light on the above, although some clues may well have been missed. Binding sites are often comprised of conserved residues that are not contiguous (continuous in a sequence), which will require further and more detailed study, although subsequences of 2 to 6 amino acid residues in length are worthy of a quick preliminary study because they are commonly involved in ligand interactions. The matches involved in here as judged by BLASTp and Clustal Omega are not statistically significant, but one might think of weak matches as ligand binding site predictions in much the same way that one thinks of epitope predictions. In much of the present paper, the structure of emodin, carbenoxolone and related compounds have involved discussion of aromatic rings and hence phenylananine (F), tyrosine (Y) and tryptophan (W) and more generally amino acid residues with hydrophobic character. Very polar subsequences are also strong binders of charged ligands, or have a role for charged molecules or inorganic ions in some way. As far as such subsequences in the coronavirus spike protein are concerned, very polar charge-pattern motifs such DRETS and DREDS are common in

ligand binding some of the molecules that may be of interest as antagonizing SARS entry, activation or replication in some way, in the present author's experience. Specifically, motifs like this were of initial interest in the present project because SARS-CoV) nonstructural proteins have zinc finger motifs, and RET and especially RED are common in PROSITE motifs at https://prosite.expasy.org/cgi-bin/prosite/prosite_search_full.pl, including zinc-finger motifs. This is not considered directly relevant to the spike protein but, for example respectively in GenBank entry AIA62240.1 and DREDS and DRETS align with SRLDKV in three-way Clustal Omega alignment with SRLDK of the original Wuhan spike protein sequence MN908947.3 and DRLDT of NP_073551.1 spike protein. However, these and many similar alignments also illustrate considerable sequence variation, and the weak matches are not close in the sequence to the subsequences of interest neither for coronavirus spike protein nor human proteins of potential discussed above. As far as ACE2 is concerned, the closes match with DRETS and DREGS is DRKKPS, but this weak match again lays well away from regions of current interest, e.g. in the sequence from the region that interacts with the spike glycoprotein. DTETA and DRFIN do occur in the C-terminal half of human 11β-hydroxysteroid dehydrogenase type 1, but again these are expected to be coincidental matches.

5. Discussion and Conclusions.

5.1. Convenient Herbal Solutions.

Like the first vaccines [77] therapeutics too have, of course, been drawn directly or almost directly from nature, until the late 19th Century when chemical synthesis became a science, and only in the 1970s did use begin to made of computers for rational drug design. The advantages of still seriously considering herbal remedies is that they tend to be tolerated by cells because they are produced in cells, they are already subjected to hundreds of years of human trial, are often economic solutions for bulk production, and are leads for further drug development and discovery. The principal non-peptide compounds considered above as possible therapeutics have such convenient and herbal sources. As reviewed previously [3], the herbal extract emodin is a convenient product extracted from rhubarb, buckthorn, and Japanese knotweed, and several fungi. The previous paper [3] also noted that emodin had certain molecular similarities with anti-inflammatory drugs such as carbenoxolone, derived from an extract, glycyrrhizic acid, from liquorish (licorice), that variously inhibit or are believed to inhibit human11β-hydroxysteroid dehydrogenase type 1. The above does not guarantee efficacy of emodin carbenoxolone against SARS-CoV-2, not least because even the emodin studies concerned SARS-CoV not SARS-Cov-2, and the case for the dehydrogenase is circumstantial, but these and related substances are worthy of investigation. Indeed, this paper has described a number of compounds that bind the

dehydrogenase. Importantly, however, recall that the weak binders that are also ACE inhibitors may be more dangerous and promote infection because they upregulate ACE2 [75]. Nonetheless, that situation is not resolved, as follows.

5.2. Interesting Circumstantial Clues and Need for Further Research

The above considerations as to the action and possible usefulness are empirical observations that are largely independent of bioinformatics and molecular computation and they are even independent of whether the correct human protein targets discussed here are correct and relevant; however, it would be valuable to know what might relate ACE2 and 11β-hydroxysteroid dehydrogenase type 1, and ideally also understand why both enzymes might "benefit" the coronavirus by interaction with them. Also, this paper could not provide any evidence of an evolutionary relationship between these proteins, despite certain similarities, or with TMPRSS2. So far, there is no obvious relationship with the dehydrogenase, and some other studies by the author on other transmembrane serine proteases do not, as yet, suggest any relationship. Without such connections, the dehydrogenase can only be considered as a rather arbitrary model pharmacophore. As such, it is possibly meritorious as correctly representing an ensemble of multiple targets, but that may be fortuitous, and hence only to be used until refuted by evidence.

Nonetheless, possible clues as to mutual relevance of these human protein targets might come noting their tissue distribution and considering how this may relate to their biological role. As regards ACE2, its mRNA is known to be present in virtually all organs. Studying SARS entry into human cells, Hamming et al. [78] considered their most remarkable finding to be the substantial surface expression of ACE2 protein not only on lung alveolar epithelial cells but also enterocytes of the small intestine, as in arterial and venous endothelial cells and arterial smooth muscle cells in all organ studied (oral and nasal mucosa, nasopharynx, lung, stomach, small intestine, colon, skin, lymph nodes, thymus, bone marrow, spleen, liver, kidney, and brain). There is the attractive prospect that several many herbal remedies considered as laxatives interact with ACE2 and inhibit SARS-CoVid-2 entry. Recall that emodin is an antagonist of both ACE2 [59-61] and 11*β*-hydroxysteroid dehydrogenase type 1 [62]. In the past, 11*β*-hydroxysteroid dehydrogenase type 1 has been considered to be distributed mainly in the human liver, with no detectable levels in the intestine or kidney, mostly membrane-bound and retained in the liver microsomal fraction [79]. This was not however the finding of Bruley et al. 80]. They found it to be highly expressed in glucocorticoid target tissues including liver and notably the lung, and modest levels in the brain. It was also found in modest levels in adipose tissue where it is of medical interest that selective increase expression occurs in obese humans and rodents and is likely to be of pathogenic importance in the metabolic syndrome [80]. Lung expression appears to be managed differently: a new promotor that the authors discovered and called P1

predominated in lung while the previously known promotor predominated in liver, adipose tissue, and brain [79].

Researchers therefore need to sort out an intriguing web of information. It is possible that a complex web of laxative and anti-inflammatory effect may provide clues by somehow relating to the body's attempts to reject and eject viruses of this kind, and the virus's attempts to resist. It is well known that some COVID-19 patients complain of stomach upsets and diarrhea. To recapitulate the essential themes in terms of action in the alimentary tract, recall again that emodin had earlier been shown by several groups of researchers (e.g. ref [79]) to inhibit SARS-CoV entry into cells (apparently initially by binding ACE2), and emodin is taken as a herbal laxative. Licorice has, conversely, been sometimes taken as a soother for alimentary disorders, and carbenoxolone has been used commercially in the past specifically to treat peptic ulcers. Intriguingly, carbenoxolone is also known to influence the renin-angiotensin system involving ACE2, so at least there appears to be a connection in terms of networks of physiological control. As noted above, while traditionally 11β-hydroxysteroid dehydrogenase type 1 has been thought of as a liver enzyme, many researchers have indicated that both ACE2 and the dehydrogenase are available in both lung and intestinal tract (e.g. refs [78-80]). This all hints also that some of the other laxatives that work in a similar stimulatory way might block viral entry on ACE2 and perhaps other targets, and should be explored. Of course, absolutely nothing should be done by patients without physician direction, because dosages are difficult matters (not least in herbal products) and there are potentially serious side effects on such as salt balance and blood pressure, and some might cause birth defects, all potentially worse than COVID-19 would be, for most people. But more worryingly still, the situation is not settled, and physicians and patients could take action in the wrong direction. Gurwitz [81] has emphasized that the picture is even more complex. He examined reports from China suggesting that a mechanism of production of lung injury during the viral infection may be due to excess free angiotensin-II, which might be displaced from ACE2 by the SARS virus particles. If so, then increasing the amount of ACE-2 could be desirable and administering angiotensin receptor antagonists could beneficially upregulate the production of ACE-2. It now becomes important to examine medical records of patients who have, and who have not, been infected by SARS-CoV-2, with a particular eye on who is, and who is not, taking ACE inhibitors.

5.3. Comments on Use of the Proposed Synthetic Peptides.

As noted in Section 1.4, Merrifield developed first solid phase peptide synthesis on crosslinked polystyrene beads in 1963 [12]. Somewhat like natural compounds discussed above, Peptides and petidomimetics are potentially important first steps in more detailed rational design of small organic molecules convenient as traditional "in a

pill" drugs. However, as in the present paper, the ability to propose specific peptides and peptidomimetics does depend on bioinformatics, and benefits from some computational chemistry. Note that in this case, one is thinking largely not of screening natural products, but now considering truly novel molecules using theoretical methods because they do not yet exist. The variations in the KRSFIEDLLFNKV motif that might be appropriate to synthetic vaccine and peptidomimetic antagonist design suggest the following where the amino acid residues in square brackets [ ] represent alternatives. (G?) means an optional glycine insertion.

[KR](G?)S[AILF][AILF]ED[IL]LF[ANDS]KV

The above is also valid as a regular expression, i.e. a match query in operating systems and software. More generally and colloquially,

(positive charge)-(optional glycine)-serine-hydrophobic-hydrophobic-glutamate-aspartate-hydrophobic-leucine-phenylalanine-(hydrophilic or alanine)-lysine-valine

Considerations at the N-terminus and C-terminus to design a synthetic vaccine, and the retro-inverso approach for a peptidomimetic agonist, are described in ref [3].

Other variations appear as the strain becomes more distant; there is not a universal clear indication of any sharp point of departure, although the above glycine (G) insertion is evidence that a significant jump can happen. One may therefore ask what variations should be included. With the emphasis on SARS-CoV-2, only closely and medium distance relatives are of interest, with the purpose of prevent mutations that escape from vaccines and antagonists, and elude diagnostics. As far as SARS-CoV-2 is concerned, KRSFIEDLLFNKV is a satisfactory basis because a large number of coronaviruses significantly different from SARS-CoV-2 preserve it, or in a few cases have very conservative substitutions. It may well be that the fact that residues are, for example, hydrophobic or positively charged is sufficient to for the approach to be applicable to other mammalian coronavirus diseases, if successful for the above basic motif form. Attempting to tackle the common cold is not a priority. In other words, it may well be that an immune response against KRSFIEDLLFNKV will also illicit a response against the motif variants, providing of course that KRSFIEDLLFNKV elicits a response itself. It remains that this motif is one of very few subsequences that still recognizable when moving on to rather distantly related coronaviruses such as those of the common cold.

5.4. Comments on Potential Therapeutic Antagonists.

One feature of both Figs. 5 to 7 is of course the constant appearance of aromatic rings, and this is also noticeable in many of the studies of antagonist's against SARS virus binding and activation. Of course, the aromatic (i.e. benzene) ring makes copious appearance in many pharmaceutical agents in any event, because they provide rigid scaffolds for added groups supported by many long-established recipes for synthesis. The compounds in Figs. 5 to 7 should be distinguished from those such as Lopinavar, Ritonar, Promazine and particularly Niclosamide that have been explored for SARS viruses in the past, because these are targeted by drug designers against the SARS virus own protease required for maturation of the assembling virus. Nonetheless, some of these do have a visual similarity to the compounds in Fig. 3, particularly Niclosamide (which is normally a medication used to treat tapeworm infestation). Also, in the present case, prevalence of aromatic rings in Figs 1 to 3 is hardly surprising, since carbenoxolone and derivatives shown in Fig. 1 were the starting point for their evolution or selection from the ZINC data base. Nonetheless, there is, in principle, nothing to constrain the evolution to aromatic chemistry [50] and later unpublished studies did produce molecules departing from aromatic chemistry. However, these bound relatively weakly.

With the possible importance of aromatic rings and avoidance of escape mutations by the coronavirus in mind, a question is whether occasional loss of phenylalanine (F) from the KRSFIEDLLFNKV motif discussed above contests the tentative hypothesis that the peptidomimetic candidates derived from KRSFIEDLLFNKV bind to a similar site as the smaller organic ligands considered here, because of two phenylalanine residues (F) in the original motif and a tendency to several benzene rings in the case of organic ligands. The answer is: perhaps. There seems to be a need to have one aromatic ring present in the motif, and no match with a coronavirus in GeneBank was detected by the author by BLAST-p using queries with no phenylalanine (F), e.g. RSAIEDLLLDKV, RSAIEDLLIDKV, RSAIEDLLADKV, RSAIEDLLMDKV, RSAIEDLLWDKV, and RSAIEDLLYDKV as queries, though as also noted above, the search has not been exhaustive. It would not be too contradictory to any of the current main hypotheses if some examples were found. The fact that tyrosine (Y) does not seem to readily substitute here for phenylalanine (F) (from which it differs only by a hydroxyl -OH, i.e. phenolic group) suggests an important hydrophobic feature of the pharmacophore at that point.

Of course, many or most drug-like molecules contain at least one aromatic ring and this is almost certainly because they can form especially strong stacking interactions in the binding site. One very relevant report in the same month of writing the present paper emphasizes that the use of protein and other fragments to characterize binding pocket and determine the strengths of ligand-protein interactions is common in both a computational and experimental approach, and that aromatic interactions are

both strong and need special attention [76]. Because of resonance and the special nature of the π orbitals, the strength of stacking is best calculated using high level quantum mechanical approaches, not empirical force fields [76]. However, as these calculations are performed in vacuum, solvation properties are neglected, and this led to the proposal of a Grid Inhomogeneous Solvation Theory (GIST) to describe the properties of individual heteroaromatics and complexes; this gave good correlation for the estimated desolvation penalty and the experimental binding free energy, and prediction of binding sites [76].

5.6. Final Comment.

A main conclusion is that peptide KRSFIEDLLFNKV remains of special interest as well conserved across coronaviruses. Other sites and other proteins of the virus may, of course, emerge as the solutions to this formidable problem. All aspects of the virus must be considered. However, even the ACE2 binding domain is significantly more prone to accepted mutations. The recurrence of the core features of the KRSFIEDLLFNKV motif over so many diverse species reminds us of zoonotic origins, and it might be recalled that Jenner, the inventor of vaccination, consider that many and perhaps all plagues of mankind might ultimately be of animal origin [77].

References.

1. Masters, P.S. , The molecular biology of coronaviruses , Advances in Virus Research . 66:193-292, (2006).
2. 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, W. J., Wang, D., Xu,W., Holmes, E. C., Gao, G. F., Wu, G, Chen, W., Shi, W., Tan, W., Genomic characterization and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, www.thelancet.com Published online January 29, 2020 https://doi.org/10.1016/S0140-6736(20)30251-30258, (2020).
3. 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, Computers in Biology and Medicine, published online 26 February 2020, 103670, (2020).
4. 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, circulated and published on ResearchGate DOI: 10.13140/RG.2.2.18275.09761, (2020).
5. Li, F., Structure, Function, and Evolution of Coronavirus Spike Proteins, Annual Reviews in Virology,3(1), 237–261, (2016).

6. Kam, Y.W., Okumura, Y., Kido,.H., Ng, L. F. P. Bruzzone, R, and Altmeyer, R., Cleavage of the SARS Coronavirus Spike Glycoprotein by Airway Proteases Enhances Virus Entry into Human Bronchial Epithelial Cells In Vitro Published: PLOSONE November 17 (2009), https://doi.org/10.1371/journal.pone.0007870 (last accessed 1/26/2020)
7. Belouzard, S., Chu, V. C. and Whittaker, G. R., Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites, Proceedings of the National academy of Science, 106(14, 5871-5876; https://doi.org/10.1073/pnas.0809524106 (last accessed 1/26/2020)
8. Gui, M., Song, W., Zhou, H., Xu, J., Chen, S., Xiang, Y., Wang, X., Entity 1 containing Chain A, B, C SARS-CoV spike glycoprotein, Cell Research 27, 119-129 (2017)
9. Liu, I. j., Tsai, W. T., Hsieh, L. E., and Chueh, L. L., Peptides Corresponding to the Predicted Heptad Repeat 2 Domain of the Feline Coronavirus Spike Protein Are Potent Inhibitors of Viral Infection, PLoS One, 8(12): e82081, (2013).
10. Forni, D., Filippi, G., Cagliani, R. De Gioia, L., Pozzoli, U., Al-Daghri, N., Clerici, M, and Sironi, Manuela, The heptad repeat region is a major selection target in MERS-CoV and related coronaviruses, *Scientific Reports* volume 5, 4480. (2015).
11. Berend, J. B., Rossen, J. W. A.,Bartelink, W., Zuurveen, C., de Hann, A. M., Duquerroy, Boucher, C. A. B., and Rottier, P. J.,Coronavirus Escape from Heptad Repeat 2 (HR2)-Derived Peptide Entry Inhibition as a Result of Mutations in the HR1 Domain of the Spike Fusion Protein, Journal of Virology, March, 2580–2585, (2008).
12. Robson, B. and Garnier, J, Introduction to Proteins and Protein Engineering, Elsevier Press, Second Edition, (1998)
13. Sachdeva, S. Peptides as 'Drugs': The Journey so Far, International Journal Peptide Research and Therepeutics (2017) 23: 49. https://doi.org/10.1007/s10989-016-9534-8 (2017)
14. Li, W., Joshi, M. D., Singhania, S., Rasey, K. H., and Murthy, A. K., Peptide Vaccine: Progress and Challenges, Vaccines (Basel) 2(3): 515–536 (2014).
15. Robson, B., Fishleigh, R. V., and Morrison, C. A. , Prediction of HIV Vaccine, Nature, 4, 325, 395 (1987).
16. Fishleigh, R. V. and Robson, B, Synthetic Peptides Related To HIV-Env Proteins, patent Patent: EP00371046A1, (1990).
17. Fishleigh, R. V., Robson, B. and Aston, R., Synthetic Polypeptides Derived From The HIV Envelope Glycoprotein.patent : EU0636145, (1995),
18. Fishleigh, R. V. and Robson, B, Fragments Of Prion Proteins, patent EP00636145A1, (1995),
19. Fishleigh, R. V. and Robson, B, and P. Mee, Fragments of prion proteins (1998) patent US05773572 (1998)
20. Citywire, https://citywire.co.uk/new-model-adviser/news/protherics-mad-cow-test-goes-international/a220861 (last access 1/28/2020)
21. Robson, B. From Zika to Flu and Back Again. CAVIRC (Report of the Caribbean Anti-Virus Informatics Research Center), https://www.researchgate.net/publication/296667599_From_Zika_to_Flu_and_Back_Again_CAVIRC_Caribbean_Anti-Virus_Informatics_Research_Center,

DOI: 10.13140/RG.2.1.5000.6808, (2016).
22. Robson, B. , Computer Aided Peptide and Protein Engineering, in Applied Biotechnology, Proceedings of Biotech 86' Europe, held in London, 1, B9-B14, (1986).
23. Robson, B., Ward, D. J., and Marsden, A. , The EPSITRON concept of peptide and protein engineering. Applications of computer-aided molecular design", Chemical Design Automated News 1, (7) 9-11, (1986)
24. Robson, B., Platt, E., Marsden, A., and Millard, P., An expert system for protein engineering. Its application in the study of chloramphenicol acetyltransferase and avian pancreatic polypeptide", (1987 J. Mol. Graphics. 5, 8-17, (1987).
25. Fishleigh, R.V., Robson, B., Garnier, J. and Finn, P. W. , Studies on rationales for an expert system approach to the analysis of protein sequence data - preliminary analysis of the human epidermal growth factor receptor", FEBS Letts. 2, 4, 219-225, (1987)
26. Garnier, J., Gibrat, J.F., Levin, J., and Robson, B. Modélisation des polypeptides: application aux oligopeptides vaccinants" INRA/EUC Rapport FRT - 85 T 0606, (1988)
27. Ball, J., Fishleigh, R. V., Greaney, P. J., Marsden, A. Platt, E., Pool, L. J., and Robson, B, A Polymorphic Programming Environment for the Chemical Pharmaceutical and Biotechnology Industries", Chemical Information Systems - Beyond the Structure Diagrams. Eds. D. Bawden and E. M. Mitchell, Ellis Horwood Press, 107-123 (1990).
28. Robson, B., Platt, E. and Li, J. , Computer Aided Design of Biomolecules: The Big Hammer Approach, in Theoretical Biochemistry and Molecular Biophysics 2 Proteins, Adenine Press, Eds. David L Beveridge and Richard Lavery, 207-222, (1992).
29. Clark, D. E., Frenkel, D., Levy, S. A., Murray, C. W., Robson, B., Waszkowycz, B. and Westhead, D. R (1995). "PRO_LIGAND: An Approach to De Novo Molecular Design. 1. Application to the Design of Organic Molecules", J. Comp.Aided.Mol.Des. 9, 13-32
30. Waszkowycz, B', Clark, D. E., Frenkel, D., Li, J., Murray, C. W., Robson, B., and Westhead, D. R., PRO_LIGAND: An Approach to De Novo Design. 2. Design of Novel Molecules from Molecular Field Analysis (MFA) Models and Pharmacophores, J. .Med. Chem., 37, 3994-4002, (1994).
31. Westehad, D. R., Clark, D. E., Frenkel, D., Li, J., Murray, C. W., Robson, B., and Waszkowycz, B., PRO_LIGAND: An Approach to De Novo Molecular Design. 3. A Genetic Algorithm for Structure Refinement, B. J. Comp.A ided. Mol. Des. 9, 139-148, (1995).
32. Frenkel, D., Clark, D. E., Li, J., Murray, C. W., Waszkowycz, B., and Westhead, D. R. , PRO_LIGAND: An Approach to De Novo Molecular Design. 4. Application to the Design of Peptides, D. Frenkel, D. E. Clark, J. Li, C. W. Murray, B. Robson, B. Waszkowycz and D. R. Westhead. (1995). J. Comp. Aided Mol. Des. 9, 213-225, (1995).
33. Kumar, S. P., Receptor pharmacophore ensemble (REPHARMBLE): a probabilistic pharmacophore modeling approach using multiple protein-ligand complexes, Journal of Molecular Modelin, 15;24(10):282 , DOI: 10.1007/s00894-018-3820-7, (2018).
34. Robson, B., Caruso, T, and Balis, U. G. J. (2013),Suggestions for a Web Based Universal Exchange and Inference Language for Medicine", Computers in Biology and Medicine, 1;43(12):2297-310.
35. Robson, B. and 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, Computers in Biology and Medicine, 66, 82-102, (2015).

36. Robson B. (2016), 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", Computers in Biology and Medicine, 79, 299–323.
37. Robson B. and 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, Computers in Biology and Medicine, 95,147-166. (2018)
38. Robson., B. , Extension of the Quantum Universal Exchange Language to precision medicine and drug lead discovery. Preliminary example studies using the mitochondrial genome , Computers in Biology and Medicine, 117, in press, (2020).
39. The Biology Workbench http://workbench.sdsc.edu/ (last access 1/28/2020)
40. U.S. National Library of Medicine, National Center for Biotechnology information, National Institutes of Health, BLASTP SUITEhttps://blast.ncbi.nlm.nih.gov/Blast.cgi (last accessed 1/28/2020).
41. Garnier, J. and Robson, B., The GOR Method for Predicting Secondary Structure in Proteins"in 'Prediction of Protein Structure and the Principles of Protein Conformation,' Ed. G. D. Fasman, Plenum Publishing Corp. 417-465, (1989).
42. Rhone-Alpes Institute of Biology and Protein Chemistry, GOR IV, https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_gor4.html (last accessed 1/28.2020).
43. De Lima, D. P., Synthesis of angiotensin-converting enzyme (ACE) inhibitors: an important class of antihypertensive drugs, Quimica Nova, 22(3), (1999).
44. B. Robson, Beyond Proteins, Trends in Biotechnology, 17:8, 311-315 (1999). B. Robson, "Doppelganger Proteins as Drug Leads", B. Robson (1996), Nature Biotechnology, 14, 892-893 (1966).
45. G.M. Figliozzi. M.A. Siani. .LE. Canne. B. Robson, and R. J. Simon. *Chemical synthesis and activity of D, superoxide dismutase*. Protein Science 5. suppl. 1 . 72, 15 (1996).
46. B. Robson, Pseudoproteins: Non-protein Protein-like Machines, The Sixth Foresight Conference on Molecular Nanotechnology (1998) https://foresight.org/Conferences/MNT6/Abstracts/Robson/index.html (last accessed 9/72019)
47. Rai, J., Peptide and protein mimetics by retro and retroinverso analogs, Chemical Biology and Drug Design, 93(5), 724-736, (2019).
48. Hagler, A. T., Osguthorpe, D. J., and Robson, B, Monte Carlo Simulation of Water Behaviour around the Dipeptide N-Acetylalanyl-N'Methylamide", Science 208, 599-601, (1980).
49. Robson, B., Some Views of Solvation Effects in the Light of a Monte Carlo Simulation", The Biophysics of Water, J, Ed. F. Franks and F. S. Mathias. John Wiley & Sons Ltd, pp. 66-7, (1982).
50. B. Robson, R. Dettinger, A. Peters, and S.K.P. Boyer, Drug discovery using very large numbers of patents: general strategy with extensive use of match and edit operations" J. Computer Aided Molecular Design 25(5) 427 (2011).
51. Soria-Guerr, R. E., Nieto-Gomez, R., Govea-Alonso, B. O, and Rosales-Mendoza, S., An overview of bioinformatics tools for epitope prediction: Implications on vaccine development, Journal of Biomedical Informatics, 53, 405-414 (2015).

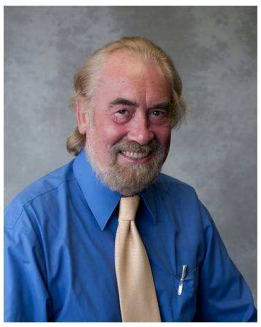
52. Kao, D. J. and Hodges R. S., Advantages of a synthetic peptide immunogen over a protein immunogen in the development of an anti-pilus vaccine for *Pseudomonas aeruginosa*. Chemical Biology and Drug Design, 74, 33–42, (2009).
53. Palatnik-de-Sousa1,C. B., Soares, I. D. S, and Rosa, D. S. Frontiers in Immunology, 18 April 2018 | https://doi.org/10.3389/fimmu.2018.00826, (2018).
54. Grifoni, A.,Sidney, J., Zhang, Y., Scheuermann, R. H., Petersnd, B., and Sette, A.,Candidate targets for immune responses to 2019-Novel Coronavirus (nCoV): sequence homology- and bioinformatic-based predictions, https://www.biorxiv.org/content/10.1101/2020.02.12.946087v1.full.pdf.(2020).
55. Wan, Y., Shang, j., Graham, R., Baric, R. S., Li, F., Receptor Recognition by the Novel Coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS 3, JVI Accepted Manuscript Posted Online 29 January 2020 J. Virol. doi:10.1128/JVI.00127-20 (2020).
56. Katz B. A., Luong, C. Ho J. D., Somoza, J. R., Gjerstad, E., Tang, J., Williams, S. R. Verner, E., Mackman, R. L., Young, W.B., Sprengeler, P,A,, Chan, H., Mortara , K., Janc, J.W., McGrath, M. E., Dissecting and designing inhibitor selectivity determinants at the S1 site using an artificial Ala190 protease (Ala190 uPA), J. Molecular Biology 19;344(2):527-47, (2004).
57. Lennart, M. R. Reinke,M., Spiegel m., Plegge, t., Hartleib, A., Nehlmeier, I., Gierer, S., Hoffmann, M., Hofmann-Winkle H., Winkler, M., and Pöhlmann1, S., Different residues in the SARS-CoV spike protein determine cleavage and activation by the host cell protease TMPRSS2, PLoS One. 12(6): e0179177, (2017).
58. Barr'e, O., Dufour, A. Eckhard, U., Kappelhoff, R., Béliveau,F., Leduc, R., and Overall, C.M., Cleavage Specificity Analysis of Six Type II Transmembrane Serine Proteases (TTSPs) Using PICS with Proteome-Derived Peptide Libraries, PLoS One. 2014; 9(9): e105984,doi: 10.1371/journal.pone.0105984 (2014).
59. Ho, T. Y., Wu, S. l., Chen, J. C, Hsiang, C.Y, Emodin blocks the SARS coronavirus spike protein and angiotensin-converting enzyme 2 interaction, Antiviral Research, 74(2), 92-101, (2007)
60. Schwarz, S., Wang, K. Yu, W., Sun, B., Schwarz, Emodin inhibits current through SARS-associated coronavirus 3a protein, Antiviral Research, ;90(1), 64-9, (2011).
61. Adedeji A. O., Severson, W., Jonsson, C., Singh, K., Weiss, S. R., Sarafian, S. G., Novel Inhibitors of Severe Acute Respiratory Syndrome Coronavirus Entry That Act by Three Distinct Mechanisms, J. virology, 87(14), 8017–8028, (2013).
62. Feng, Y., Huang, S-L, Dou, W., Zhang, S. Chen, J-H, Shen, Y, Shen, J-H, and Leng, Y, Emodin, a natural product, selectively inhibits 11β-hydroxysteroid dehydrogenase type 1 and ameliorates metabolic disorder in diet-induced obese mice, British Journal of J Pharmacology, 161(1): 113–126. (2010).
63. Westphal, U. Hydrophobicity and Hydrophilicity of Steroid Binding Sites, In: Steroid-Protein Interactions II. Monographs on Endocrinology, vol 27. Springer, (1986).
64. Kaliamurthi, S.; Selvaraj, G.; Chinnasamy, S.; Wang, Q.; Nangraj, A.S.; Cho, W.C.; Gu, K.; Wei, D.-Q. Exploring the Papillomaviral Proteome to Identify Potential Candidates for a Chimeric Vaccine against Cervix Papilloma Using Immunomics and Computational Structural Vaccinology. *Viruses 11*, 63, (2019). https://doi.org/10.3390/v11010063

65. Mehmood, A., Aman, C. K. , Wei, D-Q, Prediction and validation of potent peptides against herpes simplex virus type 1 via immunoinformatic and systems biology approach, Chemical Biology and Drug Design, 94(5), 1868-1883 (2019). https://doi.org/10.1111/cbdd.13602
66. Chu, Kaushik, A. C Wang, X. , Wang, W., Zhang, Y., Shan, X Russell, D., Salahub, Xiong, Y., Wei, D-Q, DTI-CDF: A cascade deep forest model towards the prediction of drug-target interactions based on hybrid features, *Briefings in Bioinformatics*, , bbz152, https://doi.org/10.1093/bib/bbz152.
67. Kaushik, A.C., Mehmood, A., Upadhyay, A.K. , Shalinee, P., Srivastava, S., Prayuv , M., Xiong, Y., Dai, X., Wei, D-Q, Sahi. S., CytoMegaloVirus Infection Database: A Public Omics Database for Systematic and Comparable Information of CMV,Interdiscip Sci Comput Life Sci (2019).
68. Kaushik, A .C., Mehmood, A., Peng, S., Zhang, Y. J., Dai, X., Wei, D-Q,A-CaMP: a tool for anti-cancer and antimicrobial peptide generation, Journal of Biomolecular Structure and Dynamics, published online, 6 Jan, (2020). DOI: 10.1080/07391102.2019.1708796
69. Khan, M. Y. and Kumar, V., Mechanism & inhibition kinetics of bioassay-guided fractions of Indian medicinal plants and foods as ACE inhibitors, Journal of Traditional and Complementary Medicine, 9(1): 73–8, (2019).
70. ZINC15 database. https://zinc.docking.org/ (last acessed March 9 2020).
71. Niespodziana, K., Napora, K., Cabauatan, C., Focke-Tejkl, M., Keller, W., Niederberger, V., Tsolia, M., Christodoulou, I., Papadopoulos, N. G., and Valenta, R., Misdirected antibody responses against an N-terminal epitope on human rhinovirus VP1 as explanation for recurrent RV infections, The FASEB Journal, Published Online:25 Nov 2011https://doi.org/10.1096/fj.11-193557 (2011).
72. Morris, G. E. Parker, L. C., Ward, J. R., Jones, E. C., Whyte, M. K. B. Brightling, C. E., Bradding, P., Steven K. Dower, and Ian Sabroe, Cooperative molecular and cellular networks regulate Toll-like receptor-dependent inflammatory responses, The FASEB Journal, Published Online:25 Aug 2006, https://doi.org/10.1096/fj.06-5910fje (2006).
73. Ma-Lauer, Y., Carbajo-Lozoya J., Hein, m. y. . Müller, M. A., Deng, W., J Lei, J., Meyer, B., Kusov, Y., von Brunn, B., Bairad, D. R., Hünten, S., Drosten, C., Hermeking, H., Leonhardt, H., Mann, M., Hilgenfeld, R., and von Brunnb, A., p53 down-regulates SARS coronavirus replication and is targeted by the SARS-unique domain and PL^{pro} via E3 ubiquitin ligase RCHY1, Proceedings of the National Academy of Sciences, Micorbiology, 113(35): E5192–E5201, (2016)
74. Hoffmann, M., Kleine-Weber, H,, Schroeder, S., Krüger, N., Herrler, T., Nai-Huei Wu, N-H, , A., Müller, M. A., Drosten, C., Stefan Pöhlmann, S., SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor, https://doi.org/10.1016/j.cell.2020.02.052, Available online 5 March 2020, Cell Press,(2020)
75. Di Guardo, Rapid Response (Comment): SARS-CoV-2, hypertension and ACE inhibitors, theBMJ, https://www.bmj.com/content/368/bmj.m606/rr-10 (2020).
76. Loeffler, J. R., Fernandez-Quintero, M., Schauperl. M., and Liedl M. K. R., STACKED – Solvation Theory of Aromatic Complexes as Key for Estimating Drug binding, J. Chem. Inf. Model. (in press) https://doi.org/10.1021/acs.jcim.9b01165
77. Fulford, T. and Lee, D., The Jenneration of Disease: Vaccination, Romanticism, and Revolution, Studies in Romanticism, The Johns Hopkins University Press, 39(1), 139-163, (2000).

78. Hamming, I., Timens., W, Bulthuis, M. L., Lely, A. T., Navis, G, van Goor, H.,Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis, Journal of Pathology. 2203(2):631-7 (2004).
79. Yang, X., Hua, W,, Ryu, S., Yates, P., Chang, C., Zhang, H., Di, L., 11β-Hydroxysteroid Dehydrogenase 1 Human Tissue Distribution, Selective Inhibitor, and Role in Doxorubicin Drug Metabolism and Disposition, 46(7):1023-1029 (2018). doi: 10.1124/dmd.118.081083. (2018).
80. Bruley, C., Lyons, V., Worsley, A. G. F., Wilde, M. D. Darlingtion, G. D., Morton, N. M., Seckl, J. R., Chapman, K. E., A Novel Promoter for the 11β-Hydroxysteroid Dehydrogenase Type 1 Gene Is Active in Lung and Is C/EBPα Independent,Charlotte Bruley, Val Lyons, Alan G. F. Worsley, Margaret D. Wilde, Gretchen D. Darlington, Nik M. Morton, Jonathan R. Seckl, Karen E. Chapman, *Endocrinology*, 147: 6, Pages 2879–2885, https://doi.org/10.121, (2006).
81. Gurwitz, D., Angiotensin receptor blockers as tentative SARS-CoV-2 therapeutics, DDR, Wiley, https://onlinelibrary.wiley.com/doi/pdf/10.1002/ddr.21656, DOI: 10.1002/ddr.21656 (2020).

- This paper “drills down” into the studies of the author’s previous COVID-19 paper.
- Designing vaccine and drugs must seek to avoid escape mutations.
- Subsequence KRSFIEDLLFNKV seems recognizable across many coronaviruses.
- The ACE2 binding domain is a target, but shows variation.
- A steroid dehydrogenase is argued to remain an interesting model pharmacophore.

This paper is provided to the community to promote the more general applications of the thinking of Professor Paul A. M. Dirac in human and animal medicine in accordance with the charter of The Dirac Foundation , to emphasize the advantages and simplicity of the basic form of the Hyperbolic Dirac Net, to encourage its use, and to propose at least some of the principles of the associated Q-UEL, a universal exchange language for medicine, as a basis for a standard for interoperability. These mathematical and engineering principles are used, amongst many others in an integrated way, in the algorithms and internal architectural features of the BioIngine.com, a distributed system developed by Ingine Inc. Cleveland, Ohio, for the mining of, and inference from, Very Big Data for commercial purposes.



Barry Robson BSc(Hons) PhD DSc Professor Emeritus
Harvard-Macy Dip. Med. Ed.
Dist. Sci. (admin) U. Wisconsin-Stout,
IBM Dist. Eng. Alumnus.

Biography

Dr. Barry Robson was five years as Chief Scientific Officer IBM Global Healthcare, Pharmaceutical, and Life Sciences and, prior to that, six years as the Strategic Advisor at IBM Global Research Headquarters (T. J. Watson Research Center). For most of those 11 years he held the prestigious title of IBM Distinguished Engineer. According to Barry's two page biography written by journalist Brendan Horton in Nature (389,418-420, 1997), Barry was a pioneer in bioinformatics, protein modeling, and computer-aided drug design. He is the recipient of several honors including the Asklepios Award for Outstanding Vision in Science and Technology at the Future of Health Technology Congress at M.I.T. in 2002. He has helped start up several other companies or divisions in the UK and USA. Barry continues as CEO of The Dirac Foundation in the UK, and Distinguished Scientist at the University of Wisconsin-Stout Department of Mathematics, Statistics, and Computer Science. While continuing to work for, and then collaborate with IBM, he was also University Research Director and Professor of EBM, Biostatistics and Epidemiology at St. Matthew's University School of Medicine which he helped established in its earliest days in the Cayman Islands. Immediately prior to joining IBM in 1998 he was hired as Principal Scientist at MDL Information Systems in California to help put together the technology for the multimillion sale of a bioinformatics system to the holding company forming Craig Venter's Celera Genomics that produced the first draft of the human genome. Prior to that, he was CSO of Gryphon Sciences (later Gryphon Pharmaceuticals) in South San Francisco, California, a bio-nanotechnology ultrastructural chemistry start-up largely held and then acquired by SmithKline Beecham. Before moving to the US, Barry was the scientific founder of Proteus International plc in the UK, designing and leading the development of the PROMETHEUS Expert System and its underlying GLOBAL Expert System, bioinformatics and simulation language for drug, vaccine, and diagnostic discovery. It sold for the equivalent of $9.4 million to the pharmaceutical industry in the mid-1990s. At Proteus, he also led the team that used the above Expert System to invent and patent several diagnostics and vaccines including the Mad Cow disease diagnostic subsequently marketed worldwide by Abbott. He has some 300 scientific publications including some 50 patents and two books: "The Engines of Hippocrates. From the Dawn of Medicine to Medical and Pharmaceutical Informatics" Robson and Baek, 2009, Wiley, 600 pages)" and "Introduction to Proteins and Protein Engineering" (B. Robson and J. Garnier, 1984, 1988, Elsevier, 700 pages). He has contributed to several reports to governments including Panels of the National Innovation Initiative including "Innovate America" Published by The Council on Competitiveness, Washington D.C. (2004) as a whitepaper to the President of the United States. For five years, Barry was a Nature "News and Views" Correspondent on biomolecules.

Recent Papers

1. Robson, B., Li, J., Dettinger, R., Peters, A., and Boyer, S.K. (2011), Drug discovery using very large numbers of patents. General strategy with extensive use of match and edit operations. Journal of Computer-Aided Molecular Design 25(5): 427-441 (2011)

2. Robson, B. (2012) “Towards Automated Reasoning for Drug Discovery and Pharmaceutical Business Intelligence”, Pharmaceutical Technology and Drug Research, 2012 1: 3 (27 March 2012)

3. Robson, B. (2013)“Towards New Tools for Pharmacoepidemiology”, Advances in Pharmacoepidemiology and Drug Safety, 1:6, http://dx.doi.org/10.4172/2167-1052.100012

4. Robson, B (2013) “The Concept of Novel Compositions of Matter. A Theoretical Analysis.” Intellectual Property Rights , Intel Prop Rights 1:108. doi: 10.4172/ipr.1000108

5. Robson, B., Caruso, T, and Balis, U. G. J. (2013)”Suggestions for a Web Based Universal Exchange and Inference Language for Medicine”, Computers in Biology and Medicine, 1;43(12):2297-310. doi: 10.1016/j.compbiomed.2013.09.010. Epub 2013 Sep 20.Also found in preliminary form, with permission of the Editor-in-Chief, at the US Government S&I website:- http://wiki.sifframework.org/file/view/UELRobson102corrections.pdf/451304614/UELRobson102corrections.pdf

6. Robson, B. (2014) “ Hyperbolic Dirac Nets for Medical Decision Support. Theory, Methods, and Comparison with Bayes Nets” *Computers in Biology and Medicine*, 51:183-97.

7. Robson, B. (2014) “POPPER, a Simple Programming Language for Probabilistic Semantic Inference in Medicine.” *Computers in Biology and Medicine*, 56: 107-123

8. Robson, B., Caruso, T, and Balis, U. G. J. (2014) “Suggestions for a Web Based Universal Exchange and Inference Language for Medicine. Continuity of Patient Care with PCAST Disaggregation. *Computers in Biology and Medicine*, 56: 51–66.

9. Deckelman, S., and Robson, B. (2015). “Split-Complex Numbers and Dirac Bra-Kets”, Vol. 14:3, 135-149, Communications in Information and Systems (CIS).

10. Robson, B. and Boray, S. (2015). “Implementation of a web based universal exchange and inference language for medicine. Sparse data, probabilities and inference in data mining of clinical data repositories”, Computers in Biology and Medicine, 66, 82-102.

11. Robson, B. and Boray, S (2015). “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 , 1397-1404, IEEE.

12. Robson, B. and Boray, S (2016)” 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”, Computers in Biology and Medicine, 73:71-93

13. Robson, B. and Boray, S. (2016) “Studies of the Role of a Smart Web for Precision Medicine Supported by Biobanking”, Personalized Medicine, FTG (accepted subject to satisfactory revision).

14. Robson B. (2016), “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”, Computers in Biology and Medicine, 79, Pages 299–323.

15. Robson B. and Boray, S. (2016), “Studies in the Extensively Automatic Construction of Large Odds-Based Inference Networks from Structured Data. Examples from Medical, Bioinformatics, and Health Insurance Claims Data”, Computers in Biology and Medicine, in press (2018)