4.1运用神经网络学习突变蛋白的局部结构功能关系

蛋白质从序列到功能的映射非常复杂，它涉及数千个分子在时间和空间尺度上的相互作用。而神经网络模型十分适合完成这项工作，因为它可以在很有限的信息假设下，从数据集中学习复杂的关系.

Sam Gelman团队提出了一个有监督的深度学习框架，训练有监督的神经网络学习序列到功能的映射函数。[Anastasiya V. Kulikova](https://link.springer.com/article/10.1007/s10867-021-09593-6" \l "auth-Anastasiya_V_-Kulikova-Aff1)团队使用了3DCNN，研究了能否良好预测野生型残基和进化上分化的同源物中的残基，并关注那些高准确率预测的氨基酸分布。

4.1.1关于蛋白质的特征工程及模型训练

在Sam Gelman团队的研究中，他们对蛋白质序列进行编码，编码信息包含了每个氨基酸在每个位置的物理和化学性质。分别采用One hot编码用于标记特定位置的氨基酸，以及AAindex编码捕获氨基酸的物理性质和化学性质，并采用PCA技术将特征向量降到19维，最后将两个编码连接。团队共测试了全连接、数列卷积、图卷积三种神经网络。卷积神经网络中，卷积层可以学习识别β链中常见的极性和非极性氨基酸，使网络能评估整个输入序列的β链倾向，并将这些信息与蛋白质功能关联。

而在[Anastasiya V. Kulikova](https://link.springer.com/article/10.1007/s10867-021-09593-6" \l "auth-Anastasiya_V_-Kulikova-Aff1)的团队中，他们在蛋白质数据集中随机取样残基，以创建一个反映每种氨基酸自然丰度的微环境数据集。对每个蛋白质链，团队最多选取50个残基，并且选取的残基数不超过总数的50%[4]，以免模型偏向那些分子量大的蛋白质。利用这些元数据，团队生成了微环境的体素化表示，具体由 3D 空间 （x，y，z） 和 7 个辅助通道组成。辅助通道编码了体素中存在的原子类别以及部分电荷和溶剂可及表面积的信息。之后通过两对3D卷积层提取特征，再进入分类块，得到隐蔽残基处20个氨基酸的概率向量。

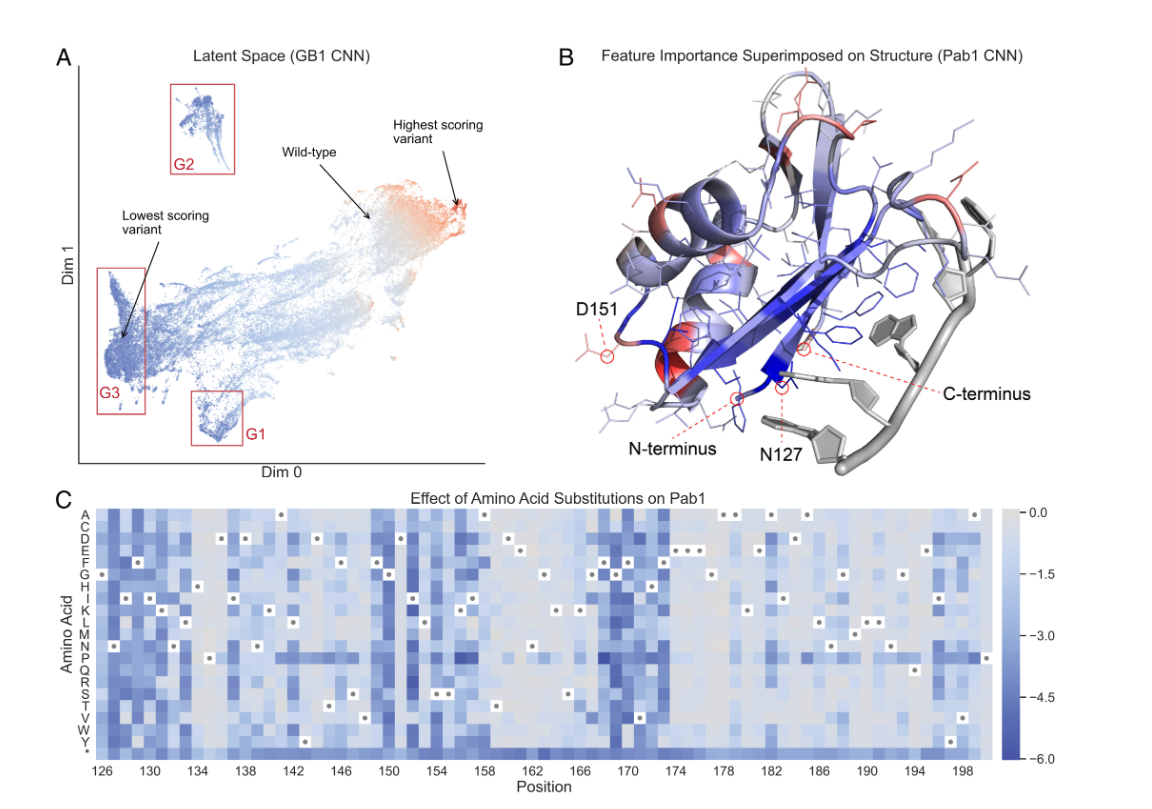
4.1.2 模型评估与应用

Sam Gelman团队运用他们的模型，结合统一流行近似和投影[5]，将网络中的连接层可视化成蛋白质的三维结构（图1[6]），并根据其得分给蛋白质着色，从而了解哪些序列位置对蛋白质的影响最大，哪些区域是突变不耐受的。

[Anastasiya V. Kulikova](https://link.springer.com/article/10.1007/s10867-021-09593-6" \l "auth-Anastasiya_V_-Kulikova-Aff1)的团队在包含130个结构的独立数据集PSICOV[4]数据集上评估他们的模型,野生型序列的预测准确率平均达到60%。将生化相似的氨基酸分组后，预测氨基酸组别的能力高于预测特定氨基酸的能力达到71%。接下来，团队询问了该网络在多序列比对 （MSA） 中预测位点的共有氨基酸的能力。该预测可以反映给定位点周围的微环境在同源结构中的保守程度，为蛋白质进化方向提供信息。

4.1.3 讨论与展望

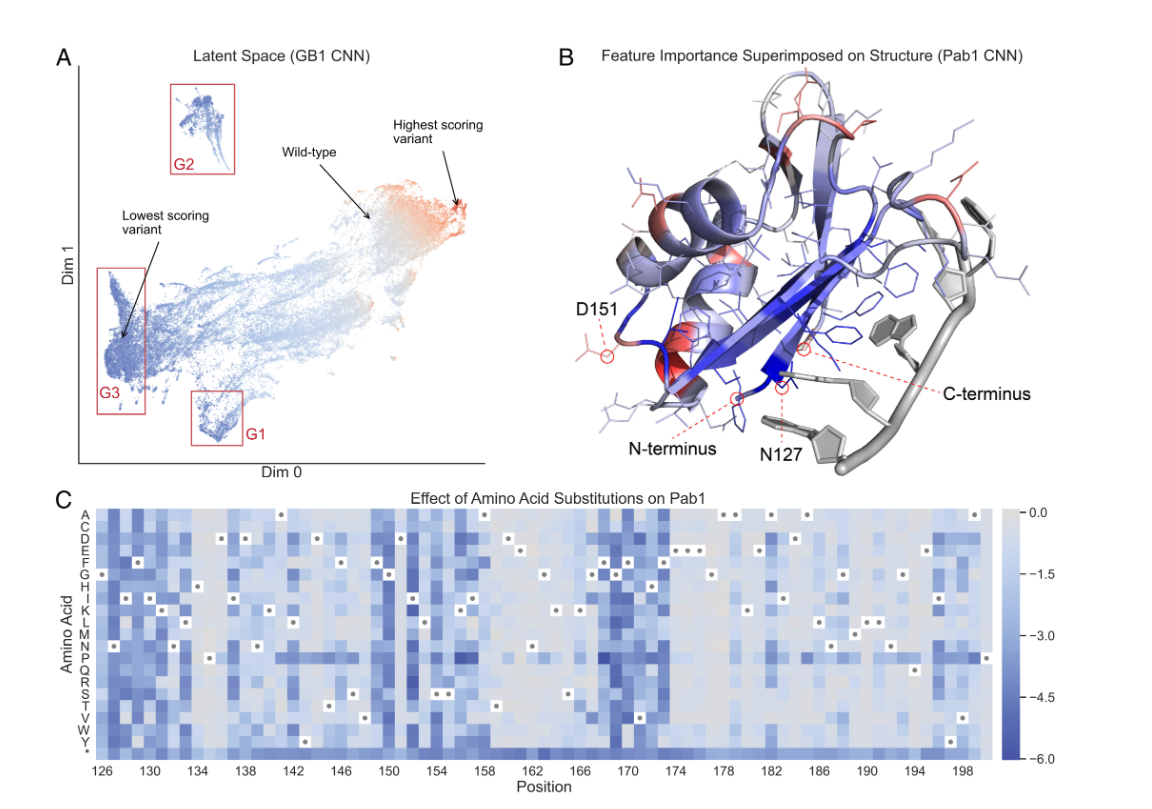
关注那些与野生型蛋白不符的氨基酸预测，这些错误预测也许为蛋白质工程提供了一种新的视角。具体而言，当模型“自信”地预测出与现有野生型不同的氨基酸时，这大概率意味着该位置具有较高的突变潜力，而模型错误预测的特定氨基酸则可能是实现结构稳定或功能增强突变的理想选择。这一发现不仅加深了我们对蛋白质进化机制的理解，还为未来的研究提供了宝贵的指导方向。



图一

参考文献

1. T. A. Hopf et al., Mutation effects predicted from sequence co-variation. Nat. Biotechnol. 35, 128–135 (2017). [Crossref](https://doi.org/10.1038/nbt.3769). [PubMed](https://pubmed.ncbi.nlm.nih.gov/28092658/).
2. A. J. Riesselman, J. B. Ingraham, D. S. Marks, Deep generative models of genetic variation capture the effects of mutations. Nat. Methods 15, 816–822 (2018). [Crossref](https://doi.org/10.1038/s41592-018-0138-4). [PubMed](https://pubmed.ncbi.nlm.nih.gov/30250057/).
3. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D.: Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589
4. Jones, D.T., Buchan, D.W.A., Cozzetto, D., Pontil, M.: PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28, 184–190 (2011)
5. L. McInnes, J. Healy, UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv [Preprint] (2020). <https://arxiv.org/abs/1802.03426> (Accessed 18 September 2020).
6. Sam Gelman, Sarah A. Fahlberg, Pete Heinzelman, Philip A. Romero & Anthony Gitter. [Proceedings of the National Academy of Sciences](https://doi.org/10.1073/pnas.2104878118) • [code](https://github.com/gitter-lab/nn4dms) • Nov 2021
7. **Learning the local landscape of protein structures with convolutional neural networks.** Anastasiya V. Kulikova, Daniel J. Diaz, James M. Loy, Andrew D. Ellington & Claus O. Wilke. [Journal of Biological Physics 47.4 (2021)](https://link.springer.com/article/10.1007/s10867-021-09593-6)

4.1 Using Neural Networks to Learn the Local Structure-Function Relationships of Mutant Proteins  
  
The mapping of proteins from sequence to function is very complex, involving the interactions of thousands of molecules on temporal and spatial scales. Neural network models are particularly well-suited for this task because they can learn complex relationships from datasets under very limited information assumptions.  
  
The Sam Gelman team proposed a supervised deep learning framework to train supervised neural networks to learn the mapping function from sequence to function.Anastasiya V. Kulikova's team used 3DCNN to investigate whether they could accurately predict residues in wild-type and evolutionarily diverged homologs, focusing on the distribution of amino acids with high prediction accuracy.  
  
4.1.1 Feature Engineering and Model Training for Proteins  
  
In Sam Gelman's team's research, they encode protein sequences, with the encoded information containing the physical and chemical properties of each amino acid at each position.One hot encoding was used to mark the amino acids at specific positions, and AAindex encoding was used to capture the physical and chemical properties of the amino acids. PCA technology was then employed to reduce the feature vectors to 19 dimensions, and finally, the two encodings were concatenated.The team tested three types of neural networks: fully connected, sequential convolution, and graph convolution.In convolutional neural networks, the convolutional layers can learn to identify common polar and non-polar amino acids in the β-chain, enabling the network to assess the β-chain tendency of the entire input sequence and associate this information with protein function.  
  
In Anastasiya V. Kulikova's team, they randomly sampled residues from the protein dataset to create a microenvironment dataset that reflects the natural abundance of each amino acid.For each protein chain, the team selected up to 50 residues, and the number of selected residues did not exceed 50% of the total number of residues[4], to avoid biasing the model towards larger molecular weight proteins.Using this metadata, the team generated a voxelized representation of the microenvironment, specifically consisting of 3D space (x, y, z) and 7 auxiliary channels.The auxiliary channels encode the atomic types present in the voxel, as well as information on partial charges and solvent-accessible surface area.Afterwards, features are extracted through two pairs of 3D convolutional layers, then they enter the classification block to obtain the probability vector of 20 amino acids at the concealed residues.  
  
4.1.2 Model Evaluation and Application  
  
Sam Gelman's team used their model, combining uniform popular approximation and projection [5], to visualize the connection layers in the network as the three-dimensional structure of proteins (Figure 1 [6]), and colored the proteins based on their scores to understand which sequence positions have the greatest impact on the proteins and which areas are mutation intolerant.  
  
Anastasiya V. Kulikova's team evaluated their model on the independent dataset PSICOV[4], which contains 130 structures, achieving an average prediction accuracy of 60% for wild-type sequences.After grouping biochemically similar amino acids, the ability to predict amino acid groups surpassed the ability to predict specific amino acids, reaching 71%.Next, the team inquired about the network's ability to predict the shared amino acids at the sites in multiple sequence alignments (MSA).This prediction can reflect the degree of conservation of the microenvironment around the given site in homologous structures, providing information on the direction of protein evolution.  
  
4.1.3 Discussion and Outlook  
  
Focus on the predictions of amino acids that differ from the wild-type protein; these erroneous predictions may offer a new perspective for protein engineering. Specifically, when the model confidently predicts amino acids that differ from the existing wild-type, it likely indicates that the position has a high mutation potential, and the specific amino acids incorrectly predicted by the model could be ideal candidates for achieving structural stability or functional enhancement mutations. This finding not only deepens our understanding of protein evolution mechanisms but also provides valuable guidance for future research.  
  
 Figure 1  
  
References

1. T. A. Hopf et al., Mutation effects predicted from sequence co-variation. Nat. Biotechnol. 35, 128–135 (2017). [Crossref](https://doi.org/10.1038/nbt.3769). [PubMed](https://pubmed.ncbi.nlm.nih.gov/28092658/).
2. A. J. Riesselman, J. B. Ingraham, D. S. Marks, Deep generative models of genetic variation capture the effects of mutations. Nat. Methods 15, 816–822 (2018). [Crossref](https://doi.org/10.1038/s41592-018-0138-4). [PubMed](https://pubmed.ncbi.nlm.nih.gov/30250057/).
3. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D.: Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589
4. Jones, D.T., Buchan, D.W.A., Cozzetto, D., Pontil, M.: PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28, 184–190 (2011)
5. L. McInnes, J. Healy, UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv [Preprint] (2020). <https://arxiv.org/abs/1802.03426> (Accessed 18 September 2020).
6. Sam Gelman, Sarah A. Fahlberg, Pete Heinzelman, Philip A. Romero & Anthony Gitter. [Proceedings of the National Academy of Sciences](https://doi.org/10.1073/pnas.2104878118) • [code](https://github.com/gitter-lab/nn4dms) • Nov 2021
7. **Learning the local landscape of protein structures with convolutional neural networks.** Anastasiya V. Kulikova, Daniel J. Diaz, James M. Loy, Andrew D. Ellington & Claus O. Wilke. [Journal of Biological Physics 47.4 (2021)](https://link.springer.com/article/10.1007/s10867-021-09593-6)