Optimizing human SIRT6 protein with deep learning of 3D structures based on maximum lifespan

Yanfei Zhang*, Zhenchao Ruan

Shanghai Starriver Bilingual School, Shanghai, China

* Corresponding Author Email: phoebezhang07@qq.com

Abstract. Aging is the largest risk factor for many diseases, and it is thus of great social and medical importance to delay the appearance of aging traits or promote healthy aging. The protein SIRT6's function promotes longevity, but it is still unclear whether the function of SIRT6 can be further optimized through protein sequence engineering. Here, we developed a comprehensive framework of sequence optimization methods for functional proteins utilizing both sequence and structure. We used RoseTTAFold to predict the structural variance of SIRT6 sequences across 142 mammalian species and then analyzed the association between the maximum lifespan (MLS), sequence, 3D structures, and amino acid selection of SIRT6. By fine-tuning the ESM model, we developed a model that is capable of accurately predicting the MLS of the species from SIRT6 sequence (Pearson's r = 0.818, MAE = 8.608 years). We further generated mutant sequences of the human SIRT6 using ProteinMPNN and analyzed different sites' importance in longevity based on their impact on predicted MLS. 37 sites that play a key role in sequence function were identified. Among them, 20 sites located in the NAD+ binding area and β1-sheet, followed by 9 sites on the α1-helix, were found to have the greatest impact on MLS. We then tested the enzymatic efficiency of 2 novel SIRT6 sequences (with predicted MLS increase of 16.5% and 20.1%) in vitro, and showed that their DNA double-strand break repair efficiency is indeed higher than the human ortholog SIRT6 sequence. Our study demonstrates that despite being highly conservative in evolution, there is still room for optimization in human SIRT6. We not only found a high correlation between MLS and protein sequence or structure, generated experimentally-validated optimized strains, but also identified key sites affecting protein function that have varying importance, thereby providing novel insights into potential anti-aging interventions.

Keywords: SIRT6, Maximum lifespan, protein design, Convolutional Neural Network, RoseTTAFold, Elastic net, Protein language model, ESM model, ProteinMPNN, Immunofluorescence.

1. Introduction

1.1. Research background

Aging is the largest risk factor for many diseases such as cancer, cardiovascular disease, and neurodegenerative diseases [1], it is thus of great social and medical significance to study the mechanism of aging and reduce the occurrence of related diseases, thereby reducing the disease burden of the population and improving the life quality of the elderly population [2]. The progressive accumulation of genetic damage may be most critical to aging, because failure to repair DNA damage may not only causes genomic instability but also epigenetic changes, leading to global gene transcriptional dysregulation, which is a major cause of cell senescence and aging [3–7]. Sirtuin 6 (SIRT6), a member of the SIRT protein family [8, 9], plays an important role in regulating DNA repair (especially in double strand breaks (DSB) repair [10]), and is thus an important factor in the process of aging. SIRT6 function indirectly promotes longevity [11, 12]. Transgenic mice overexpressing SIRT6-exhibited an increase in lifespan of up to 30% [13]. Cohort and case-control studies conducted in natural populations also revealed that certain mutations in the SIRT6 locus are associated with longevity, resulting in an increased average life expectancy of approximately 5 years [14]. Therefore, SIRT6 sequences with higher biological function and thus correspond to longer lifespan in mammals. Such sequences have broad application prospects in gene therapy and the exploration of novel drug targets.

Maximum lifespan (MLS) varies widely among mammalian species, ranging from 16 months in shrews to 211 years in bowhead whales [15]. This makes comparison of genetic differences between species (comparative genomics) a valuable field of research. Tian [10] compared the sequence of SIRT6 of 18 species of rodents with different lifespans and found that the DSB repair efficiency of SIRT6 protein is highly correlated with lifespan of species. Differences in 5 amino acids (AA) sites located on the protein surface led to the diverse maximum lifespan of different rodents, from 3 years to 32 years. However, larger-scope cross-species comparisons, such as those involving DNA repair protein mutations across whole mammalian lineages, have not yet been reported.

Traditionally, protein structure is obtained by using cryo-electron microscopes [16], therefore research utilizing the protein's 3D structure information in analysis was less common. However, bioinformatic methods have seen explosive progress in the last 2 years, models such as AlphaFold 2 [17], RoseTTAFold [18–20], and ESMFold [21, 22] are currently capable of high-precision protein structure prediction using only AA sequences, which made possible high-throughput obtaining of protein structure using deep learning only. Moreover, generative models [23, 24] have been developed for de novo protein design for a desired structure. The progress of methodologies provides new opportunities for SIRT6 research, i.e., designing optimized human SIRT6 protein variants using deep neural network generative models based on cross-mammalian sequence and lifespan analysis.

1.2. Research content

Our study assumption is as follows: the force of natural selection declines as animals live past their reproductive age, and selection rarely acts on deleterious mutations in old age [25]. Therefore, maximum lifespan-associated proteins, such as SIRT6, must have been subject to relatively less selective pressure and have not been optimized to their fullest efficiency. Through the mutation of crucial sites, it is possible to design optimized human SIRT6 sequences that has higher enzymatic efficiency and is more conducive to longevity. Corresponding site variation patterns for longevity might also be found.

Based on the above assumptions, the study obtained mammalian SIRT6 sequence and Maximum lifespan (MLS) data from network data platforms such as NCBI and used the RoseTTAFold to predict the 3D structure of all sequences. Using original sequence information and predicted 3D structure information, deep learning models such as CNN and ESM were used to construct a model to predict MLS. The ProteinMPNN model [26] was used to generate human SIRT6 mutant sequences, and the established MLS prediction model was applied to screen out optimized mutant sequences. After that, DNA repair experiments were applied to verify their effect, and patterns of key site variation for SIRT6 optimization were summarized based on optimized sequences.

2. Materials & Methods

2.1. Data collection and pre-processing

SIRT6 ortholog sequence data for mammalian was downloaded from NCBI ortholog database (https://www.ncbi.nlm.nih.gov/gene/51548/ortholog/?scope=40674&term=SIRT6). MLS were retrieved from the database AnAge [15] (a sub-library of HAGR, https://genomics.senescence.info/species/index.html) and supplemented with the study "Universal DNA methylation age across mammalian tissues" [27], which provides further adjustments of AnAge.

Sequence and MLS data were matched according to the Latin standard names of the species. When sequences for two species with different MLSs were identical, only species with a higher MLS was retained. Finally, sequence and MLS data of 142 species were obtained for analysis. Muscle (MUltiple Sequence Comparison by Log-Expectation) algorithm [28] was used for MSA, and the length of SIRT6 after MSA is 685 amino acids (AA). Among them, 355 sites correspond to the position of human SIRT6 sites.

The distribution of MLS is positively skewed, and Human MLS is significantly larger than MLS of all other species, so to ensure the accuracy of the MLS prediction models, human sequence's

(NP_057623.2) will only be used to generate mutant sequences. For rest of 141 sequences, MLS stratified 5-fold cross-validation (CV) was applied. The correlation coefficients between predicted and measured MLS values of the validation set were calculated for each fold, and the mean value of the correlation coefficients of the 5 folds was used for model comparison. Considering the skewed distribution of MLS, both Pearson's correlation coefficient and Spearman's rank correlation coefficient were calculated for evaluation.

2.2. Modeling of MLS and protein sequence

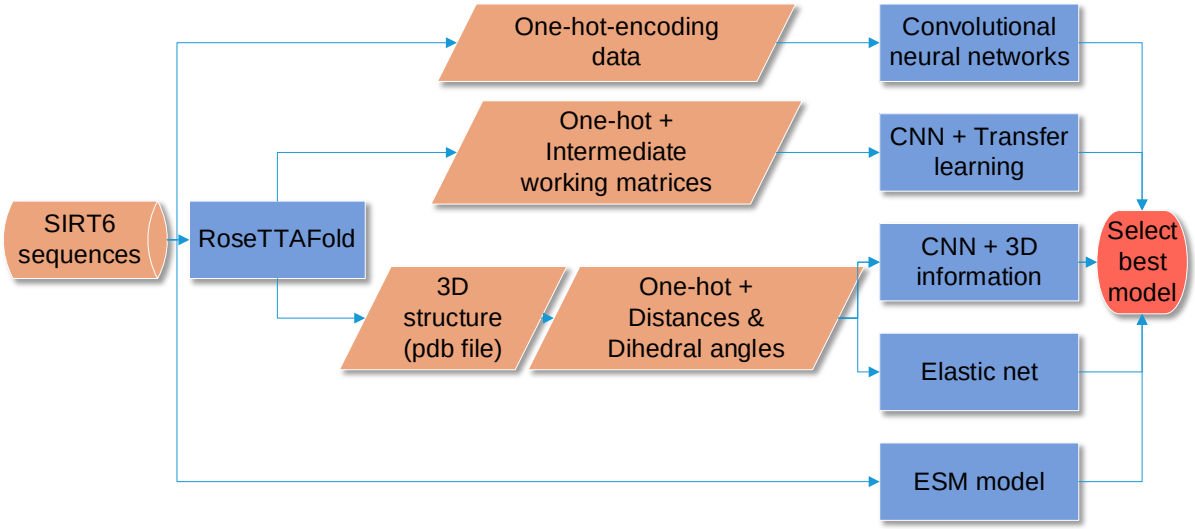


Fig. 1 Flowchart of establishing MLS prediction model

Considering the complex relationship of MLS with protein sequence, several statistical models were applied in this research (Fig 1). Convolutional Neural Networks (CNN) [29] were applied firstly, various model architectures including multi-layer convolutional layer were tried until the best is discovered. Using RoseTTAFold[18–20], distance and dihedral angles between sites of all sequences were calculated, which were used to train enhanced CNN model and ENET [30]. Intermediate working matrices of RoseTTAFold were extracted and trained in CNN using transfer learning. Finally, all sequences were inputted into the ESM model [21, 22], which utilizes both sequence and structure in its model architecture. The model with the best performance is selected. RoseTTAFold was accessed through the officially provided notebook on the Google-provided Colaboratory(Colab) platform [31], and all other models were implemented by customizing Python script based on pytorch 2.0 and scikit-learn 1.2.0 under Python 3.9.13 environment.

2.3. Searching optimized human SIRT6 sequences

In order to find mutant sequences more efficiently, we developed a sequence optimization method framework (Fig 2).

2.3.1 Screening candidate mutation sites

Since possible mutation of sequences when replacing all human SIRT6's functional region (27-272 sites) using ProteinMPNN is too huge to be fully explored in this research, candidate mutation sites were screened firstly by excluding sites that are conserved or don't impact MLS greatly using 3 methods: entropy (reflects relatively conserved sites), MSE (reflects cross entropy) and elastic net coefficient (reflects the importance to MLS). 159 candidate mutation sites were screened out for further analysis.

2.3.2 Generating mutant sequence

ProteinMPNN [26] was used for human SIRT6 mutant generation, which was implemented on Colab platform. The most effective MLS prediction model will be applied to screen out optimized mutant sequences, predicted MLS increase (PMI), or percent of MLS increase between the predicted

value of the new sequence and human SIRT6, was used as a standard for evaluating the optimization degree of mutant sequences, and PMI was averaged for 5 CV models to balance the influence of 5 CV models in evaluation.

2.3.3 Ranking of site importance in mutation

Based on generated optimized sequences, 159 candidate sites were ranked using backward stepwise elimination (BSE) to screen out crucial sites that increase MLS: sites were replaced back to the corresponding AA of human sequence separately in each sequence, and mean change of PMI after replacement is calculated. A negative mean change (decrease after replacement back) suggests that mutation at this site does improve PMI, while a positive value means that the site should not be mutated. The site with the most positive PMI change was considered first for removal (be replaced back), then sequentially other sites were replaced back and removed based on PMI change. Since the replacement started from the most positive site, the mean PMI of all sequences should first increase and then decrease with the decrease of mutated sites number. The highest point of the parabola corresponds to the subset of optimal subsets of variation sites. To search for the most concise group of crucial sites, several smaller subsets were also selected for analysis based on the change in descent speed of the parabola.

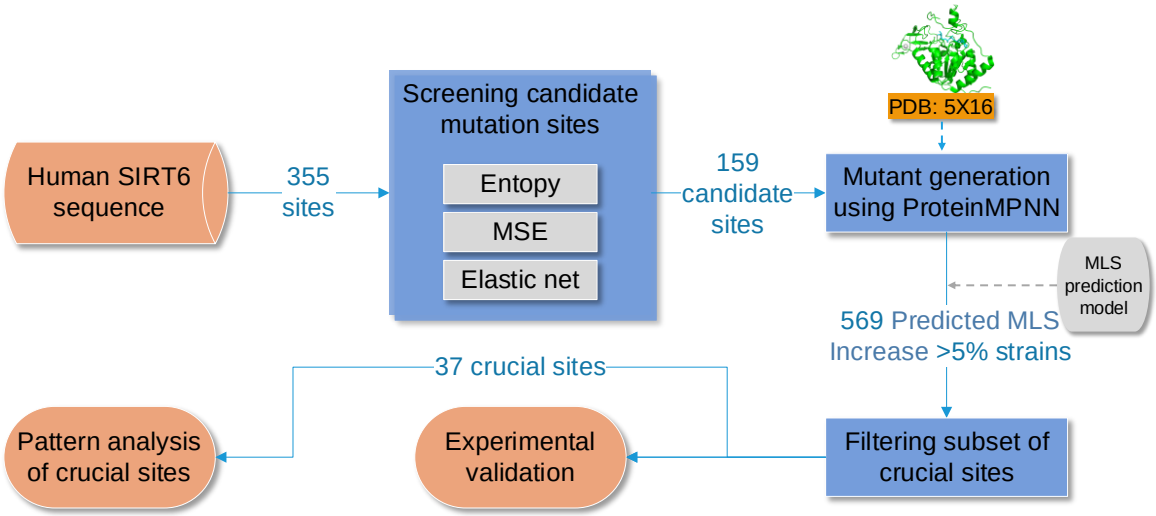


Fig. 2 Flowchart of searching optimized SIRT6 mutant sequences

2.4. Experimental confirmation of mutant SIRT6 efficiency

The designed novel sequences were translated back to nucleotide sequences based on the human codon usage bias. The resulting sequences were synthesized and cloned into pEGFP-N1 plasmid, replacing the EGFP sequence. The recombinant plasmids were then transformed into competent bacteria (E. coli strain DH5α) and the plasmids were isolated using a plasmid extraction kit. The cell line HEK293 was cultured in an appropriate growth medium at 37°C in a 5% CO2 incubator. At 70-80% confluency, transfect cells with the SIRT6 expression were constructed using a suitable transfection reagent. After 24-48 hours, selective pressure using G418 was applied to isolate cells that had incorporated the construct. The gRNAs targeting the endogenous SIRT6 gene were designed using sgRNA Scorer 2.0 and cloned into a suitable CRISPR/Cas9 vector, LentiCRISPR V2. The SIRT6 knockout vectors were co-transfected into the cell lines, then genotyping PCR was applied to identify clones where the SIRT6 gene has been successfully knocked out. The following 4 groups of cell lines were used for further experimentation:

- OE-NC: Primitive HEK293 cells + blank plasmids transfected.
- Seq-WT: SIRT6 knockout HEK293 cells + plasmids of human ortholog sequence (NP_057623.2) transfected.
- Seq-16: SIRT6 knockout HEK293 cells + plasmids of mutant sequences (predicted PMI = 16.5%) transfected.

• Seq-20: SIRT6 knockout HEK293 cells + plasmids of mutant sequences (predicted PMI = 20.1%) transfected.

Western blotting was applied to verify successful expression of SIRT6 proteins (raw human SIRT6 sequence or novel sequence), immunofluorescence and foci quantification was applied to compare the DSB repair efficiency of SIRT6 in each group. DSBs were induced by adding DNA damage inducer Methotrexate (10μM) to the cells, and then immunofluorescence staining with γ-H2AX antibodies was performed to visualize DNA damage response after 24h. The intensities per cell were normalized to the average OE-NC foci intensity, and differences between treatments were determined by one-way analysis of variance (ANOVA) using GraphPad Prism 8 software, as well as the least significant difference test (LSD) for post hoc test. A probability level of 5% ($p < 0.05$) was considered statistically significant.

3. Results

3.1. Establishing MLS predict model using SIRT6 sequences information

To perform high-throughput screening in silico on designed SIRT6 sequence, a model that can accurately predict MLS from SIRT6 sequence information needs to be developed in advance. CNN models were first trained based on all 685 AA after MSA of 141 species' SIRT6, and later from only the 355 AAs corresponding to human SIRT6. As seen in Table 1, the Pearson's r of the models is 0.685-0.701. Next, 3D structures of SIRT6 sequences (spatial distances & dihedral angles between sites), as well as RoseTTAFold's 3 intermediate working matrices in PDB structure prediction, were calculated and added into the CNN model. Moreover, ENET model was trained to see whether screening variables in the model can improve the result, but the performance of both models was lower than the original CNN model even with a more complex model architecture. Finally, esm2_t33_650M_UR50D of the ESM model was used in the analysis. The result shows that the prediction accuracy of ESM is much higher than the original CNN model ($r = 0.818$, $\rho = 0.750$,). So this model will be used as the criteria for determining the effect of mutant sequences on prolonging MLS in subsequent studies.

Table 1. Performance of different models for predicting MLS

Model Type & Variables included in Model	Model parameter settings	Metrics of validation set		
		MAE	Pearson's r	Spearman's ρ
CNN: 685 sites	Flatten/512/256/128/1	10.072	0.686	0.729
CNN: 685 sites	Conv2d(32,3)/Maxpoll2 /Flatten/512/256/128/1	10.466	0.678	0.722
CNN: Human 355 sites	Flatten/512/256/128/1	9.872	0.701	0.719
CNN: Human 355 sites	Conv2d(32,3)/Maxpoll2 /Flatten/512/256/128/1	10.030	0.685	0.709
CNN: human 355 sites + distance & dihedral angle of 355 sites	Flatten/2048/1024/512 /256/128/1	13.546	0.483	0.494
CNN: human 355 sites + transfer learning matrix from RoseTTAFold	Flatten/1024/512/256 /128/1	10.831	0.652	0.534
Elastic net: human 355 sites + distance & dihedral angle of 355 sites	L1_ratio = 0.0275, α= 1.1721	10.019	0.667	0.635
ESM: human 355 sites	**esm2_t33_650M_UR50D**	**8.608**	**0.818**	**0.750**

3.2. Generating optimized human SIRT6 sequences

Using 159 candidate mutation sites, 942,080 sequences were generated using ProteinMPNN based on human SIRT6 crystal structure (PDB: 5X16). 569 of them have been considered as optimized for their PMI > 5% and PMI of all 5 folds > 0%. Based on 569 optimized sequences, 159 sites were ranked on their importance to MLS using backward stepwise elimination based on PMI change, and

subsets of 104, 74, 50, and 37 AA sites respectively (Table 2, Fig 3A) were selected based on the descent speed's change of the PMI parabola curve (which reflects their importance to protein function). For each subset, 102,400 sequences were generated using ProteinMPNN, and their PMI distribution was summarized. The result (Fig 3B) shows that the median PMI of subset 159 was negative, suggesting that deleterious mutations were still included in this subset. For smaller subset, all their median PMI became positive, and their maximum PMI were all higher than subset 159, indicating that using smaller subsets is indeed more advantageous for sequence optimization. Moreover, as subset size decreases, the percentage of sequence for PMI > 0% increased, and it reached 98.2% for subset 37, indicating that subset 37 indeed includes sites that are most beneficial in improving MLS. However, the maximum PMI of 20.1% belonging to subset 74 instead of the expected subset 37. This suggests that relying solely on a small number of crucial sites might not be enough to generate the most optimized sequence, as sequence optimization may require the participation of multiple sites to form the most favorable spatial structure for its biological functions.

Table 2. Site list for different site subset

# of sites in subset	Site list
104 (optimal subset)	Subset 74 sites + 30, 32, 62, 80, 92, 144, 154, 160, 162, 163, 165, 185, 186, 206, 223, 225, 229, 233, 234, 237, 240, 243, 247, 250, 251, 254, 262, 270, 272, 273
74	Subset 50 sites + 34, 41, 66, 84, 90, 120, 151, 156, 157, 161, 169, 180, 198, 207, 208, 209, 212, 213, 220, 232, 235, 238, 239, 259
50	Subset 37 sites + 37, 40, 100, 102, 152, 167, 168, 171, 194, 196, 211, 217, 260
37	28, 29, 31, 33, 36, 38, 39, 43, 44, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 58, 60, 61, 63, 64, 65, 68, 76, 79, 82, 142, 145, 146, 205, 215, 219, 228, 256

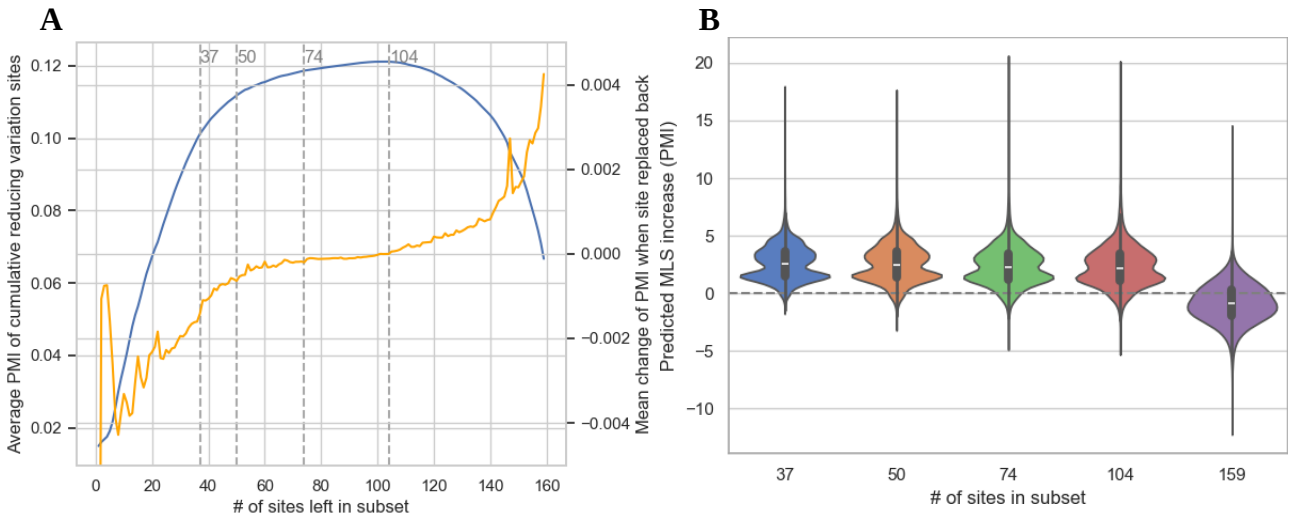


Fig. 3 Splitting of AA site subsets, and PMI distribution for mutant sequence of site subset. (A) Parabola curves of average PMI of cumulative reducing variation sites correspond to the No. of sites left in the subset. (B) Violin plot of PMI for a mutant sequence of different site subsets.

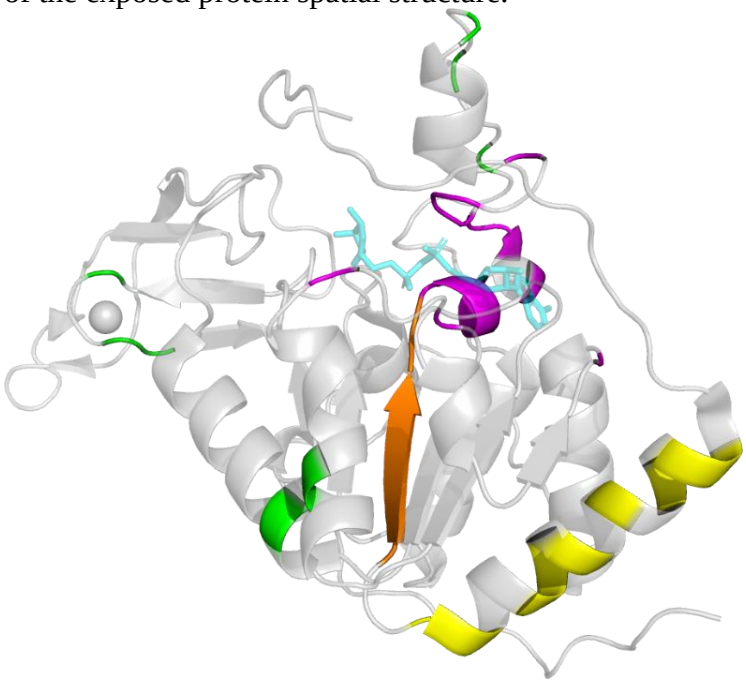
3.3. Distribution pattern of crucial sites in sequence optimization

3.3.1 Spatial location of crucial sites

The location of the crucial sites in subset 37 (the most concise subset) can be divided into 4 subregions (Fig 4).

- NAD+ binding domain sites: This binding domain involves 14 sites located in the periphery of the binding domain in terms of 3D structure, including the minor structural region of α2-α3 (53-68) that constitutes the back of the binding domain, the region of the lower edge of the binding domain (213-220) that constitutes by the Rossmann folding, and site 256 at the edge of the binding domain.
- β1-sheet: 6 consecutive sites 47-52 forms β1-sheet in Rossmann fold internal.

• α1-helix: A total of 9 sites were screened in the 26-44 region where the α1-helix is located.
• Other area: sites 142, 145, and 146 are located around the flexible loop of the SIRT6-specific zinc ion binding domain, while the remaining 5 sites are scattered around the periphery of the long and wide pocket region of the hydrophobic channel of the sirt6 protein, as well as in various portions of the exposed protein spatial structure.



For sites belong to subset 37:
Yellow: belongα1-helix.
Magenta: belong NAD+ binding area.
Orange: belong β1-sheet.
Green: belong other area.

Gray: sites not belong to subset 37.
Cyan: ADP-ribose.

Fig. 4 The position of the subset 37's site in the spatial structure of SIRT6

3.3.2 Estimated importance of each spatial subregion of subset 37

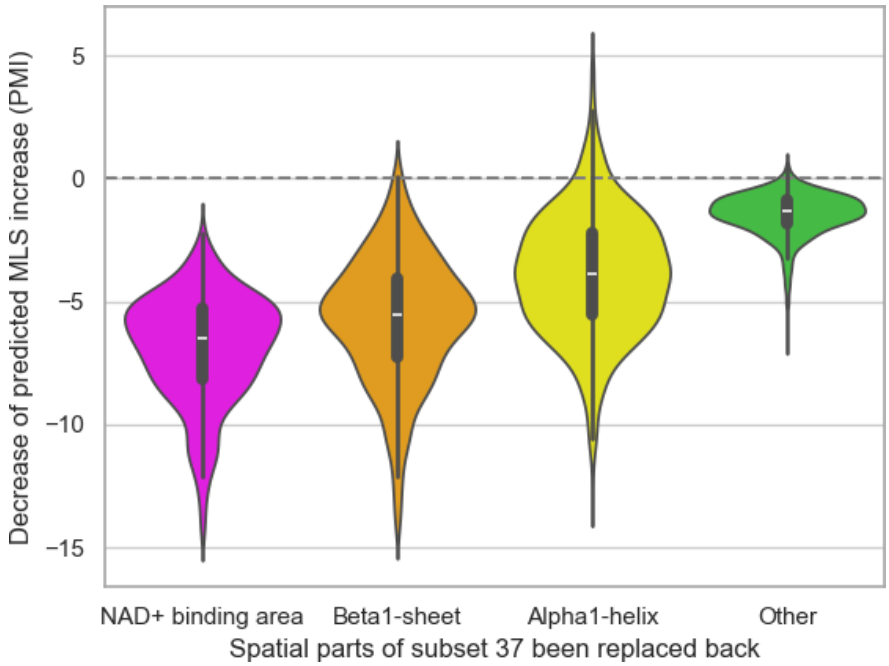


Fig. 5 Violin plot of PMI decreases for replacing back of different spatial subregions of subset 37

Based on 569 optimized sequences, the importance of subset 37's 4 subregions in sequence optimization were estimated by replacing them back to the original human AA separately. Fig 5 shows that when the NAD+ banding area or β1-sheet is replaced back, the mean PMI decrease is higher than 5%. The change in α1-helix led to relatively weaker decrease in mean PMI, and decrease caused by other sites is even weaker than that of α1-helix. These results fully demonstrate that in subset 37, NAD+ banding area and β1-sheet most significantly affects sequence optimization.

3.4. Experimentation proves increased DSB repair efficiency for optimized sequences

```
                *       20        *       40        *       60        *       80        *      100
Seq-WT : MSVNYAAGLSPYADKGKCGLPEIFDPPEELERKVWELARLVWQSSSVVFHTGAGISTASGIPDFRGPHGVWTMEERGLAPKFDTTFESARPTQTHMALVQ : 100
Seq-20 : ........................QR.R.N..Q.QK...L....CA..PS.VT.S..L.ATTS.D.......Q..K..P.P..............  : 100
Seq-16 : ........................D..N....T..E...EA...NE.STS.N..S....ALN..D.......T..L.....................  : 100

                *      120        *      140        *      160        *      180        *      200
Seq-WT : LERVGLLRFLVSQNVDGLHVRSGFPRDKLAELHGNMFVEECAKCKTQYVRDTVVGTMGLKATGRLCTVAKARGLRACRGELRDTILDWEDSLPDRDLALA : 200
Seq-20 : .A.................I.....................S..LR....EK...KR...R.....LRP.D........I.............T.A.T..  : 200
Seq-16 : ........................................T..NV.....R.............E...D.....................Q......  : 200

                *      220        *      240        *      260        *      280        *      300
Seq-WT : DEASRNADLSITLGTSLQIRPSGNLPLATKRRGGRLVIVNLQPTKHDRHADLRIHGYVDEVMTRLMKHLGLEIPAWDGPRVLERALPPLPRPPTPKLEPK : 300
Seq-20 : ....S.....E.......QE.......K...S..K.........................Q.....................................  : 300
Seq-16 : ....S.........Q.Q.L........Y.........................A...D.....................................  : 300

                *      320        *      340        *
Seq-WT : EESPTRINGSIPAGPKQEPCAQHNGSEPASPKRERPTSPAPHRPPKRVKAKAVPS : 355
Seq-20 : ..................................................... : 355
Seq-16 : ..................................................... : 355
```

Fig. 6 Alignment of Seq-20, Seq-16 and human SIRT6 sequences (Seq-WT)

To further confirm the effect of sequence optimization, we took 2 sequences with PMI of 20.1% (Seq-20) and 16.5% (Seq-16) for experimental validation (Fig 6). The number of sites mutated in Seq-20 and Seq-16 are 51 and 33, respectively, with 30 and 27 sites belonging to subset 37, which further suggested that those included in subset 37 are the most important sites.

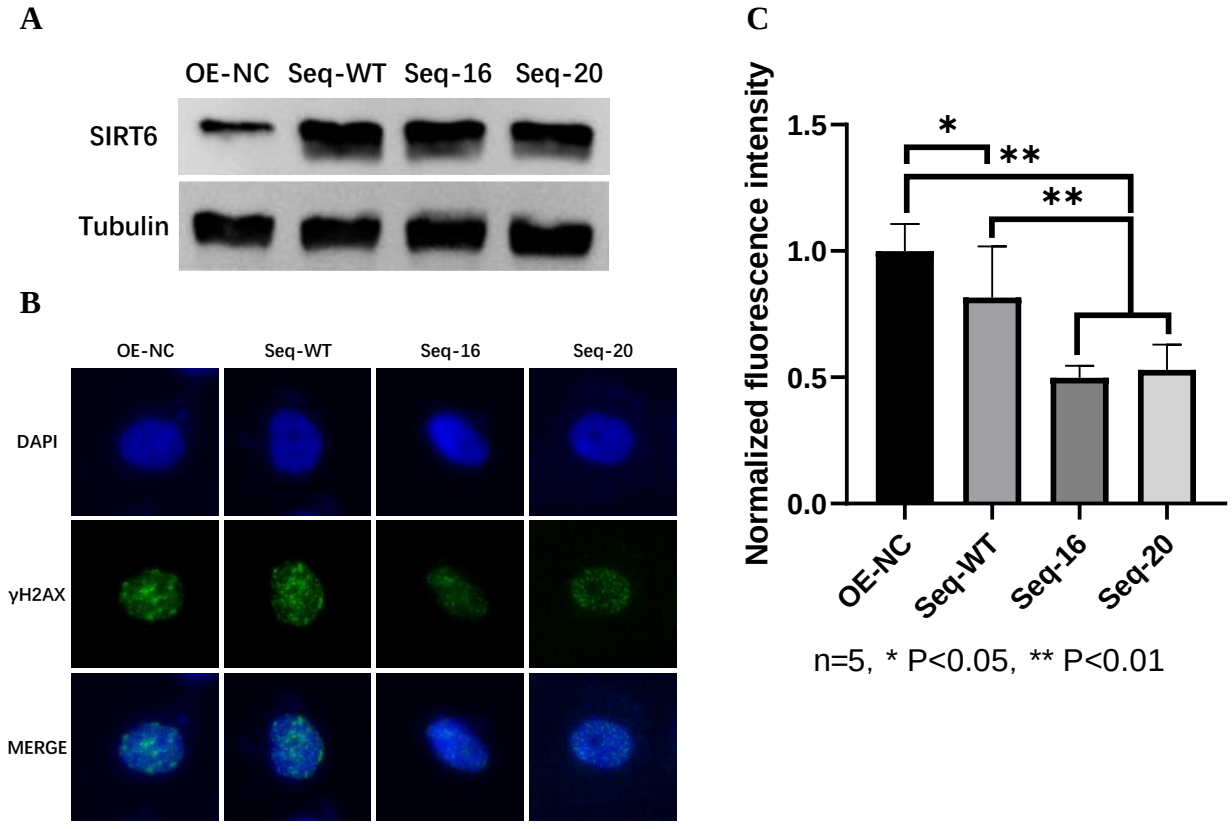


Fig. 7 Comparison of DNA damage repair by γ-H2AX immunofluorescence. (A) Result of western blotting for 4 groups of HEK293. (B) Representative fluorescent micrographs of HEK293 cells at 24h after being treated with Methotrexate. (C) Comparison of normalized fluorescence intensity

Because a full length lifespan assay is not plausible in the scope of this research, to assess the enzymatic activity of the optimized SIRT6, we decided to investigate its DSB repair efficiency, which also relates to its maximum lifespan[10]. The result of western blotting (Fig 7A) shows that all SIRT6 proteins (raw sequence or novel sequence) in each group were successfully expressed. DSB repair efficiency was tested using γ-H2AX immunofluorescence and foci quantification at 24h after treating

HEK293 cells with Methotrexate (Fig 7B and 7C). Compared to the normalized fluorescence intensity of OE-NC and Seq-WT (1.000 and 0.815 separately), lower normalized mean fluorescence intensity was observed in the cells of Seq-16 and Seq-20 (0.498 and 0.528 separately, $n = 5$, $P < 0.01$), indicating that the 2 novel sequences are more potent at repairing Methotrexate-induced DNA damage, therefore having higher DSB repair efficiency than raw human SIRT6 sequence.

4. Discussion

Analysis shows that for spatial regions of subset 37, NAD+ binding area and β1-sheet play the most important role in sequence optimization. Variation of sites located in the NAD+ binding domain directly changes the spatial structure of the binding domain [9], thus affecting the deacetylase activity of SIRT6. 6 consecutive sites in subset 37 are in β1-sheet, which constitutes the Rossmann fold internal region [32]. Although it does not have direct contact with ligands, β1-sheet is in the middle of four large alpha helixes, serves as one of the three pleated sheets at the center of the protein, and is connected directly to NAD+ binding domain, so mutations in this part of the sequence might affect the overall structure of the protein or lead to changes in the NAD+ binding domains. It is worth pointing out that 11 sites in subset 37 are located on the cofactor binding loop (β1-α2 loop) [33], which is highly dynamic in structure and plays an important role in catalyzing reactions. The concentrated distribution of crucial sites on the cofactor binding loop also indicates that our findings are consistent with the results of existing research. Further in-depth analysis of how each specific site will affect protein's function will be a topic worthy of research.

Although immunofluorescence using γ-H2AX indicates that 2 generated sequences did increase the DSB repair efficiency of HEK293 cells, limitations of the lab results should be noted. Since SIRT6 efficiency for all 142 species is relatively hard to obtain and previous studies have found a strong correlation between DSB repair efficiency and MLS ($r2=0.76$) [10], maximum lifespans are used to reflect protein function. Ideally, an experiment on the lifespan of organisms (e.g., transgenic mouse overexpressing optimized Human SIRT6 genes [34]) should be conducted, but this is not possible to implement in our study due to limited time and resources. Therefore, currently conducted experiments measure DSB repair efficiency in one single cell type, but the discrepancy between MLS and DSB repair efficiency should still be pointed out. Further research may focus on conducting lifespan assays on organisms. On the other hand, only 2 of the hundred generated sequences with a large PMI were experimentally validated due to resource constraints. Although the 2 sequences did improve DSB repair efficiency, this does not suggest that all generated sequences have improved functions for false positives, dysfunctional folds, and inaccurately predicted sequences might exist due to limitations in algorithms and predictions. If all sequences can be batch-tested and verified, the experimental data could be used to validate proposed models, discover patterns between predicted values and actual values, and investigate the relationship between mutated subsets and protein function, therefore helping researchers gain a deeper understanding of pattern sequence optimization.

5. Conclusion

Using information from both sequence, structure, and amino acid selection, we trained a model that accurately predicts species' MLS from SIRT6 sequence ($r=0.818$). Despite the highly conserved nature of SIRT6 throughout evolution, our research reveals that there is still potential for its function optimization and generated 2 novel sequences with experimentally validated increases in DSB repair efficiency. Patterns in the SIRT6 function and AA selection were also summarized, and a subset of 37 AAs were found to play important roles in MLS optimization. Among them, 20 sites in the NAD+ binding domain and β1-sheet have the most significant impact on SIRT6 function. This study not only identified key sites correlated with human SIRT6 sequence optimization but also designed optimized SIRT6 proteins with longer MLS and higher efficiency, thereby providing novel insights into potential anti-aging or anti-cancer interventions. Moreover, the study developed a comprehensive

framework of sequence optimization methods, which can be systematically applied to the optimization research on other functional proteins whose efficiency is relatively harder to obtain or measure directly.

References

[1] Chang AY, Skirbekk VF, Tyrovolas S, et al. Measuring population ageing: an analysis of the Global Burden of Disease Study 2017. Lancet Public Health, 2019, 4(3):e159–67.

[2] Aronson L. Healthy Aging Across the Stages of Old Age. Clin Geriatr Med, 2020, 36(4):549–58.

[3] MacRae SL, Croken MM, Calder RB, et al. DNA repair in species with extreme lifespan differences. Aging, 2015, 7(12):1171–82.

[4] Ui A, Chiba N, Yasui A. Relationship among DNA double-strand break (DSB), DSB repair, and transcription prevents genome instability and cancer. Cancer Sci, 2020, 111(5):1443–51.

[5] Gorbunova V, Seluanov A. DNA double strand break repair, aging and the chromatin connection. Mutat Res Mol Mech Mutagen, 2016, 788:2–6.

[6] López-Otín C, Blasco MA, Partridge L, et al. The Hallmarks of Aging. Cell, 2013, 153(6):1194–217.

[7] Kowalczyk A, Partha R, Clark NL, et al. Pan-mammalian analysis of molecular constraints underlying extended lifespan. eLife, 2020, 9:e51089.

[8] Guo Z, Li P, Ge J, et al. SIRT6 in Aging, Metabolism, Inflammation and Cardiovascular Diseases. Aging Dis, 2022, 13(6):1787–822.

[9] Sharma A, Mahur P, Muthukumaran J, et al. Shedding light on structure, function and regulation of human sirtuins: a comprehensive review. 3 Biotech, 2023, 13(1):29.

[10] Tian X, Firsanov D, Zhang Z, et al. SIRT6 Is Responsible for More Efficient DNA Double-Strand Break Repair in Long-Lived Species. Cell, 2019, 177(3):622-638.e22.

[11] You Y, Liang W. SIRT1 and SIRT6: The role in aging-related diseases. Biochim Biophys Acta BBA - Mol Basis Dis, 2023, 1869(7):166815.

[12] Tasselli L, Zheng W, Chua KF. SIRT6: Novel Mechanisms and Links to Aging and Disease. Trends Endocrinol Metab, 2017, 28(3):168–85.

[13] Roichman A, Elhanati S, Aon MA, et al. Restoration of energy homeostasis by SIRT6 extends healthy lifespan. Nat Commun, 2021, 12(1):3208.

[14] TenNapel MJ, Lynch CF, Burns TL, et al. SIRT6 Minor Allele Genotype Is Associated with >5-Year Decrease in Lifespan in an Aged Cohort. PLoS ONE, 2014, 9(12):e115616.

[15] Tacutu R, Craig T, Budovsky A, et al. Human Ageing Genomic Resources: Integrated databases and tools for the biology and genetics of ageing. Nucleic Acids Res, 2012, 41(D1):D1027–33.

[16] Callaway E. 'It opens up a whole new universe': Revolutionary microscopy technique sees individual atoms for first time. Nature, 2020, 582(7811):156–57.

[17] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596(7873):583–89.

[18] Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science, 2021, 373(6557):871–76.

[19] Mansoor S, Baek M, Juergens D, et al. Accurate Mutation Effect Prediction using RoseTTAFold. 2022[Online] Biochemistry 2022.

[20] Mirdita M, Schütze K, Moriwaki Y, et al. ColabFold: making protein folding accessible to all. Nat Methods, 2022, 19(6):679–82.

[21] Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 2023, 379(6637):1123–30.

[22] Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci, 2021, 118(15):e2016239118.

[23] Pearce R, Huang X, Omenn GS, et al. De novo protein fold design through sequence-independent fragment assembly simulations. Proc Natl Acad Sci, 2023, 120(4):e2208275120.

[24] Tong X, Liu X, Tan X, et al. Generative Models for De Novo Drug Design. J Med Chem, 2021, 64(19):14011–27.

[25] Giaimo S. Medawar and Hamilton on the selective forces in the evolution of ageing. Hist Philos Life Sci, 2021, 43(4):124.

[26] Dauparas J, Anishchenko I, Bennett N, et al. Robust deep learning–based protein sequence design using ProteinMPNN. Science, 2022, 378(6615):49–56.

[27] Lu AT, Fei Z, Haghani A, et al. Universal DNA methylation age across mammalian tissues. Nat Aging, 2023, 3(9):1144–66.

[28] Edgar RC. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. Nat Commun, 2022, 13(1):6968.

[29] Montesinos López OA, Montesinos López A, Crossa J. Convolutional Neural Networks. In: Multivariate Statistical Machine Learning Methods for Genomic Prediction Springer International Publishing: Cham 2022, pp. 533–77.

[30] Zou H, Hastie T. Regularization and Variable Selection Via the Elastic Net. J R Stat Soc Ser B Stat Methodol, 2005, 67(2):301–20.

[31] Bisong E. Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform Apress: Berkeley, CA 2019, pp. 59–64.

[32] Beauharnois JM, Bolívar BE, Welch JT. Sirtuin 6: a review of biological effects and potential therapeutic properties. Mol Biosyst, 2013, 9(7):1789.

[33] Pan PW, Feldman JL, Devries MK, et al. Structure and Biochemical Functions of SIRT6. J Biol Chem, 2011, 286(16):14575–87.

[34] Zhang Z, Tian X, Lu JY, et al. Increased hyaluronan by naked mole-rat Has2 improves healthspan in mice. Nature, 2023, 621(7977):196–205.