参赛队员姓名：张晏菲

中学：上海星河湾双语学校

省份：上　海

国家/地区：中　国

指导教师姓名：阮振超

指导教师单位：上海星河湾双语学校

论文题目：Optimizing human SIRT6 protein with deep learning of 3D structures based on maximum lifespan

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

Yanfei Zhang

Abstract

DNA damage, particularly double-strand breaks (DSB), plays an important role in aging, carcinogenesis, and other diseases. The efficacy of the DSB repair protein SIRT6 is known to correlate with the maximum lifespan (MLS) across species. However, it is still unclear whether the function of SIRT6 can be further optimized through protein sequence engineering. Here, we used RoseTTAFold to predict the structural variance of SIRT6 sequences across 142 mammalian species. We then analyzed the association between the MLS, sequence, 3D structures, and amino acid selection of SIRT6 and found that sequence and spatial information are correlated with MLS. By fine-tuning the ESM-fold model, we were able to accurately predict 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 based on their impact on predicted MLS. A subset of 37 amino acids sites that play a key role in sequence function was found, and among them, 20 sites located in the NAD+ banding area and β1-sheet were found to have the greatest impact. We then tested the DNA repair efficiency of 2 novel SIRT6 sequences (with predicted MLS improvement of 16.5% and 20.1%), and immunofluorescence showed that their DSB repair efficiency is indeed higher than human ortholog SIRT6 sequence. Together, our study demonstrates that despite being highly conservative in evolution, there is still room for optimization of human SIRT6. We not only identified crucial sites correlated with sequence optimization but also designed optimized SIRT6 protein with longer MLS, thereby providing novel insights into potential anti-aging or anti-cancer interventions. This study also developed a comprehensive framework of sequence optimization methods for functional proteins whose efficiency is relatively harder to obtain or measure directly.

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

Table of Contents

1. Introduction

1.1 Research background

Aging is the largest risk factor for many chronic diseases[1]. In 2020, it was estimated that there were approximately 700 million individuals aged 65 or older worldwide and the population is still growing rapidly. 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. Among the various causes of aging, the progressive accumulation of genetic damage is critical, and failure to repair DNA damage contributes to genomic instability, which is a major cause of cell senescence and aging[2–5]. The efficiency of double-strand breaks (DSB) pathway repair may be most critical for aging because DSBs not only alter gene sequences but also cause epigenetic changes by altering higher-order chromatin structures, leading to global gene transcriptional dysregulation. Not only does DSB contributes significantly to ageing, it also leads to the development of cancer[6], and the efficiency of DSB repair pathways is positively correlated with maximum lifespan[7,8].

Sirtuin 6 (SIRT6), a member of the SIRT protein family, possesses the enzymatic activities of NAD+-dependent histone deacetylase and mono-ADP ribosyltransferase and plays a role in long-chain lipid deacylation[9]. Research found that SIRT6 is an important factor in the regulation of the DSB repair pathway, and its function indirectly promotes longevity[10,11]. Direct activation of SIRT6 through agonists has been shown to extend lifespan, and transgenic mice overexpressing SIRT6-exhibited an increase in lifespan of up to 30%[12]. 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[13]. Therefore, SIRT6 sequences with higher efficiency repair DSB at a faster rate and thus correspond to longer lifespan in mammals. Such sequences have broad application prospects in gene therapy, new drug design, and the exploration of novel drug targets.

Maximum lifespan varies widely among mammalian species, ranging from 16 months in shrews to 211 years in bowhead whales[14]. This makes comparison of genetic differences between species (comparative genomics) a valuable field of research. In a previous study[15], Tian et al. 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 difference in protein activity that resulted in 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.

1.2 Progress in methodology

Traditionally, protein structure is obtained by synthesizing the corresponding sequence and then observing molecule structure using cryo-electron microscopes. Although the obtained structure is precise[16], the approach is costly and inefficient, making large-scale protein structure determination almost impossible. Therefore, previous research focused primarily on the relationship between sequence and function, while research utilizing the protein's 3D structure information in analysis was less common. However, bioinformatic methods for predicting protein structure based on deep neural networks 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. The accuracy of the predicted protein backbone structure is similar to that of cryo-electron microscopy, which made possible high-throughput obtaining of protein structure using deep learning only. Based on the above model, generative models[23–26] have been developed for de novo protein design for a desired structure, or generating new folds of artificial proteins containing a desired functional site. The rapid progress in methodologies described above 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.

1.3 Research content

This study is based on the following assumption: the force of natural selection declines as animals live past their reproductive age, and selection rarely acts on deleterious mutations in old age[27,28]. 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 with higher DSB repair efficiency and 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 (Figure 1). 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[29] (MPNN) 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.

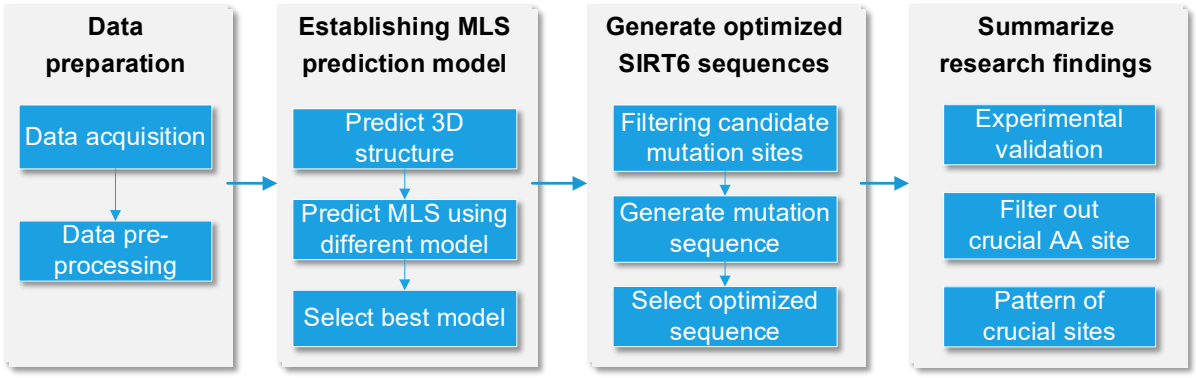


Figure 1. Overall flowchart of the research. After data preparation, the 3D structure of SIRT6 was predicted using RoseTTAFold. Different models were then fitted to predict MLS, and the most accurate model was selected to evaluate the most optimized sequence. When generating new sequences, possible mutation sites that greatly impacted MLS were filtered out and mutated using MPNN to generate new sequences. The selected MLS prediction model was applied to screen out sequences with higher functions. Finally, experiments were applied to verify their optimization, and patterns of crucial site variation for SIRT6 optimization were summarized.

2. Materials & Methods

2.1 Data collection and pre-processing

2.1.1 Data acquisition

The sequence data used in this study was downloaded from NCBI. Only ortholog sequences of each mammalian species were used. 203 SIRT6 sequences were retrieved from NCBI ortholog database (https://www.ncbi.nlm.nih.gov/gene/51548/ortholog/?scope=40674&term=SIRT6).

Because DSB repair efficiency for each species is hard to obtain, maximum lifespan (MLS) data, which positively correlates with DSB repair efficiency (r^2=0.76), was used in this study to reflect protein function. MLS were retrieved from the database AnAge[14] (a sub-library of HAGR, https://genomics.senescence.info/species/index.html) and supplemented with the

study "Universal DNA methylation age across mammalian tissues"[30], which provides further adjustments to some data in AnAge.

After sequences and MLS data were retrieved, they were matched according to the Latin standard names of the species. Each sequence was assigned a maximum lifespan of its corresponding species. When sequences for two species with different MLSs were identical, only the species with a higher MLS was retained. Finally, sequence and MLS data of 142 species were obtained for analysis.

2.1.2 Data pre-processing

Multisequence alignment (MSA) and data encoding

Muscle (MUltiple Sequence Comparison by Log-Expectation) algorithm[31] 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. After MSA, the AA sequences were transferred to a 0/1 matrix for modeling using the "one-hot encoding" technique.

Processing of human sequence

Human MLS is significantly larger than MLS of all other species for it reflected not only the natural DNA repairing abilities of enzymes but also advanced technology and medical care of the modern world. To avoid the human sequence's (NP_057623.2) interference with the model's ability to learn natural patterns, it was excluded from both the train and test set of all models and will only be used to generate mutant sequences. A remaining 141 species data was used in the MLS prediction model to ensure the robustness of the model results.

2.1.3 Data splitting

The distribution of MLS is positively skewed (Figure S1), so to ensure the accuracy of the prediction models and fully utilize the 141 sequences, MLS stratified 5-fold cross-validation (CV) was applied for the data.

To facilitate the comparison of different model results, the correlation coefficients between the 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.

2.2 Modeling of MLS and protein sequence

Considering the complex relationship of MLS with protein sequence, several statistical models ranging from the classical Elastic net model to the latest protein language model were applied in this research (Figure 2). RoseTTAFold was accessed through the officially provided notebook, 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.

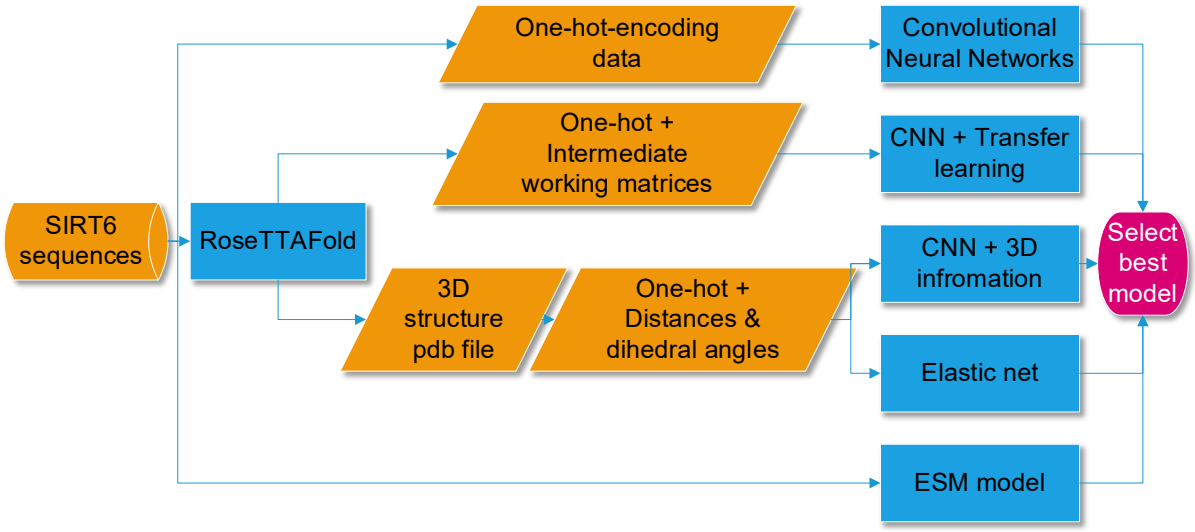


Figure 2. Flowchart of establishing MLS prediction model. Sequence information from 141 mammals was directly used to train CNN models. Using RoseTTAFold, distance and dihedral angles between sites of all sequences were calculated, which were used to train enhanced CNN models and Elastic net models. Intermediate working matrices of RoseTTAFold were extracted and trained in CNN using transfer learning. Finally, all sequences were inputted into the ESM model, which utilizes both sequence and structure in its model architecture. The model with the best performance is selected.

2.2.1 CNN model

Convolutional Neural Networks (CNN)[32] were applied, and various model architectures including multi-layer convolutional layer were tried until the best is discovered Since complex model architecture might increase the risk of overfitting, early stopping was also used, and the model stopped fitting if the loss function of the validation set decreases 10 times in a row.

2.2.2 RoseTTAFold

Using RoseTTAFold[18–20], the 3D PDB structures of all sequences were calculated. RoseTTAFold model was accessed through the officially provided notebook file RoseTTAFold.ipynb, but was further modified to loop through the execution and to obtain intermediate working matrices required for transfer learning[33] (Link of notebook code:

https://colab.research.google.com/drive/1whVfMQ-syuXFCzv7RokTMt6_g6B6JEep?usp=sharing). All sequences were computed using the same version of the code on the Google-provided Colaboratory (Colab) platform[34] over consecutive days to ensure that both versions of the RoseTTAFold model used in the analysis and the parameter settings were the same.

The PDB structural file, the pLDDT values corresponding to each site of the PDB, as well as 3 intermediate working matrices (feature_extractor, c6d_predictor, and refine) were saved. Then, spatial distances between sites as well as dihedral angles φ and ψ between sites were calculated based on 3D structure. For sequences that have missing sites after MSA, interpolation was used to fill in the missing values of distances and dihedral angles.

2.2.3 Elastic net

CNN is useful in extracting data information but not in screening variables, which becomes ineffective when information in the data is too sparse. The Elastic net model (ENET)[35,36], which combines Ridge regression and Lasso regression, can quickly screen out effective variables from a large number of candidate variables for analysis, and its results may be useful for finding the crucial mutant site. Optimal values of model parameter L1_ratio and α were determined using grid search by dividing data into training/validation set in a 7:3 ratio. The ENET was then retrained in 5-fold CV using the obtained optimal parameter settings.

2.2.4 ESM model

The ESM model[21,22], a protein language model developed by Meta, is a migratory application of language modeling in the biological field. Its underlying principle is that the protein sequence of an organism is not a random permutation of amino acids, but is subject to natural selection. Thus, its statistical patterns should imply structural information, and if a locus conservatively selects a certain amino acid or amino acids, it suggests that only the biochemical properties of these few amino acids can be adapted to the structure here. Therefore, this study also considered the ESM model for transfer learning. esm2_t33_650M_UR50D, the largest ESM model that can be used to predict continuous variables, was implemented in this research.

2.3 Searching optimized human SIRT6 sequences

In order to find mutant sequences more efficiently, we developed a sequence optimization method framework (Figure 3). Firstly, candidate mutation sites were screened out using several methods. Secondly, mutation sequences were generated using MPNN, and optimized sequences were selected based on predicted MLS increase. Finally, crucial subset of AA site

was further filtered based on the optimized sequences, and the mutation pattern of crucial sites was summarized.

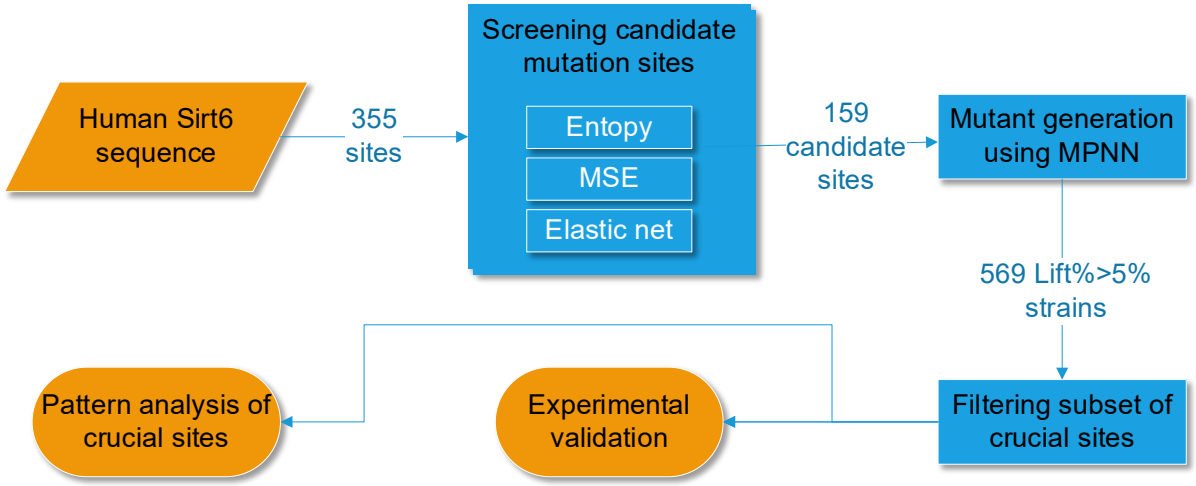


Figure 3. Flowchart of searching optimized SIRT6 mutant sequences. To reduce the computational complexity, the 355 sites corresponding to human SIRT6 were first screened based on entropy (reflects relatively conserved sites), MSE (reflects cross entropy) and elastic net coefficient (reflects the importance to MLS). The 159 candidate mutation sites were then inputted into MPNN to generate mutant sequences, and crucial sites whose mutation improves MLS in mutant sequences were screened. After obtaining subsets of crucial sites with the greatest impact on MLS, MPNN was used to generate two new sequences with predicted MLS increase of 16.5% and 20.1%, which was later validated *in vivo*. Patterns with crucial sites were also concluded and found.

2.3.1 Screening candidate mutation sites

Since site 27-272 of Human SIRT6 has been confirmed as SIRT6's functional region, sites will only be mutated in the slightly expanded 26-276 region. Since the total possible mutation of sequences when replacing all sites using MPNN is too huge to be fully explored in this research, 3 methods were applied to reduce the number of candidate mutation sites by first excluding sites that are conserved or don't impact MLS greatly.

1. Entropy: The entropy value for AA selection of each 26-276 site was calculated, and sites with entropy <= 0.4 were considered as conservative and excluded. 127 sites with entropy > 0.4 were considered as candidate mutation sites (Figure S2).
2. MSE: The Mean Square Error (MSE, which is the equivalence of cross entropy in screening sites) against the MLS is calculated for AA selection of each of the 26-276 sites. 131 sites with MSE < 335 were considered as candidate mutation sites (Figure S2).

3. Elastic net: Since ENETS screened out useful variables in modeling, 46 sites corresponding to important coefficients were selected as candidate mutation sites based on the regression coefficients of the ENET.

Finally, the sites from the above 3 methods were merged and a total of 159 candidate mutation sites (Table S1) were used in further analysis.

2.3.2 Generating mutant sequence

ProteinMPNN (MPNN)[29] is a protein design engine released in June of 2022 by Baker Lab. The tool effectively combines Rosetta's 10+ years of experience in protein design with deep learning methods by complementing physically based methods with deep learning-based approaches trained on large numbers of protein structures, which greatly outperforms similar tools in terms of computational speed and prediction accuracy. MPNN was used for human SIRT6 mutant generation, which was implemented on Colab using notebook (https://colab.research.google.com/drive/1EpHMqmEp1d8_ufBuDa2zN4kEksGNLYRX?usp = sharing).

The 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 (sequences with PMI > 5% and PMI of all 5 folds > 0%), 159 candidate sites were ranked using backward stepwise elimination (BSE)[37] 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 (the replacement of the last site kept in the replacement cycle of the next step) 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.

2.4 Experimental confirmation of mutant SIRT6 efficiency

2.4.1 Cell line generation of *de novo* SIRT6[15,38]

The designed novel sequences were translated back to nucleotide sequences based on the human codon usage bias[39]. The resulting sequences were synthesized and cloned into a 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.

2.4.2 Protein expression level assessment

Expression levels of WT and novel SIRT6 sequences were assessed using western blotting. Proteins were lysed on ice for 30 min using RIPA lysis buffer with a protease inhibitor cocktail and PMSF. Lysates were subjected to sonication and then centrifuged at 12,000 rpm for 20 min at 4°C. The supernatants were collected. Protein concentration was measured using a BCA

protein assay kit. The protein extracts were supplemented with 5×sodium dodecyl-sulfate (SDS) loading buffer and boiled for 10min. Proteins were separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and then transferred to polyvinylidene difluoride (PVDF) membranes, which were blocked in 5% BSA for 1h at room temperature. The membranes were incubated with primary antibodies overnight at 4°C and then washed. The secondary antibodies were added to the membrane for 1h at room temperature and then washed extensively. Subsequently, the immunoreactive bands were detected using a chemiluminescence reagent.

2.4.3 Immunofluorescence and foci quantification[40]

DSBs were induced by adding DNA damage inducer Methotrexate (1μM) to the cells, and then immunofluorescence staining with γ-H2AX antibodies was performed to visualize DNA damage response after 24h. Cells were seeded and cultured in a 12-well plate. After treatment, cells were fixed in 4% paraformaldehyde for 30 min at room temperature. The cells were then permeabilized with 0.5% Triton X-100 for 10 min. After that, slides were blocked in 5% BSA for 1h and then incubated with primary antibodies in 5% BSA overnight at 4 °C, followed by fluorescent secondary antibodies at room temperature for 1 h. Cell nuclei were stained with DAPI. Fluorescence images were captured using Echo Revolve microscope.

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), as well as the least significant difference test (LSD) for post hoc test. A probability level of 5% ($p < 0.05$) was considered 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. Various model architectures including multi-layer convolutional layer were tried, and as seen in Table 1, the Pearson's r of the models is 0.685-0.701. This suggests that there is some correlation between sequence's one-hot-encoding data and MLS directly, though the correlation is not strong enough to predict MLS precisely.

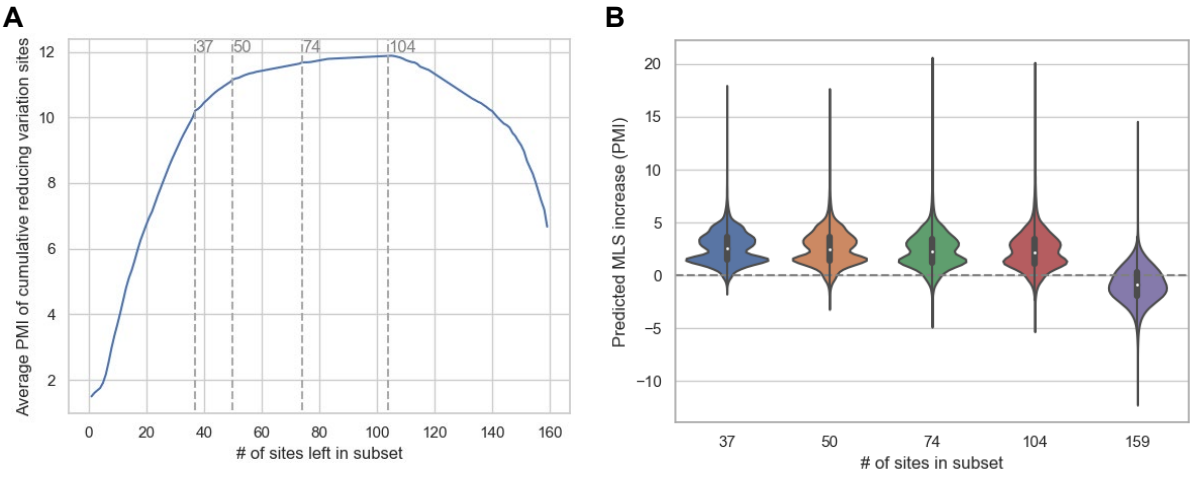
Next, 3D structures of SIRT6 sequences (spatial distances & dihedral angles between sites), as well as RoseTTAFold's 3 intermediate working matrices in PDB structure prediction, were calculated and added into the CNN model, but the performance of both models was lower than the original CNN model even with a more complex model architecture. Considering that the large amounts of inputted variables included in the above CNN models might lead to useful information being drowned out by data noise, an elastic net model was trained to see whether screening variables in the model can improve the result, but the obtained Pearson's r (r=0.667) is still slightly lower than the basic CNN model. Results of models listed above indicate that the calculated spatial distance and dihedral angle data do provide some additional information for MLS prediction, but inputting these data directly into models is inefficient in extracting useful information.

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**

Finally, esm2_t33_650M_UR50D of the ESM model, a protein language model that can extract information from protein structure as well as utilize the underlying biochemical properties of AAs, secondary and tertiary structures by simply inputting sequence and comparing it with millions of actual protein sequences, 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, ρ = 0.750, Figure S3). Since the ESM model has the best performance in MLS prediction, it will be used as the criteria for determining the effect of mutant sequences on prolonging MLS in subsequent studies.

3.2 Generating optimized human SIRT6 sequences



C

Subset	# of sequence	PMI		% of sequence for PMI		
		Mean	Max	> 0%	> 5%	> 10%
159	942080	-0.8	14.3	30.8	0.1	<0.1
104	102400	2.4	19.8	93.3	6.2	0.5
74	102400	2.4	20.1	94.3	5.5	0.3
50	102400	2.6	17.3	96.1	6.4	0.2
37	102400	2.6	17.7	98.2	5.5	0.1

Figure 4. 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. The curve first rises rapidly with the increase of site number. After forming a small platform of around 37 sites, the rising speed significantly slowed down, and finally reaches the highest platform between 104 and 74 sites, then declined with the increase of site number. (B) Violin plot of PMI for a mutant sequence of different site subsets. Negative median PMI for sequences of subset 159 suggests that deleterious mutations were still included in this subset. The median PMI for sequences of all other subsets is positive. With the decrease in subset size, the percentage of sequences with PMI > 0% increased, and the minimum PMI also increased, but the maximum PMI (novel protein with the highest efficiency) of the subset decreased. (C) PMI summary of different site subsets. The result indicates although smaller subsets do contain more important sites, sequence optimization may require the participation of multiple sites simultaneously, and relying solely on a small number of crucial sites might not generate the most optimized sequence.

As introduced in section 2.3, a 3-step approach was applied to search for optimized sequences efficiently: Firstly, 159 candidate mutation sites (which excludes conserved sites and unimportant sites) were selected from 355 sites of human SIRT6 for mutation based on entropy, MSE, and elastic net coefficient. Secondly, 942,080 sequences were generated using MPNN 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%. Finally, using 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 (Figure 4A, Table S1) 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 MPNN, and their PMI distribution was summarized.

The result (Figure 4B and 4C) shows that the median PMI of subset 159 was negative, suggesting that deleterious mutations were still included in this subset. For each smaller subset, all their median PMI became positive, and their maximum PMI, percent of PMI > 5% or PMI >10% 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 percentage of sequences with PMI > 10% decreased with a smaller subset size after subset 104, with the maximum PMI of 20.1% belonging to subset 74 instead of the expected subset 37. This suggests that although smaller subsets do contain more important sites, 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.

3.3 Distribution pattern of crucial sites in sequence optimization

3.3.1 General correlation between site distance and MLS

Figure 5 displays the heatmap of the correlation between site distance and MLS. For the functional region of sites 27-272, two key areas could be noted.

1. Site 27-65: These sites form the α1-helix, β1-sheet, and α2-helix[41]. Generally, the distance between sites in the same α/β structure is negatively correlated with MLS, but the distance in these structures, as well as with sites outside 27-65, is positively correlated with MLS, suggesting that the more compact the internal structure and more loosely spaced out of the structure, the higher the MLS.
2. Site 65-85 vs. site 135-185: These two sites are negatively correlated, suggesting the more compact between these two areas, the higher the MLS. The former corresponds to the small structural domains that form the upper part of the NAD-binding domain (α2-helix and periphery), while the latter corresponds to the zinc ion-binding module and the peripheral flexible ring. It is possible that the compact structure of this region would allow the binding

domain to bind to the corresponding molecule more rapidly, thus further affecting the catalytic efficiency of the protein.

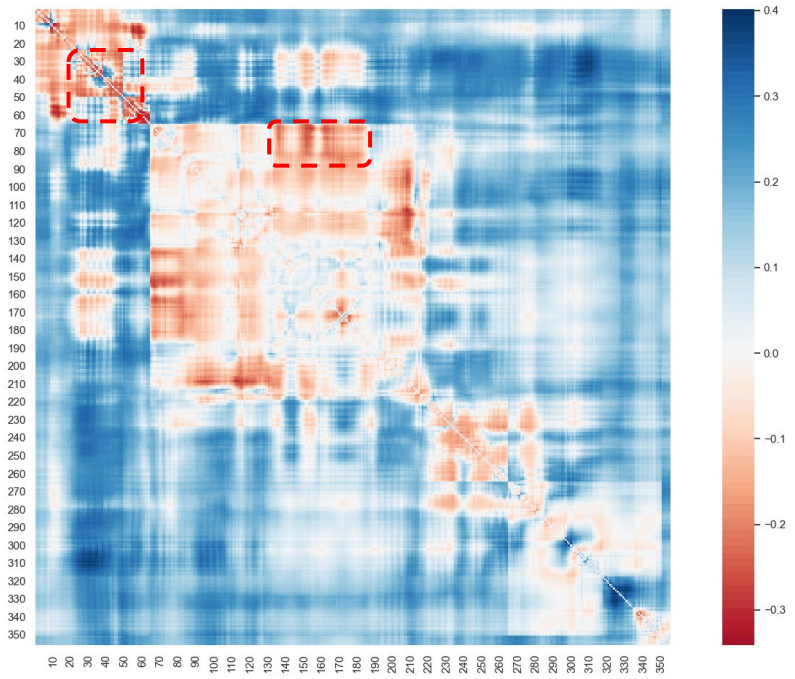


Figure 5. Correlation heatmap of distance and MLS. The relatively strong correlation between distance and MLS was observed in Site 27-65, and Site 65-85 vs. Site 135-185 (Marked with red dashed lines in the figure).

3.3.2 Spatial location of crucial sites

The location of the crucial sites in subset 37 (the most concise subset) can be divided into 4 subregions (Figure 6).

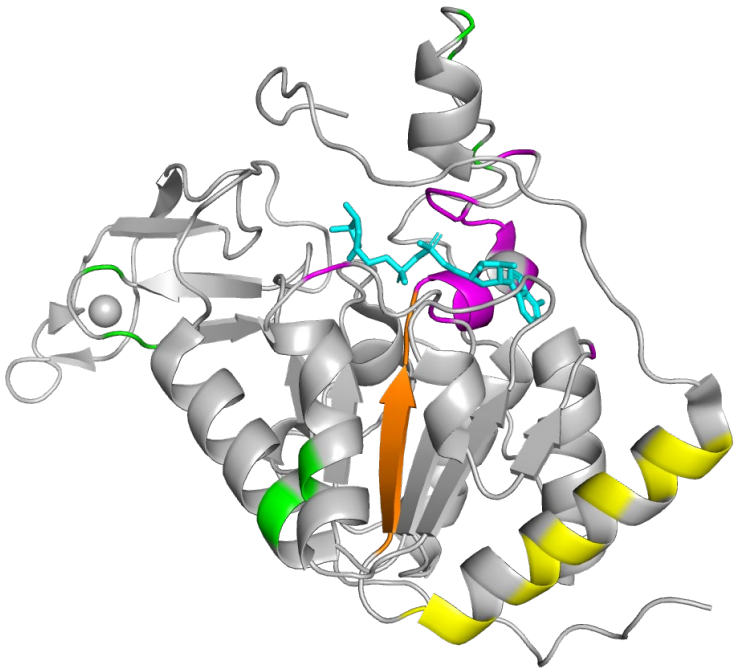
1. NAD+ binding domain sites: This binding domain involves 14 sites that are sequentially distant from each other but are located in the periphery of the binding domain in terms of 3D structure, including the minor structural region of α2-α3 (53-68) that constitutes the back of the binding domain, the 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.

2. β1-sheet: Subset 37 contains 6 consecutive sites (sites 47-52), which forms the β1-sheet in Rossmann fold internal.

3. α1-helix: A total of 9 sites were screened in the 26-44 region where the α1-helix is located.

4. Other sites: 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.

A



B

Structure	AA site range
α1-helix	26-44
β1-sheet	47-52
α2-helix	53-60
α2-α3 connection	61-68
α3-helix	69-75
α3-α4 connection	76-92
Zinc-binding area	125-190
α6-helix	193-207
β7-α7 connection	213-220
α7-helix	221-232
β9-α8 connection	254-257

Figure 6. The position of the subset 37's site in the spatial structure of SIRT6. (A) Position of the subset 37's site in SIRT6 structure (PDB ID: 5X16). Gray: Sites not belong to subset 37. Cyan: ADP-ribose. Yellow: Sites belong to α1-helix in subset 37. Magentas: Sites belong to NAD+ binding area in subset 37. Orange: Sites belong to β1-sheet in subset 37. Green: Sites belong to other area in subset 37. (B) Correspondence of AA site to spatial structure.

3.3.3 Estimated importance of each spatial subregion 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. Figure 7 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.

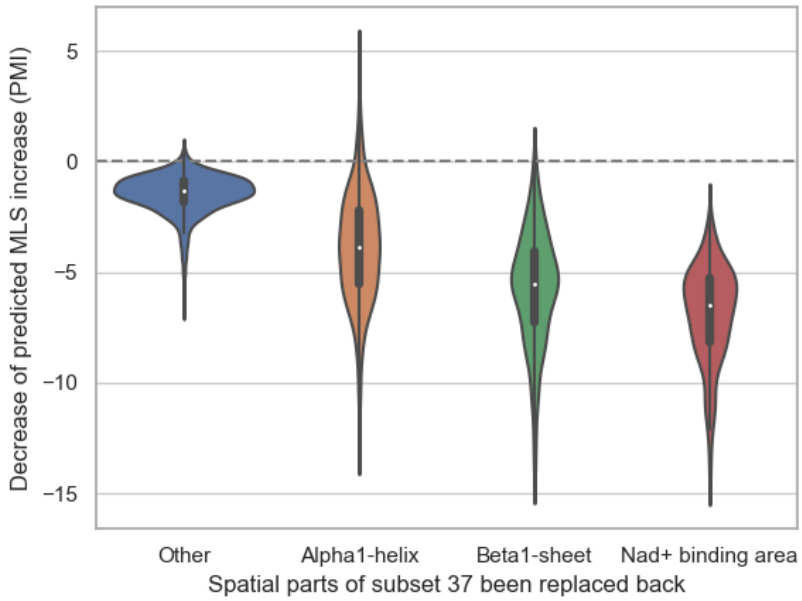


Figure 7. Violin plot of PMI decreases for replacing back of different spatial subregions of subset 37. The change in NAD+ banding area or β1-sheet led to a mean PMI decrease higher than 5%. The change in α1-helix also led to decreases in mean PMI, but the distribution range of PMI decrease significantly widened, indicating that the PMI change varied greatly among generated sequences. 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
```

Figure 8. 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. Seq-20 comes from subset 74 and is the aforementioned generated sequence with the highest PMI, while seq-16 is generated from subset 50, contains fewer changes, and is less susceptible to becoming dysfunctional due to too much change in sequence. The number of sites mutated in Seq-20 and

Seq-16 are 51 and 33, respectively (Figure 8), with 30 and 27 sites belonging to subset 37, which further suggested that those included in subset 37 are the most important sites.

The result of western blotting (Figure 9A) shows that all SIRT6 proteins (raw human SIRT6 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 (Figures 9B and 9C). 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.

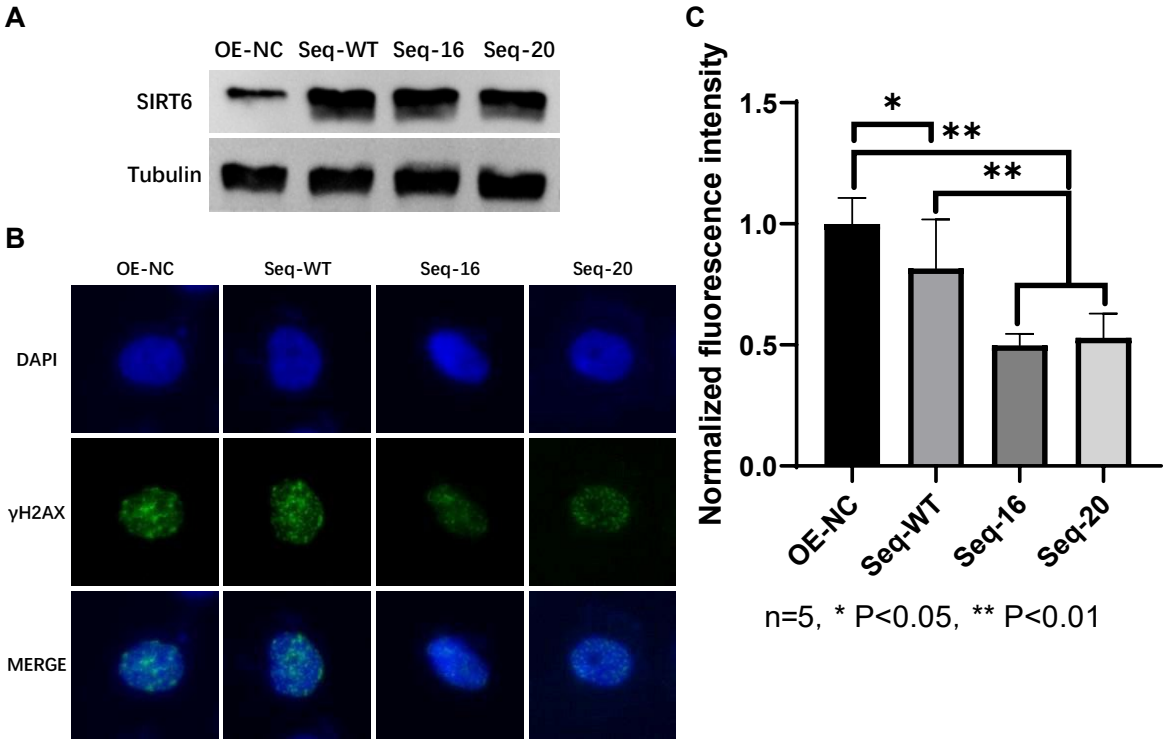


Figure 9. Comparison of DNA damage repair by γ-H2AX immunofluorescence. (A) Result of western blotting for 4 groups of HEK293. (B) Representative fluorescent micrographs of HEK293 cells at 24h after being treated with Methotrexate. (C) Comparison of normalized fluorescence intensity, where all intensities were normalized to the average OE-NC foci intensity. The normalized mean fluorescence intensity of Seq-16/20 is statistically significantly lower than OE-NC / Seq-WT, indicating higher DSB repair efficiency for the novel sequences.

4. Discussion

4.1 Extreme value of human MLS leads to challenges with prediction models

The relatively limited sample size of this study made it important to fully extract all information while eliminating bias in data. MLS data displays a right-skewed distribution (Figure S1), with most species having MLS of less than 60 years and human's 122.5 years being the largest of all mammals. Human MLS reflected not just the natural repairing abilities of the enzyme but also medical and technological advances that artificially prolonged lifespan. For instance, SIRT6 of human and gorilla sequences differs in only 2 AA out of 355 sites, but their MLS is 122.5 and 60.1 years respectively. There are also billions of human age data, which far exceeded other mammals, making extreme values of human MLS more likely to be used for this study. Thus, human SIRT6 is excluded from the train and test set to reduce bias in the prediction model. Using human SIRT6 sequence for only the generation of novel sequence and not prediction of sequence function ensured the robustness of the prediction model.

4.2 Information hidden in the sequence goes beyond distance and dihedral angle

In essence, all information used in this study, from amino acid sequence to 3D structure and amino acid properties, are derived from protein sequence. However, the relatively poor prediction performance of the CNN baseline model (Pearson's r=0.648) that only uses protein sequences indicates that using the sequence's "one-hot encoding" data is not enough to extract all useful information in sequences. Additional degrees of information, such as protein structure and biochemical characteristics of amino acids, should be considered to build more complex models. We then focused on the spatial structure of the protein and considered adding the spatial distances and dihedral angles to the model to enhance the CNN model, but surprisingly the model performance did not improve. A possible explanation may be that the total site, distance, and angle variables exceed 70,000, resulting in the inability to extract effective information through the CNN network. Transfer learning of RoseTTAFold also encountered similar problems, where flattening the working matrices may dilute the effective information. Therefore, to improve the effectiveness of the model, it is necessary to consider conducting variable screening in advance. The slightly improved performance of the ENET model, which filters variables as well as learns from them, leads to similar conclusions. Therefore, an effective combination of variable screening methods with the CNN model will be the next topic worth discussing. The protein language model ESM is capable of utilizing

sequence, structure, and underlying patterns of amino acid biochemical properties to discover how statistical trends in sequence may relate to protein structure and function. The greatly improved performance of ESM model indicates that the information contained in both the protein's spatial structure and its amino acid selection needs to be fully utilized to better predict protein performance. However, the black box nature of ESM makes it difficult for us to conduct an in-depth analysis of this issue, and it is very possible that if biochemical characteristics of amino acid on each site could be extracted as information matrix and inputted into models, the performance of previous models like CNN and ENET may be improved. Such results will also help us have a deeper understanding of the exercise between protein function and structure.

4.3 Limitations of experimental results

Although immunofluorescence using γ-H2AX indicates that 2 generated sequences did increase the DSB repair efficiency of HEK293 cells, several 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 (r^2=0.76)[15], 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[42]) 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 two of the hundred generated sequences with a large PMI were experimentally validated due to resource constraints. Although the two 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.

4.4 Importance of spatial location for crucial sites in sequence optimization

Analysis shows that sites of subset 37 are mainly located into several spatial regions. Among them, NAD+ binding area and β1-sheet play the most important role in sequence

optimization. Multiple sites are located in the NAD+ binding domain since variation of these sites directly changes the spatial structure of the binding domain[43], thus affecting the deacetylase activity of SIRT6. 6 consecutive sites in subset 37 are in β1-sheet, which constitutes the Rossmann fold internal region[44]. Although it does not have direct contact with ligands, β1-sheet is located 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 a total of 11 sites in subset 37 are located on the cofactor binding loop (β1-α2 loop)[41], 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.

Apart from the two subregions discussed above, other sites in subset 37 also have meaningful spatial locations. 9 sites are located on the α1-helix, which is at the outer edge of the Rossmann folded structure and may be related to chromatin binding efficiency (which in turn affects the activity of H3K9 and H3K56 deacetylation) due to the proximity of this helix to the N-terminus. The other sites are located around the flexible loop of the SIRT6-specific zinc ion binding domain, the periphery of the long and wide pocket region of the hydrophobic channel of the sirt6 protein, or in various portions of the exposed protein spatial structure. All these sites may play a role in affecting protein function directly or indirectly. Further in-depth analysis of how each specific site will affect protein's function will be a topic worthy of research.

4.5 Further methods of improving protein function

The maximum PMI of the mutated sequence found in this study is around 20%, and further approaches may be taken to exceed this value. First, site variations outside the functional region 27-272 can be considered. The N-terminus of SIRT6 is essential for chromatin association and intrinsic histone 3 lysine 9 (H3K9) and H3K56 deacetylation activity, whereas the C-terminus is required for the nuclear localization and recognition of nucleosomal DNA[44]. However, since the spatial structure of these sites is not fixed (pLDDT < 0.3 for most sites in prediction), they were not included in this study, and more attention needs to be paid to avoiding statistical artifacts in the analysis. Second, the sampling temperature used in MPNN can be increased. When the temperature is 0, the model takes the AA with the highest probability at the site

according to the current 3D structure, and when it is much higher than 1, the AA at the site is taken randomly. The author of MPNN points out[29] that adding noise to the backbone and using high sampling temperature can increase the diversity of sequences, therefore making the protein design task more effective. The highest sampling temperature used in our study is 1.0 (the highest value of the model by default), but we also tried different values of the parameter between 1.0 and 2.0. Results show that the higher the temperature, the larger the mean PMI and the maximum PMI reaches 27% when the temperature is 2.0. Although the risk of dysfunctionality is also higher due to the new sequence deviating from the original 3D structure at this time, it is possible to raise the temperature parameter while decreasing the number of mutation sites to prevent introducing too much disruption into the sequence function, therefore finding more optimized sequences.

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, Kassebaum NJ, Dieleman JL. Measuring population ageing: an analysis of the Global Burden of Disease Study 2017. *Lancet Public Health*. 2019;4(3):e159-e167. doi:10.1016/S2468-2667(19)30019-2
2. Hasty P, Campisi J, Hoeijmakers J, van Steeg H, Vijg J. Aging and Genome Maintenance: Lessons from the Mouse? *Science*. 2003;299(5611):1355-1359. doi:10.1126/science.1079161
3. Kowalczyk A, Partha R, Clark NL, Chikina M. Pan-mammalian analysis of molecular constraints underlying extended lifespan. *eLife*. 2020;9:e51089. doi:10.7554/eLife.51089
4. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The Hallmarks of Aging. *Cell*. 2013;153(6):1194-1217. doi:10.1016/j.cell.2013.05.039
5. MacRae SL, Croken MM, Calder RB, et al. DNA repair in species with extreme lifespan differences. *Aging*. 2015;7(12):1171-1182. doi:10.18632/aging.100866
6. 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):1443-1451. doi:10.1111/cas.14404
7. Gorbunova V, Seluanov A. DNA double strand break repair, aging and the chromatin connection. *Mutat Res Mol Mech Mutagen*. 2016;788:2-6. doi:10.1016/j.mrfmmm.2016.02.004
8. Yang JH, Hayano M, Griffin P, et al. Loss of Epigenetic Information as a Cause of Mammalian Aging. *SSRN Electron J*. Published online 2021. doi:10.2139/ssrn.3951490
9. Guo Z, Li P, Ge J, Li H. SIRT6 in Aging, Metabolism, Inflammation and Cardiovascular Diseases. *Aging Dis*. 2022;13(6):1787-1822. doi:10.14336/AD.2022.0413
10. You Y, Liang W. SIRT1 and SIRT6: The role in aging-related diseases. *Biochim Biophys Acta BBA - Mol Basis Dis*. 2023;1869(7):166815. doi:10.1016/j.bbadis.2023.166815
11. Tasselli L, Zheng W, Chua KF. SIRT6: Novel Mechanisms and Links to Aging and Disease. *Trends Endocrinol Metab*. 2017;28(3):168-185. doi:10.1016/j.tem.2016.10.002
12. Roichman A, Elhanati S, Aon MA, et al. Restoration of energy homeostasis by SIRT6 extends healthy lifespan. *Nat Commun*. 2021;12(1):3208. doi:10.1038/s41467-021-23545-7
13. TenNapel MJ, Lynch CF, Burns TL, et al. SIRT6 Minor Allele Genotype Is Associated with >5-Year Decrease in Lifespan in an Aged Cohort. Nazir A, ed. *PLoS ONE*. 2014;9(12):e115616. doi:10.1371/journal.pone.0115616
14. 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-D1033. doi:10.1093/nar/gks1155
15. 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. doi:10.1016/j.cell.2019.03.043
16. Callaway E. 'It opens up a whole new universe': Revolutionary microscopy technique sees individual atoms for first time. *Nature*. 2020;582(7811):156-157. doi:10.1038/d41586-020-01658-1

17. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2

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):871-876. doi:10.1126/science.abj8754

19. Mansoor S, Baek M, Juergens D, Watson JL, Baker D. *Accurate Mutation Effect Prediction Using RoseTTAFold*. Biochemistry; 2022. doi:10.1101/2022.11.04.515218

20. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nat Methods*. 2022;19(6):679-682. doi:10.1038/s41592-022-01488-1

21. Lin Z, Akin H, Rao R, et al. *Evolutionary-Scale Prediction of Atomic Level Protein Structure with a Language Model*. Synthetic Biology; 2022. doi:10.1101/2022.07.20.500902

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. doi:10.1073/pnas.2016239118

23. Goverde CA, Wolf B, Khakzad H, Rosset S, Correia BE. De novo protein design by inversion of the ALPHAFOLD structure prediction network. *Protein Sci*. 2023;32(6):e4653. doi:10.1002/pro.4653

24. Pearce R, Huang X, Omenn GS, Zhang Y. De novo protein fold design through sequence-independent fragment assembly simulations. *Proc Natl Acad Sci*. 2023;120(4):e2208275120. doi:10.1073/pnas.2208275120

25. Tong X, Liu X, Tan X, et al. Generative Models for De Novo Drug Design. *J Med Chem*. 2021;64(19):14011-14027. doi:10.1021/acs.jmedchem.1c00927

26. Humphreys IR, Pei J, Baek M, et al. Computed structures of core eukaryotic protein complexes. *Science*. 2021;374(6573):eabm4805. doi:10.1126/science.abm4805

27. Giaimo S. Medawar and Hamilton on the selective forces in the evolution of ageing. *Hist Philos Life Sci*. 2021;43(4):124. doi:10.1007/s40656-021-00476-6

28. Kirkwood TBL. Evolution of ageing. *Nature*. 1977;270(5635):301-304. doi:10.1038/270301a0

29. Dauparas J, Anishchenko I, Bennett N, et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*. 2022;378(6615):49-56. doi:10.1126/science.add2187

30. Lu AT, Fei Z, Haghani A, et al. Universal DNA methylation age across mammalian tissues. *Nat Aging*. Published online August 10, 2023. doi:10.1038/s43587-023-00462-6

31. Edgar RC. *High-Accuracy Alignment Ensembles Enable Unbiased Assessments of Sequence Homology and Phylogeny*. Bioinformatics; 2021. doi:10.1101/2021.06.20.449169

32. Montesinos López OA, Montesinos López A, Crossa J. Convolutional Neural Networks. In: *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer International Publishing; 2022:533-577. doi:10.1007/978-3-030-89010-0_13

33. Farahani A, Pourshojae B, Rasheed K, Arabnia HR. A Concise Review of Transfer Learning. Published online April 5, 2021. Accessed August 2, 2023. http://arxiv.org/abs/2104.02144

34. Bisong E. Google Colaboratory. In: *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Apress; 2019:59-64. doi:10.1007/978-1-4842-4470-8_7

35. Zou H, Hastie T. Regularization and Variable Selection Via the Elastic Net. *J R Stat Soc Ser B Stat Methodol*. 2005;67(2):301-320. doi:10.1111/j.1467-9868.2005.00503.x

36. Buch G, Schulz A, Schmidtmann I, Strauch K, Wild PS. A systematic review and evaluation of statistical methods for group variable selection. *Stat Med*. 2023;42(3):331-352. doi:10.1002/sim.9620

37. Sauk B, Sahinidis NV. Backward Stepwise Elimination: Approximation Guarantee, a Batched GPU Algorithm, and Empirical Investigation. *SN Comput Sci*. 2021;2(5):396. doi:10.1007/s42979-021-00788-1

38. Mao Z, Hine C, Tian X, et al. SIRT6 Promotes DNA Repair Under Stress by Activating PARP1. *Science*. 2011;332(6036):1443-1446. doi:10.1126/science.1202723

39. Madeira F, Pearce M, Tivey ARN, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res*. 2022;50(W1):W276-W279. doi:10.1093/nar/gkac240

40. Baldock RA, Day M, Wilkinson OJ, et al. ATM Localization and Heterochromatin Repair Depend on Direct Interaction of the 53BP1-BRCT 2 Domain with γH2AX. *Cell Rep*. 2015;13(10):2081-2089. doi:10.1016/j.celrep.2015.10.074

41. Pan PW, Feldman JL, Devries MK, Dong A, Edwards AM, Denu JM. Structure and Biochemical Functions of SIRT6. *J Biol Chem*. 2011;286(16):14575-14587. doi:10.1074/jbc.M111.218990

42. 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. doi:10.1038/s41586-023-06463-0

43. Sharma A, Mahur P, Muthukumaran J, Singh AK, Jain M. Shedding light on structure, function and regulation of human sirtuins: a comprehensive review. *3 Biotech*. 2023;13(1):29. doi:10.1007/s13205-022-03455-1

44. Beauharnois JM, Bolívar BE, Welch JT. Sirtuin 6: a review of biological effects and potential therapeutic properties. *Mol Biosyst*. 2013;9(7):1789. doi:10.1039/c3mb00001j

Supplementary materials

Table S1. Site list for different site subset

# of sites in subset	Site list
159	26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 68, 76, 79, 80, 81, 82, 84, 88, 90, 92, 93, 99, 100, 102, 104, 106, 108, 120, 122, 136, 138, 142, 143, 144, 145, 146, 149, 151, 152, 154, 155, 156, 157, 160, 161, 162, 163, 165, 167, 168, 169, 170, 171, 175, 180, 182, 185, 186, 191, 194, 196, 198, 199, 204, 205, 206, 207, 208, 209, 211, 212, 213, 215, 216, 217, 218, 219, 220, 223, 224, 225, 227, 228, 229, 230, 231, 232, 233, 234, 235, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276
104 (optimal subset)	28, 29, 30, 31, 32, 33, 34, 36, 37, 38, 39, 40, 41, 43, 44, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 58, 60, 61, 62, 63, 64, 65, 66, 68, 76, 79, 80, 82, 84, 90, 92, 100, 102, 120, 142, 144, 145, 146, 151, 152, 154, 156, 157, 160, 161, 162, 163, 165, 167, 168, 169, 171, 180, 185, 186, 194, 196, 198, 205, 206, 207, 208, 209, 211, 212, 213, 215, 217, 219, 220, 223, 225, 228, 229, 232, 233, 234, 235, 237, 238, 239, 240, 243, 247, 250, 251, 254, 256, 259, 260, 262, 270, 272, 273
74	28, 29, 31, 33, 34, 36, 37, 38, 39, 40, 41, 43, 44, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 58, 60, 61, 63, 64, 65, 66, 68, 76, 79, 82, 84, 90, 100, 102, 120, 142, 145, 146, 151, 152, 156, 157, 161, 167, 168, 169, 171, 180, 194, 196, 198, 205, 207, 208, 209, 211, 212, 213, 215, 217, 219, 220, 228, 232, 235, 238, 239, 256, 259, 260
50	28, 29, 31, 33, 36, 37, 38, 39, 40, 43, 44, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 58, 60, 61, 63, 64, 65, 68, 76, 79, 82, 100, 102, 142, 145, 146, 152, 167, 168, 171, 194, 196, 205, 211, 215, 217, 219, 228, 256, 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

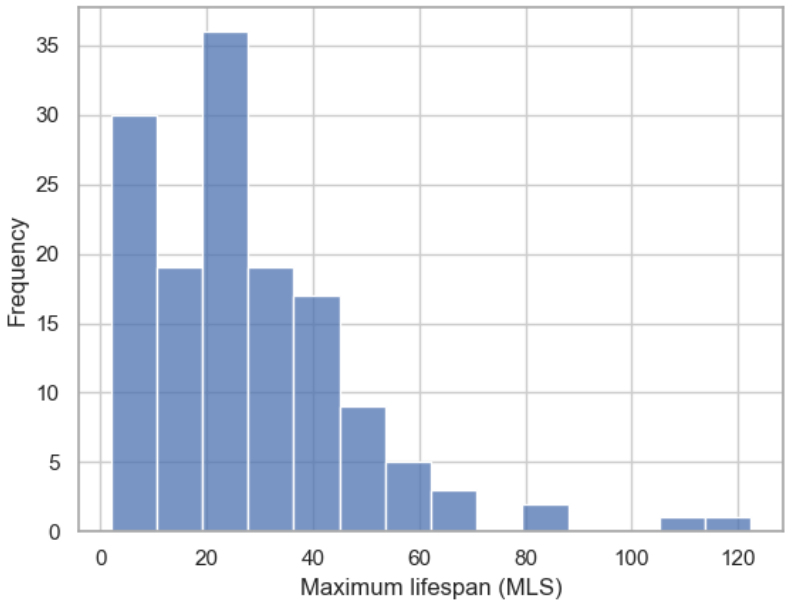


Figure S1. Histogram of maximum lifespan (MLS)

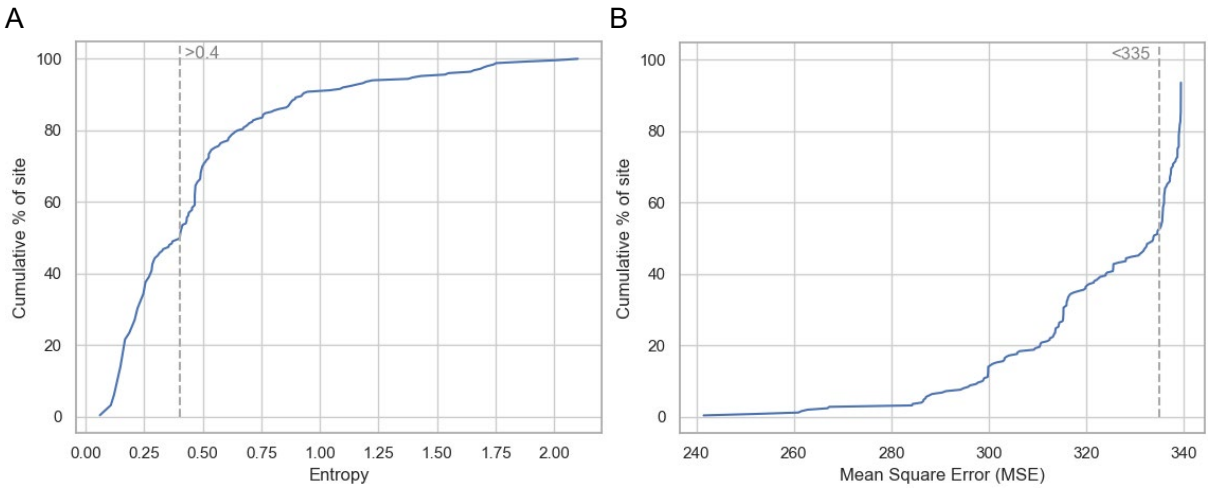


Figure S2. Line plot of Entropy and MSE for sites 26-276. (A) Line plot of Entropy. (B) Line plot of MSE.

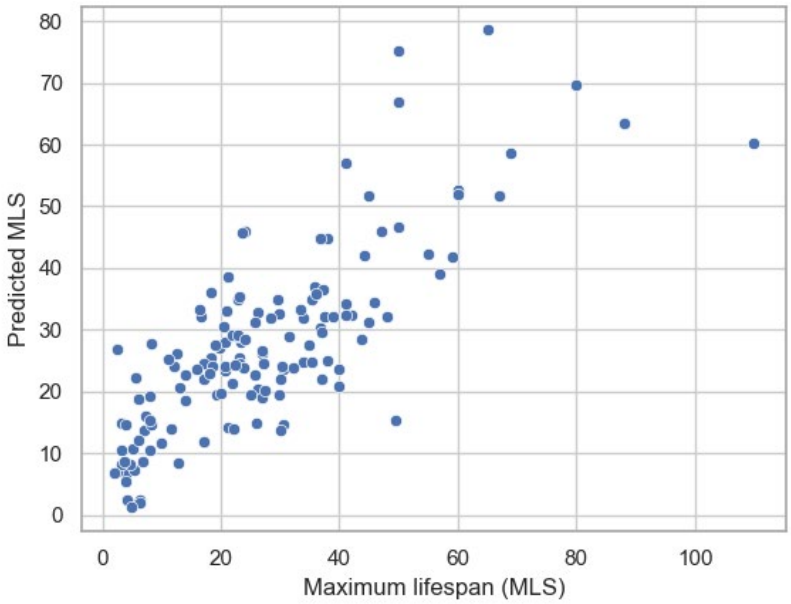


Figure S3. Scatter plot of MLS vs. predicted MLS for esm2_t33_650M_UR50D model

Acknowledgements

本研究的最初想法来自于 2021 年中 AlphaFold2 面世的新闻，在生物信息学工具飞速发展的今天，我们能否充分利用这些新的工具去实现许多曾经被认为不可能的实验和工程呢？比如通过底物结构直接设计全新的酶，又或者基于大量数据分析去优化蛋白质的结构与功能？在 2022 年参加 iGEM 项目时，我开始对衰老相关研究产生了兴趣，在对该领域有了更多了解后，最终确定了利用蛋白质的三维结构信息进行深度学习，进而对 SIRT6 蛋白序列进行优化设计这一方向。

本课题的指导教师为本校生物教师兼班主任阮振超老师，非常感谢她在课题方向选择、技术路线确定、理论指导、论文写作等多方面无偿给予的帮助。我还想感谢实验室老师在验证研究结果的湿实验过程中提供的协助和操作指导。回顾整个研究过程，我还要感谢每一次的失败与挫折，之后艰难寻找解决方案的煎熬，以及解决问题后的如释重负，这些都是我科研能力成长过程中的宝贵经历。此外，我还想感谢这个时代的开放与互联，丰富的网络资源使得我有可能自学各种分析方法和工具；诸多顶尖科学家对知识、数据、分析平台与代码的开放共享，更使得像我这样的高中生能够直接利用最先进的计算工具来开展感自己兴趣的科学研究。感谢这个时代，也希望我能不辜负这个时代。