DMS Topic 2: TP53

Frido Petersen, Dario Prifti, Enno Schäfer und Maximilian Fidlin

Group: 2: 02

E-Mail Adressen:
frido.peteresen@stud.uni-heidelberg.de
enno.schaefer@stud.uni-heidelberg.de
maximilian.fidlin@stud.uni-heidelberg.de
dario.prifti@stud.uni-heidelberg.de

Data analysis project 2023

Supervisors: **Prof. Dr. Dominik Niopek, Dr. Jan Mathony, Benedict Wolf**

Based on the Article:
Giacomelli, A., Yang, X., Lintner, R., McFarland, J., Duby, M., Kim, J., Howard, T., Takeda, D., Lý, H., Kim, E., Gannon, H., Hurhula, B., Sharpe, T., Goodale, A., Fritchman, B., Steelman, S., Vazquez, F., Tsherniak, A., Aguirre, A., and Hahn, W. (2018). Mutational processes shape the landscape of TP53 mutations in human cancer. Nature Genetics *50*, doi: 10.1038/s41588-018-0204-y

Contents

Abbreviations

AA	amino acid
BD	Basic domain
DBD	DNA binding domain
DMS	Deep mutational scanning
PC	principal component
PCA	principal component analysis
SMN	single mutated nucleotide
T1	Transactivation domain 1
T2	Transactivation domain 2
TA	Transactivation domain
TD	Tetramerization domain

1. Introduction

This project analyzes Deep mutational scanning (DMS) data of the tumor suppressor protein p53 which is encoded by the TP53 gene. p53 is highly relevant due to it being mutated in the majority of human cancers (Kotler *et al.* (2018)). To understand the job of p53, it is important to know that as a consequence of various cellular stress signals p53 is activated and can transactivate various genes that encode for mechanisms like the induction of the cell cycle arrest, DNA repair, senescence and apoptosis (Zhu *et al.* (2020)).
The data we based our work on shows different variations of p53 with and without activators. That means that every mutation yields a protein that is either fitter or less fit than the wildtype form. This information is contained in the so-called DMS score.
The data we used originates from two different research papers, the first one being Giacomelli *et al.* (2018), containing 3 different datasets, and the second source being Kotler *et al.* (2018), which had one dataset. To grasp what information is contained in these datasets, we looked at extreme DMS scores and performed analyses like dimension reduction and clustering on our data.
To better understand the fitness of a protein, one has to look at it's domains. For p53, the exact locations of those regions vary depending on the source, so we chose to determine domains according to the article of Harms and Chen (2006). p53 has two Transactivation domains (TAs), from around positions 1 to 40 and 40 to 90, with a proline rich reason inside the second TA, a DNA binding domain (DBD) from position 102 to 292, a Tetramerization domain (TD) located around positions 320 to 360 and a Basic domain (BD) (also called regulatory domain) at the c-terminal end at around 360 to the last position (393). The TAs activate transcription by interaction with the transcriptional machinery. Together with the Transactivation domain 2 (T2) the proline rich region has a proapoptotic function and also helps in enhancing p53's transcriptional activity (Harms and Chen (2006)). The DBD plays an important role in the sequence-specific function of p53 as a transcription factor. The predominant genetic locus in the TP53 gene for cancer-associated mutagenesis lies in the DBD (Rivlin *et al.* (2011)). The majority of the DBD forms a supporting scaffold in the interaction with DNA, while some specific locations in the DBD directly interact and bind to specific DNA sequences. Giacomelli *et al.* (2018) suggest that the DBD is most commonly mutated in cancer. Most of those mutations tend to be mis-/ and nonsense mutations. While the TD mediates the tetramerization which is needed to bind DNA with a high affinity and for it to be functional in transcription activation (Harms and Chen (2006)), the BD exists for regulatory purposes. (Harms and Chen (2006)).
Based on the domain information and our basic model for likeliness of each amino acid (AA) mutation, we wanted to investigate whether the DBD, as a very essential domain of p53, is less prone to mutations that are more likely to occur by chance.

2. Materials and Methods

As a baseline package, we used the Python package pandas (pandas development team, 2020). All self-written functions and methods are derived from basic pandas operations and we simply applied them to the given data. Visualisations were generated using seaborn and matplotlib (Waskom, 2021).
To compare our findings on specific AAs, we also looked at an additional, external dataset which contains the chemical properties of each AA. For the comparison of the chemical properties with the DMS-Scores, the use of hierarchical and k-means clustering, principal component analysis (PCA) and the euclidean-distances, we used functions from scikit-learn (Pedregosa *et al.*, 2011). To determine the optimal amount of clusters, we wrote our own function
By using BLAST (Altschul *et al.*, 1990), we were able to generate all the codons that are within range of a single mutation single mutated nucleotide (SMN) from the original codon for each specific position. We thereby implemented a basic model for the probability of each mutation, since SMNs are more likely to occur in a codon, compared to double or triple nucleotide mutations. It allowed us to selectively pick the possible AAs for SMNs out of all mutations provided in the datasets. The sequences were translated into an AA sequence, using a self-written library.
The scikit-learn.stats package was used, to perform statistical comparisons and tests (Pedregosa *et al.*, 2011).

3. Results

3.1. Comparability of the datasets

We examined different datasets, including p53 NULL etoposide, p53 NULL nutlin, and p53 wildtype nutlin by Giacomelli *et al.* (2018), as well as the human p53 dataset by Kotler *et al.* (2018). To gain a first understanding of trends and possibly identify p53's domain patterns, we initially visualized the datasets using heatmaps. (see fig. 1, 2 and 10a in appendix).

At a first glance it was evident that the datasets from the different research papers we used are inherently different. While the Giacomelli *et al.* (2018) datasets has a DMS-score for every possible mutation at every single position of the Protein, the Kotler *et al.* (2018) dataset only ranges from Postions 102 to 292 and does not include every possible mutation. Also there is a small difference in the wildtype protein sequence at position 72 where the Giacomelli *et al.* (2018) datasets contain an Arginine while the Kotler *et al.* (2018) dataset contains a Proline. While this does not directly impact the comparability of our datasets (since of position 72 not playing a role in the dataset of Kotler *et al.* (2018)) we have to keep in mind that this difference could have had an impact on the folding or general structure of p53. Additionally, Giacomelli *et al.* (2018) used drug resistance as a metric for assessing DMS scores while Kotler *et al.* (2018)

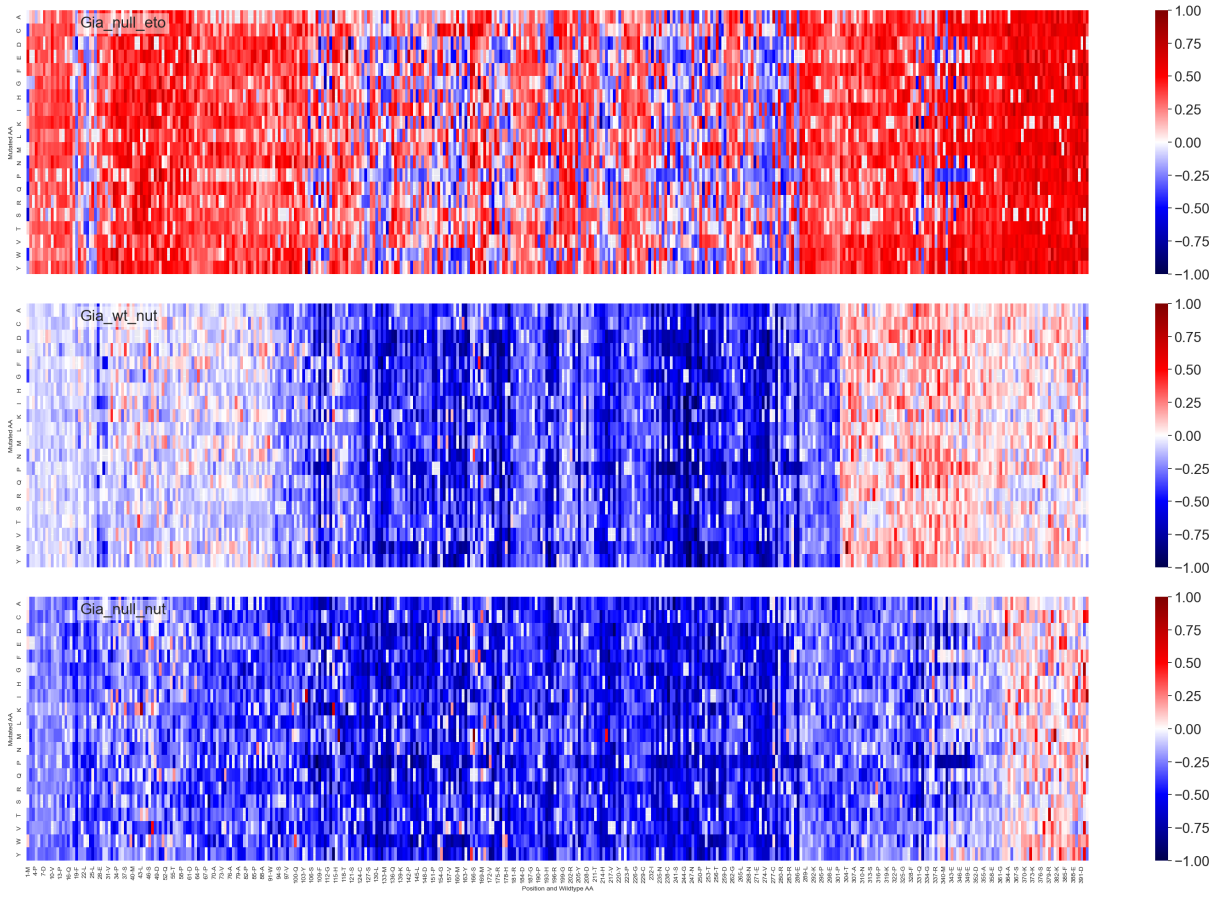


Figure 1: Heatmaps of all Giacomelli *et al.* (2018) datasets

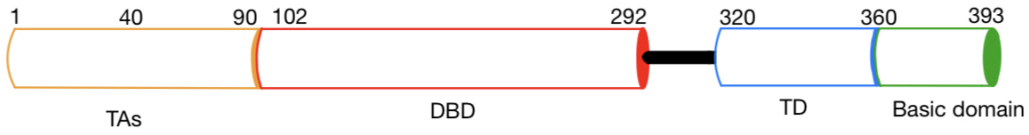


Figure 2: A rough outline of p53 domains

used growth as a scale for measuring protein-fitness. However, the Giacomelli wildtype nutlin and Giacomelli NULL nutlin datasets exhibited some similarities in terms of trends and values, which cannot be said when we compared them to the Giacomelli NULL etoposide dataset. This disparity is likely due to the use of different p53 activating agents across the datasets, namely nutlin-3 and etoposide. One notable observation across all datasets was that AAs in the range of approximately 100-300 generally display a negative effect caused by mutations.

As mentioned in our initial presentation, the datasets were created using different methods of obtaining and evaluating the DMS scores which has to be taken into account for comparison.

The first metric of comparison we chose to apply onto our datasets was to look at the AAs in the original sequence that, when replaced, caused the most negative DMS scores throughout the whole Protein. This could mean that these specific AAs serve a specific function that cannot be replaced well by other AAs (see Table 1 in appendix). We have to note that this comparison does not take into account that these values disregard positional information.

Furthermore, we decided to examine the AAs that, when mutated to, resulted in the most significant decreases in the DMS score, indicating a substantial impact on protein function (see Table 2 in appendix). Afterwards, we illustrated the trends of substitution by creating an overview showing a mean value for each substitution. These mean values were calculated for the whole length of the p53 protein (see fig. 10b in appendix).
To obtain some positional information, we found out the exact locations in each dataset which held the lowest mean DMS scores (see Table 3 in appendix).
As a final and conclusive way to show the differences of the datasets we chose to visualize our datasets as linegraphs in one plot. This plot shows the mean DMS scores for each position and lets us quickly seek out positions that are greatly affected by mutation and those that are not. By summing up all DMS scores and dividing this sum by the number of values summed we can also create a rough comparability to the Kotler dataset which is also visible in this graph (see fig. 3).

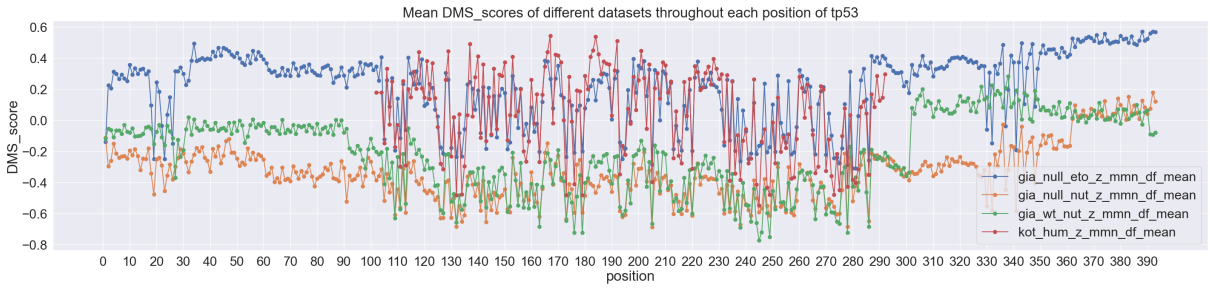


Figure 3: Linegraph of mean DMS scores at each position

For the following analyses we focused mainly on Giacomelli NULL etoposide as our main dataset, since it is easier to show what we have done with one representative dataset. We decided to choose Giacomelli NULL etoposide because it shows a clear separation of regions that are badly affected by DMS and those that are not.

3.2. Data exploration

The dendrogram plot based on the additional dataset containing the chemical properties of AAs showed that leucine and isoleucine were closest to one another, with valine being close by. Aspartate and glutamate were also grouped together and their distance to one another matched the distance between asparagine and glutamine, Serine and threonine as well as phenylalanine and tyrosine. As seen in the dendrograms, arginine and tryptophan were the most unique AAs. (see fig. 4)

Plotting distances between AAs from the distance matrices, calculated from DMS scores, showed which AAs have a similar impact on protein fitness. We learned that mutating phenylalanines had the same unique effects on DMS scores as mutating valines, which is interesting as they do not resemble each others structures. Based on the dendrogram plots, the pairs histidine and threonine, methionine and tryptophan as well as tyrosine and cysteine

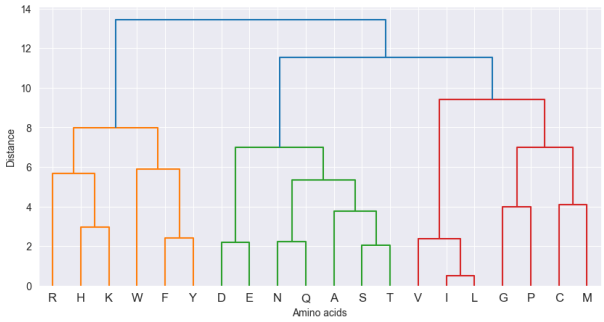


Figure 4: Dendrogram on chemical properties of AAs - This plot shows the dendrogram of a ward clustering. The letters represent AAs in the one letter code.

Figure 5: PCA results of wild-type AA distances - The plot shows wild-type AAs that were clustered with the ward method based on a distance matrix. The letters represent AAs in the one letter code.

had the same effects when being mutated. Mutating proline or alanine had the most similar influence on the DMS scores (see fig. 12a in appendix). The general trends could also be seen in the plots that included the PCA (see fig. 5), with every AA belonging to the same cluster as in the dendrogram plot. However, some pairs, like proline and alanine, that appeared to be closest to one another in the dendrogram had other AAs in between them when plotted after PCA.

Afterwards, we compared our findings on mutating the wild-type sequence with the trends of the AAs being used to mutate with. While phenylalanine being mutated had the same unqiue impact as valine being mutated, mutating with phenylalanine was by far not as unique and more closely resembled the effects of mutating with isoleucine (see fig. 12b in appendix). Still, when mutating with valine, valine and phenylalanine were part of the same cluster. That shows some similarity to the results of the AAs that were mutated.

However, proline has the most unqiue impact on p53's fitness when it is used to mutate with. The similarities proline had to alanine, glutamate and aspartate when mutated completely vanish. Due to their structural and therefore functional similarity, glutamate and aspartate were still quite close when used to mutate with, but alanine was part of a different cluster (see fig. 11 in appendix).

To comment on the effect the clustering method had on the clustering, we performed the same analyses with a k-means clustering. The amount of optimal clusters changed from four to six, so the clustering with K-Means was different. When manually changing the optimal number of clusters to six, the identical clustering was achieved.

3.3. Single mutated nucleotides (SMNs)

Since the variation of AAs of a protein is based on it's nucleotide sequence, we wanted to bias all DMS scores with their corresponding probability. We started out by using the SMNs and implemented a function, where only DMS scores of AA that can are within range of a SMN are selected, as described in 2. We then modularised our code to give the opportunity to extend the variation matrices in order to accommodate for double and triple mutated codons.
When analysing the differences of DMS scores between SMNs and double as well as triple mutated codons, we found single mutation DMS scores to be more positive. Furthermore, we compared probability biased DMS scores, we called severity scores, with their corresponding DMS scores (see fig. 6). In this heatmap, the distinct dark blue and dark red lines are way less present for the severity scores. This means, we observed significantly less scores diverging from zero as the severity scores show less extreme values when compared to the DMS scores. Especially, tryptophan, phenylalanine, and aspartic acid are rarely observed among SMNs. Hence, they are less likely to be inserted into the protein through unbiased mutation.
Furthermore, most severity scores are more positive, relatively to their corresponding DMS scores. Except for arginine, which gets assigned more negative severity scores. Additionally, we were still able to identify domains, described earlier.

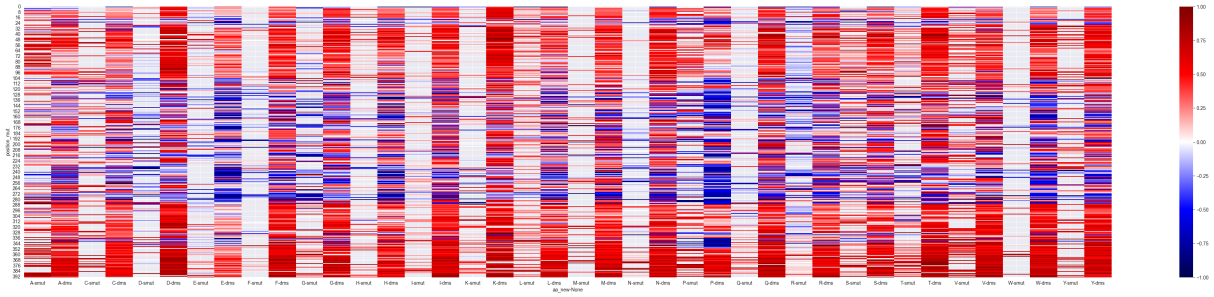


Figure 6: Comparison of DMS and SMN scores - SMN were generated biasing with their probability (severity scores). Only discussed AAs were selected.

3.4. Domain comparison

Our goal was to compare the domains of p53 by focusing on their DMS scores. We started out by slicing our initial dataset into the domains described in the introduction and plotted domain-wise DMS scores into a histogram (see fig. 7). This showed the frequency of DMS scores of the

different domains to all peak between 0.25 and 0.50. Most domain's DMS scores tended to be more negative while being asymmetrically distributed. The DBD stands out because it has a second peak at around -0.25 which is almost as high has the main peak. These results match our findings from 3.1 Comparability of the datasets, with the DBD having the most negative areas. We then applied the mean substitution matrices (as described in 3.1 Comparability of the datasets) onto the domains separately. By this, we could get a better overview of AAs that are very important for the function of the protein or respectively cannot be replaced. In the first TA phenylalanine and tryptophan show a very distinct dark line (see fig. 8a). Another anomaly can be seen in the DBD. In this case, replacing isoleucine results in very negative DMS scores (see fig. 8b). This is especially noticable for AAs with different chemical properties. For example, an exchange with other small hydrophobic AAs, such as valine, leucine or methionine results in comparatively positive DMS scores. Similar trends can also be seen for the TD.
We then applied the same functions but left out all AAs not reachable by a single mutation, as described in section 3.3. When looking at the histograms, this did not impact the distribution in most of the cases (see fig. 9a). It did however considerably impact the distribution of DMS scores in the DNA binding region. Before, the right peak was about 1.5x higher than the peak below zero. With only SMNs, the positive peak is more than twice as high (see fig. 9b). These changes cannot be observed in any of the other domains.
To further investigate this, we generated a new dataset containing random codons instead of SMNs codons. We investigated whether the means of those two datasets differ significantly using the Mann-Whitney U test. This test showed the DBD to be the only one to differ significantly when taking out single mutations compared to random codons.
We tried to replicate these results by taking a closer look at the other Giacomelli datasets. The histograms of all DMS scores again shows all domains, except the DBD, to follow a bell shape (see fig. 13a, 13b in appendix). The distribution of the DMS scores for the DBD has again more or less two local maxima. These can be seen better without the fitted curves (see fig. 13c in appendix). If we take out the non single mutations again, the more negative DMS scores get reduced significantly (see fig. 13d in appendix). For the NULL nutlin dataset, the left maximum went down to almost a third of the right maximum. Again, those drastic changes are exclusive to the DBD.

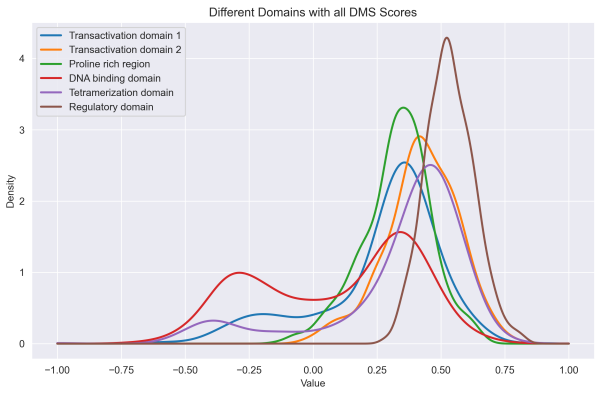
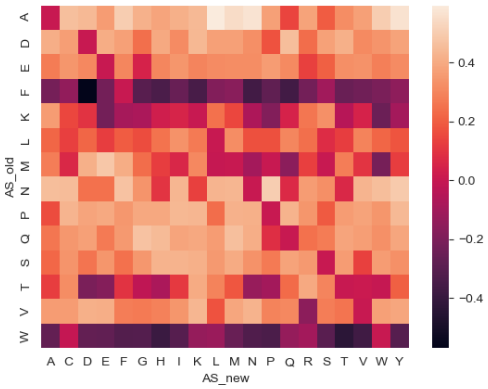
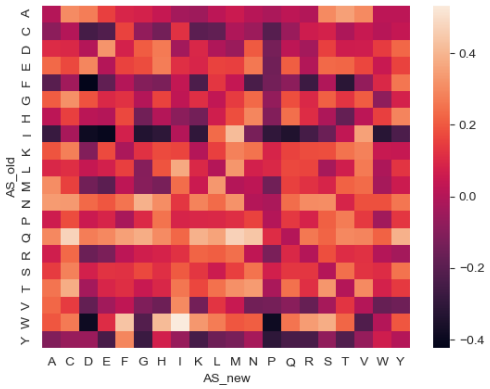


Figure 7: Domain-wise distribution of DMS scores - This diagram shows the frequency of DMS scores for each domain in p53.

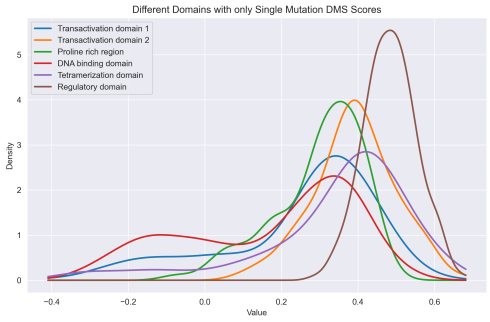


(a) Mean substitutions TA 1

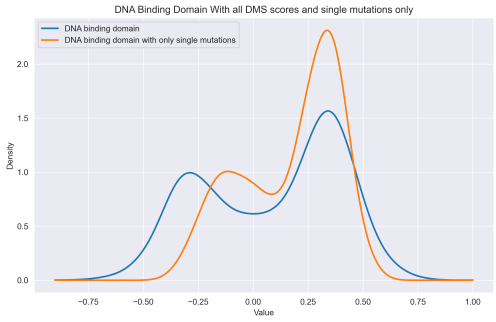


(b) Mean substitutions DBD

Figure 8: Mean DMS score for each AA substitution - This figure shows two examples for the mean DMS scores of the p53 domains. Subfigure (a) is for the TA 1, (b) for the DBD. AS_old is the AA in the wildtype protein and AS_new is the mutated AA. Each value describes the mean DMS score for that exact exchange of AAs in the corresponding domain of p53.



(a) Domain-wise distribution of DMS scores with SMN



(b) Comparison of the DBD with all DMS scores and only SMN

Figure 9: Domain-wise distribution of DMS scores with SMN - This figure shows (a) the domain-wise distribution of DMS scores with SMN and (b) the comparison of the DBD with all DMS scores and only SMN

4. Discussion

In the first part of comparability of the datasets, we were able to replicate some findings of Giacomelli *et al.* (2018). This included showing the importance of the DBD for protein fitness through heatmaps. Our goal was to determine the differences between the datasets and to determine one dataset that we based all further analyses on. We chose the "Giacomelli null etoposide" as our main data set, due to it giving us the clearest outlines of the different domains. Research of Harms and Chen (2006) reveals that mutant p53 can be categorized in various types of mutations having different hotspots. Some of those hotspots are R248, R273, R175, G245, R249, and R282. As Table 3 shows, we were able to reproduce some of these results, for example finding that the G245 mutation caused the lowest DMS scores in two different datasets. Also R280 is a finding that can be related to the R282 hotspot mutation of Harms and Chen (2006). We then continued by looking into the chemical properties and distances of AAs substitutions in p53. The results for chemical properties were as expected since each of the mentioned pairs (e.g. asparagine and glutamine) shows the same structural characteristics. To take a closer look at AAs substitutions, we created distance matrices based on the mean substitution values for the wildtype and mutated AAs. After using dimension reduction we clustered the AAs to compare it to the chemical properties.
In Figure 5 showing the PCA results, proline and alanine were not as closely grouped as the dendogram plot 12a suggested. Instead, proline seemed to be closest to serine and glutamate. This is due to these other AAs that showed almost the same influence on protein fitness. These differences are a result of matrix diagonalization, which PCA is based on. Through this analysis, the original 20 dimensional coordinate system is rotated so that the explained variance by the first two principal components (PCs) is maximal. The PCs are the dimensions of the resulting, rotated coordinate system (Pearson, 1901). To further look into this, in the next step we would have looked at which original dimensions contribute the most to PC1 and PC2. The greater distances between the AAs in the external dataset on chemical properties might be due to the fact that during the creation of the distance matrices, we lost position-related information due to iterating over the whole length of the protein.
To summarize, it became visible that DMS scores are severely affected by both the AA that is mutated to as well as the wildtype AA meaning that the trends trends did not correlate strongly. Furthermore, judging from the plots, the effects on DMS scores are not directly connected with structural similarity. Some correlation can be implied as aspartate and glutamate always appeared in each others proximity, but the general trends are not explainable by structural similarity. While isoleucin showed the most unique effects on protein fitness when mutated, proline exhibited similar effects when it was used to mutate with. As an outlook, the next step would be to perform statistical tests to quantify the significance of the observed trends. To show that our code runs on DMS data from any protein, we succesfully applied the hierarchical clustering (with and without PCA) on the E. coli ß-Lactamase dataset.

We then selected SMN AAs from our dataset. By this, we wanted to investigate, whether the probability of a mutation has influence on the impact of that mutation. One aspect of future

research is to investigate whether SMNs really are the most common mutations. We contacted several research groups asking for statistical data. This was without success, due to privacy policies.

Another interesting aspect would be to extend the code on SMNs to consider multiple base mutations to then compare the results. We also have to keep in mind that the datasets provided only included single AA exchanges. Therefore many interactions that may only be possible through multiple mutations remain unknown.

We then took a look at the domains separately. We started out by plotting the DMS scores into histograms for each domain. This showed the DBD to be the only one with a second peak at lower DMS scores, which further underlines how important DNA binding is for protein fitness.

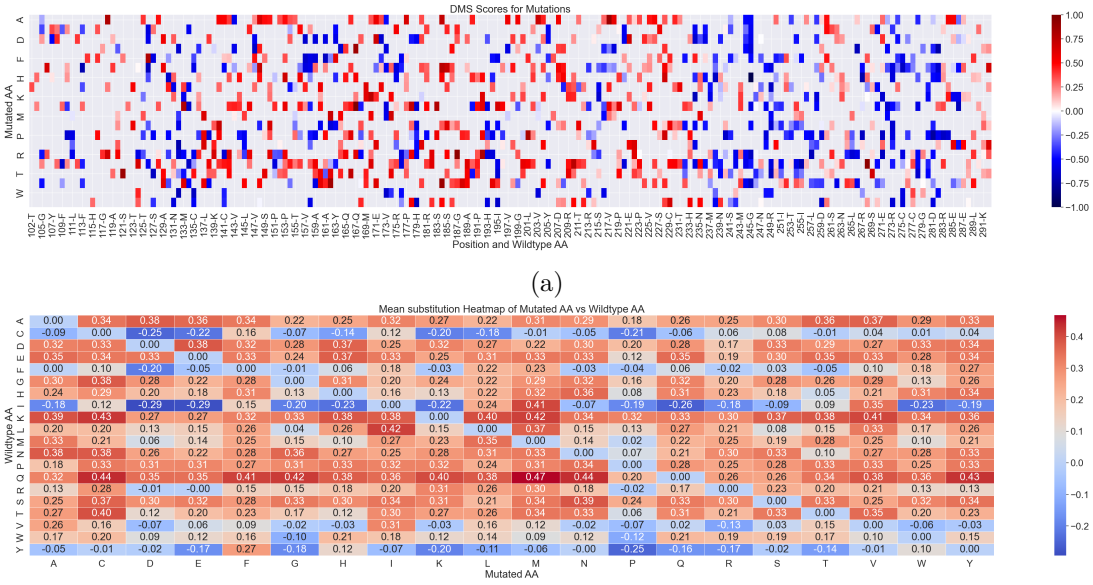
We then applied the mean DMS score matrices on each domain. For the first TA, this showed two very distinct dark lines for phenylalanine and tryptophan. This suggests a very unique and important role of these two AAs, being the only aromatic, non-polar, hydrophobic AAs. Those features might be important since one of the known functions of this domain is binding non-specifically to DNA (Baptiste *et al.*, 2002). We also confirmed some of the findings considering chemical properties with isoleucine being exchanged relatively well with leucine, valin or methionine. All of those AAs have short and hydrophobic residues. Why those findings are more of an exception rather than a rule remains unknown and to be investigated further.

We then applied the single mutation filter on each domain separately. This showed unique trends for DBD. It is the only domain who's DMS scores improve and show a different distribution. This supports our idea of the DBD to be less prone to mutations that are, to our knowledge, more likely to occur by chance. We further compared taking random AAs out of the data set with taking out single mutations only. Using the Man-Whitney U test, we showed that this led to differing results only for the DBD, further supporting our hypothesis. The significance of these results remain unclear, since the generation of random codons might not be the best method when trying to find outliers in our data. When looking at the DBD in the other Giacomelli data set, we could find similar trends. This again aligns with our idea of the DBD being less prone to single mutations in the DNA sequence.

To our best knowledge, those findings have not been reported before. Therefore, the validity of the applied methods and the corresponding results remain to be thoroughly examined. This could be done by checking other DMS datasets for the same or different proteins or by consulting patient data.

A. Supplementary Information

A.1. Comparability of the datasets



(a)

(b)

Figure 10: (a) Heatmap of the Kotler *et al.* (2018) dataset. (b) Mean substitutions of AAs in *Giacomelli null etoposide*

Table 1: Wildtype AA causing the lowest DMS scores- sorted by Giacomelli null etoposide

Wildtype AA	GNE	GNN	GWN	KH
I	-0.067487	-0.558694	-0.405441	-0.146787
Y	-0.058534	-0.572673	-0.491159	-0.358660
C	-0.051440	-0.545096	-0.543561	-0.145375

Table 2: Mutated AA causing the lowest DMS scores- sorted by Giacomelli null etoposide

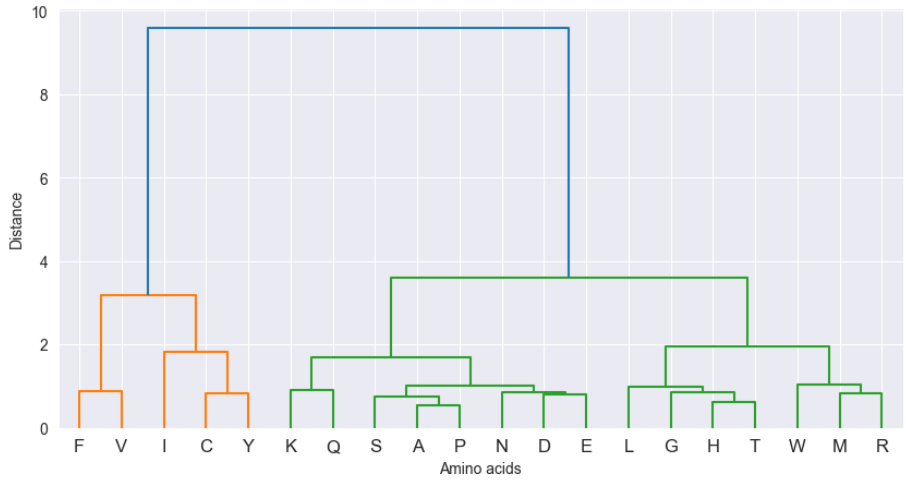
Mutated AA	GNE	GNN	GWN	KH
P	0.096191	-0.449589	-0.275914	-0.143151
D	0.177935	-0.380285	-0.229204	-0.050507
G	0.188055	-0.378543	-0.229989	-0.028291

Table 3: Position of lowest mean DMS-scores in each dataset

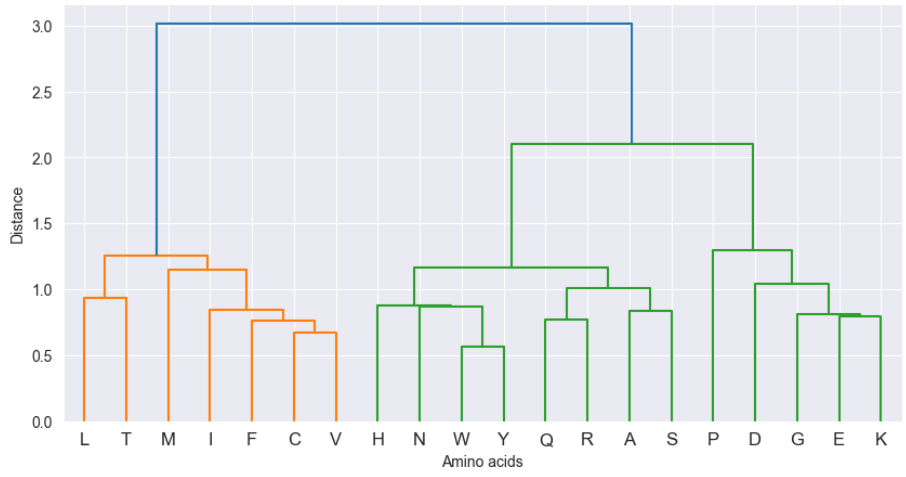
Name of the Dataset	Location of the lowest mean DMS score	mean	Original AA
Giacomelli Null Etoposide	280	-6.190289	R
Giacomelli NULL Nutlin	205	-13.762829	Y
Giacomelli WT Nutlin	245	-15.419176	G
Kotler	245	-6.568038	G

A.2. Data exploration, clustering

Figure 11: PCA results of mutated AA distances - The plot shows mutated AAs that were clustered with the ward method based on a distance matrix. The letters represent AAs in the one letter code.



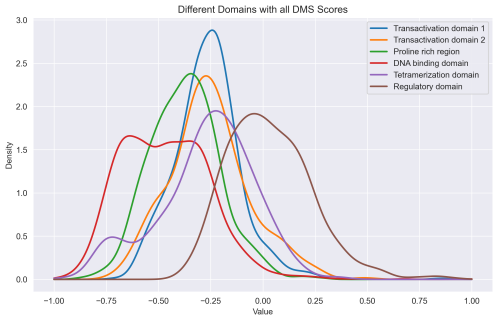
(a)



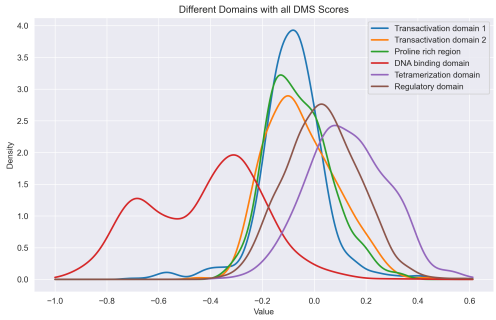
(b)

Figure 12: (a) Dendrogram on distances of p53 wild-type AAs - This plot shows the dendrogram to a ward clustering performed on a distance matrix. The letters represent AAs in the one letter code. (b) Dendrogram on distances of mutated p53 AAs - This plot shows the dendrogram to a ward clustering performed on a distance matrix. The letters represent AAs in the one letter code.

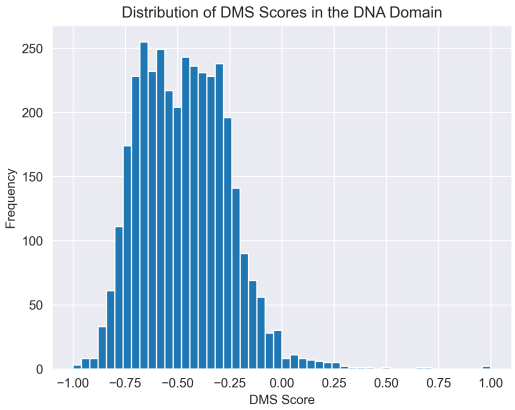
A.3. Domain comparison



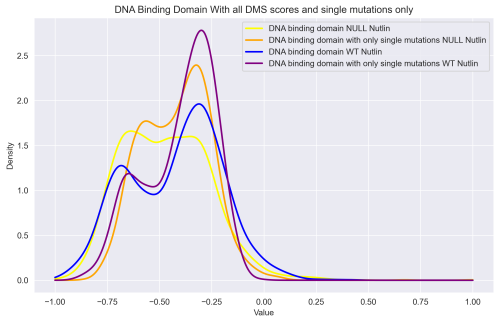
(a) Giacomelli NULL nutlin



(b) Giacomelli WT nutlin



(c) Giacomelli NULL nutlin



(d) Comparison of all DMS scores and only SMN

Figure 13: Results of domain comparison for the other Giacomelli datasets - This figure shows (a) the domain-wise distribution of DMS scores for the Giacomelli NULL nutlin dataset, (b) the domain-wise distribution of DMS scores for the Giacomelli WT nutlin dataset, (c) the distribtution of DMS scores in the DBD in the Giacomelli NULL nutlin dataset and (d) the comparison of all DMS scores and only SMN in the DBD in the other Giacomelli datasets.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology *215*, 403–410, doi: 10.1016/s0022-2836(05)80360-2.

Baptiste, N., Friedlander, P., Chen, X., and Prives, C. (2002). The proline-rich domain of p53 is required for cooperation with anti-neoplastic agents to promote apoptosis of tumor cells. Oncogene *21*, 9–21, doi: 10.1038/sj.onc.1205015.

Giacomelli, A., Yang, X., Lintner, R., McFarland, J., Duby, M., Kim, J., Howard, T., Takeda, D., Lý, H., Kim, E., Gannon, H., Hurhula, B., Sharpe, T., Goodale, A., Fritchman, B., Steelman, S., Vazquez, F., Tsherniak, A., Aguirre, A., and Hahn, W. (2018). Mutational processes shape the landscape of TP53 mutations in human cancer. Nature Genetics *50*, doi: 10.1038/s41588-018-0204-y.

Harms, K. L., and Chen, X. (2006). The functional domains in p53 family proteins exhibit both common and distinct properties. Cell Death and Differentiation *13*, 890–897.

Kotler, E., Shani, O., Goldfeld, G., Lotan-Pompan, M., Tarcic, O., Gershoni, A., Hopf, T. A., Marks, D. S., Oren, M., and Segal, E. (2018). A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. Molecular Cell *71*, 178–190.e8, doi: https://doi.org/10.1016/j.molcel.2018.06.012.

Pearson, K. (1901). LIII. iOn lines and planes of closest fit to systems of points in space/i. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science *2*, 559–572, doi: 10.1080/14786440109462720.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research *12*, 2825–2830.

Rivlin, N., Brosh, R., Oren, M., and Rotter, V. (2011). Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis. Genes & Cancer *2*, 466–474, doi: 10.1177/1947601911408889. PMID: 21779514.

pandas development team, T. (2020). pandas-dev/pandas: Pandas. doi: 10.5281/zenodo.3509134.

Waskom, M. L. (2021). seaborn: statistical data visualization. Journal of Open Source Software *6*, 3021, doi: 10.21105/joss.03021.

Zhu, G., Pan, C., Bei, J.-X., Li, B., Liang, C., Xu, Y., and Fu, X. (2020). Mutant p53 in Cancer Progression and Targeted Therapies. Frontiers in Oncology *10*, doi: 10.3389/fonc.2020.595187.