Visualizing Amino Acid Substitutions in a Physicochemical Vector Space

Louis R. Nemzer, Ph.D.

Department of Chemistry and Physics, Halmos College of Arts and Sciences, Nova Southeastern University, Ft. Lauderdale, FL USA
email: lnemzer@nova.edu

"Just as no building lacks an architecture, so no choice lacks a context"
-Richard H. Thaler[1]

"The species we see today are not those that have been buffered from evolutionary change, but those that have successfully adapted to changing environments."[2]

Abstract

A three-dimensional representation of the twenty proteinogenic amino acids in a physicochemical space is presented. Vectors corresponding to amino acid substitutions are classified based on whether they are accessible via a single-nucleotide mutation. It is shown that the standard genetic code establishes a "choice architecture" that permits nearly independent tuning of the properties related with size and those related with hydrophobicity. This work sheds light on the metarules of evolvability that may have shaped the standard genetic code to increase the probability that adaptive point mutations will be generated. Illustrations of the usefulness of visualizing amino acid substitutions in a 3D physicochemical space are shown using recent datasets collected regarding the SARS-CoV-2 receptor binding domain. First, the substitutions most responsible for antibody escape are almost always inaccessible via single nucleotide mutation, and also change multiple properties concurrently. Second, it is shown that assays of ACE2 binding by sarbecovirus variants, including the viruses responsible for SARS and COVID-19, are more easily understood when plotted with this method. The results of this research can extend our understanding of certain hereditary disorders caused by point mutations, as well as guide the development of rational protein and vaccine design.

Keywords: DNA; Genetic Code; Amino Acids; Principal Component Analysis; Natural Selection; Evolution; Rational Protein Design; COVID-19

Introduction

The canonical genetic code, which maps three-letter DNA codons into amino acids, has long been recognized to possess intrinsic error-correcting mechanisms. In particular, the encoding is nearly optimal[3] in minimizing the chance that a single nucleotide mutation will cause a drastic change in the physicochemical properties of the resulting residue. One celebrated example is that codons with the middle nucleotide U almost always produce a hydrophobic amino acid. However, the specific constraints of allowed substitutions, which lead to new variants, have received comparatively less attention. There is evidence that the genetic code has been shaped by objectives besides robustness to error.[4] In addition to the adaptive value in avoiding nonconservative mutations, natural selection may favor a system in which the most important[5] physical and chemical properties of proteins can be adjusted independently. One reason this topic has been less studied is that it raises philosophical meta-questions surrounding how to think about the selection pressure on systems to exhibit "evolvability," which is the ability to adapt efficiently.[6] It should be remembered, however, that the standard codon table is itself a product of evolution,[7] albeit over a much longer timescale[8] compared with the evolution of individual genes. Selection pressures on this "choice architecture"[9] - the system of presented options - would favor systems with the capacity to generate novel heritable variation[10,11] by appropriately constraining the impact of any single point mutation. That is, potentially valuable changes should generally significantly alter at most one physicochemical property. Conversely, pathogens attempting to disguise themselves to evade the immune system of the host would benefit from amino acid substitutions in antigens that greatly change their binding with antibodies. This work is among the first assess the canonical genetic code in this light.

Background

The physical and chemical properties of the twenty standard amino acids provide the raw materials for all biological protein structure. Several hereditary disorders are linked to specific single amino acid substitutions that break or substantially change the function of the resulting protein. Recent advances in genetic editing, such as CRISPR, have opened the possibility for targeted therapeutic interventions. Additionally, the products of rational protein design, including the SARS-CoV-2 prefusion stabilized spike protein[12] are already in clinical use, but this method depends on detailed knowledge the effects of amino acid substitutions.

Complex adaptive systems, from organisms to societies, require "evolvability," which is an ability to produce new traits,[13] but almost always[14] via bounded tinkering.[15,16] The guide rails that constrain these changes are generally stable over long periods. As an analogy, in many political systems, a written Constitution provides the metarules by which regular statutes may be altered. This Constitution may itself be modified with via amendments, but much more slowly and deliberately. Similarly, the canonical codon table and fixed amino acid properties provide the

context for the subset of mutations that are accessible by a single-nucleotide DNA mutation.[17] The near universality of the current genetic code evinces its slow rate of changes.[18]

The sum of these constraints comprises a choice architecture. In human-designed systems, a choice architecture can include everything from limits on the legislative branch, mandatory failsafe interlocks, or simply requiring extra clicks to get a custom software installation.[19] From a genetic perspective, there is an adaptive advantage to using a codon code in which the most important physical and chemical properties can be independently adjusted. Protein evolvability may be quantified in various ways, such as the proportion of sites shown to be under positive selection multiplied by the average rate of adaptive evolution at these sites.[20]

While error-resistance is clearly an important feature - since beneficial variants are likely to be similar to existing proteins[21] - a purely static genome would not be conducive to long term evolutionary success in constantly shifting environments.[22] Some previous work on protein evolvability under the genetic code[23] examined the efficiency of allowed mutations as a "search algorithm"[24] that explores a larger fraction of the space of functional variants when compared with random codes.[25] It has also been shown that the substitutions accessible via point mutation in the TEM-1 gene for beta-lactamase were more likely to be adaptive than inaccessible changes.[26] Also, the stepwise evolvability of proteins depends on the connectivity[27] of amino acid sequence space.[28] Some previous research considered decision trees[29] or random forest algorithms[30] for point mutations using physicochemical distance as part of the prediction[31] of deleterious effects. However, this research is one of the only projects to consider the directionality of the vectors, since edges oriented (nearly) parallel to one of the axes would primarily change only one physicochemical aspect at a time. As an example of the power of this framework, recent data on the amino acid substitutions in the SARS-CoV-2 receptor binding domain associated with immune escape are analyzed. In addition, when visualized in a similar space, the results for site saturation mutagenesis ACE2 binding assays with sarbecoviruses can be interpreted as either idiosyncratic – such as significantly reduced binding when a charged amino acid is introduced – or more general, as when based on size and hydrophobicity.

Method

This work builds on previous results, including a principal component analysis (PCA) used to generate the 3D representation in physicochemical space (figure 1),[32] as well as the classification system for DNA mutations.[33] Here, only nonsynonymous missense mutations are considered, and when combining forward and backward changes together, there are (20 x 19)/2 = 190 possible amino acid substitution pairs. Of these, 75 are accessible via point mutations in the canonical genetic code, and 115 are not. This leads to a vector space (see table 1) in which the x-, y-, and z-axes are assigned to size, hydrophobicity, and charge, respectively. The undirected edges represent substitutions separated into "possible" and "not possible," as shown in figure 2. The thickness each edge in the "possible" network represents the number of permutations that can

give rise to that substitution (table 2). The coloring of the edges shows how damaging the change is predicted to be, on average, based on the BLOSUM62 substitution matrix.[34] To further motivate the usefulness of this vector space representation, some notable single-nucleotide mutations are shown in figure 3. These include those in β-globin gene causing sickle cell anemia, the HFE gene causing hereditary hemochromatosis, the altered EphB2 gene in all crested pigeons,[35] and a SARS-CoV-2 variant noted for its ability to evade the immune system. The possible substitutions are then separated by the mutated nucleotide and mutation type (figure 4). The insets provide the legend for the mutation types, based on the property shared by the original and new nucleotide (puRine, pYrimidine, Weak, Strong, aMino, or Keto). The order is based on the importance of the codon position,[36] in which the middle nucleotide is the most determinative, followed by the first and then the third. The joint histograms of the BLOSUM62 and Euclidian PCA distance is separated into possible and not possible (figure 5).

In order to perform a more quantitative analysis that captures the "directedness" of amino acid substitutions in the physicochemical space independent of distance, a measure of information entropy is used. The entropy will be zero if the change is entirely along one direction (PC1, PC2, or PC3), while it will be maximized if the change is equal in all three simultaneously. First, the components are normalized by the vector length squared:

$$x = \frac{(\Delta PC1)^2}{(\Delta PC1)^2+(\Delta PC2)^2+(\Delta PC3)^2}\ ,\ y = \frac{(\Delta PC2)^2}{(\Delta PC1)^2+(\Delta PC2)^2+(\Delta PC3)^2}\ ,\ z = \frac{(\Delta PC3)^2}{(\Delta PC1)^2+(\Delta PC2)^2+(\Delta PC3)^2}$$

This ensures that x + y + z = 1. Then, the Shannon entropy[37] is calculated:

$$Entropy = -[x\log_2(x) + y\log_2(y) + z\ \log_2(z)]$$

Finally, the entropy is divided by the maximum value, $-\log_2(\frac{1}{3}) \approx 1.58$, so the normalized entropy will always be between zero and one. Figure 6a separates the amino acid substitutions based on the number of codon pairs that differ by one nucleotide. For example, each of valine's four codons (GUU, GUC, GUA, GUG) can be mutated to code for alanine (GCU, GCC, GCA, GCG) by changing the middle nucleotide from U to C. Therefore, the V <-> A substitution will be classified as having 4 codons. Substitutions that are not accessible have zero codons. The accessible changes are further separated into "low codon" for 1, 2, or 3 possible codons, and "high codon" for 4 or 6 codons. The normalized entropy distributions for the three different categories are shown in figure 6(b). As a test case for using the tools introduced here, a recent dataset[38,39,40] with 175,297 amino acid substitutions in the SARS-CoV-2 receptor binding domain is employed. For this visualization, the mean viral escape value for each substitution is calculated across all listed sites. Now, there are 337 datapoints, since forward and backwards substitutions are counted separately, but not every change appears in the dataset. The mean escape versus the normalized entropy, or change in PC2 (the principal component that measures hydrophobicity) are plotted in figures 7a and 7b, respectively. Next, the physicochemical vectors representing substitutions associated with high and low mean escape are shown (figure 8). The edges are colored by the normalized entropy for each change. Finally, select results are visualized

from another recent dataset[41] by the same research group produced via site saturation mutagenesis experiments.[42] Now, the change in binding of sarbecovirus spike proteins, including those of SARS-CoV1 (SARS) and SARS-CoV2 (COVID-19), with human ACE2 receptors are analyzed. In these assays, all possible amino acid substitutions at six analogous sites are measured.

Discussion of Results

The physicochemical vector space defined by the PCA shows a clear clustering of properties based on the middle nucleotide of the codon. This modularity is an initial hint that the genetic code should be viewed as a kind of choice architecture. Notably, if the middle letter is not changed by a mutation, it is likely that the substitution will be conservative, or even synonymous.

Figure 2 shows how the modularity extends to the accessible amino acid substitutions. There are many short connections between similar residues, and fewer long connections between amino acid families. This is even more noticeable when compared with the substitutions that are not possible via a single point mutation. There is also a significant difference in the orientation of the edges, in which possible substitutions correspond to lines more likely to horizontal or vertical, but not diagonal. This is associated with physicochemical changes confined to size or hydrophobicity, respectively, while a diagonally oriented edge changes both simultaneously.

Because both positively (R, K) and negatively (E, D) charged residues are naturally hydrophilic and tend to be large, substitutions will be automatically constrained mostly within the z direction. As an illustration, K <-> E is possible and represents a very large change in charge, but has a BLOSSUM62 of +1, which indicates a benign change.

The single-nucleotide mutations shown in figure 3 help illustrate the effects of accessible substitutions. In some cases of heterozygous advantage, as with carriers of one sickle hemoglobin gene who are more resistant to malaria,[43] a significant change in properties occurs even though it is caused by a point mutation. However, the change is limited in that sickle hemoglobin differs enough from the wild type to confer increase malaria resistance, but is similar enough so that symptoms only appear in homozygous individuals, and even then, only under low-oxygen conditions. In figure 4, the edges are separated by nucleotide position and mutation type. Here, mutations in the middle nucleotide (fig 4a) tend to be more vertically directed, while those of the first nucleotide are horizontally oriented (fig 4b). The sparsity of edges in fig 4c shows how few nonsynonymous mutations are possible by changing the third nucleotide. These findings lend themselves to the interpretation that the canonical genetic code is arranged such that the size and hydrophobicity of the residues can be changed mostly independently. Specifically, changing the middle nucleotide primarily impacts the hydrophobicity, while the size is controlled by the first letter of each codon.

Figure 5 looks at the Euclidean distance for possible and not possible substitutions. This approach is comparable to the Miyata[44] or Grantham[45] distance metrics. However, these measures use volume and polarity (along with composition in the Grantham case), which are correlated with each other, while here orthogonal principal components are employed. As expected, the distance

and BLOSUM62 values are inversely related, showing that swapping very dissimilar amino acids is likely to be deleterious. Interestingly, the modular structure of the vector space enforced by the genetic code is apparent here too. The possible substitutions show distinct peaks for the short intrafamily and longer interfamily edges, while they are muddled into a single blob for the inaccessible substitutions. In figure 6, the impact of distance is removed, and instead, the directedness along one axis is quantified using the normalized entropy. The "high codon" substitutions differ in that a much larger fraction of them have low entropy, as compared with the inaccessible and "low codon" categories.

The substitutions showing the highest escape in figure 7a, as indicated by the arrows, all tend to have large normalized entropy values, which means they change multiple properties at once. In particular, they tend to have large changes in PC2, as seen in figure 7b. In the case of the dataset used, the highest escape substitutions involve glutamic acid (E) or phenylalanine (F). They also had high entropy, as indicated by the diagonal lines and colors. With phenylalanine, there may also the effect of altering the aromatic cage[46] binding.

These results can be interpreted as to mean that the mutations that can best escape the immune system tend to change both the size and hydrophobicity at the same time. They also are not usually accessible by a single nucleotide mutation, as is the case with most large changes in physicochemical properties. Additionally, it may be adaptive for the host immune system to expend less resources on preventing escape in these situations, since the requirement that at least two mutations occur makes them inherently less likely.

Finally, the novel "fireworks" diagrams shown in figures 9 and 10 illustrate the usefulness of this approach to visualizing amino acid substitutions in physicochemical space. Binding with the ACE2 receptors in lung cells is often a crucial step in the infection cycle or respiratory viruses. Here, blue arrows represent strong binding, while red arrows represent impaired binding ability. The dashed lines show amino acid substitutions that are not accessible by a single nucleotide mutation, and the thickness of each solid line indicates the number of associated codons. Even though the pathogens responsible for SARS (SARS-CoV-1) and COVID-19 (SARS-CoV-2) are closely related, it is known that mutations at analogous sites can have very different effects on ACE2 binding affinity. As seen in figure 9, mutations at position 501 that produce a positively charged amino acid (R or K) significantly reduce binding between the SARS virus and ACE2, as will a mutation to Y. In contrast, the binding of some SARS-CoV-2 variants is greatly enhanced by a N501Y substitution.[47] In both cases, the downward blue lines indicate that substitutions to more hydrophobic amino acids do not inhibit binding (except for L).

To further highlight some comparisons of analogous sites in the dataset, figure 10 places sites 455 and 494 side-by-side for SARS-CoV-1 and SARS-CoV-2. It can be seen that the SARS virus is susceptible to reduced binding if Y455 is replaced by any of a cluster of amino acids that are large and hydrophilic. This not that case for the COVID-19 virus at the analogous site (L455), since it will suffer significantly reduced binding only if replaced with a charged amino acid. Interestingly, many of these changes are precluded from occurring by a single point mutation. At site 494, both

SARS and COVID-19 viral binding apparently suffer only due to specific substitutions, as opposed to sectors of amino acids with similar properties. For example, D494V is accessible by a single-nucleotide mutation, and has a large negative impact on the binding between SARS-CoV-1 and ACE2. A similar particularized substitutions for SARS-CoV-2 would be S494I.

The current work focuses specifically on missense point mutations allowed by the canonical genetic code to elucidate this particular set of explicit constraints. Other types of mutations, such as nonsense mutations that lead to premature termination of the protein, as well as base deletions or insertions, are not considered here, since they depend on many contextual factors, such as adjacent codons. Similarly, there is increasing evidence that even synonymous mutations can create variation subject to selection via differences in gene expression and protein folding.[48] It has also been proposed that "hidden stops"[49] serve as an additional fail-safe mechanism to halt the translation of faulty proteins in the event of mutation. These would be another example of a choice constraint intrinsic to the canonical genetic code, but are beyond the scope of the current work. Nonstandard codes are minor modifications stemming from standard code, and often involve assigning an amino acid to a stop codon.[50]

Conclusions

The canonical genetic code places constraints on the amino acids substitutions that are accessible via single-nucleotide point mutations. Just as transcribing text with a typewriter[51] is more likely to lead to usable variation compared with either a perfect fidelity copy-machine or a leaky pen, the choice architecture encountered by random mutations is crucial to the generation of bounded adaptation. Here, it is proposed that in addition to simple error minimization, the genetic code contributes to the evolvability of new variation by permitting mostly independent control over the size and hydrophobicity of amino acids, which are the two most important physicochemical properties. This can have clinical applications for heritable disorders, as well as implications for evolutionary theory. In the current era of large genomic data, the use of machine learning, including the identification of heritable conditions due to single-nucleotide polymorphisms, can be informed by the findings of this work. The constraint that mutational trajectories[52] consist only of changes accessible via point mutations may make the evolution of certain traits more predictable.[53] Also, the rational design of new proteins and vaccines can potentially be accomplished more efficiently, leading to new therapeutic interventions. In particular, the immune escape of viral variants with single or multiple mutations can be more easily predicted and prevented. Finally, new methods of visualizing and presenting data, as with the novel "fireworks" diagrams, can help make sense of large datasets, such as those produced by site saturation mutagenesis or other modern bioinformatics techniques.

Acknowledgements

I would like to express my sincere appreciation to Rees Kassen for many enlightening discussions. This work was supported by NSU PFRDG #335472.

Figure Captions

Figure 1: (a) 3D-Printed visualization of the physicochemical vector space generated by principal component analysis (PCA). The first principal component (PC1) is identified with the size of each amino acid, with smaller values to the right. PC2 is assigned to the hydrophobicity, with more hydrophilic residues at the top. The vertical height of each bar represents PC3, which shows the charge the amino acid, with more positive values being taller. **(b)** Two-dimensional projection of the same data. In both panels, the color indicates the middle nucleotide of the codon that encodes for each amino acid.

Figure 2: PCA diagrams in which the undirected edges display all possible amino acid substitutions. PC1 and PC2 are represented by the x- and y-axes, respectively. The PC3 value is indicated by the color of the node, shifted so the minimum is zero (see colorbar). The left figure shows substitutions that are accessible via single-nucleotide (point) mutations, with the thickness showing the number of possible codons associated with it. Conversely, the figure on the right has the remaining inaccessible substitutions that require more than one nucleotide change. The colors of the edges represent the impact of the substitution based on the BLOSUM62 matrix, with darker lines being more potentially harmful.

Figure 3: Selected notable point mutations. In some cases, a single missense mutation destroys the function of the resulting protein (e.g., in the kinase domain of EphB2 in crested pigeon), while in others the behavior is significantly altered (e.g., sickle-cell hemoglobin). One SARS-CoV-2 variant shown to have high immune escape[54] is also included.

Figure 4: Substitutions sorted by mutation type. The change of base is classified by the property preserved. For example, a U <-> A mutation is Y because both are pyrimidines. The positions of the mutations are shown in decreasing order of importance: (a) middle, (b) first, and (c) last nucleotide of the codon. The color of the nodes shows the PC3 value of each amino acid. The width and color of the edges show the likelihood of occurrence and change of deleterious effect, respectively in a database of oncogenes.[55]

Figure 5: Heatmap of possible and not possible substitutions. The x-axis represents the length of the edge in the physicochemical space, while the y-axis shows the corresponding BLOSUM62 value, with negative values being more damaging.

Figure 6: (a) Separation of amino acid substitutions based on the number of associated codons. “Not accessible” substitutions have zero codon pairs for the original and substituted amino acids that differ by a single nucleotide. If 1, 2, or 3 codon permutations exist, it is classified as “low codon,” and “high codon” if 4 or 6 exist. **(b)** The distributions of normalized entropy for each category of substitutions. Small values of normalized entropy correspond with substitutions that change just one direction at a time. On the other hand, large values show that multiple properties changed at the same time, which show up as diagonal lines in the physicochemical space visualizations.

Figure 7: Mean escape values for SARS-CoV-2 receptor binding domain versus **(a)** normalized entropy of the substitution or **(b)** change in PC2, which represents the hydrophilic nature of the amino acids. The arrows, which indicate the substitutions associated with the largest immune escape, tend to have large normalized entropies and changes in PC2. They are also not accessible via single nucleotide mutations.

Figure 8: PCA visualizations showing the substitutions **(a)** above or **(b)** below the midpoint of the mean escape values in the dataset (0.073). The high-escape substitutions change multiple properties at once, as indicated by the diagonal lines and large normalized entropy values.

Figure 9: "Fireworks graphs" from site saturation mutagenesis experimental data. Blue and red lines show high and low binding affinity with human ACE2, respectively. Thicker lines indicate more associated codon permutations, and dashed lines show substitutions that are not accessible by single nucleotide mutations. Here, a **(a)** SARS-CoV-1 (SARS) and a **(b)** SARS-CoV-2 (COVID-19) variant at site 501 are shown. In the case of SARS, substitutions to large, and especially positively charged, amino acids disrupt binding the most. For COVID-19, the most damaging changes are to the positively charged R or K, and only the latter is accessible via point mutation.

Figure 10: Fireworks graphs for SARS-CoV-1 and SARS-CoV-2 variants at sites 455 and 494. Sometimes, binding inhibition is similar across sectors of related amino acids, as for the large hydrophilic residues at site 455. In other situations, the changes are limited to particular substitutions, as seen for the data for site 494.

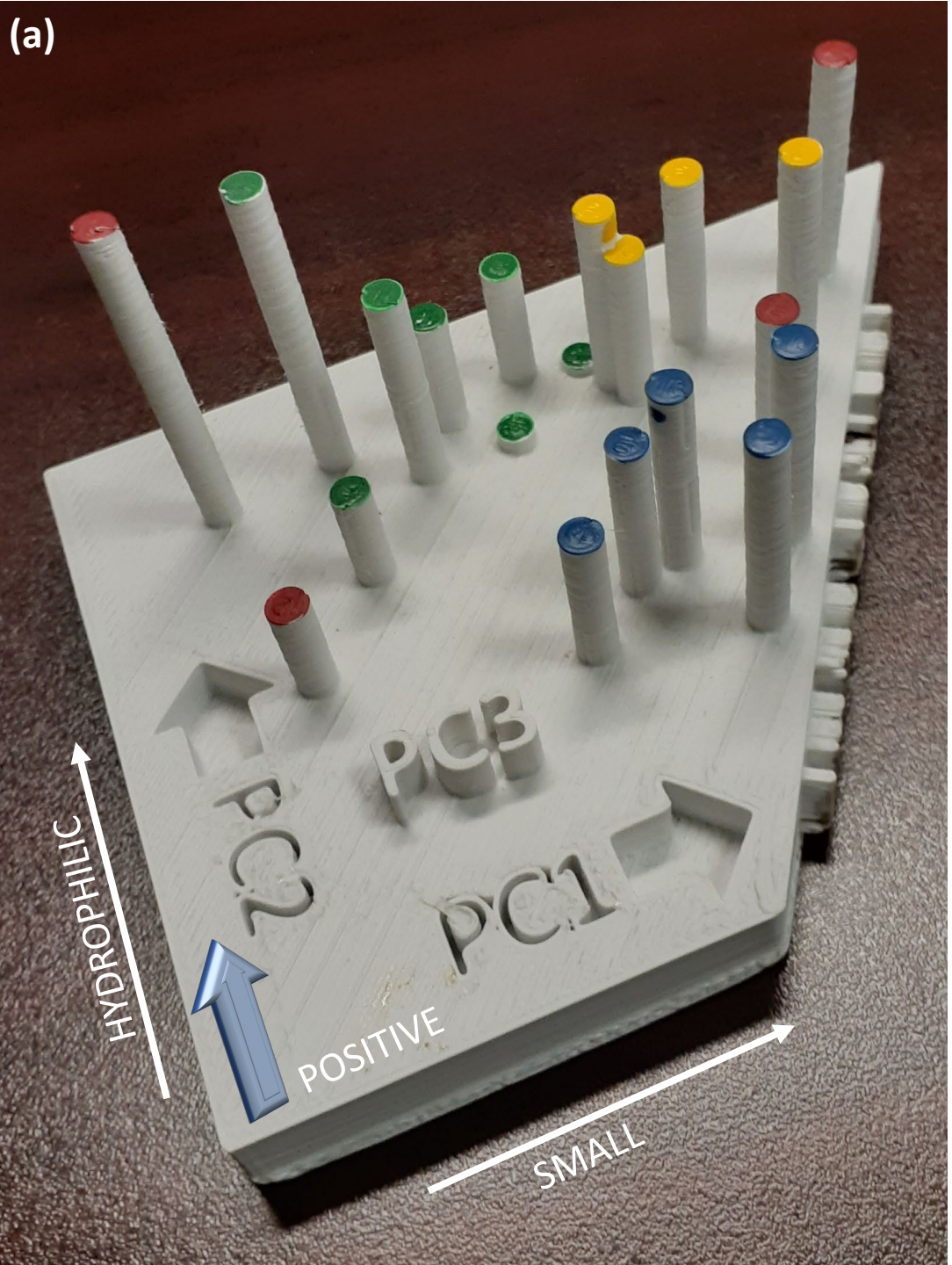
Table 1: Amino acid names, abbreviations, and PCA values. For underlying data, see [ref 32]

Table 2: Number of mutations connecting each substitution pair.

References

[1] Thaler, Richard H., and Cass R. Sunstein. Nudge: Improving decisions about health, wealth, and happiness. Penguin, 2009.
[2] Zhu, Wen, and Stephen Freeland. "The standard genetic code enhances adaptive evolution of proteins." Journal of theoretical biology 239.1 (2006): 63-70.
[3] Morgens, David W., and Andre RO Cavalcanti. "An alternative look at code evolution: using non-canonical codes to evaluate adaptive and historic models for the origin of the genetic code." Journal of molecular evolution 76.1-2 (2013): 71-80.
[4] Judson, Olivia P., and Daniel Haydon. "The genetic code: what is it good for? An analysis of the effects of selection pressures on genetic codes." Journal of molecular evolution 49.5 (1999): 539-550.
[5] Knight, Christopher G., et al. "Global analysis of predicted proteomes: functional adaptation of physical properties." Proceedings of the National Academy of Sciences 101.22 (2004): 8390-8395.
[6] Dawkins, Richard. "The evolution of evolvability." On growth, form and computers (2003): 239-255.
[7] Koonin, Eugene V., and Artem S. Novozhilov. "Origin and evolution of the genetic code: the universal enigma." IUBMB life 61.2 (2009): 99-111.
[8] Knight, Robin D., Stephen J. Freeland, and Laura F. Landweber. "Rewiring the keyboard: evolvability of the genetic code." Nature Reviews Genetics 2.1 (2001): 49-58.
[9] Thaler, Richard H. “Nudge, not sludge.” Science 361 (2018): 431
[10] Kirschner, Marc, and John Gerhart. "Evolvability." Proceedings of the National Academy of Sciences 95.15 (1998): 8420-8427.
[11] Burch, Christina L., and Lin Chao. "Evolvability of an RNA virus is determined by its mutational neighbourhood." Nature 406.6796 (2000): 625-628.
[12] Hsieh, C.L. et al., Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. Science 369, 1501–1505 (2020)
[13] Sikosek, Tobias, and Hue Sun Chan. "Biophysics of protein evolution and evolutionary protein biophysics." Journal of The Royal Society Interface 11.100 (2014): 20140419.
[14] Katsnelson, Mikhail I., Yuri I. Wolf, and Eugene V. Koonin. "On the feasibility of saltational evolution." Proceedings of the National Academy of Sciences 116.42 (2019): 21068-21075.
[15] Taleb, N. N., et al. "The precautionary principle (with application to the genetic modification of organisms). Extreme risk initiative—NYU school of engineering working paper series." (2014).
[16] Jacob, François. "Evolution and tinkering." Science 196.4295 (1977): 1161-1166.
[17] Pines, Gur, et al. "Refactoring the genetic code for increased evolvability." MBio 8.6 (2017).
[18] Koonin, Eugene V., and Artem S. Novozhilov. "Origin and evolution of the universal genetic code." Annual review of genetics 51 (2017): 45-62.
[19] Thaler, Richard H., Cass R. Sunstein, and John P. Balz. "Choice architecture." The behavioral foundations of public policy (2013): 428-439.
[20] Rorick, Mary M., and Günter P. Wagner. "Protein structural modularity and robustness are associated with evolvability." Genome biology and evolution 3 (2011): 456-475.
[21] Kassen, Rees. "Experimental evolution of innovation and novelty." Trends in ecology & evolution 34.8 (2019): 712-722.
[22] Zhu, Wen, and Stephen Freeland. "The standard genetic code enhances adaptive evolution of proteins." Journal of theoretical biology 239.1 (2006): 63-70.
[23] Pines, Gur, et al. "Refactoring the genetic code for increased evolvability." MBio 8.6 (2017).
[24] Freeland, Stephen J. "The Darwinian genetic code: an adaptation for adapting?" Genetic Programming and Evolvable Machines 3.2 (2002): 113-127.
[25] Tripathi, Shubham, and Michael W. Deem. "The standard genetic code facilitates exploration of the space of functional nucleotide sequences." Journal of molecular evolution 86.6 (2018): 325-339.
[26] Firnberg, Elad, and Marc Ostermeier. "The genetic code constrains yet facilitates Darwinian evolution." Nucleic acids research 41.15 (2013): 7420-7428.
[27] Harms, Michael J., and Joseph W. Thornton. "Evolutionary biochemistry: revealing the historical and physical causes of protein properties." Nature Reviews Genetics 14.8 (2013): 559-571.

[28] Hammerling, Michael J., et al. "Expanded genetic codes create new mutational routes to rifampicin resistance in Escherichia coli." Molecular biology and evolution 33.8 (2016): 2054-2063.
[29] Hamasaki-Katagiri, Nobuko, et al. "A gene-specific method for predicting hemophilia-causing point mutations." Journal of molecular biology 425.21 (2013): 4023-4033.
[30] Wang, Mingjun, et al. "FunSAV: predicting the functional effect of single amino acid variants using a two-stage random forest model." PloS one 7.8 (2012): e43847
[31] Li, Xidan, et al. "PASE: a novel method for functional prediction of amino acid substitutions based on physicochemical properties." Frontiers in genetics 4 (2013): 21.
[32] Nemzer, Louis R. "A binary representation of the genetic code." Biosystems 155 (2017): 10-19.
[33] Nemzer, Louis R. "Shannon information entropy in the canonical genetic code." Journal of theoretical biology 415 (2017): 158-170.
[34] Henikoff, Steven, and Jorja G. Henikoff. "Amino acid substitution matrices from protein blocks." Proceedings of the National Academy of Sciences 89.22 (1992): 10915-10919.
[35] Shapiro, Michael D., et al. "Genomic diversity and evolution of the head crest in the rock pigeon." Science 339.6123 (2013): 1063-1067.
[36] Bofkin, Lee, and Nick Goldman. "Variation in evolutionary processes at different codon positions." Molecular Biology and Evolution 24.2 (2007): 513-521.
[37] Shannon, Claude Elwood. "A mathematical theory of communication." The Bell system technical journal 27.3 (1948): 379-423.
[38] https://jbloomlab.github.io/SARS2_RBD_Ab_escape_maps/
[39] Greaney, Allison J., et al. "Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition." Cell host & microbe 29.1 (2021): 44-57.
[40] Greaney, Allison J., et al. "Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies." Cell host & microbe 29.3 (2021): 463-476.
[41] https://github.com/jbloomlab/SARSr-CoV_homolog_survey/blob/master/results/final_variant_scores/mut_variant_scores.csv
[42] Starr, Tyler N., et al. "ACE2 binding is an ancestral and evolvable trait of sarbecoviruses." bioRxiv (2021). https://doi.org/10.1101/2021.07.17.452804
[43] Bunn, H. Franklin. "The triumph of good over evil: protection by the sickle gene against malaria." Blood, The Journal of the American Society of Hematology 121.1 (2013): 20-25.
[44] Miyata, Takashi, Sanzo Miyazawa, and Teruo Yasunaga. "Two types of amino acid substitutions in protein evolution." Journal of Molecular Evolution 12.3 (1979): 219-236.
[45] Grantham, Richard. "Amino acid difference formula to help explain protein evolution." Science 185.4154 (1974): 862-864.
[46] Dong, Jinhui, et al. "Genetic and structural basis for recognition of SARS-CoV-2 spike protein by a two-antibody cocktail." BioRxiv (2021).
[47] Kupferschmidt, Kai. "Fast-spreading UK virus variant raises alarms." (2021): 9-10.
[48] Bailey, Susan F., Aaron Hinz, and Rees Kassen. "Adaptive synonymous mutations in an experimentally evolved Pseudomonas fluorescens population." Nature Communications 5.1 (2014): 1-7.
[49] Seligmann, Hervé, and David D. Pollock. "The ambush hypothesis: hidden stop codons prevent off-frame gene reading." DNA and cell biology 23.10 (2004): 701-705.
[50] Freeland, Stephen J., et al. "Early fixation of an optimal genetic code." Molecular biology and evolution 17.4 (2000): 511-518.
[51] Wagner, Günter P., and Lee Altenberg. "Perspective: complex adaptations and the evolution of evolvability." Evolution 50.3 (1996): 967-976.
[52] Blank, Diana, et al. "The predictability of molecular evolution during functional innovation." Proceedings of the National Academy of Sciences 111.8 (2014): 3044-3049.
[53] Weinreich, Daniel M., et al. "Darwinian evolution can follow only very few mutational paths to fitter proteins." science 312.5770 (2006): 111-114.
[54] Zhang, Wenjuan, et al. "Emergence of a novel SARS-CoV-2 variant in Southern California." Jama 325.13 (2021): 1324-1326.
[55] NCI Genomic Data Commons (GDC)


(a)
PC2
PC3
PC1
HYDROPHILIC
POSITIVE
SMALL

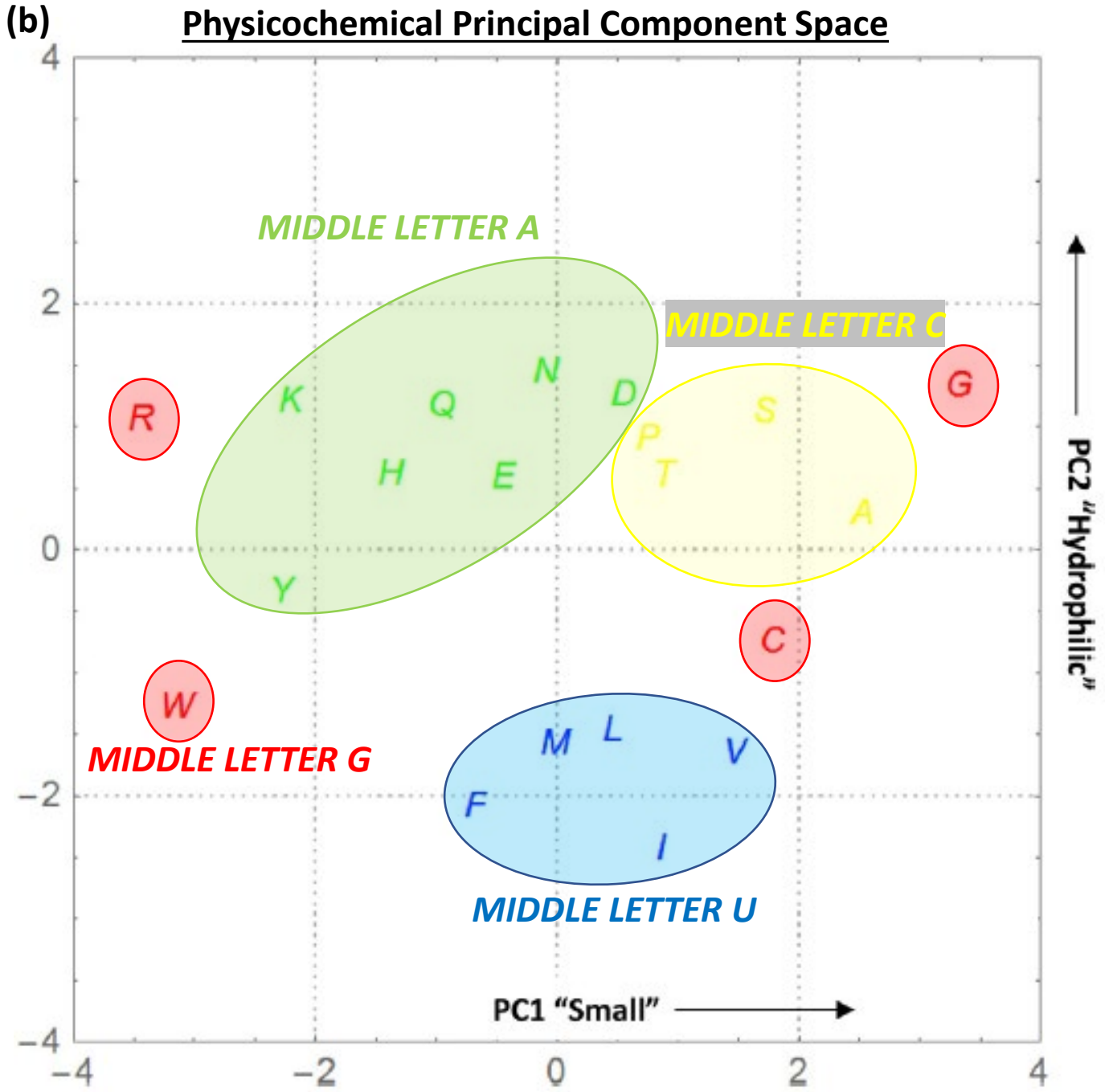

(b)
Physicochemical Principal Component Space
MIDDLE LETTER A
MIDDLE LETTER C
MIDDLE LETTER G
MIDDLE LETTER U
R
K
Q
N
D
H
E
Y
S
P
T
A
G
C
W
M
L
V
F
I
PC1 "Small"
PC2 "Hydrophilic"
4
2
0
−2
−4
−4
−2
0
2
4

Figure 1

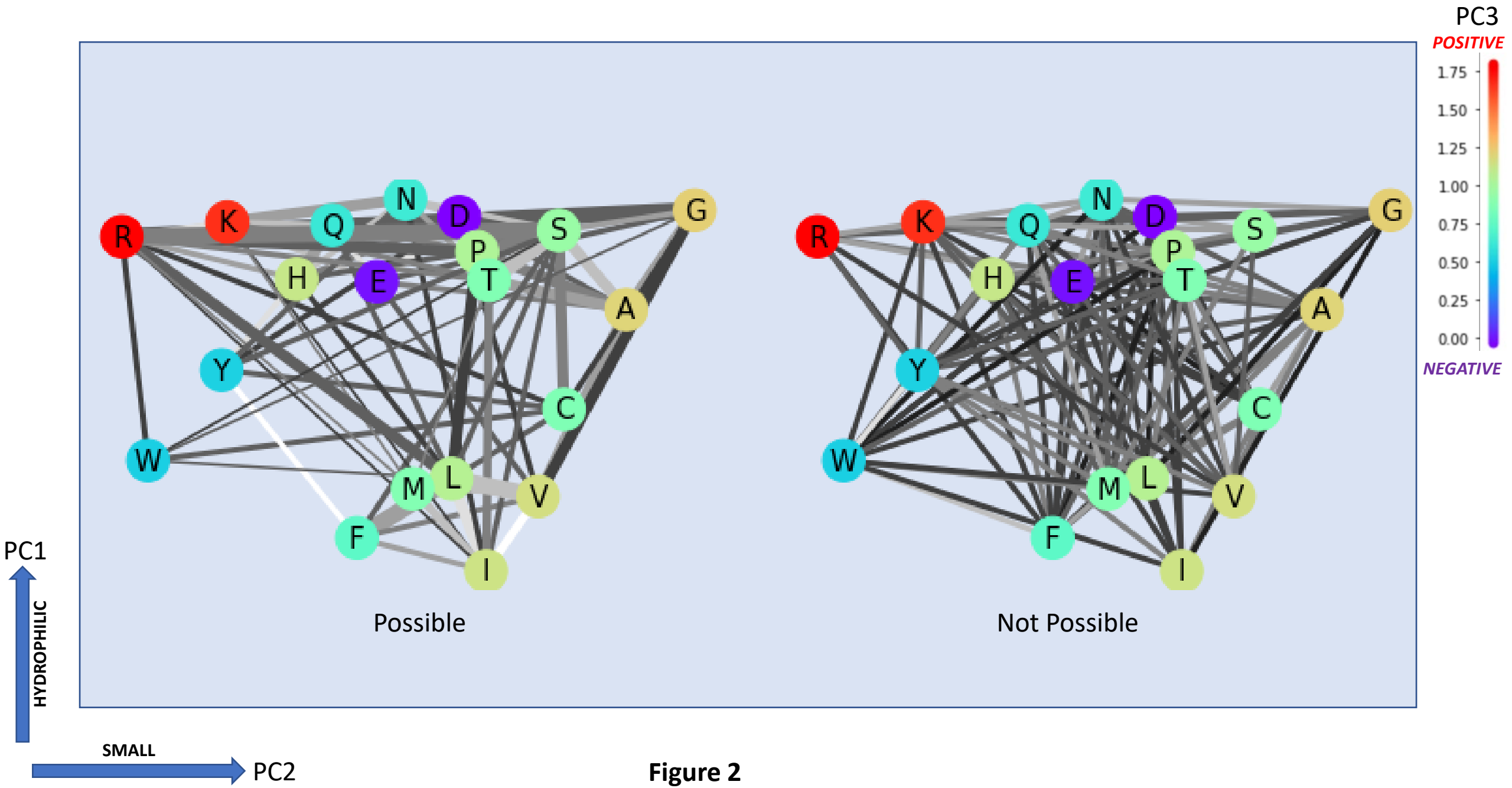

PC3
POSITIVE
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
NEGATIVE
R
K
Q
N
D
S
G
P
T
H
E
A
Y
C
W
M
L
V
F
I
Possible
Not Possible
PC1
HYDROPHILIC
SMALL
PC2

Figure 2

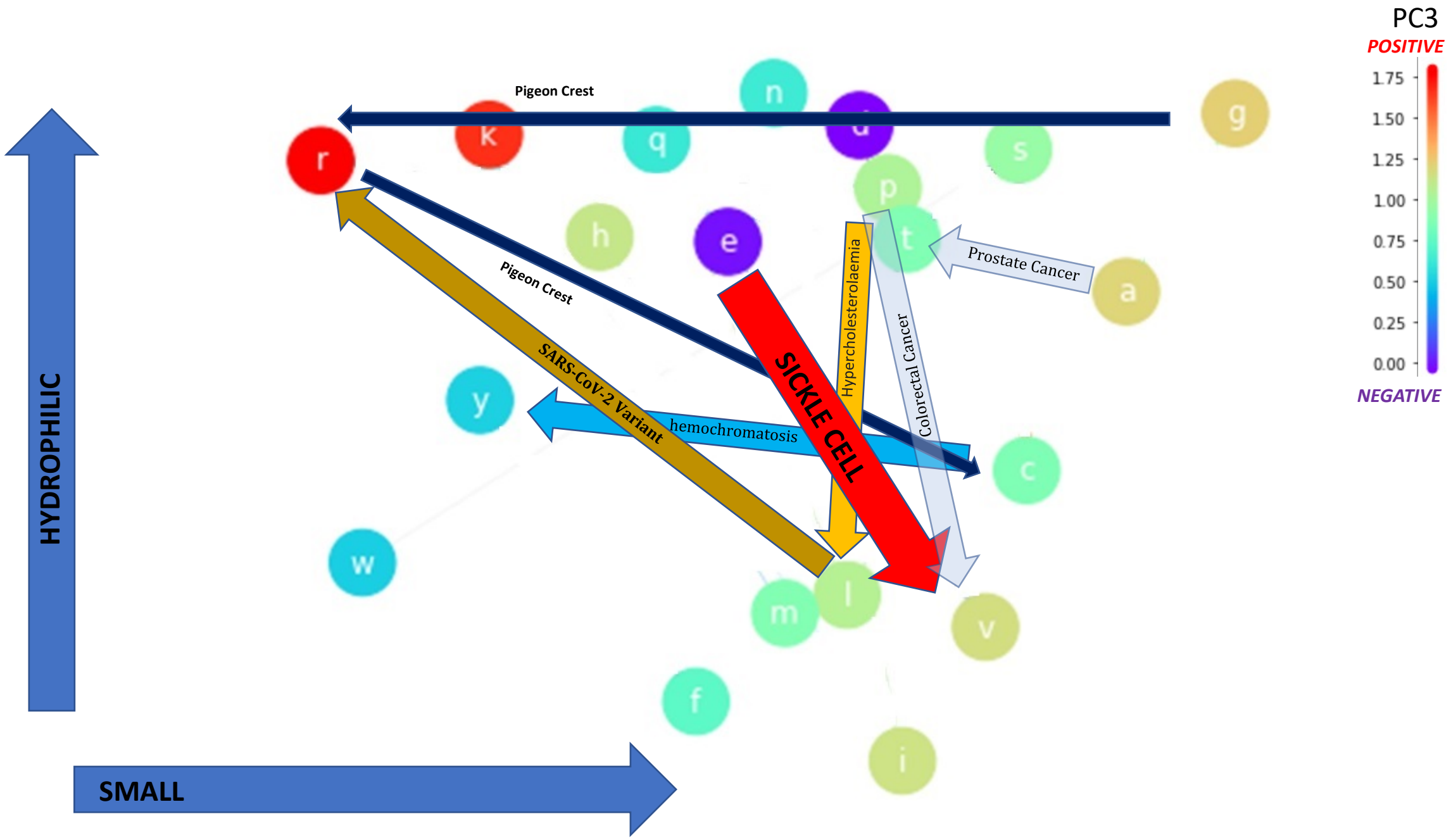
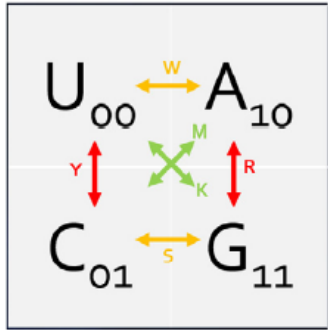
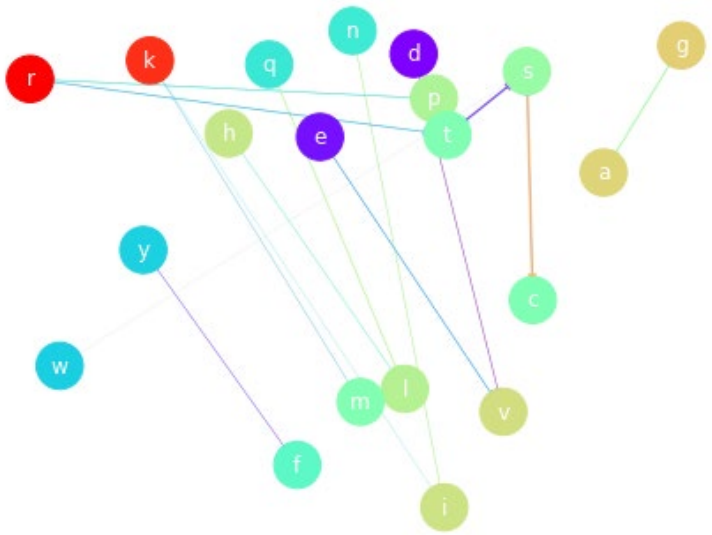

PC3
POSITIVE
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
NEGATIVE
HYDROPHILIC
SMALL
Pigeon Crest
Pigeon Crest
SARS-CoV-2 Variant
hemochromatosis
SICKLE CELL
Hypercholesterolaemia
Colorectal Cancer
Prostate Cancer
r
k
q
n
d
s
g
p
t
h
e
a
y
c
w
m
l
v
f
i

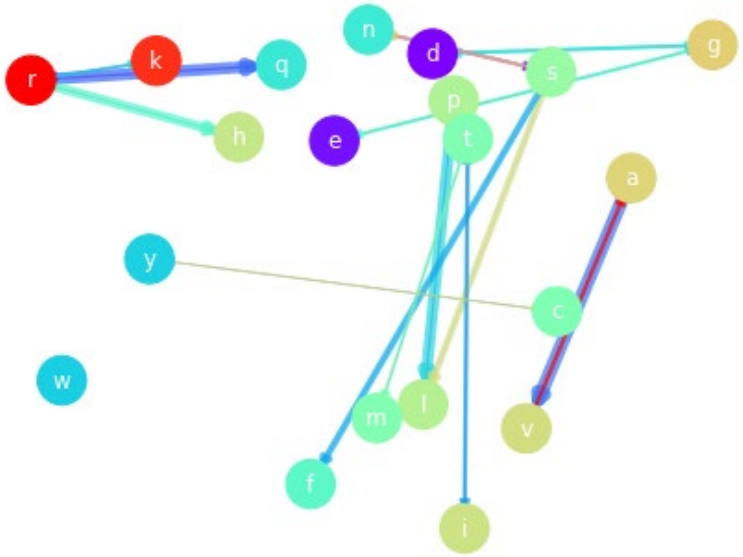
Figure 3



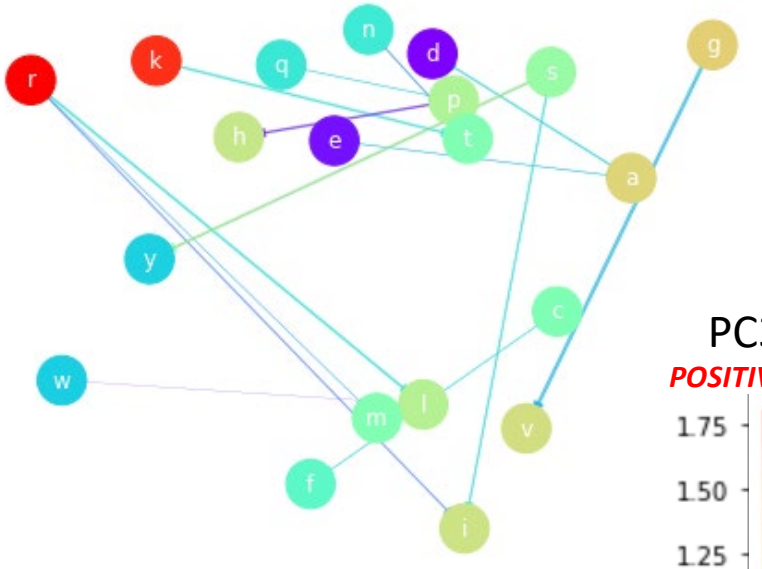
Middle Letter



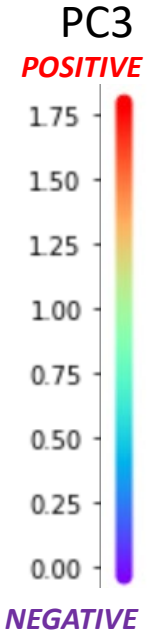
Keeps YR

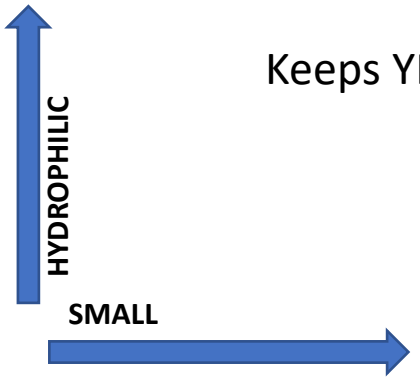


Keeps WS



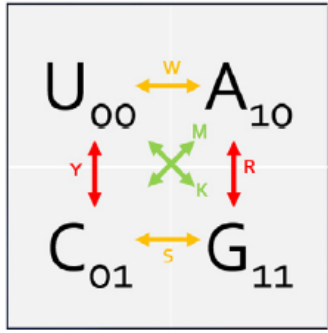
Keeps MK



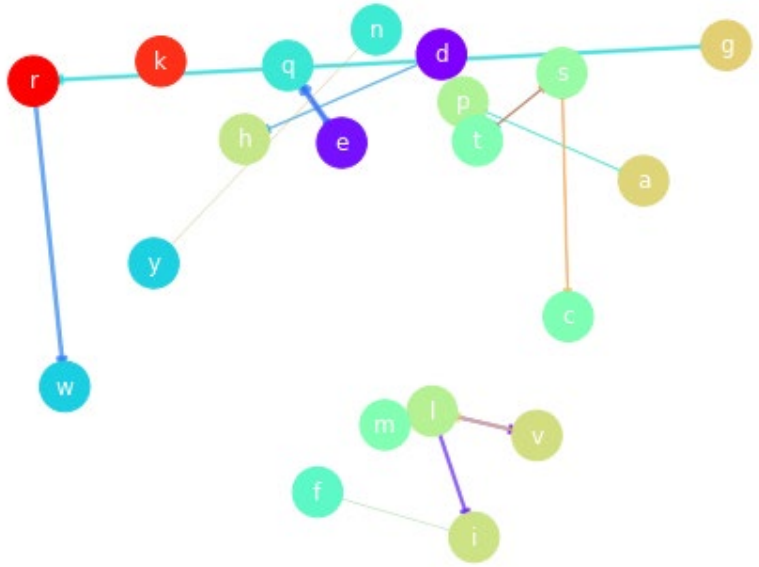


VERTICAL

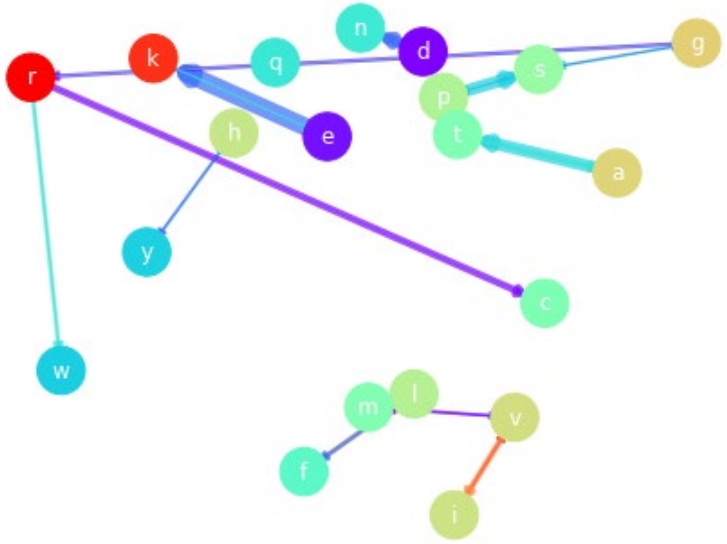
Figure 4(a)



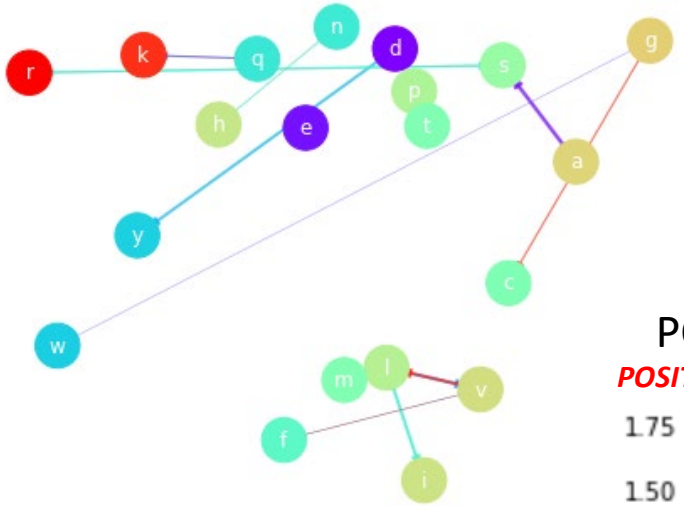
First Letter



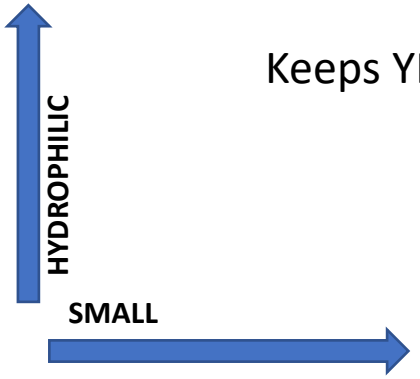
Keeps YR



Keeps WS



Keeps MK



HORIZONTAL

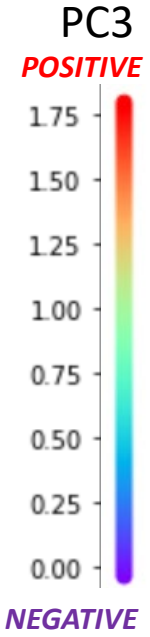
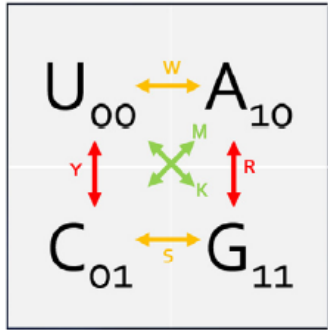


Figure 4(b)



Third Letter

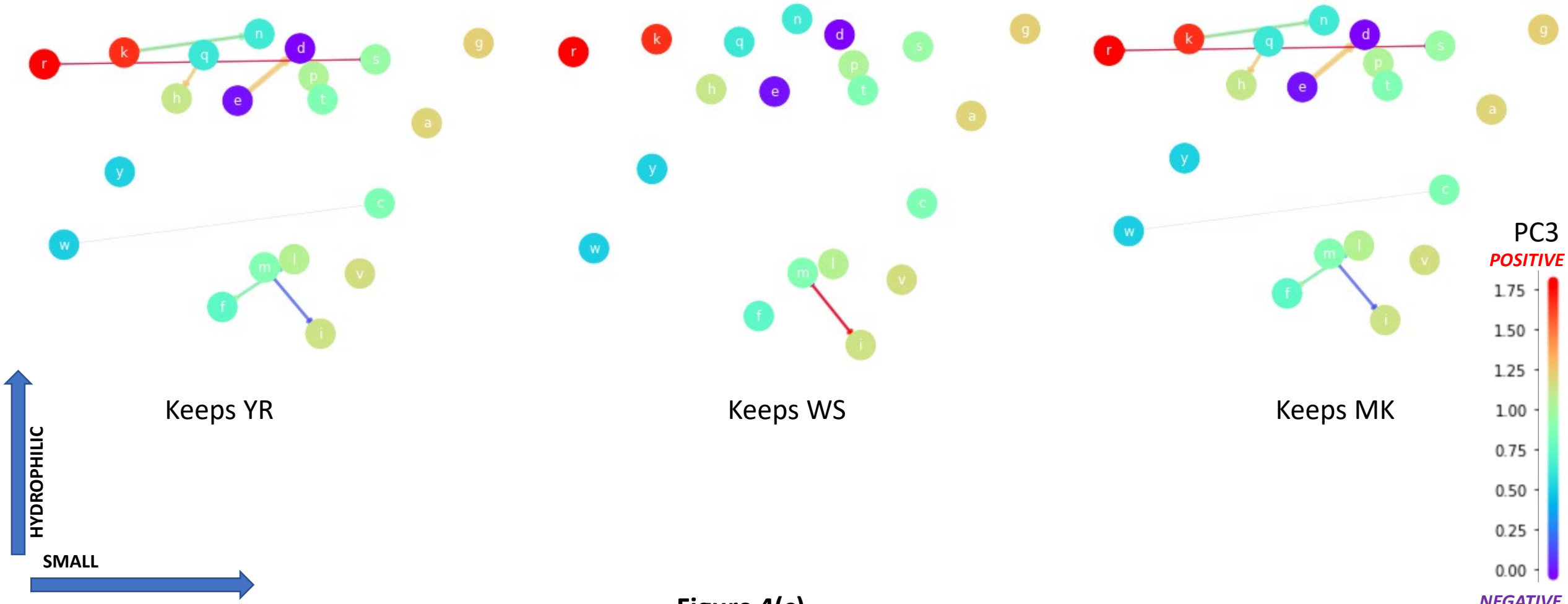
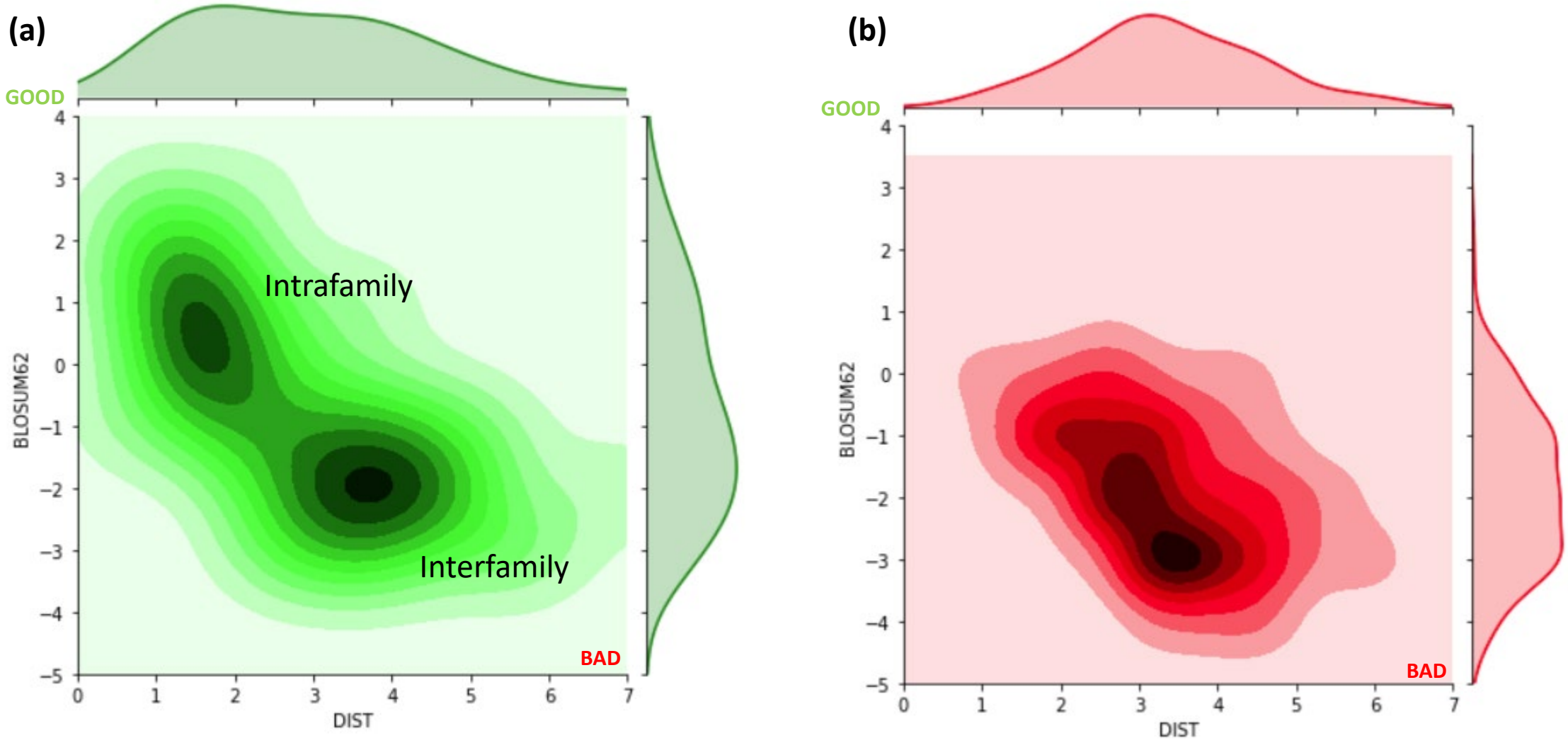


Figure 4(c)



POSSIBLE **NOT POSSIBLE**

Figure 5

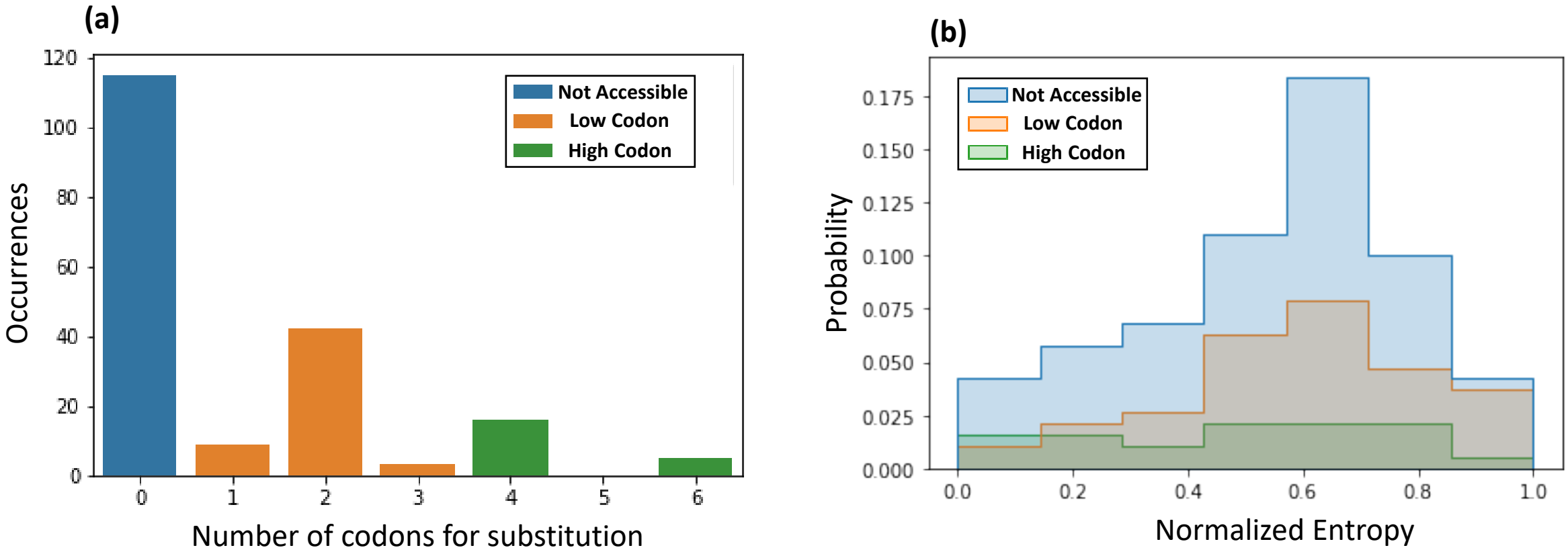

(a)
Occurrences
Number of codons for substitution
Not Accessible
Low Codon
High Codon
0
20
40
60
80
100
120
0
1
2
3
4
5
6
(b)
Probability
Normalized Entropy
Not Accessible
Low Codon
High Codon
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.0
0.2
0.4
0.6
0.8
1.0

Figure 6

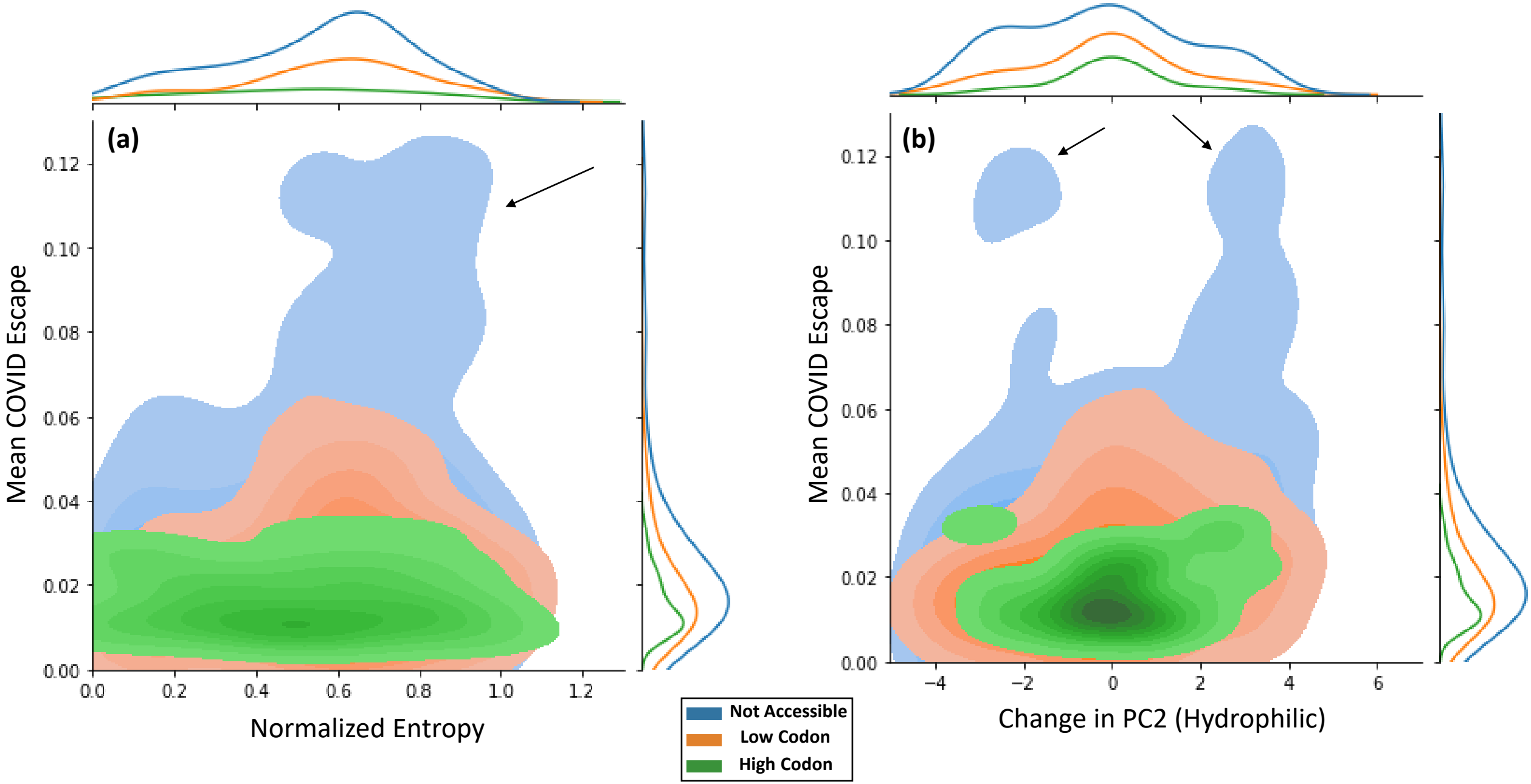
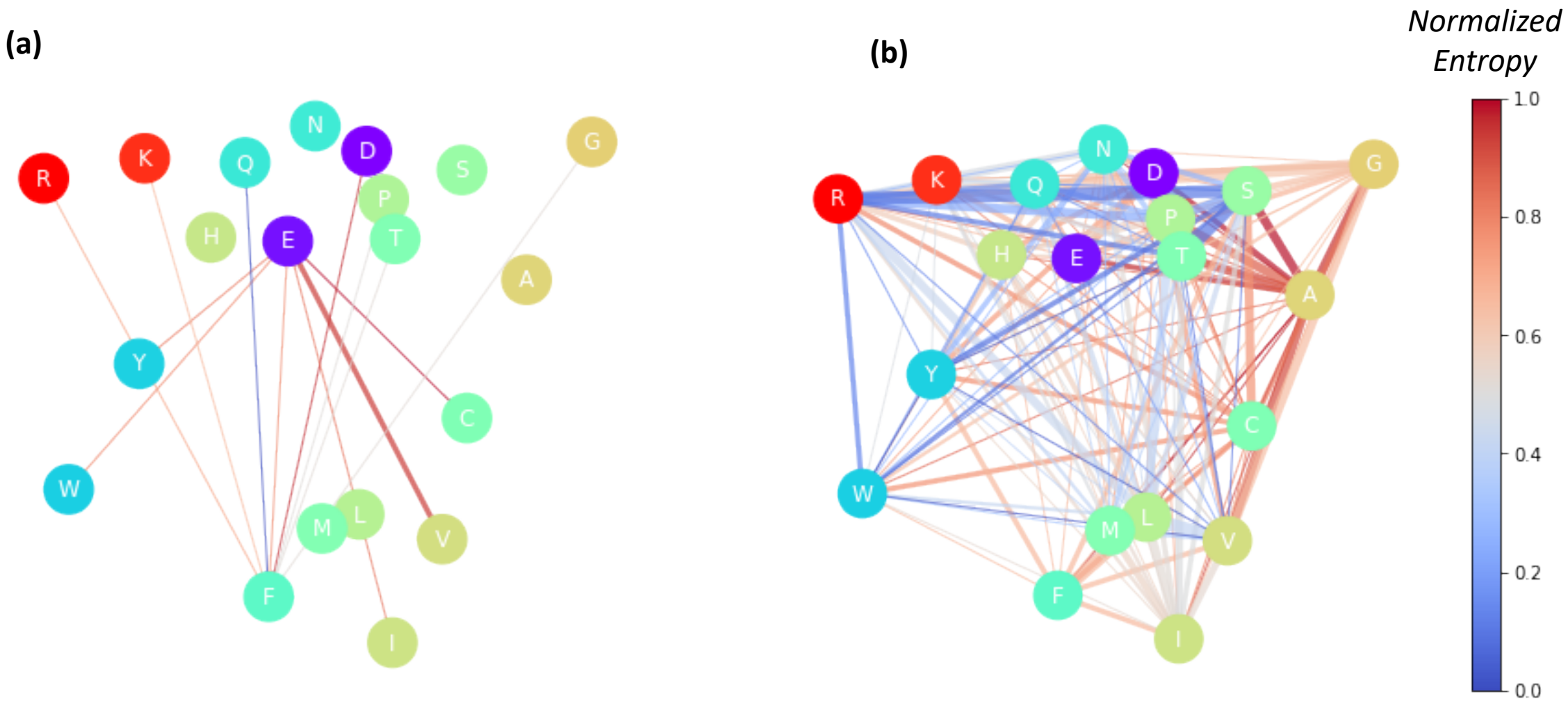

(a)
(b)
Mean COVID Escape
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Normalized Entropy
−4
−2
0
2
4
6
Change in PC2 (Hydrophilic)
Not Accessible
Low Codon
High Codon

Figure 7

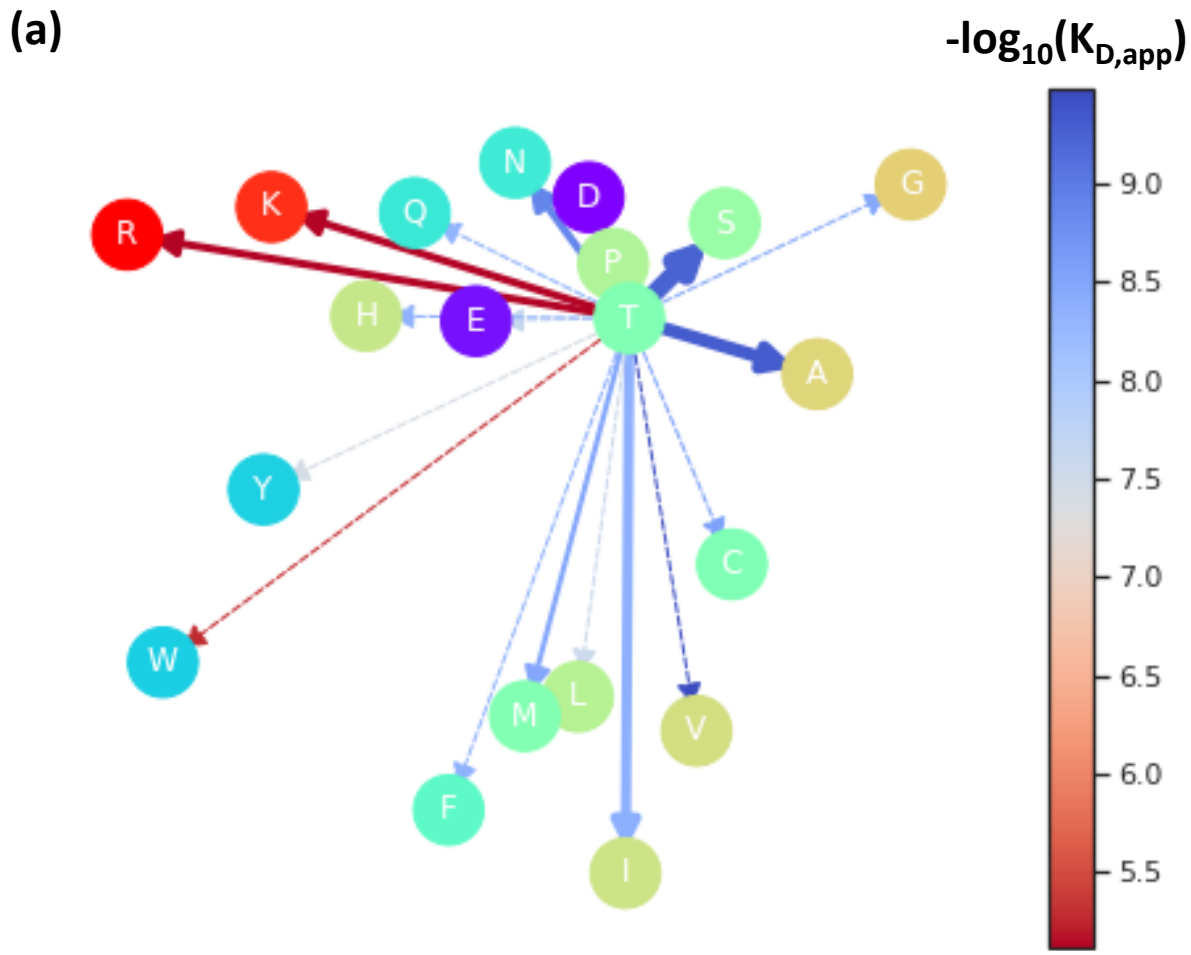


SARS-CoV-2
High Escape (>0.073)

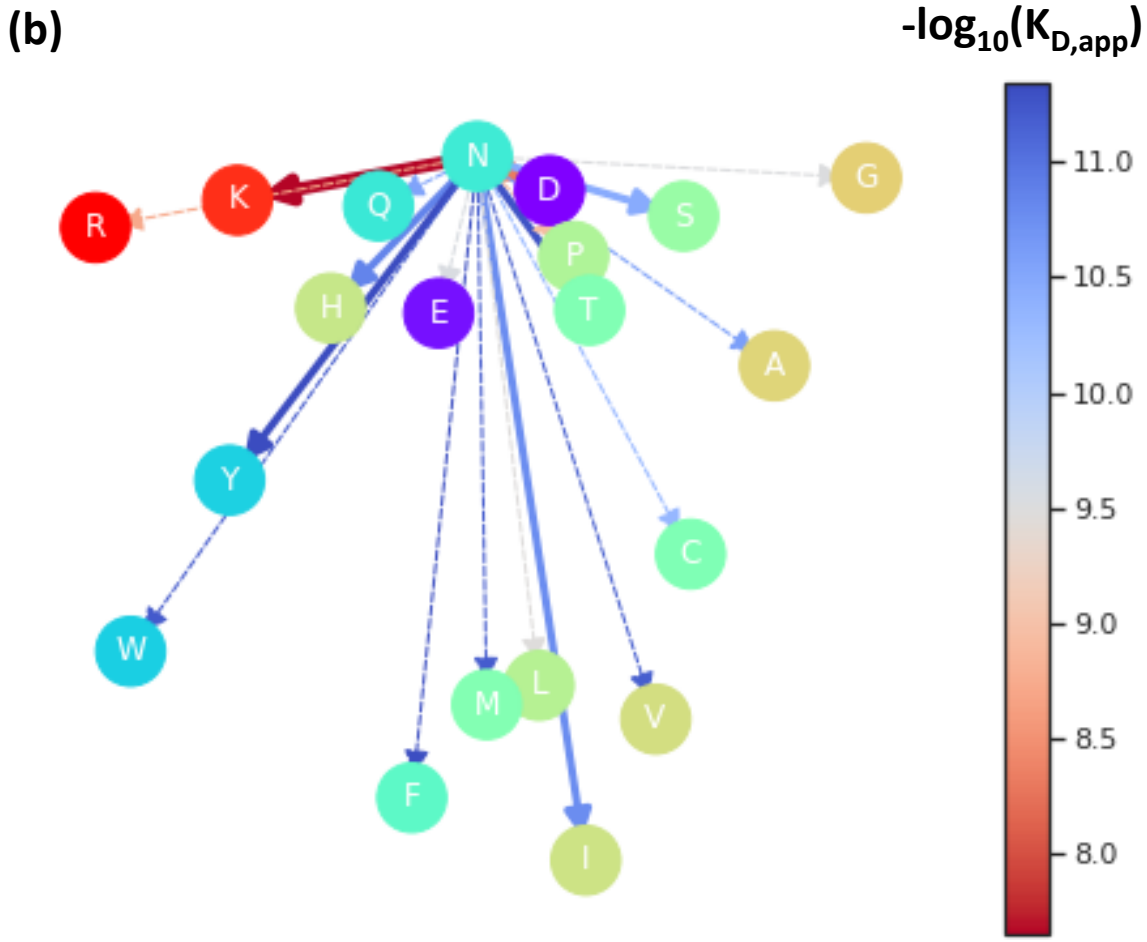
SARS-CoV-2
Low Escape (<0.073)

Figure 8

Binding with Human ACE2



SARS T501 (wt = 10.4)
SARS-CoV-1



COVID N501 (wt = 10.8)
SARS-CoV-2

Figure 9

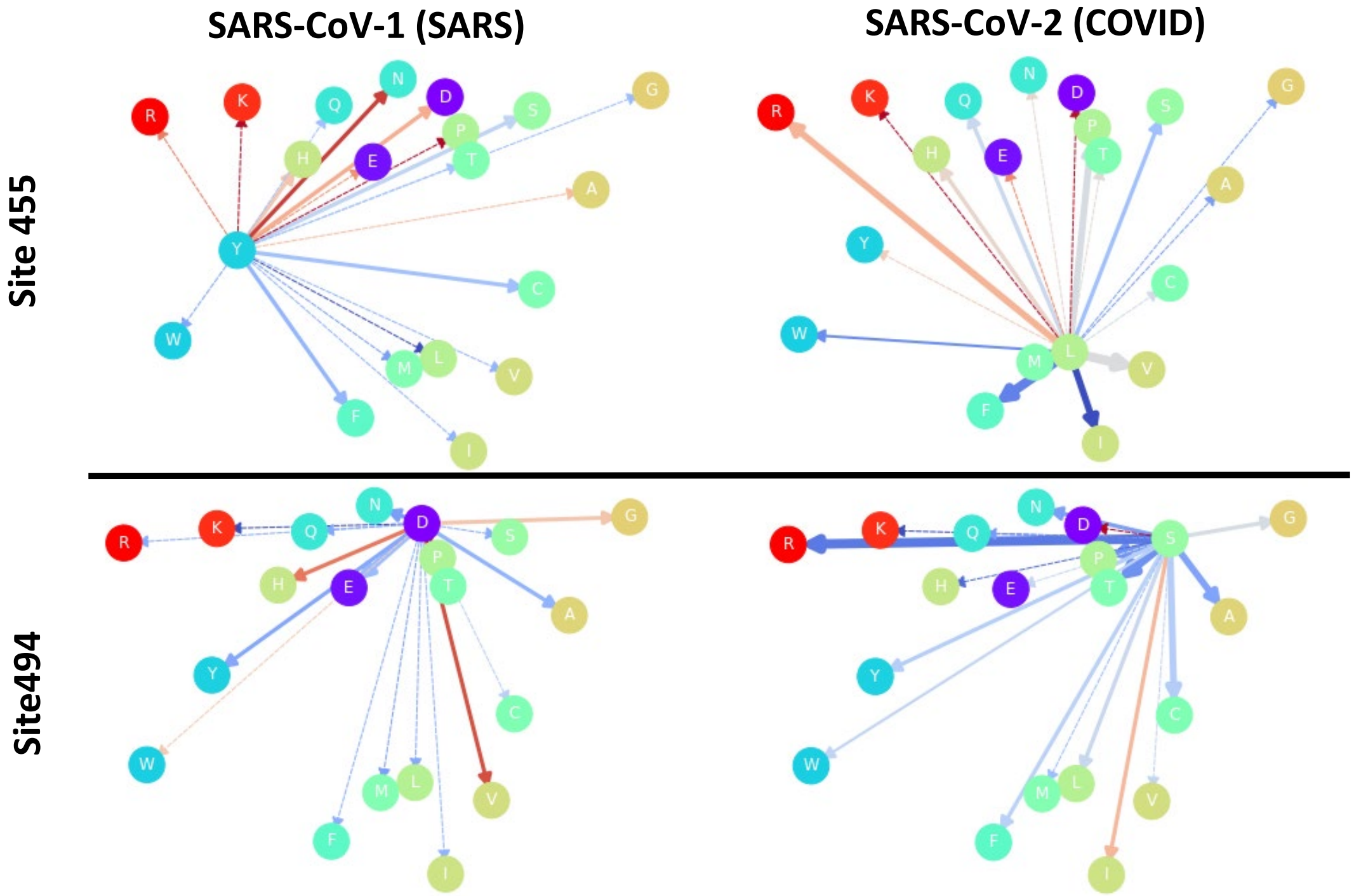

SARS-CoV-1 (SARS)
SARS-CoV-2 (COVID)
Site 455
Site494
R
K
Q
N
D
S
G
H
E
P
T
A
Y
C
W
M
L
V
F
I

Figure 10

Single Letter	Amino Acid	Abbr	PC1 ("Small")	PC2 ("Hydro")	PC3 ("Positive")
A	Alanine	Ala	2.53	0.294	0.623
C	Cysteine	Cys	1.794	-0.745	-0.067
D	Aspartic Acid	Asp	0.556	1.261	-1.911
E	Glutamic Acid	Glu	-0.418	0.584	-1.837
F	Phenylalanine	Phe	-0.654	-2.088	-0.306
G	Glycine	Gly	3.337	1.33	0.673
H	Histidine	His	-1.367	0.612	0.451
I	Isoleucine	Ile	0.874	-2.433	0.495
K	Lysine	Lys	-2.185	1.207	1.561
L	Leucine	Leu	0.464	-1.469	0.332
M	Methionine	Met	0.004	-1.575	-0.032
N	Asparagine	Asn	-0.079	1.449	-0.533
P	Proline	Pro	0.767	0.897	0.277
Q	Glutamine	Gln	-0.946	1.175	-0.563
R	Arginine	Arg	-3.431	1.056	1.792
S	Serine	Ser	1.732	1.119	0.127
T	Threonine	Thr	0.907	0.6	-0.057
V	Valine	Val	1.488	-1.655	0.542
W	Tryptophan	Trp	-3.123	-1.281	-0.795
Y	Tyrosine	Tyr	-2.253	-0.337	-0.771

Table 1

	a	c	d	e	f	g	h	i	k	l	m	n	p	q	r	s	t	v	w	y
a	12	0	2	2	0	4	0	0	0	0	0	0	4	0	0	4	4	4	0	0
c	0	2	0	0	2	2	0	0	0	0	0	0	0	0	2	4	0	0	2	2
d	2	0	2	4	0	2	2	0	0	0	0	2	0	0	0	0	0	2	0	2
e	2	0	4	2	0	2	0	0	2	0	0	0	0	2	0	0	0	2	0	0
f	0	2	0	0	2	0	0	2	0	6	0	0	0	0	0	2	0	2	0	2
g	4	2	2	2	0	12	0	0	0	0	0	0	0	0	6	2	0	4	1	0
h	0	0	2	0	0	0	2	0	0	2	0	2	2	4	2	0	0	0	0	2
i	0	0	0	0	2	0	0	6	1	4	3	2	0	0	1	2	3	3	0	0
k	0	0	0	2	0	0	0	1	2	0	1	4	0	2	2	0	2	0	0	0
l	0	0	0	0	6	0	2	4	0	18	2	0	4	2	4	2	0	6	1	0
m	0	0	0	0	0	0	0	3	1	2	0	0	0	0	1	0	1	1	0	0
n	0	0	2	0	0	0	2	2	4	0	0	2	0	0	0	2	2	0	0	2
p	4	0	0	0	0	0	2	0	0	4	0	0	12	2	4	4	4	0	0	0
q	0	0	0	2	0	0	4	0	2	2	0	0	2	2	2	0	0	0	0	0
r	0	2	0	0	0	6	2	1	2	4	1	0	4	2	18	6	2	0	2	0
s	4	4	0	0	2	2	0	2	0	2	0	2	4	0	6	14	6	0	1	2
t	4	0	0	0	0	0	0	3	2	0	1	2	4	0	2	6	12	0	0	0
v	4	0	2	2	2	4	0	3	0	6	1	0	0	0	0	0	0	12	0	0
w	0	2	0	0	0	1	0	0	0	1	0	0	0	0	2	1	0	0	0	0
y	0	2	2	0	2	0	2	0	0	0	0	2	0	0	0	2	0	0	0	2

Table 2