Title of the PhD project:

Alternative splicing-inspired protein design

PhD Supervisor

Name, first name : Laine Elodie
Email : elodie.laine@sorbonne-universite.fr
Phone : 06 71 22 69 41
Title/Employer : Maître de Conférences / Sorbonne Université
Address : 4, place Jussieu, 75005 PARIS
Laboratory : Laboratoire de Biologie Computationnelle et Quantitative (LCQB), UMR 7238, IBPS, CNRS-Sorbonne Université
Title of the team : Equipe projet MASSIV dans l'équipe de Génomique Analytique
Team leader (*if different*) *:* Alessandra Carbone
Doctoral School : EDITE (HDR soutenue le 2 octobre 2020)

Overview of the scientific projects of the team

L'équipe-projet MASSIV a été créée début 2018, avec un financement ANR JCJC. Elle se compose de 2 enseignants-chercheurs, E. Laine et H. Richard. Un post-doctorant a travaillé sur le projet sur la période 02/2018-09/2020. L'équipe accueille régulièrement des stagiaires de niveau L à M2 (6 depuis le début). Nos activités concerne l'évolution et l'impact structural de l'épissage alternatif (développement de méthodes et application à large échelle).

Main publications since January 1er, 2016

Ait-hamlat A., DJ. Zea, A. Labeeuw, L. Polit, H. Richard* and **E. Laine***. (2020) Transcripts' evolutionary history and structural dynamics give mechanistic insights into the functional diversity of the JNK family. *J Mol Biol.* **432**:2121-2140

Corsi F., R. Lavery, **E. Laine*** and A Carbone*. (2019) Multiple protein-DNA interfaces unravelled by evolutionary information, physico-chemical and geometrical properties. *accepted in PLoS Comput Biol* **doi:** https://doi.org/10.1101/743617

Laine E.*, Y. Karami and A. Carbone*. (2019) GEMME: a simple and fast global epistatic model predicting mutational effects. *Mol Biol Evol.* 36:2604–2619

Dequeker C., **E. Laine*** and A. Carbone*. (2019) Decrypting protein surfaces by combining evolution, geometry, and molecular docking. *Proteins*. **87**:952-965

Karami Y., T. Bitard-Feildel, **E. Laine*** and A. Carbone*. (2018) "Infostery" analysis of short molecular dynamics simulations identifies highly sensitive residues and predicts deleterious mutations*, Scientific Reports.* **8** :16126

+ equal contributions *co-corresponding authors, and students underlined

PhD Co-Supervisor

Name, first name: Grudinin Sergei
Email : sergei.grudinin@inria.fr
Phone : 04 38 78 16 91
Title/Employer: CNRS
Address : Antenne Inria Giant, Minatec Campus, 17 rue des Martyrs, 38054 Grenoble Cedex
Laboratory: Laboratoire Jean Kuntzmann (LJK), UMR 5224, Inria Grenoble Rhône-Alpes - CNRS
Title of the team : Nano-D
Team leader *(if different):*
Doctoral School: MSTII Grenoble

Overview of the scientific projects of the team

Les activités de l'équipe Nano-D concernent la manipulation et la modélisation des structures protéiques. Nous avons développé des méthodes pour caractériser et prédire les assemblages macromoléculaires (interactions protéine-protéine/ligand), évaluer la qualité des modèles 3D des protéines et leur complexes, et prédire la flexibilité et les transitions fonctionnelles des protéines. Ces méthodes sont très rapides et peuvent être appliquées à très large échelle. L'équipe se distingue régulièrement dans des compétitions internationales du domaine (CASP, CAPRI…). Nous collaborons avec P. Chacon (Université de Madrid) pour la flexibilité et les petites molécules (drug design), C. Venclovas (Vilnius University) pour la prédictions de structure et interactions, E. Laine (SU) pour les aspects évolutifs, et J Mairal (Inria Grenoble) et Y. Bengio (MILA, Canada) pour les méthodes d'apprentissage profond.

Main publications since January 1er, 2016

Kadukova M, Machado KDS, Chacón P, **Grudinin** S. (2020) KORP-PL: a coarse-grained knowledge-based scoring function for protein-ligand interactions. *Bioinformatics.* (*in press*)
Karasikov M, Pagès G, Grudinin S. (2019) Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. *Bioinformatics*. 35:2801-2808.
Pagès G, **Grudinin** S. (2019) DeepSymmetry: using 3D convolutional networks for identification of tandem repeats and internal symmetries in protein structures. *Bioinformatics*. **35**:5113-5120.
Pagès G, Charmettant B, **Grudinin** S. (2019) Protein model quality assessment using 3D oriented convolutional neural networks. *Bioinformatics*. **35**:3313-3319.
Derevyanko G, **Grudinin** S, Bengio Y, Lamoureux G. (2018) Deep convolutional networks for quality assessment of protein folds. Bioinformatics. **34**:4046-4053

Doctoral Project

Title: Alternative splicing-inspired protein design

Abstract:
By generating multiple transcripts from the same gene, alternative splicing has the potential to greatly expand eukaryotic proteomes. In this doctoral project, we propose to leverage the growing body of available transcriptomics and proteomics data to generate new protein functional diversity. We will use the notion of evolutionary conservation to identify a set of alternative-splicing-induced sequence variations likely relevant to protein function. We will carefully cross this information with transcript expression and proteomics data. We will then map the identified sequence variations onto protein structures and interactions, at the level of protein domain families. We will build a probabilistic model that will learn « rules » from this curated resource to determine where and how to target a protein in order to modulate its function. We will represent the input data as graphs and will develop a suitable deep learning architecture (e.g., variational auto encoder). We will produce a knowledge base for alternative splicing and new methods for graph learning applied to proteins. The expected outcome will improve our understanding of protein functioning and help to guide protein design.

Context and objective:

Eukaryotes have evolved a transcriptional mechanism that can augment the protein repertoire without increasing genome size. A gene can be transcribed, spliced, and matured into several transcripts by choosing different initiation/termination sites or by selecting different exons [1]. Alternative splicing (AS) concerns almost all multi-exon genes [2], and it can produce protein isoforms with different shapes [3], interactions partners [4], and functions [5-6]. Hence, the generative potential of AS is fascinating.

In recent years, in-depth surveys of the splicing complexity across species and tissues have been made possible by high-throughput sequencing (HTS) technologies like RNA-Seq. However, reconstructing full-length transcripts from short reads is difficult, and evaluating how many of the detected transcripts are translated and functional in the cell remains challenging [16-17]. As a consequence, there is a long-standing debate in the field about the functional impact of AS [18-20]. This has stimulated the development of « long read » sequencing technologies [13], and of integrative approaches combining gene annotations, RNA-Seq data and also data generated by other high-throughput techniques (*e.g.,* mass spectrometry) [14-15,21-23]. These efforts have contributed to better characterise AS landscape complexity and assess its phenotypic outcome.

Evolutionary conservation is a widely recognised indicator of function, and we expect that the AS-induced variations selected over millions of years of evolution comply with physical and environmental constraints and thus are likely functional. Over the past years, the team of EL has developed a couple of efficient methods to assess the evolutionary conservation of AS events and transcripts and to model the impact of AS on protein 3D structures (Zea *et al. submitted,* Ait-Hamlat *et al. J Mol Biol* 2020). These tools integrate gene annotations (from

Ensembl[1]) and RNA-Seq splice junctions (from Bgee[2]). As a proof-of-concept, we used them to date known functional AS events in the c-Jun N-terminal kinase family and identify residues responsible for AS phenotypic outcome (Ait-Hamlat *et al. J Mol Biol* 2020). We further showed a clear link between the functional relevance, tissue-regulation and conservation of AS events on a set of 50 genes (Zea *et al. submitted).* We scaled up the analysis to the whole human protein-coding genome, leading to the identification of a few thousands of conserved AS events.

Artificial intelligence (AI), and more specifically deep learning (DL), has recently emerged as a powerful approach to exploit the massive amount of protein sequence and structure data available nowadays to predict mutational outcomes [24], fold proteins in 3D [25], design mutants with the desired properties [26], predict protein binding modes [27] and classify AS events [28], among others. While most of the previously designed DL architectures use 1D or 2D as input, there is a growing interest for defining suitable data representations and convolution operations for 3D objects. The team of SG has been developing pioneer and innovative methods and architectures for DL applied to protein 3D structures (Igashov *et al. submitted*, Igashov *et al. in revision*, Pages and Grudinin 2019, Pages *et al.* 2019, Derevyanko *et al.* 2018). We have tackled the problems of 3D model quality assessment and tandem repeats detection, representing molecules as 3D grids or graphs. We have also contributed to improve the learning efficiency and accuracy of the networks, by introducing oriented 3D representations (Pages *et al.* 2019) and spherical convolution operators for graphs (Igashov *et al. submitted).*

The main goal of this Ph.D. project is to exploit the pool of moves selected through AS in the protein sequence space to learn about the determinants of protein interactions and functions, and to guide protein design. What can AS tell us about where and how to target proteins in order to modulate their activities? The specific objectives are the following:

1) Confront and complement our set of conserved AS events (AS-induced sequence variation: mutations, insertions, deletions) identified at the human genome scale with transcript-level RNA expression data (GTex[3], BioProject[4]) and mass-spectrometry data (PRIDE[5], ProteomicsDB[6]). This step will produce a curated set of protein isoforms likely to play a functional role in the cell.
2) Create an AS atlas for protein domains. We will map the events defined in (1) onto the corresponding protein domain families (seed multiple sequence alignments, MSAs, from the Pfam classification[7]) and will annotate them with structural information coming from the Protein Data Bank[8]. This step will produce a comprehensive description of where and how AS impacts a given protein domain, on its sequence (represented by an MSA), its fold (represented by an ensemble of structures) and its interactions (with proteins and ligands). We will primarily focus on the domains comprising a large body of structural data and known interacting partners, *e.g.* PF00071, PF00069, PF00071, and PF00069.

[1] https://www.ensembl.org/
[2] https://bgee.org
[3] www.gtexportal.org
[4] www.ncbi.nlm.nih.gov/bioproject
[5] www.ebi.ac.uk/pride
[6] www.proteomicsdb.org
[7] http://pfam.xfam.org
[8] www.rcsb.org

3) Develop a probabilistic model (variational graph autoencoder or deep neural network-powered autoregressive model [29-31]) that will learn the underlying AS « rules » to generate new protein functional diversity. In its simplest form, each training example coming from (2) will be a graph encoding a 3D protein structure, the associated sequence and an AS event (Fig. 1). We will enrich this basic representation by incorporating structures of the interacting molecules and sequence profiles representing homologs, and also for data augmentation purpose (by transferring information from one member of the family to another). We will rely on *active learning* to sample the huge space of possible modifications[9]. The model will modify the network by changing the labels of the nodes and the topology of the network on the fly. One of the challenges will be to go from continuous parameter space search to testable suggestions for protein editing and design.

The project will produce a knowledge base along with web-services, some easy-to-use computational tools and new architecture(s) for graph learning dealing with heterogenous data. We expect the outcome to improve our understanding of the genotype-phenotype relationship and of the underlying determinants of molecular recognition and conformational plasticity.

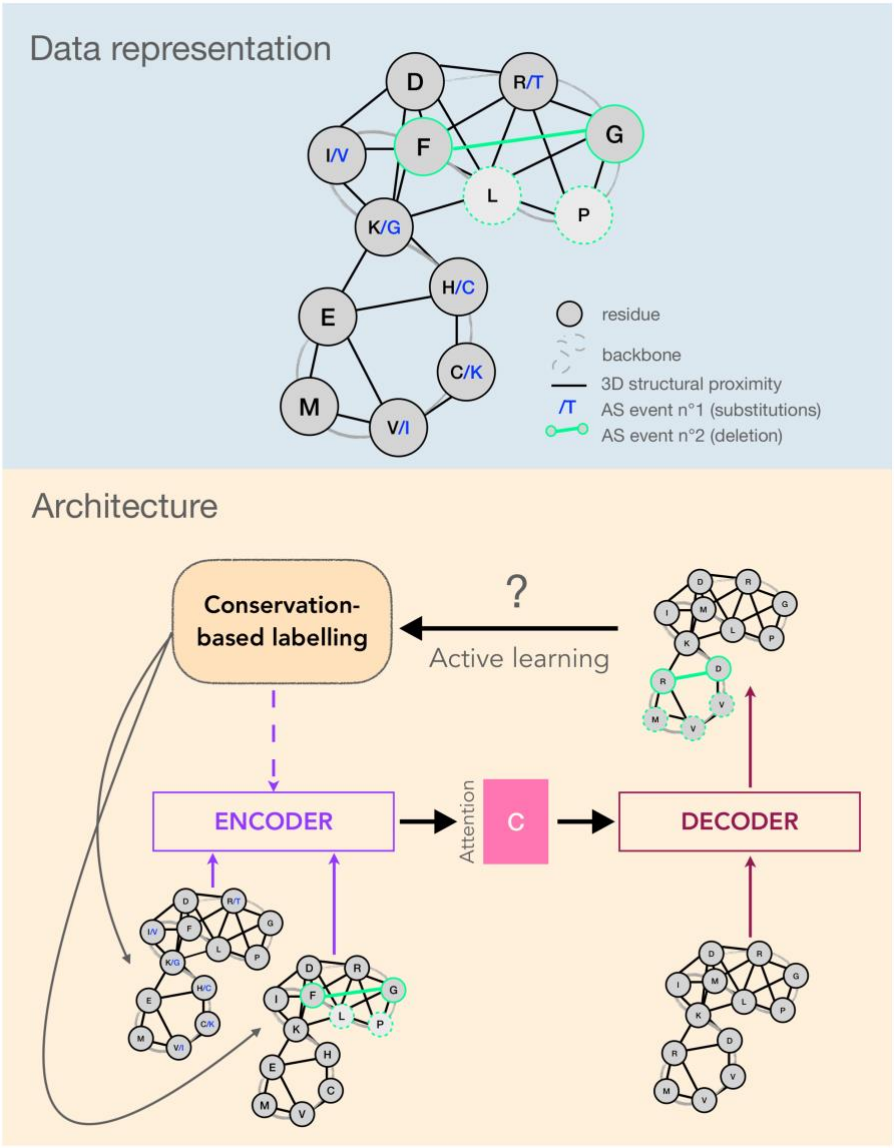

Fig. 1: Sketch of the data representation and the network architecture.

[9] For a protein domain of 100 residues, the number of possible combinations of 5 to 15 substitutions is of the order of 10^36.

Justification of suitability for *i-Bio*:

The proposed project is highly interdisciplinary, at the interface between biology, mathematics, computer science, and physics. Specifically, it lies at the cross-talk of genomics/transcriptomics sequence analysis, protein structural dynamics, evolution, and AI.

The project embeds original concepts about the relationship between genotype and phenotype. The idea that the way living organisms generate protein diversity through AS can inform us about the determinants of protein functioning and can guide protein design is new. To test it, we propose to use cutting-edge AI methods and also to contribute to their development and improve their interpretability.

The role of alternative splicing in the development of diseases like cancer has been firmly established, and we are starting to get a glimpse of how AS natural variations shape human populations' disease susceptibility. Computational methods able to deal with the massive amounts and complexity of the transcriptomics and proteomics data that will be accessible in the coming years will become instrumental in the design of biological interventions improving human health.

We ambition that the outcomes of the Ph.D. project will put the host laboratories and institutions in a leading position in the fields of transcriptomics and graph learning.

Role of each supervisor / skills provided:

EL and SG will co-supervise the student, at 70 and 30% respectively. EL will mostly contribute with her expertise in alternative splicing, protein evolution, and sequence analysis. SG will bring his expertise in machine learning applied to proteins, and more specifically in geometric deep learning. Both EL and SG have some expertise in the analysis and modelling of protein structures and motions. In the past, they have collaborated to develop approaches combining sequence- and structure-based information to predict protein structures and complexes (joint participation to CASP[10] and CAPRI[11]), and on the prediction/description protein functional transitions (Grudinin, Laine and Hoffmann 2020).

EL research activities have been centred on the sequence-structure-function relationship for many years. She has contributed to the development of computational methods exploiting evolutionary and/or structural information to predict protein interfaces (with other proteins and nucleic acids) and binding affinities, to identify protein cellular partners, and to predict mutational effects. She is currently in charge of a project (MASSIV, ANR-17-CE12-0009, 2018-2021) assessing the impact of alternative splicing on protein structures in evolution. She has published over 30 research articles (h-index 14), has deposited two patents and has (co-)supervised two post-docs and four PhD students. She has also been involved in a pedagogical initiative that led to the publication of an education article.

Previous works in the team of EL directly related to the subject:

- Zea DJ, Laskina S, Baudin A, Richard H and Laine E (2020) Assessing Conservation of Alternative Splicing with Evolutionary Splicing Graphs biorxiv 2020.11.14.382820; doi: https://doi.org/10.1101/2020.11.14.382820
- Ait-hamlat A, Zea DJ, Labeeuw A, Polit L, Richard H and Laine E (2020) Transcripts' evolutionary history and structural dynamics give mechanistic insights into the functional diversity of the JNK family *J Mol Biol* 432:2121-2140.

[10] https://predictioncenter.org/index.cgi

[11] https://www.capri-docking.org

SG has been developing about 20 highly efficient computational methods relying on physical principles and machine learning to predict protein-protein and protein-ligand complexes structures, to compare molecular shapes and to predict protein functional motions. In recent years, he has acquired very unique expertise in the development of deep learning architectures for protein 3D structures. This expertise is recognised both at the national level (invitation to present DL for structural biology at the prospective colloquium « Science des données, IA et biologie », Dec 2 2020) and at the international level (invitation to animate the CASP14 round table on deep learning, Dec 4 2020; invitation to present DL for structural biology at the 25th Congress and General Assembly of the International Union of Crystallography, Prague, 2021). He has published over 60 research articles (h-index 22), has deposited one patent and has (co-)supervised four post-docs and eight PhD students.

Previous works in the team of SG directly related to the subject:

- Igashov, I., Pavlichenko, N., & Grudinin, S. (2020). Spherical convolutions on molecular graphs for protein model quality assessment. *arXiv preprint arXiv:2011.07980*.
- Igashov, I., Olechnovic, K., Kadukova, M., Venclovas, C., & Grudinin, S. (2020). VoroCNN: Deep convolutional neural network built on 3D Voronoi tessellation of protein structures. *bioRxiv*.
- Pagès G, Charmettant B, Grudinin S. (2019) Protein model quality assessment using 3D oriented convolutional neural networks. *Bioinformatics*. 35:3313-3319.
- Derevyanko G, Grudinin S, Bengio Y, Lamoureux G. (2018) Deep convolutional networks for quality assessment of protein folds. Bioinformatics. 34:4046-4053.

Profile of the desired student:

The candidate should have a a solid background in computer science or applied mathematics, very good programming skills (C++ and Python) and deep knowledge in linear algebra. S/he should have some knowledge in biology and some familiarity with biological objects such as protein sequences and structures. Experience with -omics data (transcriptomics, proteomics) and some knowledge in evolution are a plus. Teamwork and communication skills are essential for the achievement of the project.

References

1. Graveley, B. R. (2001) Alternative splicing: increasing diversity in the proteomic world. Trends Genet, 17, 100–107

2. Wang, E. T.; Sandberg, R.; Luo, S.; Khrebtukova, I.; Zhang, L.; Mayr, C.; Kingsmore, S. F.; Schroth, G. P.; Burge, C. B. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature, 456, 470–476.

3. Birzele, F.; Csaba, G.; Zimmer, R. (2008) Alternative splicing and protein structure evolution. Nucleic Acids Res. 36, 550–558.

4. Yang, X. et al. (2016) Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. Cell, 164, 805–817.

5. Baralle, F. E.; Giudice, J. (2017) Alternative splicing as a regulator of development and tissue identity. Nature reviews Molecular cell biology, 18, 437.

6. Kelemen, O.; Convertini, P.; Zhang, Z.; Wen, Y.; Shen, M.; Falaleeva, M.; Stamm, S. (2013) Function of alternative splicing. Gene, 514, 1–30.

7. Climente-González, H.; Porta-Pardo, E.; Godzik, A.; Eyras, E. (2017) The functional impact of alternative splicing in cancer. Cell reports, 20, 2215–2226.

8. Scotti, M. M.; Swanson, M. S. (2016) RNA mis-splicing in disease. Nature Reviews Genetics, 17, 19

9. Lim, K. H.; Ferraris, L.; Filloux, M. E.; Raphael, B. J.; Fairbrother, W. G. (2011) Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. Proc. Natl. Acad. Sci. U.S.A., 108, 11093–11098

10. Ward, A. J.; Cooper, T. A. (2010) The pathobiology of splicing. J. Pathol., 220, 152–163

11. Wang, G.-S.; Cooper, T. A. (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. Nature Reviews Genetics, 8, 749–761.

12. Park, E.; Pan, Z.; Zhang, Z.; Lin, L.; Xing, Y. (2018) The expanding landscape of alternative splicing variation in human populations. The American Journal of Human Genetics, 102, 11–26.

13. Byrne, A.; Cole, C.; Volden, R.; Vollmers, C. (2019) Realizing the potential of full-length transcriptome sequencing. Philosophical Transactions of the Royal Society B, 374, 20190097.

14. Sterne-Weiler, T.; Weatheritt, R. J.; Best, A. J.; Ha, K. C.; Blencowe, B. J. (2018) Efficient and accurate quantitative profiling of alternative splicing patterns of any complexity on a laptop. Molecular cell, 72, 187–200

15. Denti L, Rizzi R, Beretta S, Vedova GD, Previtali M, Bonizzoni P. (2018) ASGAL: aligning RNA-Seq data to a splicing graph to detect novel alternative splicing events. BMC Bioinformatics. 19:444.

16. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, et al. (2014) A draft map of the human proteome. Nature. May;509(7502):575–581.

17. Wang X, Codreanu SG, Wen B, Li K, Chambers MC, Liebler DC, et al. (2018) Detection of Proteome Diversity Resulted from Alternative Splicing is Limited by Trypsin Cleavage Specificity. Mol Cell Proteomics. 17:422–430.

18. Gonzalez-Porta, M., Frankish, A., Rung, J., Harrow, J., and Brazma, A.. (2013) Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol., 14:R70.

19. Ezkurdia I, Rodriguez JM, Carrillo-de Santa Pau E, Vazquez J, Valencia A, Tress ML. (2015) Most highly expressed protein-coding genes have a single dominant isoform. J Proteome Res. 14:1880–1887.

20. Weatheritt RJ, Sterne-Weiler T, Blencowe BJ. (2016) The ribosome-engaged landscape of alternative splicing. Nat Struct Mol Biol. 23:1117–1123.

21. Marti-Solano M, Crilly SE, Malinverni D, Munk C, Harris M, Pearce A, Quon T, Mackenzie AE, Wang X, Peng J, Tobin AB, Ladds G, Milligan G, Gloriam DE, Puthenveedu MA, Babu MM. (2020) Combinatorial expression of GPCR isoforms affects signalling and drug responses. Nature. Nov 4. doi: 10.1038/s41586-020-2888-2. Epub ahead of print. PMID: 33149304.

22. Tranchevent, L. C., Aube, F., Dulaurier, L., Benoit-Pilven, C., Rey, A., Poret, A., Chautard, E., Mortada, H., Desmet, F. O., Chakrama, F. Z., Moreno-Garcia, M. A., Goillot, E., Janczarski, S., Mortreux, F., Bourgeois, C. F., and Auboeuf, D. (2017) Identification of protein features encoded by alternative exons using Exon Ontology. Genome Res., 27:1087–1097

23. de la Fuente L, Arzalluz-Luque Á, Tardáguila M, Del Risco H, Martí C, Tarazona S, Salguero P, Scott R, Lerma A, Alastrue-Agudo A, Bonilla P, Newman JRB, Kosugi S, McIntyre LM, Moreno-Manzano V, Conesa A. (2020) tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. Genome Biol. 21:119.

24. Riesselman AJ, Ingraham JB, Marks DS. (2018) Deep generative models of genetic variation capture the effects of mutations. Nat Methods. 15:816-822.

25. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D. (2020) Improved protein structure prediction using potentials from deep learning. Nature. 577:706-710.

26. Greener JG, Moffat L, Jones DT. (2018) Design of metalloproteins and novel protein folds using variational autoencoders. Sci Rep. 8:16189.

27. Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Boscaini, D., Bronstein, M. M., & Correia, B. E. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nature Methods, 17, 184-192.

28. Louadi Z, Oubounyt M, Tayara H, Chong KT. (2019) Deep Splicing Code: Classifying Alternative Splicing Events Using Deep Learning. Genes (Basel). 10:587.

29. Kipf, T. N., & Welling, M. (2016). Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

30. Riesselman, A. J., Shin, J. E., Kollasch, A. W., McMahon, C., Simon, E., Sander, C., ... & Marks, D. S. (2019) Accelerating Protein Design Using Autoregressive Generative Models. bioRxiv, 757252.

31. Ingraham, J., Garg, V., Barzilay, R., & Jaakkola, T. (2019). Generative models for graph-based protein design. In Advances in Neural Information Processing Systems (pp. 15794-15805).