Introduction

Synthetic biology is a field of biology which consist of creating new biological systems like cells, proteins or organism. These new biological systems are inspired by natural systems and then using engineering these natural systems will be modified to present new functions or improved functions.[1-2] Synthetic biology is not only useful in medicine[3] but also in industry[4] or agriculture[5]. Protein engineering is the field of synthetic biology which consist of creating new recombinant protein[6] with improved properties such as better stability, thermostability or catalytic activity and so forth.[7] The creation of new variant is challenging du to the large number of possibilities and the difficulties to accurately estimate which part of the protein is important in the function that we want to modify. Directed evolution[8] is a technic of protein engineering that consist of mutating a small number of amino acids and then observe the properties modification. It can be useful to identify some important amino acids but the protein that it can create are really close to the original protein. When the structure of a protein is well define we can also use computational rational design[9], which consist in computing all the available information for a protein to identify interesting mutation.[10] But these two technics have shown limited success and cannot be used for all proteins.

Machine Learning[11] is a field of Data Science that can be used to exploit big database by letting the computer learn the distribution underlying the data. For protein database we almost always don’t know the importance or effect of each amino acids modification which render supervised learning methods[12] difficult to use since they need data with label to perform. But some Machine Learning’s model can perform without needing labelled data which is the unsupervised Learning. Among unsupervised Machine Learning methods, generative model can be used to generate new sample similar to the data used during the training phase. It has been used in recent year to create new music[13], images[14] or even small molecules.[15] Generative model can then be used to create new protein variants that will kept important function but that will have modified properties by identifying in the protein sequence the amino acids that are well conserved in our different protein sequence and the amino acids that can be modify. If information about the protein’s properties are available they can also be used during our model training to create protein with desired properties.[16]

Most of generative model like Generative adversarial network[17] or Autoencoder(AE)[18] need input of the same size to learn. So we need to use aligned protein’s sequence to train them. But multiple sequence alignment precision decrease when the number of protein to align and the size of the protein increase. That can be an issue because we want to use generative model for every situation despite the size of the protein and we want to use as many protein as we can to improve the model training. It exist variation of GAN and AE that can accept different input size but we will rather use Transformer[19] to test unaligned protein sequence.

In recent years generative models were used to work on different protein engineering problem like generate new protein variant[20–24] or protein folding[25]. As students our main objective during this project was to discover generative model. We decided to use them to create protein to have a concrete example to work on and that is link to our studies field bioinformatics. Generative models can be used in other field of biology so it was useful knowledge for our future.[26–28] We decided to work on creating variants for the protein Tau which is responsible for Alzheimer disease.[29] Our objective was to create variant that will kept the interaction with the tubulin which is the main function of protein Tau. Our new variants would possibly keep all protein Tau’s function while loosing the ability to cause Alzheimer disease.

State of art

**Generative models**

Generative models are model able to extract characteristics from a dataset and then create new data presenting these characteristics and some variations. These new data are then similar to our original dataset but present some variations so they are new data.

We can find these kind of generative model on this [website](https://thispersondoesnotexist.com/)  where we create a face of a person that does not exist, that face does not come from a database containing photographs of faces taken from an actual person. At each connexion we can create a different face.

We used this kind of model to create new proteins that will keep the principal functions of a family of proteins but presenting new functions or less secondary effects.

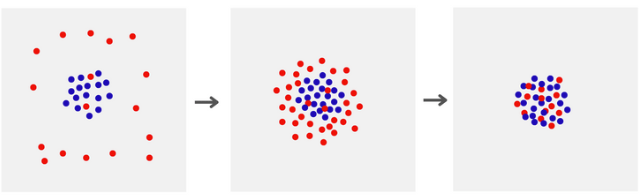
We have identified 2 models able to generate new data the Generative Adversarial Network (GAN) and the Variational Autoencoder (VAE).

**GANs : generative adversarial network**

GANs[17], [30] model are in reality constituted of two neural network. The first one called generator, is responsible for creating the new data. The second called discriminator is responsible for identifying if a data is real or invented.

When we train a GANs our purpose is that the generator is able to create data so close to real data that our discriminator identified them as real data.

To train a GANs model we first train the discriminator by providing it the real data so it learn to recognize real data. Then we ask the generator to create random data and we ask the discriminator to classify it as real or invented data. Then the discriminator give a feedback to the generator. With this feedback the generator try to create new data that are better than the first batch of invented data. We will repeat this operation until the invented data are good enough and are able to fool the discriminator. When the discriminator is not able to classify invented data as invented data then it mean that our data are closed enough from reality. **(Fig 1)**



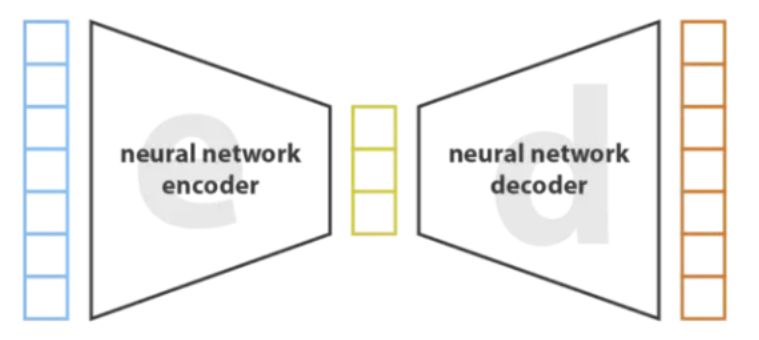
***Fig 5 :*** *visualization of invented data during GANs's training*

*In blue the real data and in red the invented data. Left is the beginning of GAN's training and right the end of the training.*

**VAE : Variational Autoencoders**

VAE are a derived from Autoencoders (AE)[18], a neural network that can be used in other fields than data creation. AE model are also constituted of two neural network the encoder and the decoder.

Encoder mission is to reduce the number of features used to describe the data without losing information. We passed from a data's description with X feature to a description with Y features with X > Y. Decoder's mission is the reverse we passed from a data's description with Y features to a description with X features with Y < X.

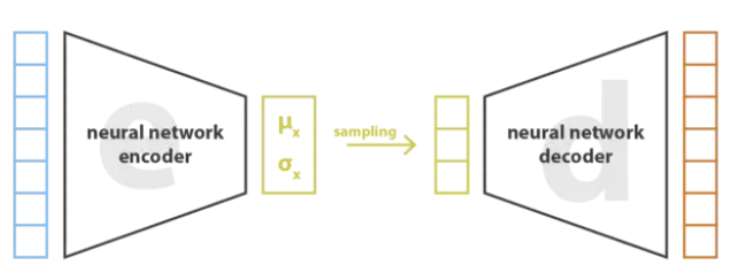


***Fig 6 :*** *Autoencoder architecture*

When we train an AE model the goal is to have for input a data with X features (***Fig 6*** : in blue) then reduce the feature's number to Y features (***Fig 6 :*** *in yellow*) using the encoder. Then with the decoder we try to recreate the exact same data with X features. The objective is to find a representation with Y features without losing any information compared to the X feature representation.

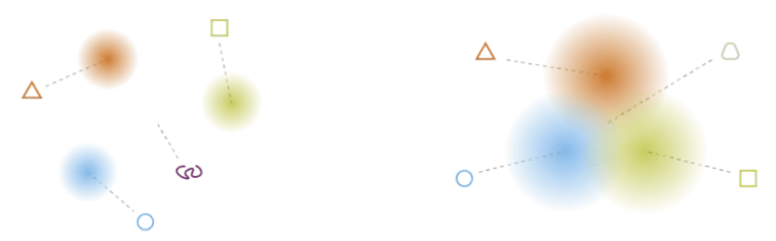
To create new data using Autoencoder we will first train the Autoencoder using the dataset of real data that we have. When the training is over and our AE is able to reduce the feature's number (encoder) and then expand the feature's number to recreate the original data (decoder), we keep only the decoder part. Then we give to the decoder random data with Y feature and then the decoder will transform it to a data in X feature similar to the real data.

But the Autoencoder have an issue when creating data which is that we can create aberrant data that are not similar to the real data. These aberrant data are due to the random number that we provide to the decoder. If these number are too different from the real data the decoder will not be able to transform it in a data similar to the real data. That's why Variational Autoencoders have been created. VAE model add a regularization step between the encoder and the decoder. (***Fig 7***)



***Fig 7 :*** *Variational Autoencoder architecture*

This normalization allow us to greatly reduce the possibility of aberrant data because even if we give to the decoder random number that are not similar to the real data the regularization step will transform these number to number closer to the real data and then limit the creation of aberrant data. (***Fig 8***)



***Fig 8*** *: Examples of data without regularization (left) and with regularization(right) If the created data is in the red, blue or yellow color a good data is created il it's in the white an aberrant data is created. We can see that with the regularization the probability of aberrant data is reduced.*

References :

[1] L. Serrano, “Synthetic biology: Promises and challenges,” *Molecular Systems Biology*, vol. 3. p. 158, 2007. doi: 10.1038/msb4100202.

[2] P. Shapira, S. Kwon, and J. Youtie, “Tracking the emergence of synthetic biology,” *Scientometrics*, vol. 112, no. 3, pp. 1439–1469, Sep. 2017, doi: 10.1007/s11192-017-2452-5.

[3] Z. Kis, H. S. A. Pereira, T. Homma, R. M. Pedrigi, and R. Krams, “Mammalian synthetic biology: Emerging medical applications,” *Journal of the Royal Society Interface*, vol. 12, no. 106. Royal Society of London, May 06, 2015. doi: 10.1098/rsif.2014.1000.

[4] S. J. King, A. Jerkovic, L. J. Brown, K. Petroll, and R. D. Willows, “Synthetic biology for improved hydrogen production in Chlamydomonas reinhardtii,” *Microbial Biotechnology*, vol. 15, no. 7. John Wiley and Sons Ltd, pp. 1946–1965, Jul. 01, 2022. doi: 10.1111/1751-7915.14024.

[5] J. Ke, B. Wang, and Y. Yoshikuni, “Microbiome Engineering: Synthetic Biology of Plant-Associated Microbiomes in Sustainable Agriculture,” *Trends in Biotechnology*, vol. 39, no. 3. Elsevier Ltd, pp. 244–261, Mar. 01, 2021. doi: 10.1016/j.tibtech.2020.07.008.

[6] L. A. Palomares, S. Estrada-Mondaca, and O. T. Ramírez, “Production of Recombinant Proteins Challenges and Solutions.” [Online]. Available: www.phrma.org

[7] C. You, R. Huang, X. Wei, Z. Zhu, and Y. H. P. Zhang, “Protein engineering of oxidoreductases utilizing nicotinamide-based coenzymes, with applications in synthetic biology,” *Synthetic and Systems Biotechnology*, vol. 2, no. 3. KeAi Communications Co., pp. 208–218, Sep. 01, 2017. doi: 10.1016/j.synbio.2017.09.002.

[8] J. D. Bloom and F. H. Arnold, “In the light of directed evolution: Pathways of adaptive protein evolution.” [Online]. Available: www.nasonline.org/SacklerDarwin.

[9] M. K. Tiwari, R. Singh, R. K. Singh, I. W. Kim, and J. K. Lee, “Computational approaches for rational design of proteins with novel functionalities,” *Comput Struct Biotechnol J*, vol. 2, no. 3, p. e201204002, 2012, doi: 10.5936/csbj.201209002.

[10] A. Nandy and S. C. Basak, “A brief review of computer-assisted approaches to rational design of peptide vaccines,” *International Journal of Molecular Sciences*, vol. 17, no. 5. MDPI AG, May 04, 2016. doi: 10.3390/ijms17050666.

[11] B. Mahesh, “Machine Learning Algorithms-A Review Machine Learning Algorithms-A Review View project Six Stroke Engine View project Batta Mahesh Independent Researcher Machine Learning Algorithms-A Review,” *International Journal of Science and Research*, 2018, doi: 10.21275/ART20203995.

[12] V. Nasteski, “An overview of the supervised machine learning methods,” *HORIZONS.B*, vol. 4, pp. 51–62, Dec. 2017, doi: 10.20544/horizons.b.04.1.17.p05.

[13] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, “Jukebox: A Generative Model for Music,” Apr. 2020, [Online]. Available: http://arxiv.org/abs/2005.00341

[14] A. Grigorev *et al.*, “StylePeople: A Generative Model of Fullbody Human Avatars.”

[15] N. de Cao and T. Kipf, “MolGAN: An implicit generative model for small molecular graphs,” May 2018, [Online]. Available: http://arxiv.org/abs/1805.11973

[16] Ł. Maziarka, A. Pocha, J. Kaczmarczyk, K. Rataj, T. Danel, and M. Warchoł, “Mol-CycleGAN: A generative model for molecular optimization,” *J Cheminform*, vol. 12, no. 1, Jan. 2020, doi: 10.1186/s13321-019-0404-1.

[17] I. Goodfellow *et al.*, “Generative adversarial networks,” *Commun ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.

[18] D. Bank, N. Koenigstein, and R. Giryes, “Autoencoders,” Mar. 2020, [Online]. Available: http://arxiv.org/abs/2003.05991

[19] A. Vaswani *et al.*, “Attention Is All You Need,” Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.03762

[20] A. Strokach and P. M. Kim, “Deep generative modeling for protein design,” *Current Opinion in Structural Biology*, vol. 72. Elsevier Ltd, pp. 226–236, Feb. 01, 2022. doi: 10.1016/j.sbi.2021.11.008.

[21] N. Anand and P.-S. Huang, “Generative Modeling for Protein Structures.”

[22] P. Tian *et al.*, “Design of a Protein with Improved Thermal Stability by an Evolution‐Based Generative Model,” *Angewandte Chemie International Edition*, Dec. 2022, doi: 10.1002/anie.202202711.

[23] A. Hawkins-Hooker, F. Depardieu, S. Baur, G. Couairon, A. Chen, and D. Bikard, “Generating functional protein variants with variational autoencoders,” *PLoS Comput Biol*, vol. 17, no. 2, Feb. 2021, doi: 10.1371/JOURNAL.PCBI.1008736.

[24] Z. Wu, K. E. Johnston, F. H. Arnold, and K. K. Yang, “Protein sequence design with deep generative models,” *Current Opinion in Chemical Biology*, vol. 65. Elsevier Ltd, pp. 18–27, Dec. 01, 2021. doi: 10.1016/j.cbpa.2021.04.004.

[25] H. Ma *et al.*, “Deep Generative Model Driven Protein Folding Simulation,” Aug. 2019, [Online]. Available: http://arxiv.org/abs/1908.00496

[26] L. Seninge, I. Anastopoulos, H. Ding, and J. Stuart, “VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics,” *Nat Commun*, vol. 12, no. 1, Dec. 2021, doi: 10.1038/s41467-021-26017-0.

[27] X. Zeng *et al.*, “Deep generative molecular design reshapes drug discovery,” *Cell Rep Med*, p. 100794, Dec. 2022, doi: 10.1016/j.xcrm.2022.100794.

[28] R. Lopez, A. Gayoso, and N. Yosef, “Enhancing scientific discoveries in molecular biology with deep generative models,” *Mol Syst Biol*, vol. 16, no. 9, Sep. 2020, doi: 10.15252/msb.20199198.

[29] M. Kolarova, F. García-Sierra, A. Bartos, J. Ricny, and D. Ripova, “Structure and pathology of tau protein in Alzheimer disease,” *International Journal of Alzheimer’s Disease*. 2012. doi: 10.1155/2012/731526.

[30] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F. Y. Wang, “Generative adversarial networks: Introduction and outlook,” *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4. Institute of Electrical and Electronics Engineers Inc., pp. 588–598, Oct. 01, 2017. doi: 10.1109/JAS.2017.7510583.