DeepNEU: a Machine Learning Stem Cell simulation platform for evaluating the impact of Loss of Function and Gain of Function mutations in the SARS-CoV-2 genome

Sally Esmail
123Genetix https://orcid.org/0000-0001-9595-4779
Wayne R Danter (wdanter@123Genetix.com)
123Genetix

Research Article

Keywords: SARS-CoV2 genome, Machine Learning, Stem cell simulations, functional mutations prediction, Viral virulence predictions

Posted Date: December 1st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-116683/v1

DeepNEU: a Machine Learning Stem Cell simulation platform for evaluating the impact of Loss of Function and Gain of Function mutations in the SARS-CoV-2 genome

Sally Esmail, PhD and Wayne R Danter MD

123Genetix, 1595 Dyer Drive,

London, Canada N6G 0T7

RUNNING TITLE

Impact of LOF and GOF mutations in the SARS-CoV-2 genome

ABSTRACT

The global pandemic caused by the SARS-C0V-2 virus continues to spread. Infection with the SARS-CoV-2 causes a disease of variable severity known as COVID-19. It is certain that the viral genome has already mutated and will continue to mutate in unknown directions. These mutations can (1) have no significant impact (they are silent), result in (2) a loss of virulence/function (LOF) or (3) a gain in virulence/function (GOF). Research involving GOF mutations remain especially controversial and highly regulated because accidental release or misuse of a more lethal virus could have catastrophic global effects.

The primary purpose of this project was to evaluate the ability of the DeepNEU machine learning stem-cell simulation platform to enable rapid and efficient assessment of the impact of viral LOF and GOF mutations on SARS-CoV-2 virulence. The data generated from this project confirm that (1) SARS-CoV-2 infection can be simulated in human alveolar type lung cells, (2) these simulated infected lung cells can be used to assess the impact of LOF and GOF mutations in a viral genome, (3) a new and simple four-factor virulence measure, the DeepNEU Case Fatality Rate (dnCFR) based on NSP3, Spike-RDB, N protein and M protein can be used to assess the impact of LOF and GOF mutations and (4) the platform combined with the dnCFR measure successfully identified specific and potentially beneficial mutations (LOF) as well as deleterious mutations (GOF) that potentially increase the virulence of the SARS-CoV-2 virus. We conclude that the DeepNEU platform and the dnCFR measure should be urgently developed, further validated and applied to SARS-CoV-2 and other viral pathogens as important tools for future pandemic preparedness.

INTRODUCTION

The continuing evolution of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome remains a major obstacle to developing effective antiviral and vaccine therapies (1,2). The potential to better understand and predict this evolution will assist in the early detection of drug-resistant strains and facilitate the development of effective antiviral drugs and vaccines (3).

One important focus in the field of virology is to develop a better understanding of the impact of genetic mutation on virulence (4,5). Predicting viral evolution is a fundamental goal in virology and this is especially true for pathogenic viruses (6). Genomic information about current viral pathogens like SARS-CoV-2 and their continued evolution will provide a better understanding of the dynamics of future virus evolution and shed light on effective strategies to contain future outbreaks (7,8).

Even though RNA viruses have a limited genome and a relatively limited evolutionary capacity, they do present unique challenges to predicting the impact of evolutionary changes. RNA viruses are known for their high mutation rates (around 1 mutation in 1,000 bases) and frequent recombination that can produce novel genotypes from co-circulating strains (6). RNA viruses additionally undergo frequent mutations as they circulate in the population as in response to host factors. The feasibility of predicting viral evolution relies upon on the breadth and scale of well posed questions and calls for cautious optimism (9).

Machine learning-based predictions of the impact of genetic mutations has been efficiently utilized in the field of viral genetics for many years, and have been mostly focused on the prediction of viral mutations that are associated with drug resistance (9). Given the current global SARS-CoV-2 pandemic and the lack of effective anti-viral therapies and vaccines, it is highly desirable to have a fast, reliable and efficient machine learning platform for the prediction of the viable mutations and study their potential effects on the virulence of SARS-CoV-2 as well as future viral pathogens for which we are not prepared.

An important goal of this study is to simulate the natural evolution of the post-pandemic strain of SARS-CoV-2 by systematically introducing both gain of function (GOF) and loss of function mutations (LOF) into the viral genome. The overarching objective of this study is to identify

therapeutic targets and improve preparedness for future epidemic/pandemic outbreaks of new strains of SARS-CoV-2 and other viral pathogens. In this study we predict the impact of novel GOF and LOF mutations in the SARS-CoV-2 genome and we have developed a case fatality rate (CFR) based measure for estimating changes in viral virulence. Our literature validated Deep machine learning platform, DeepNEU v5.0, has successfully identified the virulence potential of SARS-CoV-2 mutations well ahead of when they could occur and be identified in nature. These discoveries will offer the possibility of improving viral pandemic preparedness and better targeting surveillance between and during epidemics/pandemics.

METHODS

The DeepNEU stem cell simulation platform is a literature validated hybrid deep-machine learning system with elements of fully connected recurrent neural networks (RNN), cognitive maps (CM), support vector machines (SVM) and evolutionary systems (GA). The detailed methodology for simulation development and validation plus the description of the current database (DeepNEU v5.0) used in these experiments has been described elsewhere in detail (10-12).

The DeepNEU simulations

The main goal of this project was to extend our previous DeepNEU based research into SARS-CoV-2 infection by evaluating the potential impact of simulated LOF and GOF mutations in the viral genome on virulence. As described previously (12) we first created computer simulations (aiPSC) of human induced pluripotent stem cells (iPSC) and lung (aiLUNG) cells. Once validated, the aiLUNG simulations were exposed to simulated SARS-CoV-2 viremia by turning on extracellular Spike-RBP (Receptor Binding Domain) in the presence of active Transmembrane Serine Protease 2 (TMPRSS2) (12). The simulated SARS-CoV-2 infection of AT1 and AT2 lung cells (aiLUNG-COVID-19) was confirmed using a profile of genotypic and phenotypic features from the published literature (12). Finally, the validated aiLUNG and aiLUNG-COVID-19 simulations were used to evaluate an inclusive set of factors derived from the published SARS-CoV-2 genome (Accession number: NC_045512.2; https://www.ncbi.nlm.nih.gov/sars-cov-2/) regarding their ability to affect an increase or decrease in virulence. A summary of the fifteen SARS-CoV-2 gene/gene products evaluated in the current experiments are presented in Table 1.

Table 1: Summary of evaluated LOF and GOF mutations in the SARS-CoV-2 genome (N = 15 x 2)

SARS-CoV-2 Target	Loss of Function	Gain of Function
aiPSC-WT	**N/A**	**N/A**
aiLUNG (i.e. Uninfected)	**N/A**	**N/A**
aiLUNG + SARS-CoV-2	**N/A**	**N/A**
Spike-RBD Mutation	**-1, Locked OFF**	**+1, Locked ON**
Furin Mutation	**-1, Locked OFF**	**+1, Locked ON**
NSP12 Mutation (RdRP)	**-1, Locked OFF**	**+1, Locked ON**
orf1ab Mutation	**-1, Locked OFF**	**+1, Locked ON**
orf10 Mutation	**-1, Locked OFF**	**+1, Locked ON**
(N)ucleoprotein Mutation	**-1, Locked OFF**	**+1, Locked ON**
(M)embrane Mutation	**-1, Locked OFF**	**+1, Locked ON**
NSP3 Mutation	**-1, Locked OFF**	**+1, Locked ON**
orf7a Mutation	**-1, Locked OFF**	**+1, Locked ON**
orf8 Mutation	**-1, Locked OFF**	**+1, Locked ON**
NSP5 Mutation	**-1, Locked OFF**	**+1, Locked ON**
(S)pike Mutation	**-1, Locked OFF**	**+1, Locked ON**
(E)nvelope Mutation	**-1, Locked OFF**	**+1, Locked ON**
NSP13 Mutation (Helicase)	**-1, Locked OFF**	**+1, Locked ON**
orf3a Mutation	**-1, Locked OFF**	**+1, Locked ON**

NSP = Non-Structural Protein, RdRP = RNA-dependent RNA polymerase, orf = open reading frame, TMPRSS2 = Transmembrane protease, serine 2, aiLUNG = Wild Type (uninfected), aiLUNG + SARS-CoV-2 = aiLUNG-COVID-19

Prior to the application of simulated LOF and GOF mutations as described above, the predictions from the wild type aiPSC, aiLUNG-WT and aiLUNG-COVID-19 simulations regarding the expression or repression of genes and proteins and presence or absence of phenotypic features were validated directly against published data as outlined previously (12). All experiments in this

study were conducted in triplicate (n=3) using different initial conditions in the form of initial state vectors.

DeepNEU platform statistical analysis

As outlined previously the statistical analysis of aiPSC, aiLUNG and LUNG-COVID-19 simulation predictions versus the published literature used the unbiased binomial test. This test provides an exact probability, can compensate for prediction bias and is ideal for determining the statistical significance of experimental deviations from an actual distribution of observations that fall into two outcome categories (*e.g.*, agree vs disagree). A p-value <0.05 is considered significant and is interpreted to show that the observed relationship between simulated and actual unseen wet lab data is unlikely to have occurred by chance alone. The pre-test probability of a positive outcome prediction is 0.661 and the pre-test probability of a negative prediction is therefore 0.339. This system bias was used when applying the binomial test to all simulation outcomes. For other between group (e.g. LOF vs GOF) comparisons, the Mann-Whitney u test of significance was used (13). This nonparametric test was chosen because some of the data was not normally distributed.

For the purpose of this project it was necessary to create a simple method for estimating the impact of LOF and GOF mutations on virulence based on a recent paper by (14). These authors identified four gene products that were impacted in almost all the recognized mutations identified in the SARS-CoV-2 genome. These four gene products were Polyprotein-1ab (orf-1ab), Nucleocapsid protein (N), Spike protein (S) and Membrane protein (M). Refinement of the impact of these four proteins revealed that the non-structural protein cleavage products NSP3, NSP4 and NSP14 were largely responsible for mutations seen in the orf-1ab polyprotein, while the S-RBD protein appeared responsible for the majority of variation in the Spike protein (14). We therefore created a high case fatality rate measure (dnCFR) using DeepNEU estimates of NSP3 derived from orf-1ab polyprotein (NSP4 and NSP14 are not implemented in DeepNEU (v5.0)), Nucleocapsid protein (N), Receptor Binding Domain (S-RBD) and Membrane protein (M). The dnCFR measure was used to compare all LOF and GOF mutations. In addition, the dnCFR measure was correlated with the calculated Angular Cosine Distance (ACD) a validated metric for evaluating the distance between real valued vectors with values between -1 and +1(15,16).

Validation of the dnCFR measure included calculation of Cosine Similarity (CS) for all LOF and GOF mutations to establish the similarity to the wild type SARS-CoV-2 genome. Cosine Similarity is a commonly used measure for comparing the similarity of two or more real valued vectors with the same number of elements. In this study each SARS-CoV-2 genomic profile was represented as real valued vectors. As similarity between the genomic profile vectors increases, CS increases to +1 or maximum similarity. As CS similarity decreases away from the reference vector and becomes increasingly dissimilar, CS decrease towards -1 or maximum dissimilarity (7). We then used a simple mathematical transformation to derive Angular Cosine Distance (ACD) using the formula ACD = arcosine(CS)/Pi. The ACD metric was selected to evaluate the distance between wild type and mutated SARS-CoV-2 genomic vectors because (1) it conforms with all four properties of a valid distance metric, (2) sample sizes are relatively small (N<20) minimizing any influence of the curse of dimensionality and (3) it is a widely used and well validated metric for comparing bounded real valued (-1 to +1) vectors (15,16).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

RESULTS

The aiPSC and wild type (uninfected) aiLUNG simulations

As reported previously both the unsupervised aiPSC simulations and the unsupervised aiLUNG simulations converged quickly (24 iterations) to a new system wide steady state without evidence of overtraining after 1000 iterations (12). The aiPSC simulations expressed the same human hESC specific surface antigen and genomic profile as both undifferentiated human embryonic stem cells (hESC) and induced pluripotent stem cells (iPSC) (12). The probability that all (N=15) of these aiPSC-WT outcomes were correctly predicted by chance alone using the binomial test is 0.0021.

The aiLUNG simulations produced a similar genotypic and phenotypic expression profile when compared with the human wild type (ATI and ATII) lung cell specific factors taken from the literature (12). The probability that all (N=15) of these aiLUNG outcomes were correctly predicted by chance alone using the binomial test is 0.0021. Importantly, the data also indicate that the

generation of aiLUNG cells from aiPSC produces a heterogenous population of alveolar cell precursors and more mature alveolar cells consistent with previous study (17).

Simulation of SARS-CoV-2-infected aiLUNG cells (aiLUNG-COVID-19)

The next step in the experiments was to expose the aiLUNG cells to simulated SARS-CoV-2 virus. For this simulated infection the concept of SARS-CoV-2 viremia was activated (turned on). The viremia activates the viral life cycle beginning with the interaction of the viral Spike protein with its receptor protein Angiotensin-converting enzyme 2 (ACE2) and ending with exocytosis of new viral particles which completes the cycle by contributing new viral particles to the ongoing viremia (18). The SARS-CoV-2 genome consists of four structural genes, at least six non-structural genes and produces at least ten proteins. As described previously, the seventeen gene/protein expression profile was compared with the uninfected aiLUNG simulations to assess the validity of simulated COVID-19. All genes and proteins studied were expressed in the aiLUNG-COVID-19, but not aiLUNG simulations. The probability that all (N=17) of these aiLUNG-COVID-19 simulation outcomes were correctly predicted by chance alone using the binomial test is 0.0009 (12).

A phenotypic profile of aiLUNG-COVID-19 was also developed based on the published literature and has been described previously (12). These phenotypic features (N=8) include: New Extracellular Virus release, Spike-ACE2 Interface, Spike-RBD, TMPRSS2, Virus Clearance, Virus Intracellular RNA release, Virus Internalization and Virus Replication. The presence of all phenotypic features of COVID-19 was correctly predicted by the aiLUNG-COVID-19 simulations when compared with the aiLUNG simulations. The probability that all (N=8) of these aiLUNG-COVID-19 outcomes were predicted correctly by chance alone using the binomial test is 0.0364.

When we combined the genotypic and phenotypic profiles, the probability that all (N=25) features of simulated aiLUNG-COVID-19 were accurately predicted by chance alone using the binomial test is 0.00003.Evaluation of the validated aiLUNG-COVID-19 simulations for estimating the impact of LOF and GOF mutations on SARS-CoV-2 virulence.

LOF mutations:

The LOF mutations (N=15), representing fifteen genes and proteins of the SARS-CoV-2 genome listed above, were simulated by setting the gene/gene product concepts to -1 and locking them off during system development. This is the computational analogue to creating a gene deletion and therefore an absent gene product that is propagated from each iteration to the next until system convergence is achieved. All unsupervised aiLUNG-COVID-19 and aiLUNG-COVID-19 with LOF simulations converged quickly to a new system wide steady state without evidence of overtraining after 1000 iterations.

The dnCFR measure for the LOF mutations ranged from a value of -4.000 for the aiLUNG-WT (uninfected) simulations to a maximum of 1.365 (95% CI±0.521) with a theoretical maximum of +4.000. The mean dnCFR for LOF mutations was (0.152±0.521). Analysis of system outputs using the dnCFR identified eight LOF mutations that significantly decreased SARS-CoV-2 virulence ($p<0.05$) when compared with SARS-CoV-2 genome without mutations. The most impactful LOF mutation was in S-RBD (-1.812 ±0.521) followed closely by LOF mutation in the S protein cleavage enzyme Furin (-1.553±0.521). The remainder of the LOF mutations did not significantly alter virulence as estimated by the dnCFR ($p>0.05$)(Fig.1A and Fig. 4).

GOF mutations:

The GOF mutations (N=15) were simulated by setting the gene/gene product concepts to +1 and locking them on during system development. This is the computational analogue to creating a maximum increase in gene function and therefore a maximum gene product that is propagated from each iteration to the next until system convergence is achieved. All unsupervised aiLUNG-COVID-19 and aiLUNG-COVID-19 with GOF simulations converged quickly to a new system wide steady state without evidence of overtraining after 1000 iterations.

The dnCFR measure for the GOF mutations ranged from a value of -4.000 for the aiLUNG-WT simulations to a maximum of 2.156 (95% CI±0.131) with a theoretical maximum of +4.000. The mean dnCFR for LOF mutations was (1.652±0.131). Analysis of system outputs using the dnCFR measure identified six GOF mutations that significantly increased virulence ($p<0.05$) when compared with aiLUNG-COVID-19 without mutations. The most impactful GOF mutation was in the N protein (2.156 ±0.131) followed closely by GOF mutation in the M protein (2.063±0.131). The remainder of the GOF mutations did not significantly alter virulence ($p>0.05$) (Fig. 1B and Fig. 5).

Using the dnCFR measure to compare LOF and GOF mutations

The first step in assessing the dnCFR measure was to evaluate the ability of each of the four individual genomic components of the measure (NSP3, S-RBP, M, N proteins) to distinguish LOF from GOF mutations. Based on the 2 tailed Mann-Whitney u test, each component of the dnCFR measure easily distinguished LOF and GOF mutations from each other (all exact p values $\leq$ 0.000044) (Fig.2 and Fig. 6).

Next, we explored the ability of the dnCFR measure to distinguish between LOF and GOF mutations. The mean dnCFR (±95%CI) for LOF mutations was 0.152±0.521 and 1.652±0.131 for GOF mutations. The exact Mann-Whitney u test p value for the direct comparison was 1.55E-07 suggesting that the dnCFR measure was better at distinguishing between LOF and GOF mutations than any single element of the measure (Fig. 2A and Fig. 6).

Finally, we wished to compare the dnCFR measure with a widely used and validated distance metric, the Angular Cosine Distance (ACD). The ACD was used to estimate the distance between genomic profiles. As we described previously, we first calculated the Cosine Similarity (CS) measure for each LOF and GOF mutation profile. This value was then converted to the ACD using the equation ACD = arcosine(ACD)/Pi. ACD values were calculated for each LOF and GOF mutation. Similar to the dnCFR measure, the ACD metric also easily distinguished between the LOF and GOF mutations (Mann-Whitney u test p=3.87E-07). We next compared the ACD metric directly with the dnCFR measure using the Spearman rank correlation coefficient. The Spearman coefficient for the LOF mutations was 0.973 (critical value (N=15) is 0.779, $p<0.001$) (Fig. 3A). The Spearman coefficient for the GOF mutations was 0.903 (critical value (N=15) is 0.779, $p<0.001$) (Fig. 3B). These data indicate that the dnCFR measure and ACD metric are both able to accurately distinguish between LOF and GOF mutations and are strongly positively correlated with each other.

DISCUSSION

Recently, we evaluated the capability of the DeepNEU (v5.0) machine learning platform to simulate SARS-CoV-2 infection in simulated Type 1 (AT1) and Type 2 (AT2) alveolar lung cells (aiLUNG-COVID-19) (12). In our most recent research, we reported the ability of the DeepNEU

platform to enable the rapid identification of therapeutic targets and drug repurposing specifically for treating both the cytokine storm and coagulopathy that frequently complicate severe COVID-19 (manuscript under review). While we have used the same approach that we reported in the previous two papers ((12), manuscript under review), the primary purpose of this project was to extend our previous work by evaluating the ability of the DeepNEU platform to enable rapid and efficient assessment of the impact of LOF and GOF mutations in the SARS-CoV-2 genome. As of this writing and with a few exceptions, the diversity and impact of known mutations in the SARS-CoV-2 genome are relatively unknown and this is particularly true for GOF mutations that have increased potential to amplify SARS-CoV-2 virulence.

While there is some variation in the definition of virulence, we have defined it here to mean the severity or harmfulness of a disease. Specifically, regarding SARS-CoV-2 this means the proportion of people who die from COVID-19 which is also known as the case fatality rate (CFR). For the purposes of this research we have created a new measure called dnCFR. This new measure represents a logical extension of the insights into mutations in the SARS-CoV-2 genome provided in (14). These authors identified four gene products (proteins) that result from the most common mutations in the SARS-CoV-2 genome. These mutated proteins are (1) Non-Structural Protein 3 (NSP3), (2) Spike-Receptor Binding Domain (S-RBD), (3) Membrane (M) protein and (4) Nucleocapsid (N) protein. All four of these proteins were implemented in DeepNEU (v5.0) and could be evaluated individually and in combination to constitute the dnCFR measure. Individually each of the four mutated proteins could easily distinguish LOF from GOF mutations ($p<4.4E-05$). Importantly, when combined the dnCFR measure appeared to be about 75 times better at distinguishing between LOF and GOF mutations ($p=1.55E-07$). The dnCFR measure also performed well when compared with a validated and widely used metric, the Angular Cosine Distance (ACD) which we used to measure the distance between two real valued genomic vectors. Specifically, we wanted to measure the distance between the un-mutated SARS-CoV-2 genome and the mutated genomes using the ACD. The calculated Spearman correlation coefficient was greater than 0.900 ($p<0.001$) for all comparisons supporting the existence of a strong positive correlation between the two.

The dnCFR measure has a few important advantages namely its simplicity and ease of use. It is easily calculated by adding four values whereas the ACD calculation requires a two-step calculation involving a trigonometric transformation. The dnCFR can also be directly linked to the global, regional or local CFR estimate. As of this writing the number of global SARS-CoV-2 infections is 22,200,000 with 783,000 deaths producing a CFR of 0.035. This global CFR of 3.5% or 35,270 deaths per million is associated with a dnCFR estimate of 1.457 for the virulence of the wild type SARS-CoV-2 genome. For example, a GOF mutation that produces a dnCFR of 2 or a 1.373 increase in virulence would result in a CFR of 0.048 or 48,044 per million people infected or 12,774 excess deaths per million. Perhaps most importantly the dnCFR measure can be modified for any viral genome for which there are validated insights into the genome mutational landscape. While the dnCFR is based on sound logic and mathematical principles there are a few drawbacks. These potential drawbacks of the current version of the dnCFR measure are related to its newness and lack of more robust validation and widespread use.

dnCFR and LOF mutations

When we applied the dnCFR measure to each of the fifteen LOF mutations evaluated, eight significant mutations were identified. All these mutations produced a significant decrease in SARS-CoV-2 virulence compared with the wild type genome. The most significant LOF mutation was in S-RBD with a dnCFR of -1.812±0.521 representing a 224.37% decrease in virulence. This LOF mutation produces a loss of virulence as evidenced by a CFR of 0.00% or 0.00 per million people infected and a decrease of 35,270 deaths per million. There are three other LOF mutations that produce a negative dnCFR. These other mutated proteins are Furin (-1.553±0.521), NSP12/RdRP (-1.124±0.521) and orf1ab (-0.846±0.521) suggesting that SARS-CoV-2 virulence is dependent on each of these mutations. The least, but still significant LOF mutation was in the NSP3 protein with a dnCFR of 0.321±0.521 representing a 17.9% decrease in virulence. This LOF mutation produces a decrease in virulence associated with a CFR of 2.90% or 28,949 per million people infected and an expected decrease of 6,322 deaths per million.

These data have important implications for future research focusing on SARS-CoV-2 pandemic preparedness. Importantly, the rapid identification of drug and drug combinations, monoclonal antibodies or vaccines that target one or more of these LOF mutations would be expected to

produce a LOF or LOFs situation that could reduce CFR by a minimum of 6,322 deaths per million infections.

dnCFR and GOF mutations

When the dnCFR measure was applied to the each of the fifteen GOF mutations evaluated, six significant mutations were identified. All these mutations produced a significant increase in SARS-CoV-2 virulence compared with the wild type genome. The most significant GOF mutation was in the N protein with a dnCFR of 2.156±0.131 representing a 47.98% increase in virulence. This GOF mutation produces an increase in virulence associated with a CFR of 5.18% or 52,191 deaths per million people infected and an increase of 16,921 deaths per million. The second most significant GOF mutation was in the M protein with a dnCFR of 2.063±0.131. The least, but still significant GOF mutation was in the NSP3 protein with a dnCFR of 1.743±0.131 representing a 19.63% increase in virulence. This GOF mutation produces an increase in virulence associated with a CFR of 42,732 deaths per million people infected and an expected increase of 7,012 deaths per million.

These data also have important implications for future research focusing on SARS-CoV-2 pandemic preparedness. Importantly, the rapid identification of drug and drug combinations, monoclonal antibodies or vaccines that target one or more of these GOF mutations would be expected to produce a GOF or GOFs situation that could reduce CFR by at least 7,012 and perhaps as much as 17,000 deaths per million infections. Early application of this literature validated technology could have even greater beneficial effects on future viral pandemics for which we remain unprepared.

Interestingly, when we combined all mutations, six of the mutated proteins, both LOF and GOF, significantly impacted the virulence of SARS-CoV-2 genome as estimated by the dnCFR. These six proteins were N, M, S-RBD, Orf1ab, Orf10 and NSP3.

Evolution of the SARS-CoV-2 genome so far

The evolution of the SARS-CoV-2 genome is ongoing and so far, it appears to have evolved into at least six clades defined based on a common ancestor. These currently identified clades are labelled as G, GH. GR, S, V and L plus an O clade representing Other. These clades have different geographic representation as well as mutational profiles. Worldwide the most common clades are G, GH and GR accounting for ~74% of identified mutations. Importantly, clades GH and GR are believed to be derived from the G clade. The G glade has NSP3, RdRP (NSP12) and Spike (S) mutations. The GH clade has the same mutations as G plus an ORF3a mutation and similarly, the GR clade has the same mutations as G plus a Nucleocapsid (N) mutation (19). All these individual mutations have been evaluated by the DeepNEU platform.

Beginning with Africa, the most common clade is G followed by GH, GR, and O. In Asia the largest clade is O followed by GH, S, GR and G. The most common European clade is GR followed G, V and GH. In North America the dominant clade is GH followed by S and G. The most common clade in South America is GR followed by GH and G. Finally, in Oceania GH is the most common clade followed by O, G, V and GR. The G, GH and GR clades are variably but substantially represented in all regions of the globe discussed (19-24).

The DeepNEU platform can be used to assess the impact of regionally specific SARS-CoV-2 clades by combining LOF and/or GOF mutations. For example, globally the most common clade is GR (15) and the worse-case scenario can be simulated by combining NSP3 + RdRP (NSP12) + Spike (S) + Nucleocapsid (N) GOF mutations. Of note, the current version of DeepNEU could easily handle an almost unlimited number of LOF and/or GOF mutations. The cumulative mutational impact on the GR clade CFR can be estimated by the dnCFR as outlined above. For example, GOF mutations of all four proteins of the GR clade would result in a dnCFR of 3.052 which equates to a worst case, CFR of 7.333% or 73,891 deaths per million and an expected increase in CFR of 38,621 deaths per million. Given that the average number of mutations in the SARS-CoV-2 genome so far has been >7, this scenario is unlikely but not impossible.

<u>Future viral pandemic preparedness: We are not prepared!</u>

Finally, we must act now and fortunately we can begin with the WHO list of top 10 pathogens for which we are not now prepared. Importantly, all the pathogens on this list have been recognized longer than SARS-CoV-2(2019) has. For example, the Rift Valley fever virus was officially recognized in 1931, the Zika virus was recognized in 1947, Crimean-Congo Fever in 1967, Lassa Fever in 1969, Ebola in 1976, Nipah virus in 1998 and the MERS virus in 2012. Although none of these pathogens have effective therapies, all of them have a considerable body of knowledge regarding their genome and changes over time. So far, this is the only absolute requirement for implementing other analyses like that for SARS-CoV-2. In other words, an approach that combines the DeepNEU platform and a dnCFR measure modified for a specific viral genome can be used with the other members of the WHO list of top 10 pathogens for which, midway through 2020, we are not prepared

ACKNOWLEDGMENT

The authors wish to thank Dr. Mark Poznansky for his constructive review and suggested edits that improved the final manuscript. Dr. Poznansky is the past president and CEO of the Ontario Genomics Institute (OGI) and was previously chair of the board of OGI. He is a member of the Order of Canada, a member of the Order of Ontario and was CEO, president and scientific director of Robarts Research Institute at the University of Western Ontario. He is also an accomplished author and creator of the successful blog “Saved by Science”.

AUTHOR CONTRIBUTIONS

SE and WD conceptualized, and analyzed the experimental work, wrote the manuscript, and prepared the figures. WD performed all computational simulations of COVID-19 disease and SARS-CoV2 GOF and LOF mutations modeling.

COMPETING INTERESTS

SE and WRD have uncompensated relationship with 123Genetix medical enterprise, nonetheless the authors declare that they are providing an unbiased scientific article.

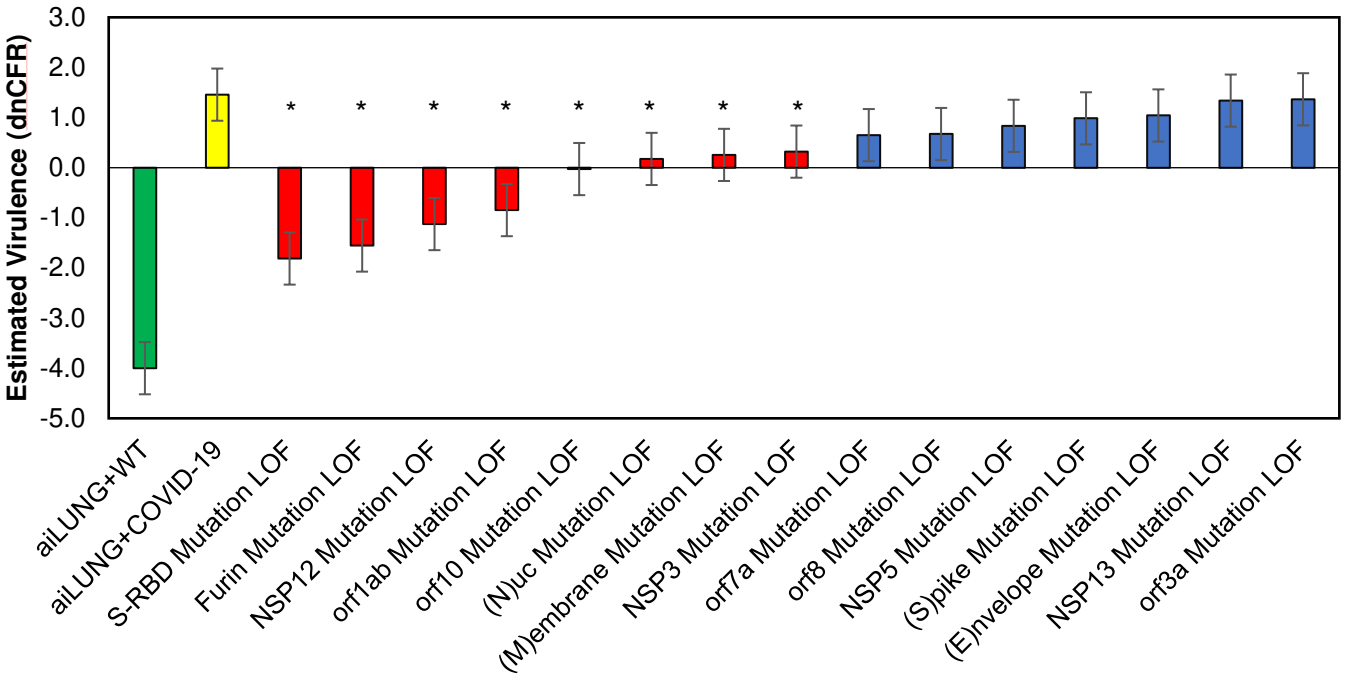
REFERENCES

1. Acter, T., Uddin, N., Das, J., Akhter, A., Choudhury, T. R., and Kim, S. (2020) Evolution of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) as coronavirus disease 2019 (COVID-19) pandemic: A global health emergency. *Science of the Total Environment*, 138996
2. Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., Masciovecchio, C., Angeletti, S., Ciccozzi, M., and Gallo, R. C. (2020) Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of Translational Medicine* **18**, 1-9
3. Liu, C., Zhou, Q., Li, Y., Garner, L. V., Watkins, S. P., Carter, L. J., Smoot, J., Gregg, A. C., Daniels, A. D., and Jervey, S. (2020) Research and development on therapeutic agents and vaccines for COVID-19 and related human coronavirus diseases. ACS Publications
4. Sanjuán, R., and Domingo-Calap, P. (2016) Mechanisms of viral mutation. *Cellular and molecular life sciences* **73**, 4433-4448
5. Geoghegan, J. L., and Holmes, E. C. (2018) The phylogenomics of evolving virus virulence. *Nature Reviews Genetics* **19**, 756-769
6. Dolan, P. T., Whitfield, Z. J., and Andino, R. (2018) Mapping the evolutionary potential of RNA viruses. *Cell host & microbe* **23**, 435-446
7. Li, H., Liu, S.-M., Yu, X.-H., Tang, S.-L., and Tang, C.-K. (2020) Coronavirus disease 2019 (COVID-19): current status and future perspective. *International journal of antimicrobial agents*, 105951
8. Zheng, J. (2020) SARS-CoV-2: an emerging coronavirus that causes a global threat. *International journal of biological sciences* **16**, 1678
9. Li, J., Zhang, S., Li, B., Hu, Y., Kang, X.-P., Wu, X.-Y., Huang, M.-T., Li, Y.-C., Zhao, Z.-P., and Qin, C.-F. (2020) Machine Learning Methods for Predicting Human-Adaptive Influenza A Viruses Based on Viral Nucleotide Compositions. *Molecular biology and evolution* **37**, 1224-1236
10. Esmail, S., and Danter, W. R. (2019) DeepNEU: artificially induced stem cell (aiPSC) and differentiated skeletal muscle cell (aiSkMC) simulations of infantile onset POMPE disease (IOPD) for potential biomarker identification and drug discovery. *Frontiers in cell and developmental biology* **7**, 325
11. Danter, W. R. (2019) DeepNEU: cellular reprogramming comes of age–a machine learning platform with application to rare diseases research. *Orphanet Journal of Rare Diseases* **14**, 13
12. Esmail, S., Danter, R Wayne. (2020) Viral Pandemic Preparedness: a pluripotent stem cell-based Machine Learning platform for simulating COVID-19 infection to enable Drug Discovery and Repurposing. *Stem Cells Translational Medicine* **In press**
13. Addinsoft. (2019) XLSTAT statistical and data analysis solution. Addinsoft Long Island, New York
14. Gussow, A. B., Auslander, N., Faure, G., Wolf, Y. I., Zhang, F., and Koonin, E. V. (2020) Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. *Proceedings of the National Academy of Sciences*
15. Chanwimalueang, T., and Mandic, D. P. (2017) Cosine similarity entropy: Self-correlation-based complexity analysis of dynamical systems. *Entropy* **19**, 652
16. Cai, S., Georgakilas, G. K., Johnson, J. L., and Vahedi, G. (2018) A cosine similarity-based method to infer variability of chromatin accessibility at the single-cell level. *Frontiers in genetics* **9**, 319
17. Li, C., Smith, S. M., Peinado, N., Gao, F., Li, W., Lee, M. K., Zhou, B., Bellusci, S., Pryhuber, G. S., and Ho, H.-Y. H. (2020) WNT5a-ROR Signaling Is Essential for Alveologenesis. *Cells* **9**, 384
18. Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., and Siddique, R. (2020) COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. *Journal of Advanced Research*

19. Mercatelli, D., and Giorgi, F. M. (2020) Geographic and Genomic Distribution of SARS-CoV-2 Mutations.
20. Kim, J.-S., Jang, J.-H., Kim, J.-M., Chung, Y.-S., Yoo, C.-K., and Han, M.-G. (2020) Genome-Wide Identification and Characterization of Point Mutations in the SARS-CoV-2 Genome. *Osong Public Health and Research Perspectives* **11**, 101
21. Laha, S., Chakraborty, J., Das, S., Manna, S. K., Biswas, S., and Chatterjee, R. (2020) Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission. *Infection, Genetics and Evolution* **85**, 104445
22. Zhao, Z., Li, H., Wu, X., Zhong, Y., Zhang, K., Zhang, Y.-P., Boerwinkle, E., and Fu, Y.-X. (2004) Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC evolutionary biology* **4**, 21
23. Nasir Abdullahi, I., Uchenna Emeribe, A., Abimbola Ajayi, O., Soji Oderinde, B., Ohinoyi Amadu, D., and Iherue Osuji, A. (2020) Implications of SARS-CoV-2 genetic diversity and mutations on pathogenicity of the COVID-19 and biomedical interventions.
24. Koyama, T., Platt, D., and Parida, L. (2020) Variant analysis of SARS-CoV-2 genomes. *Bulletin of the World Health Organization* **98**, 495

FIGURES

A



B

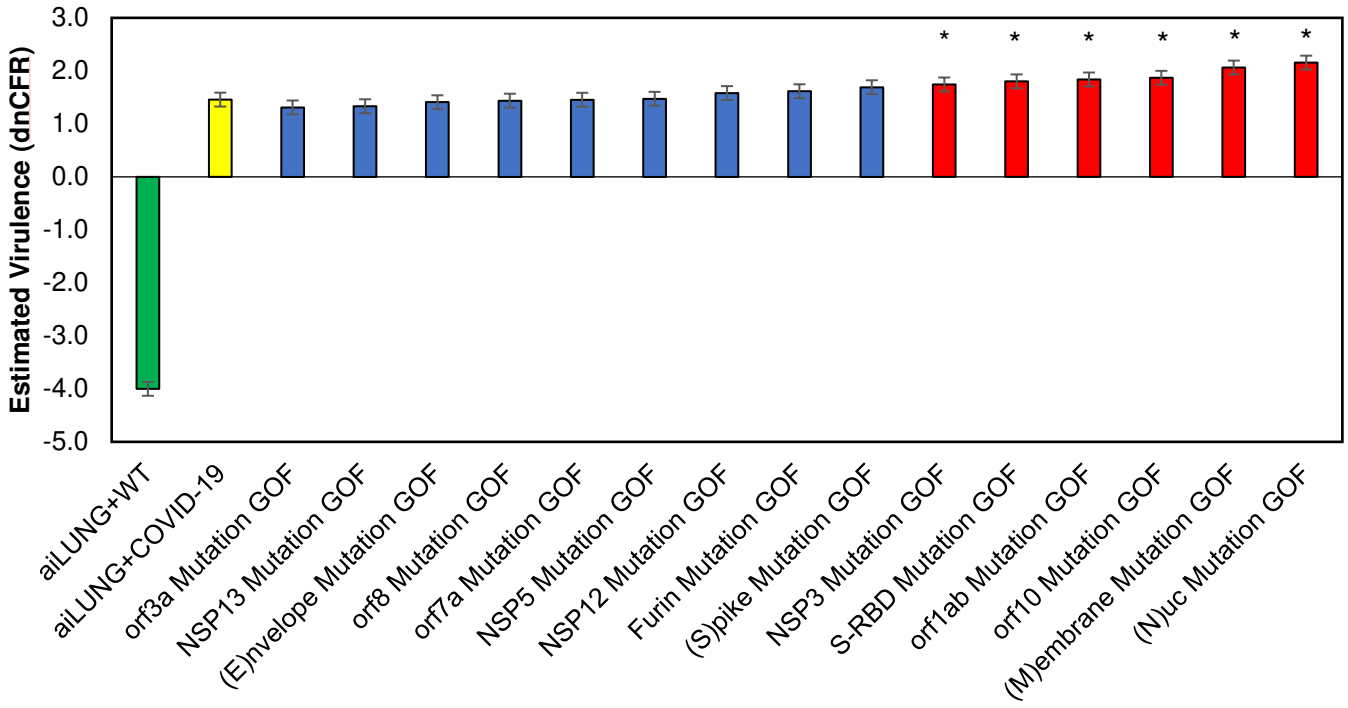


Fig. 1: **aiLUNG simulations of the effect of LOF and GOF mutations in the SARS-CoV-2 genome on viral virulence**. A, LOF simulations and their virulence estimation. Red bars represent mutations in SARS-CoV2 that result in a significant decrease in virulence. B, GOF simulations their virulence estimation. Red bars represent mutations in SARS-CoV2 that result in a significant increase in virulence. A, B, Green bar represents aiLUNG-WT; Yellow bar represents aiLUNG-COVID-19 (unmutated original SARS-CoV2 genome) simulations; blue bars represent mutations that has no significant effect on virulence. These data represent the average from 3 experiments ± the 95% confidence interval around the average.

A

Factor	ACE2	RdRP	NSP5	E	NSP13	M	N	NSP1	NSP2	NSP3	Orf10	Orf1ab	Orf3a	Orf6	Orf7a	Orf8	Plpro	Spike	S-RBD
Average LOF	-0.305	0.208	-0.076	-0.056	-0.187	-0.056	-0.056	-0.036	-0.044	0.074	0.025	-0.030	0.025	0.110	0.035	0.036	-0.075	-0.056	0.190
p-value*	>0.05 NS	<0.001	<0.01	<0.001	<0.01	<0.001	<0.001	<0.01	<0.01	<0.001	<0.001	>0.05 NS	<0.001	<0.001	<0.001	<0.001	<0.01	<0.001	<0.001
Average GOF	-0.479	0.811	0.158	0.272	0.168	0.272	0.271	0.081	0.177	0.532	0.437	0.197	0.437	0.403	0.510	0.510	0.158	0.271	0.577

B

Mutation	LOF	±95% CI	GOF	±95% CI	p value*
SARS-CoV-2/M	-0.056	0.155	0.272	0.093	3.4322E-05
SARS-CoV-2/N	-0.056	0.155	0.271	0.093	4.4044E-05
SARS-CoV-2/NSP3	0.074	0.231	0.532	0.057	1.555E-05
SARS-CoV-2/S-RBD	0.190	0.217	0.577	0.063	1.2507E-06
dnCFR measure	0.152	0.521	1.652	0.131	1.55E-07

** 2 tailed, Mann-Whitney u test*

Fig.2: DeepNEU summary analysis summary and statistics

A, DeepNEU summary results comparing the impact of LOF and GOF mutations on SARS-CoV-2 individual genotypic features. Data from 3 separate experiments. E = Envelope, M = Membrane, N = Nucleocapsid, S-RBD = Spike-Receptor Binding Domain. B, Summary results from DeepNEU simulations of the individual mutated components and composite dnCFR measure effects on SAR-CoV-2 virulence. Results are the average of 3 experiments ± the 95% CI

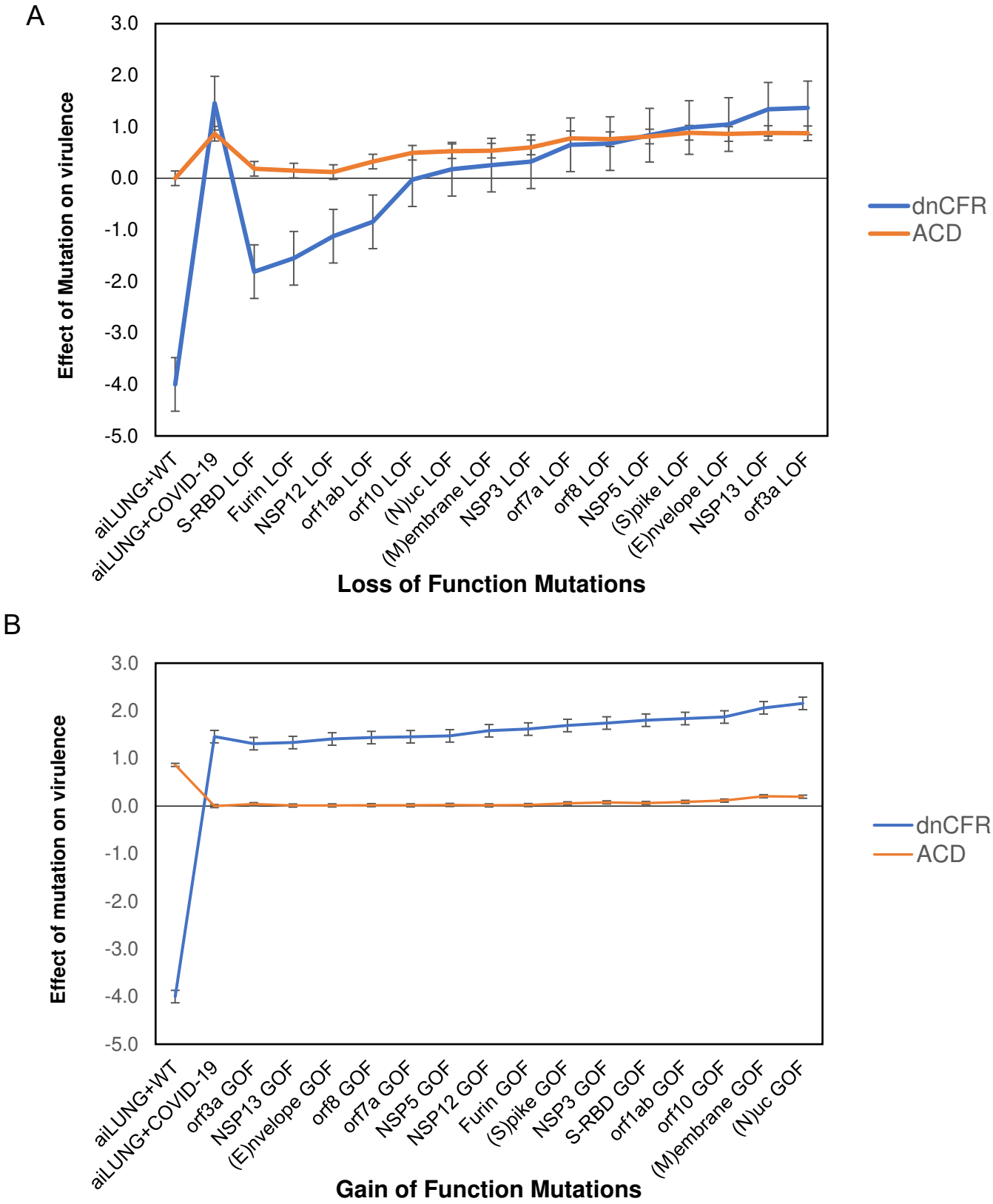


Fig. 3: DeepNEU simulations of the effect of LOF and GOF mutations on SARS-CoV2 virulence.

A, Correlation between ACD metric and dnCFR measure regarding there ability to identify LOF mutations in the SARS-CoV-2 genome. Data from from 3 experiments ± the 95% confidence around the estimates is presented.

A

SARS-CoV2	ACE2	RdRP	NSP5	E	NSP13	M	N	NSP1	NSP2	NSP3	Orf10	Orf1ab	Orf3a	Orf6	Orf7a	Orf8	Plpro	Spike	S-RBD
aiLUNG+WT	0.643	0.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000
aiLUNG+COVID-19	-0.152	0.787	0.075	0.202	0.077	0.202	0.202	0.040	0.114	0.483	0.391	0.140	0.391	0.391	0.452	0.452	0.075	0.202	0.571
S-RBD Mutation LOF	0.027	-0.722	-0.149	-0.187	-0.805	-0.187	-0.187	-0.077	-0.224	-0.539	-0.360	-0.291	-0.360	-0.360	-0.483	-0.483	-0.149	-0.187	-0.900
Furin Mutation LOF	0.316	-0.740	-0.162	-0.192	-0.808	-0.192	-0.192	-0.085	-0.244	-0.561	-0.370	-0.310	-0.370	-0.370	-0.501	-0.501	-0.162	-0.192	-0.607
NSP12 Mutation LOF	0.205	-0.900	-0.036	-0.227	-0.362	-0.227	-0.227	-0.018	-0.053	-0.473	-0.439	-0.074	-0.439	-0.439	-0.464	-0.464	-0.036	-0.227	-0.198
orf1ab Mutation LOF	0.048	-0.566	-0.439	-0.146	-0.183	-0.146	-0.146	-0.227	-0.603	-0.752	-0.283	-0.900	-0.283	-0.283	-0.638	-0.638	-0.439	-0.146	0.198
orf10 Mutation LOF	-0.595	0.742	0.049	-0.439	0.071	-0.439	-0.439	0.027	0.076	0.439	-0.900	0.090	0.376	0.376	0.417	0.417	0.049	-0.439	0.413
(N)uc Mutation LOF	-0.212	0.787	0.073	0.202	0.080	0.202	-0.900	0.039	0.112	0.480	0.391	0.135	0.391	0.391	0.450	0.450	0.073	0.202	0.393
(M)embrane Mutation LOF	-0.603	0.755	0.053	0.199	0.074	-0.900	0.199	0.029	0.083	0.449	0.381	0.098	0.381	0.381	0.425	0.425	0.053	0.199	0.508
NSP3 Mutation LOF	-0.481	0.230	0.055	0.077	0.074	0.077	0.077	-0.439	-0.394	-0.273	0.134	0.101	0.134	0.134	0.187	0.187	-0.900	0.077	0.441
orf7a Mutation LOF	-0.479	0.133	0.053	0.049	0.074	0.049	0.049	0.029	0.083	0.164	0.083	0.098	0.083	0.083	-0.900	0.135	0.053	0.049	0.388
orf8 Mutation LOF	-0.504	0.107	0.053	0.044	0.072	0.044	0.044	0.029	0.081	0.150	0.070	0.097	0.070	0.070	0.122	-0.900	0.053	0.044	0.435
NSP5 Mutation LOF	-0.343	0.272	-0.900	0.083	-0.400	0.083	0.083	0.033	0.094	0.241	0.151	0.112	0.151	0.151	0.210	0.210	0.061	0.083	0.428
(S)pike Mutation LOF	-0.773	0.725	0.033	0.194	0.062	0.194	0.194	0.018	0.050	0.411	0.368	0.060	0.368	0.368	0.396	0.396	0.033	-0.900	0.187
(E)nvelope Mutation LOF	-0.148	0.784	0.067	-0.900	0.067	0.202	0.202	0.037	0.104	0.473	0.390	0.123	0.390	0.390	0.445	0.445	0.067	0.202	0.165
NSP13 Mutation LOF	-0.424	0.756	0.062	0.200	-0.900	0.200	0.200	0.034	0.095	0.460	0.382	0.113	0.382	0.382	0.433	0.433	0.062	0.200	0.478
orf3a Mutation LOF	-0.610	0.751	0.055	0.199	0.074	0.199	0.199	0.030	0.084	0.449	0.380	0.100	-0.900	0.380	0.425	0.425	0.055	0.199	0.518

Legend	-1	-0.5	0	0.5	1

B

Phenotypic profile	ATI&ATII cells	New_ECVirus	S-ACE2 Interface	Virus Clearance	Virus IC RNA Release	Virus Internalization 000	Virus Replication	TMPRSS2
aiLUNG+WT	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-0.452
aiLUNG+COVID-19	0.096	0.399	0.240	0.985	0.263	0.499	0.501	0.699
S-RBD Mutation LOF	-0.096	-0.364	-0.438	-0.987	-0.582	-0.561	-0.588	-0.609
Furin Mutation LOF	-0.105	-0.387	-0.158	-0.984	-0.609	-0.453	-0.639	-0.639
NSP12 Mutation LOF	-0.081	-0.254	0.011	0.004	-0.155	-0.007	-0.436	-0.068
orf1ab Mutation LOF	-0.044	-0.082	0.141	0.992	0.221	0.418	-0.240	0.468
orf10 Mutation LOF	0.079	0.380	-0.087	0.984	0.167	0.311	0.651	0.472
(N)uc Mutation LOF	0.086	0.341	0.136	0.986	0.248	0.463	0.476	0.481
(M)embrane Mutation LOF	0.139	0.609	-0.040	0.983	0.181	0.337	0.670	0.474
NSP3 Mutation LOF	0.098	0.441	-0.009	0.986	0.187	0.350	0.410	0.469
orf7a Mutation LOF	0.067	0.310	-0.035	0.986	0.181	0.336	0.408	0.471
orf8 Mutation LOF	0.098	0.423	-0.024	0.986	0.180	0.337	0.385	0.469
NSP5 Mutation LOF	0.086	0.417	0.059	0.987	0.206	0.386	0.271	0.475
(S)pike Mutation LOF	0.139	0.598	-0.296	0.982	0.112	0.208	0.674	0.690
(E)nvelope Mutation LOF	-0.035	-0.165	0.046	0.966	0.223	0.419	0.333	0.481
NSP13 Mutation LOF	0.116	0.554	0.045	0.985	0.206	0.382	0.354	0.474
orf3a Mutation LOF	0.146	0.634	-0.036	0.982	0.183	0.340	0.670	0.474

Legend	-1	-0.5	0	0.5	1

Fig. 4: Heat map representation of summary DeepNEU simulations data of the effects of individual LOF mutations. A, effect of LOF on the SARS-CoV-2 genome and B, effect of LOF on phenotypic profile of simulated ATI and ATII cells. Data are the average of 3 separate experiments. Data represented as dnCFR measure +/- the 95% CI around the estimates is presented. dnCFR is the DeepNEU measure of SAR-CoV-2 virulence, where (-4) represents the maximum reduction in viral virulence and (+4) represents the maximum increase in virulence).

A

SARS-CoV2	ACE2	RdRP	NSP5	E	NSP13	M	N	NSP1	NSP2	NSP3	Orf10	Orf1ab	Orf3a	Orf6	Orf7a	Orf8	Plpro	Spike	S-RBD
aiLUNG+WT	0.643	0.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000
aiLUNG+COVID-19	-0.152	0.787	0.075	0.202	0.077	0.202	0.202	0.040	0.114	0.483	0.391	0.140	0.391	0.391	0.452	0.452	0.075	0.202	0.571
orf3a Mutation GOF	-0.144	0.782	0.073	0.198	0.056	0.198	0.198	0.037	0.110	0.475	0.386	0.137	0.900	0.386	0.446	0.446	0.073	0.198	0.438
NSP13 Mutation GOF	-0.685	0.743	0.048	0.197	0.900	0.197	0.197	0.026	0.073	0.436	0.375	0.089	0.375	0.375	0.415	0.415	0.048	0.197	0.504
(E)nvelope Mutation GOF	-0.182	0.788	0.076	0.900	0.077	0.201	0.201	0.041	0.116	0.483	0.390	0.142	0.390	0.390	0.452	0.452	0.076	0.201	0.524
orf8 Mutation GOF	-0.639	0.840	0.052	0.217	0.072	0.217	0.217	0.029	0.081	0.481	0.417	0.096	0.417	0.417	0.459	0.900	0.052	0.217	0.523
orf7a Mutation GOF	-0.640	0.843	0.053	0.217	0.073	0.217	0.217	0.029	0.081	0.482	0.418	0.097	0.418	0.418	0.900	0.460	0.053	0.217	0.538
NSP5 Mutation GOF	-0.703	0.886	0.888	0.227	0.473	0.227	0.227	0.026	0.073	0.493	0.436	0.088	0.436	0.436	0.473	0.473	0.047	0.227	0.528
NSP12 Mutation GOF	-0.562	0.900	0.091	0.227	0.070	0.227	0.227	0.048	0.138	0.542	0.439	0.172	0.439	0.439	0.508	0.508	0.091	0.227	0.585
Furin Mutation GOF	-0.386	0.739	0.161	0.192	0.084	0.192	0.192	0.085	0.243	0.560	0.370	0.309	0.370	0.370	0.500	0.500	0.161	0.192	0.671
(S)pike Mutation GOF	-0.367	0.772	0.071	0.202	0.083	0.202	0.202	0.038	0.108	0.475	0.388	0.132	0.388	0.388	0.446	0.446	0.071	0.900	0.812
NSP3 Mutation GOF	-0.652	0.899	0.052	0.228	0.073	0.228	0.228	0.439	0.480	0.756	0.441	0.096	0.441	0.441	0.482	0.482	0.900	0.228	0.531
S-RBD Mutation GOF	-0.229	0.715	0.141	0.186	0.073	0.186	0.186	0.074	0.213	0.529	0.359	0.271	0.359	0.359	0.475	0.475	0.141	0.186	0.900
orf1ab Mutation GOF	-0.670	0.934	0.439	0.235	0.268	0.235	0.235	0.227	0.603	0.826	0.455	0.900	0.455	0.455	0.741	0.741	0.439	0.235	0.542
orf10 Mutation GOF	-0.634	0.754	0.054	0.439	0.073	0.439	0.439	0.029	0.083	0.449	0.900	0.098	0.381	0.381	0.425	0.425	0.054	0.439	0.543
(M)embrane Mutation GOF	-0.620	0.757	0.053	0.199	0.074	0.900	0.199	0.029	0.082	0.449	0.381	0.098	0.381	0.381	0.426	0.426	0.053	0.199	0.515
(N)uc Mutation GOF	-0.065	0.818	0.116	0.209	0.075	0.209	0.900	0.061	0.176	0.540	0.404	0.224	0.404	0.404	0.496	0.496	0.116	0.209	0.507

Legend	-1	-0.5	0	0.5	1

B

Phenotypic Profile	ATI&ATII cells	New_ECVirus	S-ACE2 Interface	Virus Clearance	Virus IC RNA Release	Virus Internalization	Virus Replication	TMPRSS2
aiLUNG+WT	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-0.452
aiLUNG+COVID-19	0.096	0.399	0.240	0.985	0.263	0.499	0.501	0.699
orf3a Mutation GOF	0.087	0.409	0.182	0.980	0.258	0.486	0.413	0.481
NSP13 Mutation GOF	0.138	0.596	-0.090	0.982	0.165	0.311	0.850	0.474
(E)nvelope Mutation GOF	0.154	0.620	0.214	0.983	0.265	0.499	0.439	0.481
orf8 Mutation GOF	0.150	0.647	-0.052	0.982	0.177	0.330	0.723	0.474
orf7a Mutation GOF	0.160	0.690	-0.044	0.982	0.179	0.333	0.723	0.474
NSP5 Mutation GOF	0.154	0.660	-0.085	0.981	0.163	0.306	0.818	0.471
NSP12 Mutation GOF	0.155	0.669	0.019	0.983	0.330	0.370	0.726	0.531
Furin Mutation GOF	0.145	0.629	0.156	0.985	0.606	0.443	0.654	0.639
(S)pike Mutation GOF	0.142	0.617	0.246	0.984	0.245	0.465	0.705	0.697
NSP3 Mutation GOF	0.154	0.667	-0.055	0.981	0.177	0.330	0.773	0.476
S-RBD Mutation GOF	0.156	0.670	0.347	0.985	0.527	0.516	0.666	0.588
orf1ab Mutation GOF	0.160	0.698	-0.060	0.980	0.176	0.328	0.818	0.476
orf10 Mutation GOF	0.163	0.699	-0.038	0.982	0.181	0.338	0.707	0.475
(M)embrane Mutation GOF	0.143	0.627	-0.046	0.982	0.180	0.336	0.709	0.476
(N)uc Mutation GOF	0.070	0.349	0.248	0.988	0.436	0.504	0.363	0.617

Legend	-1	-0.5	0	0.5	1

Fig. 5: Heat map representation of summary DeepNEU simulations data of the effects of individual GOF mutations. A, effect of GOF on the SARS-CoV-2 genome and B, effect of GOF on phenotypic profile of simulated ATI and ATII cells. Data are the average of 3 separate experiments. Data represented as dnCFR measure +/- the 95% CI around the estimates.

A

SARS-CoV-2	dn CFR (LOF)	95% CI
aiLUNG+WT	-4.0	0.521
aiLUNG+COVID-19	1.5	0.521
S-RBD Mutation LOF	-1.8	0.521
Furin Mutation LOF	-1.6	0.521
NSP12 Mutation LOF	-1.1	0.521
orf1ab Mutation LOF	-0.8	0.521
orf10 Mutation LOF	-0.02	0.521
(N)uc Mutation LOF	0.18	0.521
(M)embrane Mutation LOF	0.3	0.521
NSP3 Mutation LOF	0.3	0.521
orf7a Mutation LOF	0.7	0.521
orf8 Mutation LOF	0.7	0.521
NSP5 Mutation LOF	0.8	0.521
(S)pike Mutation LOF	0.98	0.521
(E)nvelope Mutation LOF	1.04	0.521
NSP13 Mutation LOF	1.3	0.521
orf3a Mutation LOF	1.4	0.521

B

SARS-CoV-2	dnCFR (GOF)	±95% CI
aiLUNG+WT	-4.0	0.131
aiLUNG+COVID-19	1.5	0.131
orf3a Mutation GOF	1.3	0.131
NSP13 Mutation GOF	1.3	0.131
(E)nvelope Mutation GOF	1.4	0.131
orf8 Mutation GOF	1.4	0.131
orf7a Mutation GOF	1.5	0.131
NSP5 Mutation GOF	1.5	0.131
NSP12 Mutation GOF	1.6	0.131
Furin Mutation GOF	1.6	0.131
(S)pike Mutation GOF	1.7	0.131
NSP3 Mutation GOF	1.7	0.131
S-RBD Mutation GOF	1.8	0.131
orf1ab Mutation GOF	1.8	0.131
orf10 Mutation GOF	1.9	0.131
(M)embrane Mutation GOF	2.1	0.131
(N)uc Mutation GOF	2.2	0.131

Fig. 6: DeepNEU simulations of the calculated 4 component dnCFR as a measure of SARS-CoV-2 virulence. A, calculated 4 component dnCFR measure for LOF mutations. B, calculated 4 component dnCFR measure for GOF mutations. dnCFR is the DeepNEU measure of SAR-CoV-2 virulence, where (-4) represents the maximum reduction in viral virulence and (+4) represents the maximum increase in virulence). Results are from 3 separate experiments ±95% CI.

Figures

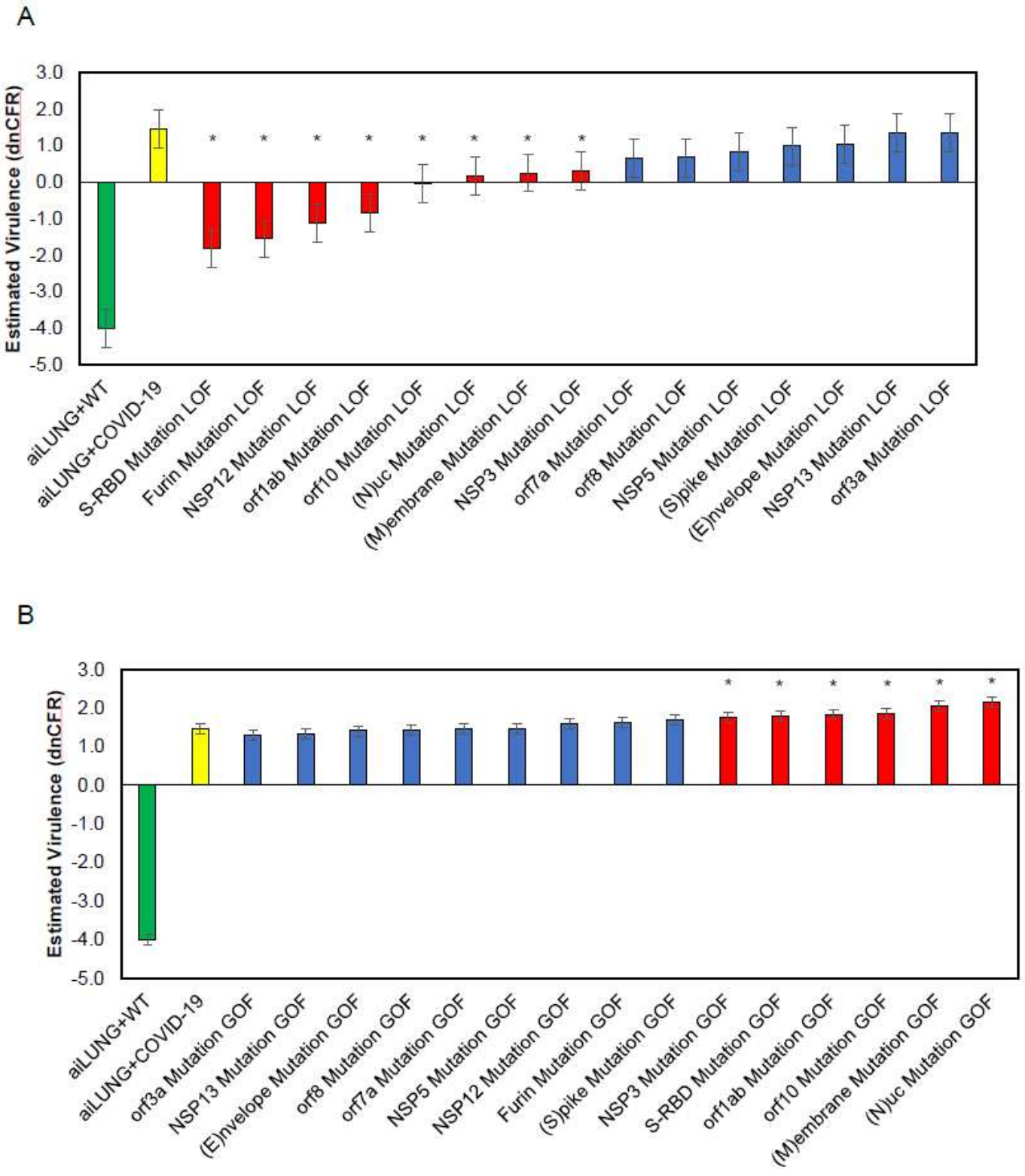


Figure 1

aiLUNG simulations of the effect of LOF and GOF mutations in the SARS-CoV-2 genome on viral virulence. A, LOF simulations and their virulence estimation. Red bars represent mutations in SARS-CoV2 that result in a significant decrease in virulence. B, GOF simulations their virulence estimation. Red bars

represent mutations in SARS-CoV2 that result in a significant increase in virulence. A, B, Green bar represents aiLUNG-WT; Yellow bar represents aiLUNG-COVID-19 (unmutated original SARS-CoV2 genome) simulations; blue bars represent mutations that has no significant effect on virulence. These data represent the average from 3 experiments ± the 95% confidence interval around the average.

A

Factor	ACE2	RdRP	NSP5	E	NSP13	M	N	NSP1	NSP2	NSP3	Orf10	Orf1ab	Orf3a	Orf6	Orf7a	Orf8	Plpro	Spike	S-RBD
Average LOF	-0.305	0.208	-0.076	-0.056	-0.187	-0.056	-0.056	-0.036	-0.044	0.074	0.025	-0.030	0.025	0.110	0.035	0.036	-0.075	-0.056	0.190
p-value*	>0.05 NS	<0.001	<0.01	<0.001	<0.01	<0.001	<0.001	<0.01	<0.01	<0.001	<0.001	>0.05 NS	<0.001	<0.001	<0.001	<0.001	<0.01	<0.001	<0.001
Average GOF	-0.479	0.811	0.158	0.272	0.168	0.272	0.271	0.081	0.177	0.532	0.437	0.197	0.437	0.403	0.510	0.510	0.158	0.271	0.577

B

Mutation	LOF	±95% CI	GOF	±95% CI	p value*
SARS-CoV-2/M	-0.056	0.155	0.272	0.093	3.4322E-05
SARS-CoV-2/N	-0.056	0.155	0.271	0.093	4.4044E-05
SARS-CoV-2/NSP3	0.074	0.231	0.532	0.057	1.555E-05
SARS-CoV-2/S-RBD	0.190	0.217	0.577	0.063	1.2507E-06
dnCFR measure	0.152	0.521	1.652	0.131	1.55E-07

** 2 tailed, Mann-Whitney u test*

Figure 2

DeepNEU summary analysis summary and statistics A, DeepNEU summary results comparing the impact of LOF and GOF mutations on SARS-CoV-2 individual genotypic features . Data from 3 separate experiments. E = Envelope, M = Membrane, N = Nucleocapsid, S-RBD = Spike-Receptor Binding Domain. B,

Summary results from DeepNEU simulations of the individual mutated components and compos ite dnCFR measure effects on SAR-CoV-2 virulence. Results are the average of 3 experiments ± the 95% CI

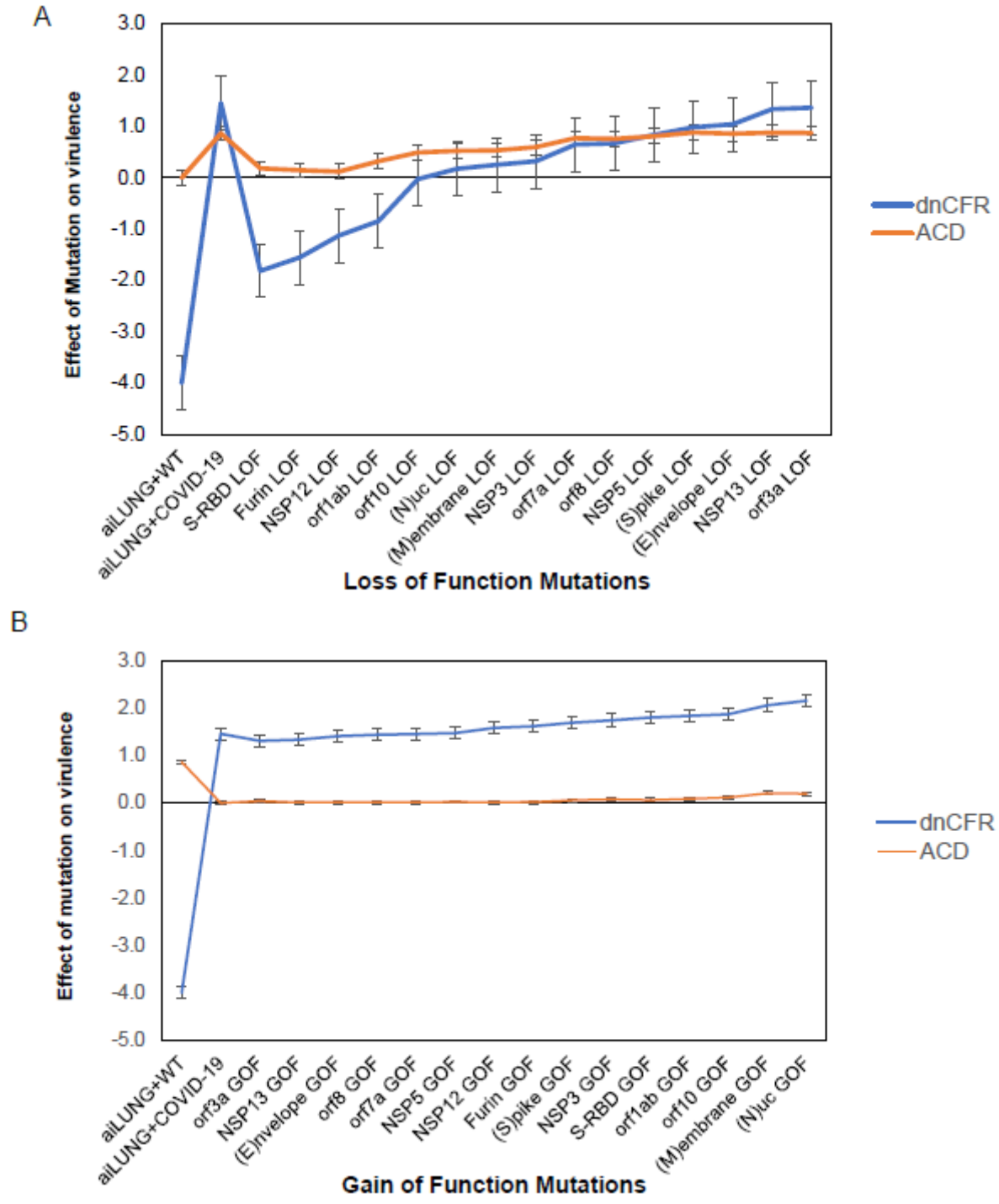


Figure 3

DeepNEU simulations of the effect of LOF and GOF mutations on SARS-CoV2 virulence. A, Correlation between ACD metric and dnCFR measure regarding there ability to identify LOF mutations in the SARS-CoV-2 genome. Data from from 3 experiments ± the 95% confidence around the estimates is presented.

A

SARS-CoV2	ACE2	RdRP	NSP5	E	NSP13	M	N	NSP1	NSP2	NSP3	Orf10	Orf1ab	Orf3a	Orf6	Orf7a	Orf8	Ftpro	Spike	S-RBD
aiLUNG+WT	0.643	0.330	1.000	1.000	-1.000	1.000	-1.000	1.000	-1.000	-1.000	-0.010	-1.000	-1.000	-1.000	1.000	-1.000	-1.000	1.000	-1.000
aiLUNG+COVID-19	-0.152	-0.787	0.025	0.202	0.077	0.202	0.202	0.040	0.114	0.483	0.381	0.140	0.391	0.391	0.452	0.452	0.075	0.202	0.571
S-RBD Mutation LOF	0.527	-0.722	-0.149	-0.187	-0.805	-0.187	-0.187	0.077	-0.224	-0.539	-0.360	-0.291	-0.360	-0.360	-0.483	-0.483	-0.149	-0.187	-0.960
Furin Mutation LOF	0.316	-0.740	-0.152	-0.192	-0.808	-0.192	-0.192	0.085	-0.244	-0.561	-0.370	-0.318	-0.370	-0.370	-0.501	-0.501	-0.152	-0.192	-0.967
NSP12 Mutation LOF	0.205	-0.969	-0.036	-0.227	-0.362	-0.227	-0.227	-0.018	-0.053	-0.473	-0.439	-0.074	-0.439	-0.439	-0.464	-0.464	-0.036	-0.227	-0.198
orf1ab Mutation LOF	0.048	-0.566	-0.439	-0.146	-0.183	-0.146	-0.146	0.227	-0.603	-0.752	-0.283	-0.906	-0.283	-0.263	-0.638	-0.638	-0.439	-0.146	0.198
orf10 Mutation LOF	0.595	0.742	0.049	-0.439	0.071	-0.439	0.439	0.027	0.076	0.439	0.040	0.090	0.376	0.376	0.417	0.417	0.049	-0.439	0.413
(N)uc Mutation LOF	-0.212	0.787	0.073	0.202	0.080	0.202	-0.900	0.035	0.112	0.480	0.381	0.135	0.391	0.391	0.460	0.460	0.073	0.202	0.393
(M)embrase Mutation LOF	-0.603	0.756	0.053	0.199	0.074	-0.900	0.199	0.029	0.083	0.469	0.381	0.098	0.381	0.381	0.425	0.425	0.053	0.199	0.508
NSP3 Mutation LOF	-0.481	0.230	0.055	0.077	0.074	0.077	0.077	-0.439	-0.394	-0.273	0.134	0.101	0.134	0.134	0.187	0.187	-0.900	0.077	0.441
orf7a Mutation LOF	-0.479	0.133	0.053	0.049	0.074	0.049	0.049	0.029	0.083	0.164	0.083	0.098	0.083	0.083	-0.900	0.135	0.053	0.049	0.200
orf8 Mutation LOF	-0.504	0.107	0.053	0.044	0.072	0.044	0.044	0.029	0.081	0.150	0.070	0.097	0.070	0.070	0.122	-0.900	0.053	0.044	0.435
NSP5 Mutation LOF	-0.343	0.272	-0.900	0.083	-0.400	0.083	0.083	0.033	0.094	0.241	0.151	0.112	0.151	0.151	0.210	0.210	0.051	0.083	0.428
(S)pike Mutation LOF	-0.773	0.725	0.033	0.194	0.052	0.194	0.194	0.018	0.053	0.411	0.368	0.068	0.368	0.368	0.396	0.396	0.033	-0.900	0.187
(E)nvelope Mutation LOF	-0.148	0.784	0.067	-0.900	0.057	0.202	0.202	0.037	0.104	0.473	0.380	0.123	0.390	0.390	0.445	0.445	0.057	0.202	0.165
NSP13 Mutation LOF	-1.424	0.756	0.052	0.200	-0.900	0.200	0.200	0.034	0.095	0.460	0.382	0.113	0.382	0.382	0.433	0.433	0.052	0.200	0.478
orf3a Mutation LOF	-0.010	0.751	0.055	0.199	0.074	0.199	0.199	0.030	0.084	0.449	0.380	0.100	-0.900	0.380	0.425	0.425	0.055	0.199	0.510

Legend	-1	-0.5	0	0.5	1

B

Phenotypic profile	ATI&ATII cells	New_ECVirus	S-ACE2 Interface	Virus Clearance	Virus IC RNA Release	Virus Internalization (00)	Virus Replication	TMPRSS2
aiLUNG+WT	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-0.452
aiLUNG+COVID-19	0.096	0.399	0.240	0.986	0.263	0.699	0.501	0.699
S-RBD Mutation LOF	-0.096	-0.364	-0.438	-0.967	-0.582	-0.561	-0.588	-0.609
Furin Mutation LOF	-0.105	-0.387	-0.158	-0.984	-0.609	-0.453	-0.630	-0.639
NSP12 Mutation LOF	-0.081	-0.254	0.011	0.004	-0.155	-0.007	-0.436	-0.058
orf1ab Mutation LOF	-0.044	-0.002	0.141	0.992	0.221	0.416	-0.240	0.466
orf10 Mutation LOF	0.079	0.380	-0.087	0.984	0.167	0.311	0.651	0.472
(N)uc Mutation LOF	0.099	0.341	0.136	0.988	0.248	0.653	0.479	0.481
(M)embrane Mutation LOF	0.139	0.605	-0.040	0.983	0.181	0.337	0.679	0.474
NSP3 Mutation LOF	0.098	0.441	-0.089	0.986	0.187	0.350	0.419	0.469
orf7a Mutation LOF	0.067	0.310	-0.035	0.986	0.181	0.336	0.408	0.471
orf8 Mutation LOF	0.098	0.423	-0.024	0.986	0.190	0.337	0.385	0.469
NSP5 Mutation LOF	0.089	0.417	0.059	0.987	0.206	0.386	0.271	0.475
(S)pike Mutation LOF	0.139	0.598	-0.296	0.982	0.112	0.208	0.674	0.690
(E)nvelope Mutation LOF	-0.035	-0.165	0.046	0.968	0.223	0.619	0.333	0.481
NSP13 Mutation LOF	0.116	0.554	0.045	0.985	0.206	0.382	0.354	0.474
orf3a Mutation LOF	0.148	0.634	-0.036	0.982	0.183	0.340	0.679	0.474

Legend	-1	-0.5	0	0.5	1

Figure 4

Heat map representation of summary DeepNEU simulations data of the effects of individual LOF mutations. A, effect of LOF on the SARS-CoV-2 genome and B, effect of LOF on phenotypic profile of simulated ATI and ATII cells. Data are the average of 3 separate experiments. Data represented as dnCFR measure +/- the 95% CI around the estimates is presented. dnCFR is the DeepNEU measure of SAR-CoV-2

virulence, where (-4) represents the maximum reduction in viral virulence and (+4) represents the maximum increase in virulence).

A

SARS-CoV2	ACE2	RdRP	NSP5	E	NSP13	M	N	NSP1	NSP2	NSP3	Orf10	Orf1ab	Orf3a	Orf6	Orf7a	Orf8	Ptpro	Spike	S-RBD
aiLUNG+WT	0.643	0.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000
aiLUNG+COVID-19	-0.152	0.787	0.075	0.202	0.077	0.202	0.202	0.040	0.114	0.483	0.391	0.140	0.391	0.391	0.452	0.452	0.075	0.202	0.571
orf3a Mutation GOF	-0.144	0.782	0.073	0.196	0.056	0.196	0.196	0.037	0.110	0.475	0.386	0.137	0.900	0.386	0.446	0.446	0.073	0.196	0.438
NSP13 Mutation GOF	0.685	0.743	0.048	0.197	0.900	0.197	0.197	0.026	0.073	0.436	0.375	0.089	0.375	0.375	0.415	0.415	0.048	0.197	0.504
(E)nvelope Mutation GOF	0.182	0.798	0.076	0.900	0.077	0.201	0.201	0.041	0.116	0.483	0.390	0.142	0.390	0.390	0.452	0.452	0.076	0.201	0.524
orf8 Mutation GOF	-0.639	0.840	0.052	0.217	0.072	0.217	0.217	0.029	0.081	0.481	0.417	0.096	0.417	0.417	0.459	0.900	0.052	0.217	0.523
orf7a Mutation GOF	-0.640	0.843	0.053	0.217	0.073	0.217	0.217	0.029	0.081	0.482	0.418	0.097	0.418	0.418	0.900	0.460	0.053	0.217	0.538
NSP5 Mutation GOF	-0.703	0.886	0.888	0.227	0.473	0.227	0.227	0.026	0.073	0.493	0.436	0.088	0.436	0.436	0.473	0.473	0.047	0.227	0.528
NSP12 Mutation GOF	-0.562	0.900	0.091	0.227	0.070	0.227	0.227	0.048	0.138	0.542	0.439	0.172	0.439	0.439	0.508	0.508	0.091	0.227	0.585
Furin Mutation GOF	-0.386	0.739	0.161	0.192	0.084	0.192	0.192	0.085	0.243	0.560	0.370	0.389	0.370	0.370	0.500	0.500	0.161	0.192	0.671
(S)pike Mutation GOF	-0.367	0.772	0.071	0.202	0.083	0.202	0.202	0.038	0.108	0.475	0.388	0.132	0.388	0.388	0.446	0.446	0.071	0.900	0.812
NSP3 Mutation GOF	-0.652	0.899	0.052	0.228	0.073	0.228	0.228	0.439	0.480	0.756	0.441	0.096	0.441	0.441	0.482	0.482	0.900	0.228	0.531
S-RBD Mutation GOF	-0.229	0.715	0.141	0.186	0.073	0.186	0.186	0.074	0.213	0.529	0.359	0.271	0.359	0.359	0.475	0.475	0.141	0.186	0.900
orf1ab Mutation GOF	-0.670	0.914	0.439	0.235	0.268	0.235	0.235	0.227	0.893	0.826	0.455	0.900	0.455	0.455	0.741	0.741	0.439	0.235	0.542
orf10 Mutation GOF	-0.634	0.754	0.054	0.439	0.073	0.439	0.439	0.029	0.083	0.449	0.900	0.098	0.381	0.381	0.425	0.425	0.054	0.439	0.543
(M)embrane Mutation GOF	-0.620	0.767	0.053	0.199	0.074	0.900	0.199	0.029	0.082	0.449	0.381	0.098	0.381	0.381	0.426	0.426	0.053	0.199	0.515
(N)uc Mutation GOF	-0.065	0.818	0.116	0.209	0.075	0.209	0.900	0.061	0.176	0.540	0.404	0.224	0.404	0.404	0.496	0.496	0.116	0.209	0.507

Legend	-1	-0.5	0	0.5	1

B

Phenotypic Profile	ATI&ATII cells	New_ECVirus	S-ACE2 interface	Virus Clearance	Virus IC RNA Release	Virus Internalization	Virus Replication	TMPRSS2
aiLUNG+WT	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-0.652
aiLUNG+COVID-19	0.096	0.399	0.240	0.995	0.263	0.499	0.501	0.699
orf3a Mutation GOF	0.087	0.409	0.182	0.999	0.258	0.486	0.413	0.481
NSP13 Mutation GOF	0.138	0.596	-0.090	0.982	0.165	0.311	0.858	0.474
(E)nvelope Mutation GOF	0.154	0.620	0.214	0.983	0.265	0.499	0.439	0.481
orf8 Mutation GOF	0.150	0.647	-0.052	0.982	0.177	0.330	0.723	0.474
orf7a Mutation GOF	0.150	0.650	-0.044	0.982	0.179	0.333	0.723	0.474
NSP5 Mutation GOF	0.154	0.660	-0.085	0.981	0.163	0.306	0.818	0.471
NSP12 Mutation GOF	0.155	0.669	0.019	0.983	0.330	0.370	0.726	0.531
Furin Mutation GOF	0.145	0.629	0.156	0.995	0.606	0.443	0.664	0.639
(S)pike Mutation GOF	0.142	0.617	0.240	0.994	0.245	0.465	0.705	0.697
NSP3 Mutation GOF	0.154	0.667	-0.055	0.981	0.177	0.330	0.773	0.476
S-RBD Mutation GOF	0.156	0.670	0.347	0.995	0.527	0.516	0.666	0.588
orf1ab Mutation GOF	0.160	0.688	-0.060	0.980	0.176	0.328	0.818	0.476
orf10 Mutation GOF	0.163	0.699	-0.038	0.982	0.181	0.338	0.707	0.475
(M)embrane Mutation GOF	0.143	0.627	-0.046	0.982	0.180	0.336	0.709	0.476
(N)uc Mutation GOF	0.070	0.349	0.248	0.988	0.436	0.504	0.363	0.617

Legend	-1	-0.5	0	0.5	1

Figure 5

Heat map representation of summary DeepNEU simulations data of the effects of individual GOF mutations. A, effect of GOF on the SARS-CoV-2 genome and B, effect of GOF on phenotypic profile of

simulated ATI and ATII cells. Data are the average of 3 separate experiments. Data represented as dnCFR measure +/- the 95% CI around the estimates.

A

SARS-CoV-2	dn CFR (LOF)	95% CI
aiLUNG+WT	-4.0	0.521
aiLUNG+COVID-19	1.5	0.521
S-RBD Mutation LOF	-1.8	0.521
Furin Mutation LOF	-1.6	0.521
NSP12 Mutation LOF	-1.1	0.521
orf1ab Mutation LOF	-0.8	0.521
orf10 Mutation LOF	-0.02	0.521
(N)uc Mutation LOF	0.18	0.521
(M)embrane Mutation LOF	0.3	0.521
NSP3 Mutation LOF	0.3	0.521
orf7a Mutation LOF	0.7	0.521
orf8 Mutation LOF	0.7	0.521
NSP5 Mutation LOF	0.8	0.521
(S)pike Mutation LOF	0.98	0.521
(E)nvelope Mutation LOF	1.04	0.521
NSP13 Mutation LOF	1.3	0.521
orf3a Mutation LOF	1.4	0.521

B

SARS-CoV-2	dnCFR (GOF)	±95% CI
aiLUNG+WT	-4.0	0.131
aiLUNG+COVID-19	1.5	0.131
orf3a Mutation GOF	1.3	0.131
NSP13 Mutation GOF	1.3	0.131
(E)nvelope Mutation GOF	1.4	0.131
orf8 Mutation GOF	1.4	0.131
orf7a Mutation GOF	1.5	0.131
NSP5 Mutation GOF	1.5	0.131
NSP12 Mutation GOF	1.6	0.131
Furin Mutation GOF	1.6	0.131
(S)pike Mutation GOF	1.7	0.131
NSP3 Mutation GOF	1.7	0.131
S-RBD Mutation GOF	1.8	0.131
orf1ab Mutation GOF	1.8	0.131
orf10 Mutation GOF	1.9	0.131
(M)embrane Mutation GOF	2.1	0.131
(N)uc Mutation GOF	2.2	0.131

Figure 6

DeepNEU simulations of the calculated 4 component dnCFR as a measure of SARS-CoV-2 virulence. A, calculated 4 component dnCFR measure for LOF mutations. B, calculated 4 component dnCFR measure for GOF mutations. dnCFR is the DeepNEU measure of SAR-CoV-2 virulence, where (-4) represents the

maximum reduction in viral virulence and (+4) represents the maximum increase in virulence). Results are from 3 separate experiments ±95% CI.