Results

TMPRSS2 has 14 exons and three isoforms (Figure 1). The promoter extends past the first exon. Isoform 1 has 529 amino acids, isoform 2 has 492 amino acids, and isoform 3 has 498 amino acids. Specifically, isoform 2 excludes exon 1. Regions that have been located include the promoter, enhancer, and the cis regulatory region. No additional information about isoform 3 was discovered.

Only nonsynonymous SNPs of TMPRSS2 were examined due to the possibility of structural changes in the shape of the protein. Of the 18 most common SNPs according to ALFA frequencies on dbSNP, the most common SNP is rs75603675, with a total frequency of 0.30337 on the alternate allele A (Table 1). This SNP occurs in exon one and is not observed in isoform 2. Only one other SNP is not observed in isoform 2, rs200291871. It was discovered that the occurrence of nonsynonymous SNPs are rare. Only 2 SNPs (rs75603675 and rs12329760) occur in frequencies larger than 0.1 (Table 1) (Figure 2). The other 16 SNP frequencies range from 10e^-5 to 10e^-3 (Table 1) (Figure 2). Global frequencies indicate that in European and African populations (Table 1). Only a small number of SNPs were found to be potentially damaging or deleterious to the function of the protein according to SIFT and PolyPhen-2. There are five SNPs that could be possibly damaging according to PolyPhen2. While SIFT determined 6 deleterious variations. The discrepancy between the two servers was rs1422446494 (Table 2). However, PredictProtein, determined that rs142244694 is probably not very damaging to the protein, as it did not meet their criteria for possible damage. The effects of point mutations at the specific amino acids for these SNPs are visualized in Figure 3.

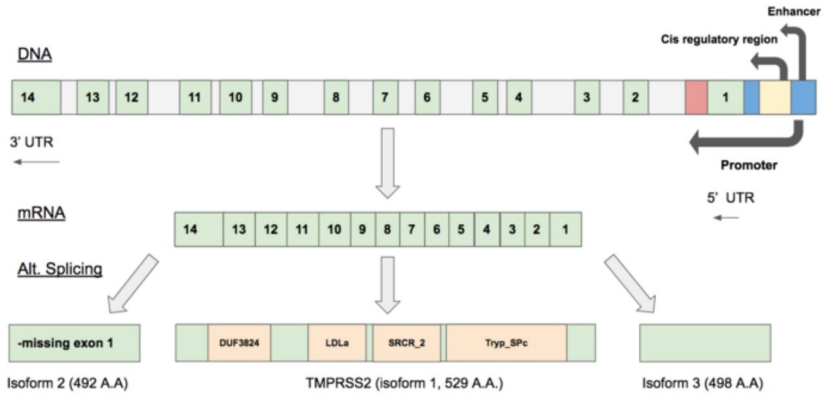

Figure 1. Gene Map displaying regulatory regions of TMPRSS2 and known isoforms. Isoform 2 excludes exon 1.

Table 1. Global frequency data obtained from NCBI dbSNP. Amino acid changes, sample size, and frequencies from European, African, Asian, Latin America, are listed. Total frequencies are listed in the rightmost column. Frequency data was compiled by NCBI from multiple databases including: ALFA, ExAC, gnomAD, GO exome sequencing project, and the PAGE study. Latin American 1 indicates those with Afro-Caribbean ancestry, while Latin American 2 indicated those with native ancestry. Asterisks indicate SNPs not found in isoform 2.

SNP ID	Amino Acid Change	Sample Size	European	African	Asian	Latin American 1	Latin American 2	Total Frequency
rs75603675*	Gly8Arg	17,922	C=0.65871 A=0.34129	C=0.9705 A=0.0295	C=1.0 A=0.0	C=1.00 A=0.00	C=1.0 A=0.0	C=0.69663 A=0.30337
rs12329760	Val160Met	295,780	C=0.779795 T=0.220205	C=0.70975 T=0.29025	C=0.6092 T=0.3908	C=0.7617 T=0.2383	C=0.8537 T=0.1463	C=0.775607 T=0.224393
rs61735793	Thr75Ile	191,500	G=0.989476 A=0.010524	G=0.9994 A=0.0006	G=1.0 A=0.0	G=0.998 A=0.002, C=0.000	G=0.9962 A=0.0038	G=0.990381 A=0.009619
rs200291871*	Gly8Arg	18,890	C=0.9887 G=0.013	C=0.9976 G=0.0024	C=1.0 G=0.0	C=1.000 G=0.000	C=1.0 G=0.0	C=0.99105 G=0.00895
rs61735791	Ala28Thr	203,412	C=0.996771 T=0.003229	C=0.9996 T=0.0004	C=0.9992 T=0.0008	C=1.000 T=0.000	C=0.999 T=0.001	C=0.996952 T=0.003048
rs148125094	Val415Ile	203,772	C=0.998865 T=0.001135	C=1.0 T=0.0	C=1.0 T=0.0	C=1.000 T=0.000	C=1.0 T=0.0	C=0.998955 T=0.001045
rs142446494	Val280Met	44,790	C=0.99939 T=0.00061	C=1.0 T=0.0	C=1.0 T=0.0	C=1.000 T=0.000	C=1.0 T=0.0	C=0.99929 T=0.00071
rs61735796	Glu260Lys	49,254	C=0.99919 T=0.00081	C=1.0 T=0.0	C=1.0 T=0.0	C=1.0 T=0.0	C=1.0 T=0.0	C=0.99933 T=0.00067
rs150554820	Phe209Ile	49,260	A=0.99928 T=0.00072	A=0.9992 T=0.0008	A=1.0 T=0.0	A=1.0 T=0.0	A=1.0 T=0.0	A=0.99933 T=0.00067
rs138651919	Pro41Leu	199,290	G=0.999588	G=0.9996	G=0.9997	G=1.0	G=1.0	G=0.999609

			A=0.000412	A=0.0004	A=0.0003	A=0.0	A=0.0	A=0.000391
rs61735790	His18Arg	199,596	T=0.999965 C=0.000035	T=0.9890 C=0.0110	T=1.0 C=0.0	T=0.991 C=0.009	T=1.0 C=0.0	T=0.999614 C=0.000386
rs768173297	Thr309Met	44,404	G=0.99966 A=0.00034	G=1.0 A=0.0	G=1.0 A=0.0	G=1.0 A=0.0	G=1.0 A=0.0	G=0.99973 A=0.00027
rs61735795	Pro375Ser	78,726	G=1.0 A=0.0	G=0.9963 A=0.0037	G=1.000 A=0.000	G=1.0 A=0.0	G=1.0 A=0.0	G=0.99982 A=0.00018
rs201093031	Val33Ala	58,202	A=0.99992 G=0.00008	A=1.0 G=0.0	A=0.994 G=0.006	A=1.0 G=0.0	A=1.0 G=0.0	A=0.99991 G=0.00009
rs147711290	Leu91Gln	107770	A= 0.99997 T=0.00000	A=0.9986 T=0.0012	A=1.000 T=0.000	A=0.999 T=0.001	A=1.000 T=0.000	A=0.999879 T=0.000074
rs114363287	Gly74Arg	199,516	C=0.999994 T=0.000006	C=0.9982 T=0.0018	C=1.0 T=0.0	C=0.998 T=0.002	C=1.0 T=0.0	C=0.999930 T=0.000070
rs147711290	Leu91Pro	107770	A= 0.99997 G=0.00003	A=0.9986 G=0.0002	A=1.0 G=0.0	A=0.999 G=0.0	A=1.0 G=0.0	A= 0.999879 G=0.000046

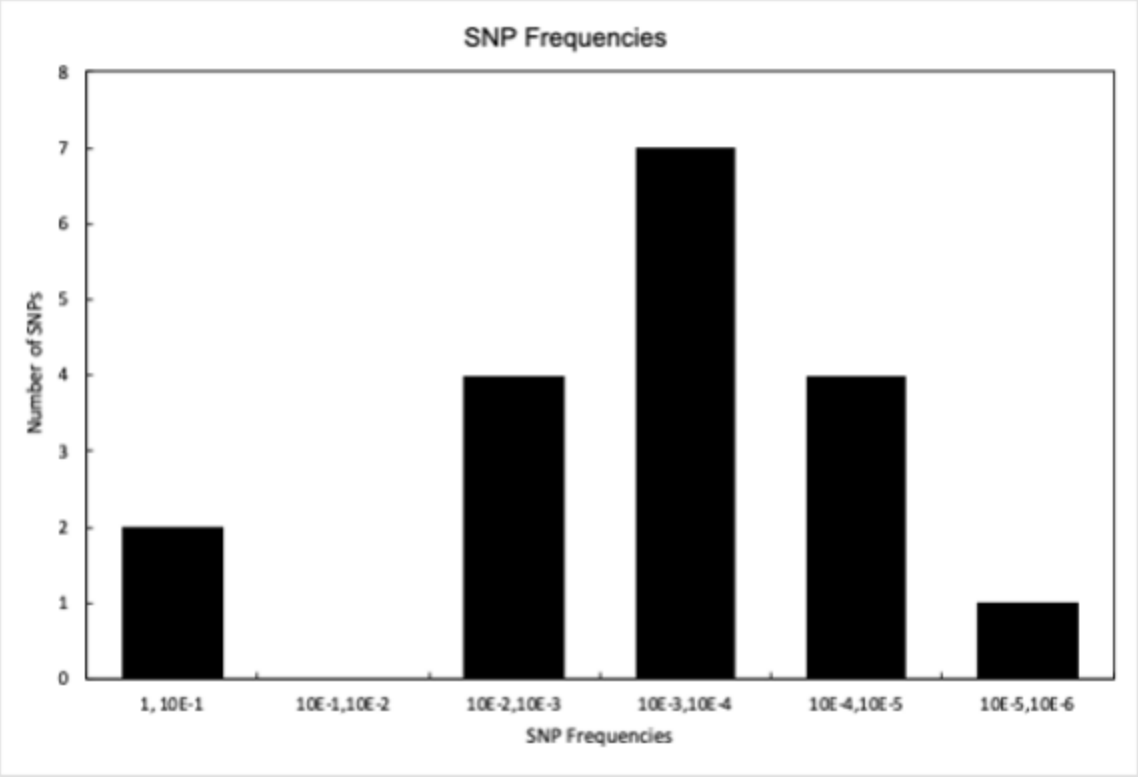

Figure 2. Histogram displaying the number of SNPs per set of frequencies. This will change to display the log bases from 10e^-6 to 1. Frequencies were obtained from dbSNP ALFA.

Table 2. SIFT and Polyphen-2 scores and predictions for chosen SNPs. Include criteria for scoring and predictions here.

TMPRSS2 SNP Predictions

rs Number	SIFT score	SIFT prediction	PolyPhen-2 score	PolyPhen-2 prediction
rs61735793	0.238	tolerated	0.015	Benign
rs75603675 G8V	0.201	tolerated	0.167	Benign
rs61735790	0.231	tolerated	0.033	Benign
rs12329760	0.009	deleterious	0.937	Probably Damaging
rs200291871	0.817	Tolerated	0.011	Benign
rs61735791	0.199	Tolerated	0.029	Benign
rs148125094	0.171	Tolerated	0.098	Benign
rs114363287	0.383	Tolerated	0.109	Benign
rs147711290 L128G	Not Found	-	0.920	Probably Damaging
rs147711290 L91P	0.005	Deleterious	1.000	Probably Damaging
rs147711290 L91R	Not Found	-	Not Found	-
rs150554820	0.004	Deleterious	0.549	Possibly Damaging
rs61735796	0.34	Tolerated	0.017	Benign
rs138651919	0.021	Deleterious	0.833	Possibly Damaging
rs61735795	0.551	Tolerated	0.086	Benign
rs142446494	0.015	Deleterious	0.294	Benign
rs201093031	1	Tolerated	0.00	Benign
rs768173297	Not Found	-	0.131	Benign

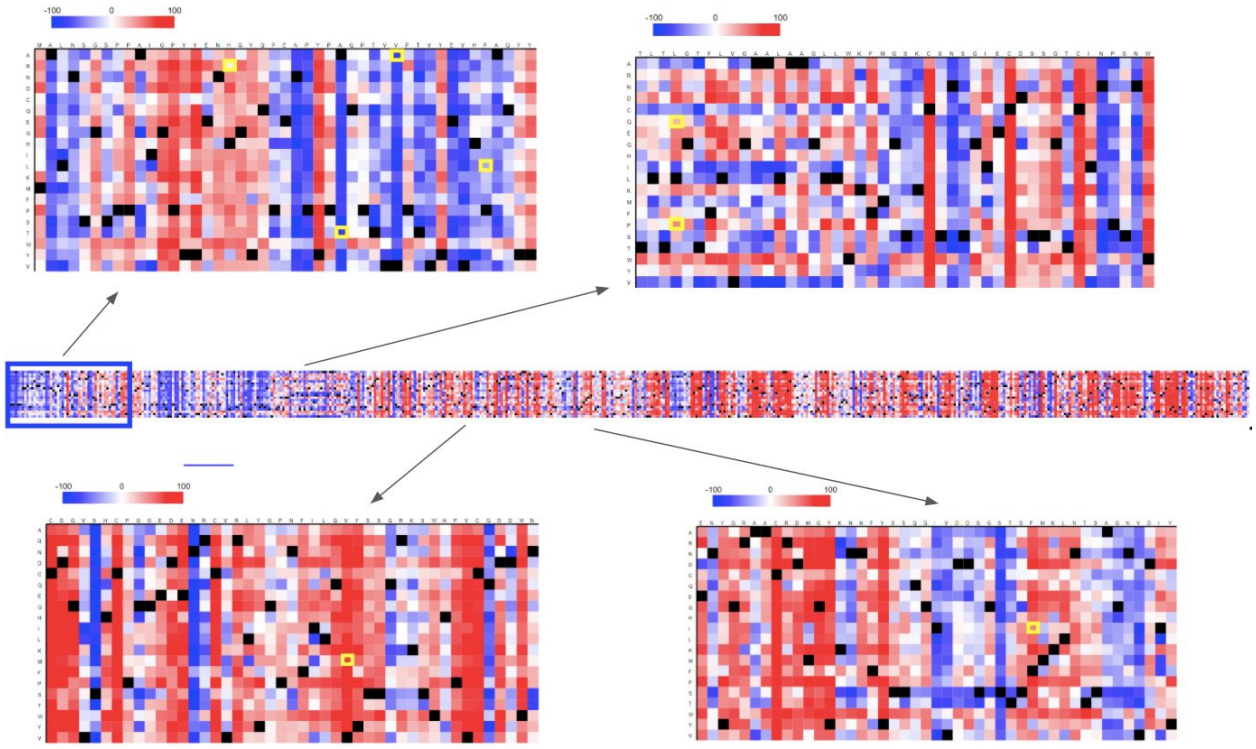

Figure 3. Heat map of the effect of point mutations in TMPRSS2. All 16 SNPs on isoform 2 were highlighted in yellow boxes. Only four panels are shown here. The four panels indicate damaging SNPs and do not account for most of the non-damaging and benign effects. Red shades indicate damaging effects, while blue shades indicate benign effects. Intensity of the colors indicates the severity of these mutations.

Due to the more extensive research done on isoform 2, the isoform 2 sequence was used to generate a predicted model for TMPRSS2, since no crystal structure has been confirmed. Several programs were used to generate models and compared through iCn3D including: SwissModel, HHPred, RaptorX, and I-TASSER.

The overall shape of the protein is globular, and appears to have more beta sheets compared to alpha helices. The numbers of each vary between each protein server. Each server matched TMPRSS2 with hepsin to build the general structures shown. Two programs, SwissModel (Figure 4) and HHPred (Figures 5) provided faulty structures that omitted parts of the sequence from their model of TMPRSS2. RaptorX (Figure 6) and I-TASSER (Figure 7)

generated complete structures. The RaptorX model appeared longer, compared to the wider, shorter structures generated by the other servers. Since I-TASSER is highly researched and included in many primary research articles, and showed the most consistent results of the four protein servers, we decided upon I-TASSER as the best model to dock TMPRSS2 with SARS-CoV-2.

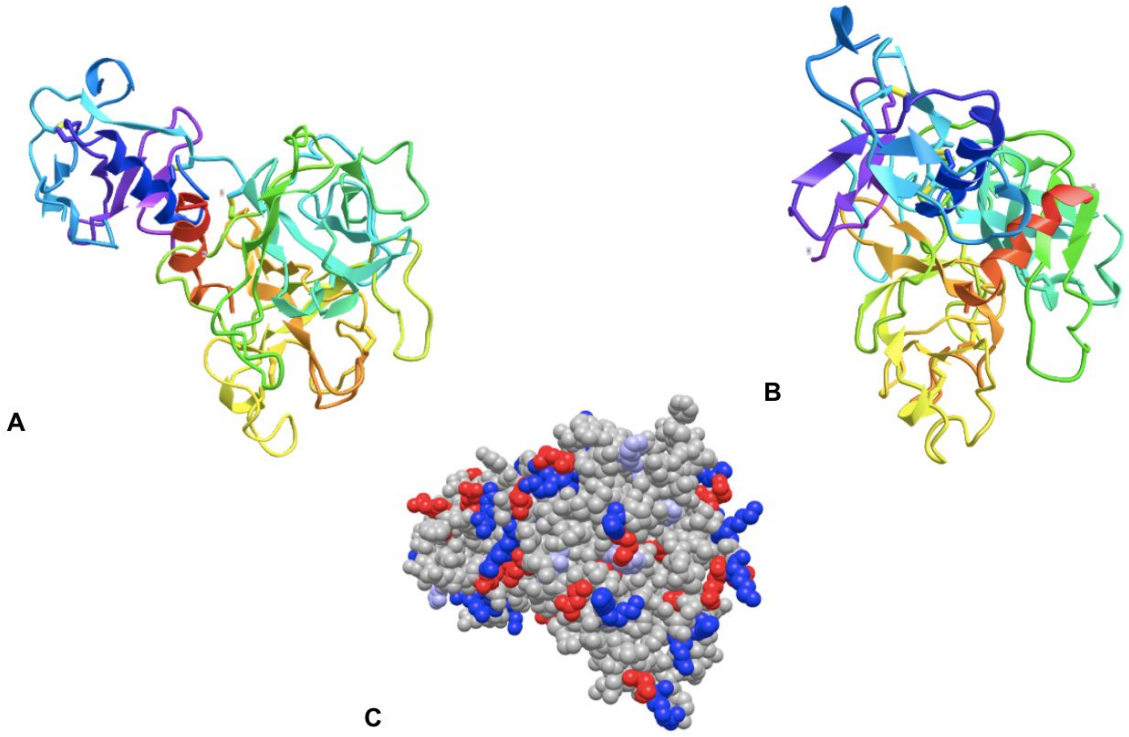

Figure 4. SwissModel structure for TMPRSS2. The original view (A), 180 degrees rotated around the Y axis (B) and space-filling model (C) are shown here. Coloration of (A) and (B) is spectrum gradient based on proximity to n or c terminii. Coloration of the space-filling model (C) is done using charge. 11 alpha helix sets and 20 beta sheets were predicted by SwissModel.

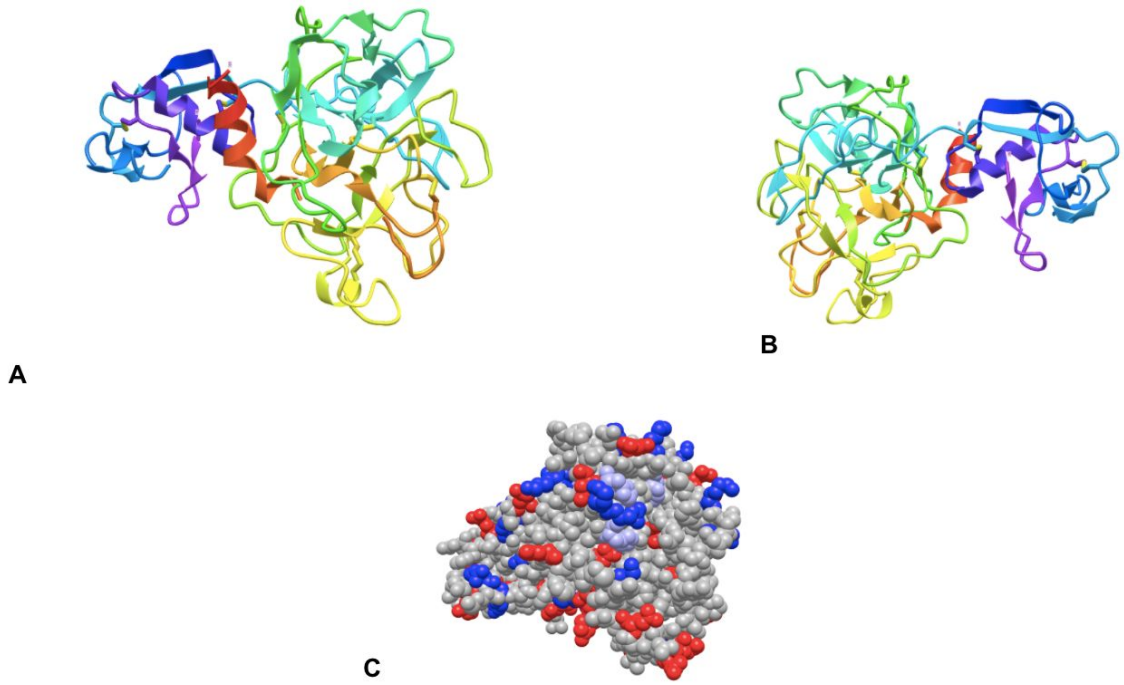

Figure 5. HHPred structure for TMPRSS2. The original view (A), 180 degrees rotated around the Y axis (B) and space-filling model (C) are shown here. Coloration of (A) and (B) is spectrum gradient based on proximity to n or c terminii. Coloration of the space-filling model (C) is done using charge. 6 alpha helix sets and 20 beta sheets were predicted by HHPred.

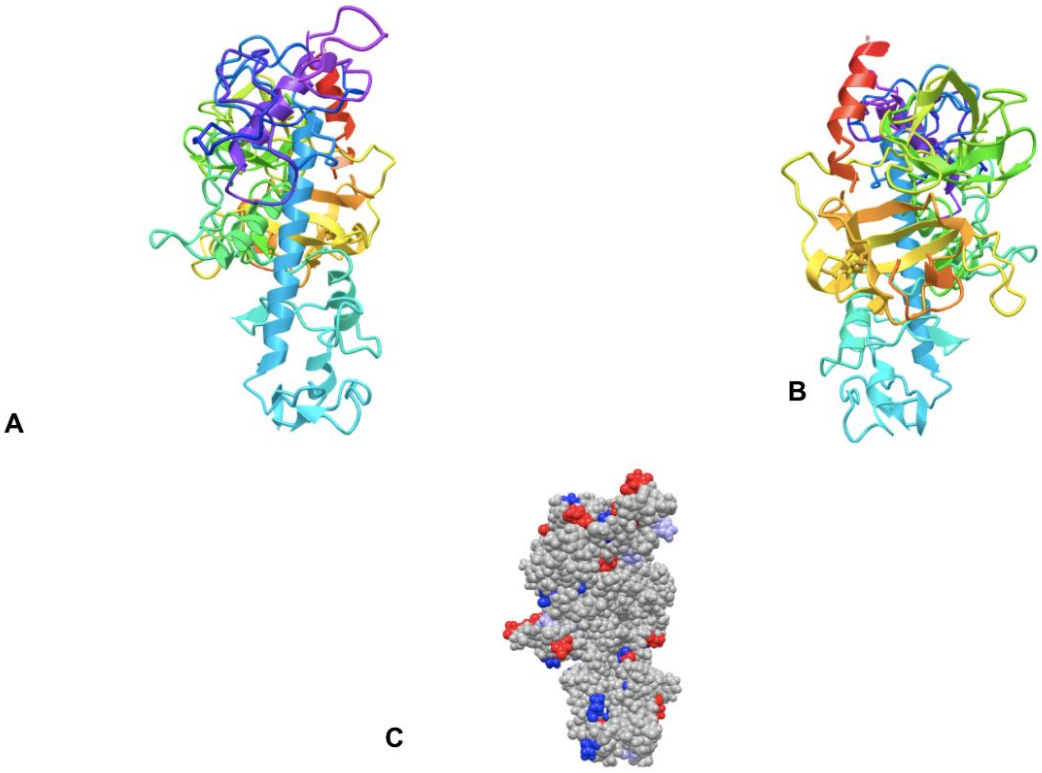

Figure 6. RaptorX structure for TMPRSS2. The original view (A), 180 degrees rotated around the Y axis (B) and space-filling model (C) are shown here. Coloration of (A) and (B) is spectrum gradient based on proximity to n or c terminii. Coloration of the space-filling model (C) is done using charge. 16 alpha helix sets and 15 beta sheets were predicted by RaptorX.

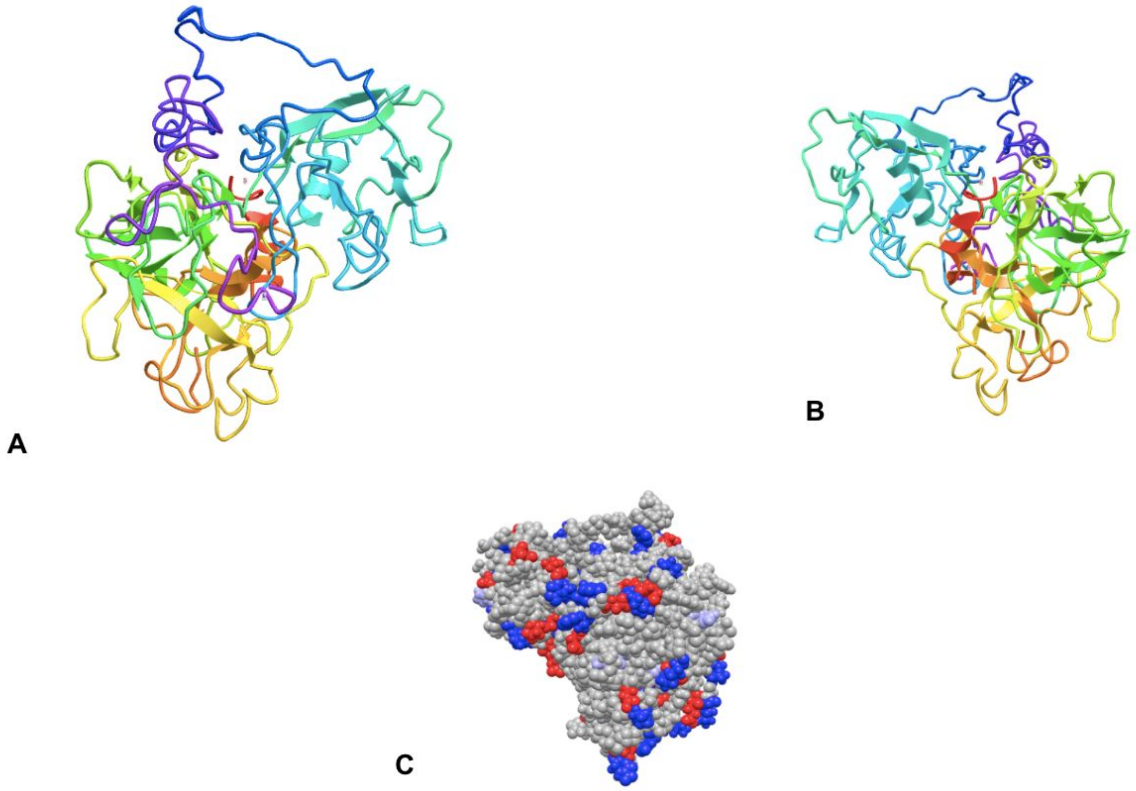

Figure 7. I-TASSER structure for TMPRSS2. The original view (A), 180 degrees rotated around the Y axis (B) and space-filling model (C) are shown here. Coloration of (A) and (B) is spectrum gradient based on proximity to n or c terminii. Coloration of the space-filling model (C) is done using charge. 4 alpha helix sets and 17 beta sheets were predicted by I-TASSER.

The predicted model of TMPRSS2 created by I-TASSER was docked with the known structure of SARS-CoV-2 S protein (PDB:7DK3) (Figure 8). 18 interaction sites were found on SARS-CoV-2 and 21 interaction sites were found on TMPRSS2 (Missing table). Two SNPs were found to be either on or close to the interaction sites. TMPRSS2 V280 interacts with SARS-CoV-2 S protein; SNP rs142446494 changes valine 280 to methionine. Four TMPRSS2 interaction sites exist from 300-317 amino acids; rs768173297 is located at 309 aa and changes threonine to methionine. It is interesting to note that the effect of point mutations for these SNPs were found to be not very damaging.

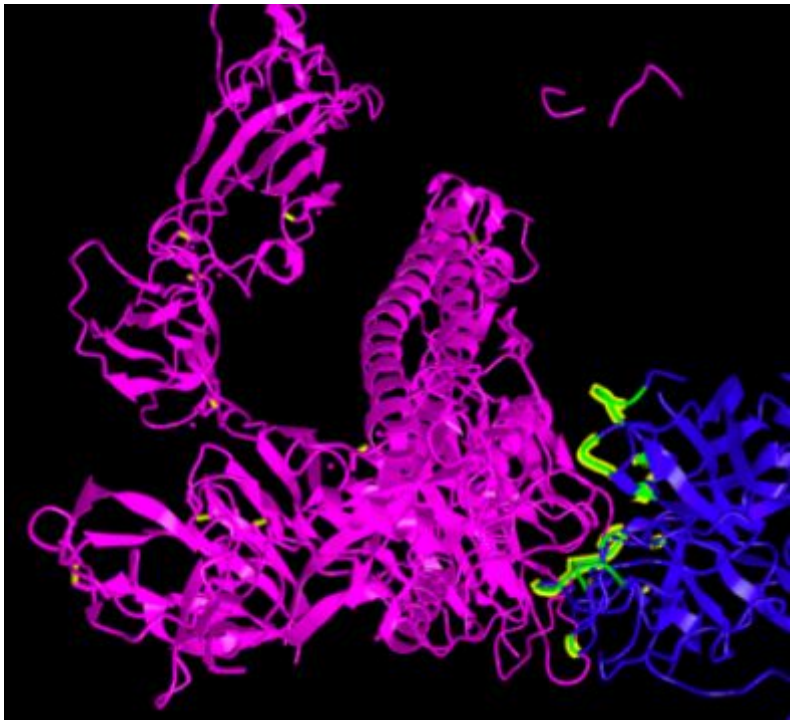

Figure 8. Interaction between TMPRSS2 and SARS-CoV-2. Need to update this with better image of interaction sites and to match other images. Green highlighted areas indicate interaction sites between the two molecules.