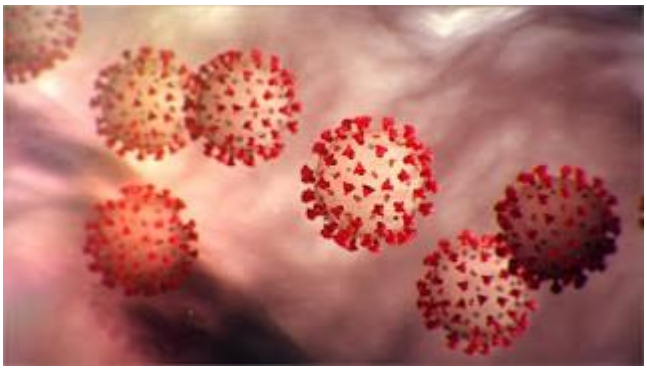


COVID-19: VARIANTS VS VACCINES

Copyright Terrence McGarty 2021

TERRENCE MCGARTY

ABSTRACT

Viral variants have been developing in COVID-19 and especially in the spike protein. In order to address these changes one must also modify the vaccines currently being produced. There are several ways to do this. One is the classic post hoc manner of monitoring what is produced and then address it. The second extreme is pre hoc, anticipating what most likely will occur and vaccinate against this anticipated variant. We present a proposal for a Bayesian pre hoc approach to vaccine development with COVID-19 variants.

TGL 186 MARCH 2021

Notice

This document represents the personal opinion of the author and is not meant to be in any way the offering of medical advice or otherwise. It represents solely an analysis by the author of certain data which is generally available. The author furthermore makes no representations that the data available in the referenced papers is free from error. The Author also does not represent in any manner or fashion that the documents and information contained herein can be used other than for expressing the opinions of the Author. Any use made and actions resulting directly or otherwise from any of the documents, information, analyses, or data or otherwise is the sole responsibility of the user and The Author expressly takes no liability for any direct or indirect losses, harm, damage or otherwise resulting from the use or reliance upon any of the Author's opinions as herein expressed. There is no representation by The Author, express or otherwise, that the materials contained herein are investment advice, business advice, legal advice, medical advice or in any way should be relied upon by anyone for any purpose. The Author does not provide any financial, investment, medical, legal or similar advice in this document or in its publications on any related Internet sites.

Furthermore, this document contains references to and quotes from papers and documents under the premise of Fair Use in order to present ideas and understandings in context. The Author has attempted to make any and all references to such material separate from those of the author per se and has referenced the source expressly in all cases. These documents are for the dissemination of ideas and have no commercial intent. The Author would appreciate any communications relating to these documents and these should be sent to:

tmcgarty@telmarc.com.

Terrence P. McGarty, Copyright © 2021, all rights reserved. This document is in DRAFT form and is solely for technical review and evaluation and it not intended for any commercial use.

Table of Contents

1 INTRODUCTION

"Prior Planning Prevents Poor Performance"

This was told me by my father early on while growing up. It has been with me ever since. In the case of COVID-19 it is clear that many entities failed grossly in any planning at all, especially such entities as the CDC, whose virus test kits were a gross failure, not to mention their monitoring network. However, what may prove even mor critical is understanding the putative mutations of the virus and anticipating what they may most likely be and take pre-emptive steps in mitigating for them. The mRNA vaccines have proven highly efficacious and also readily modifiable. Thus, if we can look at the most likely spike variants, which we contend can be accomplished, then by incorporating them in vaccines we can pre-emptively prevent future outbreaks. We examine the elements regarding this goal herein[1].

COVID-19 has become a global pandemic. With the introduction of vaccines, the initial virus may have a chance to be controlled. However, mutants or variants, are being generated naturally at a fast rate and some are getting released by and into the existing pool of infected patients and spreading as a secondary pandemic. The question posed is: What can be done to address this secondary spread of new variants?

The classic approach is to do a wait and see approach, seeing new infections, ascertain the variant, and then modify or enhance the vaccine to address that new variant. We see this in annual flu immunizations. Often this works well but there are times when one fails to see a new variant or the new variant arises after a set number of newer variants have been selected for vaccine preparations.

A second approach may be possible. We examine this approach herein. Namely, we know several facts:

1. COVID spike protein comes from an mRNA single stranded segment which is about 3000 nucleotides long generating a protein of about 1000 nucleic acids in length. We are being simplistic at this point.

2. COVID spike protein targets the ACE2 receptor on cells thus enabling entry, replication and immune responses.

3. There are multiple ACE2 receptor variants so that the spike may be more infectious in some humans than others.

4. We generally know what the pool of ACE2 receptors look like as proteins and where their binding to the spike occurs.

[1] The author is not an expert on proteins. Thus, this document is merely exploratory and may be subject to significant modifications. The author is however competent in the various elements of pattern recognition and AI applications to problems similar to what we present herein. The intent of this document is to provoke ideas regarding pre-emptive determination of viral evolutionary threats and act accordingly.

5. We now know several COVID variants that attach to ACE2 receptors. We know or can know the nature of the spike/ligand binding sites. We know or can know the profiles of COVID variants and ACE2R variants and their binding strengths.

6. We now know the putative genetic variants that are of concern and arguably we can ascertain the genetic changes that led from the wild type originally identified to each variant.

7. We have many powerful computational tools to examine proteins, asses structure from RNA elements, and consider mutational changes. The challenge here may be a bit simpler since we know the ab initio protein structures and thus we deal with a boundary value problem having COVID and its variants on one end and ACE2R and its variants on the other.

Thus, we pose the following challenge:

1. **KNOWING BOUNDARY CONDITION PROTEINS FOR SPIKE AND ACE2, AND ASSUMING STABILITY IN ACE2 VARIANTS, THEN WHAT MUTATIONS OF THE SPIKE WILL LEAD TO BINDINGS TO THESE ACE2?**

2. **FURTHERMORE, UNDERSTANDING THESE MUTATIONS CAN WE RANK ORDER THEM FROM STRONGEST BINDING DOWNWARD?**

If we can achieve the two issues noted above, then it would be reasonable to anticipate the next set of most likely mRNAs for spikes and in turn use these anticipated spikes pre-emptively in a new vaccine profile.

If correct, we can then develop a significant tamping down of infections prior to any substantial spread.

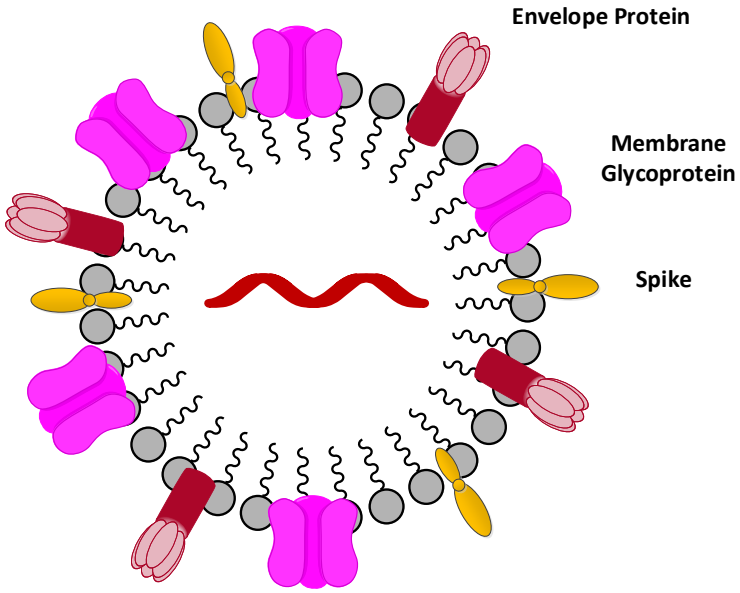
We thus examine these assumptions herein looking at what is currently available and what needs to be developed.

2 COVID-19

We here provide a brief summary of the key issues regarding COVID-19. We present a brief summary of the COVID-19 corona virus which allegedly originated in some manner from Wuhan, China[2]. The actual provenance of the virus is yet to be adequately determined. The virus has a very high level of transmissibility, virulence, and mutability. In a short period, the virus has aggressively spread globally and strangely appears to have had minimal impact in China, its source of origin.

2.1 OVERVIEW

The figure below is a graphic example of the corona virus in generic terms. It is a single stranded positive RNA virus with a well-defined spike protein on the surface. The spike protein is a putative target for immunization efforts.



We now consider some detail as regards to this virus. As Artika et al note:

The schematic diagram of coronavirus life cycle.

*The coronavirus infection is initiated by the **binding of the virus particles to the cellular receptors** leading to viral entry followed by the viral and host cellular membrane fusion.*

*After the membrane fusion event, the **viral RNA is uncoated in the host cells** cytoplasm.*

[2] In March 2020 we prepared a preliminary report on thea virus discussing what was known to that data and examining the possible pandemic dynamics. In February 4, 2020 we declared this a pandemic despite the WHO and CDC delaying any prophylactic actions, the delay thus resulting in the current global pandemic, https://www.telmarc.com/Documents/White%20Papers/173Corona.pdf

*The **ORF1a and ORF1ab** are translated to produce pp1a and pp1ab, which are subsequently processed by the **proteases encoded by ORF1a** to produce 16 non-structural proteins (nsps) which form the **RNA replicase–transcriptase complex (RTC)**.*

This complex localizes to modified intracellular membranes which are derived from the rough endoplasmic reticulum (ER) in the perinuclear region, and it drives the generation of negative-sense RNAs ((–) RNAs) through both replication and transcription.

*During replication, the **full-length (–)RNA copies of the genome are synthezied** and used as templates for the production of **full-length (þ)RNA genomes**.*

During transcription, a subset of 7–9 subgenomic RNAs, including those encoding all structural proteins, is produced through discontinuous transcription. In this process, subgenomic (–)RNAs are synthesized by combining varying lengths of the 30end of the genome with the 50 leader sequence necessary for translation.

These subgenomic (–)RNAs are then transcribed into subgenomic (þ)mRNAs.

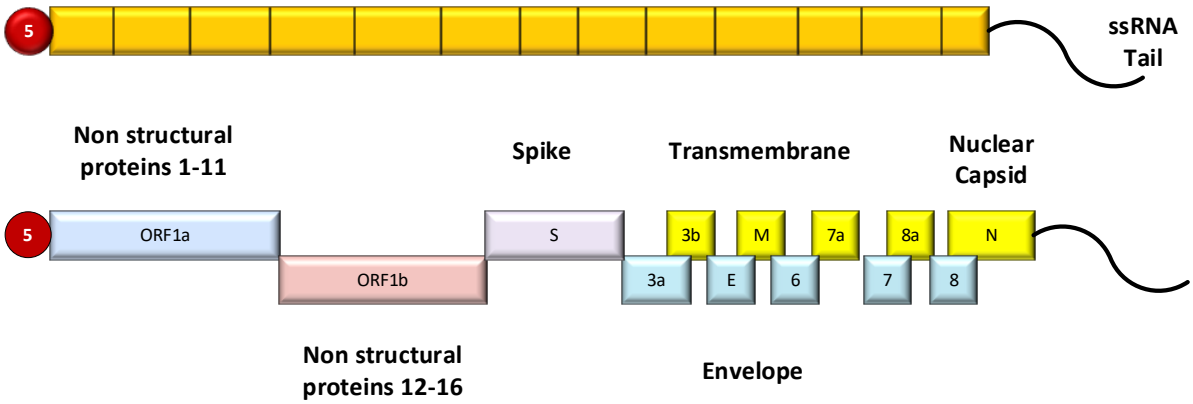
*The **subgenomic mRNAs are then translated**.*

The generated structural proteins are assembled into the ribonucleocapsid and viral envelope at the ER–Golgi intermediate compartment (ERGIC),

followed by release of the newly produced coronavirus particle from the infected cell

2.2 COVID GENE

Various authors have discussed the general structure of a corona virus gene structure and we present this below. It is a positive single stranded mRNA virus and the mRNA has a form as shown below. It is approximately 30,000 nucleotides in length and the spike protein is approximately 3,000 nucleotides in length. As with most corona viruses it contains in the mRNA, via the non-structural proteins, the ability to reconstruct itself many times over by generating the structural genes and implanting a new copy of the mRNA.



The nonstructural proteins, NSP, for the replicase transcriptase complex. There are four structure proteins:

1. S is the spike forming protein, of which we shall speak of later in detail

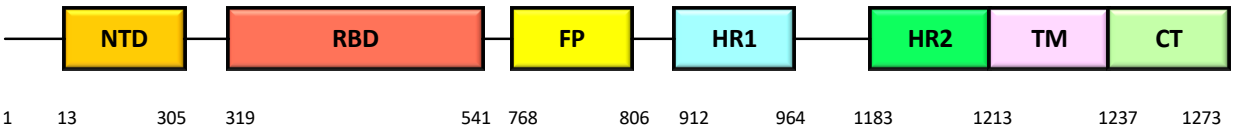
2. E is the envelope protein for the new virion

3. M is the membrane protein for the new virion

4. N is the nucleocapsid protein for the new virion

Overall, we now have the two sets; those allowing for self-reproduction and those relating to construction of the new virion.

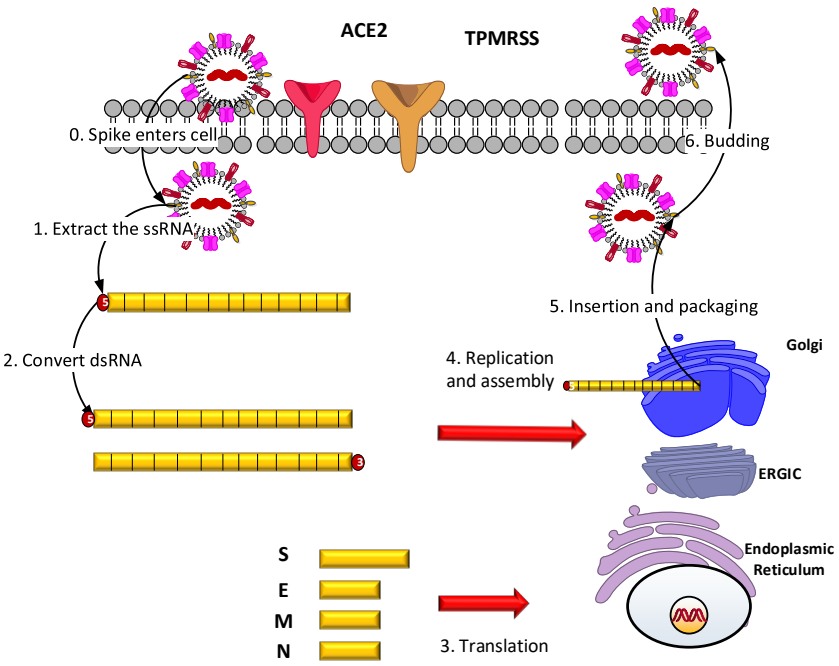
The following is the detail of the spike gene which we shall discuss later.



When examining these proteins we must note that the numbers above are those of the nucleic acids whereas the nucleotides are three times the number. Thus the 1273 nucleic acids represent almost 3900 nucleotides. Mutations in any of these nucleotides can result in a mutation of the nucleic acids and thus the protein conformation. The work by Huo et al provides a significant amount of detail in the spike protein.

2.3 COVID System

We depict the process below for the entry and reproduction of the virion. This process starts with the attachment and entry of the virion and ultimately the release of a collection of new virions until the cell is depleted and dies.



We can summarize the steps as shown above.

1. The virion spike protein attaches to the ACE2 receptor protein. As we will note later this means that there is a selective capability in this specific spike protein. As we shall note from the literature later, strangely this receptor is weak in Chinese individuals and is a strong bond in Europeans and most other ethnicities. We discuss this at length later.
2. The virion then enters the cell and disassembles. The ss mRNA is extracted
3. The structure RNA sequences are then translated in the inter ER/Golgi space into the constituent proteins.
4. The ss mRNA is completed to form a complete ds mRNA which will be used to generate multiple ss mRNAs to be inserted in other new virions. This may be a point for possible mutant changes.
5. The multiple ss mRNAs and the structural proteins are assembled in the Golgi apparatus and extracted.
6. The final result is a repackaged virion which is budded outwards.

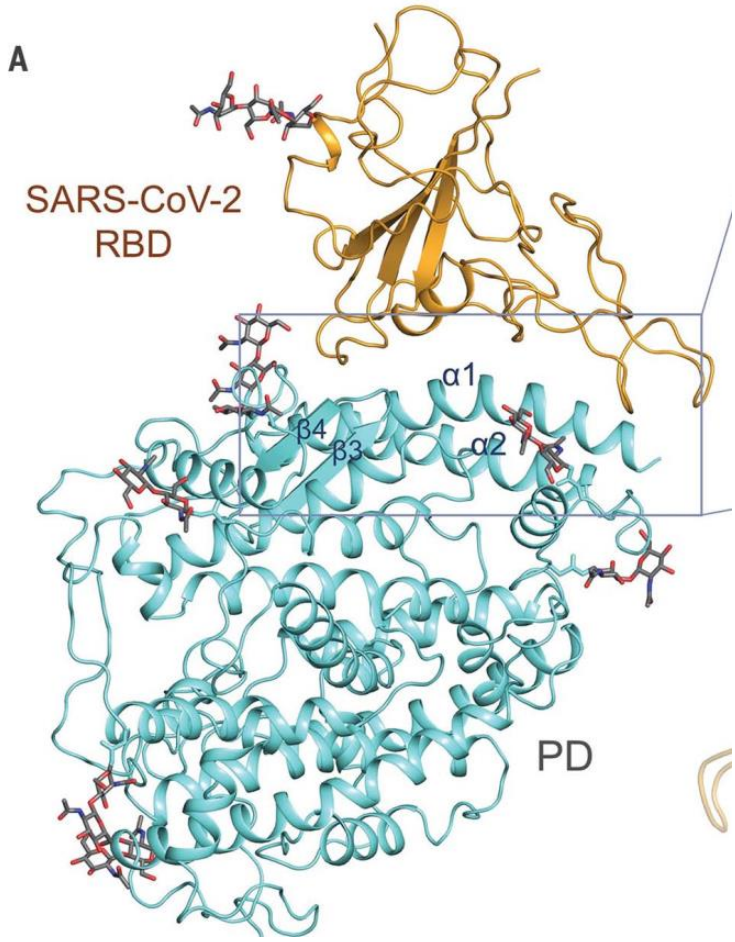
These steps are very general and there are as of yet many holes as to exactly how all this is accomplished. Yet for our purposes the processes lay out the locations of the possible mutation sites.

2.4 ACE2 AND COVID-19

It is now generally agreed that the virus enters a cell via access to the ACE2 receptor site. We shall discuss ACE2 and its functions as well as its variants in some detail in the next section but

we summarize some issue here. Jia et al noted early on in 2005 the nexus between ACE2 and corona virus infection.

In an early paper by Yan et al the authors discussed the binding of the virus to ACE2. The image of the binding sites from the Yan et al paper is shown below. RBD is the receptor binding domain of the virus spike and PD is the peptidase domain of the ACE2 protein. The detailed binding is show below and the Yan et al paper details their study on this binding.



Additional analysis has also been done by Lan et al adding significant details. Also, Samavati and Uhal have noted:

SARSCoV-2 utilizes a novel metallocarboxyl peptidase angiotensin receptor (ACE) 2 to gain entry into human cells. Similar to other CoV, during viral entry into the host cell, the spike proteins (S) on the envelope of SARS-CoV-2 are cleaved into S1 and S2 subunits. S2 does not interact with the receptor but it harbors the functional elements required for membrane fusion of the virion. The S1 protein/receptor interaction is the pivotal determinant for SARS-CoV-2 to infect a host species.

*S1 contains the **receptor binding domain (RBD)** and directly binds to the **peptidase domain (PD)** of ACE 2 to gain entry into host cells. Despite high similarity between the RBD of SARS-*

CoV and SARS-CoV-2, several amino acid variations are observed in the middle of the binding domain of SARS-CoV- 2, which provide an increased affinity to bind to ACE2 more effectively.

In our proposed study, it will be critical to understand several key issues.

1. The spike mutations from wild type may strongly influence the RBD structure by changing the binding characteristics.

2. The ACE2 variants from wild type may likewise influence the PD site binding.

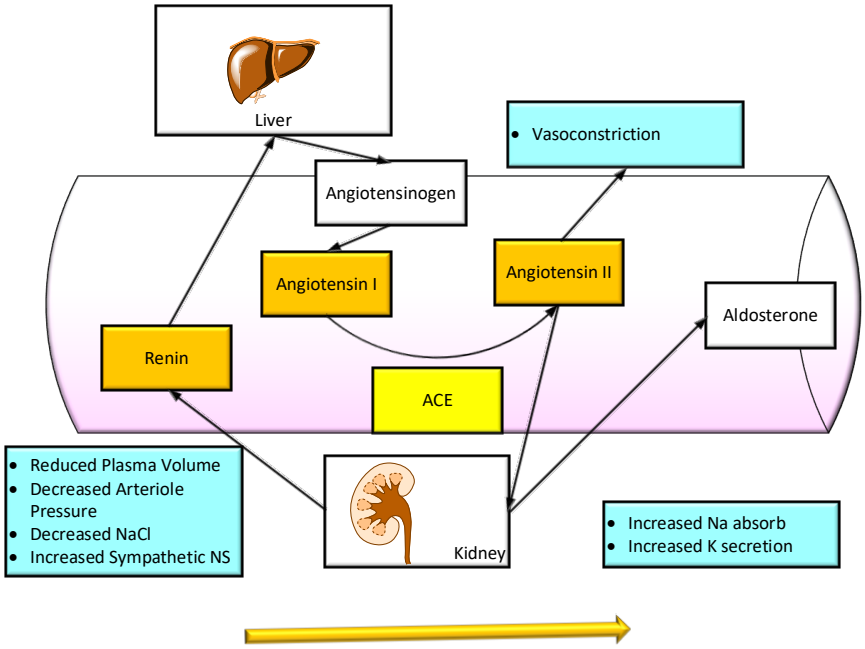
3. The aggressiveness of the virus will depend on the matching that occurs as a result of the above two.

3 ACE2 RECEPTORS

ACE and ACE2 are two key enzymes involved in homeostasis. ACE activation can result in increased blood pressure often mitigated by ACE inhibitors. ACE2 counters these effects. It is ACE2 that is targeted by COVID-19 and thus not only does it enter the cell but it inhibits the stabilizing effects of ACE2. ACE2 is one of several cellular essential protein receptors and catalysts which are involved in homeostasis. We first examine ACE and then ACE2.

3.1 ACE

We first review the ACE pathway effects and then continue on to those of ACE2. The classic renin-angiotensin-aldosterone path is shown below wherein vasoconstriction is a prominent result.

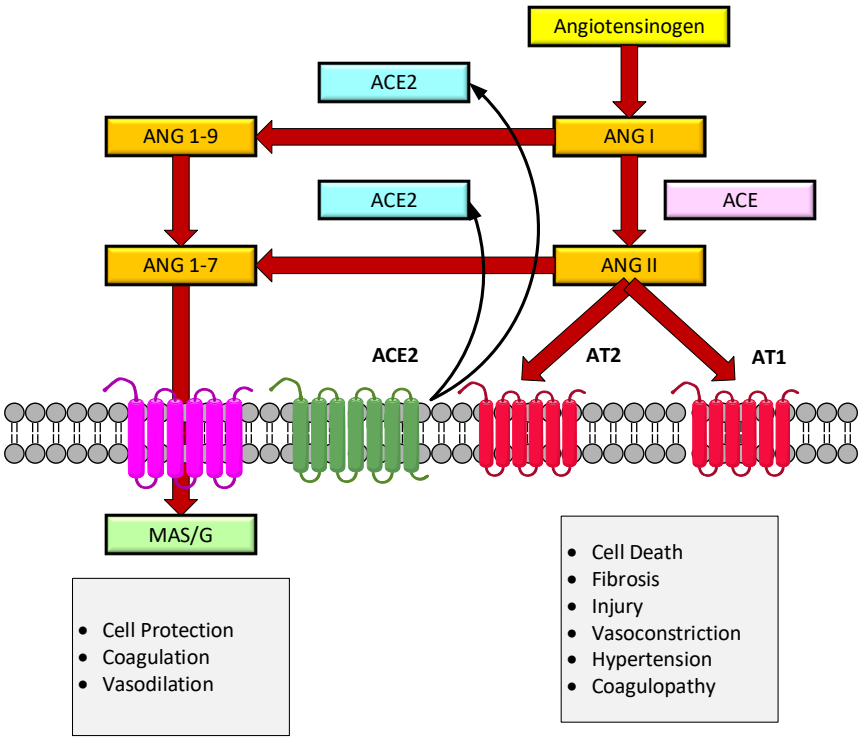


Namely, ACE, the angiotensin enzyme, facilitates the conversion of angiotensin I to angiotensin II. The latter then produces vasoconstriction and is a frequent driver for hypertension. It can also impact the renal system causing increased Na absorption and K secretion. This renin-angiotensin-aldosterone system is generally well understood. As we shall note, ACE2 provides a counterbalance to this system as well as a target for COVID-19 wild type.

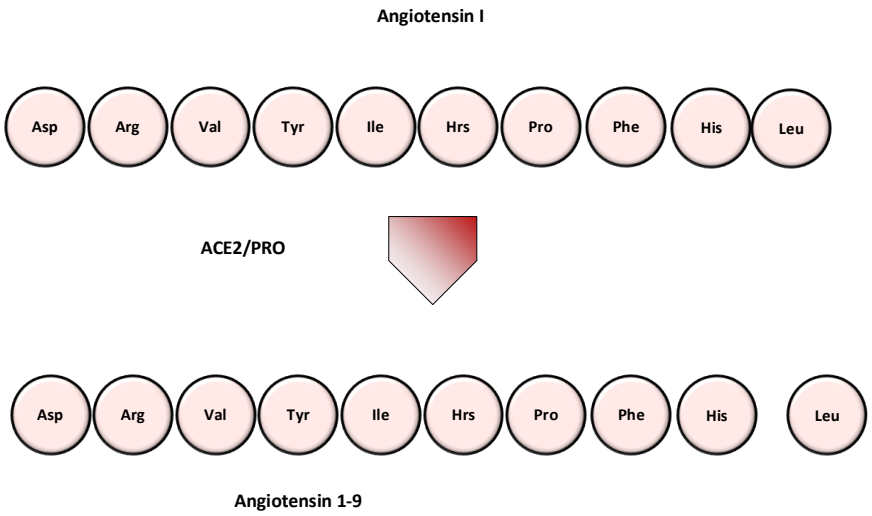
3.2 ACE2

The ACE2 gene was discovered some twenty years ago in a study of the cardiovascular system. The paper by Marian details some of the issue related to its discovery and is well worth reading. It was the ability to understand Ang 1-7 and Mas as drivers for improved homeostasis. ACE2 can

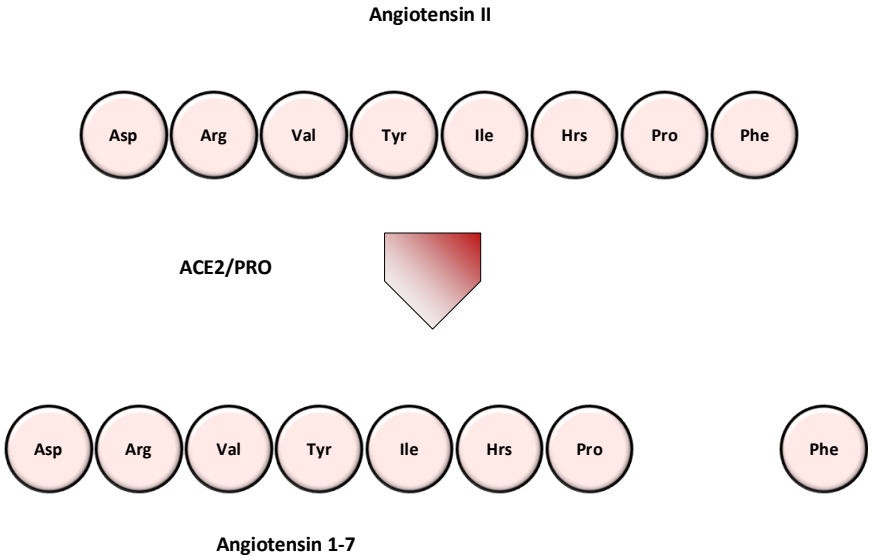
be seen as a critical gene not only for cardiovascular functions but a wide variety of systemic homeostasis. Thus it is critical to understand that any virus that attacks via ACE2 not only allows for immunological effects but also drives compromise of many other functions. Now we can view the homeostasis associated with ACE and ACE2. We show this below:



In the above we see the production of ANG 1-9 and ANG 1-7 which control MAS/G have the structures we depict below.



Following the above, we can then see how Ang 1-7 is readily obtained as shown below.



As to the above Santos et al have noted the nexus between angiotensin and the 1-7 interface with Mas via ACE2, namely:

The renin–angiotensin system plays a critical role in blood pressure control and body fluid and electrolyte homeostasis. Besides angiotensin (Ang) II, other Ang peptides, such as Ang III [Ang-(2–8)], Ang IV [Ang-(3–8)], and Ang-(1–7) may also have important biological activities. Ang-(1–7) has become an angiotensin of interest in the past few years, because its cardiovascular and baroreflex actions counteract those of Ang II. Unique angiotensin-binding sites specific for this heptapeptide and studies with a selective Ang-(1–7) antagonist indicated the existence of a distinct Ang-(1–7) receptor.

We demonstrate that genetic deletion of the G proteincoupled receptor encoded by the Mas protooncogene abolishes the binding of Ang-(1–7) to mouse kidneys. Accordingly, Mas-deficient mice completely lack the antidiuretic action of Ang-(1–7) after an acute water load. Ang-(1–7) binds to Mas-transfected cells and elicits arachidonic acid release. Furthermore, Mas-deficient aortas lose their Ang-(1–7)-induced relaxation response. ***Collectively, these findings identify Mas as a functional receptor for Ang-(1–7) and provide a clear molecular basis for the physiological actions of this biologically active peptide.***

Now ACE2 has significant properties. As Samavati and Uhal note:

In addition to its protective role in the cardiovascular system, ACE2 has a direct protective role in alveolar epithelial cells. In the lungs ACE2 has numerous physiological functions, most of which are protective against lung injury. Similar to the endothelial site, ACE2 degrades the octapeptide Ang II by removing a single amino acid from the C-terminal end of the peptide to generate the heptapeptide Ang1-7. … ACE2 protects against lung injury by: (a) degrading Ang II, which is vasoconstrictive and proapoptotic for lung epithelial cells and profibrotic, and (b) by

producing the peptide Ang1-7, which inhibits the actions of Ang II through binding to the MAS receptor. In support of this protective role for ACE2, pharmaceutical preparations of recombinant ACE2, when administered to experimental animals, protect against lung cell death, inhibit acute lung injury and prevent lung fibrosis after chronic injury to the lungs. As further evidence, the application of a specific competitive inhibitor of ACE2, DX600, to primary cultures of isolated ACEs increases the level of Ang II released into the serum-free culture medium by autocrine mechanisms, reduces the amount of released Ang1-7 and, importantly, induces apoptosis inhibitable by the AT1 receptor blocker.

Thus, functional ACE2 normally expressed by alveolar epithelial cells can be viewed as a critical survival factor for these lung cells. In addition, the enzymatic product of ACE2, the Ang1-7, itself protects against lung cells death by antagonizing that actions of Ang II.

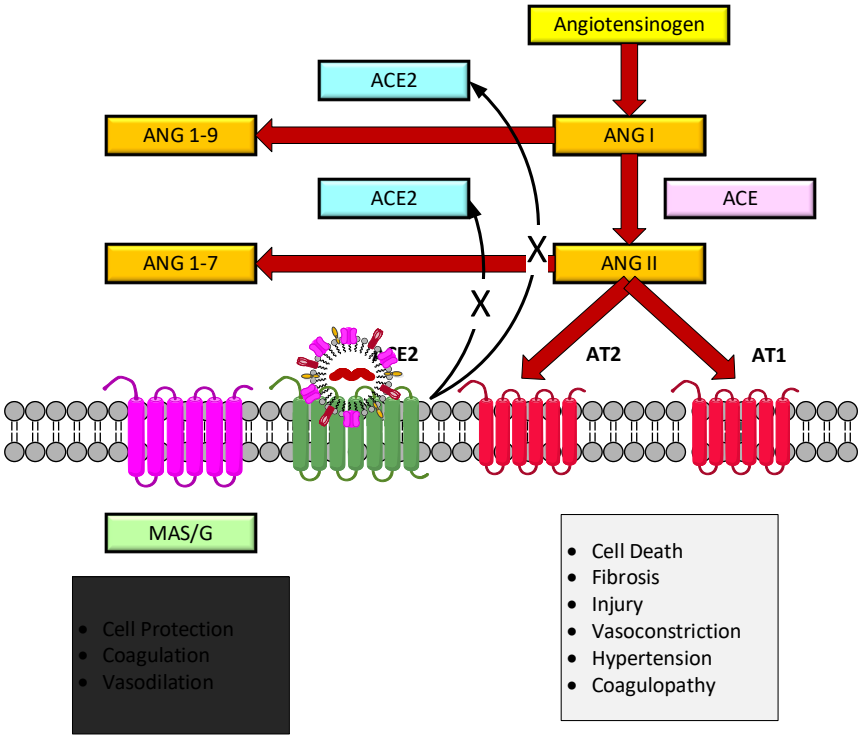
As Simoes e Silva et al note

First, two independent research groups reported simultaneously in 2000 the existence and characterization of an enzyme homologue to ACE, the ACE2, which was established later as the main Ang-(1-7)-forming enzyme.

Second, … the ***G-protein coupled Mas*** *is a functional receptor for Ang- (1-7).*

Thus, Ang-(1-7) is now considered a biologically active member of the RAS, which binds to Mas inducing many beneficial actions, such as vasodilation, inhibition of cell growth, anti-thrombosis and antiarrhythmogenic effects. Ang-(1-7) is produced mainly through the action of ACE2, which has approximately 400-fold less affinity to Ang I than to Ang II.

Therefore, Ang II is the major substrate for Ang- (1-7) synthesis. ACE2 can also form Ang- (1-7) less efficiently through hydrolysis of Ang I to Ang-(1–9) with subsequent Ang-(1-7) formation. It is important to highlight the key role of ACE2 in this new concept of the RAS since it degrades the vasoconstrictive/ proliferative peptide Ang II to form the vasodilator/ antiproliferative heptapeptide Ang-(1-7). This is a strategic finding that may be exploited for therapeutic purposes



ACE2R are cell surface proteins that bind ACE2, the angiotensin enzyme 2. There are two noted ACE2Rs recognized.

The following Table summarizes all of the above elements in some detail.

Protein	*Function*	*Location*
ACE2[3]	*The protein encoded by this gene belongs to the angiotensin-converting enzyme family of dipeptidyl carboxydipeptidases and has considerable homology to human angiotensin 1 converting enzyme. This secreted protein catalyzes the cleavage of angiotensin I into angiotensin 1-9, and angiotensin II into the vasodilator angiotensin. ACE2 is known to be expressed in various human organs, and its organ- and cell-specific expression suggests that it may play a role in the regulation of cardiovascular and renal function, as well as fertility.* ***In addition, the encoded protein is a functional receptor for the spike glycoprotein of the human coronavirus HCoV-NL63 and the human severe acute respiratory syndrome coronaviruses, SARS-CoV and SARS-CoV-2, the latter is the causative agent of coronavirus disease-2019 (COVID-19). Multiple splice variants have been found for this gene and the dACE2 (or MIRb-ACE2) splice variant has been found to be interferon inducible***	Xp22.2

[3] https://www.ncbi.nlm.nih.gov/gene/59272

Protein	*Function*	*Location*
ACE[4]	This gene encodes an enzyme involved in blood pressure regulation and electrolyte balance. It catalyzes the conversion of angiotensin I into a physiologically active peptide angiotensin II. Angiotensin II is a potent vasopressor and aldosterone-stimulating peptide that controls blood pressure and fluid-electrolyte balance. This angiotensin converting enzyme (ACE) also inactivates the vasodilator protein, bradykinin. Accordingly, the encoded enzyme increases blood pressure and is a drug target of ACE inhibitors, which are often prescribed to reduce blood pressure. This enzyme additionally plays a role in fertility through its ability to cleave and release GPI-anchored membrane proteins in spermatozoa. Many studies have associated the presence or absence of a 287 bp Alu repeat element in this gene with the levels of circulating enzyme. This polymorphism, as well as mutations in this gene, have been implicated in a wide variety of diseases including cardiovascular pathophysiologies, psoriasis, renal disease, stroke, and Alzheimer's disease. **Regulation of the homologous ACE2 gene may be involved in progression of disease caused by several human coronaviruses, including SARS-CoV and SARS-CoV-2. Alternative splicing results in multiple transcript variants encoding both somatic (sACE) and male-specific testicular (tACE) isoforms.**	17q23.3
Ang I[5]	**Angiotensin I is a ten amino acid peptide formed by renin cleavage of angiotensinogen. Angiotensin I has no direct biological function except that high levels can stimulate catecholamine production.** It is metabolized to its biologically active byproduct angiotensin II, a potent vasoconstrictor, by angiotensin converting enzyme (ACE) through cleavage of the two terminal amino acids. It has a role as a neurotransmitter agent and a human metabolite. It is a tautomer of an angiotensin I dizwitterion.	

[4] https://www.ncbi.nlm.nih.gov/gene/1636

[5] https://pubchem.ncbi.nlm.nih.gov/compound/Angiotensin-I

Protein	*Function*	*Location*
Ang II[6]	**an octapeptide that is a potent but labile vasoconstrictor.** it is produced from angiotensin i after the removal of two amino acids at the c-terminal by angiotensin converting enzyme. the amino acid in position 5 varies in different species. to block vasoconstriction and hypertension effect of angiotensin ii, patients are often treated with ace inhibitors or with angiotensin ii type 1 receptor blockers.	
Ang 1-9[7]	Angiotensin (1-9) is a **nine amino acid peptide** which is formed when angiotensin converting enzyme 2 (ACE2) hydrolyzes the carboxy terminal leucine from angiotensin I. It is an anti-cardiac hypertrophy agent. **It has a role as a human metabolite, a rat metabolite, an antihypertensive agent and a cardioprotective agent. It is a tautomer of an angiotensin (1-9) dizwitterion.**	
Ang 1-7[8]	Ile(5)-angiotensin II (1-7) is an angiotensin compound consisting of the linear heptapeptide sequence L-Asp-L-Arg-L-Val-L-Tyr-L-Ile-L-His-L-Pro. **It has a role as a vasodilator agent. It is a tautomer of an Ile(5)-angiotensin II (1-7) dizwitterion.**	
Mas[9]	This gene encodes a class I seven-transmembrane G-protein-coupled receptor. The encoded protein is a receptor for angiotensin-(1-7) and preferentially couples to the Gq protein, activating the phospholipase C signaling pathway. The encoded protein may play a role in multiple processes including hypotension, smooth muscle relaxation and cardioprotection by mediating the effects of angiotensin-(1-7).	6q25.3
AT1 or AGTR1[10]	*Angiotensin II is a potent vasopressor hormone and a primary regulator of aldosterone secretion. It is an important effector controlling blood pressure and volume in the cardiovascular system. It acts through at least two types of receptors. This gene encodes the type 1 receptor which is thought to mediate the major cardiovascular effects of angiotensin II. This gene may play a role in the generation of reperfusion arrhythmias following restoration of blood flow to ischemic or infarcted myocardium. It was previously thought that a related gene, denoted as AGTR1B, existed; however, it is now believed that there is only one type 1 receptor gene in humans. Alternative splicing of this gene results in multiple transcript variants.*	3q24

Protein	*Function*	*Location*
AT2 or ATGR2[11]	*The protein encoded by this gene belongs to the G-protein coupled receptor 1 family, and* ***functions as a receptor for angiotensin II****. It is an intergral membrane protein that is highly expressed in fetus and in neonates, but scantily in adult tissues, except brain, adrenal medulla, and atretic ovary. This receptor has been shown to mediate programmed cell death and this apoptotic function may play an important role in developmental biology and pathophysiology. Mutations in this gene are been associated with X-linked cognitive disability.* ***Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) and SARS-CoV-2 infection results in down-regulation of angiotensin converting enzyme-2 (ACE2) receptors, the effects of which, triggers serious inflammatory lesions in the tissues involved, primarily in the lungs. The inflammatory reaction appears to be mediated by angiotensin II derivatives, including the angiotensin AT2 receptor which has been found to be upregulated in bronchoalveolar lavage samples from Coronavirus disease 2019 (COVID19) patients***	Xq23

We can further summarize these as follows[12]*:*

Coronavirus replication is initiated by the binding of S protein to the cell surface receptor(s). The S protein is composed of two functional domains, S1 (bulb) which mediates receptor binding and S2 (stalk) which mediates membrane fusion. Specific interaction between S1 and the cognate receptor triggers a drastic conformational change in S2, leading to fusion between the virus envelope and the cellular membrane and release of the viral nucleocapsid into the host cell cytosol. Receptor binding is the major determinant of the host range and tissue tropism for a coronavirus.

Some human coronaviruses (HCoVs) have adopted cell surface enzymes as receptors, angiotensin converting enzyme 2 (ACE2) for SARS-CoV-1 and HCoV NL63. The receptor-bound S protein is activated by cleavage into S1 and S2, mediated by one of two of two host

[6] https://pubchem.ncbi.nlm.nih.gov/compound/172198

[7] https://pubchem.ncbi.nlm.nih.gov/compound/71745056

[8] https://pubchem.ncbi.nlm.nih.gov/#query=ang%201-7

[9] https://www.ncbi.nlm.nih.gov/gene/4142

[10] https://www.ncbi.nlm.nih.gov/gene/185

[11] https://www.ncbi.nlm.nih.gov/gene/186

[12] https://pubchem.ncbi.nlm.nih.gov/pathway/Reactome:R-HSA-9678110

proteases, the endosomal cysteine protease cathepsin L and another trypsin like serine protease. *Type II transmembrane serine proteases TMPRSS2 and TMPRSS11D have also been implicated in the activation of S protein of SARS-CoV-1. Host factors may play additional roles in viral entry (not annotated here).*

Valosin containing protein (VCP) contributes by a poorly understood mechanism to the release of coronavirus from early endosomes. Host factors may also restrict the attachment and entry of HCoV. Some interferon inducible transmembrane proteins (IFITMs) exhibited broad spectrum antiviral functions against various RNA viruses including SARS-CoV-1 while others may facilitate HCoV entry into host cells.

Now Etelvino et al have noted:

The renin-angiotensin system is an important component of the central and humoral mechanisms of blood pressure and hydro-electrolytic balance control. Angiotensin II is a key player of this system.

Twenty-five years ago the first manuscripts describing the formation and actions of another peptide of the RAS, angiotensin-(1-7), were published. Since then several publications have shown that angiotensin-(1-7) is as pleiotropic as angiotensin II, influencing the functions of many organs and systems. The identification of the ACE homologue ACE2 and, a few years later, Mas, as a receptor for angiotensin-(1-7) contributed a great deal to establish this peptide as a key player of the RAS. Most of the actions of angiotensin-(1-7) are opposite to those described for angiotensin II.

This has led to the concept of two arms of the RAS:

one comprising ACE/AngII/AT1R and

the other ACE2/Ang-(1-7)/Mas.

More recently, we have described the identification of a novel component of the RAS, alamandine, which binds to the ***Mas-related G protein coupled receptor D****. This peptide is formed by decarboxylation of the Asp residue of angiotensin-(1-7), leading to the formation of Ala as the Nterminal amino acid. Alternatively, it can be formed by hydrolysis of Ang A, by ACE2. Its effects include vasorelaxation, central effects similar to those produced by angiotensin-(1-7), blunting of isoproterenol-induced heart fibrosis, and anti-hypertensive action in SHR.*

The putative enzyme responsible for alamandine formation from angiotensin-(1-7) is under investigation. The identification of this novel component of the RAS opens new venues for understanding its physiological role and opens new putative therapeutic possibilities for treating cardiovascular diseases.

As Bader et al have noted:

*The **Mas-related G protein**–coupled receptors (Mrgprs or Mas-related genes) comprise a subfamily of receptors named after the first discovered member, Mas.*

For most Mrgprs, pruriception seems to be the major function based on the following observations:

1) they are relatively promiscuous in their ligand specificity with best affinities for itch-inducing substances;

2) they are expressed in sensory neurons and mast cells in the skin, the main cellular components of pruriception; and

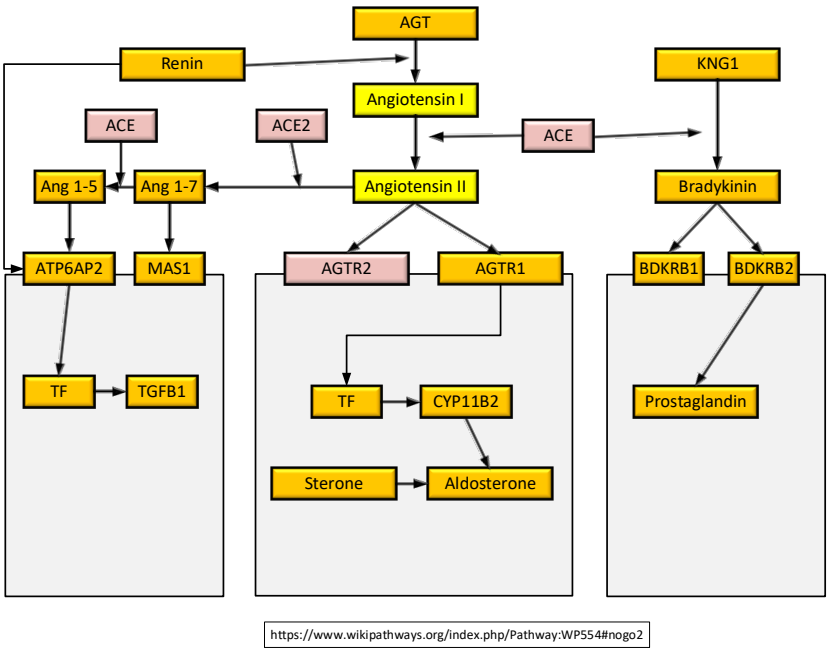
3) they appear in evolution first in tetrapods, which have arms and legs necessary for scratching to remove parasites or other noxious substances from the skin before they create harm.

Because parasites coevolved with hosts, each species faced different parasitic challenges, which may explain another striking observation, the multiple independent duplication and expansion events of Mrgpr genes in different species as a consequence of parallel adaptive evolution. Their predominant expression in dorsal root ganglia anticipates additional functions of Mrgprs in nociception.

Some Mrgprs have endogenous ligands, such as b-alanine, alamandine, adenine, RF-amide peptides, or salusin-b. However, because the functions of these agonists are still elusive, the physiologic role of the respective Mrgprs needs to be clarified.

The best studied Mrgpr is Mas itself. It was shown to be a receptor for angiotensin-1–7 and to exert mainly protective actions in cardiovascular and metabolic diseases. This review summarizes the current knowledge about Mrgprs, their evolution, their ligands, their possible physiologic functions, and their therapeutic potential.

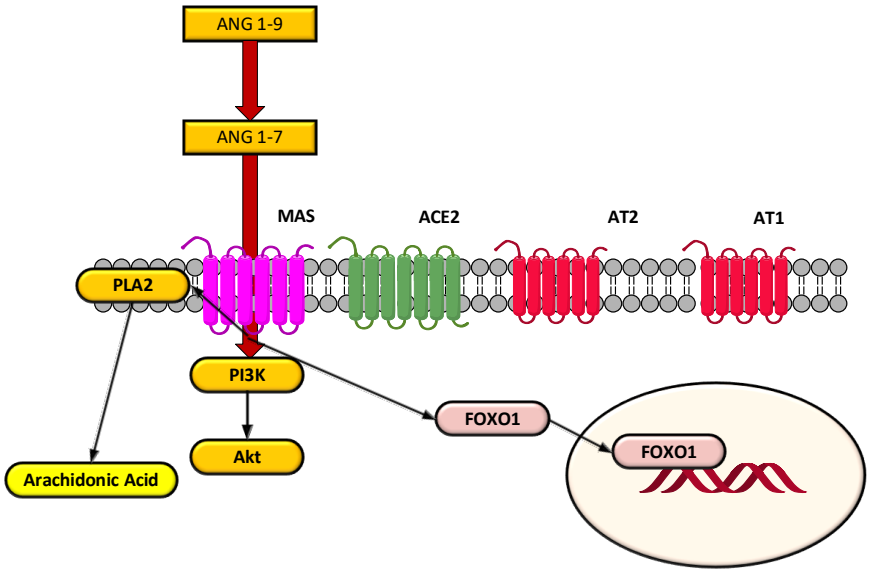
We detail some of these functions in the diagram below.



Now we further know from Bader et al that the MAS activated pathway is as shown below[13]:

[13] From Bader et al we note: *The most important known signaling pathways of Mas involve phospholipase A (PLA) to generate arachidonic acid (AA) and phosphoinositide 3 kinase (PI3K) and AKT to activate eNOS by phosphorylation at serine 1177 and dephosphorylation at threonine 495. Ang-(1–7), AVE0991, CGEN-856S, and angioprotectin have been shown to activate these pathways and A-779 antagonizes these actions. Ang-(1–7) also activates forkhead box protein O1 by dephosphorylation at serine 236. The ligands for the PLA and PI3K pathways (other than CGEN-856S), however, do not activate a phospholipase C (PLC)/Ca2+ signaling pathway observed in Mas-overexpressing cells, which in some studies were additionally transfected with promiscuous G proteins. This pathway is activated by the ligands neuropeptide FF (NPFF), AR234960, MPF7, and CGEN-856S and blocked by AR244555. Thus, different Mas ligands activate distinct signaling pathways, a phenomenon called biased agonism.*

see Bader et al, Mas and Its Related G Protein–Coupled Receptors, Mrgprs, Pharmacol Rev 66:1080–1105, October 2014



3.3 ACE2 VARIANTS

ACE2 has many variants. As Sorokina et al have noted:

To this end, we collected information regarding naturally-occurring variants of the Angiotensin-converting enzyme 2 (ACE2), an epithelial receptor that both SARS-CoV and SARS-CoV-2 use to enter the host cells. We built 242 structural models of variants of human ACE2 bound to the receptor binding domain (RBD) of the SARS-CoV-2 surface spike glycoprotein (S protein) and refined their interfaces with HADDOCK.

Our dataset includes 140 variants of human ACE2 representing missense mutations found in genome-wide studies, 39 mutants with reported effects on the recognition of the RBD, and 63 predictions after computational alanine scanning mutagenesis of ACE2-RBD interface residues. This dataset will help accelerate the design of therapeutics against SARS-CoV-2, as well as contribute to prevention of possible future coronaviruses outbreaks.

There are a multiplicity of ACE2 variants. The following Table is from that source. It depicts some of the more common variants. The details are in the reference[14].

[14] see Sorokina et al, *Mutations reported in the literature mappable on the cryo-EM structure. Table includes energetic calculations with HADDOCK. vdW – van der Waals interaction score; Elec – electrostatic interaction score; Desol – desolvation score; BSA – buried surface area; a.u. – arbitrary units of energy; Effect – The measured biochemical effect on the interaction with the SARS-CoV spike glycoprotein. A-No effect on interaction with SARS-CoV spike glycoprotein; B-Slightly inhibits interaction with SARS-CoV spike glycoprotein; C-Strongly inhibits interaction with SARS-CoV spike glycoprotein.; D-Abolishes interaction with SARS-CoV spike glycoprotein; E-Inhibits interaction with SARS-CoV spike glycoprotein.; F-About 95% loss of angiotensin I* cleavage; G-Complete *loss of enzyme activity; H- more than 50% loss of angiotensin I cleavage. *also included in the Alanine scanning mutagenesis data*

#N	Variation	Haddock Score[15]	vdW (a.u.)	Elec (a.u.)	Desol (a.u.)	BSA (Å2)	Effect
0	Wild Type	−89.9 ± 5.2	−54.7 ± 2.4	−184.5 ± 9.3	1.7 ± 4	1729 ± 34.6	WT
1	QAK24-26KAE	−95.2 ± 4.1	−52 ± 2.2	−196 ± 33	−4 ± 5.1	1709.5 ± 68.2	B
2	K31D	−85.7 ± 6.5	−56.4 ± 2.5	−176.9 ± 11.6	6 ± 6.4	1692.6 ± 10.3	D
3	E37A	−91.2 ± 3.7	−52.2 ± 1.1	−176.9 ± 30.6	−3.6 ± 7	1660.4 ± 22.9	A
4	D38A	−86 ± 8.3	−55.5 ± 2.2	−174.8 ± 33.8	4.5 ± 4.2	1685.6 ± 53.2	A
5	Y41A	−92.1 ± 4.3	−52.7 ± 1.7	−171.7 ± 17	−5 ± 6.9	1687.3 ± 26.7	C
6	K68D	−85.9 ± 5.9	−55.5 ± 1.8	−188.8 ± 21	7.4 ± 4.6	1739.4 ± 52	B
7	MYP82-84NFS	−99 ± 6.4	−53.9 ± 2.6	−176.8 ± 11.8	−9.7 ± 7.4	1750 ± 36.3	E
8	E110P	−90.2 ± 3.2	−56.2 ± 2.3	−208.3 ± 12.1	7.6 ± 6.3	1739.8 ± 32.7	A
9	PD135-136SM	−102 ± 3.7	−56.6 ± 2.1	−210.3 ± 19.8	−3.4 ± 8.8	1774 ± 53.6	A
10	E160R	−90.5 ± 2	−55.5 ± 2.5	−182 ± 12.4	1.4 ± 5.7	1721 ± 30.4	A
11	R169Q	−98 ± 4.2	−53.5 ± 2.9	−202 ± 22.8	−4.1 ± 2.5	1733.7 ± 65.7	F
12	R192D	−111.8 ± 5	−58.2 ± 2.7	−198.3 ± 11.3	−13.9 ± 6.4	1764.1 ± 12.6	A
13	R219D	−94.4 ± 4.9	−54.6 ± 4.6	−206 ± 11.9	1.3 ± 5.5	1756.8 ± 25.6	A
14	H239Q	−97.1 ± 8.2	−56 ± 1	−194.1 ± 16.6	−2.3 ± 9.4	1697.8 ± 39.8	A
15	W271Q	−95.5 ± 3.2	−54.8 ± 1.4	−208.4 ± 8.7	1 ± 4.5	1736.5 ± 15	F
16	R273Q	−104.1 ± 6.3	−55.7 ± 2.1	−215.2 ± 17.4	−5.4 ± 6.1	1767.5 ± 38.9	G
17	K309D	−91.1 ± 7.8	−57.8 ± 2.2	−180.4 ± 17.5	2.9 ± 9.2	1759.9 ± 32	A
18	E312A	−90.9 ± 2.5	−57 ± 1.9	−172.6 ± 15.8	0.7 ± 3.2	1705.4 ± 23	A
19	T324A	−99.7 ± 4.7	−54.9 ± 0.8	−189.8 ± 26.4	−6.9 ± 3.9	1726.7 ± 7 ± 1	A
20	NVQ338-340DDR	−84 ± 3.7	−54 ± 3.9	−193.2 ± 10.9	8.6 ± 4.5	1730.3 ± 28.2	A
21	H345A	−92.3 ± 4.8	−55 ± 0.8	−188.4 ± 4.7	0.4 ± 4.9	1674.9 ± 26.7	G
22	D350A	−95.6 ± 7.5	−54.7 ± 3.8	−184 ± 20.2	−4.1 ± 1.2	1691.8 ± 47.5	A
23	K353H	−104.7 ± 4.6	−57.2 ± 1.5	−174.6 ± 5.1	−12.7 ± 5.1	1749.2 ± 38.3	D
24	K353A	−93 ± 1.9	−50.7 ± 3	−160 ± 34.1	−10.3 ± 4.7	1668.1 ± 28.4	D
25	K353D	−101.9 ± 8	−53.8 ± 3.3	−201 ± 8.3	−7.9 ± 6.6	1739 ± 30.3	D
26	D355A	−96.4 ± 2	−53.9 ± 1.9	−207.4 ± 8	−1 ± 2.8	1719 ± 16.3	C
27	R357A	−101.9 ± 5.2	−58.3 ± 0.9	−199.5 ± 11.7	−3.7 ± 6.1	1759.3 ± 41.2	C
28	L359K	−98.8 ± 3.9	−53.9 ± 4.2	−174 ± 27.4	−10 ± 2.1	1728.7 ± 32.5	A
29	L359A	−90.9 ± 5.6	−54 ± 1.2	−205.2 ± 17.2	4.2 ± 7.7	1733.2 ± 38.9	A
30	M383A	−105.6 ± 4.7	−59.6 ± 3.1	−197.5 ± 13.7	−6.5 ± 7.5	1788.8 ± 25.8	B
31	P389A	−104.3 ± 9.9	−55.6 ± 5.2	−198 ± 5.9	−9.1 ± 9.3	1755.2 ± 61	B
32	R393A	−106.3 ± 1	−54.5 ± 3.4	−201.1 ± 10.8	−11.5 ± 3	1715.3 ± 43.5	B
33	SPD425-427PSN	−96.7 ± 3.9	−55.7 ± 0.2	−171.9 ± 14.3	−6.7 ± 2.7	1727.6 ± 53	B
34	KGE465-467QDK	−81.8 ± 3.1	−53.1 ± 3.7	−184.4 ± 23.9	8.2 ± 3	1753.7 ± 38.8	A
35	K481Q	−97.1 ± 3.4	−56.8 ± 1.6	−198.8 ± 9.3	−0.6 ± 1.6	1726 ± 42.2	H
36	H505A	−99.7 ± 7.8	−55.4 ± 3	−193.6 ± 12.2	−5.6 ± 7.1	1724.3 ± 26.3	G
37	R514Q	−87.6 ± 2.5	−54.2 ± 2.5	−193.1 ± 16.5	5.2 ± 3.8	1739.2 ± 36.7	H
38	R559S	−91.4 ± 7.5	−54.6 ± 1.4	−203.1 ± 14.5	3.7 ± 7.7	1744.3 ± 12.4	B
39	F603T	−92.6 ± 3.3	−58 ± 1.7	−165.9 ± 16.3	−1.4 ± 3.3	1739.6 ± 24.7	A

One can see that many are single nucleic acid changes. This may become a challenge if one must fit the mutated spikes to the variant ACE2. This is also a challenge since we generally cannot ascertain the ACE2 configuration when testing for spike presence of infection.

[15] See Vangone et al. The Haddock Score is a tool to assess binding of proteins to ligands.

We shall examine these in detail in the next section. ACE2 has substantial variability and thus should be considered as an integral elements in addressing the virus inhibition.

4 VARIANTS

We have been seeing an accumulation of COVID variants. We have examined these to some degree as they have evolved[16]. In this section we examine the current literature regarding both ACE2 and COVID spike variants. The assumption is that the spike targets ACE2. One must be cautious with this assertion since as we see mutations of spike we may very well see changes to alternative entry points rather than just ACE2. The current COVID virus has great evolutionary aggressiveness and thus we should be concerned by tamping it down as aggressively as we can.

4.1 ACE2 VARIANTS

ACE2 has been seen as a critical control element to counter ACE and its tendency to cause hypertension. Ashoor et al have detailed multiple ACE2 variants, they note:

SARS-CoV-2 infectivity is largely determined by the virus Spike protein binding to the ACE2 receptor. Meanwhile, marked infection rate differences were reported between populations and individuals. To understand the disease dynamic, we developed a computational approach to study the implications of both SARS-CoV-2 RBD mutations and ACE2 polymorphism on the stability of the virus-receptor complex. We used the 6LZG PDB RBD/ACE2 3D model, the mCSM platform, the LigPlot+ and PyMol software to analyze the data on SARS-CoV-2 mutations and ACE variants retrieved from GISAID and Ensembl/GnomAD repository.

We observed that out of 351 RBD point mutations, % destabilizes the complex according to free energy (ΔΔG) differences. *We also spotted variations in the patterns of polar and hydrophobic interactions between the mutations occurring in out of contact residues. Similarly, comparison of the effect on the complex stability of different ACE2 variants showed that the pattern of molecular interactions and the complex stability varies also according to ACE2 polymorphism. We infer that it is important to consider both ACE2 variants and circulating SARS-CoV-2 RBD mutations to assess the stability of the virus- receptor association and evaluate infectivity. This approach might offers a good molecular ground to mitigate the virus spreading …*

The interaction ACE2/SARS-CoV-2 occurs through the receptor-binding domain (RBD) sequence of the S1 chain of the spike protein. This functional domain is highly prone to mutations.

Meanwhile, the human ACE2 receptor gene that is located on chromosome X is highly polymorphic with hundreds of genetic variants identified to date in various populations . Noteworthy are the ACE2 variants reported in the Chinese populations though few variants are highly frequent and the majority of them are very rare .

Interestingly some ACE2 variants was shown to be significantly associated with the onset of diseases such arterial hypertension, diabetes mellitus, cerebral stroke, coronary artery disease,

[16] https://www.researchgate.net/publication/348248952_COVID-19_Mutations_and_Infectivity

heart septal wall thickness and ventricular hypertrophy all considered as comorbidity factors of COVID-19. The complexity of the SARS-CoV-2 S protein association with the ACE2 receptor is highlighted in a number of X-ray crystallography models.

ACE2 variants appear to have the capability to create co-morbidities. Thus, patients with variant ACE2 receptors may have a double hit; co-morbidity as well as aggressive COVID attack. In effect it may not be that the co-morbidity is the cause but a secondary factor resulting from a variant ACE2. They continue:

These are namely the 6LZG (SARS-CoV-2 Spike RBD /ACE2) complex, the 2AJF (SARS-CoV Spike RBD /ACE2) complex, the 6M0J (SARS-CoV-2 Spike RBD /ACE2) complex, and the 6VW1 (chimeric SARS-CoV/SARS-CoV-2 Spike RBD /ACE2) complex . These models have different resolution and are available at RCBS Protein Data Bank (https://www.rcsb.org/). Studies using these models have shown that the affinity of the RBD SARS- CoV-2 to the human ACE2 varies between ACE2 genetic variants suggesting that some people might be genetically protected where others are susceptible to COVID-19.

Indeed, by measuring the ACE2 affinity for SARS-CoV-2 Spike using docking simulations,.. showed that ACE2 variant SP, which is more frequent in Africans and KR more frequent in Europeans, are respectively protective and predisposing genetic factors to COVID-19.

Again the science demonstrates why we see increased morbidity and mortality. One must be cautious not to conflate social and political issue with the underlying science. We shall see this effect again and again.

Meanwhile, other studies denied the association between ACE2 genetic polymorphism and the susceptibility to SARS infection and to COVID- . Nevertheless, the virus-host-cell interaction involves specific contacts between two structurally defined molecular entities. Therefore, it is very likely that the interaction between the SARS-CoV-2 and the human cell surface receptor ACE2 and thus the virus infectivity is determined by the genetic variations of both the pathogen and the receptor interacting sequences. This can reflect a higher magnitude of complexity in the interplay between this new viral pathogen and its human host.

Now as Devaux et al note:

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has emerged in Chinese people in December 2019 and has currently spread worldwide causing the COVID-19 pandemic with more than 150,000 deaths.

In order for a SARS-CoV like virus circulating in wild life for a very long time to infect the index case-patient, a number of conditions must be met, foremost among which is the encounter with humans and the presence in homo sapiens of a cellular receptor allowing the virus to bind.

Recently it was shown that the SARS-CoV-2 spike protein, binds to the human angiotensin I converting enzyme 2 (ACE2).

This molecule is a peptidase expressed at the surface of lung epithelial cells and other tissues, that regulates the renin-angiotensin-aldosterone system. Humans are not equal with respect to the expression levels of the cellular ACE2.

Moreover, ACE2 polymorphisms were recently described in human populations. Here we review the most recent evidence that ACE2 expression and/or polymorphism could influence both the susceptibility of people to SARS-CoV-2 infection and the outcome of the COVID-19 disease. Further exploration of the relationship between the virus, the peptidase function of ACE2 and the levels of angiotensin II in SARS-CoV-2 infected patients should help to better understand the pathophysiology of the disease and the multi-organ failures observed in severe COVID-19 cases, particularly heart failure….

ACE2 limits the adverse vasoconstrictor and profibrotic effects of AngII. The hydrolysis of AngII into Ang (1e7) reduces the oxidative stress of AngII on endothelial cerebral arteries. Ang (1e7) was reported to have vasodilatory and antifibrotic actions. Disruption of ACE2 results in increased AngII levels and impaired cardiac function. Reduced levels of cardiac ACE2 have been reported in hypertension (HT) and diabetic heart disease.

Low expression of ACE2 mRNA was associated to HT, dyslipidemia and/or heart failure.

A polymorphism of ACE2 gene was first documented in the Chinese population with three ACE2 variants (rs4240157, rs4646155, and rs4830542) associated with HT,70e74 in a Nicotine Dependence in Teens Canadian cohort rs2074192, rs233575, and rs2158083 mutations were significantly associated with pathological variations of blood pressure.

ACE2 rs21068809 mutation (C > T) has been reported associated with clinical manifestations of HT.76 In Indian the study of 246 HT patients and 274 normotensive people indicated an association of HT with ACE2 rs21068809 mutation.77

In Brazilian patients, the combination of ACE I/D and ACE2 G8790A polymorphisms revels susceptibility to HT.78 The RAAS pathway can also be regulated by a polymorphism in ACE.

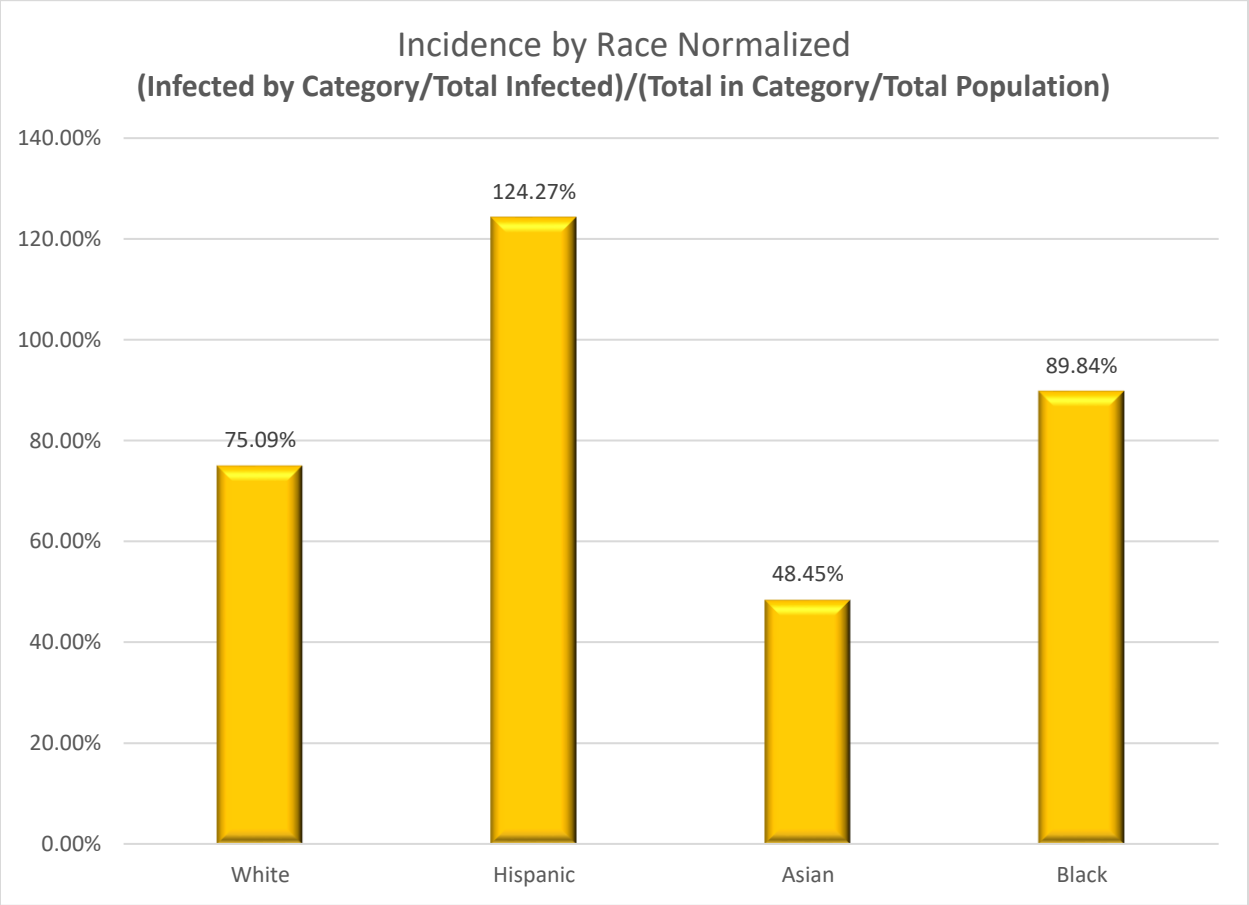
In African-American with hypertension an ACE polymorphism was reported.

Here again we see the genetic difference that result in ACE2 changes are the drivers of ethnic variations in morbidity and mortality not necessarily some social construct.

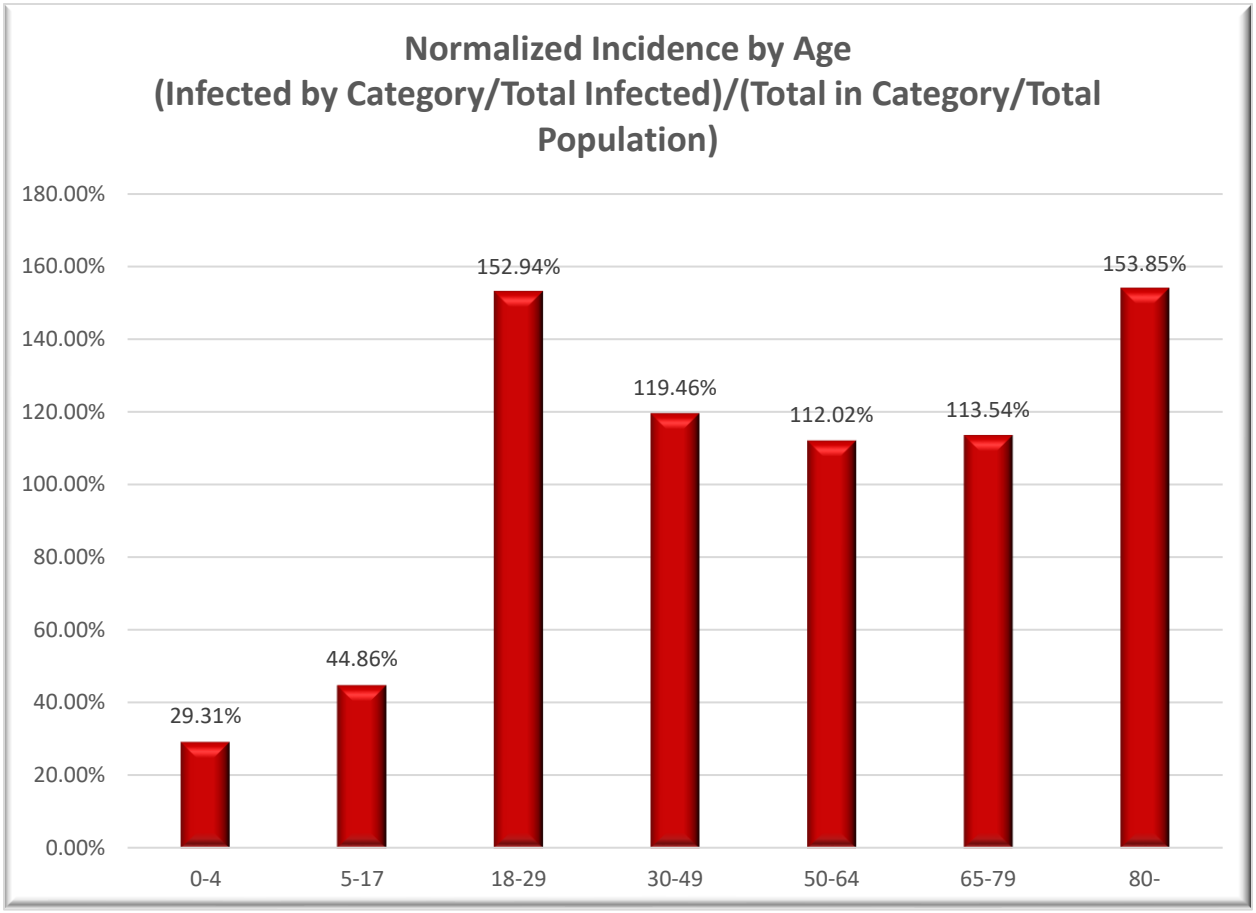
Very recently, Cao and colleagues reported the results of a large investigation (1700 variants) of coding sequences variants in ACE2 and the allele frequency differences between populations in ACE2 gene from the China Metabolic Analytics Project and 1000 Genome Project database and other large scale genome databases. They found one variant with a truncation Gln300 in China. In addition, they reported 32 variants among which seven hotspot variants in different populations.

We demonstrate some of the ethnic variations below. It is now well known that the ACE2 receptor variants are a dominant cause of ethnic variations in infectivity. We present so current data for comparison.

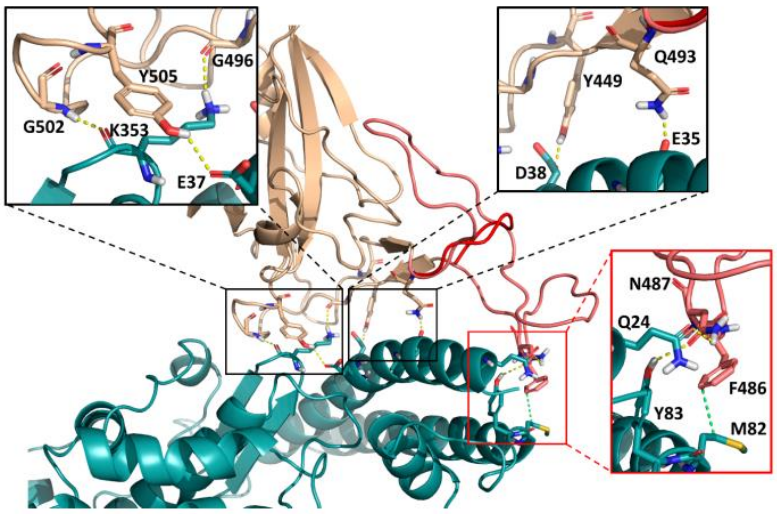
Below we show the normalized by race. Note the high incidence amongst Hispanics. This can be a result of both ACE2 receptor variants as well as socioeconomic conditions. African Americans have the most vulnerable ACE2 receptors but their normalized distribution is substantially lower. In contrast Asians have a protective ACE2 receptor and this is exhibited below.



Now by age we see where the problem is. It is the 18-29 year olds who are the spreaders. The 80+ are dominated by LTC cases most likely via the 18-29 class. Also the 5-17 year old school group has a low relative count.



Now Kemenesi et al have examined the variants of ACE with the effective binding domain as shown below:



The authors then note:

Spike mutations are of special interest due to their possible role in the emergence of variants with modified antigenicity. This altered antigenicity may bypass vaccine effectiveness, monoclonal therapeutic options, and several others.

A recent study revealed a 0.75% prevalence for Spike deleterious variants by analyzing 146,795 genome sequences. These deletions were mostly positioned within the N-terminal domain. There are several identified Spike mutants rapidly spreading, such as the D614G which is now considered as a prevalent SARS-CoV-2 variant all around the world. Sequential rounds of evolution in the context of an outbreak situation with an emerging virus was previously reported in case of SARS-CoV and during the West African Ebola outbreak. It is considered as an adaptation to the new human host and may lead to significant increase in the prevalence of certain variants. Dominant mutations may drive the main scenario of an outbreak, but recessive mutations are also present, although the identification of the latter is highly challenging.

The general understanding of SARS-CoV-2 evolution during the current pandemic situation may reveal future scenarios and can facilitate therapeutic research directions and strategies. We identified a major deletion in the RBD of the Spike protein and verified an altered receptor binding capacity via in silico methods. The variant presented in this study lacks a major part of the RBD along with several important AA positions for ACE2 binding.

We therefore hypothesized a weaker receptor binding capacity. The occurrence of such a mutation in a natural infection and the recessive nature of this deleterious variant revealed a scenario for evolutionary adaptation of the virus within a single host.

The identification of this strain may be of high interest in future studies, involving attenuated strains and thereby it may facilitate therapeutic advances. A limitation of our study is the lack of infectious isolate, since the virus isolation efforts failed to retrive infectious isolate in vitro. It is possibly due to the condition of the sample after multiple freeze-thaw cycles or even due to the lower specificity to VeroE6 cell line of this particular deleterious strain. Reverse genetics may be used in future studies for in vitro experiments with this deleterious variant to verify in silico data and to examine in vitro characteristics.

The most important take-home message of our study is that we need more studies to understand main genomic-evolutionary mechanisms of circulating SARS-CoV-2 viruses regarding their most powerful evolutionary tool, the recombination.

Sequencing efforts therefore need to be focused on the surveillance of emerging recombinant variants, not only tracking nucleotide changes in the genome.

Here, we present a naturally occurring recessive RBD deleterious SARS-CoV-2 variant. Strains with altered receptor-binding capacity may be of special interest in future studies in relation to vaccine development, therapeutic options or simply for the general understanding of evolutionary mechanisms regarding the COVID-19 pandemic.

Let us note regarding recombinant viral variants[17]:

Recombination involves the exchange of genetic material between two related viruses during coinfection of a host cell. Viral recombination occurs when viruses of two different parent strains coinfect the same host cell and interact during replication to generate virus progeny that have some genes from both parents. Recombination generally occurs between members of the same virus type (e.g., between two influenza viruses or between two herpes simplex viruses). Two mechanisms of recombination have been observed for viruses: independent assortment and incomplete linkage. Either mechanism can produce new viral serotypes or viruses with altered virulence.

Recombination by Independent Assortment: Recombination by independent assortment can occur among viruses with segmented genomes. Genes that reside on different pieces of nucleic acid are randomly assorted. This can result in the generation of viruses with new antigenic determinants and new host ranges. Development of viruses with new antigenic determinants through independent assortment is called antigenic shift.

Recombination of Incompletely Linked Genes: Genes that reside on the same piece of nucleic acid may undergo recombination. The closer two genes are together, the rarer is recombination between them (partial linkage).

Phenotypic Variation from Recombination: Development of viruses with new antigenic determinants by either type of recombination may allow viruses to infect and cause disease in previously immune hosts.

Thus variants can come about by a variety of means. Insertions, deletions, and very powerfully via recombination.

4.2 COVID VARIANTS

COVID-19 variants, s, are being understood and identified in the early stages. The fundamental issue is to stop the viral propagation before it mutates to an ever more aggressive form. The mutations follow the same means as any other genetic structure but since it is a single stranded mRNA it is susceptible to the most aggressive forms. From the CDC and the recent NYC B.1.526 variant from West et al we have[18]:

[17] https://www.ncbi.nlm.nih.gov/books/NBK8439/

[18] https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html It should be noted that CDC data is chronically late often incomplete. There are now evolving multiple independent registries for variants which may become more current and more reliable. We have found in our experience and in our opinion that the CDC across the board has been grossly deficient in reporting correctly, timely and in a useable manner.

Name	*Name Nextstrain*	*First Detected*	*Cases in US*	*Countries Reporting*	*Key Mutations*	*Transmissibility Rate*
B.1.1.7	20I/501Y.V1	United Kingdom	Y	70	69/70 deletion, 144Y deletion, N501Y, A570D, D641G, P681H	~50% increase
P.1	20H/501Y.V3	Japan Brazil	Y	>4	E484K K417N/T N501Y D614G	Not determined
B.1.351	20H/501.V2	South Africa	Y	>30	K417N, E484K, N501Y, D614G	Not Determined
B.1.526		New York	Y	>4	L5F, T95I, D253G, E484K or S477N, D614G, and A701V.	Not Determined

B.1.1.7: In the United Kingdom (UK), a variant of SARS-CoV-2 known as B.1.1.7 emerged. This variant carries a large number of mutations and has since been detected around the world, including in the United States (US). This variant was first detected in the US at the end of December 2020. In January 2021, scientists from the UK reported early evidence that suggests the B.1.1.7 variant may be associated with an increased risk of death compared with other variants. More studies are needed to confirm this finding.

B.1.351: In South Africa, another variant of SARS-CoV-2 known as B.1.351 emerged independently of B.1.1.7. According to a non-peer-reviewed preprint article, this variant shares some mutations with B.1.1.7. Cases attributed to B.1.351 have been detected outside of South Africa, and this variant was first detected in the US at the end of January 2021. Preliminary evidence from non-peer-reviewed publications suggests that the Moderna mRNA-1273 vaccine currently used in the US may be less effective against this variant, but additional studies are needed.

P.1: In Brazil, a variant of SARS-CoV-2 known as P.1 emerged; it was first identified in January 2021 in travelers from Brazil who arrived in Japan. This variant was detected in the US at the end of January 2021. The P.1 variant has 17 unique mutations, including three in the receptor binding domain of the spike protein (K417T, E484K, and N501Y), according to non-peer-reviewed preprint articles. There is evidence to suggest that some of the mutations in the

P.1 variant may affect the ability of antibodies (from natural infection or vaccination) to recognize and neutralize the virus, but additional studies are needed.

From West et al we now have what they are calling the New York Variant:

Wide-scale SARS-CoV-2 genome sequencing is critical to monitoring and understanding viral evolution during the ongoing pandemic. Variants first detected in the United Kingdom, South Africa, and Brazil have spread to multiple countries. We have developed a software tool, Variant Database (VDB), for quickly examining the changing landscape of spike mutations. Using this tool, we detected an emerging lineage of viral isolates in the New York region that shares mutations with previously reported variants. The most common sets of spike mutations in this lineage (now designated as B.1.526) are

L5F, T95I, D253G, E484K or S477N, D614G, and A701V.

This lineage appeared in late November 2020, and isolates from this lineage account for ~25% of coronavirus genomes sequenced and deposited from New York during February 2021.

They also present the following Table of other isolates:

Pattern	*Number of isolates*	*Top Locations*	*First collection date*
L5F T95I D253G E484K D614G A701V	**243**	**US(240;NY 235)**	**12/16/2020**
E484KD614G V1176F	235	Brazil(132), US(40)	4/15/2020
W152L E484K D614G G769V	49	US(32)	11/1/2020
E484K D614G P681H	37	US(37;MD27)	11/18/2020
R102I F157LV367F E484K Q613H P681R	36	England(35)	12/27/2020
Q52RA67V H69-V7O-Y144- E484K D614G Q677H F8881	36	England(22)	12/15/2020

An additional view of variants is shown below[19]:

[19] https://en.wikipedia.org/wiki/Variants_of_SARS-CoV-2#CITEREFAlm_et_al.

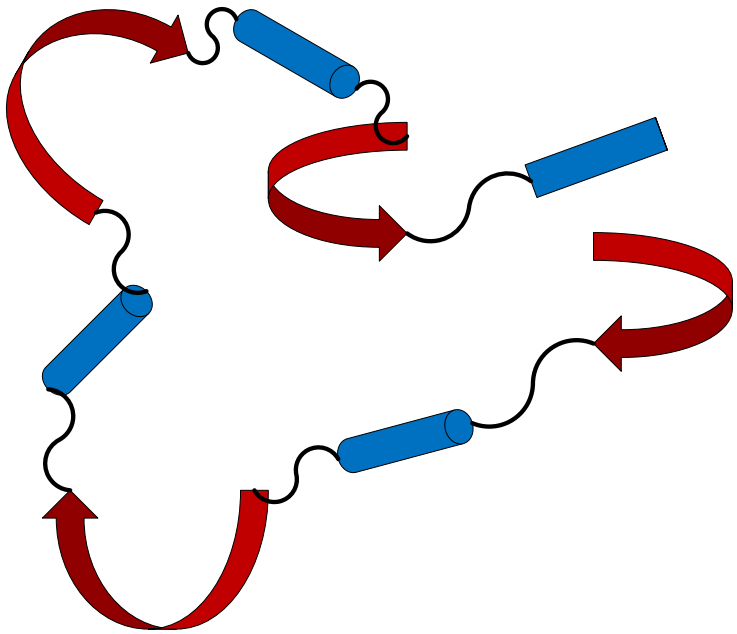
Pango lineages (by Rambaut et al.)	Notes to pango lineages (see Alm et al.)	Nextstrain clades, 2021	GISAID clades	Notable variants
A.1–A.6		19B	S	contains "reference sequence" WIV04/2019[1]
B.3–B.7, B.9, B.10, B.13–B.16		19A	L	
			O[a]	
B.2			V	
B.1	B.1.5–B.1.72	20A	G	Lineage B.1 in the Rambaut et al. system
	B.1.9, B.1.13, B.1.22, B.1.26, B.1.37		GH	
	B.1.3–B.1.66	20C		Includes CAL.20C[
		20G		Predominant in US generally, Jan '21[
		20H		Includes B.1.351 aka 20H/501Y.V2 or 501.V2 lineage
	B.1.1	20B	GR	Includes B.1.1.207
		20D		Includes P.1 and P.2[
		20F		
		20I		Includes lineage B.1.1.7 aka VOC-202012/01 or 20I/501Y.V1
	B.1.177	20E	GV	Derived from 20A

5 PROTEIN MODELLING

Protein modelling is a highly complex and evolving process. Knowing the protein sequence is generally the easy part. Finding its three-dimensional structure is complex. One can recall that the one dimensional structure is merely a sequence of amino acids. The two dimensional structure is the effect of hydrogen bonds between nucleic acids allowing them to form helices or sheets, namely long strands of matched pairs via hydrogen bonding (O-H bonds) and then the three dimensional structure amongst helices, sheets, and plain runs of nucleic acids. Now add to this a second protein with similar structure and we have the chance for protein-protein bonding.

5.1 SECONDARY STRUCTURES

Secondary structures are often key to binding. They are the helices and ribbons that establish the ultimate backbone for the subsequent three-dimensional folding. Graphically the two dimensional flat-model including strings can be exemplified as below:



Predicting this structure is the next step after having the one dimensional structure. Now as Lasfar and Bouden have noted:

The prediction of the secondary structure of proteins is one of the most studied problems in computational biology.

However, the accuracy of the predicted secondary structure is insufficient for practical utility.

Indeed, this is the necessary first step. The issue is that the problem is generated by having just a nucleic acid sequence and must then determine where the helices and ribbons are and how large they may be. But in anticipation of our analysis, we are not starting at that point for we indeed have both the two and three dimensional structure of the wild type (Wuhan virus) as well as various mutant variants. They continue:

In this paper, we propose an algorithmic approach based on Hidden Markov Models (HMM) to model the problem of prediction. Therefore, HMM are often used for data mining in bioinformatics. In this research, we have built a HMM that models the prediction problem of protein secondary structure. Moreover, two procedures for estimating the probability parameters were performed by the Maximum Likelihood Estimation (MLE) of protein sequences from a public database (Brookhaven PDB).

Finally, a new prediction approach based on a posteriori probability of hidden regimes has been implemented. Our model appears to be very efficient on single sequences, with a score of 66.6% by comparing the first results obtained with the real secondary sequence and encouraging for an improvement of the system …

The initial idea was to propose a new system for predicting the secondary structure of proteins based on the model of hidden Markov chains. The first HMM implantation applied to the prediction of secondary protein structure was carried out in 1993 by K. Asai et al. However, the precision of the unsatisfactory prediction of this approach gave rise to other work based on different methods. So it seemed reasonable to think that their decoding path (the viterbi algorithm) is inappropriate for this kind of problem. A new approach was then needed to better predict the sequence of secondary structures: the hidden state of posteriori probability decoding approach was proposed in this direction.

Moreover, work on the construction of a well-parameterized HMM model was carried out in parallel: the development of the two probability distribution estimation procedures were carried out by the maximum likelihood estimators. In addition, the overfitting problem was solved by the Laplace rule in order to regularize the resulting probabilities. Finally, this new approach of prediction makes it possible to obtain very encouraging results with a test sequence of small size. However, even if the approach seems interesting, the system still needs to be developed to deal with large sequence sets. The research perspectives consist of verifying, confirming the proposed HMM model on other protein families and using other hybrid data mining techniques to obtain higher prediction accuracy.

We believe that we can utilize a backward propagation approach using HMM learning algorithms[20].

5.2 STRUCTURE PREDICTION

[20] see https://www.researchgate.net/publication/344565580_Correlation_vs_Causation_The_Perils_of_AI In this paper we do discuss the strength and weakness of these approaches. With COVID-19 we have a reasonable set of baseline variants which we will examine shortly. Thus we are looking for most likely not dispositive causative factors.

Moving from two dimensional to full structure is ever so more difficult, As Senior et al note:

Protein structure prediction can be used to determine the three-dimensional shape of a protein from its amino acid sequence1. This problem is of fundamental importance as the structure of a protein largely determines its function; however, protein structures can be difficult to determine experimentally. Considerable progress has recently been made by leveraging genetic information. It is possible to infer which amino acid residues are in contact by analysing covariation in homologous sequences, which aids in the prediction of protein structures3.

Here we show that we can train a neural network to make accurate predictions of the distances between pairs of residues, which convey more information about the structure than contact predictions. Using this information, we construct a potential of mean force4 that can accurately describe the shape of a protein.

We find that the resulting potential can be optimized by a simple gradient descent algorithm to generate structures without complex sampling procedures.

The resulting system, named AlphaFold, achieves high accuracy, even for sequences with fewer homologous sequences. In the recent Critical Assessment of Protein Structure Prediction5 (CASP13)—a blind assessment of the state of thefield—AlphaFold created high-accuracy structures (with template modelling (TM) scores6 of 0.7 or higher) for 24 out of 43 free modelling domains, whereas the next best method, which used sampling and contact information, achieved such accuracy for only 14 out of 43 domains. AlphaFold represents a considerable advance in protein-structure prediction. We expect this increased accuracy to enable insights into the function and malfunction of proteins, especially in cases for which no structures for homologous proteins have been experimentally determined

What we believe is still lacking is a reasonably acceptable metric for effectiveness, such as binding affinity. Now as Yang et al (2020) have noted:

The prediction of interresidue contacts and distances from coevolutionary data using deep learning has considerably advanced protein structure prediction.

Here, we build on these advances by developing a deep residual network for predicting interresidue orientations, in addition to distances, and a Rosetta-constrained energy-minimization protocol for rapidly and accurately generating structure models guided by these restraints.

In benchmark tests on 13th Community-Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP13)- and Continuous Automated Model Evaluation (CAMEO)-derived sets, the method outperforms all previously described structureprediction methods. Although trained entirely on native proteins, the network consistently assigns higher probability to de novodesigned proteins, identifying the key fold-determining residues and providing an independent quantitative measure of the "ideality" of a protein structure. The method promises to be useful for a broad range of protein structure prediction and design problems …

The accurate prediction of the structure of de novo-designed proteins in the complete absence of coevolutionary signal has implications for both the model and protein design generally.

First, the model is clearly learning general features of protein structures.

This is not surprising given that the direct couplings derived by the coevolutionary analysis on a protein family are the 2-body terms in a generative model for the sequences in the family, and thus training on these couplings for a large number of protein families is equivalent to training on large sets of protein sequences for each structure in the training set.

From the design point of view, we have asserted previously that de novo-designed proteins are "ideal" versions of naturally occurring proteins; the higher probability assigned by the model to designed proteins compared to naturally occurring proteins makes this assertion quantitative. Remarkably, similar "ideal" features appear to have been distilled from native protein analysis by expert protein designers to be incorporated into designed proteins, and extracted by deep learning in the absence of any expert intervention.

Our finding that the model provides information on the contribution of each amino acid in a designed protein to the determination of the fold by the sequence suggests the model should be directly applicable to current challenges in de novo protein design.

This

representation we believe still lacks the completeness required. One must remember that in the binding domain, where we have two proteins and we seek to find the binding domain as well as some metric representing the strength of that binding, we need to begin perhaps in another direction.

We have further examined the approach taken from Lee et al:

Predicting a protein's structure from its amino acid sequence remains an unsolved problem after several decades of efforts. If the query protein has a homolog of known structure, the task is relatively easy and high-resolution models can often be built by copying and refining the framework of the solved structure.

However, a template-based modeling procedure does not help answer the questions of how and why a protein adopts its specific structure. In particular, if structural homologs do not exist, or exist but cannot be identified, models have to be constructed from scratch. This procedure, called ab initio modeling, is essential for a complete solution to the protein structure prediction problem; it can also help us understand the physicochemical principle of how proteins fold in nature.

Currently, the accuracy of ab initio modeling is low and the success is generally limited to small proteins (<120 residues). With the help of co-evolution based contact map predictions, success

in folding larger-size proteins was recently witnessed in blind testing experiments. In this chapter, we give a review on the field of ab initio structure modeling.

Our focus will be on three key components of the modeling algorithms: energy function design, conformational search, and model selection. Progress and advances of several representative algorithms will be discussed. …

Successful ab initio modeling from amino acid sequence alone is considered the "Holy Grail" of protein structure prediction , since this will mark an eventual and complete solution to the problem.

Again, our interest is not "ab initio" but an *a posteriori* approach, starting with the wild type Wuhan virus. The authors continue:

In addition to the generation of 3D structures, ab initio modeling can also help us understand the underlying principles of how proteins fold in nature; this could not be done by the template-based modeling approaches which build 3D models by copying and refining the framework of other solved structures. An ideal approach to ab initio modeling would be to treat atoms in a protein as interacting particles according to an accurate physics-based potential, and fold the protein by solving Newton's equations of motion in each step of movements. A number of molecular dynamics simulations were carried out along this line of approach by using the classic CHARMM and AMBER force fields. Although the MD based simulation is very important for the study of protein folding, the success in the viewpoint of structure prediction is quite limited.

One reason is the prohibitive computing demand for a normal size protein. On the other hand, knowledge-based (or hybrid knowledge- and physics-based) approaches making use of Monte Carlo sampling schemes appear to be progressing rapidly, producing many examples of successful low-to-medium accuracy models often with correct topology for small and medium size proteins. Although very rare, successful higher resolution models (<2–3 Å in Ca-RMSD) have been witnessed in blind experiments.

The current state-of-the-art ab initio protein structure prediction methods often utilize as much information as possible from known structures, in several different ways.

First, the use of local structure fragments directly excised from the PDB structures helps reduce the degrees of freedom and the entropy of the conformational search and yet keep the fidelity of the native protein structures.

Second, the knowledge-based potential derived from the statistics of a large number of solved structures can appropriately grasp the subtle balance of the complicated correlations between different sources of energy terms.

With the carefully parameterized knowledge-based potential terms aided by various advances in the conformational search methods, the accuracy of ab initio modeling for proteins up to 100–150 residues has been significantly improved in the last decade.

With the help of co-evolution based contact map predictions, an exciting examples has been recently reported on a free-modeling target (T0806) up to 258 residues in the most recent CASP experiment.

The use of evolutionary models may be helpful especially when one has some initial well understood starting point. They continue:

However, such performance is only possible when sufficient number of homologous sequences can be obtained to ensure the accuracy of contact predictions: this situation is rare for ab initio modeling target proteins that have no homologues in the PDB. For further improvement, parallel developments of accurate potential energy functions and efficient optimization methods are both necessary. That is, separate examination/development of potential energy functions is important; meanwhile, systematic benchmarking of various conformational search methods should be performed, so that the advantages as well as the limitations of available search methods can be explored separately.

Currently, the ab initio modeling methods solely based on the physicochemical principles of interaction are still far behind, in terms of their modeling speed and accuracy, compared with the methods utilizing bioinformatics and knowledge-based information. However, the physics-based atomic potentials have recently demonstrated their potential in refining the detailed packing of side-chain atoms and peptide backbones.

Development of composite methods using both knowledge-based and physics-based energy terms should represent a promising approach to the problem of ab initio modeling. It is important to acknowledge that with the progress in structure genomics and structural biology, the number of experimental structures in the PDB has been rapidly increasing, significantly extending the scope of the template-based protein structure predictions. Nevertheless, the traditional comparative modeling approaches can only yield model predictions with the accuracy of the templates whereas the efficiency of template structure refinements is highly correlated with our ability in ab initio protein folding, because structure refinements often involve reconstruction of part of the side-chain and local backbone structures, and sometime the global topology for the low-resolution templates.

Meanwhile, for most templates available in the PDB, a considerable portion of the sequence is either disordered or unaligned in the query-template alignments; the structures of these portions must be constructed using ab initio modeling. Finally, a very important bottleneck drawback in template-based modeling is that the alignment accuracy dramatically decreases with the sequence identity between query and template becomes low (e.g. <30%).

Most recently, it has been demonstrated that the structural models built by free modeling can be used to help identify analogous templates that are of low sequence similarity but high structural similarity to the native, by matching the low-resolution ab initio models to experimentally solved structures in the PDB and thereby improve the success rate of distant-homologous structure predictions.

Thus, the development of efficient ab initio folding algorithms will remain a major theme in the field and should have important impacts on all aspects of protein structure prediction.

It thus appears that the ab initio approach is still out of reach. But again that is not the direction we take. However the algorithms discussed above may very well prove of merit in the search we are after.

5.3 LIGAND PROTEIN INTERACTIONS

Ultimately we are focused on the two protein binding phenomenon. Our concern is to develop a useful metric to rank order protein-protein binding. Much work has been done in the pharmacology area regarding this problem, since it does go to the heart of many current therapeutics. (see Stevens p 254 and Patrick pp 116-117)

The question then is; given two well defined proteins, how best to determine their bonding propensity. As Yang et al note:

Predicting protein-ligand interactions using artificial intelligence (AI) models has attracted great interest in recent years. However, data-driven AI models unequivocally suffer from a lack of sufficiently large and unbiased datasets.

Here, we systematically investigated the data biases on the PDBbind and DUD-E datasets. We examined the model performance of atomic convolutional neural network (ACNN) on the PDBbind core set and achieved a Pearson R2 of 0.73 between experimental and predicted binding affinities.

Strikingly, the ACNN models did not require learning the essential protein-ligand interactions in complex structures and achieved similar performance even on datasets containing only ligand structures or only protein structures, while data splitting based on similarity clustering (protein sequence or ligand scaffold) significantly reduced the model performance.

We also identified the property and topology biases in the DUD-E dataset which led to the artificially increased enrichment performance of virtual screening. The property bias in DUD-E was reduced by enforcing the more stringent ligand property matching rules, while the topology bias still exists due to the use of molecular fingerprint similarity as a decoy selection criterion. Therefore, we believe that sufficiently large and unbiased datasets are desirable for training robust AI models to accurately predict protein-ligand interactions. …

State-of-the-art AI technologies represent a new paradigm in virtual screening with both opportunities and challenges for future improvement. The differences in different AI models mainly come from two aspects: one is the training dataset, and the other is the characterization method. At present work, we focused on analyzing the biases in two widely applied datasets for protein-ligand interactions.

We believe that we can indeed utilize current AI techniques as we have discussed previously. The challenge will still be the evolutionary changes and the metric for binding effectiveness. They continue:

The former is represented by PDBbind, a collection of experimentally determined proteinligand complex structures with known binding affinities, which is reliable, but the amount of data is small and arguably suffers from the data redundancy caused by the protein and ligand similarity. Our systematic investigation of ACNN models on the PDBbind datasets led to a surprising observation that the model performance was not correlated with learning essential proteinligand interactions.

Even the models trained on ligands or proteins performed as well as trained on complexes, while data splitting based on the similarity (protein sequence or ligand scaffold) clustering reduced the performance significantly.

This suggests that the model performance may rely on the similarity of atomic features existing in the training and test subsets. It is expected that the rapidly increased amount of protein-ligand binding and structural data will improve the generality of the models by sampling the much larger and diverse chemical space.

DUD-E has become a common dataset for evaluating structure based virtual screening methods, which were designed to benchmark enrichment performance by prioritizing the actives among a large amount of property-match but topology-dissimilar decoy molecules. As evidenced at present study, the topology bias is difficult to avoid when train on DUD-E. Therefore, care must be taken when using DUD-E for training AI models to predict protein-ligand interactions. However, DUD-E can still serve as an independent dataset to test the prediction power of AI models without using it for training.

The use of fingerprint for selecting topological dissimilar decoys in the DUD and DUD-E datasets introduces topology bias in cross-target, and even cross-class CV. If we want to perform crosstarget CV on DUD-like datasets for benchmarking AI models, the decoys shall be selected not only dissimilar to actives of a specific target, but also similar to actives of the other targets. Therefore, it is desirable to develop a more sophisticated approach for DUD-like decoy selection by depleting the topology bias, and such dataset may serve as a general-purpose benchmarking dataset to assess the enrichment performance of different virtual screening approaches (including AI models).

Nevertheless, it is encouraging that ACNN models have shown powerful capability for learning correlations hidden in structural data. Using the same neural network structure, ACNN was able to learn the structural similarities between ligands and between proteins. Even after protein sequence similarity clustering, ACNN still performed well in predicting ligand binding affinities. It is likely that ACNN model is well suitable for analysis of protein binding pocket, and it can be applied in protein pocket similarity analysis and protein pocket druggability prediction

Zhou et al have examined the binding issues for RNA and proteins, a slightly different issue from protein to protein binding. They note:

RNA–protein interactions permeate biology. Transcription, translation, and splicing all hinge on the recognition of structured RNA elements by RNA-binding proteins. Models of RNA–protein interactions are generally limited to short linear motifs and structures because of the vast sequence sampling required to access longer elements. Here, we develop an integrated approach that calculates global pairwise interaction scores from in vitro selection and high-throughput sequencing.

We examine four RNA-binding proteins of phage, viral, and human origin. Our approach reveals regulatory motifs, discriminates between regulated and non-regulated RNAs within their native genomic context, and correctly predicts the consequence of mutational events on binding activity. We design binding elements that improve binding activity in cells and infer mutational pathways that reveal permissive versus disruptive evolutionary trajectories between regulated motifs. These coupling landscapes are broadly applicable for the discovery and characterization of protein–RNA recognition at single nucleotide resolution.

Their approach is of interest in possibly more general bonding.

5.4 EVOLUTIONARY FACTORS

Viral evolution is a complex process. It entails global evolutionary changes and local changes[21]. Evolutionary changes are the most concerning long term. The virus can learn to adapt and spread across species. Local changes may simply be the result of simple deletions or additions just y chance. Evolutionary changes are much mor adaptive and thus more aggressive. We briefly consider evolutionary but to model them in the context of short term impact is both difficult and prone to a lack in specificity.

Let us begin with some simple first steps. As Sirovetz et al have noted:

*Protein sequences have evolved to fold into functional structures, resulting in families of diverse protein sequences that all share the same overall fold. One can harness protein family sequence data to infer likely contacts between pairs of residues. In the current study, we combine this kind of inference from coevolutionary information with a coarse-grained protein force field ordinarily used with single sequence input, the A**ssociative memory, Water mediated, Structure and Energy Model (AWSEM),** to achieve improved structure prediction.*

The resulting Associative memory, Water mediated, Structure and Energy Model with Evolutionary Restraints (AWSEM-ER) yields a significant improvement in the quality of protein structure prediction over the single sequence prediction from AWSEM when a sufficiently large number of homologous sequences are available.

Free energy landscape analysis shows that the addition of the evolutionary term shifts the free energy minimum to more nativelike structures, which explains the improvement in the quality of structures when performing predictions using simulated annealing.

[21] https://www.nature.com/scitable/topicpage/the-origins-of-viruses-14398218/

Simulations using AWSEM without coevolutionary information have proved useful in elucidating not only protein folding behavior, but also mechanisms of protein function. The success of AWSEM-ER in de novo structure prediction suggests that the enhanced model opens the door to functional studies of proteins even when no experimentally solved structures are available … While understanding the general principles of protein folding is of scientific interest, this understanding is also helpful for devising algorithms that can be used to predict protein structure.

Nevertheless, predicting detailed high-resolution structures of particular proteins remains an important but difficult task in practice. The structure prediction problem can be stated in the following way. Given the sequence of amino acids for a specific protein, without doing additional experiments on the protein, can we predict the native structure into which that protein will fold? For an ever-growing number of sequences the answer is now \yes".

With the dramatic increases in computational power and the number of experimentally known structures in the protein data bank (PDB, knowledge-based prediction methods are proving able to achieve usefully accurate predictions of structure.

For some time now, the most accurate method of protein structure prediction has been homology modeling, which involves exploiting knowledge of evolutionary descent.[5] Homology modeling involves searching for sequences that are globally similar to the target sequence and then, hopefully, finding a match to a protein whose structure has been solved already. One can generally assume that, if two sequences are sufficiently similar, they must have evolved from a common ancestor and will also share similar structures.

Homology modeling works as a consequence of evolutionary restraints: in order for a modern protein along with all its ancient ancestors and cousins to have always functioned in the same way, they all must have folded throughout history into structures which must not have changed much. As the number of solved protein structures has increased, so has the power of homology modeling based on analogy to known structures. It has become increasingly rare to find globally novel folds. Nevertheless, at least in some cases, it remains a challenge to recognize whether or not a specific protein has a homolog with a solved structure prior to the experimental determination of the structure. Such proteins are sometimes said to be in a \twilight zone" of homology inference….

Adding coevolutionary information from families of proteins can strongly improve the structure prediction capabilities of AWSEM, which ordinarily only uses a single target sequence as input. It is important to emphasize that none of the algorithms surveyed in this study were in a mode that used any structural information from homologous proteins.

Again, we must have some reasonable model for evolutionary changes. In our current problem, we can reasonably assume stable ACE2 whereas we have a highly mutable spike.

(The only exception to this is in the supplementary information where we show results of including homologue input explicitly, for comparison. Using structural homologues in the fragment memory term gives results comparable to other schemes of homology modeling.)

Including evolutionary information in structure prediction is a powerful part of the protein structure prediction toolkit, especially when no structurally solved homologues can be recognized. The incorporation of coevolutionary information into the AWSEM model now opens the door to mechanistic studies of functional proteins that do not have experimentally solved structures and are too large to simulate on biological timescales with all-atom models.

Returning to the evolutionary factor we refer to Woo and Reifman who have noted:

Viruses with RNA genomes evolve rapidly, evading selective pressure from the host immune response and adapting to changing environments.

In particular, their capacity to switch host species, emerge into new vulnerable populations, and cause outbreaks has significant implications for viral disease control.

Many such outbreaks in recent history have been attributed to viral species jumps: influenza A virus has jumped from birds and pigs to humans multiple times, the severe acute respiratory syndrome (SARS) epidemic was caused by a species jump of coronavirus from bats and palm civets to humans, and human immunodeficiency virus type 1 (HIV-1) is believed to have switched hosts from primates.

Characterizing the complex factors affecting the evolution of viruses in natural settings among diverse groups of interacting hosts is a challenging task. Valuable insights have been gained by evolutionary experiments under more controlled conditions, particularly within the context of serial-passage experiments. In a serial-passage experiment, a cell culture or live host is inoculated by viral (or other) pathogens, usually already well adapted to different cell types or hosts. A pathogen's growth under the restrictive host environment leads to within-host selection for advantageous variants, either present as a minority in the founder population or generated from error-prone replications.

After a certain amount of time (approximately days) of such growths, a small subset of the resulting pathogen population is sampled and used to inoculate a fresh new medium or host, initiating a subsequent round of the passage.

Generally, rapid adaptation to the new host environment in the form of increased fitness and virulence is observed (typically within 10 passages) along with attenuation of adaptation to the former host. In addition to revealing key adaptation strategies a pathogen can exhibit, an important application of serial passages is the production of attenuated vaccine strains capable of eliciting immune responses without virulence (16).

The recent availability of rapid and inexpensive deep-sequencing techniques has the potential to significantly improve our understanding of how key factors affect virus evolutionary dynamics, including species jumps, especially when combined with controlled experiments such as serial

passages. A major obstacle in leveraging the growing sequence data, however, is the lack of quantitative connections between such genotype data and experimentally measurable viral phenotypes, including growth properties, infectivity, virulence, and tissue tropism.

The main objective of this study was to develop, validate, and apply stochastic models of viral evolutionary dynamics that can bridge this gap by realistically modeling genomic diversification and adaptation processes during serial passages …

We may infer insights into a better understanding of evolutionary dynamics of viral pathogens in natural settings from the study of serial passages. Systematic investigation of the effects of factors influencing the relative ease of species jump…can play important roles in such interpretations. For instance, the host cell number and bottleneck size roughly correspond to the size of susceptible host populations and the degree and severity of selection present during the natural spread of a viral infection, whereas the infection and death rates characterize the population dynamics within this host environment.

In this viewpoint, the longitudinal patterns of population size and quasispecies structure … can be regarded as idealizations of what typically happens in natural infectious cycles: a viral strain ventures into a new host environment, completing multiple cycles of rapid growth followed by stagnation due to host depletion and reemergence in a fresh host population, but with factors characterizing each round highly variable and unpredictable rather than uniform as in passages.

Our study of evolutionary dynamics under passages suggests that adaptation mostly occurs in the early growth phase, an adapting viral population may frequently get stuck without a direct route to a neighboring highly fit strain*, and the* ***speed of adaptation is most sensitive to the distance and target/bottleneck sizes****. The specific instance of adaptation simulated under the empirically derived influenza A virus fitness landscape explicitly demonstrates that in reality, a quasispecies adaptation involves the stochastic evolution of both the master sequence and its clouds, instead of a simple jump from the WT genotype to a global MF sequence.*

Finally we have the evolutionary issue of interaction as discussed by Manrubia and Lázaro note:

Interactions between viruses and hosts either at the individual or at the population level have a profound influence in the evolution of RNA viruses. Part of this evolution is reflected in the topology of the corresponding phylogenetic trees.

The type of infection caused by the virus (acute or persistent), the duration of the immunity elicited by the pathogen, the rate of the replenishment of susceptible hosts, the capacity of the virus to acquire immune-escape mutations, or the transmission mode, are some of the factors that affect the evolutionary dynamics of RNA viruses.

Recent developments in molecular biology and bioinformatics have permitted to reconstruct the phylogenies of several RNA viruses along a number of years. The comparison of the phylogenetic trees obtained reveals important differences that in many cases can be explained on the basis of the epidemiological behaviour of each virus. The reconstruction of the phylogeny of

the influenza subtype H3N2 shows that the evolution of the virus follows a single track in the long term .

The tree has a main trunk with branching points representing drifting strains that become extinct in an approximately one-year period. This type of topology likely results from the continuous emergence of new virus variants that can evade recognition by the immune system through a change in the properties of their antigenic sites. In the case of influenza the renew of susceptibles comes from births of new hosts and from regained susceptibility of previously infected individuals to new influenza variants.

The short infectious periods of influenza, together with its high genetic variation, leads to a rapid strain turnover, probably one of the reasons producing the strong seasonality of influenza epidemics. An unresolved question concerns the transmission mode of influenza through respiratory droplets containing just one or a few viral particles. This strong reduction in population size could lead to the fixation of mutations during optimization of the virus inside each particular host.

The expected result should be a very high diversity among lineages (infecting different hosts) that is not observed in natural populations of the virus. Mathematical models of influenza epidemics have approached this question by taking into account the balance between the production of new strains and the competition-induced stochastic extinction of existing variants. The results obtained show that, if cross-immunity is the only form of competition between strains, there is an exponential growth of diversity. However, if a short-lived immunity that inhibits reinfection by any new strain during a short time is considered in the model, the probability of explosive diversity growth is strongly reduced.

A different type of phylogenetic tree is exhibited by viruses that provoke a strong cross-immunity, as is the case of measles. In this case the epidemic cycles arise from repeated exhaustion of susceptible hosts combined with the lifelong immunity elicited by the pathogen. An immune response equally effective against all strains does not promote selection of the best adapted genotypes. Therefore, many strains can coexist with relative frequency, and the topology of the phylogenetic tree is mainly determined by spatial-temporal dynamics.

The situation is clearly different in the case of persistent infections, as those established by HIV and hepatitis C virus. The long period of evolution experienced by these viruses inside their hosts means that both dynamics, intra- and inter-host should be reflected in their phylogenetic trees. Due to the long time between transmission events, epidemic cycles are not observable in the case of these viruses. Instead of that, what is observed is a growing epidemic trend.

The fact that the infection by a subtype does not protect against new reinfections with other subtypes of HIV or HCV results in a phylogeny that mainly reflects the demographic and spatial history of transmission. The continuous evolution of these viruses inside their hosts permits to perform intra-host phylogenies. In HIV, the highly immunogenic envelope gene presents and intra-host phylogeny similar to that exhibited by influenza A at the population level. The presence of drugs to treat the infection is another factor that strongly conditions the intra-host phylogeny of the virus.

Phylogenies of RNA viruses are very difficult to interpret due to the large number of factors involved.

To know the precise interactions between the immune response, the genetic variation of the pathogen and the inter-strain competition (whose intensity depends on the mode of transmission and on the nature of the infection–acute or persistent) does not suffice. Moreover, all these factors must be integrated in the context of population dynamics of the host at the spatial-temporal level. An important advance is being made in this field by contrasting mathematical approaches with the actual behaviour of viruses in nature. Progress in the amount of field data at hand comes from the large number of virus genome sequences available and from the increased number and precision of epidemiological data.

Thus understanding the evolutionary constraints will be critical but may be difficult to model.

6 VARIANT TARGETING

We have demonstrated that we know the wild type spike protein and the ACE2 receptor protein. We know the binding sites. Now we pose the following problem:

1. Take the wild type and consider variants based upon evolutionary mutations. Or use the wild type, WT, and using existing variants

2. Eliminate any variants which have no binding

3. Consider binding site strengths for known variant combinations.

4. Perform evolutionary variant changes and calculate binding strength.

6.1 VARIANTS

Let us examine some recent examples. From Ashoor et al we have the following Table which depicts their analysis showing the n umber of polar and hydrophobic interactions between the 18 ACE2 contact residues, on 9 ACE2 variants, and on 24 spike variants including the Wuhan wild type. The authors note:

*Table … summarizes the variations of molecular interactions numbers and patterns between the **SARS-CoV-2 RBD** mutations occurring in the 18 residues the virus engages to interact with the **ACE2 isoform1** and the eight variants found to be significantly destabilizing when tested against the Wuhan RBD sequence in comparison to hACE2 isoform1. The **number and position of polar and hydrophobic bonds**[22] in addition to the **salt bridges** characterize each interaction pattern. The most striking interactions variations concern SARS-CoV-2 RBD mutants N501T and N501S and ACE2 variant D355N.*

[22] Recall that hydrophobic bonds in proteins arise as a result of the interaction of their hydrophobic amino acids with the polar solvent, namely water. The hydrophobic amino acids are gly, ala, val, leu, ile, met, pro, phe, trp. These amino acids have hydrocarbon sidechains that, because of their non-polar chemistry, are forced into close association (hydrophobic "bonds") in an aqueous solvent. To understand the formation of hydrophobic bonds, familiarity with the energetics that drive the packing of solvent H2O molecules into liquid lattices is required. In contrast the polar bonds result from the typical OH coupling having unequal charge distribution.

RBD sequence residues		ACE2 variants																	
		Isoform 1		D355A		D355N		E35D		E35K		F4		IL		M82I		T27A		S1ꜝ	IP
		P	H	P	H	P	H	P	H	P	H	P	H	P	H	P	H	P	H
Ref. sequence	SARS-CoV‧2 Wuhan strain	12	65	12	64	13	68	13	65	12	65	12	65	12	65	12	64	11	64
Mutations	K417N	11	64	11	63	12	67	12	64	11	64	11	64	11	64	11	63	10	63
	K417R	12	68	12	67	13	71	13	68	12	68	12	68	12	68	12	67	11	67
	G446A	11	65	11	64	12	68	12	65	11	65	11	65	11	65	11	64	10	64
	Y449N	10	62	10	61	11	65	11	62	10	62	10	62	10	62	10	61	9	61
	Y453F	11	59	11	58	12	62	12	59	11	59	11	59	11	59	11	58	9	58
	L455F	12	70	12	69	13	73	13	70	12	70	12	70	12	70	12	69	11	69
	F456L	12	64	12	63	13	67	13	64	12	64	12	64	12	64	12	62	11	63
	A475V	12	68	12	67	13	71	13	68	12	68	12	68	12	68	12	66	11	66
	G476S	12	64	12	63	13	67	13	64	12	64	12	64	12	64	12	63	11	63
	G476A	12	64	12	63	13	67	13	64	12	64	12	64	12	64	12	63	11	63
	F486L	12	62	12	61	13	65	13	62	12	62	12	62	1262		12	61	11	61
	Q493R	13	66	13	65	14	69	12+ 1 salt	62	12	62	13	66	13	66	13	65	12	65
	Q493L	12	64	12	63	13	67	12	64	12	64	12	64	12	64	12	63	11	63
	G496C	12	71	12	70	13	74	13	71	12	71	12	71	12	71	12	70	11	70
	T500I	11	61	11	58	11	60	12	59	11	59	11	59	11	59	11	58	10	58
	N501T	12	66	12	65	13	69	13	66	12	66	12	66	12	66	12	65	11	65
	N501S	12	65	12	64	13	68	13	65	12	65	12	65	12	65	12	64	11	64
	N501Y	12	72	12	71	13	75	13	72	12	72	12	72	12	72	12	71	11	71
	G502D	12	69	12	68	13	72	13	69	12	69	12	69	12	69	12	68	11	68
	G502C	12	68	12	67	13	71	13	68	12	68	12	68	12	68	12	67	11	67
	G502R	12	70	12	69	13	73	13	70	12	70	12	70	12	70	12	69	11	69
	Y505H	12	61	12	60	13	64	13	61	12	61	12	61	12	61	12	60	11	60
	Y505E	12+ 1 salt	66	12+ 1 salt	65	13+ 1 salt	69	13+ 1 salt	66	12+ 1 salt	66	12+ 1 salt	66	12+ 1salt	66	12+ 1salt	65	11 + 1salt	65

The above chart provides certain starting points.

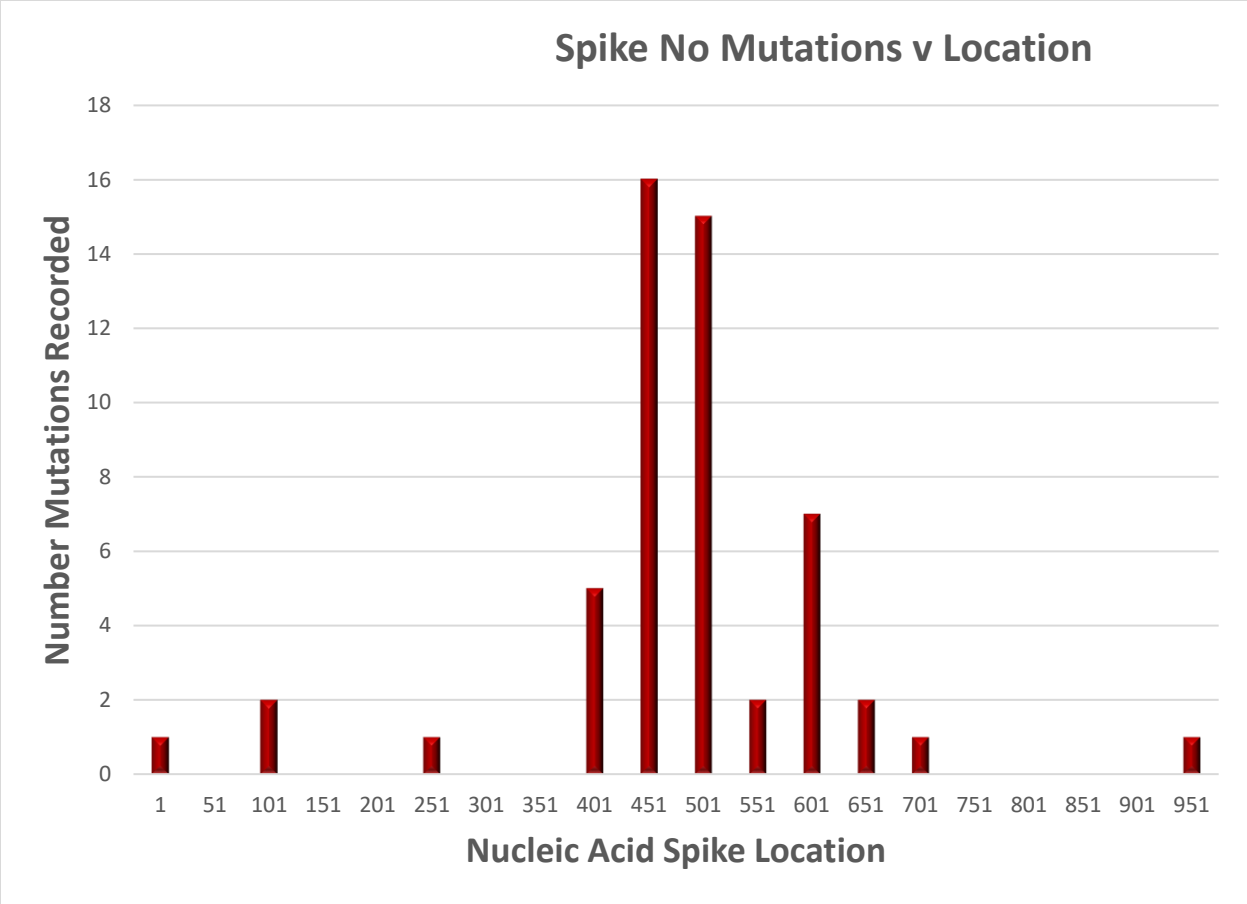
1. The spike proteins shown above have evolved along certain paths. These paths can be extended and a list of all the putative mutations and variants can be obtained.

2. The polar bonds between the two proteins are generally smaller than the putative hydrophobic bonding.

Variants or mutations may occur in a variety of ways. Our focus is on local variants, ones where we see a change of a nucleotide, a deletion or perhaps a deletion. We would not consider complex variants that do occur but may result in a dramatic change. As can be seen above, the variants may be single or multiple points. We list below the variants as shown by Lan et al and present them in a increasing location on the spike protein.

Mutation	Nucleic Acid Number	From	To
K417N	417	K	N
K417R		K	R
G446A	446	G	A
Y449N		Y	N
Y453F	453	Y	F
L455F	455	L	F
F456L	456	F	L
A475V	475	A	V
G476S	476	G	S
G476A		G	A
F486L	486	F	L
Q493R	493	Q	R
Q493L		Q	L
G496C	496	G	C
T500I	500	T	I
N501T	501	N	T
N501S		N	S
N501Y		N	Y
G502D	505	G	D
G502C		G	C
G502R		G	R
Y505H	505	Y	H
Y505E		Y	E

We then plot the number per segmented location as shown below. As will be seen the dominant mutations are centered on the protein nucleic acid sequences. This chart represents an amalgam of all known variants as of the publication date.

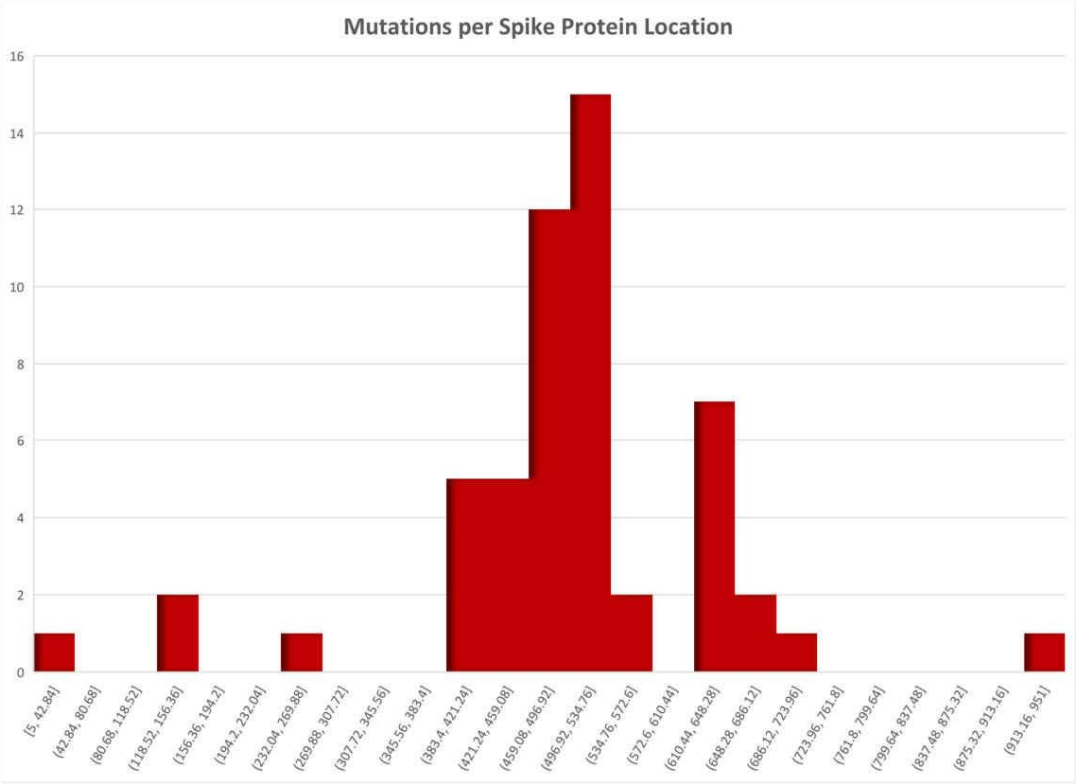


For simplicity the chart below summarizes the nucleic acid definitions and related RNA sequences so one can readily see what putative nucleotide has changed.

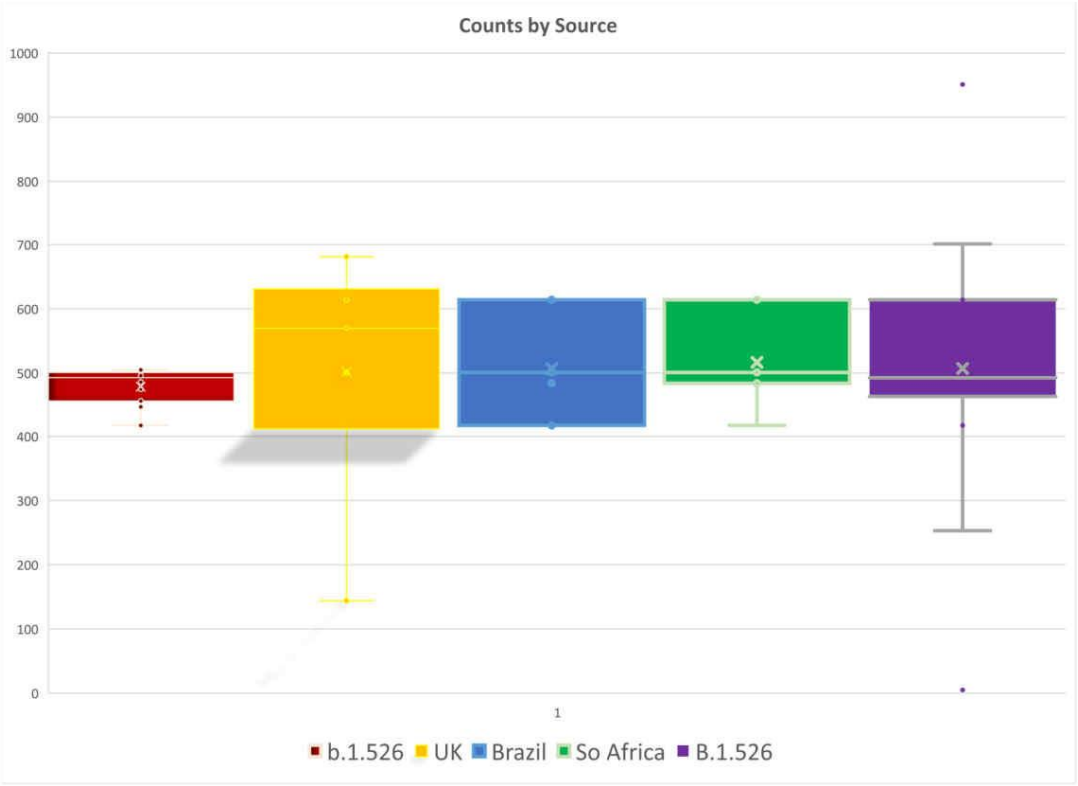
Alanine	Ala	**A**	**GCA GCC GCG GCU**
Cysteine	Cys	C	**UGC UGU**
Aspartic acid	Asp	**D**	**GAC GAU**
Glutamic acid	Glu	**E**	**GAA GAG**
Phenylalanine	Phe	**F**	**UUC UUU**
Glycine	Gly	**G**	**GGA GGC GGG GGU**
Histidine	His	**H**	**CAC CAU**
Isoleucine	lie	I	**AUA AUC AUU**
Lysine	Lys	**K**	**AAA AAG**
Leucine	Leu	**L**	**UUA UUG CUA CUC CUG CUU**
Methionine	Met	M	**AUG**
Asparagine	Asn	**N**	**AAC AAU**
Proline	Pro	**P**	**CCA CCC CCG CCU**
Glutamine	Gin	**Q**	**CAA CAG**

Arginine	Arg	**R**	**AGA AGG CGA CGC CGG CGU**
Serine	Ser	**S**	**AGC AGU UCA UCC UCG UCU**
Threonine	Thr	**T**	**ACA ACC ACG ACU**
Valine	Val	**V**	**GUA GUC GUG GUU**
Tryptophan	Trp	**W**	**UGG**
Tyrosine	Tyr	**Y**	**UAC UAU**

We now take the Lan et al variants along with the CDC current variants and plot another set of sequence counts using all current available data. Again, we see the center clusters. There are end points which now depict some changes as well. This does not reflect deletions.



We have shown below the spread of variation for each of the four data sets we have examined. The UK has the broadest spread and Lan the least. It must be remembered that Lan et al was presented early on whereas the other three are current actual variants. Note we have two spreads on the B.1.526 New York variant. Since this is a recent variant we have included both measures[23].

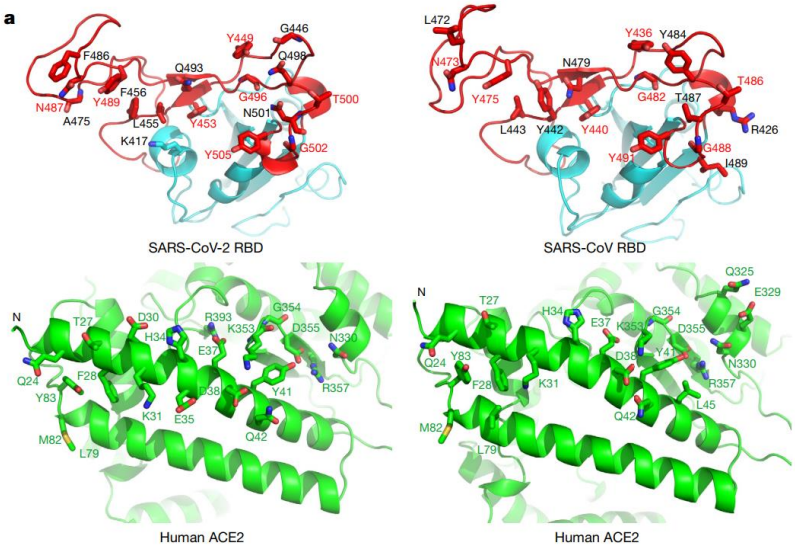


Overall, we can now make some useful observations. Namely that the variants often have center clustered mutations which means that one may just need to focus on variants centered at the spike RNA. That may allow for a lower computational load.

6.2 PROTEINS

Lan et al have examined the binding spots of SARS-CoV-2 and SARS-CoV and this is shown below:

[23] See Annavajhala et al they note: *Phylogenetic analyses of sequences in the database further reveal that this B.1.526 variant is scattered in the Northeast of US, and its unique set of spike mutations may also pose an antigenic challenge for current interventions.*



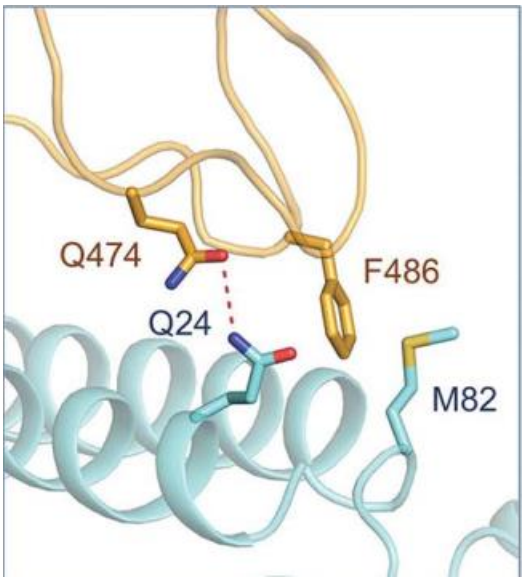
Lan et al also present an interesting Table delineating the binding sites between ACE2 WT and the current SARS and the previous version. This is shown below (as modified) and this shows some overlap between the two regarding ACE2 sites. There is a considerable amount of difference and sites not in both.

	SARS-CoV-2 RBD	Length (A)	ACE2	Length (A)	SARS-CoV RBD
Hydrogen bonds	N487(ND2)	2.6	Q24(OE1)	2.9	N473(ND2)
	K417(NZ)	3.0	D30(OD2)		
	Q493(NE2)	2.8	E35(OE2)		
			E37(OE1)	3.4	Y491(OH)
	Y505(OH)	3.2	E37(OE2)		
			D38(OD1)	3.0	Y436(OH)
	Y449(OH)	2.7	D38(OD2)	3.0	Y436(OH)
	T500(OG1)	2.6	Y41(OH)	2.8	T486(OG1)
	N501(N)	3.7	Y41(OH)	3.3	T487(N)
	G446(O)	3.3	Q42(NE2)		
	Y449(OH)	3.0	Q42(NE2)		
			Q42(OE1)	2.7	Y436(OH)
	Y489(OH)	3.5	Y83(OH)	3.3	Y475(OH)
	N487(OD1)	2.7	Y83(OH)	2.8	N473(ND2)
			Q325(OE1)	3.8	R426(NH2)
			E329(OE2)	3.0	R426(NH2)
			N330(ND2)	2.8	T486(O)
	G502(N)	2.8	K353(O)	2.6	G488(N)
	Y505(OH)	3.7	R393(NH2)		
Salt bridges[24]	K417(NZ)	3.9	D30(OD1)		
	K417(NZ)	3.0	D30(OD2)		
			E329(OE2)	3.7	R426(NH1)
			E329(OE1)	3.9	R426(NH2)
			E329(OE2)	3.0	R426(NH2)

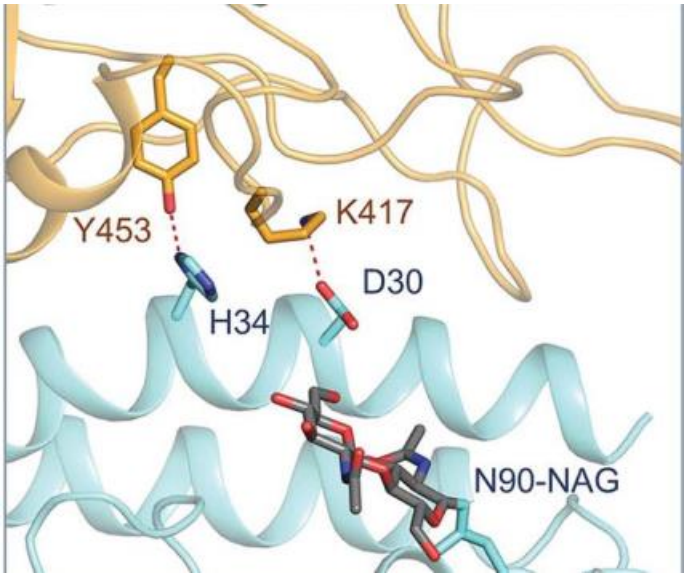
Note: ND2, nitrogen delta 2; NE2, nitrogen epsilon 2; NZ, nitrogen zeta; N, nitrogen; NH1 ،nitrogen eta 1: NH2 ،nitrogen eta 2; OH ،oxygen eta،■ O ،oxygen،· OD1 ،oxygen delta 1،· OD2،· oxygen delta 2; OG1, oxygen gamma 1,- OE1 ،oxygen epsilon 1,- OE2, oxygen epsilon 2.

Yan et al also presented their analysis of the binding elements as shown below:

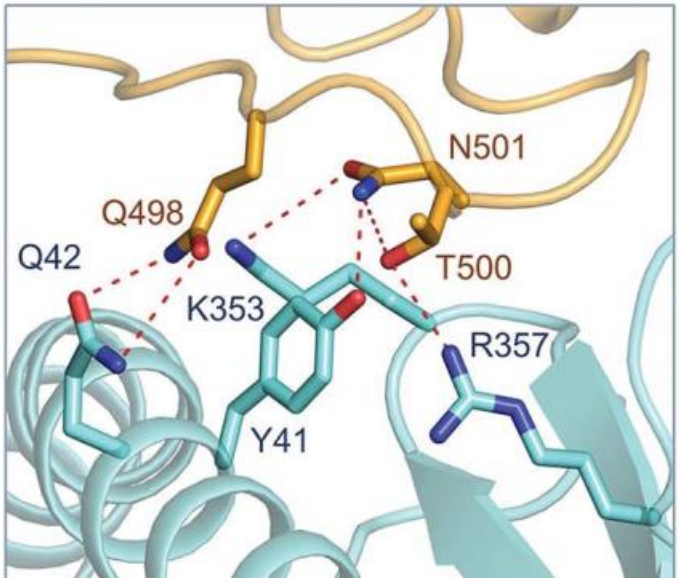
[24] A salt bridge is a complex bond consisting two non-covalent interactions, one the typical hydrogen bond and the second an ionic bond. Thus the polar is an O-H bond and the ionic may be an O-HN type bond.



In the above we see the Q474 spike bond to the Q24 ACE2 element. As with all of these bonds we should attempt to determine the metric which best defined the effectiveness of this bond. We have made several suggestions herein.



In the above we have two bonds, first the spike Y453 to the ACE2 H34 and second the spike K417 to the ACE2 D30. It is interesting to see the central nucleic acids of the spike binding to the tail ends of the ACE2.



In this case we have the spike Q498 binding is what may be a salt bridge with Q42, then the spike N501 to the K353 and apparently the R357. It also appears to have the spike T500 bonging with R357. Here we see movement on ACE2 bonding to the central region.

These multiple bonds may result in a strong bonding to the ACE2 resulting in the ability for the virion to enter the cell.

Here, in the above examples, the red elements are polar interactions. The nucleic acids and their locations on the respective proteins are shown. It is worth comparing these to the Table by Lan et al previously. Some of these linkages are at variance. However this information can be used for protein perturbation analysis.

7 IMMUNE SYSTEM ISSUES

The primary focus has been on antibodies to the spike protein. Classic vaccines used killed viruses or otherwise disabled viruses. Thus they presented to the immune system the full panoply of antigens, not only a spike but for the most part all the structural elements. From this Ab were produced and provided immunity. The question is; what were the Ab developed against, that is what was the Ag?

Recently there has been a focus on CTL, cytotoxic T lymphocytes or killer T cells. In a recent Nature article they note[25]:

Coronavirus vaccine development has largely focused on antibodies, … Antibodies, particularly those that bind to crucial viral proteins and block infection, can hold the key to 'sterilizing immunity', which not only reduces the severity of an illness, but prevents infection altogether. That level of protection is considered the gold standard, but typically it requires large numbers of antibodies …

Alongside antibodies, the immune system produces a battalion of T cells that can target viruses. Some of these, known as killer T cells (or CD8+ T cells), seek out and destroy cells that are infected with the virus. Others, called helper T cells (or CD4+ T cells) are important for various immune functions, including stimulating the production of antibodies and killer T cells. T cells do not prevent infection, because they kick into action only after a virus has infiltrated the body. But they are important for clearing an infection that has already started. In the case of COVID-19, killer T cells could mean the difference between a mild infection and a severe one that requires hospital treatment …

T cells, by contrast, can target viral proteins expressed inside infected cells, and some of those proteins are very stable, she says. This raises the possibility of designing vaccines against proteins that mutate less frequently than spike, and incorporating targets from multiple proteins into one vaccine.

7.1 B CELLS, ANTIBODIES AND KILLER T CELLS

As Abbas notes regarding cytotoxic T cells, CTLs, or more euphemistically "killer T cells", they work as follows:

The T cells expressing CD8+ proliferate and differentiate into cytotoxic T lymphocytes (CTLs), which express cytotoxic granules and can kill infected cells.

The differentiation of CD8+ T cells into functional **CTLs and memory cells** requires recognition of antigen presented by **dendritic cells**, **signals from CD4+ helper T** cells in some

[25] https://www.nature.com/articles/d41586-021-00367-7

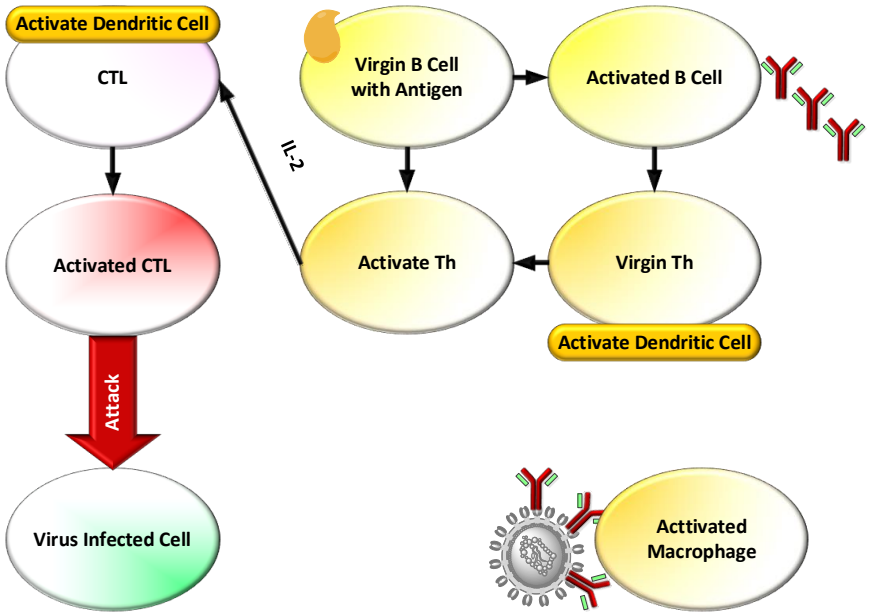
situations, costimulation, and cytokines. Differentiation to CTLs involves the acquisition of the machinery to kill target cells and is driven by various transcription factors.

In some situations of chronic antigen exposure (such as tumors and chronic viral infections), CD8+ T cells initiate a response but begin to express inhibitory receptors that suppress the response, a process called exhaustion.

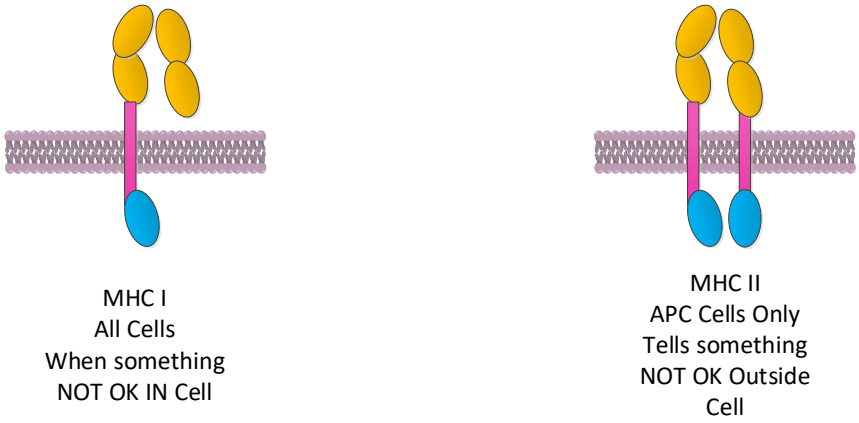
CD8+ CTLs kill cells that express peptides derived from cytosolic antigens (e.g., viral antigens) that are presented in association with class I MHC molecules. CTL-mediated killing is mediated mainly by granule exocytosis, which releases granzymes and perforin. Perforin facilitates granzyme entry into the cytoplasm of target cells, and granzymes initiate the process of apoptosis.

CD8+ T cells also secrete IFN-γ and thus may participate in defense against phagocytosed microbes and in delayed type hypersensitivity (DTH) reactions.

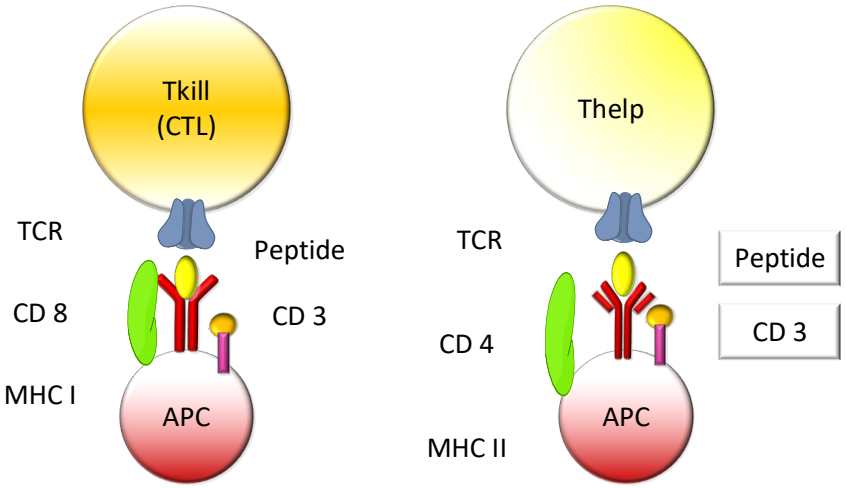
We demonstrate the overall dynamics of this process below. B cells can generate Ab from the presented Ag but then T cells clean up the identified cells. In contrast, CTL can themselves remember the Ag after it has been presented and then subsequently attack the infected substances.



The CTL has MHC I receptors as shown below. In contrast most cells have MHC II receptors to identify them as self.



The dynamics of the CTL process is then shown in the small below.



Thus CTL have memory, memory of the specific Ag without any Ab needed and then can remain active immune system defenders

7.2 Vaccines and Killer T Cells

Recent work has been done regarding CTL and COVID-19. As Tarke et al have noted:

This study presents a comprehensive analysis of the patterns of epitope recognition associated with SARS-CoV-2 infection in a cohort of approximately 100 different convalescent donors

spanning a range of peak COVID-19 disease severity representative of the observed distribution in the San Diego area. SARS-CoV-2 was probed using 1,925 different overlapping peptides spanning the entire viral proteome, ensuring an unbiased coverage of the different HLA class II alleles expressed in the donor cohort.

For HLA class I, we used an alternative approach, selecting 5,600 predicted binders for 28 prominent HLA class I alleles, representing 61% of the HLA A and B allelic variants in the worldwide population, and affording an overall 98.8% HLA class I coverage at the phenotypic level. The biological relevance of the epitope characterization studies summarized here is underlined by the use of the ex vivo AIM assay that does not require in vitro stimulation, which potentially skews the results by eliciting responses from naive cells.

The AIM assay is also more agnostic for different types of CD4+ T cells, as it measures all activated cells, regardless of T cell subset or any particular pattern of cytokine secretion. …

Finally, the functional relevance of our study was highlighted by the generation of improved epitope MPs for measuring T cell responses to SARS-CoV-2; these experimentally defined pools are associated with increased activity and lower complexity when compared to our previous MPs based on overlapping and predicted peptides.

We plan to make these epitope pools available to the scientific community at large and expect that they will facilitate further investigation of the role of T cell immunity in SARS-CoV-2 infection and COVID-19. In conclusion, we identify several hundred different HLA class I and class II restricted SARS-CoV-2-derived epitopes.

We anticipate that these results will be of significant value in terms of basic investigation of SARS-CoV-2 immune responses and in the development of both multimeric staining reagents and Tcell-based diagnostics. In addition, the results shed light on the mechanisms of immunodominance of SARS-CoV-2, which have implications for understanding host-virus interactions, as well as for vaccine design.

Now we also must remember the usefulness of memory T cells. As Omilusik and Goldrath comment:

Vaccination has substantially reduced illness and death from infectious disease by exploiting the ability of long-lived memory T cells to 'remember' a previous encounter with a specific microbe and mount a rapid response upon pathogen re-exposure. Understanding how immunological memory is established and maintained might provide insights that could enable improvements in vaccine design. … [26]

Naive T cells are those that haven't previously responded to a pathogen. When they recognize a pathogen, they rapidly divide and express molecules such as cytokine proteins that help to fight infection. These responding cells are called effector T cells (more specifically, a type of effector cell called a cytotoxic T cell) and they can migrate into inflamed tissues and kill infected cells.

[26] See Akondy et al.as well as Youngblood et al. who reveal the cell population that gives rise to memory T cells, and how the population of memory T cells evolves.

Once the pathogen is eliminated, most effector cells die, but a small pool of long-lived memory cells remains that is poised to respond rapidly if reinfection occurs. Which cells give rise to memory T cells has been extensively investigated. Two general possibilities have been proposed: the cells either arise from a subset of the effector cells that escape death, or instead descend directly from naive T cells, which could, as early as their first cell division, give rise to cells with effector-T-cell or memory-T-cell potential.

Memory T cells appear to retain immunogenicity for significant periods. They have the capability to identify infected cells and attack them as they do the virus itself. They continue:

Virus-specific memory T cells that were present between one and two years after vaccination showed little dilution of the deuterium, indicating that minimal cell division had occurred in these cells. In a comparison of cell-surface proteins, these memory cells resembled naive T cells, but the deuterium labelling reveals that they were formed from the dividing effector-T-cell population. These results support a model in which viral-specific CD8+ T cells extensively proliferate upon pathogen recognition and modify their DNA to favour expression of effector molecules. Later the cells stop dividing, stop expressing effector genes and reexpress many genes associated with the naive state, such as genes that aid T-cell survival and migration.

There has been a recent interest in these memory T cells as the backstop in the COVID vaccine scenario. However their function and underlying cellular dynamics are still only partially understood.

8 OBSERVATIONS

We can now make several observations.

8.1 ACE2 VARIANTS ARE GENERALLY WELL KNOWN.

ACE2 is the target for the spike protein. However there are a multiplicity of ACE2 receptors, many allowing a more aggressive form of viral attack. Yet their structure is understood in detail including the binding sites. Thus one end of the problem seems well enough defined to then seek the other end of the problem.

8.2 ACE2 VARIANTS CAN RESULT IN CO-MORBIDITIES AND ALSO INCREASED MORBIDITY AND MORTALITIES IN COVID INFECTIONS

Many of the ACE2 variants also are promoters of co-morbidities which are known to make viral infections of increased morbidity and mortality. Yet to date there does not appear to be any program to ascertain ACE2 structures as well as the attacking virion structure including the spike. Frankly both must be obtained and catalogued in order to effectively deal with this and future viral attacks.

8.3 COVID-19 VARIANTS ARE INCREASING

COVID-19 variants are increasing, some with greatly increased morbidity, mortality and infectivity. We have data on some of these, albeit quite limited.

8.4 COVID-19 VARIANTS LACK REQUIRED TRACKING GLOBALLY

It is clear that it is essential that there be an open transparent and timely database of COVID variants. We have argued for this for the past year but it appears that entities like the CDC prefer to keep such under wraps.

8.5 LIGAND CONNECTIONS ARE WELL KNOWN FOR WILD TYPE ACE2 AND COVID-19

We now know the ligand connections for the WT spike. As Yi et al have noted:

Coronavirus disease 2019 (COVID-19), caused by the novel human coronavirus SARS-CoV-2, is currently a major threat to public health worldwide. The viral spike protein binds the host receptor angiotensin-converting enzyme 2 (ACE2) via the receptor-binding domain (RBD), and thus is believed to be a major target to block viral entry. Both SARS-CoV-2 and SARS-CoV share this mechanism. Here we functionally analyzed the key amino acid residues located within receptor binding motif of RBD that may interact with human ACE2 and available neutralizing antibodies.

The in vivo experiments showed that immunization with either the SARS-CoV RBD or SARS-CoV-2 RBD was able to induce strong clade-specific neutralizing antibodies in mice; however, the cross-neutralizing activity was much weaker, indicating that there are distinct antigenic features in the RBDs of the two viruses. This finding was confirmed with the available neutralizing monoclonal antibodies against SARS-CoV or SARS-CoV-2. It is worth noting that a newly developed SARS-CoV-2 human antibody, HA001, was able to neutralize SARS-CoV-2, but failed to recognize SARS-CoV. Moreover, the potential epitope residues of HA001 were identified as A475 and F486 in the SARS-CoV-2 RBD, representing new binding sites for neutralizing antibodies.

Overall, our study has revealed the presence of different key epitopes between SARS-CoV and SARSCoV-2, which indicates the necessity to develop new prophylactic vaccine and antibody drugs for specific control of the COVID-19 pandemic although the available agents obtained from the SARS-CoV study are unneglectable.

8.6 Evolutionary Models for COVID-19 can be developed

As we have observed, it appears that we can begin to develop evolutionary models for the spike. A serious constraint is the impact of deletions and recombinations in this virus RNA. We have seen them occur especially in the recent variants. Recombination may be more complex in that it could create a dramatically new virus. We have a location, namely centered RNA, and we have some putative a priori types of transitions. This may afford a reasonable starting point for a Bayesian analysis. (see Manrubia and Lazaro and also Woo and Reifman)

8.7 Incremental evolutionary changes in ACE2/COVID-binding affinity can be determined

If we desire a ranking then one must be able to determine a useful metric. It can be argued that binding affinity may be such a metric. As Ou et al have noted:

The current global pandemic of COVID-19 is caused by a novel coronavirus SARS-CoV-2. The SARS-CoV- 2 spike protein receptor-binding domain (RBD) is the critical determinant of viral tropism and infectivity.

To investigate whether naturally occurring mutations in the RBD have altered the receptor binding affinity and infectivity, firstly we analyzed in silico the binding dynamics between mutated SARS-CoV-2 RBDs and the human ACE2 receptor.

Among 1609 genomes of SARS-CoV-2 strains isolated during the early transmission phase, 32 non-synonymous RBD mutants were identified and found clustered into nine mutant types under high positive selection pressure. Applying molecular dynamics simulations, three mutant types (V367F, W436R, N354D/D364Y) displayed higher binding affinity to human ACE2, likely due to the enhanced structural stabilization of the RBD beta-sheet scaffold.

The increased infectivity of one mutant (V367F) circulating worldwide was further validated by performing receptor-ligand binding ELISA, surface plasmon resonance, and pseudotyped virus

assays. Genome phylogenetic analysis of V367F mutants showed that during the early transmission phase, most V367F mutants clustered more closely with the SARS-CoV-2 prototype strain than the dual-mutation variants (V367F + D614G), which emerged later and formed a distinct sub-cluster.

The analysis of critical RBD mutations provides further insights into the evolutionary trajectory of SARS-CoV-2 under high selection pressure and supports the continuing surveillance of spike mutations to aid in the development of COVID-19 drugs and vaccines.

Furthermore, Gan et al demonstrate the measurement of binding affinities and thus this can be readily performed for all putative spike variants. Also the work by Fersht is an extensive set of quantitative metrics which may have some application.

8.8 KILLER T CELLS AND VACCINES

We have introduced the CTLs and their ability to deal with antigens and viral infections. CTL are useful in various cancer immunotherapies. However they can detect a multiplicity of Ag and this have the potential for a broad based immune response that lasts. It is worth following this field as well.

The key issue with CTL is twofold. First, what Ag do they develop upon? Second, how long will they last? Clearly the answer to the first may yield insight into a broader base of vaccine prevention. mRNA vaccines are limited to the spike as currently constructed. Simple, direct, but limited. In contrast the virion may have a multiplicity of Ags detectable via macrophages[27]. Also it seems that CTL may last longer than Abs. Thus it is worth examining this path in some depth.

[27] https://www.researchgate.net/publication/336116071_Tumor_Associated_Immune_Cells_On_the_one_hand_and_on_the_other_hand

9 REFERENCES

1. Abbas et al, Basic Immunology, 4th Ed, Elsevier, 2011
2. Akondy et al, Origin and differentiation of human memory CD8 T cells after vaccination, Nature, 28 Dec 2017
3. Alm et al, Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020, Euro Surveill. 2020 Aug 13; 25(32): 2001410.
4. Al-Samkari et al, COVID-19 and coagulation: bleeding and thrombotic manifestations of SARS-CoV-2 infection, Blood, 23 July 2020, Volume 136, Number 4
5. Annavajhala et al, A Novel SARS-CoV-2 Variant of Concern, B.1.526, Identified in New York, medRxiv preprint doi: https://doi.org/10.1101/2021.02.23.21252259
6. Artika et al, Molecular biology of coronaviruses: current knowledge, Heliyon 6 (2020)
7. Ashoor et al, A computational approach to evaluate the combined effect of SARS-CoV-2 RBD mutations and ACE2 receptor genetic variants on infectivity: The COVID-19 host-pathogen nexus, bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.352344 ; this version posted January 8, 2021
8. Bader et al, Mas and Its Related G Protein–Coupled Receptors, Mrgprs, Pharmacol Rev 66:1080–1105, October 2014
9. Beck and Aksentijevich, Susceptibility to severe COVID-19, Science 23 Oct 2020
10. Campbell and Heyer, Genomics, Proteomics, and Bioinformatics. CSHL Press, 2003
11. Chan and Weitz, COVID-19 coagulopathy, thrombosis, and bleeding, Blood 23 July 2020 | Volume 136, Number 4
12. Chowdhury et al, Immune response in COVID-19: A review, Journal of Infection and Public Health 13 (2020) 1619–1629
13. Devaux et al, ACE2 receptor polymorphism: Susceptibility to SARS-CoV-2, hypertension, multi-organ failure, and COVID-19 disease outcome, Journal of Microbiology, Immunology and Infection (2020) 53, 425e435
14. Domingo, Virus as Populations, Academic Press, 2016
15. Esperanza et al, High Affinity Nanobodies Block SARS-CoV-2 Spike Receptor Binding Domain Interaction with Human Angiotensin Converting Enzyme, bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219857 ; this version posted August 23, 2020
16. Etelvino et al, New Components of the Renin-Angiotensin System: Alamandine and the Mas-Related G Protein-Coupled Receptor D, Curr Hypertens Rep (2014) 16:433
17. Fersht, Structure and Mechanism In Protein Science: A Guide To Enzyme Catalysis And Protein Folding, WSPC; Reprint edition (June 28, 2017)
18. Gan et al, Diversity of ACE2 and its interaction with SARS-CoV-2 receptor binding domain, bioRxiv preprint doi: https://doi.org/10.1101/2020.10.25.354548 ; this version posted November 4, 2020

19. Glasgow et al, Engineered ACE2 receptor traps potently neutralize SARS-CoV-2, PNAS, November 10, 2020, vol. 117, no. 45
20. Gupta et al, Extrapulmonary manifestations of COVID-19, Nature Medicine, 2020
21. Helbert, Immunology, 3rd Ed, Elsevier, 2017
22. Huo et al, Neutralizing nanobodies bind SARS-CoV-2 spike RBD and block interaction with ACE2, Nature Structural and Molecular Biology, 2020
23. Jia et al, ACE2 Receptor Expression and Severe Acute Respiratory Syndrome Coronavirus Infection Depend on Differentiation of Human Airway Epithelia, Journal of Virology, Dec. 2005, p. 14614–14621
24. Kemenesi et al, Effect of An 84-bp Deletion of the Receptor-Binding Domain on the ACE2 Binding Affinity of the SARS-CoV-2 Spike Protein: An In Silico Analysis, Genes 2021, 12, 194
25. Kessel and Ben-Tal, Introduction to Proteins, CRC, 2011
26. Kirby et al, Identification of Contact Residues and Definition of the CAR-Binding Site of Adenovirus Type 5 Fiber Protein, Journal of Virology. Mar. 2000, p. 2804–2813
27. Lan et al, Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor, Nature, Vol 581, 14 May 2020
28. Lasfar and Bouden, A method of data mining using Hidden Markov Models (HMMs) for protein secondary structure prediction, Procedia Computer Science 127 (2018) 42–51
29. Lesk, Bioinformatics, 5th Ed, Oxford, 2011
30. Lesk, Introduction to Protein Science: Architecture, Function, and Genomics, 3rd Ed, Oxford, 2016
31. Lee et al, Ab Initio Protein Structure Prediction, from D.J. Rigden (ed.), From Protein Structure to Function with Bioinformatics, DOI 10.1007/978-94-024-1069-3_1, Springer 2017
32. Li et al, Hydrophobic interaction and hydrogen-bond network for a methane pair in liquid water, PNAS, February 20, 2007, vol. 104, no. 8
33. Li, Structure, Function, and Evolution of Coronavirus Spike Proteins, Annu. Rev. Virol. 2016. 3:237–61
34. Marian, The Discovery of the ACE2 Gene, Circ Res. 2013;112:1307-1309
35. Manrubia and Lázaro, Viral evolution, Physics of Life Reviews 3 (2006) 65–92
36. Nei and Kumar, Molecular Evolution and Phylogenetics, Oxford, 2000
37. Omilusik and Goldrath, The origins of memory T cells, Nature, 28 Dec 2017
38. Ou et al, Emergence of SARS-CoV-2 spike RBD mutants that enhance viral infectivity through increased human ACE2 receptor binding affinity, bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844 ; this version posted September 12, 2020
39. Patrick, Medical Chemistry, 5th Ed, Oxford, 2013

40. Petsko and Ringe, Protein Structure and Function, Sinauer, 2004

41. Prabakaran et al, A model of the ACE2 structure and function as a SARS-CoV receptor, Biochemical and Biophysical Research Communications 314 (2004) 235–241

42. Samavati and Uhal, ACE2, Much More Than Just a Receptor for SARS-COV-2, Frontiers in Cellular and Infection Microbiology, 1 June 2020, Volume 10, Article 317

43. Santos et al, Angiotensin-(1–7) is an endogenous ligand for the G protein-coupled receptor Mas, PNAS, July 8, 2003, vol. 100, no. 14

44. Selmi et al, The long and latent road to autoimmunity, Cellular & Molecular Immunology (2018) 15:543 – 546

45. Senior et al, Improved protein structure prediction using potentials from deep learning, Nature, 30 Jan 2020

46. Serfozo et al, Ang II (Angiotensin II) Conversion to Angiotensin-(1-7) in the Circulation Is POP (Prolyloligopeptidase)-Dependent and ACE2 (Angiotensin-Converting Enzyme 2)-Independent, Hypertension January 2020

47. Simoes e Silva et al, ACE2, angiotensin-(1-7) and Mas receptor axis in inflammation and fibrosis, British Journal of Pharmacology, 2013

48. Sirovetz et al, Protein Structure Prediction: Making AWSEM AWSEM-ER by Adding Evolutionary Restraints, *Proteins: Structure, Function, and Bioinformatics,* doi:10.1002/prot.25367, 2017

49. Sorokina et al, Structural models of human ACE2 variants with SARS-CoV-2 Spike protein for structure-based drug design, Scientific Data | (2020) 7:309

50. Stevens, Medical Chemistry, Pearson, 2014

51. Tarke et al, Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases, Cell Reports Medicine 2, 100204, February 16, 2021

52. Tramontano, Protein Structure Prediction, Wiley-VCH, 2006

53. Woo and Reifman, Quantitative Modeling of Virus Evolutionary Dynamics and Adaptation in Serial Passages Using Empirically Inferred Fitness Landscapes, Journal of Virology p. 1039–1050, anuary 2014 Volume 88 Number 2

54. Vangone et al, Sense and simplicity in HADDOCK scoring: Lessons from CASP-CAPRI round 1, Proteins, 2017; 85;417-423

55. Voit, Systems Biology, Garland, 2013

56. Weigt, et al. Identification of direct residue contacts in protein-protein interaction by message passing, PNAS, Jan 2009.

57. West et al, SARS-CoV-2 lineage B.1.526 emerging in the New York region detected by software utility created to query the spike mutational landscape, bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431043 ; this version posted February 23, 2021

58. Yan et al, Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2, Science, 27 March 2020.

59. Yang et al, Improved protein structure prediction using predicted interresidue orientations, PNAS, 21 Jan 2020, V 117, No 3

60. Yang et al, Predicting or Pretending: Artificial Intelligence for Protein-Ligand Interactions Lack of Sufficiently Large and Unbiased Datasets, Frontiers in Pharmacology | www.frontiersin.org 1 February 2020 | Volume 11 | Article 69

61. Yi et al, Key residues of the receptor binding motif in the spike protein of SARS-CoV-2 that interact with ACE2 and neutralizing antibodies, Cellular & Molecular Immunology (2020) 17:621 – 630

62. Youngblood et al, Effector CD8 T cells dedifferentiate into long-lived memory cells, Nature, 28 Dec 2017

63. Zhou et al, Global pairwise RNA interaction landscapes reveal core features of protein recognition, Nature Communications, (2018) 9:2511

10 INDEX