Protein language models transfer learning for the prediction of mutational effects and protein disorder

Gianluca Lombardi

▶ **To cite this version:**

Gianluca Lombardi. Protein language models transfer learning for the prediction of mutational effects and protein disorder. Bioinformatics [q-bio.QM]. Sorbonne Université, 2025. English. ⟨NNT : 2025SORUS618⟩. ⟨tel-05605740⟩

HAL Id: tel-05605740

https://theses.hal.science/tel-05605740v1

Submitted on 28 Apr 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

École doctorale d'Informatique, Télécommunications et Électronique, Paris

Thèse de doctorat
Sorbonne Université

Protein language models transfer learning for the prediction of mutational effects and protein disorder

Présentée et soutenue publiquement le 18 décembre 2025 par:

Gianluca LOMBARDI

Les membres du jury:

Jean–Daniel ZUCKER	Président du jury
Barbara BRAVI	Rapporteuse
Tatiana GALOCHKINA	Rapporteuse
Jun CHENG	Examinateur
Pietro LIÒ	Examinateur
Alessandra CARBONE	Directrice de thèse

Abstract

Inferring the effects of amino acids mutations in protein sequences on their functionality and interactions with other molecules is essential to understand the evolutionary landscape of proteins and to unravel the molecular mechanisms behind rare genetic or degenerative diseases. In this scope, the objective of this thesis is the development of computational methods, mainly based on deep learning, for the prediction of effects of mutations in proteins and protein-protein interactions, from protein sequences. Methods that do not rely on structures are suited to screen increasingly large databases of protein sequences, and to study disordered proteins, for which the characterisation is particularly challenging due to limited or absent structural information. Indeed, the recent introduction of protein language models (PLMs), deep transformers architectures trained on millions of protein sequences, allowed to obtain semantically rich representations of proteins, that can be adapted or serve as inputs for downstream tasks.

The first contribution of the thesis is the introduction of MuLAN, a siamese architecture that leverages transfer learning from PLMs and a local attention mechanism to predict variations in binding affinity in binary protein complexes caused by mutations. Beyond its applications as predictive tool, we demonstrated in multiple evaluation settings that local attention scores learned by the model focus on important regions of the protein sequence, either related to interaction sites or functional motifs. MuLAN was applied in different biological systems and in collaborations with several groups of experimentalists to correctly discriminate between multiple conformations of protein complexes, identify residues responsible for the interaction specificity of a protein family, and to accelerate and enhance the functional classification of proteins.

In an attempt to inspect the link between disorder and protein interactions, previous works introduced the concept of soft disorder, which was shown to correlate with interaction sites and to play an important role in assembly of protein complexes. We expanded on this work by publishing SoftDis, a dataset of soft-disordered and interacting regions based on complexes in the Protein Data Bank. In addition, we showed that embeddings extracted from PLMs could be adapted to predict soft and intrinsic disorder from protein sequences in LoRA-DR-suite framework, reaching state-of-the art performances, as demonstrated in the recent CAID3 challenge.

Concerning the prediction of mutational effects on single protein sequences, we worked on the integration of pre-trained sequence-based PLMs with other modalities and methodologies, in an attempt to improve their predictions. Two promising approaches were the combination of predictions from structure and sequence encoders with ESCOTT, a

rule-based model that relies on evolutionary conservation and global structural information, and the structure conditioning of a pre-trained autoregressive sequence model. In addition, distillation from models including multiple modalities and cross-protein information can be leveraged to accelerate predictions and correct some biases in the models. These approaches led to improved results with respect to baselines, without major computational costs.

In many of these applications, emphasis was placed on the model interpretability with respect to the biological background, which is essential in the field for trustworthy results and possible medical applications.

Résumé

Inférer les effets des mutations d'acides aminés dans les séquences protéiques sur leur fonctionnalité et leurs interactions est essentiel pour comprendre le paysage évolutif des protéines et pour dévoiler les mécanismes moléculaires responsables des maladies génétiques rares ou dégénératives. Dans ce cadre, l'objectif de cette thèse est le développement de méthodes computationnelles, basées sur l'apprentissage profond, pour la prédiction des effets des mutations dans les protéines et les interactions protéine-protéine, à partir des séquences protéiques. Les méthodes qui ne se fondent pas sur les structures sont adaptées au criblage de grandes bases de données de séquences protéiques et à l'étude des protéines désordonnées, dont la caractérisation est particulièrement difficile en raison d'informations structurelles limitées ou inexistantes. L'introduction récente des modèles de langage protéique (PLMs), des architectures profondes de type transformers entraînées sur des millions de séquences protéiques, a permis d'obtenir des représentations sémantiquement riches des protéines, qui peuvent être adaptées ou servir comme inputs pour des tâches de prédiction.

La première contribution de la thèse est l'introduction de MuLAN, une architecture siamoise qui exploite l'apprentissage par transfert à partir des PLMs et un mécanisme d'attention locale pour prédire les variations d'affinité de liaison dans des complexes protéiques causées par des mutations. Au-delà de ses applications comme outil prédictif, nous avons démontré dans plusieurs contextes d'évaluation que les scores d'attention appris par le modèle se concentrent sur des régions importantes de la séquence protéique, liées soit aux sites d'interaction soit aux motifs fonctionnels. MuLAN a été appliqué dans différents systèmes biologiques et en collaboration avec plusieurs groupes d'expérimentateurs afin de discriminer correctement entre plusieurs conformations de complexes protéiques, d'identifier des résidus responsables de la spécificité d'interaction d'une famille de protéines, et d'accélérer et améliorer la classification fonctionnelle.

Afin d'étudier le lien entre désordre et interactions protéiques, des travaux antérieurs ont introduit le concept de désordre soft, qui s'est montré corrélé aux sites d'interaction et jouant un rôle important dans l'assemblage des complexes protéiques. Nous avons approfondi ce travail en publiant SoftDis, un jeu de données de régions de soft désordre et interagissantes, basé sur des complexes dans la Protein Data Bank. De plus, nous avons montré que des représentations extraits des PLMs pouvaient être adaptés pour prédire le désordre soft et intrinsèque à partir des séquences protéiques dans le modèle LoRA-DR-suite, atteignant des performances à l'état de l'art, comme démontré dans le récent défi CAID3.

Concernant la prédiction des effets mutationnels sur des séquences protéiques, nous avons travaillé sur l'intégration de PLMs pré-entraînés basés sur les séquences avec d'autres modalités et méthodologies, dans une tentative d'améliorer leurs prédictions. Deux approches encourageantes étaient la combinaison de prédictions provenant d'encodeurs de structure et de séquence avec ESCOTT, un modèle qui s'appuie sur la conservation évolutive et l'information structurelle globale, et le conditionnement structurel d'un modèle séquentiel autoregressif pré-entraîné. En outre, la distillation à partir de modèles incluant plusieurs modalités et informations inter-protéines peut être exploitée pour accélérer les prédictions et corriger certains biais dans les modèles. Ces approches ont conduit à des résultats améliorés par rapport aux bases de référence, sans coûts computationnels élevés.

Dans beaucoup de ces applications, l'accent a été mis sur l'interprétabilité du modèle par rapport au contexte biologique, ce qui est essentiel dans le domaine pour obtenir des résultats fiables et pour des possibles applications médicales.

Contents

List of Figures **xi**

List of Tables **xiii**

Introduction **1**

1 Protein biology, mutations and disorder **3**
1.1 The building blocks of life . . . 3
1.1.1 Protein structure . . . 3
1.1.2 Protein synthesis . . . 7
1.1.3 Protein-protein interactions . . . 8
1.2 Mutational effects in protein chains and complexes . . . 9
1.2.1 Conservation, coevolution and epistasis . . . 10
1.2.2 Experimental methods . . . 12
1.2.3 Databases . . . 14
1.3 Protein disorder . . . 17
1.3.1 Disorder, binding interfaces and mutational effects . . . 19
1.3.2 Databases . . . 19

2 Computational background **21**
2.1 Deep learning and sequence modeling . . . 21
2.1.1 From perceptrons to deep networks . . . 22
2.1.2 Attention mechanism and Transformers . . . 25
2.2 Learning from multiple sequence alignments . . . 29
2.2.1 Energy-based and latent variable models . . . 29
2.2.2 Multiple attention mechanisms, AlphaFold2 and AlphaMissense . . . 31
2.3 Learning from single sequences: Protein Language Models . . . 32
2.3.1 Multimodality . . . 33
2.4 Prediction of mutational effects . . . 34
2.4.1 GEMME, ESCOTT, PRESCOTT . . . 35
2.4.2 Zero-shot scoring . . . 38
2.4.3 Supervised learning on labeled data . . . 39
2.5 Computational approaches for the prediction of protein disorder . . . 42
2.6 Evaluation metrics . . . 44

3 MuLAN **47**
3.1 Introduction . . . 47
3.2 Results . . . 49
3.2.1 MuLAN predicts binding affinity changes in binary complexes . . . 49
3.2.2 MuLAN reconstructs complex-dependent mutational landscapes based on $\Delta\Delta G$ predictions . . . 59
3.2.3 Computational details . . . 64
3.3 Methods . . . 64
3.3.1 Sequence embeddings . . . 64
3.3.2 A deep learning architecture with light attention . . . 65
3.3.3 PLMs fine-tuning . . . 66
3.3.4 Extraction of Attention weights . . . 67
3.3.5 Datasets . . . 67
3.3.6 Implementation details . . . 69
3.4 Discussion . . . 69
3.5 Data and Software Availability . . . 71

4 Biological applications of MuLAN **72**
4.1 Thioredoxin determinants in *Chlamydomonas reinhardtii* . . . 72
4.1.1 Biological context . . . 72
4.1.2 Results . . . 74
4.1.3 Discussion . . . 76
4.2 SFBB and S-RNase interactions in *Malus* and *Petunia* . . . 77
4.2.1 Biological context . . . 78
4.2.2 Results . . . 78
4.3 Positive selection in aphid gustatory and olfactory receptors . . . 81
4.3.1 Biological context . . . 82
4.3.2 Results . . . 82
4.4 Integration in PLM-View pipeline for functional classification . . . 85

5 SoftDis Dataset and LoRA-DR-suite framework **88**
5.1 Introduction . . . 88
5.2 Results . . . 89
5.2.1 A suite of models based on adapted embeddings . . . 89
5.2.2 Comparative Performance Analysis of LoRA-DR-suite on intrinsic disorder . . . 91
5.2.3 Extending the definition of disorder through the database SoftDis for soft disorder . . . 94
5.2.4 Soft disorder classification . . . 94
5.2.5 Linking soft and intrinsic disorder with pLDDT . . . 95
5.2.6 Adapted embeddings better align with disorder . . . 97
5.2.7 Adapted embeddings influence on attention scores . . . 98
5.3 Materials and Methods . . . 99

5.3.1 Intrinsic disorder datasets . . . 99
5.3.2 Construction of the soft disorder database SoftDis . . . 99
5.3.3 PLMs fine-tuning and hyperparameters selection . . . 100
5.4 Discussion . . . 101
5.5 Data and Software Availability . . . 102

6 Enhancing protein language models predictions 103
6.1 Integration of ESM models in ESCOTT pipeline . . . 103
6.1.1 ESM-IF1 scores as proxy for structural trace . . . 104
6.1.2 Late logits fusion . . . 105
6.1.3 Learned logits fusion . . . 106
6.2 AlphaMissense distillation for fast and accurate predictions . . . 109
6.2.1 Data processing . . . 110
6.2.2 Training and evaluation . . . 110
6.3 Gated structural conditioning in auto-regressive model . . . 112
6.3.1 PoET model . . . 112
6.3.2 Flamingo-inspired integration of structural information . . . 113
6.3.3 Data and training . . . 114
6.3.4 Results . . . 115
6.4 Post-processing bias correction from ranked data . . . 116
6.4.1 Learning from different experimental assays . . . 116
6.4.2 Experimental setup . . . 117
6.4.3 Positional bias enhances global structural features . . . 120

7 Conclusions 123
7.1 Key findings and outlook . . . 124
7.2 Personal feedback . . . 125

Bibliography 126

A `gemme-utils` package 156

B Additional information for MuLAN 160

C Additional information for LoRA-DR-suite 164

D List of articles 173

List of Figures

1.1 Amino acids structure and properties 4
1.2 Hierarchical levels of protein structural organization 5
1.3 Schematic representation of protein synthesis 7
1.4 Schematic representation of affinity variations induced by point mutations . 10

2.1 Transformer architecture with encoder and decoder components 26
2.2 Visual representation of ESCOTT and PRESCOTT methods. 37

3.1 Representation of MuLAN model and its tasks 48
3.2 Performance of MuLAN across evolutionary time 54
3.3 MuLAN large scale predictions of interface residues on 224,888 PDB chains 56
3.4 MuLAN attention scores variations for splicing variants 59
3.5 MuLAN predictions for TRX complexed with diversed partners 60
3.6 MuLAN analysis of the BPTI (Bovine Pancreatic Trypsin Inhibitor) complexed with trypsin 61
3.7 MuLAN predictions for Gain-of-Function (GoF) mutations in MEFV 62

4.1 Structure of the conserved TRX fold 73
4.2 ProfileView analysis of the thioredoxin family and the thioredoxin type-f subfamily. 75
4.3 MuLAN analysis of thioredoxin type f (CrTRXf2) in *C. reinhardtii*. 76
4.4 Experimental results of FBPase activity under redox treatments 77
4.5 Structural analysis of S-RNase and SFBB in *Malus* 79
4.6 Structural analysis of SRNase–SLF interactions in *Petunia* 81
4.7 Phylogenetic trees and detected subfamilies for gustatory receptors (GR) and olfactory receptors (OR). 83
4.8 MuLAN attention scores and positive selection sites in aphid gustatory and olfactory receptors. 84
4.9 Functional trees for TRX family obtained with PLM-View, with and without the integration of MuLAN scores 87

5.1 LoRA-DR-suite model architecture and intrinsic and soft disorder datasets construction 90
5.2 Structural analysis of the Bacillus subtilis Oxalate Decarboxylase OxdC protein. 95
5.3 Structural analysis of the MEFV, Synuclein and SAMP human proteins. . . 96

5.4 Variations in predicted contacts maps induced by LoRA layers. 98

6.1 Performances of models with different input modalities on ProteinGym assays, by function type. 104
6.2 Spearman correlation coefficients for different linear combinations of scores for selected examples in ProteinGym 107
6.3 Optimal lambda values for combination of ESCOTT, ESM-IF1 (λ_1) and ESM2 (λ_2) by function type. 107
6.4 Regression results for λ_1 and λ_2 in LOGO cross-validation test setting. . . . 108
6.5 AlphaMissense dataset statistics 110
6.6 Train and validation loss curves for ConvBert-AM-SaProt model on AlphaMissense human dataset. 111
6.7 Attention scheme in PoET model 113
6.8 Gating values along training phase for structure-adapted PoET model . . . 115
6.9 Spearman correlations for ESM2-650M zero-shot scores and updated scores in LOGO-CV setting for ProteinGym dataset, by E-value. 119
6.10 Spearman correlations for ESM2-650M zero-shot scores and updated scores in LOGO-CV setting for ProteinGym dataset, by tested function, hydrophobicity and position 121
6.11 Spearman correlations for ProSST-2048 zero-shot scores and updated scores in LOGO-CV setting for ProteinGym dataset, by tested function, hydrophobicity and position 122

A.1 Comparative zero-shot analysis of the nuclease activity, NucB, with different tools predicting multiple mutations 159

B1 ROC curve for MuLAN-Ankh-large attention weights over S1102 training set 161
B2 ROC curves for MuLAN-Ankh-large attention weights over 177,545 interacting proteins from the PDB, by secondary structure. 161
B3 Attention profiles for aligned TP53 isoforms (UniProt ID P04637). 162
B4 Attention profiles for isoforms aligned with attention profiles from native sequences 162
B5 Representation of the Light Attention model architecture 163

C1 Distribution of the length of soft-disordered regions in SoftDis clustered dataset 172
C2 Distribution of the number of soft-disordered regions per protein in SoftDis clustered dataset 172

List of Tables

1.1 Function types of ProteinGym DMS assays 15

2.1 Neural networks, domains and symmetries 23

3.1 MuLAN and iMuLAN performance. RMSE is measured in kcal/mol. 50
3.2 MuLAN and iMuLAN performance on benchmark datasets from SKEMPI . 51
3.3 Classification metrics for MuLAN and top five predictors on CAID3 Binding-IDR dataset 58

4.1 Putative targeting SFBB families determined for each S-RNases in the *Malus* dataset 80

5.1 Results of fine-tuned models on CAID1-DisProt and CAID2_NOX test sets 92
5.2 Results of fine-tuned models on the CAID3_NOX test set 93
5.3 Results of LoRA-DR-suite models trained on soft disorder on the SoftDis test set 95

6.1 Performance of the combination of ESCOTT, ESM2-650M and ESM-IF1 models over ProteinGym DMS benchmark 106
6.2 Performance of learned combination of ESCOTT, ESM2-650M and ESM-IF1 models in LOGO cross-validation setting over ProteinGym DMS benchmark 109
6.3 Performance of AlphaMissense distillation over ProteinGym benchmark . . 112
6.4 Performances of structure-aware PoET model in ProteinGym DMS benchmark dataset 116
6.5 LOGO-CV average SCC by selection types for ESM2-650M, ESM2-35M and ProSST-2048 model and variants with learned bias 118
6.6 Results of bias models trained with different embeddings on Human500Domains test dataset 120

B1 Benchmark datasets of changes of binding affinity upon mutation 160
B2 RMSE of MuLAN predictions for fine-tuned PLMs used as feature extractors 160

C1 Results of LoRA-DR-suite models trained on DisProt 7.0 dataset on CAID3_NOX test set. 164
C2 Proteins of CAID3_NOX dataset excluded from the reduced test set 165
C3 Results of LoRA-DR-suite models on the CAID3_PDB test set 166
C4 Results of LoRA-DR-suite models on the CAID3_Binding test set 167

C5 Results of LoRA-DR-suite models on the CAID3_Linker test set 168
C6 Proteins of CAID3_PDB, CAID3_Binding and CAID3_Linker datasets excluded from the respective reduced test sets 169
C7 Performances of logistic regression classifiers trained on model embeddings . 170
C8 Hyperparameters selected for each LoRA-DR-suite model, and ROC AUC score on the respective validation set. 171
C9 Average execution time per protein sequence for LoRA-DR-suite models on the intrinsic disorder evaluation sets 171

Introduction and outline

Understanding the effects of point mutations on protein sequences and how they affect their function and interactions with other molecules is pivotal to study the local evolutionary landscape of proteins, from a theoretical point of view, and to unravel the complex mechanisms behind rare genetic or degenerative diseases, from a clinical perspective. The advancements in speed and scalability of mutagenesis experiments, enabled by high-throughput technologies for gene synthesis and sequencing, has recently allowed to construct large libraries of protein variants with associated effects on protein properties like activity, stability and binding. However, such experiments are still expensive and time demanding, and the pathogenicity of the majority of disease-associated variants, as reported in the ClinVar comprehensive database, is still unknown. In addition, a large fraction of the proteome in living organisms is composed of disordered regions or proteins, that are challenging to detect and characterize but intervene in many cellular processes, in particular regulation and signaling, for which mutations in functionally critical hotspots, while not affecting protein stability, may disrupt entire pathways.

These difficulties encouraged the development of computational methods to predict mutational effects and protein disorder from protein sequences and structure. The recent introduction of novel deep learning architectures, such as large transformer-based protein language models (PLMs), provided a new representation framework, that allows to encode protein sequences in semantically reach embeddings, containing information on amino acids properties, conservation and coevolutionary signals. PLMs, which are trained on millions of protein sequences and, possibly, additional modalities, can capture both global and local sequence patterns, that can be leveraged for downstream prediction of protein properties, including mutational effects and disorder. How to properly extract the right piece of information from protein embeddings and adapt them for specific tasks, however, is not straightforward.

In this thesis, we aimed to develop deep learning solutions for sequence-based predictions of mutational effects and disorder, that can be applied at large scale and possibly improve on already existing methods. The first chapter of the thesis is dedicated to a biological introduction on protein sequences, structures, and interactions, with a focus on the role of disorder and the causes and effects of mutations. Next, we provide an overview of the computational methods developed to address these problems, from deep learning foundations to task-specific models. In this scope, we highlight the importance of family-specific models capable to extract information from multiple sequence alignments, and the advancements introduced by the new generation of PLMs.

Chapter 3 introduces, MuLAN, a novel deep learning model for the prediction of variation of binding free energy upon mutations in protein complexes. Its key component is a convolutional light attention module that processes input proteins embeddings by assigning normalized attention weights to each residue. In various evaluation settings, we demonstrate that high weights are associated to binding and functional regions, providing an unsupervised method for fast interface predictions that extends to disordered regions. We then show applications of MuLAN in different biological systems in Chapter 4, where it provided insights on functional residues and positive selection sites that determine protein specificity to binding targets.

In the fourth chapter we report on SoftDis, a dataset for soft disorder and binding interfaces that extends the classical notion of intrinsic disorder, and LoRA-DR-suite, an ensemble of PLMs fine-tuned with low-rank adapters to predict intrinsic and soft disorder from protein sequences. We evaluate the models' performances on different datasets and tasks related to protein disorder, and we

Lastly, Chapter 6 is dedicated to proposed lightweight extension of existing PLMs to improve zero-shot performances on the predictions of single-point mutations. They include integration of PLMs pseudo-likelihood scores in rule-based models, task-informed models ensembling, model distillation, integration of structural information in a sequence-based pre-trained model, and post-processing corrections from cross-protein data. These methods provide improvement over reference baselines, while requiring reduced computational overhead.

Chapter 1

Protein biology, mutations and disorder

1.1 The building blocks of life

Proteins are large and complex biomolecules that play fundamental roles in all living organisms, being usually referred to as the building blocks of life. They participate in a wide range of biological processes, including the structural maintenance of cells and tissues, signaling, regulation of gene expression, metabolic reactions, immune defense, transport of ions and small molecules, and muscle contraction. Importantly, a large fraction of these processes involve the interaction of proteins with other molecules, comprising other proteins, DNA, RNA and small ligands.

The study of protein structure, functions, interactions, and how they are altered by genetic mutations occurring in one or multiple sites in the protein, is therefore instrumental in understanding complex cellular processes and to develop or improve targeted medical treatments.

1.1.1 Protein structure

The basic constituents of proteins are amino acids, organic molecules that can ligate sequentially to each other to create long peptide chains, through peptide bonds. All amino acids share the same fundamental structure, which consists of a central carbon atom (C_α) with four covalent bonds that connect it to an amino group (-NH_2), a carboxyl group (-COOH), a hydrogen atom, and a distinctive side chain (R group). Peptide bonds are created by a condensation reaction between the carboxyl group of one amino acid and the amino group of the following one, which results in a covalent bond (-CO-NH-) with the production of a water molecule. When they are linked together in peptide chains, amino acids are also referred to as residues. The C_α carbon and the additional nitrogen, carbon, and oxygen atoms that form the covalent bonds connecting the individual residues define the backbone of the protein, whose 3D folded structure is influenced by the side chains.

The R group varies between the different amino acids, defining their unique physical and chemical properties. Based on that, 20 standard amino acids can be commonly found in protein chains, along with few non-standard ones, like selenocystein (an analogue of the more common cysteine with selenium in place of the sulfur atom).

Amino acids are indicated with three-letters or one-letter codes, naturally defining the proteins' alphabet. Based on their properties, amino acids can be grouped into different

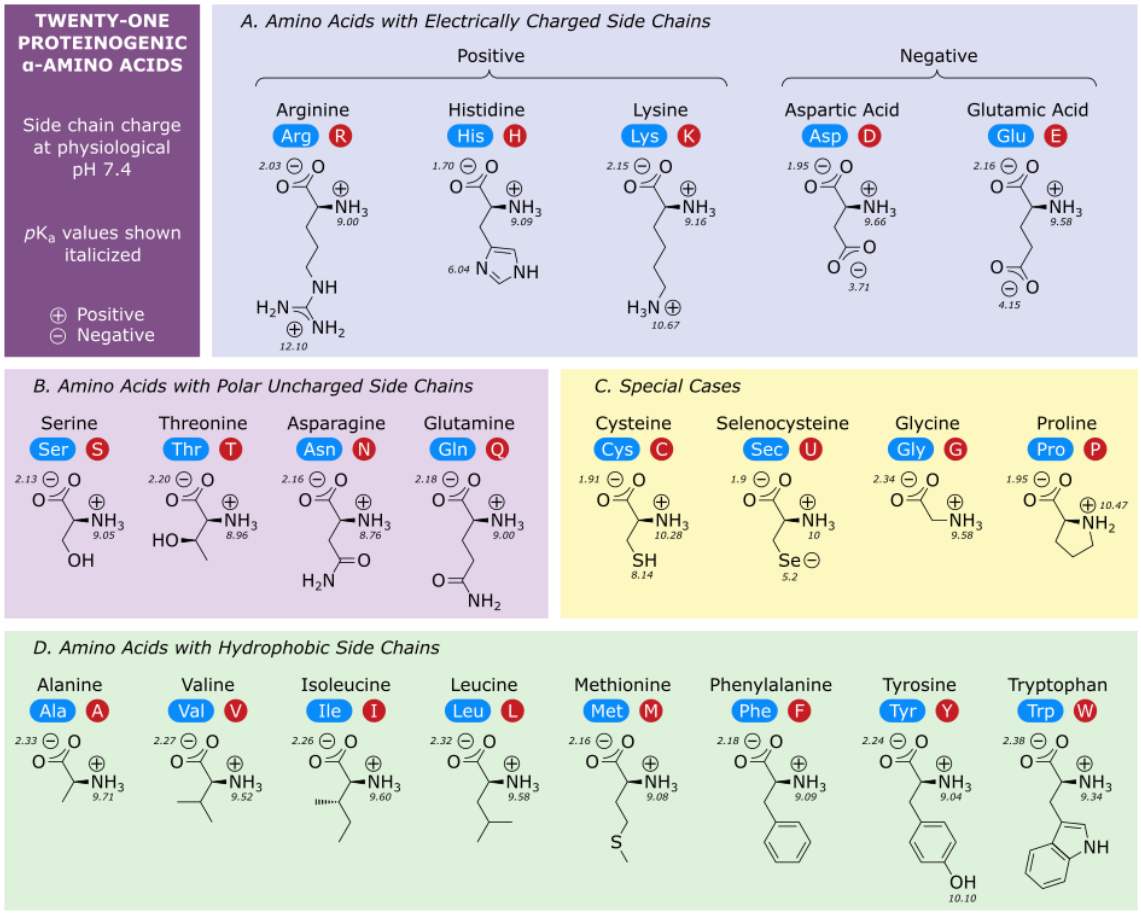

Figure 1.1: **Amino acids structure and physico-chemical properties**. Image is taken from `https://commons.wikimedia.org/wiki/File:ProteinogenicAminoAcids.svg`, and it is licensed under CC BY-SA 4.0 license.

classes (Figure 1.1), which define their structural and functional role in the protein:

– *Charged amino acids.* These residues form ionic bonds that stabilize protein structures through long-range electrostatic interactions. Positively charged amino acids include histidine, lysine, and arginine, whereas aspartic acid and glutamic acid have a negative charge.

– *Polar amino acids.* Due to their hydrophilic nature, polar residues determine protein solubility under physiological conditions. They include asparagine, glutamine, serine, and threonine, and are typically located on protein surfaces, where they interact with the aqueous environment.

– *Hydrophobic amino acids.* These residues induce the formation of protein cores by clustering together to avoid contact with water, thus stabilizing the internal structure. They comprise alanine, isoleucine, leucine, methionine, and valine, along with the aromatic subclass, which includes phenylalanine, tyrosine, and tryptophan.

– *Special cases.* Cysteine, glycine, and proline deviate structurally from standard amino

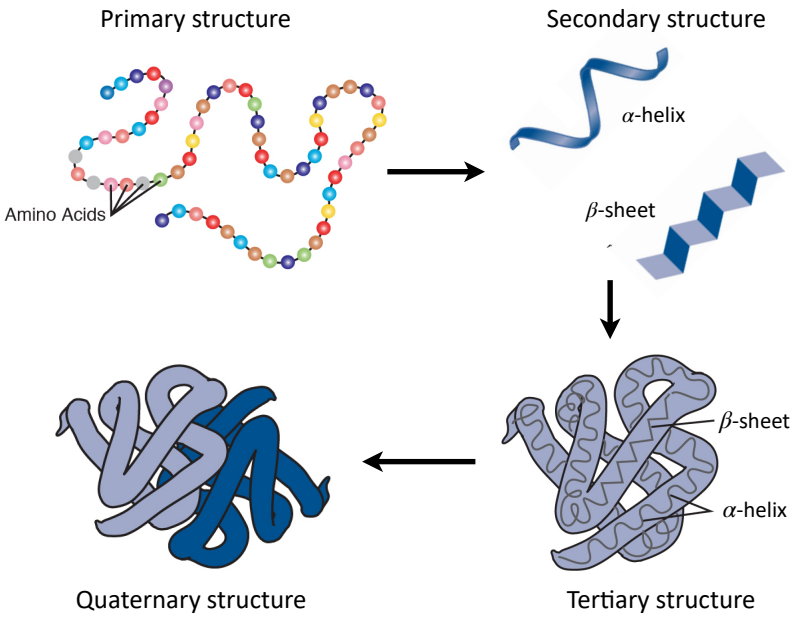

Figure 1.2: **Hierarchical levels of protein structural organization**. Image is adapted from `https://www.genome.gov/genetics-glossary/Protein`.

acids, resulting in unique functional properties. Cysteine contains a sulfur atom in its side chain, allowing the formation of stabilizing disulfide bonds. Glycine, with only a hydrogen atom as its side chain, is the smallest amino acid and provides elevated flexibility. In contrast, proline has a cyclic structure that introduces conformational rigidity in the chain.

The sequence of amino acids in a peptide chain defines its *primary structure*, which is commonly represented as a sequence of 1-letter amino acid codes. The direction conventionally goes from the free amino group (N-terminus) to the free carboxyl group (C-terminus). Through the folding process, the protein adopts its characteristic three-dimensional structure, in which different levels of structural organization can be recognized (Figure 1.2).

The first level of structural organization beyond the primary sequence is the *secondary structure*, that arises from local folding of the protein backbone [**1**]. The most common secondary structure elements are α-helices, β-sheets, and turns or loops, each with distinct properties:

– *α-helices* are coiled structures stabilized by intramolecular hydrogen bonds formed between the oxygen in a carbonyl group, and the hydrogen from the amino group of the amino acid that is four residues apart. This repeating pattern results in a compact, rod-like configuration with a pitch of approximately 5.4 Å, and side chains extending outward from the helical axis. This conformation minimizes steric hindrance and enables functional interactions. Surface-exposed α-helices are particularly abundant in transmembrane proteins, where their hydrophobic residues interact with lipid layers, and in coiled-coil motifs that mediate protein-protein interactions.

– *β-sheets* consist of extended strands aligned in a parallel or antiparallel orientation, stabilized by hydrogen bonds. The resulting sheet-like structure presents side chains that alternate above and below the plane of the sheet, producing surfaces with distinct chemical properties. β-sheets can assemble into large, stable structures such as β-barrels, which frequently form ionic channels and pores in membrane proteins.

– *Turns and loops* connect α-helices and β-strands, allowing proteins to adopt compact, globular shapes. Turns are short motifs, with a typical length of four residues, that break a secondary structure conformation and reverse the direction of the polypeptide chain, the most common type of which is the β-turn. Loops, by contrast, are longer and more flexible strands, often found on protein surfaces, where they contribute to molecular recognition and binding specificity.

Other secondary structure elements exist, like diverse types of helices and turns, or intermediate metastable states, but they are rarely observed in natural proteins. In addition, *random coils* define the class of conformations that indicate an absence of any regular secondary structure.

Experimentally, a common method to characterize the secondary structure of a protein is Circular Dichroism (CD) spectroscopy, in the far-UV range (spanning approximately 180 nm to 260 nm) [**2**].

The three-dimensional arrangement of a polypeptide chain, resulting from the folding and packing of secondary structural elements into a typically compact and globular conformation, defines the protein *tertiary structure* [**3**]. Unlike secondary structure, which is largely stabilized by local hydrogen bonds between atoms in the protein backbone, tertiary structure is maintained by long-range interactions between residues side chains. The most important include hydrophobic interactions, which guide non-polar residues towards the interior of the protein to minimize their exposure to water molecules, hydrogen bonds between polar side chains, and strong ionic interactions (salt bridges) formed between oppositely charged residues, like lysine and glutamate, that contribute to structural stability. Covalent interactions between spatially close cysteine residues define disulfide bonds, which act as molecular cross-links and further reinforce protein folding, as they are common in extracellular proteins such as insulin or immunoglobulins. Additionally, weaker Van der Waals interactions provide minor but essential stabilization by optimizing local conformations of side chains atoms.

It is at this level of structural organization that the formation of protein domains occurs. Protein domains are compact and semi-independent structural units within a polypeptide that can often fold and function autonomously. Domains typically consist of 50-250 amino acids, and their modular nature enables evolutionary flexibility, as recombination and duplication of single domain units determines the emergence of novel protein architectures and functions.

The formation of a stable tertiary structure represents the critical level at which protein sequence information translates into biological function, and even small destabilizing events can lead to loss of activity or pathological aggregation.

Finally, the *quaternary structure* refers to the spatial organization and interaction of multiple polypeptide chains into a very stable functional complex. Each polypeptide chain within a protein complex is referred to as a subunit. Complexes composed of identical subunits are termed homomeric, while those formed by different subunits are called heteromeric.

Quaternary assemblies can form simple dimers (two proteins) or highly complex multimers, such as the hemoglobin tetramer (4), where cooperative binding of oxygen to heme groups is enabled by subunits communication, which highlights the functional relevance of such conformation. Another notable example is the ribosome, a massive ribonucleoprotein complex in which protein subunits interact with RNA to enable protein synthesis.

Ultimately, quaternary structure enables functional versatility through cooperative interactions, allosteric regulation, and the ability to form macromolecular machines that are essential for cellular processes.

1.1.2 Protein synthesis

Proteins are synthesized through a tightly regulated biological mechanism that ensures the accurate expression of genetic information encoded in the deoxyribonucleic acid (DNA). In eukaryotic organisms, DNA is contained in the nucleus, while proteins are assembled in ribosomes, situated in the cell cytoplasm. To enable communication between these two regions, protein synthesis involves two steps, namely transcription and translation (Figure 1.3).

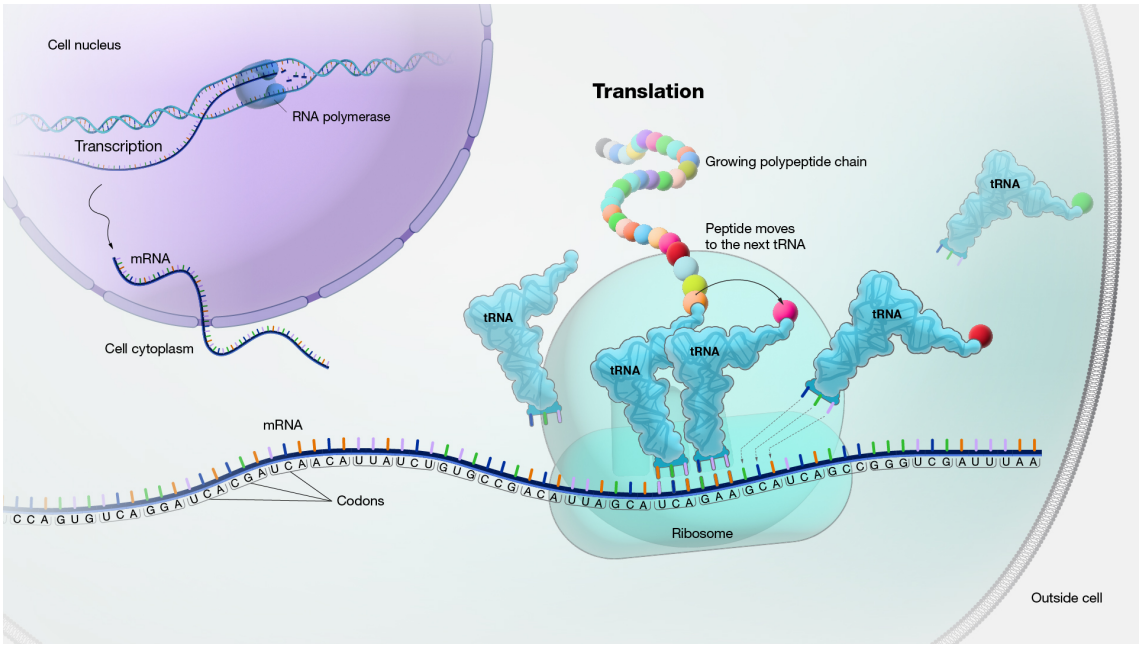

Figure 1.3: **Schematic representation of protein synthesis, highlighting the transcription and translation phases**. Image is taken from `https://www.genome.gov/genetics-glossary/Translation`.

– During *transcription*, which takes place in the nucleus of the cell, genetic information is transferred from DNA to RNA. During transcription, a segment of DNA (gene) that contains a promoter, a protein coding region, and a terminator, is copied into a complementary strand of messenger RNA (mRNA). This process is mediated by different enzymes, in particular RNA polymerase, that binds to the promoter region of the gene, unfolds the compact structure of DNA and synthesizes the mRNA transcript. In eukaryotes the new mRNA, which is called pre-mRNA at this stage, undergoes multiple post-transcriptional modifications that ensure its stability in the passage from the nucleus to the cytoplasm and may modify the coding regions. Such modifications include the addition of a 5' cap, splicing of non-coding introns, polyadenylation of the 3' end, and RNA editing. After this step, the mature mRNA strand leaves the nucleus.

– Once in the cytoplasm, the second stage of synthesis, *translation*, begins in the ribosome, where the sequence of nucleotides in the mRNA is decoded into a sequence of amino acids. The ribosome binds to the mRNA and begins to read the sequence of nucleotides in triplets, that are called codons. For each codon, a specific transfer RNA (tRNA) molecule which carries the corresponding amino acid is recruited. The ribosome then catalyzes the formation of peptide bonds between amino acids, enabling the formation of the polypeptide chain. The process continues until a stop codon is reached, signaling termination and release of the newly synthesized protein. The $4^3 = 64$ possible triplets decode for only 20 amino acids (besides start and stop codons): this redundancy is called codon degeneracy, and it accounts for the existence of synonymous mutations.

1.1.3 Protein-protein interactions

In many cases, the functionality of a protein depends on its interaction with other proteins or molecules. The stability of a protein complex is physically described by the binding free energy ΔG_{bind}, which represents the difference in Gibbs free energy between the bound and unbound states of the proteins, and determines the affinity and specificity of the interaction [**4**]. A negative ΔG_{bind} indicates that the complex is energetically favorable, and it arises from a balance of enthalpic contributions, accounting for inter-protein interactions (hydrogen bonds, electrostatic interactions, van der Waals forces) and entropic effects such as hydrophobic burial and conformational restrictions induced by the binding. ΔG_{bind} can be estimated by the thermodynamic relation

$$\Delta G_{bind} = -RT \ln K_d, \tag{1.1}$$

where K_d is the equilibrium dissociation constant, R the gas constant, and T the absolute temperature.

The ensemble of protein-protein interactions (PPI) in a closed environment defines a network, that can be analysed in order to understand how proteins cooperate to perform biological functions. PPI networks are often scale-free, with a few highly connected hub proteins mediating multiple interactions and playing essential roles in cellular processes.

1.2 Mutational effects in protein chains and complexes

Alterations and damages in the nucleotides sequence of DNA may arise spontaneously, due to errors in DNA replication, recombination, or as a consequence of external mutagens such as ultraviolet radiation, chemical agents, or ionizing radiation. To mitigate the potentially lethal effects that such damages might cause, cells have evolved many biological responses, that are classified in either DNA repair mechanisms, or damage tolerance mechanisms [**5**].

All these processes, while tightly regulated especially in eukaryotic cells, are still prone to errors with variable frequencies, even in healthy organisms. These errors ultimately result in mutagenesis, the process by which a permanent and heritable change to the DNA occurs. These changes are generally classified into single nucleotide variants (SNVs) including point mutations, insertions or deletions (indels), copy number variations, and larger-scale chromosomal rearrangements. Mutagenesis represents an essential driver of genetic variation and evolution, but it also contributes to disease susceptibility, cancer development, and aging [**6**].

Focusing on single genes, if one or more nucleotides are mutated, this variation will be expressed during protein synthesis. Due to codon degeneracy, point mutations might still lead to the same encoded amino acid: in this case of *synonymous* (or silent) mutation, the synthesized protein will not change. On the contrary, if the amino acid changes, the mutation is called non-synonymous, or *missense*, and it induces a variation in the protein sequence with respect to the wild-type. In the case it introduces a premature stop codon, the mutation is called *nonsense*: it causes the translation of the gene into a protein to terminate prematurely, resulting in a shorter, often non-functional protein. A similar outcome might be caused by frameshift mutations, induced by indels that are not multiples of three nucleotides and alter the downstream reading frames. Besides coding regions, mutations in regulatory elements such as promoters, enhancers and untranslated regions can also affect transcription initiation and efficiency, or RNA stability. Alterations in canonical splice sites may instead disrupt normal mRNA processing, leading to the production of non-functional protein isoforms.

The impact of missense mutations on protein chains depends critically on their location in the protein structure and on the characteristics of the involved amino acids. At the structural level, missense mutations can disrupt local folding and affect protein stability, especially when they alter essential hydrophobic packing, hydrogen bonding, or ionic interactions. This happens when they occur in highly conserved residues, typically in the core of the protein or in exposed functional domains. In this second case, mutations may also drastically impact catalytic activity by altering active site residues, prevent oligomerization by affecting protein-protein interfaces, or disrupt sites of post-translational modifications essential for regulation.

In the case of protein-protein interactions, missense mutations might induce variations in the binding free energy by modifying the physicochemical and geometrical properties of the interacting surface, and thus reducing or augmenting the affinity towards specific partners.

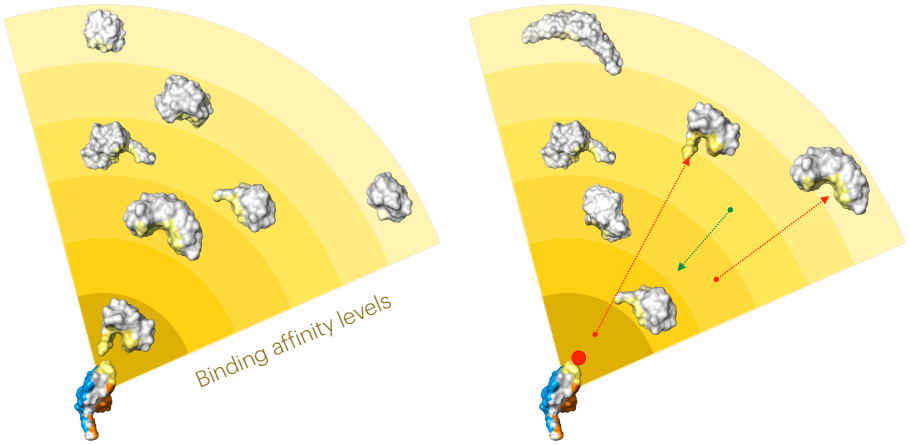

Figure 1.4: **Schematic representation of affinity variations induced by point mutations.** Mutations on protein interaction surface may change its affinity towards different partners, effectively modifying the protein-protein interaction network. Courtesy of Alessandra Carbone.

Considering the complex PPI network of an organism, perturbations induced by mutations in one protein, especially if part of a highly interconnected hub, might alter affinities with multiple partners, causing cascading effects and disrupting entire pathways (Figure 1.4). As individuals of any species carry multiple (mainly benign) mutations in their genome, the slight variations in PPIs induced by these mutations are one of the reasons why some drugs or medical treatment are effective on some patients, but null or even harmful in other cases.

In addition to the more common loss-of-function (LoF) effects, mutations can induce toxic gain-of-function (GoF) phenotypes, that amplify the physiological activity of the protein or even confer a new altered functionality that is deleterious for the organism. GoF mutations can lead to misfolding and aggregation, alteration of protein interactions, or dysregulation of critical cellular processes, as observed in neurodegenerative disorders like Huntington's disease and Amyotrophic Lateral Sclerosis (ALS). A prominent example of GoF mutations are those occurring in the MEFV gene, which are the primary cause of Familial Mediterranean Fever (FMF), an autoinflammatory disease. Some missense mutations in MEFV (e.g., M694V, M694I, and V726A), disrupt the normal autoinhibitory regulation of pyrin, leading to uncontrolled inflammatory responses that produce variable disease phenotypes [**7**]. Predicting the pathogenicity of GoF mutations is indeed a challenging task, as increased activity is often associated with an improved protein fitness, which is typically the target quantity in computational predictions.

1.2.1 Conservation, coevolution and epistasis

Protein sequences that we observe in the natural world are the outcome of evolutionary forces acting on living organisms over billions of years. Evolution preserves sequences that ensure proper folding and functionality, while most deleterious variants that do not

guarantee survival are removed. As a consequence, proteins encode both the history of random mutational events and the selective constraints imposed by their functional roles. Proteins that share a common evolutionary origin form protein families, which are groups of homologous sequences that typically display a conserved structural and functional framework, ensured by significant sequence similarity. Within a family, highly conserved residues are a direct consequence of evolutionary constraints, as substitutions at those positions are strongly suppressed. Conserved elements include catalytic sites in enzymes, residues that stabilize key secondary structure elements, or core residues essential for maintaining overall stability and functionality. In contrast, variable positions are typically located on solvent-exposed surfaces or flexible loops where mutations are more easily tolerated [**8**]. Sequence variations in such positions may contribute to environment-specific and adaptations, or may instead drive functional diversification and specificity, both across species and within a single species [**9**].

Under the homology assumption, the detection of conservation patterns across protein families is achieved by comparing similar sequences from different species using multiple sequence alignments (MSAs). MSAs introduce gaps into the sequences to align homologous positions, with the goal of creating a matrix where each line is a (gapped) sequence, and columns represent conserved features or shared functional units across the protein family. The construction of informative MSAs is a complex optimization problem, where the objective is to maximize an overall score that reflects the similarity and functional relatedness of the input sequences, and a plethora of algorithms have been developed that rely on dynamic programming, heuristics strategies, and probabilistic modelling [**10**].

Beyond conservation, evolution also determines sequence of co-variations, due to the fact that residues interact physically and functionally in the folded three-dimensional structure. A mutation that destabilizes a local structure or alters a functional site can be balanced by compensatory mutations in structurally close residues, leading to detectable *coevolution* patterns across sequences in the MSA. Strong coevolution signals between aligned positions in a protein family typically correspond to residues contacts in the folded structure. Similarly, by aligning orthologous pairs of interacting partners from multiple species, coevolution patterns that reflect the reciprocal adaptation of residues at binding interfaces can also be detected. These signals, however, are usually weaker, due to the fact that inter-protein contacts are typically fewer and have a transient nature, and the same residues may be involved in interactions with multiple partners, leading to less consistent selective pressure. Such coevolution patterns are not limited to single organisms, but appear also in paired species, for example in the co-adaptation of proteins mediating the interactions between parasites or pathogens and their hosts [**11**].

A particularly important consequence of such interdependencies inside a protein sequence is the phenomenon of *epistasis*, where the pathogenicity of a mutation depends on the presence or absence of other mutations in the sequence, that define its genetic background. Epistasis can be positive, when combined mutations confer a fitness advantage greater than expected from their individual contributions, or negative, when the combined effect is deleterious

despite the individual mutations being neutral or beneficial [**12**]. This non-additivity of mutational effects reflects the ruggedness of the protein fitness landscape, where different mutational trajectories can lead to divergent evolutionary outcomes [**13**]. For example, the adaptive evolution of antibiotic resistance enzymes such as beta-lactamases has been shown to depend on epistatic interactions that constrain or enable particular evolutionary pathways [**14**].

Protein evolution is therefore a context-dependent process shaped by intricate networks of residue interactions, where epistatic effects can amplify the selective consequences of random mutations and influence evolutionary paths by affecting proteins fitness landscapes. It is in fact the local conformation of this fitness landscape, also in relation with protein activity and interactions, that ultimately determines the pathogenicity of a missense mutation in the protein sequence.

1.2.2 Experimental methods

Several experimental techniques have been developed to systematically assess the effects of mutations on protein sequence, structure, and function. Depending on the scale and throughput, they can be included in two main categories: site-directed mutagenesis methods and deep mutational scanning.

Site-directed Mutagenesis and functional assays. Site-directed mutagenesis (SDM) enables the design of specific nucleotide substitutions, insertions, or deletions in a DNA sequence, thus altering the encoded protein at defined positions [**15**]. The classical approach of SDM, often referred to as the Kunkel method, relies on the use of synthetic oligonucleotides that carry the desired mutation to prime DNA synthesis, resulting in a plasmid containing the modified gene sequence. Polymerase Chain Reaction (PCR) is used for gene amplification, in order to exponentially increase the number of DNA segments [**16**].

Once the targeted mutation is introduced, the mutated gene is expressed in a suitable expression system, including bacterial, yeast, or mammalian cells, and the resulting mutated protein can be isolated and studied through functional assays. The particular experimental technique depends on the biological role of the protein under investigation and the physical quantity to measure. For enzymes, catalytic activity is quantified using substrate conversion assays or kinetic measurements. For binding proteins, ligand-binding affinities can be determined using surface plasmon resonance (SPR), isothermal titration calorimetry (ITC), or fluorescence-based methods. Structural and stability changes, instead, can be probed by thermal denaturation or protease sensitivity assays. Additionally, structural characterization of the protein at required resolution can be obtained through X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, or cryo-electron microscopy (Cryo-EM).

Deep Mutational Scanning. Deep Mutational Scanning (DMS) combines site-directed mutagenesis, high-throughput functional assays and next-generation sequencing to quantify the effects of thousands of mutations in parallel. While traditional mutagenesis methods typically focus on introducing and analyzing one or few mutations at a time, DMS allows to

evaluate all possible single amino acid substitutions, and even combinations of mutations, across an entire protein sequence [**17, 18**]. This approach eliminates the need to select specific mutagenesis sites based on prior hypothesis, and it provides a comprehensive approximation of the fitness landscape in the neighborhood of the wild-type sequence, that reveals how each residue contributes to protein fitness.

In the first step of DMS, a diverse library of protein variants is generated. This can be achieved through a variety of mutagenesis techniques. For single-point mutational coverage, site-saturation mutagenesis (SSM) is often used, where each codon in the gene is systematically replaced by codons encoding all 20 amino acids and, possibly, stop codons. Error-prone PCR (epPCR) can also be employed to increase the initial sequence library by introducing a broad set of random mutations [**19**]. Similarly to SDM, the resulting variant library is then cloned into an expression vector and expressed in cells where the protein encoded by the gene can carry out a function that enables a selection mechanism, which enriches cells with active protein variants and depletes those with deleterious ones.

In this second stage, functional selection links the protein's activity, stability, or interaction properties to a measurable phenotype. Depending on the biological property to be inspected, this may involve growth under selective pressure, like antibiotic resistance, toxicity, binding to a target molecule, enzymatic activity assays, or localization-based readouts. The critical requirement is that the assay must be capable of distinguishing functional variants from non-functional ones in a high-throughput manner [**20**].

Subsequently, the frequency of each variant before and after selection is quantified by next-generation sequencing (NGS). High-throughput sequencing provides a count of each variant's representation in pre-selection and post-selection populations. The relative change in frequency of a given variant provides an estimate of its fitness or functional impact compared to the wild-type protein. Typically, fitness scores are computed as log-ratios of output to input frequencies, normalized against synonymous variants or wild-type controls to limit noise [**17**]. Resulting functional scores for single-point mutations can then be represented in matrices of size $L \times 20$ (21 including stop codon), where L is the length of the protein fragment considered and each value maps the corresponding mutation to its measured outcome.

Indeed, an important limitation of *in vitro* mutagenesis experiments is that functional measures can be context dependent, for instance considering isolated proteins in non-physiological conditions, which might mask or exaggerate the effects of mutations on protein fitness, or may fail to reflect how those mutations impact fitness at the cellular or organismal level. Additionally, detecting subtle fitness changes is often limited by experimental noise, assay dynamic ranges, or sampling constraints, which potentially leads to underestimation of mutational effects and limits in the reproducibility between alternative replicas of the same experiments [**21**].

1.2.3 Databases

UniProt and PDB

While not directly related to mutations, UniProt [**22**] and the Protein Data Bank (PDB) [**23**] are two of the reference databases of curated sequence and structural information for proteins.

UniProt is the most widely used and authoritative protein sequence and functional annotation database, that integrates experimentally validated data with computational predictions and external databases to provide a unified resource for the whole community. Each entry includes rich annotations such as sequence features, taxonomic and functional annotations, domain architectures, interactions, partial or complete structures, and possible diseases associated to known variants. The current version of UniProt comprises more than 250 million protein sequences, with more than 570 thousands that are manually reviewed.

Complementary to this, the PDB is the primary database for experimentally determined three-dimensional structures of biological macromolecules, including proteins but also nucleic acids and small molecules. It hosts structural models up to atomic-level resolution obtained by experimental techniques such as X-ray crystallography, NMR spectroscopy, and the increasingly popular cryo-EM. Since its establishment in 1971, the PDB has exponentially grown into a critical infrastructure comprising now more than 240,000 deposited structures, that enables the exploration of structural motifs, binding interfaces, biological assemblies, and conformational variability, and the training and evaluation of deep learning methods of molecular folding.

ClinVar

ClinVar is a publicly available database that aggregates reports of relationships between sequence variation in human genes and clinical phenotypes, integrating submissions from clinical laboratories, research groups, expert panels, and practice guidelines [**24**]. Each record links a genetic variant to asserted clinical significance categories, namely Pathogenic, Likely Pathogenic, Likely Benign, Benign, or Variant of Uncertain Significance (VUS) [**25**]. VUSs fraction has been consistently increasing over time, and now constitute more than 70% of all the entries [**26**].

ClinVar implements a review status system that captures the level of consensus (single submitter, multiple submitters with/without conflicts, expert panel, practice guideline) that can guide researchers in assessing the reliability of variant interpretations and prioritizing variants for downstream analyses, clinical reporting, or functional validation.

Currently, the database comprises more than 3.7 million unique records with clinical significance classification, representing around 90 thousands genes. ClinVar is also integrated with complementary resources for gene and protein annotation (UniProt, Ensembl, HGNC) enabling cross-referencing of sequence variants with protein function, gene nomenclature, and genomic context.

ProteinGym benchmark

An important data source that combines data from mutational assays collected in different experiments is ProteinGym [**27, 28**], an extensive benchmark that integrates both experimental and clinical data, and introduces a robust evaluation scheme that allows comparisons of predictive models across different biological contexts. In the following, we refer to the version 1.3 of ProteinGym.

Data Subsets. ProteinGym consists of two main types of data sources: deep mutational scanning (DMS) assays and clinical variant datasets.

– *Deep Mutational Scanning (DMS).* The benchmark aggregates 217 DMS assays spanning approximately 2.4 million missense mutations across 186 unique proteins from various taxa, including humans, other eukaryotes, prokaryotes, and viruses. They comprise diverse protein families, such as kinases, ion channel proteins, G-protein coupled receptors, polymerases, and transcription factors. DMS assays include mostly single-point mutations, but multiple-point mutations are also present. In addition, the dataset includes 66 assays evaluating nearly 300,000 indels.

Data are collected from previous curated works, particularly EVMutation benchmark [**29**], DeepSequence benchmark [**30**], and thermostability data [**31**], but also expanded with additional experimental sources. DMS assays are grouped into five function categories, based on the nature of the experimentally measured properties. A summary of function (or selection) types and numbers of assays is provided in Table 1.1, derived from Table A2 in [**28**].

Function type	# Assays		Measured property
	Subs	Indels	
Activity	43	3	Catalytic or biochemical activity
Binding	14	0	Affinity or degree of binding to target
Expression	17	2	How much the protein is expressed in the cell
Fitness	77	6	Organism's growth rate
Stability	66	55	Thermostability

Table 1.1: **Function types of ProteinGym DMS assays**. For each type, the number of assays in the substitutions (Subs) and indels datasets is indicated, with a description of the measured property.

– *Clinical Variants.* The clinical benchmark dataset includes about 63,000 curated missense substitutions across 2,525 proteins, and around 3,000 indels distributed among 1,555 genes. These data are collected from authoritative resources such as *ClinVar* [**24**] and *gnomAD* v2.1.1 [**32**], which provide pathogenicity labels, population frequencies, and functional annotations. Compared to the continuous measurements from DMS, clinical datasets are discrete, distinguishing pathogenic from benign variants in pathogenic and benign categories.

For DMS data, assays were processed in a standardized way to record mutations and their corresponding experimentally measured effects. MSAs and structures are also provided

for each protein: MSAs are generated with JackHMMer [**33**], searching sequences against UniRef100 database [**34**], while structures are predicted using AlphaFold2 pipeline [**35**]. Metadata for each assay include UniProt identifiers, wild-type sequence, species information, MSA depth, related publication and assay-specific details such as binarization thresholds.

All datasets, metadata, MSAs, and structural files are made publicly available via the ProteinGym website[1] and its associated GitHub repository, to guarantee reproducibility and accessibility of the benchmark.

Evaluation Settings. Two main evaluation paradigms are considered: in the *zero-shot setting*, models are tested on unseen sequences without supervised fine-tuning, which is a particularly relevant task in cases where labels are scarcely available; in the *supervised setting*, models are trained or fine-tuned on a subset of the data and tested on held-out variants, enabling the evaluation of generalization performance and simulating protein design pipelines, where new data are progressively collected. Multiple cross-validation schemes are designed to test increasingly difficult generalization capabilities of the mdoels.

Evaluation metrics depend on the dataset. For continuous DMS assays, model predictions are evaluated using Spearman's correlation, but also area under the receiver operating characteristic curve (ROC AUC) and Matthews correlation coefficient (MCC), computed over binarized labels. Additional evaluation metrics include normalized discounted cumulative gain (NDCG), and top-k recall, important for protein design. For clinical datasets, evaluation is based primarily on ROC AUC. A summary of the metrics considered in this work is provided in Section 2.6 of the manuscript.

Different stratification strategies of results are implemented to enable extensive analysis of models performance, namely taxonomic group, MSA depth, number of mutations (single, double, multiple) and selection types.

Models Coverage. ProteinGym has been used to evaluate more than 90 models[2], spanning multiple categories: alignment-based methods, inverse folding approaches, protein language models, and clinical variant effect predictors. New baselines are continuously added to the benchmark, providing an up-to-date picture of the models landscape and highlighting the strengths and weaknesses of the different approaches on the different tasks.

MaveDB

Although not directly considered in our work, MaveDB (Multiplexed Assays of Variant Effect Database) is a public, curated repository for datasets produced by DMS and other multiplexed functional assays that quantify the effects of genetic variants at scale [**36**]. The database design involves a hierarchical structure where each study is organized into experiment sets, experiments, and score sets, providing consistency and reproducibility across datasets.

In recent years, following significant updates to the database [**37**], MaveDB has become a primary resource for compiling large-scale mutational datasets that are now widely used for benchmarking variant effect prediction methods and training machine learning

[1] `https://proteingym.org`

[2] Up-to-date leaderboard is available at `https://proteingym.org/benchmarks`

models. As a matter of fact, the current release comprises more than 7 million variant effect measurements derived from hundreds of experiment sets.

SKEMPI

SKEMPI (Structural database of Kinetics and Energetics of Mutant Protein Interactions) v2.0 is the most complete source for experimentally measured binding affinities of wild-type and mutated protein complexes [**38**], that includes the smaller datasets AB-Bind, PROXiMATE, and dbMPIKT. It includes measurements for over 7,085 mutations in 345 protein complexes from the PDB, comprising 5112 single-point mutations, 911 double-point mutations and 1062 multiple-point mutations of at least 3 residues, and up to 20. Considered proteins include protease-inhibitor (Pr/PI) assemblies, antibody-antigen (AB/AG) complexes, and assemblies formed between major histocompatibility complex proteins and T-cell receptors (pMHC-TCR).

For each entry, corresponding to a single- or multiple-point mutation, the database provides the sequences and PDB structure of the wild-type complex, the identifiers and names of the partners, the affinities of the wild-type and mutated complexes K_d^{wt} and K_d^{mut}, some related experimental measurements, details about the experimental method and conditions, and the structural region of the mutation sites, either INT, SUR, SUP, COR or RIM, according to the definitions in [**39**]. Given the affinities, the variation in binding free energy $\Delta\Delta G$ can be computed as

$$\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt} = RT \log\left(\frac{K_d^{mut}}{K_d^{wt}}\right) \tag{1.2}$$

where ΔG_{mut} and ΔG_{wt} are the binding free energies of the mutated and wild-type complex, respectively. Most of the mutations in SKEMPI are deleterious, resulting in positive $\Delta\Delta G$ values. In general, mutations happening in the interface (SUP, COR, RIM), in particular in the core (COR), induce bigger changes in binding affinity than those located in the non-interacting surface (SUR) or the interior (INT) of the protein. SKEMPI database is the reference data source for the training of MuLAN model, introduced in Chapter 3.

1.3 Protein disorder

In Section 1.1.1, we introduced the traditional description of protein structure in terms of well-defined secondary, tertiary, and quaternary conformations, that determine the protein function. However, a large fraction of protein regions do not display a stable three-dimensional fold under physiological conditions. These regions, known as *intrinsically disordered regions* (IDRs), lack a single stable conformation, and exist instead in dynamic conformational ensembles that fluctuate on multiple timescales [**40, 41**]. When the whole protein presents this property, it is refereed to as *intrinsically disordered protein* (IDP). In eukaryotes, more than 30% of proteins contain at least one IDR, with estimates reaching up to 50% for the human proteome [**42**].

In terms of amino acid composition, IDPs present a low sequence complexity, as they

are enriched in polar and charged residues and display a low content of bulky hydrophobic amino acids. Thus, their unusual sequence composition reduces their propensity to collapse into globular structures. For this reason, disordered regions cannot be studied through X-ray crystallography, as they appear as a lack of electron density in local regions. They can instead be characterized with spectroscopic methods such as NMR, fluorescence spectroscopy, or circular dichroism [**40**].

The study of disordered proteins and their activity in the cells led to the emergence of a new *disorder-to-function* paradigm, that recognizes the essential role of protein disorder in the biological processes that require flexibility and adaptability. Usually, functional regions within IDRs are composed of short stretches of adjacent amino acids, that are known as short linear motifs (SLiMs) or molecular recognition features (MoRFs). These motifs mediate transient, low-affinity contacts with diverse binding partners, enabling disordered regions to act as versatile interaction hubs in PPI networks [**43**]. The typical features of these regions, that are generally short and with few essential residues, makes it easier for them to evolve from a random peptide sequence through different mutational paths of point substitutions, compared to interaction surfaces in globular domains, for which the fitness landscape is much more complex [**44**]. Upon binding to a target, these regions may undergo disorder-to-order transitions, adopting an energetically favorable conformation and a well folded state in the protein complex. This property is particularly important in cellular signaling and transcription pathways and regulation of PPI networks, that require the ability to mediate very specific and highly dynamic interactions.

In addition to their role as binding regions, these disordered segments are also preferential sites for post-translational modifications (PTMs), as they are easily accessible for modifying enzymes. PTMs, that include phosphorylation, glycosylation, ubiquitination, and methylation, are often reversible and their occurrence is triggered by specific cellular conditions, which makes them ideal mediators for signal transmission and dynamical modulation of molecular functions [**43**].

Besides partner recognition and regulation, disordered proteins contribute to structural organization in the cell: for instance, the low-complexity of IDPs sequences or large IDRs promote weak multivalent interactions that cause a liquid-liquid phase separation (LLPS) in the cell. This leads to biomolecular condensation and the formation of membraneless compartments, or organelles, which are found both in the cytosol and in the nucleus. These organelles, that comprise stress granules, germ granules, and P bodies [**45**], help compartmentalize and regulate biochemical reactions without the need of physical membranes, and potentially inhibit the activity of other molecules by sequestering them [**46**–**48**]. The formation of such condensates is modulated in the cells by altering the concentration of phase-separated molecules, or modifying the phase separation threshold through changes in the degree of molecular assembly or intrinsic solubility. The dynamic regulation of phase separation enables the rapid adaptation of cellular responses to stress or signaling cues [**46**].

1.3.1 Disorder, binding interfaces and mutational effects

The conformational variability of disordered regions has profound consequences for how mutations affect protein function, particularly when these regions overlap with binding interfaces. From an evolutionary perspective, disordered regions are more tolerant of substitutions at the sequence level than structured domains, but conserved motifs and physicochemical patterns within disorder highlight functional constraints. Since SLiMs are generally short, single point mutations can readily disrupt binding affinity, specificity, or the capacity to undergo disorder-to-order transitions. They can also perturb PTM sites, modify transient contacts, and inhibit partner recognition.

On the other hand, mutations can also create novel motifs or increase the binding propensity of existing ones, leading to excessively stable interactions that drive disease. Thus, the lack of structural constraints in IDRs makes them evolutionarily adaptable, but also vulnerable to GoF mutational effects.

Similarly, mutations in IDPs can alter the multivalent interactions that drive LLPS, causing the formation of stable, irreversible aggregates that cannot perform their disorder-mediated function. Some of the most intensively studied IDPs are implicated in neurodegenerative diseases and cancer biology. α-synuclein, an abundant neuronal protein, is intrinsically disordered under native conditions and plays roles in synaptic vesicle regulation. Mutations such as A30P, E46K, and A53T increase the protein's aggregation propensity and cause the formation of amyloid fibrils, which are characteristic of Parkinson's disease [**49**]. In this case, the loss of conformational flexibility and the formation of stable aggregates illustrate how mutations can convert functional disorder into a pathological structured state. Similarly, the IDP protein tau, which stabilizes microtubules in neurons cytoskeleton, is prone to hyperphosphorylation and aggregation into neurofibrillary tangles, which are pathological signatures of Alzheimer's disease and are also implicated in other conditions like Parkinson's disease and dementia [**50**].

An additional link between disorder and binding regions is provided by the concept of *soft disorder*, introduced in [**51**]. Soft disorder refers to residues that adopt variable conformations across different crystal structures of the same or highly similar proteins, representing an intermediate state between intrinsic disorder and structured flexibility. In globular proteins, soft-disordered regions (SDRs) are associated with possible binding interfaces, and modification of these regions upon binding are an indicator of assembly paths [**52**]. Soft disorder will be a central concept in Chapter 5 of the manuscript.

1.3.2 Databases

The study of IDRs and IDPs has been strongly supported by the development of dedicated databases that provide experimentally verified annotations and predictions from computational mdoels. Below we report a brief overview of these resources.

DisProt

DisProt is the central reference for experimentally validated intrinsically disordered regions and proteins [**53**]. It contains manually curated annotations from the literature, with each entry providing experimental evidence supporting disorder. DisProt adopts different ontologies to categorize and describe structural and functional aspects of an IDP/IDR, and the technique or evidence associated with an annotation. Currently, DisProt contains more than 13 thousands disordered regions from 3,201 proteins across many taxonomic groups, making it the most authoritative source for experimental disorder annotations and a benchmark for developing computational predictors. An example is the Critical Assessment of Protein Disorder (CAID) challenge, a competition that periodically evaluates state-of-the-art disorder predictors on unseen and newly introduced entries in the database [**54**].

MobiDB

MobiDB is a comprehensive database that integrates disorder annotations from multiple sources, combining curated data, predictions, and indirect evidence from homologous sequences search [**55**]. it provides manually curated annotations and homology transfer results for 521 thousands proteins, which include functional regions, post-translational modifications, binding partners and biological functions, and indirect evidence, derived from PDB and AlphaFold predictions, for additional 1.4 million proteins. MobiDB also incorporates state-of-the-art disorder prediction algorithms and interactive visualization for NMR conformational ensembles and crystallographic structures.

Other resources

In addition to DisProt and MobiDB, several smaller but complementary databases provide specialized perspectives on protein disorder. IDEAL (Intrinsically Disordered proteins with Extensive Annotations and Literature) curates functionally characterized IDPs, emphasizing the mechanistic roles of disorder in signaling, binding, and post-translational regulation, as well as transitions between order and disorder [**56**]. FuzDB and MFIB focus instead on disordered protein interactions, the former collecting examples of *fuzzy interactions* where disorder persists upon binding, and the latter cataloging disorder-to-order transitions induced by binding [**57, 58**]. Finally, the Protein Ensemble Database (PED) contains experimentally determined structural ensembles of IDPs, providing representations of conformational heterogeneity [**59**].

Chapter 2

Computational background on protein modeling and variant effects prediction

According to the Anfinsen's principle, the function of a protein is determined by its three-dimensional structure, which in turn is encoded in the sequence [**60**]. The large applicability of this principle has led to the development of mathematical models and computational methods to infer the properties of proteins from their sequences, leading to the field of computational biology. Standard methods for protein sequence analysis include multiple sequence alignments (MSA), evolutionary conservation, and coevolution, which are used to infer the structure and function of proteins.

However, the improvement in experimental techniques for protein sequencing and structure determination paved the way for the advent of deep learning methods, which have revolutionized the field of protein modeling, by leveraging the increasingly large amount of available data.

This chapter introduces the concept of neural networks, which are the building blocks of deep learning, and their application to protein modeling, from energy-based models to multimodal protein language models. Lastly, I discuss some computational methods for predicting the effects of mutations on protein function and I list the metrics used for evaluation of models' performances on the different tasks encountered in this work.

2.1 Deep learning and sequence modeling

The term "artificial intelligence" (AI) was first introduced in 1956 by John McCarthy and his colleagues, who defined it as the study and development of machines capable of performing tasks that typically require human intelligence, such as learning, adapting to new situations, and improving their performance through experience [**61**]. In broader terms, AI comprises a collection of theories, methodologies, and techniques that enable computational systems to imitate and, in some cases, extend human cognitive functions such as perception, reasoning, decision-making and problem solving.

Among the subfields of AI, machine learning (ML) focuses on designing algorithms that automatically learn patterns from data, exploiting statistical regularities to produce predictions or decisions. This emphasis on data-driven prediction, as opposed to explicit modelisation of underlying mechanisms, allows ML methods to generalize to previously unseen conditions.

Based on the learning task, ML algorithms are generally divided into three broad categories:

Supervised learning. In supervised learning, the training data are annotated with labels that identify particular properties or classes. The goal is to learn a mapping function between inputs and outputs such that the model can correctly infer the labels of unseen input examples. Applications include classification, regression, and structured prediction tasks. Essential to this learning paradigm is the access to high-quality labeled data, that usually require human annotation or supervision. Especially in the biology field, this usually translates to costly and time-consuming experiments, increasing the importance of publicly available databases and community resources.

Unsupervised learning. In contrast to supervised learning, the data in this setting are unlabeled. The objective is to uncover hidden structures, groupings, or patterns in the data, without the guidance of explicit target labels. Typical applications include clustering, anomaly detection, dimensionality reduction, and pattern discovery.

More recently, *self-supervised learning* emerged as a subcategory of unsupervised learning, in which the model learns informative representations from unstructured and unlabeled data by identifying supervisory signals. Instead of relying on manually annotated labels, self-supervised approaches generate pseudo-labels from the data itself, typically by defining tasks that require the model to predict hidden or masked parts of the input. Solving these tasks drives the model to extract meaningful features which can later be transferred to downstream applications. Self-supervised methods have achieved unprecedented success in natural language processing (NLP), biological sequence analysis, and computer vision, and have enabled models to scale as more and more raw data are available. For instance, modern generative models for text are trained on a huge fraction of all available web pages in Internet, and their complexity increased to hundreds of billions of parameters.

Reinforcement learning In reinforcement learning (RL), models usually learn by interacting with an environment and receiving feedback in the form of rewards. Instead of requiring labeled datasets, RL employs a trial-and-error strategy where an agent chooses actions, observes their consequences, and adjusts its policy to maximize the cumulative reward. RL has been particularly successful in applications such as game playing and robotics, but also in the adaptation of large language models for conversational tasks.

2.1.1 From perceptrons to deep networks

Neural networks are neural-inspired nonlinear models, employed in different ML tasks. The basic computational unit of a neural network is the *perceptron*, that mimics the working principles of neuron cells [**62**]. Mathematically, it consists of a linear combination of its inputs followed by a non-linear activation function. More specifically, given an input vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the perceptron computes a weighted sum $\sum_{i=1}^{n} w_i x_i + b$ and applies a decision function, thereby allowing the model to learn and represent linear decision boundaries.

The weights w_i and bias b are iteratively updated through a learning rule that minimizes the error between the predicted and true labels. Consequently, a single perceptron can

solve binary classification tasks only when the data are linearly separable. In order to approximate more complex, non-linear functions, multiple perceptrons can be combined into layers, which are then stacked together in a hierarchical manner to form a *multi-layer perceptron* (MLP), a fully-connected feedforward neural network. In an MLP, the first layer receives the raw input and computes a set of intermediate features, which are propagated through one or more hidden layers. The parameters of all layers are learned jointly using the backpropagation algorithm [**63**], which efficiently computes the gradients of the objective loss function with respect to all weights. In this way, each layer extracts progressively higher-level features from the data, and the network learns a complicated non-linear transformation from the inputs $\mathbf{x}$ to the final output.

The development and design of neural network architectures, that exploit large datasets and their inductive biases to extract complex representations from data, is the subject of deep learning (DL). In this branch of ML, the paradigm shifts from manual feature engineering to define inputs that are relevant for the specific task, to automatic extraction of abstract patterns from large sets of raw data, that can be adapted to different tasks [**64**].

The success of deep learning, that drove the substantial revolution in the AI field of the last decade, stems from multiple factors, including improvements in hardware adapted for faster matrix operations (GPUs, TPUs), availability of larger datasets, and development of algorithms for stable training and optimization of deep networks.

Symmetries and neural networks

Many ML tasks (especially in the context of computer vision and image processing) also possess additional structure, such as locality and translation invariance. Neural network architectures that reflect the symmetries encoded in the data are more likely to produce a better hidden representation of the inputs, and thus deliver better performances in ML tasks. The relation between symmetries and networks is at the foundation of geometric deep learning [**65, 66**]. In the following, we report some standard network architectures developed for sequence modeling, with a focus on the underlying symmetries. An overview is also shown in Table 2.1.

Architecture	Domain	Symmetry
CNN	Grid	Translation
LSTM	1D Grid	Time warping
Transformer	Complete graph	Permutation

Table 2.1: Examples of common neural network architectures, domains they are applied to and the symmetry that they satisfy.

Convolutional Neural Networks

Convolutional neural networks (CNNs) constitute a class of neural network architectures specifically designed to exploit the temporal or spatial structure of data (in one, two or higher number of dimensions) through the application of learned convolutional operators.

Unlike fully-connected MLPs, where each neuron is connected to all inputs, CNNs apply a sparse local connectivity to extract local patterns in the input. The extent of this connectivity, called receptive field of the neurons, is limited in the spatial or temporal dimensions of the grid, determined by the size of the applied filters, or kernels.

The convolution operation allows the same filter to be applied across all spatial or temporal locations on the input grid. This weight-sharing property across distant locations inherently endows convolutional neural networks with translation equivariance: the feature map produced by the network depends only on the local convolutional operation, not its spatial location. Such inductive bias reduces the number of learnable parameters and makes CNNs particularly effective in settings where the location of a pattern is less important than its presence.

The first CNNs trained automatically with gradient descent algorithm were introduced in [**67**] for handwritten images recognition, and modern variants have become a fundamental building block for computer vision tasks such as object detection, semantic segmentation, and image classification [**68, 69**]. Deep CNNs, comprising multiple convolutional and pooling layers, are able to learn hierarchical representations, where early layers capture simple, low-level features, such as edges or local gradients, and deeper layers progressively build more abstract representations that are relevant for the final task.

While two-dimensional convolutions are typically used for image processing, one-dimensional convolutions are particularly well suited for modeling sequential data. Because of their ability to detect position-independent motifs, 1D CNNs have been successfully applied to biological sequence modeling, such as predicting protein binding sites and identifying sequence motifs in DNA [**70**]. Similarly, in protein sequence analysis, one-dimensional convolutions are able to identify patterns of amino acids that correspond to functional or structural motifs [**71**], providing an efficient alternative to recurrent neural networks for capturing short- and medium-range dependencies. Furthermore, the adoption of dilated convolutions can effectively extend the effective receptive field of 1D CNNs, enabling them to capture long-range interactions over sequences without increasing the number of parameters and allowing to produce representations that span entire protein sequences [**72**]. In this work, convolutional layers were employed in the light attention component of MuLAN model (Chapter 3),

Recurrent Neural Networks

Recurrent neural networks (RNNs) are class of deep learning architectures that model sequential and temporal data by introducing an internal summary vector state. This component is updated after each time step, enabling the model to exploit contextual information in the sequence. RNNs can generalize across temporal positions, by applying the same parameters at each time step. These characteristics have made RNNs particularly effective for tasks such as speech recognition, language modeling and machine translation [**73, 74**]. In computational biology, RNNs have been successfully applied to model complex dependencies across nucleotides or amino-acid sequences, for example in the prediction of transcription factor binding sites [**75**].

Due to the recurrent application of a single RNN cell to sequential inputs, the depth of

the RNN is effectively equal to the number of steps in the input, leading to instabilities during the training phase. The tendency to suffer from vanishing or exploding gradients when trained on long sequences and the exponentially smaller weights given to long-term interactions is a major limitation of RNNs [**64**]. To overcome this problem, more advanced RNN architectures were developed, where key gating mechanisms and explicit memory components were introduced. A successful implementation is the long short-term memory (LSTM) network [**76**]: by incorporating gating mechanisms, including input, forget, and output gates, the LSTM can regulate the flow of information through time steps and maintain a memory cell that is preserved between computational steps, thus being more robust in capturing long time dependencies. In [**77**], the authors demonstrate that the design of the LSTM, incorporating gating mechanisms, endows the model with a time-warping invariance. Given the time warping definition as any monotonically increasing and differentiable mapping between times (like a dilation operation), such invariance is defined as the capacity of two models of the same class of processing warped data and non-warped data in the same way. In practice, this property ensures that gated RNN models can effectively capture long-term dependencies. Besides LSTMs, another gated RNN architecture is the gated recurrent unit (GRU) [**78**], which simplifies the gating mechanism by combining the input and forget gates into a single update gate. Despite this difference, GRU often yields similar performances compared to LSTMs, while requiring fewer parameters [**79**].

2.1.2 Attention mechanism and Transformers

The sequence modeling field experienced a revolution when the transformer architecture was introduced, in 2017 [**80**], thanks to the ability of the new deep learning architecture to process information in parallel, to capture long-distance dependencies in sequential data and to be scaled and deployed on multiple devices, allowing model sizes to reach the scale of billions of parameters. These characteristics make transformers particularly suited for the development of foundation models, large-scale models trained in a self-supervised manner on vast and diverse datasets. The objective of these models is to learn general-purpose representations that capture complex statistical patterns in the data in a unified features space, and which can then be adapted to a wide variety of downstream problems [**81**].

In order to be processed by the transformer architecture, sequential data have to first be tokenized, i.e. converted into small discretized units, called tokens, that can then be mapped into high dimensional vectors through an embedding layer. Tokenization strategies are not unique, and a balance between the number of possible different tokens, the proper granularity required for the specific task, and the resulting length of the tokenized input has to be achieved to ensure model efficiency and applicability [**82, 83**]. Indeed, transformers space complexity scales quadratically with the input length, and model generalization beyond the length of training sequences (context length) requires specific architecture choices and is still an area of active research, making it the decisive bottleneck in many applications [**84, 85**].

Encoder and decoder

The original transformer architecture, represented in Figure 2.1, consists of two main components, an encoder and a decoder. Both components are a stack of multiple identical layers, that include multi-head attention, feed-forward networks, layer normalization and residual connections [**80**]. Encoder and decoder serve different purposes:

The **encoder** processes an input sequence and converts each token into a continuous, context-aware representation. Through its self-attention mechanism, the encoder can assign different weights to tokens within the sequence, enabling it to capture dependencies and relationships even across distant positions.

The **decoder** generates the output sequence, as in translation or text generation tasks. It attends both to the encoder's output representations and to the tokens it has already produced, using a combination of cross-attention and masked self-attention. The decoder operates in an auto-regressive manner, producing the output one token at a time.

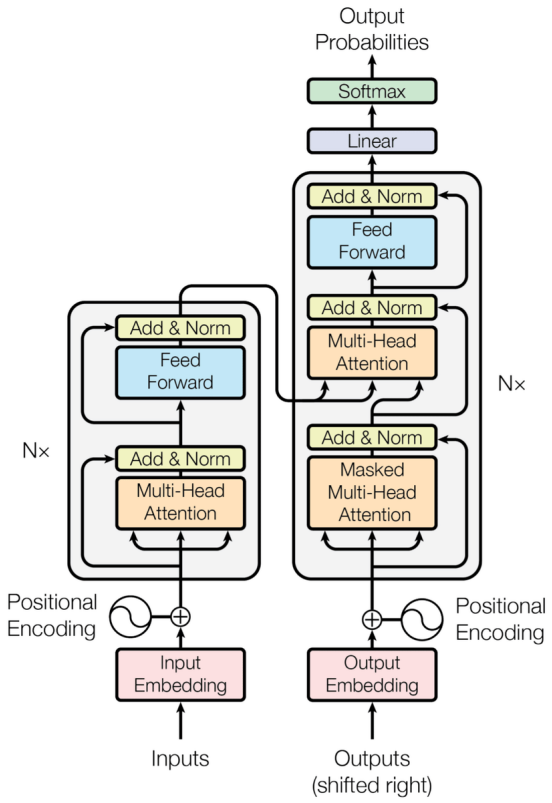

Figure 2.1: Original transformer architecture with encoder and decoder components, as introduced in [**80**], from which the image is taken.

Since their introduction, transformer architectures have been adapted to suit different task requirements, giving rise to three primary variants:

Encoder-only: Designed for tasks that involve understanding and analyzing input sequences through their contextualizaed encoding into vector representations. Such tasks

include text classification, token classification tasks like named entity recognition (NER), and integration in retrieval augmented generation (RAG) pipelines. Examples in NLP field include BERT [**86**], RoBERTa [**87**], and DistilBERT [**88**].

Decoder-only: Specialized for text generation, these models predict the next token in a sequence based on preceding tokens, excelling in language modeling, open-ended text generation, and dialogue systems. They are optimized to produce coherent and contextually appropriate text, and are also the central component of modern chatbots and agentic systems. The GPT family of models [**81, 89**] exemplifies this category, demonstrating strong performance across diverse benchmarks.

Encoder-decoder: Suited for sequence-to-sequence tasks where an input sequence is transformed into a distinct output sequence. Typical applications include machine translation, summarization, question answering and multimodal integration. In this setup, the encoder encodes the input into vector representations, and the decoder sequentially generates the output by attending to both the encoder's representations and its own previous outputs. Examples include T5 [**90**], which frames diverse problems in a unified text-to-text paradigm, and BART [**91**].

Attention

The core component of transformer architecture is the attention mechanism, that allows the model to process embeddings for single tokens simultaneously and weigh their importance with respect to each other, effectively capturing dependencies at arbitrary lengths. It is also the only component of the transformer that enables communication between different tokens, whereas all the other parameters and layer act on the single token representations in parallel [**92**].

In the attention layer, the (normalized) input embedding X is linearly projected into queries $Q = XW_q$, keys $K = XW_k$ and values $V = XW_v$ matrices, where W_q, W_k and $W_v \in \mathbb{R}^{d\times d}$ are model's learnable weights, and d is the embedding dimension. The attention output is then computed as

$$\mathrm{Att}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{2.1}$$

where the softmax function normalizes the rescaled attention scores between query and key vectors, such that each token representation is updated with a weighted sum of value vectors.

While this expression describes the operation of a single attention head, already in its early implementations transformers combined the output of multiple attention heads in parallel, each of them with their own learnable parameters. The introduction of *multi-head attention* allows each head to focus on specific patterns in the context, without increasing the model complexity.

In the bidirectional self-attention algorithm, typical in encoder architectures, attention is applied to each token of X, treating all tokens as context. In unidirectional self-attention,

instead, a triangular mask is applied such that each token can only attend to previous tokens (including itself) as context. Finally, cross-attention acts on two sequences of token representations X and Z, where attention is applied on X (projected to query matrix) treating Z as the context (key and value matrices). In the latter case, Z can have a different length with respect to X.

From its expression, it is clear that the attention operation is invariant under permutation of the tokens representations. Therefore, to properly model the sequential nature of the inputs, an encoding of positional information, either absolute or relative, is required. This is achieved through the introduction of positional embedding, that learns to represent a token's position in the sequence and is usually added to the initial token embedding. However, learned positional embedding limit the length of the inputs sequences to a fixed context. To overcome this limitation, hard-coded positional mapping are often used, as in the original transformer paper or in the more recent Rotary Positional Embedding (RoPE) implementation [**93**], whose modern variants are an area of active research [**94**–**96**].

Parameter-efficient fine-tuning and low rank adaptation

The adaptation of pre-trained models to domain-specific tasks poses significant challenges, as standard fine-tuning of a billion-parameter model requires considerable computational resources, appropriate hardware, and large annotated datasets. As major downside, it may also lead to catastrophic forgetting, meaning that the model loses part of the general-purpose knowledge acquired during pre-training, limiting its generalization capacity. To overcome these limitations, many parameter-efficient fine-tuning (PEFT) strategies have emerged, which rely on the fine-tuning of a small number of (additional) parameters, significantly decreasing computational and storage costs, while yielding performance comparable to fully fine-tuned models [**97**]. PEFT methods can be grouped in two main classes: prompt tuning and model adapters.

Prompt tuning, especially useful in generative textual or multimodal models, refers to the prepending of the input sequence with additional text that describes the task or even demonstrates one or more examples of the task (hard prompts) or the concatenation of input embeddings with learnable tensors that can be optimized to the specific dataset (soft prompts). Hard prompts do not require additional parameters to learn, but are typically restricted to the text modality and require effort to design effective prompts. Soft prompts have instead a broader applicability, at the cost of interpretability, as the learned embeddings, optimized in a continuous space, are not restricted to the discrete tokens appearing in the model's vocabulary.

Adapters include all those methods that introduce additional learnable modules in the model's architecture to actively adapt the internal representations of the model to the specific task. Such modules are typically added after the attention or fully-connected layers of the pre-trained transformer (serial adapters) or in parallel to frozen weight matrices W to enable lightweight updates ΔW.

Among tens of adapter-based methods, low-rank adaptation (LoRA) gained popularity for its efficiency and effectiveness [**98**]. LoRA represents the weight updates $\Delta W \in \mathbb{R}^{d \times k}$ with

two update matrices through low-rank decomposition, i.e. $\Delta W = BA$, with $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$. $r \ll \min(d, k)$ is the rank of the decomposition.

In [**98**], the authors demonstrate that performances of LoRA applied to different text models and tasks are comparable or even exceeding standard-fine tuning, while requiring the training of only around 0.1% of the original number of parameters. The weight updates ΔW are shown to amplify task-specific features that were already present but not strongly emphasized in the pre-trained model.

2.2 Learning from multiple sequence alignments

The introduction of transformer architecture allowed to extend the standard conservation and coevolution analysis that arises from MSAs to a data-driven approach that generalizes beyond single protein families. In this section, we briefly describe statistical and machine learning approaches that were introduced to learn patterns in protein sequences from MSAs, from energy-based methods and latent-variable models to attention-based architectures, that were instrumental in the development of AlphaFold2 model of protein folding. All these diverse approaches aim to model sequence landscapes, from which mutational effects can be inferred.

2.2.1 Energy-based and latent variable models

Energy-based models (EBMs) provide a framework to model protein sequences where the evolutionary constraints encoded in MSAs are expressed as an energy landscape over amino acid configurations. Among EBMs, Direct Coupling Analysis (DCA) has emerged as one of the most influential approaches to uncover the constraints that determine protein structure [**99**]. DCA models the joint probability distribution of a sequence of amino acids $x = (x_1, x_2, \ldots, x_L)$, where x_i denotes the amino acid at position i, using a Potts model with $q = 21$ possible states (20 amino acids plus a gap state). The probability of a sequence is then given by

$$P(x|h, J) = \frac{1}{Z} \exp\left(\sum_{i=1}^{L} h_i(x_i) + \sum_{i<j} J_{ij}(x_i, x_j)\right), \tag{2.2}$$

where $h_i(x_i)$ are position-specific bias terms (also called local fields) capturing conservation, $J_{ij}(x_i, x_j)$ are coupling parameters modeling the direct statistical dependency between residues i and j, and Z is the partition function ensuring normalization, corresponding to the sum of the exponential term over all possible configurations. This formulation assumes that the evolutionary statistics of an MSA can be explained by a maximum entropy model constrained by the observed single-site and pairwise frequencies.

Parameter inference in DCA is challenging due to the intractability of Z for realistic sequence lengths. Several approximations have been developed: mean-field approximations based on the inverse of covariance matrices [**99**], message passing algorithms [**100**], and more robust pseudo-likelihood maximization approaches that maximize the conditional

probability of each site given all others [**101**]. Additionally, parameters inference can also be achieved by training a Boltzmann machine on the multiple sequence alignment (MSA) through pseudo-likelihood maximization, which can then be used for generation of synthetic sequences [**102**]. Expanding on this, restricted Boltzmann Machines (RBMs) represent a variant that introduces hidden variables able to learn higher-order dependencies, potentially extending the expressive power of traditional DCA approaches.

The Frobenius norm $F_{ij} = \sqrt{\sum_{x,y} J_{ij}(x,y)}$ of the learned couplings J_{ij}, with a correction term that accounts for mutual information, can be used to predict intra-protein contact maps, as highest values correlate strongly with the residue contacts in protein structures of the considered family. Besides structural inference, DCA models can be used to quantify the effect of mutations on protein fitness. The negative log-probability of a sequence defines a statistical energy term, $\phi(x) = -\log(P(x))$, which approximates the evolutionary fitness of the sequence. The effect of a mutation from a sequence x to x' can be quantified as the difference in statistical energy

$$\Delta\phi = \phi(x') - \phi(x), \tag{2.3}$$

where a negative $\Delta\phi$ indicates that the mutant sequence has lower energy (thus higher probability) and is therefore more compatible with the evolutionary constraints, while a positive $\Delta\phi$ denotes deleterious effects. Since the Potts model includes pairwise couplings, the predicted impact of a mutation depends not only on the local conservation at that site but also on couplings with amino acids present at other positions, naturally capturing epistatic effects [**103**].

A complementary approach to learn the distribution of protein sequences from MSAs are latent variable models, such as Variational Autoencoders (VAEs) [**30**]. VAEs approximate the full high-dimensional distribution of sequences by learning a low-dimensional latent representation $z \in \mathbb{R}^d$ that captures the evolutionary constraints. A VAE consists of two neural networks: an encoder $q_\phi(z|x)$ that maps sequences x to a posterior distribution over latent variables, and a decoder $p_\theta(x|z)$ that reconstructs sequences from z. Training optimizes the evidence lower bound (ELBO),

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x|z)\right] - D_{KL}\Big(q_\phi(z|x) \,||\, p(z)\Big), \tag{2.4}$$

where $p(z)$ is typically taken as a standard Gaussian prior. The first term enforces accurate sequence reconstruction, while the Kullback-Leibler divergence D_{KL} regularizes the latent space, by keeping the posterior probability $q_\phi(z|x)$ close to the prior distribution.

The Evolutionary model of Variants Effect (EVE) [**104**] is a successful application of (Bayesian) VAEs for mutational effect prediction. EVE framework is trained separately on MSAs for many different proteins, learning a family-specific generative model of protein sequences. Mutational effects are computed by sampling from the approximate posterior distribution $q_\phi(z|x)$ learned by the VAE, which provides an estimation of the relative likelihood of amino acid variants with respect to the wild type. Such scores correlate with functional and clinical impact of mutations: when benchmarked against deep mutational

scanning data and ClinVar pathogenicity annotations, EVE was able to achieve state-of-the-art performance, exceeding supervised predictors. As a self-supervised method, EVE is less biased by the incomplete or noisy annotations that are typical in the experimental settings.

2.2.2 Multiple attention mechanisms, AlphaFold2 and AlphaMissense

A major limitation of statistical approaches such as DCA and family-specific generative models based on VAEs is that their capacity to capture evolutionary constraints is determined by the quality and depth of the MSAs used for parameter inference. For proteins with few homologs or with highly variable regions that are poorly covered in the alignment, reliable modeling of the fitness landscape becomes unfeasible. Moreover, these approaches are typically trained on each individual MSA to infer family-specific couplings, which makes them computationally demanding for large-scale variant assessment across diverse protein families.

The introduction of attention-based architectures specifically designed for protein MSAs has significantly advanced the extraction of evolutionary information, addressing some of these limitations. MSA Transformer [**105**] introduced a dual attention mechanism tailored to the matrix representations of MSAs. In this architecture, tied row attention operates along sequence positions within each protein, capturing residues long-range dependencies and conservation patterns. Complementarily, column attention attends to the aligned positions in different sequences, learning to select informative sequences from the alignment. Trained on 26 millions MSAs, MSA Transformer allowed contact maps inference with unprecedented precision and generalization to previously unseen protein families, providing an alignment-aware and scalable foundation model for sequence representation learning.

Similar principles underpin the design of the Evoformer module of AlphaFold2 [**35**], which processes MSA-derived features combined with pairwise residue representations, to provide informative embeddings for an end-to-end structure prediction pipeline. Each Evoformer block combines MSA attention layers, which refine evolutionary signals by iteratively applying row and column attention, with updates to the pair representation that capture residue-residue geometry. In particular, triangular multiplicative updates and triangular self-attention modules are used to propagate geometric constraints across residue triplets, enforcing consistency with three-dimensional structures. By jointly updating MSA and pair representations along the Evoformer blocks, AlphaFold2 progressively builds detailed representations that encode both evolutionary constraints and spatial relationships. The final Structure module then decodes these representations into atomic coordinates, by predicting the relative positions and orientations between each residue frame, achieving experimental-level accuracy in protein structure prediction. This synergy between attention-based MSA processing and equivariant geometric reasoning is a decisive architectural element that enabled AlphaFold2 to set a new paradigm in protein structure predictions.

AlphaMissense [**106**] demonstrated the adaptability of AlphaFold2 architecture for variant effects predictions, by leveraging its structural reasoning-oriented design. This is achieved by

combining two training strategy. First, the AlphaFold architecture is fine-tuned to predict the effects of sampled variants leveraging masked language modeling, while also keeping the structure prediction objective with reduced weight in the combined loss function. In a second step, logits-derived pathogenicity scores are further fine-tuned on human proteins for a binary classification of variants as likely benign or pathogenic, where labels are assigned based on the frequency of missense variants observed in human or primate populations. As unobserved variants (labeled as pathogenic in the training set) might in fact be benign, the authors leveraged a self-distillation procedure from intermediate checkpoints to filter out likely mislabeled variants. For large scale predictions, results are then calibrated in the [0,1] range based on a reduced set of variants from ClinVar. As in AlphaFold2, the model leverages multi-head attention across both rows and columns of MSAs to capture evolutionary conservation and coevolution signals, while also incorporating structural information inherited from AlphaFold2 pre-training. By combining architecture-level innovations from AlphaFold2 with supervision from clinical and population variant datasets, AlphaMissense achieves state-of-the-art performance in pathogenicity prediction, outperforming statistical baseline models and prior machine learning approaches. These diverse applications demonstrate the generality of attention-based MSA modeling and its utility not only for structural biology but also for clinical genomics. However, it must be noted that models performances are highly dependent on the quality and completeness of MSAs, limiting their applicability to novel or orphan proteins.

2.3 Learning from single sequences: Protein Language Models

The availability of several millions of metagenomic protein sequences allowed to apply the methodologies of NLP to protein sequences. The resulting models, mostly relying on transformer architectures, are called Protein Language Models (PLMs). This modelisation approach, based solely on single sequences, allows to overcome the limitations of MSA-based models, while also enabling faster predictions due to reduced preprocessing steps and larger inputs sizes in terms of sequence length, as MSA depth is no longer a constraint. It must be noted, however, that enriching representations with homology information, when available, improves predictions in downstream tasks. In addition, lack of homologous sequences indirectly affects also single-sequence PLMs, since correlations for these sequences are more difficult to capture, as well as generalization to novel sequences.

As in NLP, tokenization strategies have an important impact on model architecture and context length: for proteins, single amino acid tokenization is a natural choice, which allows a direct translation between model's embeddings and amino acids properties. In this way, a context length of around one thousand tokens is enough to represent the vast majority of single protein sequences, while limitations still remain when extension to protein complexes or multiple sequences is required.

PLMs architectures resemble (and in many cases, directly adopt) those of NLP, with the same distinction in *encoder-only*, *decoder-only* and *encoder-decoder*. Here we provide some examples of early PLMs implementations and their related tasks:

– Encoder-only models, such as ProtBERT [**107**] and ESM [**108**], are designed to learn contextual representations of protein sequences. These models learn the relationships between amino acids and produce high-quality embeddings, which can subsequently be used for a wide range of tasks, such as protein folding, token classification (secondary structure, disorder) or sequence regression (solubility, fluorescence, remote homology) [**109**].

When the predicted property refers to the whole protein, rather than single amino acids, a common strategy is to pool the protein representation by averaging over the length dimension, in order to obtained vector representations of fixed size, to use as input features for downstream tasks. Some works, however, show that learning the pooling strategy might be beneficial for model performances [**110, 111**].

In addition, it has been shown that attention layers learn coevolution signals between amino acids, which can be used to predict contact maps and serve as inputs for folding models, like ESMFold [**112**].

– Decoder-only models are optimized for autoregressive generation. For example, ProtGPT2 [**113**] and ProGen [**114**] are designed to model the probability distribution over amino acid sequences and to generate novel proteins one residue at a time. These architectures are particularly suited for protein design applications, where the objective is to produce novel sequences that satisfy functional or structural constraints.

– Encoder-decoder models are exemplified by ProtT5 [**107**] and Ankh [**109**], that generate informative sequence embeddings and are equipped with a bidirectional-autoregressive decoder. In practice, in most applications, only the encoder is kept for embeddings extraction, while the decoder requires additional tuning to perform sequence generation and transformation tasks [**109**].

All the models appointed previously rely only on the sequence modality, learning protein representations from patterns in amino acid sequences only. A natural extension of this approach, again derived from NLP field, is the introduction of additional data types to enhance representations or better align them to specific tasks.

2.3.1 Multimodality

Beyond single sequences inputs, multimodality has recently become a key feature in foundation models: in the NLP field, text is combined with image, video or audio data, allowing tasks such as visual question answering, video captioning, text-to-image or text-to-video generation, speech translation. In bioinformatics, multimodal models provide more complete representations of proteins, by combining information from sequence, structure, physical properties, functional annotations. Modality fusion is usually achieved according to three possible strategies: early fusion, mid-level fusion of late fusion. In the following, we provide some examples of recent multimodal PLMs, divided by fusion strategy.

Early fusion. In early fusion, embeddings from different modalities are usually concatenated or projected into a shared input space before a common encoder. A prominent representative of this class of models is indeed ESM3 [**115**], that sums embedded representations of sequences, structure, solvent accessible surface area (SASA), secondary

structure information and InterPro annotations before giving them as input to a unique transformer encoder. Different heads attend to the same processed representation, allowing decoupled decoding for each modality. This comes at the cost of increased complexity, with models ranging from 1.4 to 99 billions parameters. SaProt model [**116**] is also an example of early fusion: in this case, however, structure information is directly coupled with sequence tokenization, by defining a structure-aware tokenization strategy that uses Foldseek vocabulary [**117**]. In ProstT5 [**118**], instead, the original ProtT5 model is fine-tuned to enable bidirectional translation between amino acid tokens and the 3Di tokens, that encode local structure, learning and hidden representation that captures both modalities. A more recent example for protein encoding and scoring of mutational effects is AIDO.Protein-RAG [**119, 120**], that fine-tuned a mixture-of-experts encoder model trained on single sequences (AIDO.Protein-16B) by adding sequences retrieved from a pre-trained protein IDs generator (MSA Retriever) and structural tokens extracted from a pre-trained structure encoder (AIDO.StructureTokenizer). The fine-tuning comprised multiple rounds and endowed the model with additional evolutionary and structural modalities, leading to state-of-the-art performances in ProteinGym DMS benchmark.

Mid-level fusion. This strategy includes models that encode different modalities separately, but combines them through cross-attention layers. Examples include ProSST [**121**], that leverages a disentangled self-attention mechanism between sequence, structure, and positional embeddings in the transformer layers, and MASSA [**122**], a framework that aligns sequence, structure and functional annotations through self-attention and multiple learning objectives.

Late fusion and ensemble This category of models combine outputs of independently trained unimodal predictors at the prediction layer. For scoring mutational effects, a recent example is VenusREM [**123**], that combines ProSST outputs with matrix representation of evolutionary logits extracted from MSAs.

Multimodal PLMs benefit from complementary signals: sequence models bring evolutionary context and broad family generalization, while structural modalities provide explicit spatial constraints and geometry-aware representations. Fusion strategies that allow cross-modal attention tend to perform best on tasks that require precise spatial reasoning (binding affinity, interface design), but they are more computationally demanding and sensitive to errors in predicted structures, like misfolded loops or incorrect docking poses. Moreover, modality imbalance between larger sequence corpora and fewer high-quality experimental structures creates training and generalization challenges, partially mitigated via data augmentation with predicted structures and contrastive objectives.

2.4 Prediction of mutational effects

Many mathematical and computational methods have been developed in the recent years to address the problem of predicting the mutational effects of amino acids substitutions and, in some cases, insertions and deletions.

Based on the inputs considered, they can rely on single sequences, multiple sequence alignments, structure, local and global physico-chemical and geometric features, or any combination of them. These models can be grouped in three main categories, depending on the modelisation strategy adopted:

- Mathematical modelisation: includes all models that apply a rule-based algorithm on the inputs, based on domain knowledge. They typically involve a limited number of parameters, that are optimized on a held out validation set. An example is GEMME and the derived models ESCOTT and PRESCOTT, that I will discuss in the following section.
- Unsupervised learning: models that learn patterns or representations from unlabeled data, such as generative models or energy-based models. The predicted probability for the presence an amino acid in a given position compared to the wild type is used as a proxy for the pathogenicity of the mutation.
- Supervised learning: models trained directly on labeled datasets to predict mutational effects, often using machine learning or deep learning techniques. Inputs can either be hand-crafted features or deep representation from pre-trained models (transfer learning).

2.4.1 GEMME, ESCOTT, PRESCOTT

GEMME (Global Epistatic Model predicting Mutational Effects) algorithm predicts single-point mutational effects by explicitly modeling interdependencies between all positions in a sequence [**124**]. The algorithm infers the evolutionary relationships relating natural sequences by quantifying their global similarities. These relationships, encoded in a tree, are used to determine the extent to which each position is conserved in evolution and to estimate the evolutionary fit required to accommodate mutations. Such measure of conservation is markedly different from measures that quantify frequencies of occurrence at single positions (columns) of an alignment. Indeed, we infer conservation levels by reconstructing phylogenetic trees from global similarities between sequences. This means that the conservation level of one position embeds the covariations between this position and all other positions in the sequence. For a position to be "conserved," we ask that the letter(s) appearing at that position is/are fully conserved in subtrees of ancient origin.

The key component of the algorithm is the T_{JET} evolutionary trace, averaged over multiple phylogenetic trees reconstructed from different MSA subsamplings. For each tree, the T_{JET} trace assigned to residue r_i at position i is the relative depth at which the residue is observed in the tree, in terms of branching nodes' cardinality. A more precise definition is provided in [**125**]. For an alignment of N effective sequences, $\sqrt{N}$ trees are evaluated over $\sqrt{N}$ sequences each. Intuitively, a pathogenic mutation has an expected T_{JET} trace close to 1, since the residue is conserved from a node close to the root, while variable positions will have scores closer to 0.

The GEMME score s_G assigned to a mutation $x_i^{wt} \rightarrow x_i^{mut}$ is the weighted sum of an independent term, that measures the conservation of an amino acid in the position i of the

alignment, and an epistatic term, that integrates the T_{JET} trace:

$$s_G(x_i^{mut}) = T_{\text{JET}}(i) \cdot \left\{ \alpha K \min \left[1, \frac{s_G^E(x_i^{mut})}{\max_{j,z} \left(s_G^E(z_j) \right)} \right] + (1-\alpha) s_G^I(x_i^{mut}) \right\} \tag{2.5}$$

where K is a normalization factor depending on the sequences in the alignment, s_G^E and s_G^I are the epistatic and independent term. The independent term is defined as the log-ratio between the frequencies of the mutated and wild type amino acid for position i. Defining S_{x_i} the set of sequences displaying amino acid x at position i,

$$s_G^I(x_i^{mut}) = -\log \left[\frac{\max(1, |S_{x_i^{mut}}|)}{|S_{x_i^{wt}}|} \right] \tag{2.6}$$

The epistatic terms is instead given by the minimum evolutionary distance D_e between the query sequence q and a sequence t displaying the mutation x_i^{mut}:

$$s_G^E(x_i^{mut}) = \min_{t \in S_{x_i^{mut}}} D_e(q,t) \tag{2.7}$$

The evolutionary distance is defined as

$$D_e(q,t) = \sum_{j=1}^{n} T_{\text{JET}}(j)^2 \cdot (1 - \delta_{x_j^q, x_j^t}) \tag{2.8}$$

where n is the length of the ungapped query sequence, and δ is the Kronecker delta.

By default, the α factor balancing the contributions of epistatic and independent term is set to 0.6. Other operations, like using a reduced representation of the amino acid alphabet that groups amino acids with similar physico-chemical properties and regularization factors to avoid divergences contribute to the final score. Importantly, since the epistatic effect is partially accounted for in the s_G^E term, the score assigned to multiple-point mutations is simply the sum of the scores of the single-point mutations.

ESCOTT builds on GEMME by including structural features, in particular circular variance (CV), in the trace computation [**126**]. CV quantifies the vectorial distribution of neighboring atoms around a fixed point in three-dimensional space and reflects the accessibility from the protein surface to its interior [**127**], making it less sensitive to minor conformational changes. For a residue i, it is defined as the average over atomic CVs for all atoms $(a_i, \ldots, a_n)$ in i, with

$$\text{CV}(a_i) = 1 - \frac{1}{|N_c(a_i)|} \left\| \sum_{a_j \in N_c(a_i)} \frac{\vec{r}_{ij}}{\|\vec{r}_{ij}\|} \right\| \tag{2.9}$$

where $N_c(a_i) = \{a_j \neq a_i \,|\, \|\vec{r}_{ij}\| < r_c\}$ is the set of atoms at distance less than r_c from atom a_i. In ESCOTT, r_c is set to 7.0 Å by default. According to the definition, buried residues will have a CV score closer to 1, due to the rather homogeneous distribution of neighboring atoms, while exposed residues have a lower CV. This was motivated by the

fact that mutation happening at the interior of the protein are more likely to be disruptive for protein stability and functionality.

ESCOTT additionally introduces a ranking of the scores as a final step, where the full predicted mutational landscape is rescaled between 0 (benign) and 1 (pathogenic).

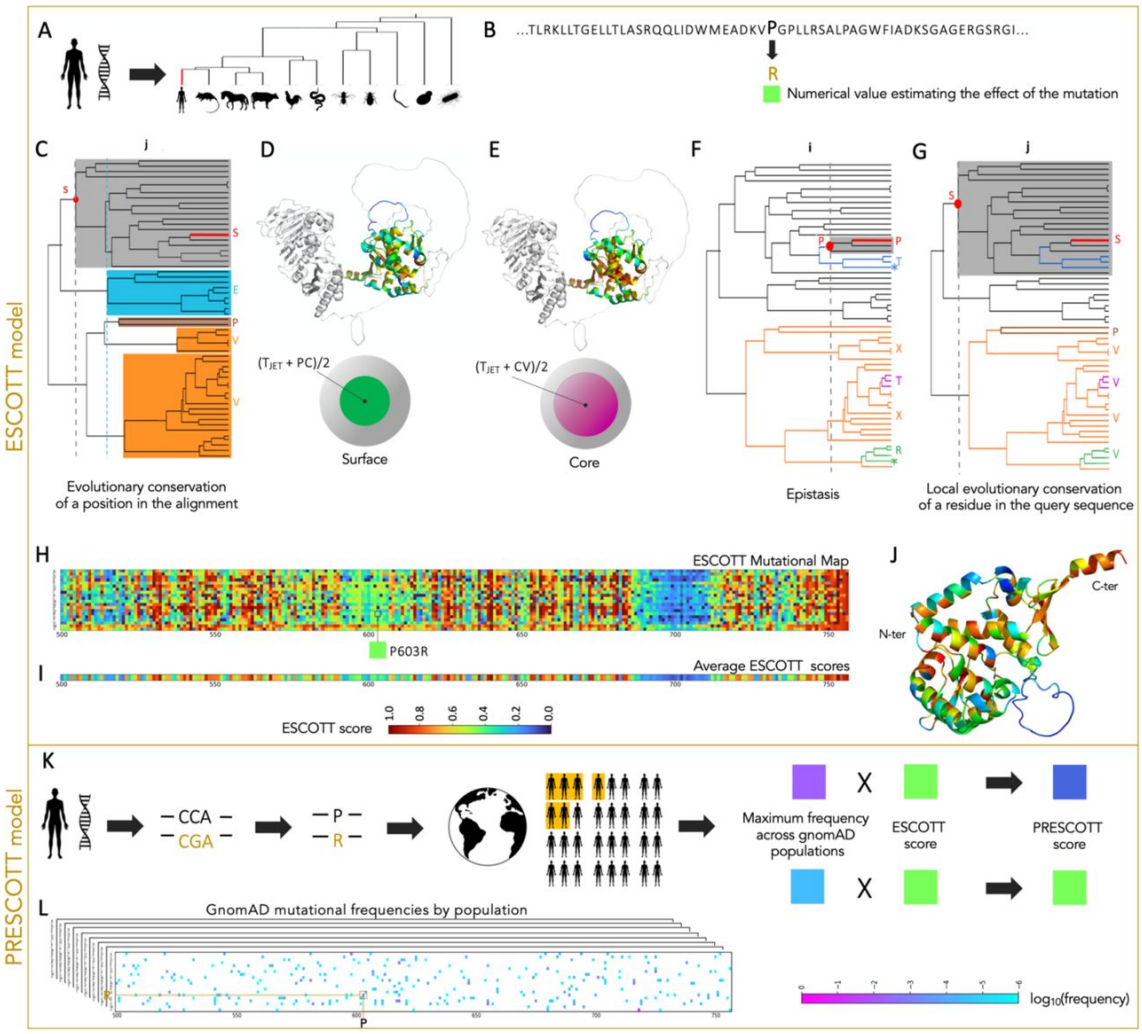

Figure 2.2: Visual representation of ESCOTT and PRESCOTT methods. The figure is taken from [**128**].

PRESCOTT adds a post-processing step for human proteins that recalibrates scores using allele frequency data from gnomAD database [**32, 129**]. High allele frequencies observed in different human genomes for a particular mutation imply that the mutation tolerance is indeed higher than predicted, therefore the corresponding score is shifted towards the benign region, with intensity depending on the relative allele frequency with respect to an identified threshold $F_c = 10^{-4}$ (for reference, mutations with frequency lower than 0.005 are considered rare). More explicitly, the PRESCOTT score s_P for a single-point mutation

is obtained as:

$$s_P = \begin{cases} s_E, & \text{if } f \leq F_c \\ \max\left[0, s_E - \left(1 - \frac{\log_{10} f}{\log_{10} F_c}\right)\right], & \text{if } f > F_c \end{cases} \tag{2.10}$$

where s_E is the pre-computed ESCOTT score with values in range [0, 1] and f is the frequency for the specific mutation extracted from gnomAD.

In this work, I relied on ESCOTT as an orthogonal high-performing method for comparison with deep learning approaches in scoring mutational effects, but I also evaluated it in different settings and contributed to the code implementation, as reported in Appendix A.

2.4.2 Zero-shot scoring

Machine learning models trained in a self-supervised fashion learn the likelihood of a protein sequence (or structure) depending on the input context and the training data, and can be used to predict mutational effect as zero-shot scores [**130**]. In previous sections, we have already described some examples of these models, from DCA-inspired generative models to attention-based protein encoders like PLMs. AlphaMissense can also be included in this category, as in its first training pipeline masked language modeling-derived logits are used to infer mutational effects, with subsequent fine-tuning and calibration steps for better alignment to clinical human data.

In practice, given a protein sequence representation $X = (x_1, x_2, \ldots, x_n)$, where x_i are (possibly multimodal) representations of single amino acids, models trained in a self-supervised framework learn probabilities $p(x_i = a \,|\, X^*)$ of observing an amino acid a at position i given the sequence context X^*. The sequence context can be either $X^* = X_{<i}$ for autoregressive models, or $X^* = X_{-i}$ for bidirectional encoder models.

The log-odds score S of observing a mutated amino acid x^{mut} instead of the wild type x^{wt} at position i is

$$S(x_i^{wt} \to x_i^{mut}) = \log \frac{p(x_i = x^{mut} \,|\, X^*)}{p(x_i = x^{wt} \,|\, X^*)} \tag{2.11}$$

For bidirectional encoder models, the above log-odds score is treated as zero-shot score for the particular mutation. The limitation is that, due to the same size required for mutated and wild-type sequences, encoder models can only account for substitutions scores.

On the contrary, auto-regressive generative models can handle sequences of variable lengths, and therefore assign scores to insertion and deletion too. For auto-regressive models, the probability of a given sequence X can be decomposed into the product of conditional probabilities

$$p(X) = p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | X_{<i}) \tag{2.12}$$

Thus, the log-odds score of observing a sequence X' that differs from X for a set of

mutations M is given by

$$S(X \to X') = \log \frac{p(X')}{p(X)} = \log \frac{\prod_{i=1}^{n'} p(x'_i|X'_{<i})}{\prod_{i=1}^{n} p(x_i|X_{<i})} = \\ = \sum_{i=1}^{n'} \log(p(x'_i|X'_{<i})) - \sum_{i=1}^{n} \log(p(x_i|X_{<i})) \tag{2.13}$$

where each term in the sum is exactly the cross-entropy loss at position i for sequence $X^{(')}$, with respect to the same sequence shifted by one position as target.

Previous studies demonstrated the ability of zero-shot scores of foundation models to capture pathogenicity of mutations for single protein sequences. Depending on the inputs and representations learned by those models, correlation with experimental data may be higher for a specific selection type, and worse for others. In ProteinGym, for example, structure-aware models excel in the prediction of stability, while sequence-based models are better at predicting protein fitness or activity [**131**].

2.4.3 Supervised learning on labeled data

The recent development of large foundation models allows proteome-wide zero-shot scoring of mutational effects. However, these scores correspond to probabilities ratios between wild-type and mutated sequences as predicted by the model, and are independent of the particular context or experimental measure considered. When sufficient curated data for a protein property are available, training a supervised model on such data shall provide a better alignment between predictions and true experimental values. In addition, design of supervised models may exploit inductive biases for the specific problems (such as modelisation of local environments, aggregation of global features, site-specific properties), resulting in performance improvements. For mutational effects, typical prediction tasks include binary classification of clinical pathogenicity, variations of folding stability and fitness for single sequences, and changes in binding affinity in protein complexes.

Mutational effects in single sequences

Supervised predictors estimate the functional impact of amino-acid substitutions by learning a mapping from variant features to labels (e.g., "deleterious/benign," clinical pathogenicity, or quantitative fitness-related properties, like protein stability). They differ mainly in the training labels used, including clinical assertions, literature-curated effects, proxy labels, the feature sets used as inputs, like evolutionary conservation, sequence context, structure/biophysics, protein networks, and the model class, from linear/ensemble learners to deep neural networks.

ML and hybrid models. One of the first ML methods is I-Mutant2.0, that trained Support Vector Machines on experimental stability $\Delta\Delta G$ data and takes either sequence or structure input to predict the direction and magnitude of stability change. Conservation-centric classifiers such as SIFT [**132**] and PolyPhen-2 [**133**] leverage MSAs or sequence profiles to quantify the likelihood of a substitution is at a given position; they are fast

and widely adopted but can struggle when MSAs are shallow or when function changes without strong conservation signals. Feature-rich ensemble methods (e.g., MutationAssessor [**134**], PROVEAN [**135**], ClinPred [**136**], MutPred2 [**137**]) integrate evolutionary descriptors with structural/biophysical terms, protein-level context, and sometimes gene-level constraint metrics, which improve models robustness. mCSM uses hand-crafted graph signatures paired with supervised random forest regressors for $\Delta\Delta G$ or functional impact assessment [**138**].

Given the large variety of models, ensemble methods were developed that pool diverse predictors into a meta-classifier/regressor. Examples include REVEL (Rare Exome Variant Ensemble Learner) [**139**], that trains a supervised meta-classifier on many constituent tools (comprising PolyPhen, SIFT, MutationAssessor) and conservation metrics to improve pathogenicity prediction on ClinVar benchmarks and CADD (Combined Annotation-Dependent Depletion) [**140**], that uses many annotations and a supervised SVM model trained on simulated vs observed variants to produce a general deleteriousness score for SNVs and indels. These ensemble methods are more robust than single predictors and mitigate their biases, but require feature extraction for all constituent tools, and when those are correlated the gains can be limited.

Deep learning approaches. These methods leverage representations of protein sequences and structures, either pre-trained or trained from scratch, and feed them to deep networks. Between them, PrimateAI [**141**] is a deep neural network that leverages sequence patterns and data for variations in primate species to predict clinical impact of single-point mutations; PrimateAI-3D variant also incorporates structural information via 3D CNNs to further improve pathogenicity prediction [**142**]. The models are trained on large datasets including primate polymorphisms for negative examples. For stability predictions, DeepDDG [**143**] trains deep neural networks on structural descriptors to predict $\Delta\Delta G$ for point mutations. These methods show strong empirical performance, but are more affected by training labels that biased by discovery pipelines, and usually lack interpretability.

Supervised DMS predictors. This class includes models with varying complexity trained on commonly heterogeneous DMS data. Envision [**144**] was among the first supervised models trained on aggregated DMS datasets, using gradient boosting on features derived from both sequence and structure. CPT-1 [**145**] trains a linear regressor on ClinVar and DMS data, using amino acid descriptors, MSA and structural features coupled with zero-shot scores from DL models (ESM, EVE) for cross-protein predictions. Among methods that predict DMS outcomes in a sequence, given a limited set of experimental results (simulated with cross-validation strategies), are ProteinNPT [**146**], that employs a transformer architecture with different attention mechanisms trained on top of embeddings from MSA Transformer, and Kermut [**147**], based instead on Gaussian processes with a specifically designed kernel that integrates sequence and structure information. These models are valuable for protein engineering and variant interpretation, but generalization across proteins and assay types remains challenging due to data heterogeneity.

Mutational effects in protein complexes

Predicting how point (or multiple) mutations change the binding free energy ($\Delta\Delta G$) of protein-protein or protein-ligand complexes is central to understanding disease variants, engineering affinity, and guiding drug design. Computational approaches for this task fall into three broad classes: physics-derived methods for energy computation like FoldX [**148**] and Rosetta [**149**], statistical and knowledge-based methods using potentials and handcrafted features processed through machine learning algorithms, and deep learning methods that learn mappings from sequence and structure-derived descriptors or pre-trained representations to $\Delta\Delta G$. Below is a brief overview of major models, with their principal strengths and limitations.

ML and hybrid models. One of the earliest ML predictors for $\Delta\Delta G$ is BeAtMuSiC [**150**], that uses coarse-grained, knowledge-based statistical potentials and structural decomposition to estimate $\Delta\Delta G$ of binding for point mutations at interfaces. It is an interpretable and relatively fast method, but requires a 3D structural model. iSEE [**151**] uses a limited number of interface, structure, evolution, and energy-based features for its predictions. SAAMBE-3D [**152**] and the sequence-only variant SAAMBE-SEQ [**153**] use instead features from structure, biophysical terms, and evolutionary conservation. The more recent mCSM-PPI2 [**154**] encodes the local atomic environment as graph-based signatures (distance patterns between atom types) and couples these with random forest regressors.

Other models combine ML regressors with physical-based descriptors. Among these, MutaBind2 [**155**] combines a small number of physically interpretable features (solvent interactions, conservation, stability terms) with ML regressors to predict $\Delta\Delta G$ for single and multiple mutations. BindProfX [**156**] similarly fuse physics/statistical potentials with ML. This approach often leads to better generalization for limited data, due to the explicit use of thermodynamic proxies. However, it is more computationally intensive and sensitive to selected features.

Deep learning approaches. MuPIPR [**157**] is a sequence-based approach that employs protein representations obtained from a LSTM architecture and processed through GRU and convolutional layers for the prediction of $\Delta\Delta G$. Among models using structural inputs, TopNetTree [**158**] integrates topological descriptors in CNN-assisted gradient boosting trees, obtaining competitive results on challenging antibody-antigen complexes. Different studies, like Deep Local Analysis (DLA) [**159**] showed that 3D CNNs operating on voxelized protein local environments can predict $\Delta\Delta G$ from structure without heavy feature engineering. Such methods can learn local 3D interaction patterns and outperform classical methods when sufficient data are available. However, they usually require self-supervised pre-training on structural data of protein complexes before being applied to downstream tasks. More recent models rely instead on graph neural networks and transformer architectures that natively operate on atomic or residue graphs, and can incorporate sequence embeddings from large pre-trained protein language models. Newer methods such as DDMut-PPI [**160**] and several transformer/GNN hybrids report state-of-the-art results on datasets from SKEMPI, by extracting both local mutant subgraphs and global context. These models are indeed more flexible, integrating multimodal inputs, but they entail reduced interpretability

compared to explicit features, and are more sensitive to training splits and dataset biases.

Physics/energy methods and hybrid approaches provide interpretable energetic components and often generalize across types of mutations when structural modeling is accurate, but they can miss subtleties of environment and statistics captured by ML and are also generally much slower in the execution. On the contrary, ML methods are faster and provide good correlations in test benchmarks, but they are usually not interpretable and depend on dataset curation, being thus more prone to biases.

Concerning models inputs, structure-based methods require experimentally determined or predicted structures, and errors in homology models or AlphaFold/ESMFold predictions can propagate and degrade performance. In addition, they are not feasible for large scale analysis, as folding predictions, especially for multimers, require considerable time and resources. Conversely, sequence-only predictors are valuable when structure is absent, providing a faster alternative; however, when an experimental structure is available, or the folding predictions have high confidence scores, structure-based methods are usually more accurate. Prediction of effects of multiple-point mutations and epistatic effects remains a challenging task for all the models.

2.5 Computational approaches for the prediction of protein disorder

The structural heterogeneity and context dependence of IDRs make experimental characterization difficult, thereby motivating the development of computational predictors. Over the last two decades, a wide range of methods have been proposed, ranging from early sequence-based machine learning classifiers to advanced deep learning architectures, usually trained on curated datasets such as DisProt [**53**].

The first generation of disorder predictors relied heavily on handcrafted sequence features such as amino acid composition, hydrophobicity, charge distribution, and low sequence complexity. Examples include IUPred [**161**] and its variants, which estimates disorder propensity by computing pairwise inter-residue interaction energies. Its central assumption is that ordered regions are stabilized by favorable contacts, while disordered regions cannot form sufficient stabilizing interactions to fold. AIUPred [**162**] is a more recent model derived from the IUPred family, that combines the original IUPred energy estimation descriptors with neural networks to improve generalization of disorder predictions.

With the increasing availability of curated IDR datasets, machine learning methods using SVMs, decision trees, or shallow neural networks became popular. For instance, DISOPRED3 [**163**] employs neural network classifiers trained on position-specific scoring matrices (PSSMs) derived from sequence alignments, improving sensitivity to evolutionary conservation and mutation tolerance. Beyond binary classification, the model also predicts putative binding sites inside disordered regions. These models demonstrated better accuracy than purely statistical approaches, especially in detecting short disordered segments.

However, their limited model capacity and dependence on pre-computed sequence profiles restricted scalability and generalization across diverse proteomes.

As individual predictors often capture complementary aspects of disorder, ensemble methods were developed to integrate outputs from multiple models. Tools like MobiDB-lite [**164**] combine sequence-based statistical predictors with machine learning scores and more recent transformer-based approaches to provide more robust estimates of disorder across different contexts. Such ensemble approaches tend to outperform single predictors by reducing model-specific biases. Nonetheless, consensus models remain limited by the accuracy of their constituent tools and may smooth over meaningful differences between predictors.

The advent of deep learning has significantly transformed disorder prediction. For example, SPOT-Disorder [**165**] uses bidirectional LSTMs combined with evolutionary profiles to capture both local sequence motifs and long-range dependencies. Similarly, flDPnn [**166**] applies fully connected neural networks and random forest classifiers enhanced with residue-level embeddings to predict disordered regions and their associated functions, like binding sites and linkers. RawMSA [**167**] uses different neural networks, including CNNs and LSTMs, for the prediction of various protein properties, including disorder, from MSAs. These models achieve higher accuracy by directly learning hierarchical sequence representations. Their limitations include high computational costs for generating MSAs and challenges in generalizing to sequences lacking homologous information.

Recent progress has been driven by PLMs, which learn contextual embeddings from large unlabeled sequence datasets. For example, predictors like SETH [**168**] leverage embeddings from ProtT5 transformer to achieve state-of-the-art performance in identifying both long disordered regions and shorter flexible linkers. The main limitations of transformer-based models are their interpretability and high computational cost, as training requires massive GPU resources and inference can be expensive for large proteomes. AlphaFold-derived descriptors, like pLDDT confidence, ensemble variability, and relative surface accessibility of predicted structures, are also successfully used as disorder descriptors.

Predictors of conformational flexibility and dynamic disorder

Beyond binary disorder prediction, some methods aim to quantify flexibility on a continuous scale. Tools such as DynaMine [**169**] predict backbone dynamics from sequence using logistic regression models, calibrated against NMR chemical shift data. Similarly, ODiNPred [**170**] is a deep neural network trained on NMR chemical-shift derived order parameters and sequence features, designed to better capture variable ordering levels. More recent neural network approaches include MEDUSA [**171**], that applies CNNs to amino-acid sequence augmented by evolutionary information from homologous sequences and physico-chemical residue descriptors to predict per-residue flexibility classes, related to crystallographic B-factor.

Limitations and outlook

Despite significant progress, disorder and flexibility prediction still face challenges. Most predictors treat disorder as a static sequence property, neglecting its context dependence on binding partners, post-translational modifications, or cellular conditions. In addition, prediction of functionality of disorder regions, like binding sites and linkers, remains challenging for all methods despite recent progresses due to scarcity of experimental data and noise of underlying biological signals. Transformer-based models offer improved accuracy, but their black-box nature hinders biological interpretability.

2.6 Evaluation metrics

In the following chapters, various metrics are used to assess the performance and confidence of the models. For easier lookup, they are defined below, grouped by related task.

Regression metrics

Mean Square Error (MSE) and Root Mean Square Error (RMSE) are used to quantify the error of a regression model. For ground truth values $\hat{y}$ and predicted values y, they are defined as

$$\begin{aligned} \mathrm{MSE} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ \mathrm{RMSE} &= \sqrt{\mathrm{MSE}} \end{aligned} \tag{2.14}$$

Pearson correlation coefficient (PCC) measures the linear correlation between two data sets, with values ranging between -1 and 1. 0 implies no linear correlation, while $(-)1$ corresponds to perfect (anti-)correlation. Given paired observations (x_i, y_i) for $i = 1 \dots n$, of variables X and Y, PCC is defined as

$$r(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.15}$$

where $\bar{x}$ and $\bar{y}$ are the sample means for the two datasets.

Spearman rank correlation coefficient (SCC) assesses how well the relationship between two variables can be described using a monotonic function, and corresponds to the PCC between the rank values of the data sets. SCC is less sensitive than PCC to outliers, and is commonly used to compare experimental mutation scores with unsupervised models predictions, since their relation is not necessarily linear. Given paired observations (x_i, y_i) of variables X and Y and their ranks $(R[x_i], R[y_i])$, SCC is computed as

$$\rho(X,Y) = r(R[X], R[Y]) \tag{2.16}$$

Classification metrics

Precision and recall, also referred to as positive predictive values (PPV) and sensitivity, are two basic metrics to evaluate binary or multi-class classifiers. Given true positives (tp), false positives (fp) and false negatives (fn) for a predefined classification threshold, the two metrics are defined as

$$\begin{aligned} \text{Precision} &= \frac{tp}{tp + fp} \\ \text{Recall} &= \frac{tp}{tp + fn} \end{aligned} \tag{2.17}$$

F1 score is the harmonic mean between precision and recall:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{2.18}$$

Matthews correlation coefficient (MCC) is a balanced measure for assessing classification performance, when equal importance is given to positive and negative samples. Its values range between -1 and 1, where a coefficient of +1 refers to a perfect prediction, 0 to a random prediction and -1 indicates total disagreement between prediction and label. It is computed as

$$MCC = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \tag{2.19}$$

where tp, tn, fp, fn are true positives, true negatives, false positives and false negatives, respectively.

Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves are used to assess the performance of a binary classifier at different classification thresholds. In both cases, to draw the curves predictions are sorted based on their scores.

ROC plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The TPR is equivalent to recall, defined as $tp/(tp + fn)$, while the FPR is defined as $fp/(fp + tn)$, representing the proportion of negatives incorrectly classified as positives. The area under the ROC curve (ROC-AUC) provides a single measure of overall classifier performance, with a value of 1 indicating perfect classification and 0.5 corresponding to random guessing.

In contrast, PR curve plots precision against recall, and the area under the curve (PR-AUC), also referred to as Average Precision, has a value of 1 for a perfect classifier, and total number of positive observations as random baseline.

Confidence metrics for protein folding

Predicted Local Distance Difference Test score (pLDDT) is a per-residue measure of local confidence, learned by the folding prediction model during the training phase. pLDDT measures confidence in the local structure, estimating how well the prediction would agree with an experimental structure. It is based on the local distance difference test Cα, which is a score that does not rely on superposition but assesses the correctness

of the local distances [**172**]. A pLDDT above 90 assures that both the backbone and side chains are typically predicted with high accuracy. In contrast, a pLDDT above 70 usually corresponds to a correct backbone prediction with misplacement of some side chains.

Predicted Template Modeling score (pTM) measures the accuracy of the global structure of the protein and is relatively insensitive to localised inaccuracies [**173**]. A pTM score above 0.5 means the overall predicted fold for the monomer (or complex) might be similar to the true structure. A pTM score below 0.5 means the predicted structure is likely wrong.

For multimer predictions, particularly useful metrics are the ipLDDT score and ipTM score, that measure per-residue local confidence in the interface regions and relative orientations between proteins involved in the interaction. For iPTM score, values higher than 0.8 represent confident high-quality predictions, while values below 0.6 suggest likely unreliable results.

Chapter 3

MuLAN: inspecting protein interactions with light attention networks

In this chapter, we report on MuLAN, a deep learning model for the prediction of binding affinity changes in protein complexes caused by mutations from sequence information only. In addition, the model provides an unsupervised framework for the identification of binding and functional regions in single sequences. The content of the chapter is a modified version of the article "MuLAN: Mutation-driven Light Attention Networks for investigating protein-protein interactions from sequences", available at `https://doi.org/10.1101/2024.08.24.609515`.

3.1 Introduction

Proteins are vital to cell function, and understanding their behavior during interactions within complex systems is key to deciphering the biological mechanisms of life. This task becomes particularly daunting when considering how mutations affect protein interactions and their binding strength. Such mutations can impact protein behavior in several ways, including altering stability, affecting functional activity, and modifying protein-protein interactions (PPIs). These changes are typically triggered by physical alterations at the interaction sites, such as changes in hydrophobic areas, electrostatic interactions, or hydrogen bonds, leading to variations in the binding free energy, ΔG, between mutated and wild-type protein complexes. The assessment of this variance, or "binding affinity change" $\Delta\Delta G$, is crucial for understanding disease predisposition and the effectiveness of drugs [**174**].

Traditionally, experimental techniques have been employed to examine protein complex binding affinities, but these methods are expensive and time-consuming [**175**]. This has spurred the development of computational strategies to bypass these drawbacks. Over the past decade, numerous computational approaches considering three-dimensional structure information have been introduced to assess binding energy differences. They include molecular dynamics simulations reaching very accurate estimations, but requiring extensive computational resources which limit their large-scale application [**176**]. To address these challenges, empirical energy functions like FoldX and Rosetta have been developed. Despite their faster predictions, these models suffer from limited conformational sampling, leading to reduced accuracy. Recently, advancements in deep learning based on structural data

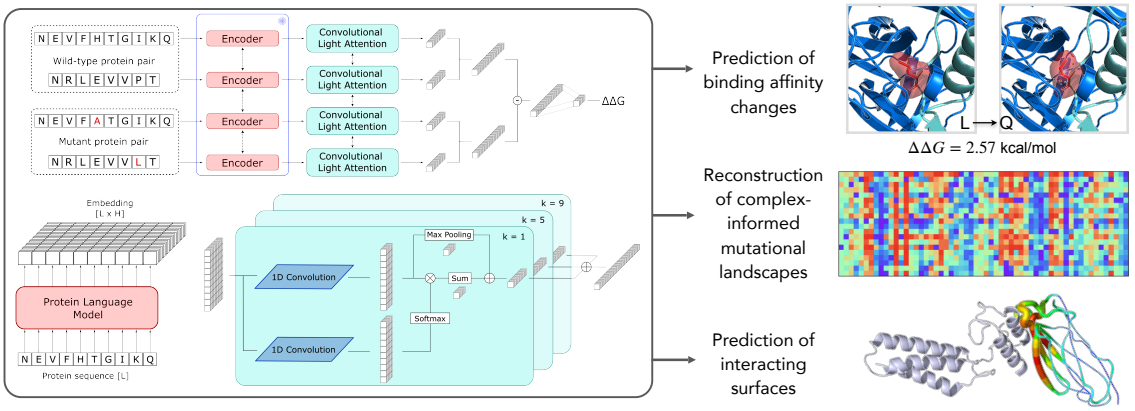

Figure 3.1: **Representation of MuLAN model and its tasks**. The model leverages a Protein Language Model encoder to extract features from wild-type and mutated sequences within a complex, and processes them through a convolutional light attention block, that produces interpretable representations for the prediction. The MuLAN architecture (left) is employed for three main goals (right): 1. to predict variations in binding affinity (top); 2. to reconstruct entire complex-informed mutational landscapes (center); 3. to identify potential interacting surfaces from the representations learned by the attention mechanism (bottom).

have shown promise, though their effectiveness heavily relies on the availability of extensive training datasets reporting experimental data for known complex structures. The creation of the SKEMPI v2.0 dataset [**38**] has significantly enhanced the precision and efficiency of methods like iSEE [**151**], mCSM-PPI2 [**154**], BindProfX [**156**], MuPIPR [**157**], MutaBind2 [**155**], GeoPPI [**177**], and DLA-Mutations [**159**]. All these methods incorporate geometrical and physico-chemical information on the complexes, and in some cases, information on the changes in binding mechanisms, involving atomic properties, which are known to directly contribute to the binding affinity. A fundamental bottleneck on the performance of all computational methods so far is the availability of high quality structural complexes for both wild-type and mutants upon which they rely. The lack of suitable complex structures affects the large scale application of the methods [**178**].

The advent of protein language models and deep learning architectures, such as transformers [**80**], presents a novel challenge: understand protein interactions from sequences. Achieving this would usher in a new era, offering fresh insights into protein behavior based solely on their sequences and providing faster computational predictions. We hypothesize that protein mutational changes are key to inferring protein interaction sites and reconstructing mutational landscapes for proteins involved in binary interactions, directly using sequence data. To address this, we introduce a deep learning approach, MuLAN, based on light attention layers that combine contextualized representations from pre-trained protein language models to encode protein complexes and predict binding affinity changes due to single-point and multiple-point mutations. As anticipated, MuLAN predictions will facilitate a detailed understanding of protein interactions (3.1).

3.2 Results

We introduce a deep learning architecture, MuLAN, which takes four sequences as input: two representing a wild-type binary complex and two representing the mutated sequences within the complex. MuLAN outputs a score evaluating the variations in binding affinity ($\Delta\Delta G$) caused by single or few-point mutations, potentially occurring in both sequences. MuLAN is central to our approach for understanding protein interactions from sequences. It achieves three main goals: predicting binding affinity changes, identifying residues at the protein interface within the complex, and reconstructing complex-informed mutational landscapes *based on* $\Delta\Delta G$ *predictions*. This last goal marks significant progress over previous methods, as it uniquely accounts for the dependency of the mutational landscape on the interacting protein partner.

3.1 illustrates the design of the MuLAN architecture (see Methods). The design is guided by the principle that predicting properties related to protein complexes solely from sequences, while maintaining plausible biological interpretations, requires a model that processes input sequences and integrates their representations without losing information about individual residues. MuLAN embodies this intuition by learning a weighted average of residue embeddings extracted from pre-trained language models for both wild-type and mutated protein sequences to predict binding affinity changes caused by mutations. To manage the high dimensionality of the embeddings, we first apply parallel convolutional filters to learn a lower-dimensional representation, effectively combining information from neighboring residues. Light attention is achieved by applying softmax normalization to the convolution output over the sequence length, with computational complexity scaling linearly with sequence length. It is important to note that long-range correlations, which cannot be captured by this approach, are already encoded in the sequence embeddings. For $\Delta\Delta G$ prediction, we use a multi-layer perceptron with linear activation on the difference of embeddings for mutated and wild-type complexes. In addition to the baseline MuLAN model, we trained a second version called iMuLAN. This enhanced model incorporates iGEMME scores [**128**] as auxiliary features, representing the sensitivity of mutations to be deleterious. These scores were concatenated to the input of the final linear layer in MuLAN. Since iGEMME scores pertain to single-point mutations, we evaluated iMuLAN exclusively for this scenario.

3.2.1 MuLAN predicts binding affinity changes in binary complexes

Binding affinity changes can be predicted from the sequences of both wild-type and mutant proteins. A comprehensive assessment of MuLAN's performance was conducted through a series of experiments involving thousands of mutations and hundreds of complexes. We used benchmark datasets for both single and multiple mutations. These experiments utilized datasets derived from SKEMPI v2.0, including S1102, S1131, S2003, and S4169 for single-point mutations in binary complexes, and S1400 for multiple-point mutations. A comprehensive description of these datasets can be found in Methods and Table B1. We used Pearson's correlation coefficient (PCC) and root mean squared error (RMSE) to compare the experimental values with those predicted by our model, following standard

Table 3.1: MuLAN and iMuLAN performance. RMSE is measured in kcal/mol.

Model	S1102		S1131		S2003		S4169		S1400	
	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE
Mutation-based splitting										
MuLAN–ESM2–3B	0.854	1.223	0.849	1.255	0.781	1.078	**0.739**	1.216	0.904	1.187
MuLAN–Ankh–large	**0.868**	**1.185**	0.866	1.211	0.774	1.068	**0.739**	1.152	**0.907**	**1.172**
iMuLAN–ESM2–3B	0.864	1.193	0.858	1.248	0.779	1.086	0.725	1.152	-	-
iMuLAN–Ankh–large	0.867	**1.185**	**0.871**	**1.196**	**0.792**	**1.059**	**0.739**	**1.134**	-	-
Complex-based splitting										
iMuLAN–ESM2–3B	0.698	**1.469**	**0.733**	**1.469**	**0.685**	**1.320**	0.564	**1.357**	-	-
iMuLAN–Ankh–large	**0.714**	1.480	0.731	1.495	0.618	1.417	**0.568**	1.419	-	-

regression analysis practices. We conducted cross-validation experiments using two distinct setups for splitting the training and testing data. In the first setup, a mutation-based split was used, where the training and testing data were divided according to individual mutations, allowing the model to encounter the same complexes in both phases but with different mutations. In the second setup, a complex-based split was employed, ensuring that the complexes used in training were entirely different from those in testing. This complex-based approach is more rigorous and provides a better assessment of the model's generalizability to new, unseen sequences.

Mutation-based splitting

Both the baseline MuLAN model (Methods) and the iMuLAN version were evaluated using two different protein language models (PLMs) for extracting embeddings: ESM2–3B [**112**] and Ankh–large [**109**]. Following previous studies, we implemented a 10-fold cross-validation for datasets containing single-point mutations, and a 5-fold cross-validation for the S1400 dataset. Results are detailed in 3.1. The Ankh–large model outperformed the baseline across all datasets on both metrics, except for the PCC in the S4169 dataset. Notably, the embedding dimensions for these models are 2560 and 1536, respectively, indicating that embedding size is not the determining factor for the model's predictive capability. Furthermore, integrating *i*GEMME scores consistently enhanced the performance metrics for both models, particularly in the S1102 and S1131 datasets.

MuLAN, which considers only the wild-type and mutated protein sequences as inputs, can easily handle multiple-point mutations — a capability often lacking in structure-based models. Notably, for the S1400 dataset, we achieved state-of-the-art results in terms of PCC and RMSE (Table 3.2). In the case of single-point mutations, our results were on par with other structure-based models for the S1102 and S1131 datasets. However, we observed lower PCC values for the S2003 and S4169 datasets, which are part of SKEMPI's second release. We hypothesised that these lower scores were due to the embeddings from PLMs not capturing enough information about the properties and structures of some proteins in these datasets. To validate this hypothesis, we used ESMFold [**112**] to predict the structures of wild type proteins within the studied datasets and assessed model confidence using pTM and average pLDDT scores. We found that the average pTM scores were 0.83 for the S1102, S1131, and S1400 datasets, but lower for S4169 (0.74) and S2003 (0.73).

Table 3.2: MuLAN and iMuLAN performance on benchmark datasets from SKEMPI. RMSE is measured in kcal/mol. Missing values ("-") are not reported by the corresponding analyses.

Model	PCC	RMSE	Model	PCC	RMSE
S1400			S2003 - mutation based		
MuPIPR	0.88	1.32	DLA-Mutation	0.81	-
MuLAN	**0.91**	**1.17**	MuLAN	0.79	**1.06**
S1102			S2003 - complex based		
MuPIPR	0.85	1.22	DLA-Mutation	**0.735**	**1.23**
iSEE	0.80	1.41	DLA-Mutation*	0.686	1.31
MuLAN	**0.87**	**1.19**	MuLAN	0.685	1.32
S1131			S4169		
TopNetTree	0.85	-	TopNetTree	**0.79**	-
Hom-ML-V2	0.857	1.28	MpbPPI	**0.79**	-
MpbPPI	0.865	-	MuLAN	0.74	**1.13**
MuLAN	**0.87**	**1.20**			

Similarly, the average pLDDT scores were higher (0.78) for S1102, S1131, and S1400, and lower for S4169 (0.74) and S2003 (0.75). These lower scores for the S2003 and S4169 datasets suggest that the predicted structures were less reliable, potentially explaining why sequence information alone was insufficient for accurately predicting protein properties in these cases. This was particularly true for proteins in the antibody-antigen classes, where we observed a consistent drop in performance.

Complex-based splitting

We conducted a 10-fold cross-validation on the single-point mutation datasets S1102, S1131, and S4169. For the S2003 dataset, we followed instead the complex-based split between training and testing sets as established in [**159**]. We excluded multiple-point mutations from this analysis and opted to evaluate using iMuLAN version, which demonstrated superior metrics in mutation-based split scenarios.

Although the evaluation metrics, particularly for the iMuLAN–Ankh–large model, tend to worsen in this setting, they still demonstrate the model's strong generalization capabilities on unseen complexes (Table 3.1). Notably, the RMSE values remain below 1.5 kcal/mol across all datasets.

Comparison with other approaches

We compared $\Delta\Delta G$ predictions of MuLAN with other recently developed approaches, previously evaluated on the same benchmark datasets extracted from SKEMPI. The majority of these methods rely on structural features for their predictions, or on a combination of sequence and structure information, while only MuPIPR [**157**] is entirely sequence-based. On S1400, we achieve an higher PCC of 0.90 and a lower RMSE of 1.17 kcal/mol, with respect to MuPIPR reported values of 0.88 and 1.32 respectively, showing an improvement in

predictions for multiple-point mutations, where other structure-based approaches typically struggle. On single-point mutations from S1102, we achieve a PCC of 0.86 and a RMSE of 1.19 kcal/mol, outperforming both iSEE and MuPIPR (3.2). On S2003, we compared MuLAN results with those from DLA–Mutation [**159**], a model that uses 3D convolutions and a self-supervised pre-training objective based on masking to model the local structural environment around the mutated residue, requiring high-resolution complexes for training and testing. In spite of a total different approach, based on global sequence features, we achieve a PCC of 0.79 in mutation-based cross validation, slightly lower than 0.81 of DLA–Mutation, and a PCC of 0.685 in the same complex-based train/test splitting considered. While being significantly lower than the 0.735 achieved by DLA–Mutation when enriched with auxiliary features (namely the structural region, evolutionary conservation, GEMME score, buriedness and physico-chemical properties of the wild-type residue), the result is comparable with the base model implementation (DLA–Mutation* in 3.2), that considers only the structural region as auxiliary feature, and outperforms the DLA-mutation ablated version that includes all additional features, but is not pre-trained with the masking objective (PCC = 0.608). These results show good generalization capabilities of the model, that do not depend on the accuracy and resolution of available structures in the region of interest. Finally, S1131 and S4169 are widely used evaluation datasets, constructed with single-point mutations from SKEMPI v1 and v2, respectively. On S1131, we outperform TopNetTree, Hom-ML-V2, and the more recent MpbPPI, all of which are based on deep learning, achieving a PCC of 0.87 and an RMSE of 1.20. However, on S4169, our model's performance falls short, with a PCC of 0.739, which is significantly lower than the 0.79 obtained with TopNetTree and MpbPPI. As mentioned above, we partially attribute this loss in performance to the less informative embedding representation of some proteins in this dataset, such as antibody-antigen complexes.

PLMs fine-tuning improves performances for smaller models

In an attempt to enhance the model prediction capabilities, we experimented with fine-tuning of backbone PLMs and applied two alternative fine-tuning strategies, with different objectives:

- *evo-tuning*: this strategy involves continual learning on the masked language modeling (MLM) task, where the model predicts a masked fraction of residues in the sequence based on the context. We restricted the training dataset to the four most common species in SKEMPI, namely *Homo sapiens*, *Bos taurus*, *Mus musculus* and *Escherichia coli*;

- *supervised fine-tuning*: this approach focuses on token classification fine-tuning for predicting residues that belong to inter-protein interaction surfaces. This information is not directly captured by the base models, despite their ability to predict intra-protein contact maps with minimal supervision [**179**].

We tested the fine-tuned models by using their embeddings as inputs to MuLAN, and we adopted a mutation-based cross-validation scheme on S1102 and S1400 datasets. The

results for RMSE, reported in Table B2, demonstrate that the *evo-tuning* strategy can effectively lead to improved performances even for smaller models, like ESM2–35M, due to the higher taxonomic proximity between sequences appearing in the training and in downstream tasks. On the other hand, *supervised fine-tuning* on interface classification generally increases the RMSE and therefore is not beneficial for the $\Delta\Delta G$ prediction.

Large-scale biological interpretation of light attention weights

Protein interaction surfaces are plastic, adapting to different partners such that the same residue at the interface might contribute differently to the binding strength of distinct interaction, reflecting both biochemical versatility and evolutionary adaptation. Identifying which residues are most relevant within a specific molecular context is therefore essential for understanding how interactions evolve and diversify.

MuLAN directly contributes to this goal by providingpartner-specific insights into both residue relevance and interaction strength. Although the model is not explicitly trained to identify interfaces, its attention weights effectively highlight interface regions, revealing which residues the network deems important for modulating binding energetics. This emergent interpretability offers a biologically grounded window into what the model has learned and enables detailed exploration of protein interaction mechanisms across evolutionary scales.

At the architectural level, MuLAN learns a weighted average of residue representations through convolutional filters responsible for its light attention mechanism. This design allows the extraction of interpretable attention weights that provide physical and biological insights about how residues contribute to binding (see Methods for attention weights extraction). The availability of experimental 3D structures in the SKEMPI database, combined with MuLAN's per-residue processing, enables a direct mapping between the network's attention outputs and structural regions on the protein. Namely, we conjectured that high-attention positions might be related to regions on the protein surface that are more likely to contribute to variations in binding affinity. In order to test this hypothesis, we computed the attention scores for the sequences involved in the S1102 dataset, for MuLAN–Ankh–large, and compared them with interfaces extracted from the corresponding Protein Data Bank (PDB) [**180**] structures with INTBuilder [**181**]. The resulting ROC AUC score of 0.757 (Figure B1)confirms that MuLAN attention maps accurately recover interface regions undergoing mutations, despite the model never being explicitly trained for this task.

Evaluation of MuLAN attention scores for interface detection in the PDB

To comprehensively evaluate MuLAN's ability to identify interaction interfaces across molecular and evolutionary diversity, we performed a large-scale analysis using structures from the PDB. The dataset included 224,888 protein chains across 123,613 entries, encompassing protein complexes with proteins, nucleic acids, and small molecules. For each molecular partner, residues were independently labeled as interacting when the central atoms of residues were within 8Å of one another. We computed the mean ROC AUC for all

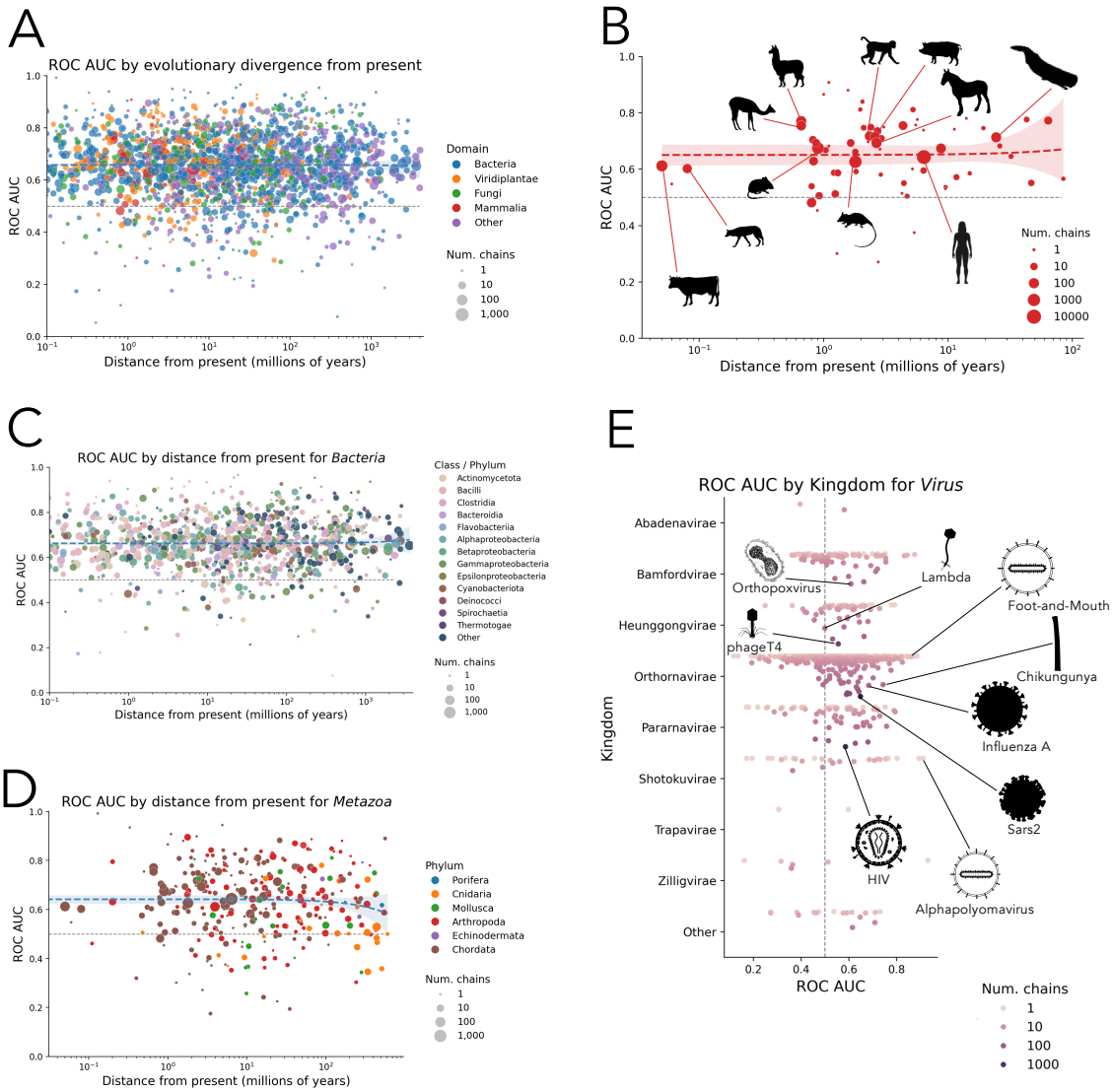

Figure 3.2: **Performance of MuLAN across evolutionary time**. ROC AUC for predicted interface residues, averaged per species, as a function of evolutionary divergence from the present. Each point represents the mean ROC AUC for all protein chains from a given species, with color indicating taxonomic domain and point size reflecting the number of protein chains analyzed. Species are organized in: (A) domains: Bacteria, Viridiplantae, Fungi, Mammalia, Others; (B) mammalian species; (C) classes and phyla of bacterial species (see inset legend); (D) metazoan phyla: Porifera, Cnidaria, Mollusca, Anthropoda, Echinoderma, Chordata; (E) viral kingdoms.

protein chains belonging to a given taxon, enabling us to quantify MuLAN's generalization across deep evolutionary time by assessing model performance on a per-species basis.

Figure 3.2A-D summarizes these results as a function of evolutionary divergence time. MuLAN maintains consistently high interface-prediction performance across species separated by hundreds of millions of years, with the mean AUC across all taxa remaining above 0.6. Remarkably, no systematic decline in performance is observed with increasing divergence, indicating that MuLAN's sequence representations capture deeply conserved biochemical and structural principles underlying molecular recognition. Comparable performance is maintained for viral proteins across viral kingdoms (Figure 3.2E).

This evolutionary robustness becomes even more evident when examining how MuLAN performs across diverse classes of molecular partners. As shown in Figure 3.3A, MuLAN achieves an overall ROC AUC of 0.642, with an AUC of 0.612 specifically for protein-protein interactions, and a notably higher classification accuracy of 0.763 on DNA/RNA interfaces. The comparatively lower AUC for protein-protein interactions is consistent with their intrinsic complexity: proteins often interact with multiple partners, sometimes through partially overlapping interfaces, which can complicate evaluation due to the shared or degenerate nature of binding regions. To probe this hypothesis, we represented each protein as a residue-level contact graph, linking residues with C_α-C_α distances below 7Å. For each residue, we computed a smoothed attention score by averaging its own score with those of its graph neighbors. Recomputed metrics revealed modest but consistent gains in performance (AUC=0.663 for ligands, 0.767 for DNA/RNA, 0.615 for proteins, and 0.645 overall), indicating that spatially proximal residues tend to share similar functional relevance as captured by attention scores.

For small-molecule ligands, Figure 3.3A reports an overall AUC of 0.660 across the full PDB. A refined "activome" analysis based on UniProt annotations further improves performance, reaching an AUC of up to 0.77. Specifically, beginning from all PDB structures, we aggregated ligand-binding sites from chains mapping to the same UniProt entry. The initial 224,888 PDB chains were mapped to 20,696 UniProt sequences, and we retained only proteins with ≤ 1024 amino acids and at least one annotated ligand-binding site; proteins lacking any positive labels were excluded. We then evaluated ROC AUC across varying thresholds t, where t denotes the minimum number of structures in which a binding site must appear to be considered a true positive. As summarized in Figure 3.3E, performance consistently improves at higher thresholds, in line with the expectation that frequently observed binding sites correspond to more stable and biologically recurrent interactions.

The high AUC values obtained especially for nucleic acids and ligands suggest that the attention mechanism emphasizes functionally critical sequence positions—typically localized hotspots involved in recognition or catalysis.

To further dissect the origins of this broad generalisation capacity, we examined specific PDB-derived examples and interaction types, assessing how MuLAN's attention scores behave across diverse molecular partners and structural contexts. Interestingly, the regions highlighted by MuLAN attention are not directly related to structural motifs or domains; they also include coiled and flexible regions that are observed to mediate interactions in experimentally determined structures from the PDB. Using DSSP [**183**], we assigned

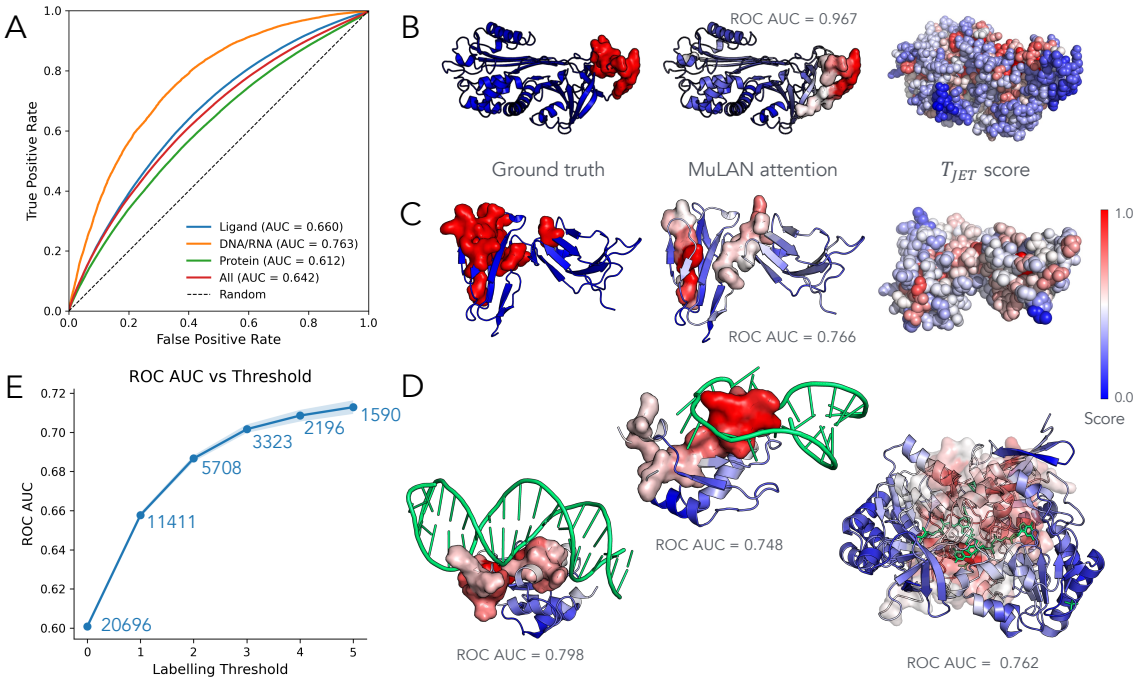

Figure 3.3: **MuLAN large scale predictions of interface residues on 224,888 PDB chains**. (A) ROC curves of MuLAN-Ankh-large model tested on PDB chains for interface predictions. Results are reported separately for interactions with small ligands, DNA/RNA, other proteins, or all interactions combined. (B) Experimental interaction surface (left), MuLAN scores (center) and T_{JET} scores (right) for Alaserpin protein (PDB ID 1K9O_I), example of protein interface that is correctly identified by MuLAN, despite its low evolutionary conservation (low T_{JET} scores). MuLAN ROC AUC for the protein is reported. (C) Experimental interaction surface, MuLAN scores and T_{JET} scores for Low affinity immunoglobulin gamma Fc region receptor III-B (PDB ID 1E4K_C). T_{JET} scores are taken from [**182**]. (D) Examples of MuLAN attention predictions for interactions with DNA (PDB ID 7RH2), RNA (PDB ID 1A9N), and small ligands (PDB ID 2V3V). red(E) ROC AUC values for UniProt ligand-binding sites, grouped by the number of known small-molecule interactors. The threshold t specifies the minimum number of molecules required for inclusion, and $t = 0$ corresponds to using all molecules. The number of proteins considered at each t is reported in the figure.

secondary structure classes for 177,545 chains and evaluated model performance across helices, β-strands, and coils. As shown in Figure B2, interaction detection varies by secondary structure and interaction type: DNA/RNA and ligand interfaces are more accurately identified in helical regions (AUC = 0.799 and AUC = 0.683, respectively), whereas protein-protein interactions are more effectively predicted in β-strand regions (AUC = 0.639). Despite these differences, overall performance remains comparable across secondary structure types, with AUC values for all interactions ranging between 0.650 and 0.662.

We next asked how these attention patterns relate to evolutionary conservation. Not all accurately detected interfaces correspond to evolutionarily conserved regions. As illustrated in Figure 3.3BC, MuLAN correctly identifies functional interfaces even in regions of low-conservation, based on the evolutionary conservation measure T_{JET} [**125**] applied to homologous sequences. This demonstrates that the model captures the functional and structural heterogeneity of interface regions, including RNA- and DNA-binding surfaces as well as small molecules pockets (Figure 3.3D).

Taken together, these results demonstrate that MuLAN captures the evolutionary grammar of molecular interactions. Trained solely on protein sequence data, the model generalizes across molecular partners, secondary-structure classes, and even across species separated by hundreds of millions of years of divergence. Its attention mechanisms reveal how the statistical signals learned from sequences correspond to the physical and evolutionary logic of molecular recognition—linking mutation, structure, and function in a unified framework.

MuLAN attention scores identify binding motifs in disordered regions

The large scale analysis of attention scores accuracy in the detection of binding interfaces in PDB structures provides an evaluation of the generalization capabilities of such framework. However, one interesting application of sequence-based approaches such as MuLAN is their ability to infer properties about disordered regions, which might be overlooked by structural approaches. In particular, we evaluated MuLAN's ability to correctly identify binding motifs in disordered regions. For that, we considered labeled sequences taken from the Binding-IDR dataset from the third edition of the Critical Assessment of Intrinsic protein Disorder (CAID) challenge, including 52 experimentally annotated proteins recently included in the DisProt database [**54, 184**]. As in the previous evaluation, we computed attention scores for all test proteins using the provided sequences, and we considered AUROC and AUPRC metrics for comparison with other top predictors, as reported in CAID results (Table 3.3). MuLAN achieves an AUROC of 0.590 and an AUPRC of 0.472, positioning itself in the top five predictors for both metrics. It is important to notice, however, that the other methods were specifically trained on the task, while MuLAN achieves comparable performances in an unsupervised setting. This result demonstrates that MuLAN's attention predictions can be leveraged to identify disordered binding motifs on unknown proteins from single sequences only.

Model	AUROC	AUPRC
bindEmbed211DR-rawGeneral	0.641	0.514
bindEmbed211DR-idrGeneral	0.635	0.497
LIPNet	0.619	0.505
ProBiPred-Protein	0.613	0.472
MuLAN	0.591	0.472
AIUPred-2-Binding	0.590	0.463

Table 3.3: Classification metrics for MuLAN and top five predictors on CAID3 Binding-IDR dataset. Results are taken from `https://caid.idpcentral.org/challenge/results`

Attention scores and splicing variants

Alternative splicing regulatory mechanism contributes to transcriptomic and proteomic complexity in eukaryotes, allowing the generation of multiple distinct protein sequences (isoforms) from a single gene locus [**185**]. While many splicing events fall into variable regions of protein structures or affect complete domains, some of them affect structured and well-conserved regions of the corresponding protein family, effectively modifying the protein folding and possibly allowing for new functional and interaction properties [**186**]. To inspect whether MuLAN attention scores would be sensitive to this behavior, we compared results for four TP53_HUMAN isoforms (UniProt IDs P04637-1, P04637-2, P04637-4, P04637-8), for which the splicing event only affects intrinsically disordered regions, keeping the structured domain almost intact, with isoforms for four other proteins taken from [**186**], namely DNMT2_HUMAN, AUHM_MOUSE, HER1_WORM and TIP30_HUMAN, for which the splicing event induces a non-trivial rearrangement of the protein structure. For each protein, we locally aligned the different variants and computed the Pearson correlation coefficient between MuLAN attention scores on aligned sites, as reported in Figure 3.4. While for TP53_HUMAN we considered four splicing variants, for the other proteins we considered the identified non-trivial variant and the sequences corresponding to the closest folds to the native and modified structures in SCOP database [**187**], labeled by their astral ID [**188**]. For TP53_HUMAN variants, correlations are higher than 0.96, indicating an essentially identical interaction behavior for the structured domain (scores for aligned sequences are reported in Figure B3). However, for the other proteins, correlations between native protein and isoform are lower, ranging from 0.47 for DNMT2_HUMAN to 0.91 for AUHM_MOUSE (see Figure B4 for corresponding alignments), while they are higher than 0.96 between the native protein and its SCOP structure. Variations are more marked close to the splicing site, and in particular they are evident in DNMT2_HUMAN, for which the splicing variant is much smaller than the native sequence. Interestingly, correlations between spliced variants and closest SCOP fold reflect their similarity in structure and potential interactions, with moderate values in the range 0.27 to 0.60, despite high sequence divergence (minimum Blastp e-value is 0.25 between Q9BUP3-2 and d1gsoa2). This analysis reveals MuLAN attention ability to identify potential functional differences in isoforms from their single sequence only, opening the possibility of extending it for the screening of large databases.

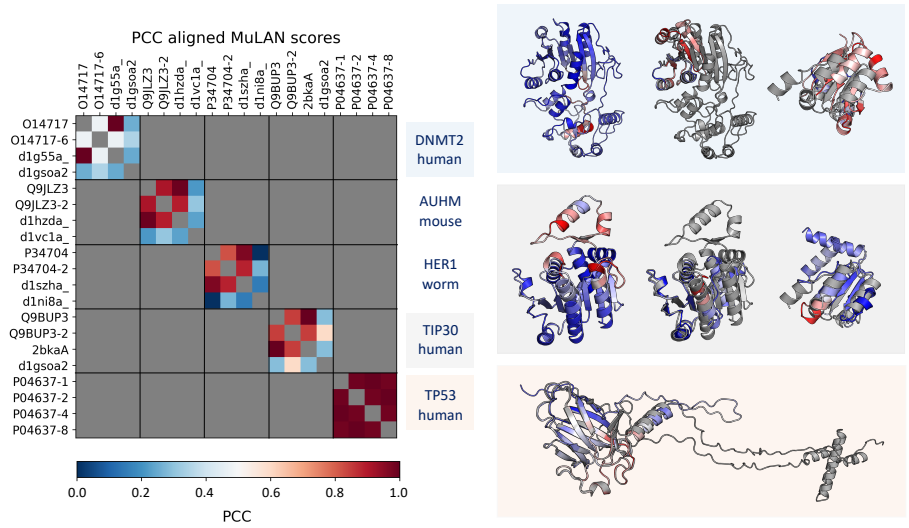

Figure 3.4: **MuLAN attention scores variations for splicing variants**. On the left, Pearson correlation coefficients between MuLAN attention scores for isoforms and closest SCOP structures, on aligned positions. In the last subsquare, four isoforms of TP53_HUMAN are considered, labeled by their UniProt ID, while for the others subsquares native protein, splicing variant, and closest SCOP structures, identified by ASTRAL ID are reported, respectively. On the right, aligned structures colored by MuLAN attention score. For DNMT2_HUMAN protein (top frame), structures are reported for UniProt ID O14717 (left), spliced variant O14717-6 aligned with O14717 (center) and O14717-6 aligned with d1gsoa2. Analogously, for TIP30_HUMAN, structures are reported for Q9BUP3, spliced variant Q9BUP3-6, and again digsoa2. Last structure represents TP53_HUMAN protein, with P04637-8 isoform aligned with P04637-1.

3.2.2 MuLAN reconstructs complex-dependent mutational landscapes based on $\Delta\Delta G$ predictions

The ability to rapidly predict $\Delta\Delta G$ variations for any given mutation in a protein complex using only sequence inputs enables the reconstruction of Deep Mutational Scanning experiments that measure binding affinities in complex interactions. These reconstructed matrices not only assess the variation of this property for the query protein upon mutation, but are also informed by the interacting partner, resulting in key differences in predictions depending on the specific binary complex.

We explored the model capabilities in this context by conducting three distinct experiments, each comparing MuLAN's predictions with those obtained from ESCOTT [**128**], a computational model designed to predict mutational effects in a protein based on sequence and structure, thereby enabling a biological interpretation of the observed differences. All experiments used the MuLAN–Ankh–large model (thus without addition of *i*GEMME scores, to avoid possible biases), trained on the S1102 dataset. To minimize information leakage, we ran BLASTp and removed any complexes where one of the partners shared more than 40% sequence identity with training proteins. The three experiments included: (i) identifying interacting residues in Bovine Pancreatic Trypsin Inhibitor, (ii) identifying

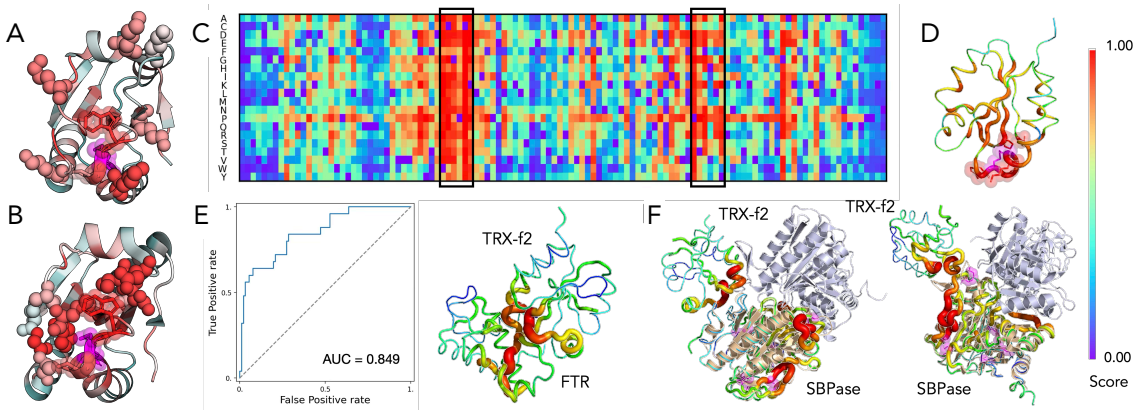

Figure 3.5: **MuLAN predictions for TRX complexed with diversed partners**. (A) Selectivity determinants for the *C. reinhardtii* type-f2 TRX protein. (B) Selectivity determinants for the *C. reinhardtii* type-h TRX protein. (C) Mutational landscape predicted by MuLAN for TRX-f2 in interaction with Ferredoxin-thioredoxin reductase (FTR), where highlighted regions are mostly involved in the interaction. (D) Average predicted values projected on TRX-f2 structure; the residues associated with highest values are represented as spheres, among which cysteins involved in redox catalyzation are highlighted in magenta. (E) Interface predictions for TRX-f2–FTR interaction, with associated ROC curve and projection on AlphaFold-Multimer predicted structure (RMSD=0.374 from homologous complex PDB ID 2PVO, in *Spinacia oleracea*). 27 residues in TRX-f2 were positively labeled as part of the interface, based on a 5.0 Å distance cutoff. (F) TRX-f2 complexed with SBPase, as predicted by AlphaFold-Multimer (on the left) and ESMFold (on the right), with projected interface predictions. SBPase in dimeric form (PDB ID 7B2O) is aligned to the predicted structure. Cysteines on SBPase, that suggest possible interaction sites, are highlighted in magenta.

selectivity determinants and conformational states of complexes involving thioredoxin in *C. reinhardtii*, and (iii) analyzing predictions for gain-of-function mutations in the MeFV gene and other inflammation-related proteins, where unsupervised mutational effects predictors typically underperform [**128**].

Residues involved in BPTI interactions identified as highly mutation-sensitive by MuLAN

We considered Bovine Pancreatic Trypsin Inhibitor (BPTI) interacting with three homologous proteins in binary complexes, namely its coevolved biological target bovine trypsin (BT), the non-cognate trypsin paralogs bovine α-chymotrypsin (ChT) and human mesotrypsin (MT), taken from [**189**]. Sequences for the interacting proteins were taken from the corresponding PDB entries 3OTJ, 1CBW and 2R9P. BPTI protein presents a loop involved in the interaction, for which mutational effect are difficult to predict for tools based on single sequences.

We ran the trained model and obtained predictions for all possible single-point mutations in BPTI, and we ranksorted the predictions in order to compare them with ESCOTT, that assigns scores in the range [0, 1], with higher values related to more impactful mutations.

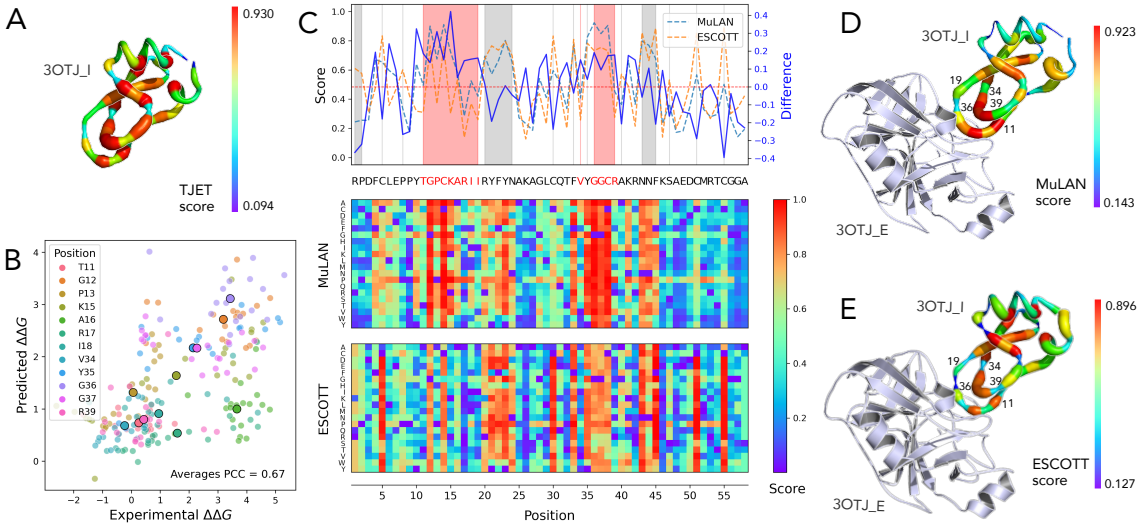

Figure 3.6: **MuLAN analysis of the BPTI (Bovine Pancreatic Trypsin Inhibitor) complexed with trypsin**. (A) T_{JET} evolutionary conservation scores for residues in BPTI, as computed by JET [**125**] from the multiple sequence alignment, plotted on structure PDB ID 3OTJ_I. (B) MuLAN $\Delta\Delta G$ predictions for mutations considered in [[**189**]], for which experimental values are available. The different colors refer to the positions of the mutated amino acid, and the highlighted dots represent the averages over all mutations in that specific position. (C) MuLAN and ESCOTT positional average scores (left axis) are plot together with their differences (right axis). Interface residues correspond to positions 11-19, 34 and 36-39 in the sequence and are indicated with red characters. They are highlighted either by a red line or a red background in the plot. They are identified from the complex structure PDB ID 3OTJ. Positions highlighted by a grey line or a grey background correspond to residues with average score, over all mutations, $> .5$ in ESCOTT, hence predicted to be highly sensitive to mutations. MuLAN (center) and ESCOTT (bottom) landscape reconstructions, where the color scale from 0 (low effect) to 1 (high effect) corresponds to MuLAN/ESCOTT scores. (C) MuLAN and ESCOTT positional average scores (left axis) are plot together with their differences (right axis). Interface residues correspond to positions 11-19, 34 and 36-39 in the sequence and are indicated with red characters. They are highlighted either by a red line or a red background in the plot. They are identified from the complex structure PDB ID 3OTJ. Positions highlighted by a grey line or a grey background correspond to residues with average score, over all mutations, $> .5$ in ESCOTT, hence predicted to be highly sensitive to mutations. MuLAN (center) and ESCOTT (bottom) landscape reconstructions, where the color scale from 0 (low effect) to 1 (high effect) corresponds to MuLAN/ESCOTT scores. (D) Residues in the BPTI protein are colored with positional averages from the MuLAN prediction matrix in (C). (E) Residues in the BPTI protein are colored with positional averages from ESCOTT prediction matrix in (C).

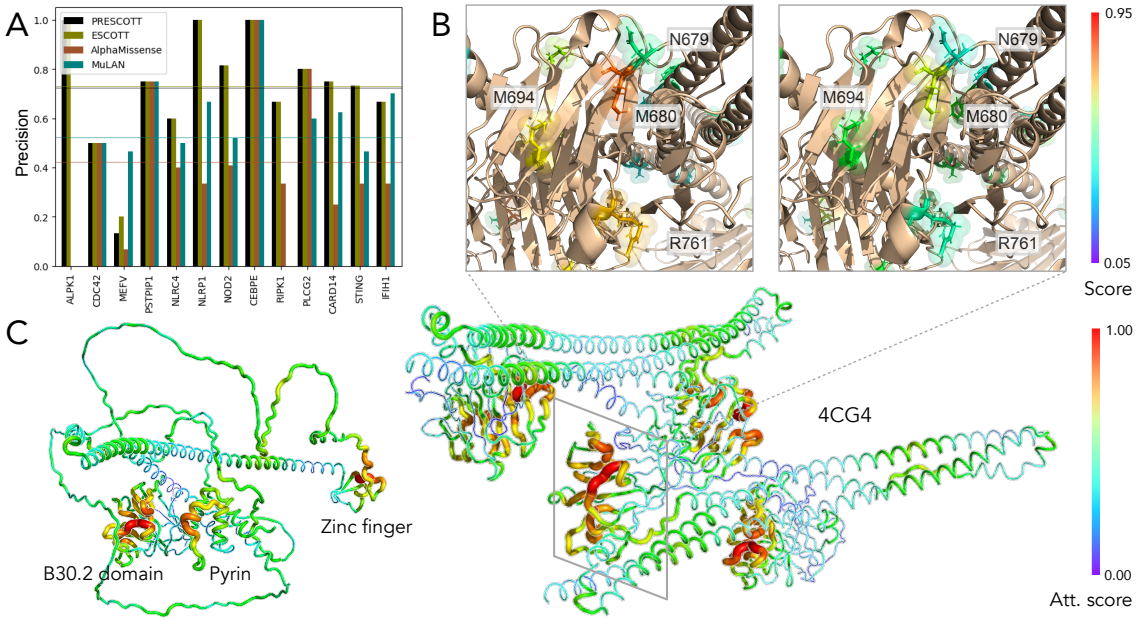

Figure 3.7: **MuLAN predictions for Gain-of-Function (GoF) mutations in MEFV**. (A) Fraction of correctly predicted pathogenic mutations from the GoF dataset [**128**], comprising 13 proteins involved in inflammation. (B) Normalized average predictions from MuLAN (left) and ESCOTT (right) for some of the position considered for GoF mutations in the MEFV protein. (C) Interface scores computed from MuLAN and projected on AlphaFold predicted structure of the MEFV protein (left), where the different domains are highlighted, and on homodimeric complex assembly through the CHS-B30.2 domain (PDB ID 4CG4), corresponding to residues 414-781 of MEFV. MuLAN analysis of MeFV was performed on the dimeric complex.

Our findings are summarized in 3.6, where we report MuLAN predictions as well as their comparison with ESCOTT. We also report a comparison with T_{JET} scores [**125**] measuring the evolutionary conservation levels of the amino acids (3.6A). While the model fails at capturing the differences in experimental binding affinities for the three complexes due to high sequence similarity in interacting proteins, it is able to capture additional information on single amino acid variants predictions caused by interaction. In particular, MuLAN results tend to enhance scores for residues belonging to the interacting surface (see red residues and regions in 3.6C, top), and decrease the peaks for other regions sensible to mutations see grey residues and regions in 3.6C, top. The predictions also show general similarities with evolutionary conservation scores (3.6A), which is an information that is expected to be captured by PLMs [**130**]. However, while a highly conserved region is found in the interior of the protein, and only a subset of residues involved in the interaction is labeled as conserved, MuLAN extrapolates beyond this conserved context, by assigning the highest scores to residues on the interface (in particular residues 34-38 and 11-15). This example, combined with the unsupervised analysis of interfaces from MuLAN representations, proves that the model can effectively learn information on the protein complex from single sequences.

MuLAN enhances the prediction of gain-of-function mutations in the MeFV gene and other proteins involved in inflammation

Gain-of-function mutations are a class of missense mutations leading to enhanced protein activity. Most of the tools for classification of mutations, including PRESCOTT and AlphaMissense, struggle with the identification of pathogenic gain-of-function mutations, in particular for those found in the MEFV gene, involved in various autoinflammatory diseases. This protein is of particular interest since many single-point gain-of-function mutations cause Familial Mediterranean Fever (FMF) and no current computational approach performs well on it. In an attempt to predict the effects of such mutations, we reconstructed the mutational landscape matrix with MuLAN for MEFV protein, by considering its homodimeric interaction (3.7BC). For comparison, we tested the same procedure also on other proteins from the GoF dataset known to be involved in autoinflammatory diseases. The dataset contains single-point mutations for 13 proteins, in variable proportions (from the single mutation R219H of CEBPE to the 30 mutations of IFIH1). We compared the results obtained with PRESCOTT, ESCOTT and AlphaMissense on this set of mutations, and extracted the attention scores from MuLAN for interface prediction (3.7A). MuLAN improves over all computational methods for MeFV and outperforms AlphaMissense on almost all other proteins in the dataset. 3.7B illustrates MuLAN evaluation of a MeFV region where a number of known pathogenic mutations is localised and the corresponding ESCOTT evaluation which misses to identify highly sensitive residues.

MuLAN identifies selectivity determinants and facilitates the selection of structural conformations in thioredoxin

Thioredoxins are widely spread across the tree of life and play a key role in controlling the redox status of protein disulfide bonds in all non-parasitic organisms [**190**]. Their activity is ensured by the presence of a solvent-exposed motif (most commonly Trp-Cys-Gly-Pro-Cys) containing two cysteine (Cys) residues that are able to catalyze protein disulfide reduction [**191**]. In plant chloroplasts, five phylogenetically distinct classes or types (f/m/x/y/z) differentially contribute to photosynthetic functions [**192**] while cytosolic and mitochondrial isoforms (types h/o) control redox signaling in plants like in all organisms. The TRX fold is ancient, highly stable, and conserved, making it difficult to attribute TRX/target complementarities. In Chlamydomonas, targets have been identified solely for types f/x, while types m/y/z remain completely uncharacterized. MuLAN identified several residues as critical for TRX activity and binding in both f- and h-type thioredoxins. This was achieved by averaging the predicted impact of all possible mutations at each position in the sequence. By integrating MuLAN scores with an independent functional analysis of TRX homologous sequences [**193**], we identified distinct selectivity determinants: for TRX-f2, a cluster of six electropositive residues exposed around the redox site (Figure 3.5A); and for TRX-h, residues enabling interaction with a broader range of substrates (Figure 3.5B). These predictions guided the rational design of a synthetic recoding of TRX-h, serving as a scaffold to graft the TRX-f2 selectivity motifs. The resulting synthetic sequence—engineered to exhibit type-f functionality—was experimentally validated and shown to recognize type-f

substrates with enhanced efficacy [**194**], as explained in Chapter 4.

In addition, we examined TRX-f2 interactions with various partners in *C. reinhardtii*, including ferredoxin-thioredoxin reductase (FTR), sedoheptulose-bisphosphatase (SBPase), fructose-1,6-bisphosphatase (FBPase), and Phosphoglycerate kinase (PGK). These interactions are particularly challenging to study as the structures of these complexes remain undetermined. Similarly to the BPTI case, we predicted $\Delta\Delta G$ values for all single-point mutations in TRX-f2 in complex with the considered partners (3.5C for the interaction with FTR). We compared the regions highlighted by this analysis (3.5D) and by the attention mechanism, with the different interaction surfaces identified by AlphaFold–Multimer (3.5EF). We found that the model identified the same interacting surface of TRX-f2, not only according to the MuLAN attention that do not depend on the partner, but also according to the $\Delta\Delta G$ for the different complexes. For the TRX-f2-SBPase complex, we generated the AlphaFold–Multimer and ESMFold complex structures (3.5D). Based on the mutational landscapes computed for both TRX-f2 and SBPase, MuLAN can easily indicate the most favorable complex, which in this case turns out to be predicted by ESMFold, due to the co-localisation of the interaction surfaces and the presence of necessary cysteines in the TRX-f2 binding region.

3.2.3 Computational details

Executing the MuLAN model requires neither accurate structures of the complexes nor the pre-extraction of specific features from the input, which could prove a bottleneck for structure-based approaches. On the contrary, the only required inputs are the sequences of the interacting protein complex, that are then tokenized and embedded in the PLM, and the specific mutations to evaluate. Once the embedding for mutated proteins are computed, the average evaluation time for the downstream model is around 0.015 seconds on a A100 GPU for one mutation, scaling linearly with the length of input sequences. The computation of embeddings is more time-consuming, leading to an average execution time of 1 second per mutation for the end-to-end model. However, for the reconstruction of entire mutational landscapes, time performances improve by pre-computing embeddings for the protein of interest in batches. In practical terms, generating embeddings and reconstructing full mutational landscapes takes about one hour for proteins with 1,000 amino acids, and around one minute for 200-residue proteins. By contrast, extracting attention weights (including embedding generation) for the entire human proteome takes only 40 minutes on a 24GB NVIDIA GeForce RTX 3090 GPU.

3.3 Methods

3.3.1 Sequence embeddings

The primary structure of a protein of length L is a sequence of amino acids $S = (a_1, \ldots, a_L)$. In order to use it as input to a deep learning algorithm, it has to be encoded in a numerical format. The most unbiased and straightforward way to encode sequences without any prior knowledge injected is by representing each amino acid as a 1-hot vector, where the

protein becomes a concatenation of L such vectors. In more recent years, the development of increasingly Large Language Models (LLMs) trained on protein sequences allows for contextualized embeddings. These protein Language Models (PLMs), often based on the transformer architecture [**80**], and capable of scaling up to billions of parameters, utilize a multi-head attention mechanism to capture long-distance correlations between amino acids. By predicting a fraction of masked residues from the context given by the rest of the sequence over millions of sequences and iterations, the model learns high dimensional representations (size H, typically of the order of 10^3) for each amino acid. Their concatenation defines the protein embedding, of dimension $L \times H$.

In many protein-level downstream tasks, such as solubility, localization or fitness prediction, an embedding that is independent of the length of the protein is required, which is usually obtained by global pooling over the length dimension [**108**]. Dimensional reduction techniques applied to the extracted embeddings reveal that the models effectively learn some local features of single amino acids, for example hydrophobicity, charge, polarity, size, and global features of the whole proteins, such as conserved regions, location in the cell, and even structure and function to some extent [**112, 179**]. Similarly to LLMs, the performances of PLMs in zero-shot or few-shot downstream tasks are a measure of the information contained in the extracted embeddings and, ultimately, of the effectiveness of the encoder model itself. Up to date, PLMs have been developed in order to encode (encoders) or generate (decoders) single protein sequences, with performances that generally scale with the size of the model, but how to combine these representations in order to encode and predict properties for a protein complex still remains a challenge, and may likely depend on the considered task.

3.3.2 A deep learning architecture with light attention

To extract and combine information from the embeddings of both wild-type and mutated proteins, we developed an architecture incorporating a Siamese encoder and a feature-combination module to perform the final regression. The model is structured into three parts: (i) a pre-trained protein language model with frozen weights, employed to extract contextualized embeddings from the input sequences; (ii) an attention block that processes the embeddings to generate vectors whose size does not depend on the length of the sequence; (iii) a fusion block designed to combine the encoded inputs for predicting the target property (B5).

The model processes the four input sequences independently, by applying a Light Attention Block [**195**] on each. The architecture is built upon two parallel 1D convolutional layers applied to each sequence (B5C). Of these layers, C_1 and C_2, one applies a dimensionality reduction of the input representation, cutting down the latent dimension (corresponding to the number of filters in convolution) from around 10^3 to 64, and the other is followed by a SoftMax function applied across the length dimension, thus assigning a normalized weight to each residue within each latent dimension separately. We proceed to perform a point-wise product of the two outputs, summing over the length dimension,

according to the equation

$$Y = \sum_{i=1}^{L} C_1(X) \cdot \text{SoftMax}_L(C_2(X))_i \tag{3.1}$$

The result is a local linear attention mechanism, the scope of which can be varied by modifying the kernel size k. Testing different values for k, we discovered optimal performance with blocks applied in parallel for $k = [1, 5, 9]$, and by joining their outputs together. Additionally, we applied a global MaxPooling over the length dimension for each convolutional head, which is then concatenated with the representation extracted directly from the Light Attention mechanism. The resulting flattened vector is then fed into a multilayer perceptron (MLP) layer equipped with a LeakyReLU activation function prior to entering the fusion block.

In the last layer, the outputs for the two interacting proteins are merged using both the dot product and the absolute value of the difference of the embedding vectors, in order to ensure that the resulting representation remains symmetrical regardless of the order in which the two proteins are considered. The final embeddings for both the wild-type and mutated complex are then subtracted from one another and fed into a linear layer to predict $\Delta\Delta G$. A schematic representation of the complete model is depicted in B5.

The training goal is to reduce the mean squared error (MSE) between the observed experimental values and the model's predictions. The evaluation of the models involves calculating the Pearson correlation coefficient (PCC), which assesses the linear correlation between the experimental outcomes and the predictions, and the root mean squared error (RMSE).

3.3.3 PLMs fine-tuning

To enrich the information content of the embeddings derived from PLMs, we designed two fine-tuning strategies using smaller-scale models: masked language modeling and binary classification at the single residue level. In the first approach, following [**157**], we adopted the same self-supervised task that was used for pre-training the model. However, we narrowed down the training dataset to sequences exclusively from the most frequently occurring species in our dataset, specifically *Homo sapiens*, *Bos taurus*, *Mus musculus*, and *Escherichia coli*. This strategy aimed to tailor the model's focus towards sequences more closely related to those found in the protein complexes under study. We named it "evo-tuning".

The second fine-tuning strategy, named "supervised fine-tuning", involves supervised binary classification, distinguishing whether residues are part of the protein's interacting surface. To label the training data, we utilized LEVELNET software [**196**] on all the proteins in the dataset, excluding those listed in SKEMPI v2.0 [**38**]. LEVELNET identifies interface regions from online databases for each biological entity, and maps them to similar sequences through homology transfer. Hence, for a protein, labels are defined on the union of the interface regions known to LEVELNET. Out of nearly 77,000 sequences evaluated, about 28,000 were found to lack an interacting surface, possibly due to insufficient data, and

were therefore excluded from the dataset. This process resulted in approximately 49,000 sequences, which were randomly divided into training and test datasets using an 80/20 split ratio. This approach is driven by the observation that mutations significantly impacting $\Delta\Delta G$ often occur at the interface. Incorporating this information into the embeddings is expected to enhance the training efficiency of the downstream model.

3.3.4 Extraction of Attention weights

The model's design is crafted to promote interpretability by retrieving attention weights from its layers. Outputs from the attention heads are extracted and concatenated along the position dimension, such that each residue is encoded in a vector of dimension $N \times C$, where N is the number of heads and C is the number of filters in the convolution (in our case, 3 and 64 respectively). By averaging over the obtained vectors, and applying MinMax normalization to the result, we allocate a singular scalar weight to each residue. This weight reflects the significance of that residue's representation in determining the final $\Delta\Delta G$ prediction.

3.3.5 Datasets

$\Delta\Delta G$ experimental data

The main source of our data is the SKEMPI v2.0 database, which is widely used in the context of the prediction of binding affinity changes upon mutations [**38**]. It is the most complete database for experimentally measured binding affinities for wild type and mutated complexes, for which the structure is solved and available in the Protein Data Bank (PDB) [**180**]. SKEMPI v2.0 reports measurements for 7085 single-point and multiple mutations obtained from 345 different complexes, and it includes other smaller databases such as AB-Bind [**197**], PROXiMATE [**198**], and dbMPIKT [**151**]. For each entry, the PDB structure of the wild-type complex is provided, as well as the names of the considered interacting partners in the complex, the mutated residue(s), the experimental methods employed for each measurement, and the associated affinities for wild-type and mutated complex, k_{wt} and k_{mut}. From them, ΔG can be computed as

$$\Delta G = -RT\ln(k) \tag{3.2}$$

where R is the molar gas constant and T is the absolute temperature.

The structural region of the mutation site is also reported, as belonging to one of the five classes COR, SUP, RIM (corresponding to core, support and rim regions at the interface), INT or SUR (for interior or surface) [**199**]. The tendency is towards higher variations in binding affinity for mutations located in the interacting surface, where most of them tend to be deleterious rather than beneficial.

The intrinsic asymmetry of the database must be considered during the model development. This can be addressed either by deliberately overlooking the asymmetry, resulting in a predisposition for positive predictions, or by augmenting the database with inverse mutations, as in [**158**], coupled with incorporating a distinct term in the loss function

to attain symmetry. Although this method can enhance performance in cross-validation assessments, it may lead to higher model redundancy, thereby reducing its ability to generalize to unseen data.

The SKEMPI v1.0 and v2.0 databases have been a source for numerous benchmark datasets used in recent years to train and evaluate a variety of models grounded in physics, statistics, and more recently, machine learning and deep learning approaches. We have also utilized some of these datasets for our training and evaluation, including S1400 featuring both single-point and multiple-point mutations, and S1102, S1131, S2003, and S4169 featuring only single mutations. The dataset names' numbers reflect the count of data points, with each representing a complex associated with a $\Delta\Delta G$ value linked to one or several mutations in one or both sequences. Table B1 reports the number of complexes, the number of mutations and the source of each dataset.

PDB Dataset, Interactome and Activome

To construct the Interactome dataset—capturing all experimentally resolved interaction types present in the PDB—for the large-scale evaluation of MuLAN attention scores, we relied on biological assembly data for all structures deposited in the PDB before 20-3-2024 that contained at least two entities, including proteins, nucleic acids and small molecules, and that had already been pre-processed for the construction of Boltz-1 datasets [**200**]. For each complex, we extracted the atomic coordinates of the biological entities, and computed distances between the C_α atoms of protein residues and the reference atoms of other entities. We used a distance threshold of 8Å to define positive interaction labels, and categorized interactions into three classes: interactions between two proteins, interaction between proteins and RNA or DNA, and interactions between proteins and small molecules (ligands). After removing all entries in which no interaction was found, the final dataset comprised 224,924 labeled protein chains derived from 123,624 PDB files.

The Activome dataset was subsequently derived from the PDB dataset by mapping each PDB chain to its corresponding UniProt entry and retrieving ligand-binding sites using the RCSB GraphQL API. Each residue was then assigned an integer value representing the number of distinct ligands observed to bind that position in any of its associated PDB chains. These procedure resulted in a collection of 32,297 UniProt sequences, 22,062 of which contained at least one experimentally determined ligand-binding site. Finally, we removed all sequences longer than 1024 residues, obtaining 20,696 entries for our analysis.

Enhancing PLMs with targeted fine-tuning

For the fine-tuning of PLMs, we leveraged datasets from two distinct sources. The dataset for masked language-modeling fine-tuning was sourced from STRING database [**201**], which reports over 20 billion experimentally measured and inferred protein-protein interactions across approximately 12,000 different organisms, amounting to more than 59 million protein sequences. We selected only proteins belonging to four species (*Homo sapiens*, *Bos taurus*, *Mus musculus* and *Escherichia coli*), resulting in a total of 66235 sequences for training and validation.

3.3.6 Implementation details

All the models, training and evaluation scripts are written in Python using the PyTorch library. Pre-trained PLMs are available in the HuggingFace Hub and can be easily downloaded following the related documentation.

For efficient model training, we pre-computed the embeddings from PLMs for every wild-type and mutated protein in the training dataset. The single embeddings were then loaded in batches and dynamically zero-padded to the longest sequence length in the batch.

All downstream models are trained until convergence, using AdamW optimizer with default parameters, a batch size of 32 and an initial learning rate of 5×10^{-4}, that is reduced by half on validation loss plateau (with patience of 5 epochs). The random seeds for dataset splitting in cross-validation experiments and the initialization of model parameters were set to 42 for reproducibility.

For PLMs fine-tuning, instead, we ran the training for 3 epochs with data parallelism across 4 A100 GPUs and gradient accumulation, with a total batch size of 256 and a learning rate of 5×10^{-5}, with AdamW optimizer. Training data for fine-tuning tasks were randomly split in training and validation set, following a 0.8 - 0.2 ratio.

3.4 Discussion

Understanding the physical interactions between proteins and other molecules is paramount for unraveling the molecular mechanisms that explain phenotypes. Our study focuses on protein-protein interactions leveraging deep learning to analyze sequence-derived protein representations, offering a robust alternative to structure-based methods. Deep learning enables us to tackle essential questions such as reconstructing interaction networks, identifying binding interfaces, and estimating changes in binding affinity. The sequence-based approach is especially advantageous for studying disordered and flexible regions involved in transient interactions, which are challenging to model with structural data. By extracting information from sequences, deep learning leverages evolutionary constraints that shape protein structure and function. MuLAN makes significant strides in this area, in three main ways: 1. it estimates the strength of interaction with binding partners, crucial for understanding disease susceptibility and drug efficacy. 2. it reconstructs a deep mutational scanning matrix for each interacting partner, incorporating information from both proteins. 3. it enhances the identification of residues in sequences that are involved in protein interaction sites, addressing the complex problem of varying interaction sites with different partners.

MuLAN brings protein interactions into sharp focus

By analyzing a pair of interacting protein sequences, and by coupling it with the same pair "perturbed" with mutations at one or more amino acids, MuLAN can estimate changes in binding affinity. It does it at large-scale since it does not require precise complex structures. This approach enables the reconstruction of mutational effects on each protein and identify interaction sites, providing deeper insights into protein function

and interactions. Notably, the mutational landscape of a protein varies when analyzed with different partners, demonstrating MuLAN's ability to capture distinct interaction details for each protein pair. This capability closely reflects biological contexts, enhancing predictive accuracy beyond existing computational methods.

Crucially, our large-scale analyses demonstrate that MuLAN's predictive power is evolutionarily stable. Even when applied to proteomes separated by hundreds of millions of years of divergence, the model maintains consistent accuracy in identifying interface residues (mean AUC > 0.6 across all taxa). This indicates that the learned sequence representations and attention mechanisms capture deeply conserved biochemical and structural principles underlying protein interactions. In effect, MuLAN's internal representations act as a computational analogue of the evolutionary fitness landscape, where residue relevance reflects long-term selective constraints that govern molecular recognition.

The significance of MuLAN is twofold. Firstly, it enhances the identification of residues in sequences that are *likely to be* part of protein interaction sites, offering *a fast and structure-agnostic* method to address a major challenge in the field. Secondly, and most importantly, MuLAN introduces an innovative approach to reconstructing a deep mutational scanning matrix that considers protein interactions and incorporates information from both interacting partners. This hypothesis, previously unexplored, brings us closer to accurately modeling biological contexts. Unlike existing computational methods that overlook specific partner interactions, limiting their predictive capabilities, our approach identifies key differences by incorporating these interactions, thereby enhancing predictive accuracy and advancing our understanding of protein dynamics.

In summary, our study represents a significant advancement in the field of protein interaction research. By combining deep learning with sequence-based analysis, we provide a more comprehensive and accurate understanding of protein interactions and their underlying mechanisms. This not only enhances our ability to predict binding affinity changes but also opens new avenues for studying the complex nature of protein interactions in biological systems.

High-throughput proteome-wide analyses with MuLAN

Owing to MuLAN's remarkable efficiency—extracting attention weights, including embedding generation, for the entire human proteome in just 40 minutes on a single 24GB NVIDIA GeForce RTX 3090 GPU—and its broad applicability across arbitrary species, the model offers a powerful and scalable means to identify residues critical for functional activity and binding in any proteome. This positions MuLAN as a practical tool for rapid, proteome-wide functional annotation and comparative analysis across diverse organisms.

Challenges in $\Delta\Delta G$ prediction accuracy

As it is usually the case for machine learning regression models, extrapolation is challenging, as it requires the model to generalize well beyond the training data, and it is particularly true in the case where the training set is limited. In our endeavor to predict $\Delta\Delta G$, the training dataset predominantly feature wild type and mutant complexes with binding

differences under 10 kcal/mol (absolute value), with most data points below 4 kcal/mol and a handful of outliers. Additionally, there is a notable experimental preference for data on deleterious mutations over beneficial ones, leading to data asymmetry. These two factors contribute to the model's inclination to predict lower positive values for unseen data, potentially misrepresenting the true $\Delta\Delta G$ values. Nonetheless, we anticipate accurate ranking of predictions, where higher values signify deleterious mutations, and lower values indicate neutral or beneficial mutations. The pooling with other orthogonal methods, as we did here with the addition of iGEMME scores or, as recently shown [**202**] with AlphaFold3 ranking scores [**203**], may help mitigate this limitation.

3.5 Data and Software Availability

MuLAN is available at `http://gitlab.lcqb.upmc.fr/lombardi/mulan` under the CC-BY-NC-SA 4.0 License. Predictions of attention scores for the human proteome are available at `https://zenodo.org/records/15719109`.

Chapter 4

Biological applications of MuLAN

The ability of MuLAN attention scores to capture diverse biological signals in proteins from their sequence only led to multiple applications to various biological systems, where they served as an *in silico* validation method, alone or combined with other modelisation tools, or as an intermediate step in complex pipelines.

In this chapter we report on these different contributions, providing in each case a brief summary of the relevant biological context and highlighting the role of MuLAN in the specific case and its interplay with additional models.

4.1 Thioredoxin determinants in *Chlamydomonas reinhardtii*

In this work, in collaboration with collegues from the Synthetic and Systems Biology of Microalgae group in the laboratory, we applied MuLAN for the identification of functional determinants of specificity of thioredoxin TRXf2, a protein expressed in the chloroplast of the microalga *Chlamydomonas reinhardtii*, towards one of its targets. The results were then tested *in vitro* by the experimental team, who designed the functional switch of another type of thioredoxin, specifically the TRXh, by expressing the computationally identified functional determinants in its sequence. As a result, the recoded hybrid TRXh gained selectivity towards the same protein targeted by TRXf2. This section is based on the article "Functional recoding of *Chlamydomonas reinhardtii* thioredoxin type-h into photosynthetic type-f by switching selectivity determinants", published in *Frontiers in Plant Science* in 2025 [**194**].

4.1.1 Biological context

Thioredoxins (TRXs) are a universally conserved family of small redox-active proteins that act as central regulators of thiol-disulfide exchange in diverse biological systems. Their fundamental role in catalyzing the reversible reduction of disulfide bonds makes them essential for maintaining protein redox balance, responding to oxidative stress, and modulating redox-sensitive signaling pathways [**204, 205**].

Structurally, TRXs adopt a compact, globular fold, which consists of a central mixed β-sheet of four strands arranged in the order 4-3-1-2, flanked on both sides by two α-helices, forming a three-layered $\alpha/\beta/\alpha$ sandwich [**206**], that can be recognized in Figure 4.1. This fold is not only conserved across all TRXs but is also shared with related oxidoreductases

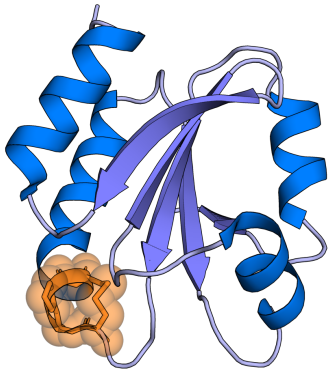

Figure 4.1: Structure of the conserved TRX fold, here displayed for TRXf2 in *Chlamydomonas reinhardtii* (PDB 6I1C). Secondary structure elements are highlighted with different colors, with the active site CGPC highlighted in orange.

such as glutaredoxins and protein disulfide isomerases, reflecting its evolutionary success as a redox-active scaffold.

The active site cysteines are surface-accessible and positioned within a loop between a β-strand and an α-helix. In the reduced state, the thioredoxin active-site tetrapeptide CGPC is exposed to the solvent, enabling the first cysteine residue to engage in the reduction of disulfide bonds in target proteins [**207**].

The *Viridiplantae* lineage has undergone notable diversification of its TRX gene family, giving rise to a seemingly complex array of isoforms [**208**]. In the chloroplast of the model green microalga *Chlamydomonas reinhardtii*, distinct types of TRXs are encoded and expressed: TRXf1, TRXf2, TRXm, TRXx, TRXy, and TRXz [**209**]. Structural comparisons based on NMR spectroscopy, crystallographic studies, and computational modeling have confirmed a high degree of structural conservation across *Chlamydomonas reinhardtii* TRX isoforms, while also revealing variations in surface electrostatic potential distribution [**191,210,211**]. These localized electrostatic differences support the hypothesis that TRX-target specificity may be governed by a "selectivity code" based primarily on electrostatic complementarity [**212**].

Functionally, several chloroplast enzymes involved in photosynthetic metabolism are regulated via redox-dependent activation by TRX-mediated disulfide bond reduction [**213,214**]. This includes key enzymes of the Calvin-Benson-Bassham cycle (CBBC) responsible for carbon fixation, whose activity is tightly linked to the reducing power generated by photosystems under light conditions [**215**–**217**]. Among these, fructose-1,6-bisphosphatase (FBPase) and sedoheptulose-1,7-bisphosphatase (SBPase) have been the subject of detailed *in vitro* studies demonstrating the preferential interaction and reduction by TRXf isoforms [**218**]. This electrostatic-based specificity is further supported by proteomic evidence showing enrichment of FBPase in TRXf-based affinity purification experiments [**219**], reinforcing the view that electrostatic complementarity underlies selective TRX-target

interactions in the chloroplast.

In this study, our objective was to properly identify the functional determinants of TRXf specificity for FBPase, and test them in a synthetically recoded TRXh.

4.1.2 Results

Designing enzymatic switches presents a considerable challenge, primarily due to the difficulty in identifying the residue signatures that govern enzymatic activity and substrate specificity. To address this, we employed ProfileView, a computational framework designed to extract functional signatures from protein sequences [**193**]. Applied to the TRXf2 isoform, ProfileView analyzes a diverse repertoire of naturally occurring thioredoxin (TRX) sequences to identify functionally distinct subgroups. By leveraging evolutionary information embedded in natural sequence variation, ProfileView classifies TRX proteins based on inferred functional properties.

In this study, ProfileView was applied to a dataset of over 10,000 TRX sequences, leading to the identification of eight major functional classes, each predicted to represent distinct enzymatic roles. Among these, one class comprising 301 sequences included all known type-f TRXs from different species, along with several characterized type-o and type-h isoforms. Figure 4.2A illustrates the functional tree generated by ProfileView, depicting the organization of these 301 sequences within a functional, non-phylogenetic sequence space. Within this tree, all annotated type-f sequences formed a compact cluster, from which a subgroup of 34 type-f-like sequences was delineated.

From this subset, ProfileView revealed a conserved electropositive surface motif in CrTRXf2, comprising six lysine residues (K78, K106, K126, K134, K143, and K164) that form a positively charged crown surrounding the TRX catalytic site (see sequence logo in Figure 4.2B and structural model in Figure 4.2C). These lysines were prioritized as candidate positions for mutagenesis in the design of our functional switch.

To refine the selection of residues for targeted mutation, we applied MuLAN to the CrTRXf2-CrFBPase interaction, generating a full mutational sensitivity map for CrTRXf2 (Figure 4.3A). High-impact positions—those predicted to most affect binding—were concentrated in the protein core, in the active site motif (WCGPC), and in a proximal PTF motif (residues 146-148) near the active site.

The averaged position-wise scores from MuLAN were mapped onto the crystal structure of CrTRXf2 (PDB: 6I1C) by averaging the predicted impact of all possible mutations at each position, as shown in Figure 4.3B. Notably, the six lysine residues identified via ProfileView also received high sensitivity scores in the MuLAN analysis, indicating their functional relevance and mutational sensitivity. These lysines were consistently scored higher than surrounding residues, further supporting their role as key determinants of function and validating their selection for experimental mutagenesis. As an independent validation of the MuLAN results, we used AlphaFold3 to predict the structure of a binary complex between TRX and FBPase (Figure 4.3C), where highlighted regions according to MuLAN attention scores are indeed predicted to be in contact.

ProfileView and MuLAN confirmed the previously documented electropositive character

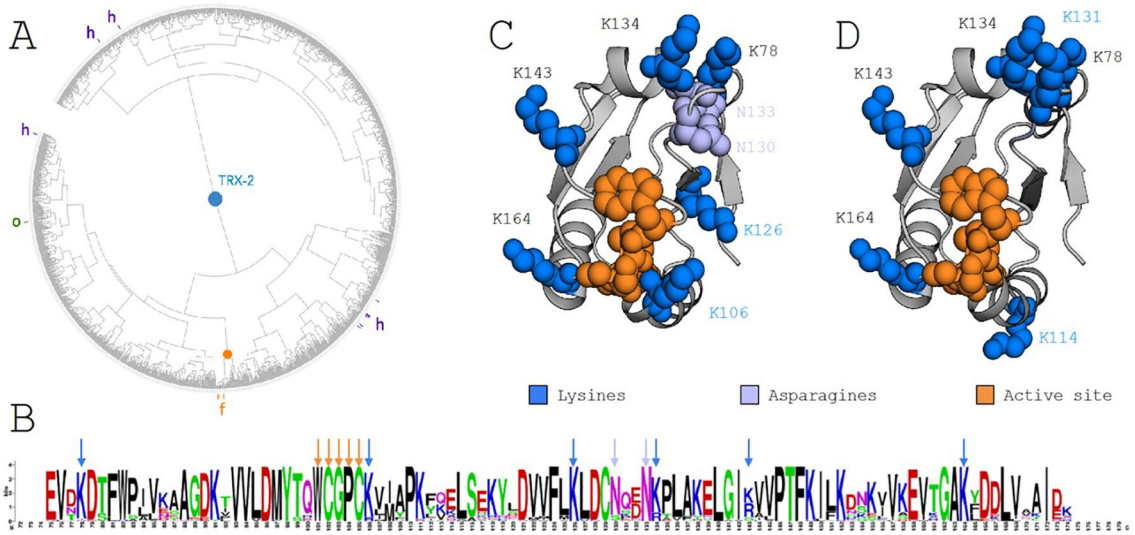

Figure 4.2: **ProfileView analysis of the thioredoxin family and the thioredoxin type-f subfamily**. (A) ProfileView tree of the TRX subfamily, comprising 301 sequences. The orange root marks the position of known TRX type f (TRXf2) sequences from C. reinhardtii. The tree was visualized using iTOL (`https://itol.embl.de`). (B) Motif derived from the alignment of sequences within the subtree highlighted by the orange root in (A), generated using WebLogo (`https://weblogo.berkeley.edu/logo.cgi`). Orange arrows denote the conserved binding motif WCGPC. Blue arrows highlight conserved lysines (K) present in the *C. reinhardtii* TRXf2 sequence, while purple arrows indicate two conserved asparagines (N). (C) The TRXf2 structure (PDB 6I1C) showing residues involved in the functional determinant of the C. reinhardtii sequence. The extended disulfide bond motif (WCGPC) is highlighted in orange. A positively charged crown surrounding the active site, composed of six lysines (blue) and two asparagines (purple), is shown, with all other residues in gray. (D) Residues mutated in the switch experiment are highlighted in blue. The extended WCGPC motif (orange) indicates the location of the mutated residues forming the positive electrostatic crown. Note that K114 and K131 were experimentally mutated in place of K106 and K126, as shown in (C).

that correlates with f-type TRX [**218, 220**]. We substituted 5 lysines defining TRX f-type onto cytosolic TRXh1, at the positions identified by computational analyses and in agreement with visual comparison of the crystal structures of CrTRXf2 and CrTRXh1 [**191, 221**]. After purification of resulting TRXχ1, a chimeric TRXh holding the f-type recognition module, we could demonstrate *in vitro* that these mutations enable TRXχ1 to functionally mimic CrTRXf2 in FBPase redox activation, as shown in Figure 4.4, that reports experimental activation of CrFBPase under four different reducing mixes, with respect to the untreated case. The different mixes contained the reducing agent dithiothreitol (DTT), that slightly activates FBPase alone, and TRXf2/TRXh1/TRXχ1, in separate assays. Details on experimental conditions for reproducibility are reported in [**194**]. This minimally reconstituted system provides the simplest evidence that electropositive surface around the nucleophile cysteine is the actual functional determinant of f-type TRX. Interestingly, TRXχ1 appeared more efficient than wild-type CrTRXf2 for CrFBPase activation,

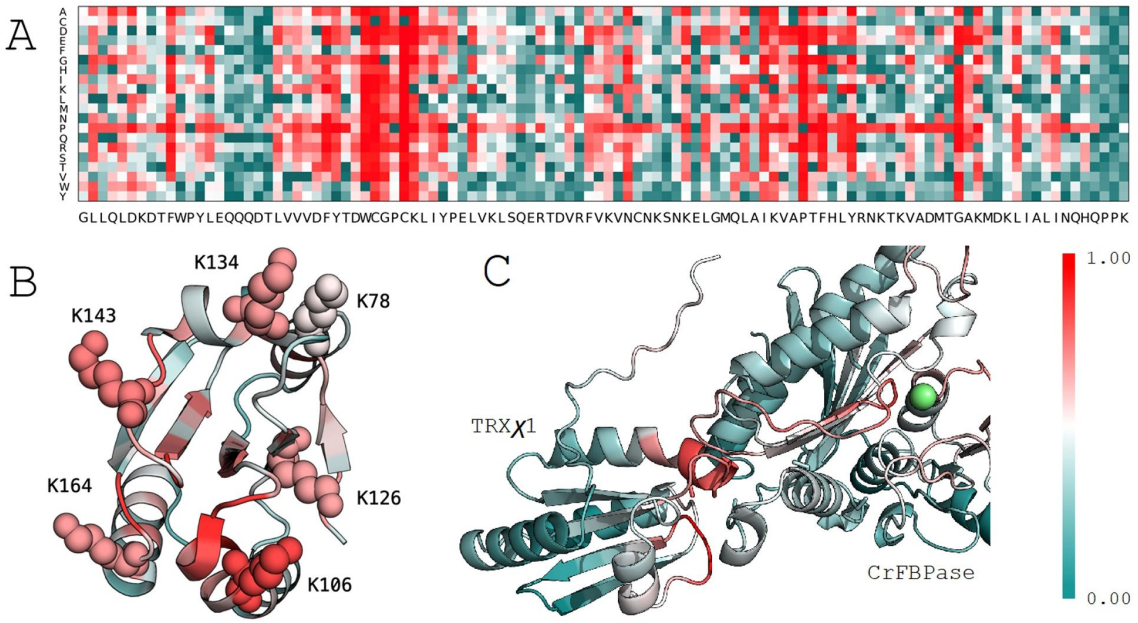

Figure 4.3: **MuLAN analysis of thioredoxin type f (CrTRXf2) in *C. reinhardtii*.** (A) MuLAN reconstruction of the mutational landscape for CrTRXf2. Each residue in the sequence is mutated, and the predicted effect is scored and visualized with a color gradient (red: high effect, green: no effect). (B) Highlighting of the six lysines identified by ProfileView, with their corresponding MuLAN scores, averaged by position. (C) MuLAN attention scores mapped onto the AlphaFold structural model of the TRX-FBPase complex. The magnesium ion is depicted in green.

suggesting that natural TRX may not operate at biochemical optimum.

4.1.3 Discussion

This work highlighted the utility of combining ProfileView, a tool for functional classification of protein sequences, and MuLAN, designed to predict variations in binding affinity and interacting residues, in identifying crucial functional signatures for protein activity, in this case in the selective recognition of a cognate protein surface.

By mapping residues in a target sequence to those conserved within functionally similar sequences identified by ProfileView, we pinpointed key lysines essential for the function of CrTRXf2. This demonstrates not only the effectiveness of ProfileView in elucidating protein-specific features, but also its general potential for studying proteins with other diverse and specialized functions. While CrTRXf2, a well-characterized type-f sequence, served as our model, the broader thioredoxin family and in particular other *Chlamydomonas reinhardtii* TRX sequences remain relatively unexplored (ie. TRXx, TRXy, TRXz) but could extend beyond in the thioredoxin superfamily (glutaredoxins, TRX-like, etc.).

MuLAN confirmed the importance of the six lysines identified by ProfileView and offered a mechanistic perspective by linking these residues to their role in the recognition of FPBase canonical target. Integrating the two computational methods allowed to derive

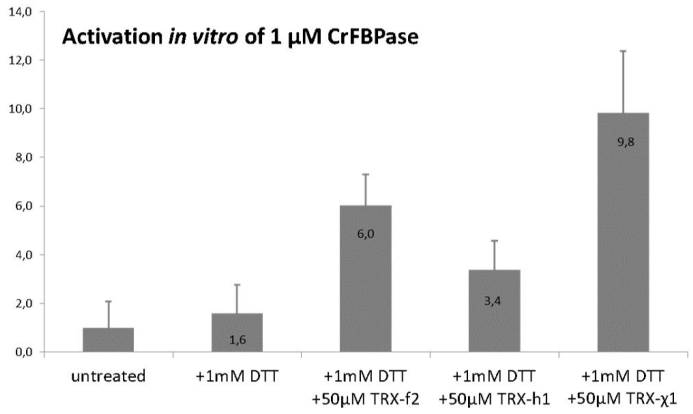

Figure 4.4: Experimental results of FBPase activity under redox treatments, measured as relative enrichment with respect to untreated activity. The reducing solutions included 1 mM dithiothreitol (DTT), 1 mM DTT and 50 µmol/L of CrTRXf2/CrTRXh1/TRX*chi*1, respectively. Error bars report standard deviation over 4 replications of the same experiment.

comprehensive insights into protein structure-function relationships. Looking ahead, the synergy between these methods provides a promising framework for studying not only enzymes but a wide range of proteins.

A practical expansion of our computational/biochemical approach is to achieve the identification of all TRX types on the register of their redox targets, and to test in vivo each functional determinant on a generic TRX-folded protein chassis. Further research might focus on expanding these methods to more complex protein systems, integrating experimental validation, and applying them to other protein families to deepen our understanding of functional determinants and interaction mechanisms.

4.2 SFBB and S-RNase interactions in *Malus* and *Petunia*

This research project was conducted in collaboration with the Amandine Cornille's group of the New York University of Abu Dhabi, with the objective of characterizing the complex interactions between different haplotypes of S-RNase and SFBB proteins encoded in the S-locus of wild and domesticated apples (*Malus*). We also compared the results with those for *Petunia*, a better characterized organism that also contains a similar self-incompatibility (SI) system, discovering similarities in the patterns for the two systems.

For this work, we combined MuLAN attention scores for the identification of functional residues with structure prediction models, such as Boltz-1, a folding model based on denoising diffusion processes, to generate and inspect the 3D structures of experimentally unresolved complexes [**200**]. The related article, "Genomic architecture of the self-incompatibility locus in apple provides insights into the evolution of collaborative non-self recognition", is accessible on *BioRxiv* [**222**].

4.2.1 Biological context

In hermaphrodite plants, self-incompatibility (SI) is a widespread genetically-encoded mechanism that promotes outcrossing by preventing self-fertilization [**223, 224**]. These systems generally rely on the interaction between a pistil-expressed molecule and one or more pollen-expressed factors, which together mediate compatibility or rejection between gametes. SI systems occur in major plant families, including crops like potatoes [**225**], cultivated pear and apple trees [**226**], or coffee [**227**].

The SI system rely on interactions between a pistil-expressed S-RNase (a ribonuclease that degrades RNA in pollen tubes) and multiple pollen-expressed F-box proteins, which act as non-self antitoxins [**228**–**230**]. These proteins specifically recognize and neutralize non-self S-RNases through the ubiquitin-mediated proteasome pathway, allowing compatible cross-fertilization while blocking self-fertilization [**229, 231, 232**]. The pollen-expressed genes have different names depending on the lineage, for instance, SLF (S-locus F-box) in *Petunia*, or SFBB (S locus F-box brothers) in *Malus*. The tightly regulated interaction ensures that pollen from genetically similar plants is rejected, thus reinforcing outcrossing and maintaining population-level diversity.

The pollen and pistil-expressed genes are physically clustered within a genomic region called the S-locus, one of the most polymorphic loci described in flowering plants due to long-term balancing selection and suppressed recombination [**233, 234**]. Each S-haplotype is defined by a specific combination of pistil and pollen determinants, inherited together as a single unit. At the core of SI systems lies the S-allele, which serves as a genetic identity tag determining mating compatibility.

The objective of the project was to leverage a manually annotated dataset of 27 haplotype-phased *Malus* genomes (13 wild and 14 cultivated) to investigate the genomic architecture and evolution of the S-locus in apple tree. In particular, we focused on predicting the genome regions responsible for self-incompatibility specificity between SFBB and S-RNase across S-alleles.

4.2.2 Results

S-RNase polymorphism has a central role in determining specificity, with a known hypervariability of this gene sequence across S-alleles. Collaborators identified 8 codon positions as evolving under positive selection and simultaneously conserved within each S-RNase allele. Inference was run using CodeML from the PAML package [**235**], and identified residues had posterior probability higher than 0.95, based on Bayes Empirical Bayes analysis. These residues represent strong candidates for driving functional divergence between alleles, and were distributed along the entire S-RNase coding sequence, suggesting that allele-specific interactions may involve multiple contact points. Several of these positively selected sites did not overlap with the hypervariable regions previously predicted to mediate recognition specificity on the protein surface [**236, 237**]. We run Boltz-1, to predict the tertiary structure of different S-RNase, using MMSeqs2 server for MSA retrieval and 5 different noise initializations for each prediction. The predicted structures revealed that most positively selected residues are nonetheless surface-exposed, indicating potential

accessibility for molecular interactions. However, they were spread across the surface, suggesting that their functional relevance might not be related to specificity towards SFBBs only. A similar conclusion might also be derived from attention score predictions with MuLAN, that did not particularly highlight the identified positions (Figure 4.5A).

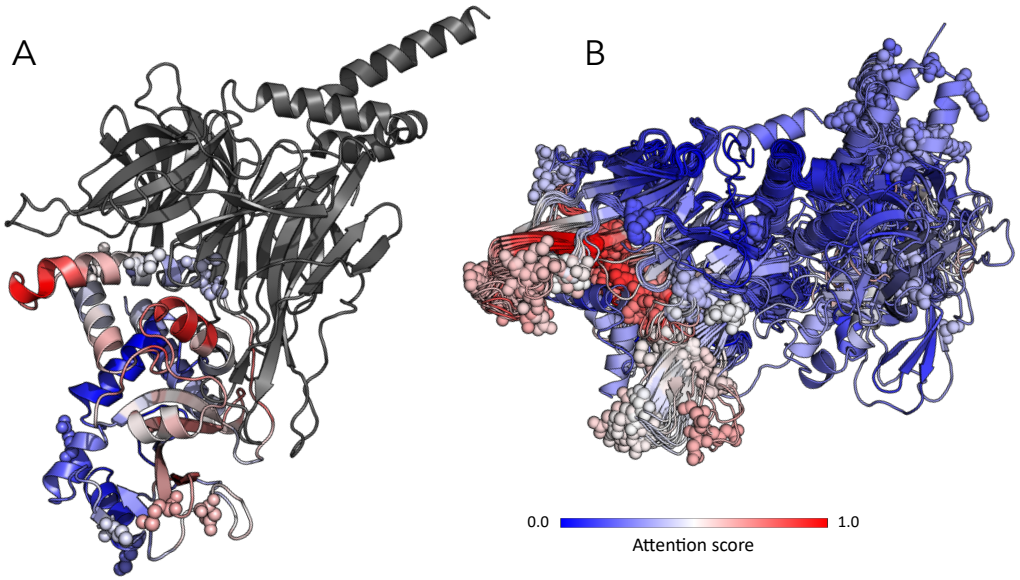

Figure 4.5: **Structural analysis of S-RNase and SFBB in *Malus*.** (A) Boltz-1 prediction of representative S-RNase protein structure (MdS2, S-haplotype 14), showing the typical α/β fold, in interaction with a SFBB from the SFBB2 family, in gray. The spheres indicate residues inferred to be under positive selection across S-haplotypes, while colors are assigned according to MuLAN attention scores. (B) Aligned Boltz-1 predictions of representative SFBB proteins for each S-haplotype, with highlighted positive selection sites and MuLAN attention scores coloring.

An analogous analysis of SFBB families revealed higher confidence in predicted structures, with positive selection sites found on two opposite sides of the protein surface (Figure 4.5B). Again referring to the Figure, positive selection sites highlighted on the left side were consistent across S-haplotypes, while those on the right side were variable and had lower confidence in CodeML analysis. Interestingly, the same highlighted surface was recognized as functionally important by MuLAN in all the considered structures, with small differences between them. In addition, the region was also identified by Boltz-1 as the interacting surface, when S-RNAse–SFBB complexes were considered.

To identify the SFBB family likely responsible for targeting a given S-RNase, we first examined which SFBB families were consistently absent across all haplotypes carrying the same S-RNase allele. For example, SFBB2 was absent from all MdS2 haplotypes (h14, h13, h3, h10, and h23), and SFBB11 was missing from all MdS4 (h20 and h22) and MdS5 (h2 and h26) haplotypes. These patterns suggested that the corresponding absent SFBB families could serve as the pollen-side determinants that recognize and detoxify these S-RNases. To refine this inference, we further investigated whether amino acid variants deviating from family-level conservation patterns could be associated with specific S-RNase

alleles. In total, each S-RNase allele was associated with the absence or mutation of at least one specific SFBB family (Table 4.1).

S allele (S-RNase)	Putative targeting SFBB family (supported by)
MdS2 (5 haplotypes)	SFBB2 (absence from all haplotypes), SFBB3 (consistent AA variations on sites 299, 336, 421 and 476)
MdS4 (2 haplotypes)	SFBB11 (absence from all haplotypes)
MdS5 (2 haplotypes)	SFBB11 (absence from all haplotypes)
MdS9 (3 haplotypes)	SFBB12 (consistent AA variation on site 543), SFBB13 (consistent AA variation on sites 297, 336, 543, 549), SFBB16 (consistent AA variation on site 336)
MdSf (2 haplotypes)	SFBB5 (consistent AA variation on site 336), SFBB7 (consistent AA variation on site 476)
MdSe (2 haplotypes)	SFBB9 (absence from all haplotypes)

Table 4.1: **Putative targeting SFBB families determined for each S-RNases in the *Malus* dataset.** Conclusions are based on the absence of SFBB families or on SFBB proteins mutations in S alleles.

To evaluate the predicted interactions between specific SFBB and S-RNase pairs, we used Boltz-1 model to compute confidence scores. Our goal was to compare scores between pairs hypothesized to interact and those not expected to do so. However, confidence scores (pLDDT, pTM, i-pLDDT, i-pTM) yielded low confidence for *Malus* (pLDDT < 0.6, pTM < 0.3), precluding clear discrimination of expected versus unexpected interactions, as predicted structures were not reliable. In particular, low scores were attributed to S-RNase in the complexes, while structures for SFBBs were generally reliable (pLDDT > 0.8).

To assess whether Boltz-1 is able to capture biologically relevant interactions in a comparable system, we applied the same analysis to the SI system of *Petunia*, where interactions between SLF (S-locus F-box) proteins and S-RNases have been experimentally validated [**229**]. In this case, confidence scores were globally higher (pTM > 0.6, Figure 4.6A), and the highest scores corresponded to experimentally supported interactions. For example, SLF-1 has been shown to interact with S9 and S17 S-RNases; SLF-2 with S9, S11, and S19; SLF-3 with S7; and SLF-9 with S19 [**229**]. Consistent with these findings, Boltz-1 predictions yielded the highest confidence score for the SLF-1/S9 S-RNase pair. Furthermore, positive selection sites, detected in Petunia S-RNase sequences, are not only located at the surface of the tertiary protein structure, but also overlap with higher attention scores from MuLAN model (Figure 4.6B).

Taken together, our results highlight the complexity of S-RNase functional diversification and its interplay with SFBB gene evolution. While several codon positions show signatures of positive selection and are conserved within S-RNases, their contribution to allele-specific recognition remains only partially captured by sequence-based models such as MuLAN. Similarly, although patterns of SFBB family absence or consistent amino acid divergence suggest targeted evolution driven by S-RNase alleles, *in silico* interaction predictions using Boltz-1 failed to clearly differentiate functional SFBB-S-RNase pairs in *Malus*. In contrast,

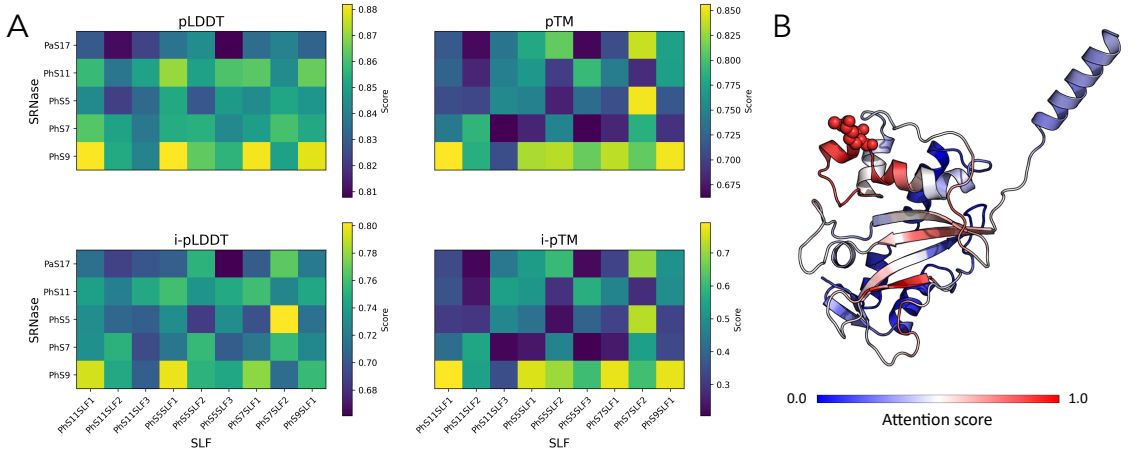

Figure 4.6: **Structural analysis of SRNase–SLF interactions in *Petunia*.** (A) Confidence scores (pLDDT, pTM, i-pLDDT and i-pTM) for Boltz-1 predictions of S-RNase and SLF complexes (see Section 2.6 for metrics definitions). (B) Predicted structure of a representative S-RNase, with the only positive selection residue highlighted as spheres. Color scale reflects MuLAN attention scores.

the same predictive approach successfully recovered known SLF-S-RNase interactions in *Petunia*, a species with a comparable gametophytic self-incompatibility system. While this discrepancy might be primarily attributed to a reduced experimental characterization of the *Malus* system, that affects the confidence of folding models, it may also reflect differences in structural or regulatory mechanisms underlying pollen-pistil recognition in *Malus*, potentially involving more subtle or indirect modes of interaction. One possible avenue is the involvement of long non-coding RNAs (lncRNAs), whose presence has been demonstrated in the intron of S-RNase in *Fragaria viridis*, suggesting a role in regulation of S-RNase activity [**238**]. These findings underscore the need to integrate structural, evolutionary, and regulatory layers to fully unravel the molecular logic of self-incompatibility in *Malus*.

4.3 Positive selection in aphid gustatory and olfactory receptors

This last project was again conducted in collaboration with the Amandine Cornille's group of the New York University of Abu Dhabi. In this case, we used MuLAN for *in silico* prediction of positive selection sites in aphid receptors, that were individuated from sequences analysis.

The related article, titled "Comprehensive annotation of olfactory and gustatory receptor genes and transposable elements revealed their evolutionary dynamics in aphids" has been accepted for publication in *Molecular Biology and Evolution* [**239**].

4.3.1 Biological context

Understanding the genomic basis of parasite adaptation to their hosts offers key insights into the evolutionary mechanisms driving population adaptation [**240–242**] and can inform strategies to address challenges such as parasite outbreaks [**243, 244**]. Such investigations require identifying both the number and nature of genes involved, as well as their evolutionary trajectories, which may be shaped by genetic drift [**245, 246**] or natural selection [**247, 248**]. Increasing attention has also been given to the role of transposable elements (TEs) in generating variation at adaptive loci and fueling host-parasite evolutionary arms races [**249–252**].

While TE-driven genome evolution is well documented in plant pathogens, its role in shaping genes underlying insect-plant interactions is less understood [**241,253,254**]. Insects exhibit extensive variation in both TE content and chemosensory gene repertoires [**255, 256**]. Several chemosensory gene families are critical for host recognition in herbivorous insects, with olfactory (OR) and gustatory receptors (GR) acting as key signal-transducing components in olfactory and gustatory pathways.

Positive selection in OR and GR genes supports a role for host-driven adaptation, as in *Spodoptera* species [**257**]. Similarly, hemipterans such as the pea aphid show rapid OR and GR expansions with evidence of positive selection, consistent with host specialization [**258**]. In contrast, in dipterans such as *Drosophila*, OR and GR diversity appears shaped mainly by drift, likely due to endemism and smaller effective populations [**259**]. Despite these advances, studies on OR and GR evolution remain concentrated on a few model species. Progress is further constrained by high sequence divergence across and within species, which hampers automated annotation and necessitates extensive manual curation.

In this study, a newly assembled genome of the rosy apple aphid (*Dysaphis plantaginea Passerini*) and 11 publicly available aphid genomes were used to systematically annotate OR and GR genes, classify TEs and examine their associations. Our contribution was to offer alternative *in silico* predictions based on MuLAN analysis to evaluate the functional potential of OR and GR residues under selection.

4.3.2 Results

To evaluate the selective pressure on GR and OR genes, 365 OR amino-acid sequences and 381 GR amino-acid sequences from the 12 considered genomes were first classified in subfamilies, building Maximum Likelihood trees with IQ-TREE v2.0 [**260**]. The GR genes clustered into 12 subfamilies and the OR genes clustered into 24 subfamilies respectively, some of which corresponded to previously known groups (Figure 4.7).

The high number of OR subfamilies relative to GR subfamilies and the star-like topology of the phylogenetic tree suggested that OR genes are evolving faster than GR genes or may have diverged more recently, as previously suggested [**261, 262**].

CodeML program from PAML package v4 [**235**] was used to estimate the substitution rate and the ratio of nonsynonymous (dN) to synonymous (dS) substitutions of the protein sequences of the 12 and 24 GR and OR gene subfamilies. Results showed that, of the 381 GR sequences analyzed, 57.7% evolved under purifying selection, 21.62% under positive

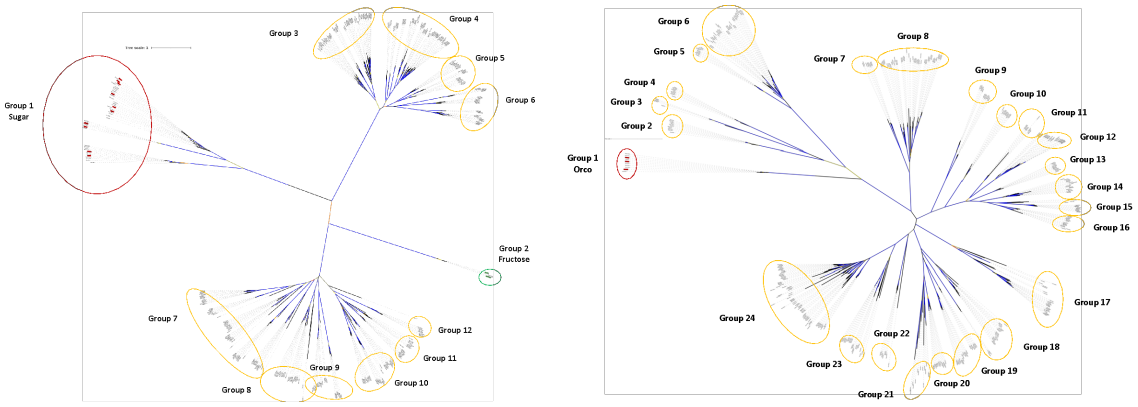

Figure 4.7: Phylogenetic trees and detected subfamilies for gustatory receptors (GR) on the left and olfactory receptors (OR) on the right. Trees were obtained using IQ-TREE v2.0 software on 381 and 393 sequences from considered aphid genomes, detecting 12 and 24 subfamilies, respectively. Known sugar and fructose groups and conserved odorant co-receptor (Orco) group are highlighted with distinct colors.

selection, and 20.61% neutrally. Of the 365 OR sequences, 46.11% evolved under purifying selection, 25.47% under positive selection, and 27.34% under neutral evolution. Additionally, the signatures of selection were not OR/GR subfamily-specific: each subfamily presented all the three signatures of selection.

When compared to other genomic regions, the OR and GR genes exhibited significantly weaker purifying selection. The results reveal that GR and OR genes mainly evolve under purifying selection, with nonsignificant differences in selection pressure on OR genes than compared onto GRs. A negative correlation was also observed between the dN/dS ratio and divergence among species, indicating that genes with the lowest divergence were under the highest evolutionary rates. This, together with evidence of recent positive selection in 4% of OR and 37% and GR genes, suggests that these genes undergo rapid episodes of positive selection, followed by sustained periods of purifying selection, which aligns with a pattern of diversifying selection.

In particular, three GR (GR03, GR04, GR06) and one OR (OR20) subfamilies consistently showed signatures of positive selection in CodeML models. We compared the identified sites with those highlighted by MuLAN attention scores, obtained from single sequences analysis, and we observed that positively selected sites consistently displayed significantly higher attention scores, except in the OR20 subfamily (Figure 4.8A), compared to neutrally evolving sites.

Notably, regions under positive selection in GR families overlapped with high-attention surface patches (Figure 4.8B), suggesting adaptive diversification of functionally important residues. In contrast, OR20 exhibits fewer positively selected sites and a more diffuse distribution of functional patches, potentially reflecting weaker adaptive constraints or insufficient statistical power due to the limited number of positively selected sites. This significant association suggests that diversifying selection preferentially targets functionally essential residues in GR receptor genes, supporting a mechanistic link between adaptive

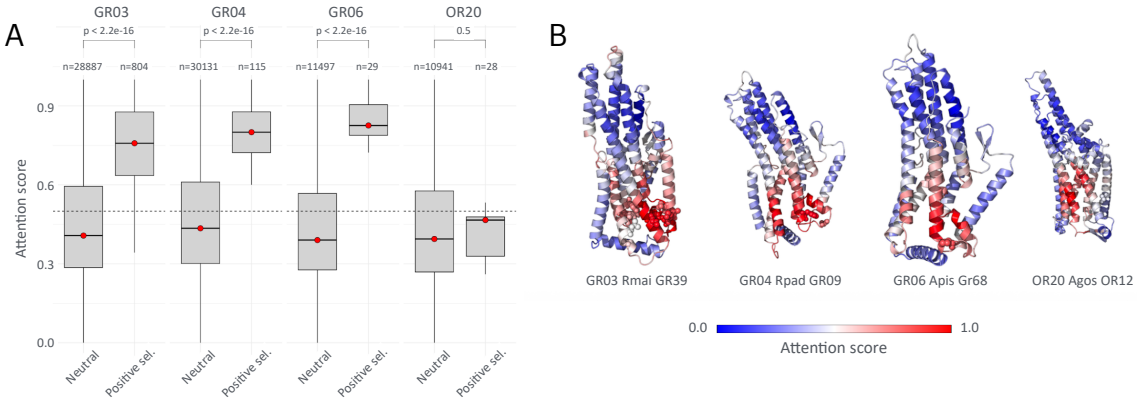

Figure 4.8: **MuLAN attention scores and positive selection sites in aphid gustatory and olfactory receptors.** (A) Distributions of MuLAN scores over receptors families with strong selection signals, for residues targeted by positive selection or neutral. (B) Predicted structures for representative sequences in each family with Boltz-1 model; structures are colored according to MuLAN attention scores, and positive selection sites are highlighted by sphere representation.

evolution and receptor functional diversification.

To investigate the role of TEs in shaping the diversity of GR and OR genes in aphids, TE enrichment in their vicinity was also assessed. TEs were detected using the LOLA tool [**263**], and their relationships to gene locations were evaluated using TEgrip [**264**]. A stronger enrichment of TEs near OR genes compared to GR genes was detected when considering 2 kb and 10 kb flanking regions. In addition, TE age was estimated via Kimura distance as a proxy for sequence divergence [**265**], finding that TEs associated with OR genes were significantly younger than those linked to GR genes or other regions of the genome.

However, MuLAN analysis did not reveal any significant differences between TE-associated and non-TE-associated genes, which is consistent with evidence of purifying selection detected across TE-associated OR and GR genes. This suggests that recent TE insertions around OR genes have not driven widespread adaptive selection. Taken together, these findings reveal a mechanistic difference in how TE activity influences the evolution of OR and GR genes. OR loci appear evolutionarily younger—marked by more recent TE insertions, lower synteny, and star-like phylogenetic topologies—yet they show fewer adaptive signatures. Conversely, older GR loci retain adaptive hotspots linked to functionally important encoded residues.

To conclude, the interaction between host plants and herbivorous insects can indeed generate the inferred evolutionary dynamics, with recurrent episodes of positive selection followed by long periods of purifying selection that drive the diversification of chemosensory genes. In host-parasite systems, such as those involving aphids and their host plants, variation in these genes is critical for adapting to plant-produced defensive volatiles or tastant compounds [**241, 266**]. Aphids with enhanced olfactory capabilities are better equipped to detect and navigate toward suitable hosts from a distance, thereby increasing

their fitness. Once on the plant, improved gustatory perception enables aphids to identify optimal feeding sites [**267**]. The faster evolutionary rate observed in OR genes compared to GR genes likely reflects their primary role in long-distance host recognition, whereas GRs function at a later stage, facilitating feeding after host location. However, genome-wide scans are still needed to fully characterize these adaptive processes over shorter evolutionary timescales and to reveal the generality of these patterns [**241**].

4.4 Integration in PLM-View pipeline for functional classification

As a consequence of the success of MuLAN model in the identification of sites in protein sequences related to functional specificity and binding, we experimented with its integration in models for functional classification of proteins. Finding solutions for this task is indeed a central area of research in the Analytical Genomics group: as an example, in collaboration with other colleagues, we built a supervised model for the functional classification of proteins that also identifies the functional motifs in the sequences from limited training samples, based on ESM2 encoder and BERT span detection scheme for extractive question answering task [**86**]. The article, titled "Scaling the Profile of Life by Function with SPIN", has recently been published in *Bioinformatics Advances* [**268**].

In addition to that, other colleagues are working on PLM-View, a pipeline that relies on PLM-based alignment and clustering for a fine-grained unsupervised functional classification of entire protein families. The method works as follows: given a set $\mathcal{P}$ of protein sequences to classify, a reduced set of representative anchors $\mathcal{A}$ is obtained by clustering with MMseqs2 [**269**], in order to reduce the number of pairwise comparisons. Then, for each sequence, the protein embedding $X \in \mathbb{R}^{L \times H}$ is constructed using Ankh PLM [**109**], where L is the length of the protein and H is the latent dimension.

An alignment between each pair of anchor and sequence is computed, applying Needleman-Wunsch algorithm on a similarity matrix S computed from residue-wise embeddings similarities. More specifically, given embeddings $X = [x_1, \dots, x_{L_a}]$ for the anchor $a \in \mathcal{A}$ and $Y = [y_1, \dots, y_{L_p}]$ for the protein $p \in \mathcal{P}$, the elements of $S \in \mathbb{R}^{L_a \times L_p}$ are given by

$$S_{ij} = e^{-||x_i - y_j||^2} + \frac{x_i \cdot y_j}{||x_i||\,||y_j||}, \tag{4.1}$$

and gap penalties are set to 0. Note that, by definition, all entries are strictly positive, and in particular $S_{ij} \in (0, 2]\ \forall\ i, j$.

MuLAN scores associated to the anchor a are used to reweight the columns of S as

$$\tilde{S}_{ij} = m_i L_a S_{ij} \tag{4.2}$$

where $m_1, \dots, m_{L_a}$ are MuLAN raw attention scores (without MinMax rescaling), such that $\sum_{i=1}^{L_a} m_i = 1$. In addition, a filtering step is applied to anchor positions with score $m_i L_a < t$, where t is a threshold set to 1 by default (corresponding to the average of the distribution). As the distribution of MuLAN scores is left-skewed, with values typically

ranging between 0.7 and 2, this filtering step removes, on average, more than 50% of residues from each anchor.

The resulting pairwise alignments are then used to construct a matrix $R \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{P}|}$, where each element defines a similarity score between an anchor and a sequence. Consequently, each column contains similarities between a given protein p and all the anchors, providing a vector representation of proteins in the anchor-defined space. Hierarchical clustering algorithms can then be applied to these representations to build the dendogram.

The integration of MuLAN into the pipeline allows for an accelerated screening of important residues, dramatically reducing the required computational time and, in some cases, improving the classification in terms of topology of the resulting tree. As an example, Figure 4.9 illustrates three subtrees obtained with PLM-View for 28,583 proteins of the thioredoxin family: the first computed without any MuLAN integration, the second with MuLAN attention weighting and the third with additional MuLAN filtering to reduce the effective length of the sequences. We specifically examined the discrimination between f-type and h-type TRXs, for which experimental evidence supports functional diversification. Among the three trees, the one constructed with MuLAN weighting but without filtering (Figure 4.9B) achieves the clearest separation between the two classes, while preserving intra-class grouping. In contrast, completely removing MuLAN scores leads to a mixing of both types, whereas applying additional filtering results in an intermediate outcome, with strong clustering of f-type TRXs, but partial dispersion of the h-type. In the latter case, however, the reduced precision is compensated by the halving of computing time.

Additional applications and analyses are ongoing, in particular for fine-grained classification of Carbohydrate-active enzymes (CAZymes) [**270**].

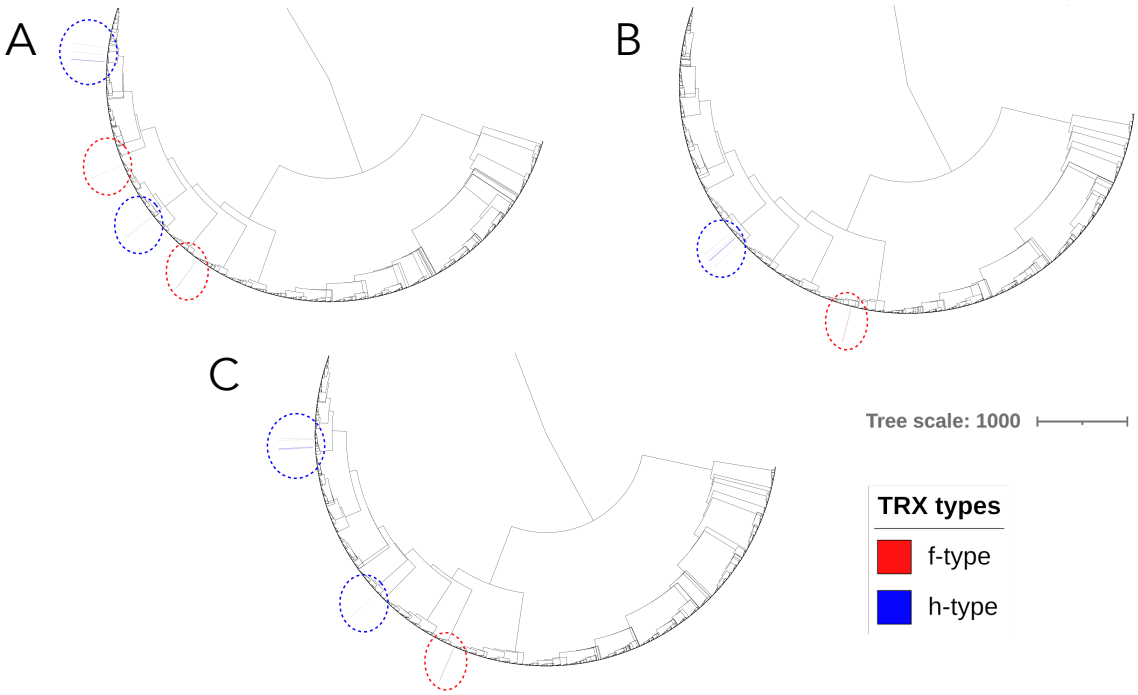

Figure 4.9: **Functional trees for TRX family obtained with PLM-View, with and without the integration of MuLAN scores.** The branch containing TRXs for f-type and h-type is displayed for each tree, using iTOL software [**271**], and leaves corresponding to the two types of TRXs are highlighted with different colors. The trees are constructed without MuLAN scores (A), with MuLAN scores used as weights (B) and with MuLAN scores used for both weighting and filtering (C).

Chapter 5

SoftDis Dataset and LoRA-DR-suite framework

Intrinsic disorder and conformational flexibility are key determinants of protein function, as disordered regions often mediate transient interactions and regulatory processes. In addition, soft disorder plays a crucial role in identifying binding regions in protein surfaces and assembly pathways. Therefore, the availability of comprehensive datasets of protein disorder is essential for training and evaluating predictive models, and to analyze the different properties of the various degrees of disorder.

In this context, this chapter reports on the article "LoRA-DR-suite: adapted embeddings predict intrinsic and soft disorder from protein sequences", published in *Bioinformatics* [**272**]. In the article, we construct a new dataset for soft disorder from crystallographic structures in the PDB, and evaluate PLMs on intrinsic and soft disorder prediction tasks on benchmark datasets.

5.1 Introduction

While many proteins rely on a well-defined three-dimensional structure to carry out their functions, a substantial fraction of the proteome in any organism consists of polypeptide segments that lack a stable and reproducible ordered structure [**273**–**276**]. Remarkably, these structurally flexible or amorphous regions often retain full functionality. Such unstructured and frequently underexplored segments are known as intrinsically disordered regions (IDRs). When the intrinsic disorder spans the entirety of a protein sequence, these proteins are classified as intrinsically disordered proteins.

IDRs are typically identified using structural techniques such as circular dichroism spectroscopy. In X-ray crystallography, they appear as missing residues or highly flexible regions, often indicated by elevated B-factor values [**277**]. However, disorder is not always absolute—a residue may be absent in one structure yet resolved in another, reflecting a dynamic disorder-to-order transition. This inherent flexibility complicates the classification of disordered regions in protein structures. To address this challenge, the concept of *soft disorder* was introduced in [**51**], providing a unified framework to capture the ambiguity between ordered and disordered states. Large-scale statistical analyses of protein chains in the Protein Data Bank (PDB) revealed a strong correlation between soft disordered regions (SDRs) and protein interaction sites [**51**]. These regions appear to play a crucial role in

the assembly and stabilization of protein complexes [**52**]. Beyond this, this study provided evidence of allosteric effects, where SDRs undergo spatial rearrangements upon binding, facilitating the formation of new interaction interfaces in larger, more complex assemblies.

Given their structural flexibility, IDRs and SDRs play essential roles in molecular recognition, signaling, and regulation and are implicated in numerous diseases [**278**–**281**]. Accurately identifying and characterizing these regions is crucial for advancing protein research, therapeutic development by uncovering novel biological functions and deciphering complex molecular pathways [**282**]. A systematic exploration of IDRs and SDRs could offer critical insights into the mechanisms of poorly understood diseases and guide therapeutic strategies targeting disordered regions and their interaction networks.

Given the fundamental role of disordered regions, accurate predictive models are essential. Numerous computational tools identify IDRs, intrinsically disordered-binding regions, and flexible segments [**283, 284**], with deep learning-based approaches increasingly dominant. To enhance their performance and practical applicability, recent efforts have systematically benchmarked these predictors using previously unresolved sequences, providing a robust assessment of their reliability and generalization capabilities [**54, 285, 286**]. With the rapid production of genomic and metagenomic sequences, these predictive tools serve as a crucial link between sequence data and functional annotation, delivering structural insights without the cost, time, and limitations of experimental methods.

In this work, we present LoRA-DR-suite, a novel predictor for identifying IDRs and SDRs in protein sequences. The method leverages adapted embeddings and is built on a simple deep learning architecture.

5.2 Results

5.2.1 A suite of models based on adapted embeddings

We designed (Figure 5.1A) a suite of classification models relying on hidden representations encoded by protein language models (PLMs). Sequences are tokenized at the single amino acid level and embedded through multiple transformer layers, generating highly-dimensional vector representations for each residue.

Exploiting the relatively large training datasets ($\sim 10^5 - 10^6$ distinct tokens for intrinsic disorder, $\sim 10^7$ for soft disorder), we implemented low-rank adaptation (LoRA) of the model, by introducing additional trainable layers. These layers learn low rank decomposition of large weight matrix updates [**98**], enabling effective modification of the model internal representations. This approach is analogous to fine-tuning but involves training only a small fraction of the model parameters. By freezing pre-trained parameters, LoRA prevents catastrophic forgetting while retaining the ability to disable adapter layers and revert to the original model. For all models, LoRA layers were added to the linear transformations in every multi-head attention module of the architecture. Details on hyperparameter choices, specific to the pre-trained model and classification task, are provided in Materials and Methods.

In this scope, the detection of disordered residues is formulated as a binary token

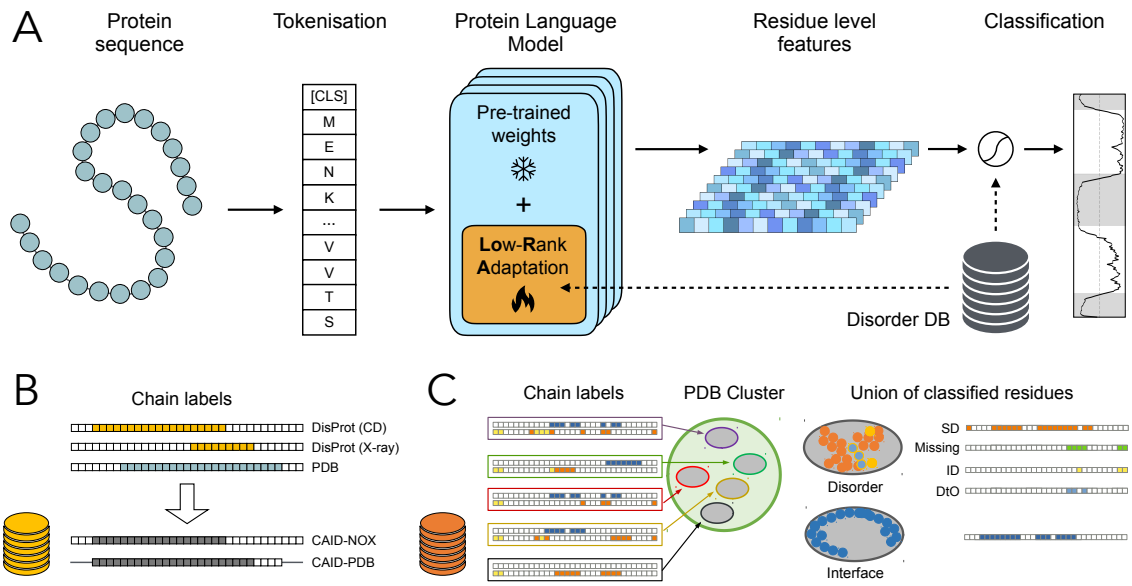

Figure 5.1: **Model architecture and datasets construction**. A. Taking as input a protein sequence, after tokenisation, its adapted embedding is generated by a protein language model, with frozen weights, coupled with several LoRA layers. A linear classification head with dropout is used to generate a disorder profile for the protein, where scores of disorder are computed for each residue and used to construct a disorder profile for the entire sequence. B. Three types of data are used to define intrinsic disorder. They are either coming from circular dichroism (CD) experiments, or X-ray crystallography (X-ray), or missing residues in a crystal (PDB). The CAID-PDB annotation is defined as the union of residues identified as intrinsically disordered by circular dichroism (CD) and X-ray crystallography, while the CAID-NOX annotation includes only residues identified by CD. Residues labeled as disordered or not by a method are represented by colored or white squares, respectively. Notably, missing residues in X-ray crystallographic data are unlabeled in the CAID-PDB annotation but assigned a negative label in the CAID-NOX annotation. C. In the SoftDis database, soft disordered residues (SD), missing residues (Missing), intrinsically disordered residues (ID), and disorder-to-order transition residues (DtO) are identified by clusters of chains with a very similar protein sequence (see Materials and Methods). Each protein sequence associated to a PDB structure in a cluster (left) contributes information on missing residues (yellow squares), residues with high B-factor (orange square) and interface residues (blue squares). The union of these classified residues allows to label residues as SD, Missing, ID, DtO, and interface residues for the cluster and for its sequences (right).

classification task. Using pre-trained transformer architectures, we add a linear classification head with dropout, applied in parallel to the representation of each residue. As a result, LoRA-DR-suite generates a profile of intrinsic/soft disorder for the input sequence, assigning each residue a probability score indicating its likelihood of being disordered (Figure 5.1A, right).

In this study, we employed well-established PLMs as pre-trained models, including two versions of the Evolutionary Scale Model (ESM2) with 35M and 650M parameters [**112**], ProtT5 from the ProtTrans suite [**107**], and Ankh [**109**]. All models were trained using cross-entropy loss with balanced class weights to address class imbalance in the intrinsic and soft disorder datasets. These four models are integrated into the LoRA-DR-suite package.

5.2.2 Comparative Performance Analysis of LoRA-DR-suite on intrinsic disorder

The performance of LoRA-DR-suite models was evaluated against several established methods for intrinsic disorder prediction across three benchmark datasets: CAID1-DisProt [**285**], CAID2_NOX [**54, 286**], and CAID3_NOX (`https://caid.idpcentral.org/challenge`). These datasets were introduced by the Critical Assessment of Intrinsic protein Disorder (CAID) challenge, a biannual competition to benchmark the state-of-the-art in predicting IDRs in proteins: CAID1-DisProt contains the NOX subset of the first CAID round consisting of 646 protein sequences, CAID2_NOX contains 210 protein sequences from CAID2, and CAID3_NOX comprises 148 protein sequences from CAID3. Protein sequences in the NOX datasets are labeled based on information derived from circular dichroism experiments (Figure 5.1B). Key performance metrics include ROC AUC, F1, MCC, and PR AUC (see Section 2.6), allowing for a comprehensive comparison of model efficacy. For each dataset, we compare with the best performing tools that participated to the corresponding CAID round. For the CAID1 and CAID2 evaluation, our models were trained exclusively on data from DisProt 7.0 database (models denoted with the suffix "DisProt7"). For the CAID3 evaluation, we retrained our models by expanding the training set with data from both CAID1 and CAID2 (models labelled with the suffix "ID"). Further details on training data and methodology are provided in the Materials and Methods section. The performance of models trained on DisProt 7.0 database in the CAID3 evaluation is summarized in Table C1.

On CAID1-DisProt test set (Table 5.1), ESM2_650M-LoRA achieved the best overall performance (ROC AUC 0.833, PR AUC 0.510), indicating strong discriminative power and precision-recall trade-off. ESM2_35M-LoRA and Ankh-LoRA closely followed (ROC AUC 0.812 and 0.819), with marginal differences in F1 and MCC scores, underscoring the robustness of LoRA-enhanced models. These models notably outperformed approaches such as SPOT-Disorder2 [**165**] and RawMSA [**167**], which lagged in all metrics. This highlights the advantage of using PLMs with low-rank adaptation for intrinsic disorder prediction.

The CAID2_NOX results (Table 5.1) further demonstrate the competitiveness of

LoRA-enhanced PLMs. Ankh-LoRA, ProtT5-LoRA, and ESM2_650M-LoRA achieved comparable ROC AUC scores (0.837-0.838), with Ankh-LoRA slightly leading in F1 (0.548) and MCC (0.427). flDPnn2 [**287**] and Dispredict3 [**288**] performed well in ROC AUC and PR AUC, remaining competitive alternatives. Again, LoRA-enabled transformers show effectiveness in balancing precision and recall across diverse datasets.

Model	ROC AUC	F1	MCC	PR AUC
CAID1-DisProt				
ESM2_650M-LoRA-DisProt7	**0.833**	**0.498**	**0.394**	**0.510**
ESM2_35M-LoRA-DisProt7	0.812	0.463	0.350	0.505
Ankh-LoRA-DisProt7	0.819	0.490	0.380	0.499
ProtT5-LoRA-DisProt7	0.817	0.482	0.373	0.463
flDPnn	0.814	0.483	0.370	0.475
flDPlr	0.793	0.452	0.330	0.421
rawMSA	0.780	0.445	0.328	0.413
SPOT-Disorder2	0.760	0.470	0.349	0.340
CAID2_NOX				
ESM2_650M-LoRA-DisProt7	0.837	0.535	0.407	0.594
ESM2_35M-LoRA-DisProt7	0.813	0.534	0.406	0.571
Ankh-LoRA-DisProt7	0.838	**0.548**	**0.427**	0.556
ProtT5-LoRA-DisProt7	0.837	0.544	0.420	0.577
flDPnn2	0.838	0.526	0.414	**0.596**
flDPnn	0.835	0.521	0.409	0.581
Dispredict3	**0.842**	0.517	0.404	0.579
DisoPred	0.824	0.484	0.364	0.503
SPOT-Disorder2	0.760	0.506	0.303	0.544

Table 5.1: Results of fine-tuned models on CAID1-DisProt and CAID2_NOX test sets, compared with best performing tools for each challenge, based on ROC AUC metric. See `https://caid.idpcentral.org/challenge/results`.

To ensure a fair comparison on the CAID3_NOX benchmark (Table 5.2), we filtered models based on the fraction of residues (positive or negative) they scored during the competition, as some lacked predictions for all proteins. We then reported metrics on CAID3_NOX dataset for the 5 best performing models with full residue coverage, and on the common subset of proteins scored by the top 10 models (115 proteins out of 148) corresponding to the 53.8% residue coverage. It is worth noticing that sequences in the restricted test set are shorter (median length 314 residues) than those removed (median length 1053 residues; Table C2). Both experiments reveal the strong performance of LoRA-enhanced models, which predict intrinsic disorder across entire protein sequences. Among them, notably ESM2_650M-LoRA achieved top scores in ROC AUC (0.880), F1 (0.649), MCC (0.524), and PR AUC (0.721). ESM2_35M-LoRA also performed competitively, highlighting scalability across transformer sizes.

An analogous analysis was performed for the other CAID3 subsets—CAID3_PDB, CAID3_Binding, and CAID3_Linker—reported in Tables C3, C4, and C5 respectively, where some proteins have been removed (Table C6). While task-specific methods outperform

LoRA models on CAID3_PDB and CAID3_Linker, LoRA-enhanced models achieve superior performance in binding prediction, surpassing dedicated predictors particularly in PR AUC (0.72 for ProtT5-LORA-ID) and Max F1 score (0.618 for esm2_650M-LORA-ID). Notably, for all LoRA models except Ankh-LoRA-ID, the threshold maximizing F1 in binding prediction exceeds 0.85, significantly higher than the 0.50-0.62 range observed for generic disorder prediction. This suggests that LoRA models can reliably identify putative binding regions without requiring task-specific training.

Across all benchmarks, ESM2_650M-LoRA consistently demonstrated strong performance, particularly on CAID3_NOX, where additional data from CAID1 and CAID2 were used for training (Materials and Methods). Adding LoRA layers significantly enhanced the capacity of pre-trained language models to predict intrinsically disordered residues with precision and robustness, even with class imbalance. Models like SPOT-Disorder2 and RawMSA lagged behind, whereas hybrid models such as flDPnn showed competitive but narrower advantages.

Model	ROC AUC	F1	MCC	PR AUC	Max F1
CAID3_NOX (100% residue coverage)					
ESM2_650M-LoRA-ID	**0.880**	**0.649**	**0.523**	**0.721**	0.656
ESM2_35M-LoRA-ID	0.868	0.645	0.518	0.689	**0.659**
Ankh-LoRA-ID	0.861	0.637	0.507	0.684	0.638
ProtT5-LoRA-ID	0.862	0.628	0.494	0.698	0.633
rawMSA-disorder	0.858	0.504	0.399	0.640	0.640
flDPnn3a	0.851	0.587	0.436	0.687	0.642
rawMSA	0.847	0.606	0.462	0.656	0.614
UdonPred-combined	0.841	0.620	0.478	0.634	0.626
flDPnn3b	0.836	0.602	0.456	0.645	0.615
CAID3_NOX (53.8% residue coverage)					
ESM2_650M-LoRA-ID	**0.896**	**0.745**	**0.606**	**0.800**	**0.760**
ESM2_35M-LoRA-ID	0.880	0.730	0.583	0.765	0.752
Ankh-LoRA-ID	0.866	0.717	0.559	0.751	0.718
ProtT5-LoRA-ID	0.884	0.731	0.581	0.789	0.732
DisorderUnetLM	0.881	0.704	0.561	0.778	0.713
flDPnn3a	0.871	0.676	0.499	0.767	0.720
ESMDisPred-2PDB	0.864	0.572	0.459	0.740	0.728
DisoFLAG-IDR	0.863	0.690	0.538	0.761	0.701
ESMDisPred-2	0.857	0.605	0.477	0.738	0.718

Table 5.2: Results of fine-tuned models on the CAID3_NOX test set. F1 and MCC scores were calculated for comparative tools as they were not provided on the official results page `https://caid.idpcentral.org/challenge/results`. Performance is reported separately for models covering 100% of residues and for a reduced test set comprising 53.8% of residues, see main text. See `https://caid.idpcentral.org/challenge/results`.

5.2.3 Extending the definition of disorder through the database SoftDis for soft disorder

The concept of soft disorder was introduced in [**51**] as a general term for regions in a protein identified as flexible (characterized by a high B-factor) or intermittently missing across different X-ray crystal structures of the same sequence. Formally, this definition is derived from an extensive analysis of clusters of alternative structures for the same protein sequence in the Protein Data Bank (PDB): based on the same procedure introduced in [**51**], we newly analyzed 229,370 crystallographic structures and clustered their 484,044 chains into 64,285 clusters to create the SoftDis database. A residue is classified as soft-disordered if it is flexible or missing in at least one crystal structure within a cluster, excluding residues that are consistently missing across all structures. Based on this definition, we statistically demonstrated a significant correlation between SDRs and interface sites observed in protein complexes of identical sequences [**51**]. This analysis also revealed that disordered regions often undergo structural shifts upon binding, facilitating the formation of new interfaces that arise during the assembly pathway [**52**]. Since all chains were retained during database construction, SoftDis encompasses short segments previously identified as MoRFs [**289**] and LIPs [**290**]. While these segments are typically labeled as soft disorder, the soft disorder category covers a broader range. Figures C1 and C2 illustrate SDR lengths (from 1 to > 300 residues) and show that the number of SDRs per PDB chain can reach several dozen.

The SoftDis database is available on the HuggingFace datasets Hub (see Data and Software Availability). It includes labels for representative sequences in clusters, covering soft disorder, missing residues, and interfaces as identified in the PDB. Reduced datasets based on similarity clustering and non-redundant splits designed for learning tasks are also included along with detailed data for chains within each cluster.

5.2.4 Soft disorder classification

Results for intrinsic disorder classification demonstrate PLMs fine-tuning effectively identifies disorder markers, prompting us to extend these models to the broader definition of soft disorder. In this scope, we evaluated the same suite of models on our newly constructed soft disorder dataset (details in Materials and Methods). For learning and evaluation, we first reduced redundancy of training data by clustering representative sequences for each PDB cluster at 50% sequence identity, then randomly splitting the remaining sequences into train, validation and test sets. Sequences in validation and test sets were further filtered at a 30% sequence identity threshold to ensure generalizability. For each fine-tuned PLM of LoRA-DR-suite, the hyperparameters search was performed according to the same procedure as in the case of intrinsic disorder, but using a smaller subset (15%) of all training data for feasibility. Models with selected parameters were then re-trained on the complete dataset.

Table 5.3 provides a performance comparison of LoRA-DR-suite models for predicting soft disordered residues. The results reveal a trade-off between model complexity and predictive performance: ESM2_650M-LoRA, with its larger size and higher computational demand, achieves the highest accuracy, whereas the smaller ESM2_35M-LoRA model

Model	ROC AUC	F1	MCC	PR AUC
ESM2_650M-LoRA-SD	**0.839**	**0.655**	**0.476**	**0.729**
ESM2_35M-LoRA-SD	0.812	0.626	0.430	0.690
Ankh-LoRA-SD	0.831	0.646	0.468	0.716
ProtT5-LoRA-SD	0.823	0.638	0.451	0.702

Table 5.3: Results of LoRA-DR-suite models trained on soft disorder on the SoftDis test set.

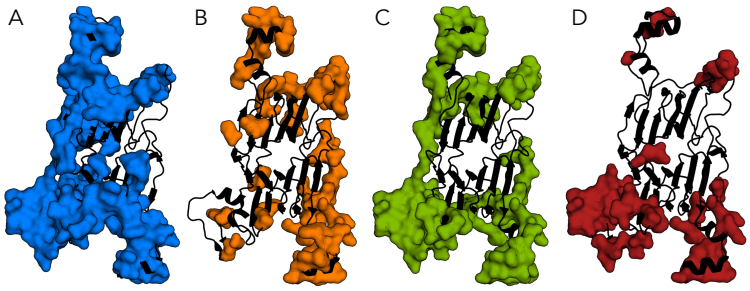

Figure 5.2: **Structural analysis of the Bacillus subtilis Oxalate Decarboxylase OxdC protein.** (A) Residues involved in binding sites across any complex in the SoftDis cluster containing PDB ID 2UYA are highlighted in dark blue. (B) Soft disordered residues identified in all PDB complexes, as in A, are shown in orange. (C) Residues predicted as soft disordered by ESM2_650M-LoRA are highlighted in green. (D) Residues with pLDDT scores below the chain's average are highlighted in red.

remains computationally efficient with slightly reduced performance. Ankh-LoRA offers a balanced intermediate option. These results align with intrinsic disorder classification results, underscoring the predictive power of adapted embeddings in recognizing disorder patterns. Notably ESM-derived models excel despite their lower architectural complexity.

To illustrate the performance of LoRA-DR-suite on soft disorder predictions, we analysed the Bacillus subtilis Oxalate Decarboxylase OxdC protein (PDB ID 2UYA). Experimental data from the SoftDis database were used to compare the union of interface residues (Figure 5.2A) and the union of soft disordered residues (Figure 5.2B) computed on all PDB complexes containing this chain, to the predicted soft disorder residues (Figure 5.2C). Notably, the predicted soft disordered residues show a high correlation with the experimental data. Additionally, low pLDDT values from the AlphaFold structural model of the protein (Figure 5.2D) align with the predicted soft disordered residues [**35**], further supporting the robustness of the LoRA-DR-suite's predictions.

5.2.5 Linking soft and intrinsic disorder with pLDDT

We conducted a large scale analysis on the 2009 chains from our SoftDis testing dataset that have structural models available in the AlphaFold database. This analysis reveals that lower pLDDT values—indicating lower structural confidence or higher disorder—correlate positively with soft disorder predictions and the frequency of soft disorder observed in similar protein chains (Figure 5.3A). These results suggest that regions with low pLDDT scores align well with both predicted and observed soft disorder patterns, effectively

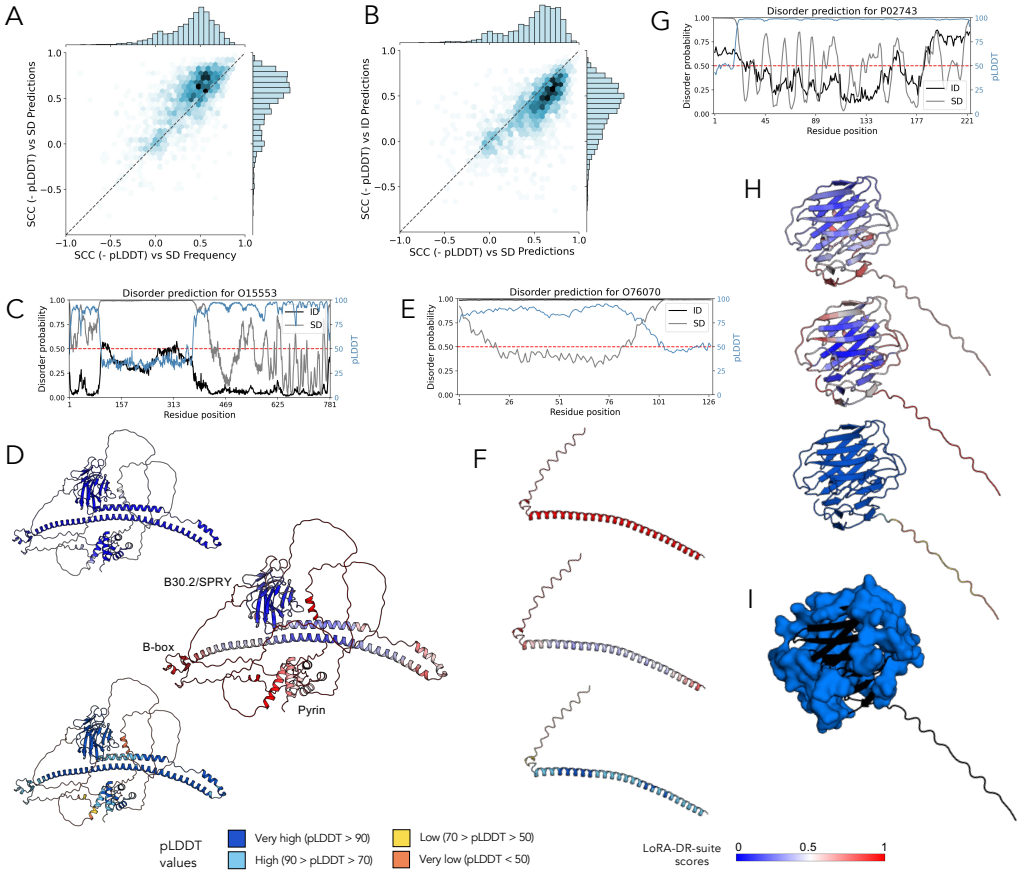

Figure 5.3: **Structural analysis of the MEFV, Synuclein and SAMP human proteins.** (A) Heatmap of similarity (measured by the Spearman Correlation Coefficient) between pLDDT scores and ESM2_650M-LoRA scores for soft disorder predictions, and between pLDDT scores and the frequency of a residue being soft disordered within the PDB cluster in the SoftDis database. The correlations are calculated on the negative of the pLDDT scores. (B) As in A, heatmap between pLDDT scores and ESM2_650M-LoRA scores of intrinsic disorder, and pLDDT and ESM2_650M-LoRA scores of soft disorder predictions. (C) Profiles of ESM2_650M-LoRA scores of soft disorder, ESM2_650M-LoRA scores of intrinsic disorder, and pLDDT for the MEFV protein (UniProt ID O15553). (D) MEFV structural model in AlphaFold Database colored with ESM2_650M-LoRA intrinsic disorder scores (top), ESM2_650M-LoRA soft disorder scores (middle), and pLDDT confidence scores (bottom). (E) Profiles of ESM2_650M-LoRA scores of soft disorder, ESM2_650M-LoRA scores of intrinsic disorder, and pLDDT for the Synuclein protein (UniProt ID O76070). (F) Synucleins structural model in AlphaFold Database colored with ESM2_650M-LoRA intrinsic disorder scores (top), ESM2_650M-LoRA soft disorder scores (middle), and pLDDT confidence scores (bottom). (G) Profiles of ESM2_650M-LoRA scores of soft disorder, ESM2_650M-LoRA scores of intrinsic disorder, and pLDDT for the Serum amyloid P-component protein (SAMP; UniProt ID P02743). (H) AlphaFold model structure of the SAMP protein with residues colored by predicted scores of intrinsic disorder (top), predicted scores of soft disorder (middle), and pLDDT values (bottom). Corresponding color bars are shown on the right. (I) PDB structure 1GYK, chain A, with residues highlighted in blue to indicate interface residues identified in SoftDis. These residues were computed using the cluster represented by chain C of PDB structure 1SAC, which contains 90 chains.

capturing structural variability and flexibility. Figure 5.3B further shows that pLDDT scores align with LoRA-DR-suite predictions for both intrinsic and soft disorder. However, the correlation is stronger for soft disorder, highlighting key distinctions in how these two disorder types relate to structural confidence [**291**].

Figures 5.3C and 5.3D focus on the MEFV (Mediterranean Fever) human protein gene, an important modulator of innate immunity. The soft disorder and pLDDT profiles exhibit a strong Spearman correlation (0.865 with -pLDDT), with soft disorder effectively capturing flexibility differences across annotated domains that are not detected as intrinsic disorder. For example, the pyrin domain (1-92) and B-box domain (370-412) are identified as highly flexible, whereas the B30.2/SPRY domain (580-775) appears much more stable. Additionally, the long helices (420-582), required for homotrimerization and pyroptosome induction, exhibit variable flexibility: stable at the center (low soft disorder, high pLDDT) but highly flexible at the helix extremes (high soft disorder, low pLDDT). These patterns are undetected by intrinsic disorder predictions, which correlate weakly with -pLDDT (0.796).

Figures 5.3EF analyze the Synuclein protein. While our intrinsic disorder predictor classifies it as highly disordered, the soft disorder profile reveals a more detailed perspective: the helix's extremes are identified as highly flexible, while the central region is stable. -pLDDT strongly aligns with soft disorder (correlation of 0.816), but weakly with intrinsic disorder (0.505), underscoring differences between disorder types.

Figures 5.3G and 5.3H analyze the Serum amyloid P-component protein. This protein exhibits consistently high pLDDT values across its sequence, with correlations of 0.591 and 0.829 observed between -pLDDT and intrinsic disorder and soft disorder, respectively. The soft disorder profile identifies several regions with high scores, indicative of flexible zones. These regions are localized on the protein surface (Figure 5.3H, center) and correspond to interaction sites, as highlighted by the blue residues in Figure 5.3I, derived from the SoftDis database.

These observations underscore the relevance of soft disorder predictions in identifying structurally flexible, functionally important regions, involved in protein-protein interactions. The strong correlation with -pLDDT further highlights the potential of combining structural confidence scores with soft disorder profiles to enhance our understanding of protein flexibility, structural variability and interaction dynamics.

5.2.6 Adapted embeddings better align with disorder

LoRA layers alter the internal representations and output embeddings of the models. To evaluate how these changes affect disorder classification, we trained logistic regression classifiers on embeddings from both the pre-trained and LoRA-adapted ESM2_650M models. For intrinsic disorder, we used the CAID3_NOX dataset; for soft disorder, we selected a random subset of 500 proteins from the SoftDis test set to avoid biases from the LoRA training data. Each dataset was split: 70% of sequences were used for 5-fold cross-validation (optimized via ROC AUC across varying regularization strengths), and 30% were reserved for independent testing. As shown in Table C7, LoRA-adapted

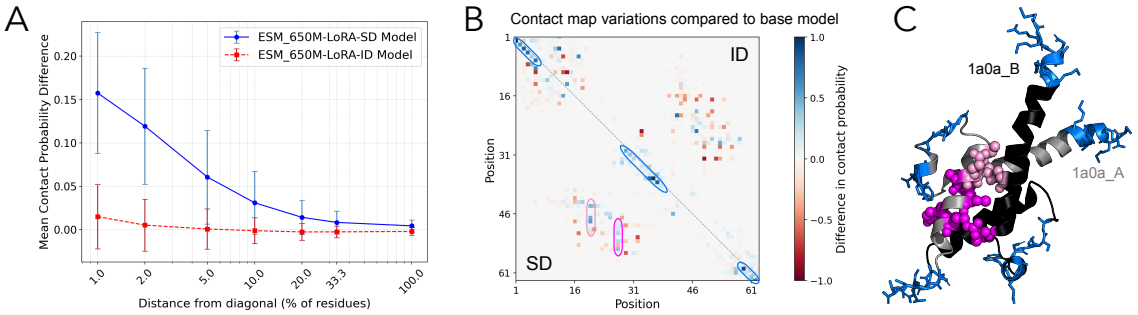

Figure 5.4: **Variations in predicted contacts maps induced by LoRA layers.** (A) Average variation in contacts probability scores returned by ESM2_650M contact prediction head, between models fine-tuned on soft disorder (ESM2_650M-LoRA-SD) and intrinsic disorder (ESM2_650M-LoRA-ID) with respect to the ESM2_650M base model, at different distances from the principal diagonal (measured as fraction of total protein length). (B) Example of contact map variations induced by fine-tuned models on soft disorder (SD, below diagonal) and intrinsic disorder (ID, above diagonal) for PDB chain 1A0A_B. (C) Enhanced predicted contacts from soft disorder fine-tuning are plotted on the dimeric PDB structure (1A0A; DNA chains are removed). Positions with increased scores near the diagonal are colored in blue on both chains. Residues showing enhanced contacts with positions 20 and 27 from chain 1A0A_B are colored in pink and magenta, respectively, in chain 1A0A_A.

embeddings substantially improve classification performance on both tasks, with a ROC AUC of 0.881 and 0.817 for intrinsic and soft disorder respectively, versus 0.832 and 0.784 for ESM2_650M.

5.2.7 Adapted embeddings influence on attention scores

LoRA-DR-suite models exploit adapted embeddings to extract and distinguish fundamental differences between disorder signals. To explore whether intrinsic and soft disorder residues are primarily influenced by global or local effects learned via adapter layers, we used contact prediction as an indirect approach to analyze changes in attention maps after model fine-tuning [**179**].

We performed a large scale analysis of variations in predicted contact maps across 2,512 chains from the SoftDis test set, excluding those longer than 1,024 residues, to assess the impact of LoRA layers compared to the base model ESM2_650M. Adapted embeddings were generated by training with information from either intrinsic or soft disorder, while contact probabilities were derived from the pre-trained ESM2_650M contact prediction head on attention outputs. Figure 5.4A shows that ESM2_650M-LoRA for intrinsic disorder induces variations primarily at very short ranges. In contrast, training on soft disorder leads to variations observed at much greater distances, extending up to a third of the protein's length. Although the number of long-range variations decreases with increasing distance, they remain consistently present. Figure 5.4B presents an in-depth analysis of the phosphate system positive regulator protein PHO4 (chain B), whose dimer is known to form a complex with DNA. The variability in contact probabilities is significantly more pronounced when training on soft disorder, with notable differences observable (see

blue squares) at short range compared to embeddings trained on intrinsic disorder. For intrinsic disorder, adapted embeddings tend to loose contact information at larger ranges (see red squares), a phenomenon also present for soft disorder but to a lesser extent. In Figure 5.4C, positions with enhanced contact scores for the model trained on soft disorder are shown. High scores near the principal diagonal (blue) are mostly found in coiled regions, while those at larger distances reveal contacts with a partner chain in the dimeric complex. Notably, residues predicted to contact alanine at position 20 (pink) and leucine at position 27 (magenta) are highlighted. These results emphasize the unique influence of soft disorder training on capturing contact probabilities through adapted embeddings.

5.3 Materials and Methods

5.3.1 Intrinsic disorder datasets

For model selection and training, we used the dataset from [**166**], derived from DisProt 7.0 [**53, 184, 292**] following identical preprocessing. The original dataset comprised 745 experimentally labeled sequences, split into training (445 sequences), validation (100 sequences) and test (200 sequences) sets. To ensure low similarity between training and test sequences, the combined training and test sets were clustered at 25% sequence identity using the CD-HIT algorithm [**293**]. Test sequences belonging to the same clusters as training sequences were excluded, leaving a final test set of 176 low-similarity sequences.

To evaluate the generalization capabilities of the models, we also used datasets from the first, second and third editions of the CAID challenge. The CAID1-DisProt dataset contains 646 protein sequences that were unknown to the contestants before the evaluation, and whose labels follow the DisProt format. Proteins were sourced from diverse species, including humans and other mammals (368 sequences) as well as prokaryotes (77 sequences). Similarly, the CAID2_NOX dataset consists of 210 sequences from the Disorder_NOX subset of the second challenge edition, while the CAID3_NOX includes 148 sequences from the Disorder_NOX subset of the third edition. Moreover, in the CAID3 evaluation, we added all sequences from CAID1-DisProt and CAID2_NOX, as well as the initially held out set, to the training data, while keeping the original 100 sequences for validation. Sequences in the training and validation sets were truncated to a maximum length of 1024 residues in all the experiments. See the legend of Figure 5.1B for the definition of intrinsic disorder, CAID_NOX and CAID_PDB.

5.3.2 Construction of the soft disorder database SoftDis

To construct the SoftDis database, we retrieved protein structures from the PDB archive (snapshot as of 15-10-2024) and clustered sequences using MMSeqs2 [**269, 294**] at 90% sequence identity and 90% coverage. This process yielded 64,285 clusters encompassing a total of 484,044 chains belonging to 229,376 structures. The average number of structures per cluster was 7.53, with a median of 3. Nearly half of the clusters (31,412) contained only 1 or 2 sequences, while the largest cluster included 1,727 homologs.

The representative chain for each cluster was selected from the experimental structure with the highest resolution or best R-value. For protein complexes, individual chains were analyzed separately and assigned to their respective clusters.

Residues in each chain were labeled as missing if annotated under REMARK 465 in the PDB file. A residue was classified as soft disordered if it was either missing or had a normalized B-factor $b_i = (B_i - \overline{B})/\sigma_B > 3$, where B_i represents the B-factor of the C_α atom, and $\overline{B}$ and σ_B are the mean and standard deviation of B_i values within the chain. Additionally, residues were labeled as interface if they participated in protein-protein or protein-DNA/RNA interactions. Protein-protein binding sites were identified using the INTerface Builder tool [**181**], where residue contacts are defined by C_α atoms within 5 Å. Protein-DNA/RNA binding sites were determined as residues showing decreased accessible surface area, measured using Naccess [**295**] with a 1.4 Å probe, upon binding.

For each site in the representative sequence of a cluster, we recorded how often it was labeled as soft disordered across all chains in the cluster, excluding consistently missing sites. Similarly, we noted how often each site was labeled as an interface. Figure 5.1C illustrates this labeling process for soft disordered residues, missing residues, intrinsically disordered residues, and disorder-to-order transition residues in the SoftDis database.

To adapt the dataset for machine learning pipelines, sequences shorter than 20 or longer than 2,048 amino acids were removed, and redundancy was reduced by clustering sequences at 50% identity and 80% coverage. The resulting 38,218 sequences were randomly split into training (70%), validation (10%), and test (20%) sets. Validation and test sets were then pruned at different sequence identity thresholds (0.9, 0.7, 0.5 and 0.3) with respect to the training set, enabling progressively challenging evaluations. For training and evaluation, we used the dataset pruned at the strictest (0.3) threshold, resulting in 26,752 training, 1155 validation, and 2523 test sequences.

To generate binary soft disorder labels, residues are assigned 1 if marked as soft disordered in at least one sequence in the cluster, and 0 otherwise. On average, 32% of residues per sequence are labeled as soft disordered across splits. We also provide the fraction of chains where each residue is identified as soft disordered, enabling model uncertainty quantification. A similar analysis applied to residues at protein-protein or protein-DNA/RNA interaction sites supports multi-label classification at each position.

5.3.3 PLMs fine-tuning and hyperparameters selection

Pre-trained PLM checkpoints used in our evaluation were taken from the HuggingFace Hub, enabling easy customization for parameter-efficient fine-tuning strategies such as LoRA on attention layers. Hyperparameter optimization was performed using the HuggingFace Trainer and Optuna backend [**296**], with the following possible choices across all experiments:

– Discrete parameters:
  – LoRA rank $r \in \{8, 16, 32\}$;
  – dropout probability $p \in \{0.1, 0.2, 0.3\}$;
  – scheduler warm up ratio $\in \{0.1, 0.2\}$;

  – optimizer weight decay $\in \{0.1, 0.01, 0.001\}$.

– Maximum learning rate $lr \in [10^{-5}, 10^{-3}]$, with uniform probability on logarithmic scale.

– Attention layers with LoRA adapters, $\{Q, K, V\}$ or $\{Q, V\}$.

Optimizer was set to AdamW [**297**] with default parameters and cosine scheduler decay. We used an effective batch size of 8 for training on Disprot 7.0, and of 16 for training on the full intrinsic disorder and soft disorder datasets.

For each experiment, the models with the highest ROC AUC on the validation set were selected from 40 trials using random parameter search (Table C8). Models were trained for up to 10 epochs, with early stopping regularization based on validation ROC AUC and patience set to 2. Training and evaluations were performed on a single computing node equipped with an Intel Xeon Gold 6330 processor (4 cores, 64 GB RAM reserved) and a NVIDIA A100 PCIe GPU with 80GB RAM (see Table C9 for evaluation times on CAID datasets).

5.4 Discussion

By incorporating LoRA adaptations into PLMs, LoRA-DR-suite achieves performance that matches or surpasses state-of-the-art methods, positioning itself as a highly competitive tool for intrinsic disorder prediction. Its versatility allows it to effectively address both intrinsic and soft disorder, broadening its applicability and paving the way for new research directions in the field.

Why a Predictor of Soft Disorder? By predicting regions of flexibility or transient disorder across ensembles of related proteins, LoRA-DR-suite models provide valuable insights in the following areas:

1. Identification of protein interactions sites. Protein interfaces are critical regions, often sensitive to mutations and linked to pathogenic variants. IDRs enable diverse binding modes, facilitating regulation, recognition, and signaling through high-specificity, low-affinity interactions [**278**]. SDRs further enrich interface complexity, with residues of defined interface roles—transient or permanent—that can dynamically shift between structured and disordered states [**52**]. Hence, differentiating intrinsic disorder, soft disorder, and other residues is key to understanding protein interfaces, and the integration of these concepts into models of protein-protein interactions and protein dynamics holds the potential for significant breakthroughs.

2. Link between disorder and assembly pathways. Soft disorder reveals how structural flexibility shapes protein interactions and functional networks [**52**]. Thus, soft disorder-informed predictions can serve as a foundation for designing network-targeted therapeutics exploiting the unique properties of SDRs. In addition, the connection between soft disorder and assembly may reveal how proteins exploit flexibility and transient disorder in multivalent interactions.

3. Guidance for experimental studies. Accurate and fast sequence-based predictions help to efficiently screen large sequence datasets and prioritize candidate residues for validation, reducing reliance on costly and time-intensive methods for structural determination.

Expanding protein disorder research through multiple disorders. Integrating intrinsic disorder—defined via circular dichroism on isolated protein chains—with soft disorder—identified through flexibility and missing residues across similar X-ray structures— broadens disorder research in many ways.

Indeed, identifying factors underlying the overlap and divergence between the two notions of disorder, as predicted by LoRA-DR-suite and annotated in SoftDis, could reveal their distinct functional roles in signaling, regulation, and scaffolding, and uncover the interplay between structural and environmental factors that shape disorder dynamics. Furthermore, exploring the evolutionary constraints shaping disorder types across species or protein families could highlight distinct selective pressures driving protein functional diversity [**298**].

Future research in these directions is critical to understanding how intrinsic and soft disorder contribute to human diseases, where disruptions in disorder-mediated interactions drive pathological states [**278**]. Studying intrinsically disordered and hybrid proteins—comprising ordered domains and functional IDRs exhibiting missing, disordered-to-ordered, consistently flexible, or flexible-to-rigid residues— may yield insights into disease mechanisms and novel therapeutic targets.

Repurposable adapted embeddings. Although not explored here, the adapted embeddings produced by the models can be extracted, analyzed, or repurposed for other tasks related to intrinsic or soft disorder in proteins. The prediction of conditions or partners that induce disorder-to-order transitions in SDRs is a promising research direction that would likely benefit from LoRA-DR-suite models, connecting predictive modeling with functional insights.

5.5 Data and Software Availability

LoRA-DR-suite models and SoftDis dataset are available at `https://huggingface.co/CQSB`, with sample scripts for model deployment, dataset loading and processing. Scripts for hyperparameters optimization and models testing are available at `http://gitlab.lcqb.upmc.fr/lombardi/LoRA-DR-suite`.

Chapter 6

Enhancing protein language models predictions with hybrid modelisation and multiple modalities

This chapter is dedicated to different attempts of improving zero-shot model scores for the estimation of effects of mutations on single proteins through different methodologies, that include combination of models encoding different modalities (Section 6.1), distillation of state-of-the-art models for fast and accurate predictions (Section 6.2), multimodality augmentation of an homology-based generative model through structural information (Section 6.3) and a post-processing position-dependent bias learned from multiple mutational scanning experiments (Section 6.4).

All the introduced methodologies were evaluated on ProteinGym dataset, that in its version 2.0 provides millions of experimental mutational data over 217 protein assays [**28**]. The leaderboard of predictive models on the ProteinGym dataset can be found at `https://proteingym.org/benchmarks`, and is constantly updated. Therefore, I will provide the context in which each method was introduced, and what were the best alternatives at that time.

All the introduced methodologies improved the results on benchmark data over baseline methods from which they derived, and offer interesting insights that can be exploited in future models development.

6.1 Integration of ESM models in ESCOTT pipeline

Introduced in section 2.4, ESCOTT model is an algorithm that computes scores associated to single-point mutations for single protein chains, extracting information mainly from the input multiple sequence alignment and, marginally, from the protein structure [**128**]. While performances on ProteinGym benchmark proved its efficacy, our goal was to improve its predictions by combining them with zero-shot scores from PLMs that encoded complementary information, namely ESM2 for single sequences and ESM-IF1 for structures [**112, 299**].

The motivation is that, as reported in [**131**], different modalities are better suited for specific properties predictions, as represented in Figure 6.1.

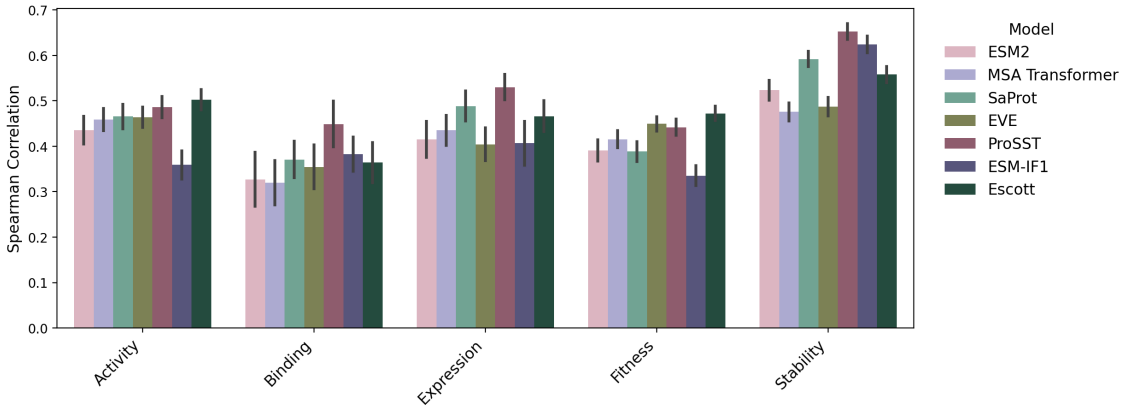

Figure 6.1: Performances of models with different input modalities on ProteinGym assays, by function type.

6.1.1 ESM-IF1 scores as proxy for structural trace

Structural information is encoded in ESCOTT by the circular variance term in the trace computation, that enhances scores for buried residues, while not affecting the exposed ones. However, other structural factors can be determinant in a mutation pathogenicity, such as the local conformation of the protein, that are not captured by the CV.

To mitigate this limitation, we introduced a score extracted from ESM-IF1 to substitute the CV in ESCOTT computation. As the model encodes complex structural features and decodes them to discrete amino acids for each position, the likelihood of an amino acid highly depends on its local structural environment, and it is quantified by the decoded (log-)probabilities [**293**]. As the CV depends only on the amino acid position and not on the specific substitution considered, we used the average loss over position i, with a normalization over the sequence, as a proxy for the circular variance. Since the model uses only the structure as input and thus is agnostic on the specific amino acid at position i, we also include the wild type amino acid in the average computation:

$$\mathrm{CV}_i \to T_{\mathrm{IF1}} = \frac{\min\{0, \mathbb{E}_a[\log p(x_i = a)]\}}{\min_{j \in 1 \dots L} \mathbb{E}_a[\log p(x_j = a)]}. \tag{6.1}$$

The resulting T_{IF1} (IF1 trace) scores are in the [0, 1] range, and have an interpretation similar to CV: values closer to 1 are associated to positions where substitutions are, on average, pathogenic (resulting in large negative log-probabilities), while values towards 0 indicate benign substitutions or no effect, from a structural point of view.

The effect of IF1 trace on the overall performance of ESCOTT is limited, with an average SCC of 0.479 compared to the ESCOTT baseline of 0.474, with large improvement for stability assays (0.591 instead of 0.557) and binding (0.419 instead of 0.381), and an average drop of 0.015 points for expression, fitness and activity assays.

Indeed, this hybrid modelisation strategy is limited by the fact that the structural modality, while capturing additional features, does not always provide complementary information for all mutation types or protein contexts. Thus, the overall improvement is

modest and depends on the specific assay and protein under consideration.

6.1.2 Late logits fusion

Another simple, yet effective approach, is to combine the predictions for models with different data sources. Many multimodal zero-shot models have been developed in the last three years [**115, 116, 121, 300**], but simple (linear) combinations of already existing models can usually achieve similar performance.

An illustrative example is given by the recent RSALOR model [**301**], that computes the effect of a mutation $x_i^{wt} \to x_i^{mut}$ as

$$S(x_i^{mut}) = \Big(1 - \mathrm{RSA}(x_i^{wt})\Big) \cdot \left(\log \frac{f(x_i^{wt})}{1 - f(x_i^{wt})} - \log \frac{f(x_i^{mut})}{1 - f(x_i^{mut})}\right) \tag{6.2}$$

where $\mathrm{RSA}(x_i^{wt})$ is the relative solvent accessible surface area computed from the experimental or predicted 3D structure, and $f(x_i)$ is the frequency of a specific amino acid at position i in the MSA. Despite its simplicity, the model achieves an average Spearman correlation of 0.465, better than all single-modality PLMs of any size. This is particularly interesting as no epistatic effects are considered in this model, that acts independently on each site. In addition, many models based on single sequences or structure are adapted to incorporate MSA information following TranceptEVE strategy [**302**] to improve their performance [**123, 303, 304**].

Similarly, we considered the combination of complementary models ESCOTT, ESM2 and ESM-IF1. We computed the zero-shot scores for all the models as reported in Section 2.4. For ESCOTT, we did not apply ranksorting of the scores, but we applied instead a rescaling strategy to ensure similar ranges for the three measures. Indeed, log-odds ratios between models' output probabilities do not have a pre-defined range, but for PLMs expected values are between -20 and +5 with respect to wild type, with negative values associated to more impactful mutations. Values higher than 0 are particularly interesting as they are expected to correspond to mutations that improve protein fitness over the wild type. In ESCOTT, normalization terms in the models set output values for single-point mutations between -10 and +2, with similar distribution to PLM approaches. Another similarity is that for proteins where uncertainty is higher, i.e. PLM representations are less informative and MSA is more shallow, output values have a lower variance and are closer to 0.

Therefore for each protein, given ESCOTT predicted landscape $S_E \in \mathbb{R}^{L\times 20}$ and PLM zero-shot landscape S_{PLM} of same size, we compute the rescaling multiplicative factor for PLM values as

$$\alpha_{\mathrm{PLM}} = \frac{\min S_E}{\min S_{\mathrm{PLM}}}. \tag{6.3}$$

This rescaling, while being sensitive to possible outliers, ensures that values range is controlled by ESCOTT predictions and preserves their sign, keeping discrimination between increased and reduced fitness. For multiple-point mutations, scores are obtained as the sum of single-point mutation scores, with the same rescaling strategy applied after the sum.

Model	Avg. SCC	SCC by Function				
		Activity	Binding	Expression	Fitness	Stability
ESCOTT + ESM	0.516	0.505	0.418	0.508	0.488	0.639
AIDO Protein-RAG	**0.518**	**0.517**	0.426	0.522	**0.491**	0.635
VenusREM	**0.518**	0.495	**0.454**	**0.533**	0.459	0.650
ProSST	0.507	0.476	0.445	0.530	0.431	**0.653**
ESCOTT	0.476	0.499	0.389	0.468	0.466	0.557
ESM-IF1	0.422	0.368	0.389	0.407	0.324	0.624
ESM2-650M	0.414	0.425	0.337	0.415	0.368	0.523

Table 6.1: Performance of the combination of ESCOTT, ESM2-650M and ESM-IF1 models over ProteinGym DMS benchmark, compared with the three current best models in ProteinGym leaderboard (20-09-2025) and the three models included in the logits fusion.

Rather surprisingly, a simple average of the rescaled scores achieves an average SCC of 0.516 over all the ProteinGym assays, comparable to the latest state-of-the-art models VensuREM [**123**] and AIDO Protein-RAG [**119, 120**], that combine a multimodal single sequence and structure model with MSA processing strategies. Performance over single selection types is reported in Table 6.1.

This analysis reveals that, while modality early fusion is instrumental in producing richer embeddings for many downstream tasks, for mutational effect predictions combinations of results from simpler yet specialized models (late fusion) seems to be more effective. Notably, the combined model achieves better performance than any of ESCOTT, ESM-IF1 and ESM2 for all the functions considered, demonstrating the complementarity of the models and the improved robustness and generalization of the late fusion. The efficiency of late fusion is also highlighted by the comparison with ESM3-1.4B model [**115**] a representative example of early fusion as the embeddings in the transformer residual stream are exactly the sum of those for single modalities. ESM-1.4B achieves an overall SCC of 0.466, lower than all models that apply logits fusion.

It is worth noticing, however, that such mutation score can only be computed for amino acids substitutions, requiring other methods, like decoder models, to evaluate insertions and deletions.

6.1.3 Learned logits fusion

Due to the significant improvement in performance obtained with an average pooling of zero-shot scores, we worked on a strategy to learn a better combination of models' scores. In fact, we found that for some proteins one model was considerably outperforming the others, and the combination in such cases was deleterious. Some examples can be found in Figure 6.2, where we plotted the SCC for different values of λ_1 and λ_2 for the linear combination of scores

$$s = \lambda_1 \alpha_{\text{IF1}} s_{\text{IF1}} + \lambda_2 \alpha_{\text{ESM2}} s_{\text{ESM2}} + (1 - \lambda_1 - \lambda_2) s_E. \tag{6.4}$$

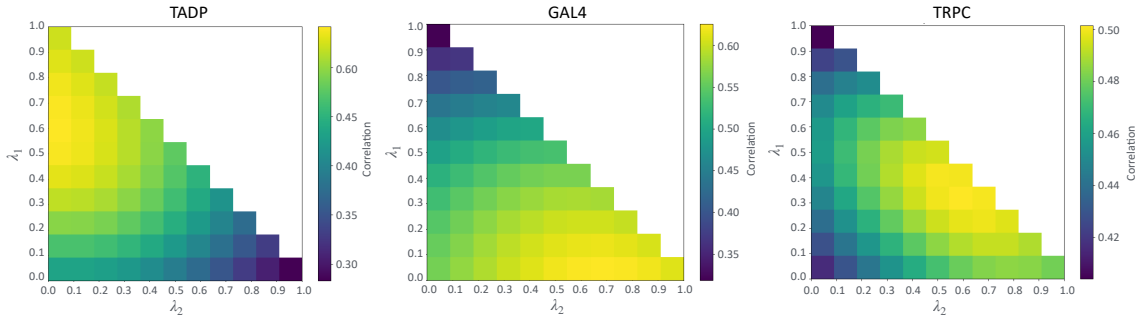

Figure 6.2: Spearman correlation coefficients between experimental scores and combinations of ESCOTT, ESM-IF1 and ESM2 scores, for varying values of λ_1 and λ_2, for three ProteinGym assays.

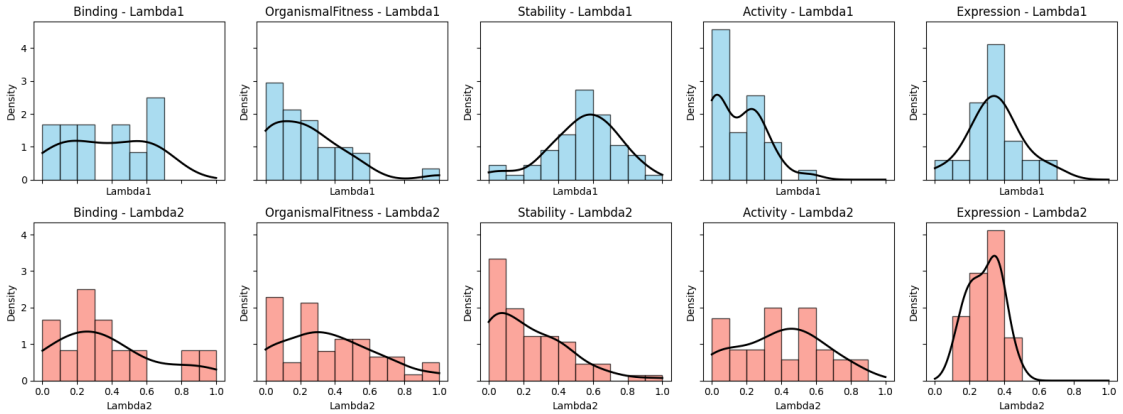

Figure 6.3: Optimal lambda values for combination of ESCOTT, ESM-IF1 (λ_1) and ESM2 (λ_2) by function type.

For instance, experimental scores for TADP protein assay display better correlation with ESM-IF1 scores, opposite to GAL4 protein, for which best results are obtained using solely ESM2 predictions. In the general case, best results are obtained as combinations of scores, with non-zero λ_1 and λ_2 coefficients, as in TRPC protein.

By repeating the procedure over all possible (λ_1, λ_2) values for each protein, we can define the optimal values as

$$(\lambda_1^*, \lambda_2^*) = \arg \max_{\substack{\lambda_1+\lambda_2=1 \\ \lambda_1,\lambda_2\in[0,1]}} \mathrm{SCC}(\mathbf{s}(\lambda_1, \lambda_2), \mathbf{s}_{\mathrm{exp}}) \tag{6.5}$$

where $\mathbf{s} = (s_1, \ldots, s_N)$ are the mutation scores defined above and $\mathbf{s}_{\mathrm{exp}}$ are the corresponding experimental scores. These optimal lambdas can then be used as proxies for the learning task. As illustrated in Figure 6.3, distributions for optimal lambdas change with the considered function type, which is indeed an important feature in the determination of lambda value.

As additional features, since the lambda values are single scalars for the entire protein, we consider global properties of the inputs that might be informative in the choice of one

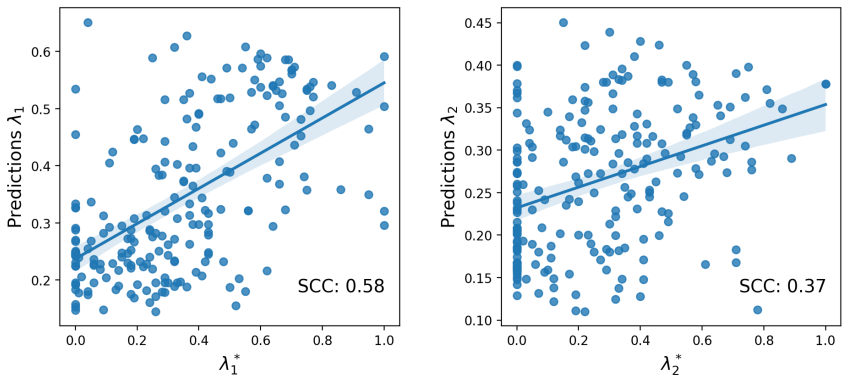

Figure 6.4: Regression results for λ_1 and λ_2 in LOGO cross-validation test setting.

method above the other. In particular, we used average coverage and conservation from the input MSA, discretized in buckets of range 0.2, and fraction of residues associated to each of the 8 secondary structures classes extracted from DSSP [**183**]. Assay function type was instead encoded as 1-hot vector.

This gives a total of 23 features on which we trained standard machine learning models, including linear regression, decision trees, random forest, gradient boosting, with different hyperparameters and regularization strategies. For model and hyperparameters selection, we performed 5-fold cross-validation by randomly dividing the experimental assays. We relied on AutoGluon framework [**305**] for the automation of the experiments pipeline. For testing, we adopted leave-one-group-out (LOGO) cross-validation, where each group is defined by all the assays associated to the same protein. This setting also describes possible use cases, for additional assays not included in the ProteinGym dataset.

We found that extremely randomized trees performed best in 5-fold cross validation, achieving SCC of 0.630 for λ_1 and 0.395 for λ_2. Results of the best models in the LOGO cross-validation test setting, reported in Figure 6.4, display a limited reduction in regression performance, with SCC = 0.58 for λ_1 and 0.366 for λ_2, showing that the method is quite robust. It must be noticed, however, that regression performances are quite bad overall, due to the fact that features might not be informative enough and that the optimal interplay between models also depends on the experimental data collected, and more broadly on the training procedure of the PLMs. Nonetheless, learned values show a significant correlation with optimal ones, in particular for λ_1. For λ_2 correlation is lower, but the fact that the average learned value is around 0.33 (0.27 ± 0.08) demonstrates why a simple average of the scores was already a good approximation for the learned combination.

With these learned coefficients, the overall SCC raises to 0.521, compared to 0.516 for the simple average, and 0.547 for optimal λ_1^* and λ_2^*. Detailed performances are reported in Table 6.2, showing improvements over simple average for all functions. More moderate improvements are obtained for Binding and Expression functions, that are also those with fewer experimental assays (13 in both cases).

In an attempt to refine the logits fusion strategy, we also experimented with multiple lambda vectors associated to different chunks of the same protein. However, we found that

Model	Avg. SCC	SCC by Function				
		Activity	Binding	Expression	Fitness	Stability
ESCOTT + ESM (LOGO)	**0.521**	**0.522**	0.419	0.512	**0.500**	**0.653**
ESCOTT + ESM (avg.)	0.516	0.505	0.418	0.508	0.488	0.639
AIDO Protein-RAG	0.518	0.517	0.426	0.522	0.491	0.635
VenusREM	0.518	0.495	**0.454**	**0.533**	0.459	0.650
ProSST	0.507	0.476	0.445	0.530	0.431	**0.653**
ESCOTT + ESM (opt.)	0.547	0.547	0.447	0.531	0.529	0.679

Table 6.2: Performance of learned combination of ESCOTT, ESM2-650M and ESM-IF1 models in LOGO cross-validation setting over ProteinGym DMS benchmark. Results are compared with the three current best models in ProteinGym leaderboard (20-09-2025) and the average and optimal combination of the considered models.

the complexity of the optimization space and the instability of the coefficients with respect to the chunks definition and length made the problem intractable in the considered setting.

6.2 AlphaMissense distillation for fast and accurate predictions

AlphaMissense is a method built on the AlphaFold2 pre-trained model [**106**], that already encodes sequence-to-structure information, to infer pathogenicity scores for single-point mutations, calibrated in the range [0, 1]. AlphaMissense predictions are available in the AlphaFold Database [**306**] for all human proteins, but the weights for the model were not publicly released.

However, labeled data make it possible to build a fast distillation model that maps input semantic-rich embeddings from PLMs to the corresponding AlphaMissense scores, similarly to how VespaG model was distilled from GEMME [**307**] (even if this project was carried out before VespaG publication). Broad generalizability is possible thanks to the following characteristics of training data:

– predicted scores are homogeneous across diverse proteins as they are calibrated over a large validation set [**106**];

– predictions for the human proteome are provided as full mutational landscape, with no gaps or missing data;

– as the task is framed as a multi-output regression at the token level, the number of training samples greatly exceeds the dimensionality of the embeddings, which reduces risks of overfitting and enables stable parameter estimation.

For our experiments, we considered ESM2 (sequence-only) and SaProt (structure-enriched vocabulary), both in the 650M parameters version, to extract the embeddings. When the work was carried out, SaProt-650M was the best available multimodal model in terms of average SCC, while ESM2-650M was the best model for single sequences. SaProt uses Foldseek alphabet to encode local structural features [**117**], and we extracted them using

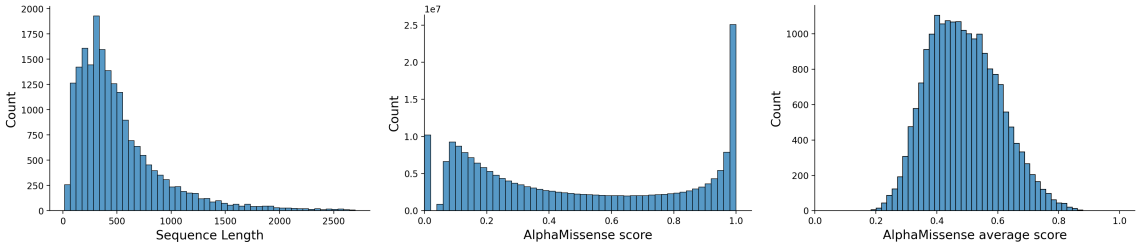

Figure 6.5: Distributions of protein lengths, AlphaMissense predicted scores and per-protein average score over 19,801 sequences in the processed AlphaMissense human dataset.

structures provided in AlphaFoldDB. In positions for which pLDDT confidence is lower than 50, the structure token is masked [**116**].

6.2.1 Data processing

AlphaMissense predictions are available for 20,516 human proteins, from which we removed 715 sequences due to mismatches between sequences obtained from AlphaFold2 and AlphaMissense predictions, or lack of predicted structures. The remaining 19,801 sequences spanned lengths between 16 and 2699 amino acids, with an average value of 516 residues. Protein lengths and AlphaMissense scores distributions over the dataset are shown in Figure 6.5.

To reduce sequence similarity between train and validation data, we clustered the sequences at 30% sequence identity and 80% coverage with MMseqs2, obtaining 12,753 clusters, among which the vast majority only contained single sequences, with a maximum cluster population of 377 sequences. We then selected 90% of clusters (17,938 sequences) for training, ensuring that the bigger ones were part of the set, and kept the remaining 10% of clusters (1,863 sequences) for validation. We used ProteinGym as external test set of experimental data.

As AlphaMissense predictions are in the [0, 1] range and follow an asymmetric distribution towards the two extremes, we applied a logits transformation, such that a score s is mapped to $s' = \log(1-x)/\log(x+\epsilon)$, with ϵ set to $10^{-}4$ to avoid divergences.

6.2.2 Training and evaluation

As downstream trainable modules, we compared a simple two-layer perceptron and a stack of ConvBert layers [**308**], as adopted in [**109**], with different number of layers. While MLP acts on single token representations, ConvBert layers enable further inter-token communication through convolution and attention mechanism.

The training objective was the minimization of the Mean Square Error loss, estimated over all possible amino acids substitutions in each position. All models are trained for 10 epochs using AdamW optimizer with weight decay of 0.01, cosine decay scheduler with 2.5 epochs of linear warm up, maximum learning rate of 2e-4 and early stopping with patience of 3 epochs. For ConvBert model, we set convolution kernel size to 7, number of attention heads to 4, intermediate MLP dimension to twice the embeddings size, and number of

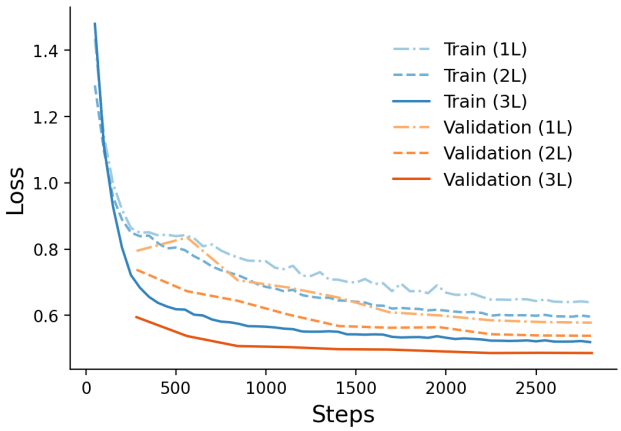

Figure 6.6: Train and validation loss curves for ConvBert-AM-SaProt model on AlphaMissense human dataset.

layers between 1 and 3. As additional regularization, we applied a 0.2 dropout to attention and MLP outputs.

We considered SaProt as backbone model for model classification, and applied the same best ConvBert downstream model to ESM2-650M. Loss curves for ConvBert-AM-SaProt models are shown in Figure 6.6, and results on ProteinGym test set are reported in Table 6.3, where we considered SCC, AUC and MCC as evaluation metrics. Positive and negative labels are defined from binarization values of experimental distributions provided for each assay [**28**]. In all the metrics considered, increasing downstream model's complexity up to three ConvBert layers improves performances. In particular, ConvBert-AM-SaProt (3L) reaches an average by selection type of 0.474 (0.488 overall), with a state-of-the-art average of 0.526 in activity assays. However, SCC for stability assays degrades from 0.592 to 0.512, denoting better alignment of AlphaMissense scores with other fitness-related measures. In addition, input structural information is indeed beneficial, as demonstrated by the difference in SaProt- and ESM2-derived models. Still, in both cases, metrics significantly improve over the model baseline.

The model is also much faster than PLMs at scoring the mutations: as it is trained directly on embeddings from wild type sequences, it does not require multiple forward passes to compute zero-shot scores for all positions. Instead, it only requires a single forward pass to generate the embeddings, and a single forward pass in the downstream model to compute the outputs for all possible substitutions, allowing for fast screening of large datasets. As a reference, given precomputed Foldseek tokens (which is not a computationally expensive step, if the protein structure is provided), scoring all mutations for the 217 assays in ProteinGym took less than 40 seconds on a RTX 3090 GPU, including processing of multiple-point mutations, computed as the sum of single scores.

Model	SCC	AUC	MCC
SaProt-650M	0.457	0.751	0.359
MLP-AM-SaProt	0.470	0.762	0.371
ConvBert-AM-SaProt (1L)	0.471	0.762	0.372
ConvBert-AM-SaProt (2L)	0.472	0.763	0.374
ConvBert-AM-SaProt (3L)	**0.474**	**0.764**	**0.375**
ESM2-650M	0.414	0.729	0.327
MLP-AM-ESM2	0.442	0.748	0.351
ConvBert-AM-ESM2 (3L)	0.451	0.752	0.357

Table 6.3: Performance of downstream models trained on AlphaMissense human predictions over ProteinGym benchmark, for SaProt-650M and ESM2-650M backbone models. Scores are computed according to ProteinGym evaluation scheme, by averaging metrics results for each selection type.

6.3 Gated structural conditioning in auto-regressive model

The pre-training of a foundation PLM of hundreds of millions (if not billions!) of parameters is computationally very expensive and requires systems of multiple GPU nodes, as well as large datasets of several milions of proteins, making it usually unfeasible. This is the reason why various models were developed, integrating pre-trained PLMs with modality-specific adapters, trained on smaller datasets.

Here, we consider PoET [**309**], a transformer decoder model that already enables sequence conditioning for protein generation, similarly to the more recent ProtMamba state space model [**310**]. In this project, we trained a model adapter that merges sequence and structure modality, to enable also structural conditioning in the inference phase. It is important to mention that this project was carried out in November 2024, before the public release of PoET-2 [**304**].

6.3.1 PoET model

PoET model (which stands for Protein Evolutionary Transformer) introduces architectural innovations to the standard transformer decoder that bridge autoregressive sequence modeling with evolutionary context integration.

It operates over sets of homologous sequences drawn from a protein family. Each input is represented as a sequence of sequences, a context prompt comprising multiple homologous proteins, and the model learns to generate or evaluate target sequences in this joint context. Crucially, this is achieved without the need for explicit multiple sequence alignments (MSAs), allowing PoET to natively handle insertions and deletions.

The model architecture is based on a transformer with two forms of attention (Figure 6.7):

– Intra-sequence causal attention, which models sequential dependencies within each individual protein in the autoregressive fashion.

– Inter-sequence attention, which enables the model to incorporate contextual signals across

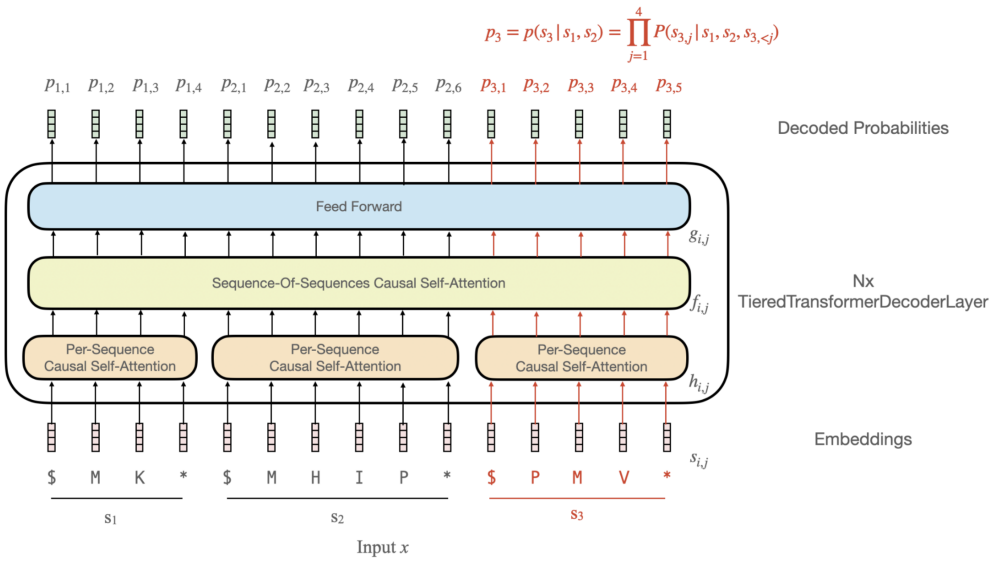

Figure 6.7: Attention scheme introduced in PoET to model both intra-sequence and inter-sequence context. The image is taken from [**309**].

homologous sequences without depending on their ordering or alignment. This is achieved through an inter-sequence relative positional encoding that is invariant to ordering of sequences

This hybrid attention design facilitates parameter sharing across both positional and homologous dimensions, enabling scalable modeling of protein families with varying levels of sequence similarity and diversity.

PoET contains 200M parameters and is trained on a large corpus of unaligned sequence clusters derived from [**34**], using standard auto-regressive cross entropy over amino acid tokens. The training objective operates over entire protein sequences conditioned on the prompt set, allowing the model to learn coevolutionary relationships and family-specific constraints implicitly.

PoET's learned representations have been evaluated on deep mutational scanning benchmarks, where it achieves strong zero-shot variant effect prediction performance, outperforming previous alignment-based models and large-scale PLMs (e.g., ESM2, MSA Transformer). In addition, the model allows prompt-conditioned generation of novel protein sequences, and it can scale to diverse protein families without requiring MSAs or supervised fine-tuning [**309**].

6.3.2 Flamingo-inspired integration of structural information

Prompt-conditioning on retrieved homologous sequences allows restricting protein generation on specific regions in the sequence landscape, designing proteins that are more likely to display a desired property, such as thermostability or solubility in a certain environment. However, including structural information for the desired fold can increase structural diversity [**311**] and improve the zero-shot scoring of mutations, as already proved

in ESCOTT and in other multimodal models.

Our integration of structural information is inspired by Flamingo gated fusion [**312**], where information from a different modality (images or video for Flamingo, protein structure in our case) is extracted through a pre-trained encoder and injected into the pre-trained language decoder model through gated cross attention and dense layers for modality fusion. The gating mechanism is achieved through a $\tanh \alpha$ factor, where α is a learnable parameter, that modulates the new layers' addition to the residual stream and ensures that the original decoder is kept intact at initialization, to improve training stability. Structure is encoded through ESM-IF1 pre-trained encoder [**313**], and cross-attention from the encoded structure(s) attend to all sequences individually. Gated layers are only inserted before the intra-protein attention layer. In this way, the model learns how to decode single sequences from a single associated fold, while keeping the pre-trained inter-sequence signal as it is.

In addition to the cross-attention and MLP gated layers in the decoder, a two-layers MLP encoder adapter is added to project the embeddings from ESM-IF1 in the latent space of PoET embeddings. This single projection is then mapped into K and V matrices for cross attention with layer-specific weight matrices. During training, all pre-trained weights are kept frozen, except for the final decoding layer.

6.3.3 Data and training

As training and validation dataset, we relied on sequences and associated structures from CATH v4.3.0 [**314, 315**]. CATH is a public database that provides information on the evolutionary relationships of protein domains from the PDB, based on a hierarchical classification scheme: At the Class (C) level, domains are categorized based on their secondary structure composition, such as all-α, all-β, α/β mixtures, or minimal secondary structure content. The Architecture (A) level incorporates the overall spatial arrangement of these secondary structure elements to define domain organization; at the Topology/Fold (T) level, the specific patterns of connectivity and organization among secondary structure elements are considered; finally, assignment to the Homologous Superfamily (H) level is made when there is strong evidence of evolutionary relatedness between domains, indicating homology. Within each superfamily, sequences are further classified into Functional Families (FunFams) based on sequence patterns. Relatives within these FunFams are likely to share highly similar structures and functions.

For each FunFam, a reference structure is provided, as well as an MSA containing all the sequences for the specific FunFam. From CATH v4.3.0, we extracted 63.335 FunFams that we randomly split into train and validation sets, with a 90% - 10% ratio. It is important to emphasize that the majority of sequences in the dataset are not full protein sequences, but single domains; however, the training objective is to learn a mapping between MSA and associated fold, rather than learning a full protein sequence/structure embedding, which is suitable for single domains.

As our primary objective is to assess the effect of structure inclusion in the evaluation of mutational effects, we kept ProteinGym as external test set for this task. For sequences

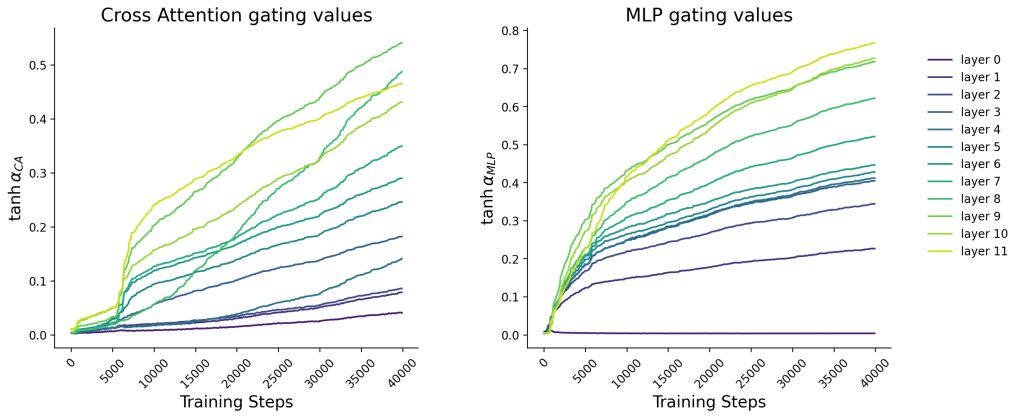

Figure 6.8: Gating values along training phase for structure-adapted PoET model, for cross attention and MLP component in each transformer layer.

retrieval and score attribution, we used the same strategy as in the original PoET paper, picking sequences for prompting at inference time according to different similarity thresholds with respect to the wild type sequence. The associated structure is instead generated with AlphFold2. Scores for mutated sequences are computed following the procedure described in Section 2.4 for autoregressive models.

We trained the model for a total of 40 thousands steps (10 days), with AdamW optimizer, maximum learning rate of 5e-4, inverse square root decay with 1'000 warm up steps, default weight decay of 0.01, and an effective batch size of 128 samples. The context length, that defines the maximum number of tokens and, thus, of homologous sequences in the sample, was set to 8'192, with sequences extracted randomly from the input MSA. As reported in PoET paper, the sequence order (by length or similarity) does not have much effect on output logits [**309**], therefore we applied random sampling over sequences in each FunFam.

6.3.4 Results

We ran different experiments to select structure encoder and training hyperparameters. Here, we report the results for the best model in terms of validation loss, for which we adopted a two-layer MLP with hidden dimensions of 1024 and GELU activation. We also set a dropout probability of 0.1 for gated MLP and attention outputs during the training phase.

In Figure 6.8, we report the $\tanh \alpha_i$ gating values for all layers during the training procedure. As in the Flamingo implementation, we see that the values naturally increase from 0 during training, reflecting the effective incorporation of structural features into the model. For MLP layers, α values are higher for the last layers, while they tend to 0 for the first ones, where parameters encode complex features related to input sequences. A similar behavior is also observed is cross attention layers, but with exceptions for layers 8 and 9, for which α is higher than the last two layers. Additionally, there seems to be a phase transition during training around step 5.000, where α values for cross attention layers increase rapidly, before resorting to the previous growing rate.

Model	Avg. SCC	SCC by Function				
		Activity	Binding	Expression	Fitness	Stability
PoET + Structure	0.475	0.464	0.386	0.477	0.444	0.601
PoET	0.470	0.494	0.396	0.466	0.475	0.519
ProSST	0.507	0.476	0.445	0.530	0.431	0.653
ESCOTT	0.476	0.499	0.389	0.468	0.466	0.557
ESM-IF1	0.422	0.368	0.389	0.407	0.324	0.624

Table 6.4: Performances of structure-aware PoET model in ProteinGym DMS benchmark dataset.

In terms of evaluation of mutational effects, the new model slightly improves overall performance over the PoET baseline, especially concerning mutations that alter protein stability, as shown in Table 6.4. Results, however, show that models pre-trained with multimodal data like ProSST are indeed better for this task.

An interesting analysis that is missing, that we leave for future work, is to assess the quality in terms of folding and diversity of generated sequences, and the comparison between sequences generated with sequence conditioning only, as in the original PoET model, and with additional structure conditioning. Finally, the model would surely benefit from an improved training dataset containing also structures and domains from the updated version of AlphaFold database, that has been recently released cite [**316**].

Nonetheless, this project introduces a different way of inserting an additional modality into a pre-trained model with relatively limited data and computational resources, compared to the pre-training of a multimodal PLM.

6.4 Post-processing bias correction from ranked data

6.4.1 Learning from different experimental assays

Data from mutational experiments, such as DMS, are usually obtained with different experimental techniques, resulting in varying distributions and values. While methods have been developed to predict experimental outcomes from few data points in an assay [**146,147**], or use them to better align zero-shot prediction scores with experimental values, integrating data from other experiments is more challenging, due to distribution shifts and differences in labels attributions. This is the reason why self-supervised foundation models are usually better in the assessment of mutational effects on large scales and diverse proteins. A prior work on this task can be found in [**145**], where the authors trained a logistic or linear regression on clinical or DMS data from few proteins, respectively, using information from the MSA, the predicted structure, zero-shot scores and amino acid properties as input features, finding substantial improvement but requiring a costly computational overhead. For DMS data, to overcome the heterogeneity problem, they considered in the training set only assays from the same lab, employing the same selection strategy.

Here, we designed a simple procedure to leverage experimental data from different experiments and embeddings extracted from the last layer of the protein language model

to learn a single value to bias the zero-shot predictions of foundation models, without requiring inputs from additional sources or modalities, besides those already encoded. This is a simplification of the problem with respect to the supervised learning of entire landscapes, and does not require additional inputs besides those already used in the models, which enables an easy application as post-processing step.

In practice, given a protein sequence representation $X = (x_1, x_2, \ldots, x_n)$, where x_i are (possibly multimodal) representations of single amino acids, models trained in a self-supervised framework learn probabilities $p(x_i = a \mid X^*)$ of observing an amino acid a at position i given the sequence context X^*. The sequence context can be either $X^* = X_{<i}$ for autoregressive models, or $X^* = X_{-i}$ for bidirectional encoder models. Zero-shot log-odds score S of observing a mutated amino acid x^{mut} instead of the wild type x^{wt} at position i are defined as

$$S(x_i^{wt} \to x_i^{mut}) = \log \frac{p(x_i = x^{mut} \mid X^*)}{p(x_i = x^{wt} \mid X^*)}. \tag{6.6}$$

Our goal is to learn a bias to the score that depends on the model representation (embedding) at position i:

$$S'(x_i^{wt} \to x_i^{mut}) = S(x_i^{wt} \to x_i^{mut}) + \beta(x_i) \tag{6.7}$$

where x_i depends on the context X^*. Importantly, the representation x_i does not depend on the wild-type amino acid in position i.

Also, we learn a bias that is purely positional, which means that it is independent of the specific mutation considered. As learning objective, we used a differentiable version of the Spearman correlation, therefore optimizing the ranking between predictions and normalized experimental values. Differentiable ranking was achieved with the implementation provided in [**317**]. We limited the analysis to single-point mutations, and we normalized the experimental scores by applying ranksorting for each protein, and averaging the ranks for each position.

6.4.2 Experimental setup

For the selection of model and strategy, we randomly split the 217 assays from ProteinGym dataset in train and test sets, counting 177 and 40 assays respectively, and we defined the validation set taking 20% of mutations contained in the training data. Also, due to the problem definition, we discarded multiple-point mutations, insertions and deletions.

As PLMs, we evaluated the protein encoders ESM2, in the versions with 35M and 650M parameters, and ProSST, that includes also structure information and is the best performing model in terms of SCC when logits fusion is not considered. We tried different simple networks to learn β, namely a single linear layer, a 2-layer MLP with ReLU activation function. We also considered learning an additional expanding factor $\alpha(x_i)$ such that

$$S'(x_i^{wt} \to x_i^{mut}) = \alpha(x_i) S(x_i^{wt} \to x_i^{mut}) + \beta(x_i) \tag{6.8}$$

to account for possible logits rescaling, similar to a position-wise temperature parameter.

However, we found that the simple linear layer outperformed the other approaches on validation data, and therefore we selected this model implementation. Training

We evaluated the method in two different settings: leave-one-group-out cross-validation (LOGO-CV) on ProteinGym data, and assessment on independent test sets Human500 [**318**], that includes full mutational landscapes for 500 human domains, where stability was measured. In the LOGO cross-validation, groups are assigned based on UniProt ID of wild-type proteins, implying that different experiments for the same protein are grouped in the same fold. Results for ESM2-650M, ESM2-35M and ProSST-2048 are reported in Table 6.5, averaged by selection type. As one could expect, more significant improvements are obtained for models with a worse zero-shot performance, while they tend to saturate for better models.

Selection	Base SCC	SCC by pos.	Updated SCC	Updated SCC by pos.
ESM2-650M model				
Activity	0.431	0.487	0.464	0.543
Binding	0.382	0.433	0.424	0.468
Expression	0.409	0.376	0.437	0.421
Fitness	0.381	0.456	0.403	0.479
Stability	0.496	0.577	0.562	0.664
Average	0.420	0.466	0.458	0.515
Average All	0.428	0.491	0.468	0.542
ESM2-35M model				
Activity	0.310	0.338	0.377	0.460
Binding	0.315	0.352	0.350	0.371
Expression	0.337	0.301	0.373	0.354
Fitness	0.210	0.244	0.290	0.364
Stability	0.387	0.470	0.479	0.592
Average	0.312	0.341	0.374	0.428
Average All	0.300	0.343	0.375	0.452
ProSST-2048 model				
Activity	0.486	0.577	0.520	0.630
Binding	0.527	0.617	0.513	0.599
Expression	0.527	0.545	0.523	0.551
Fitness	0.441	0.531	0.476	0.574
Stability	0.638	0.723	0.649	0.730
Average	0.524	0.598	0.536	0.617
Average All	0.522	0.605	0.544	0.632

Table 6.5: LOGO-CV average Spearman correlation coefficient by selection types for ESM2-650M, ESM2-35M and ProSST-2048 model and variants with learned bias. For each model, the base SCC for all mutations, the base SCC for per-position averaged scores (SCC by pos.), and their updated scores with learned bias are reported. The "Average" metrics correspond to the average between SCC values for each selection type, while "Average All" is the averge over all proteins in the dataset.

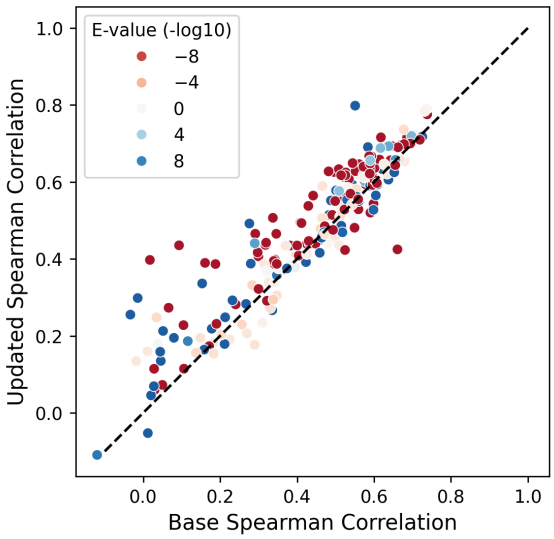

Figure 6.9: Spearman correlations for ESM2-650M zero-shot scores and updated scores in LOGO-CV setting for ProteinGym dataset, by E-value.

To test if the observed improvements should be primarily attributed to the sequence similarity in the ProteinGym data, we computed the minimum E-value between each protein and all the others in the dataset, and plotted the Spearman correlations highlighting the corresponding E-values. As shown in Figure 6.9 for ESM2-650M model, the updated scores improve correlations for single-point mutations for the vast majority of experimental assays, and the improvement is not correlated with E-values E (PCC=-0.09 between variation in SCC and $-\log_{10}(E)$). A similar behavior is observed for ProSST-2048 and ESM_35M too, with correlations PCC=-0.01 and PCC=-0.22, respectively.

For an independent evaluation, we trained the model separately on the selected 177 assays from ProteinGym and on stability assays only, and we considered test data from stability experiments in Human500 dataset, comprising short sequences in the range 19-97 amino acids. Results, reported in Table 6.6, show an improvement on the average correlation for ESM2-650M and ESM2-35M, while performances are similar or slightly lower for ProSST-2048. Training on stability assays only appears to be slightly beneficial or neutral when considering single mutations, but leads to reduced performances on average per-position scores. The fact that results of updated model are lower for ProSTT-2048 indicate a small reduction in generalization of the model, which is currently the state-of-the-art in this task. On the contrary, cross-protein transfer substantially increases results for ESM2 models, with ESM2-650M SCC by position (0.580) surpassing original ProSST-2048 performance (0.567), as structural features are enhanced. Another interesting aspect is that, for all models, original performances are higher when only the short mutated sequence, rather than the full protein sequence from UniProt, is considered, highlighting the importance of more local correlations in stability assays.

Selection	Base SCC	SCC by pos.	Updated SCC	Updated SCC by pos.
ESM2-650M	0.490	0.531	0.513 (0.510)	0.580 (0.566)
ESM2-650M	0.462	0.500	0.473 (0.478)	0.532 (0.531)
ProSST-2048	0.518	0.567	0.508 (0.518)	0.573 (0.572)

Table 6.6: Spearman correlation of bias models trained with different embeddings on Human500Domains test dataset, compared with original performances. For updated correlations, the first values are obtained with models trained on 177 ProteinGym assays from all selection types, while values in brackets are obtained with models trained only on stability assays.

6.4.3 Positional bias enhances global structural features

As the higher improvements are obtained for stability assays, we hypothesize that the model learns a bias related to structural features. To test this hypothesis, we plotted the distributions of predicted scores for ProteinGym test proteins with ESM2-650M, with and without the learned bias, and compared them with ranksorted experimental scores (Figure 6.10), based on amino acids hydrophobicity and position in the protein structure. Amino acids belonging to the interior of the protein and to its surface were separated based on a SASA value of 30.0 $Å^2$, computed on predicted structures provided in ProteinGym. We found that the effects of the bias on the updated scores distributions is twofold. First, it reduces their variance with respect to the base scores, and second, more importantly, it improves the separation between distribution for internal and surface residues, which is a property observed for experimental data, in particular for stability assays. The improved separation between the two distributions is indeed one of the reasons for the gain in correlation induced by the addition of the RSA term in RSALOR [**301**], as well as for the learned bias. For the same reasons, learned bias is less impactful on ProSST-2048 scores, as the model can already better discriminate between internal and surface residues, due to its structural encoder, which encodes local structures of amino acids including their forty nearest neighbors based on the distances of C_α atoms. This can be seen from Figure 6.11, where distributions for base scores and updated scores are very similar between each other, as opposed to those for ESM2-650M model.

Overall, the application of this simple model, that does not use any additional feature other than PLMs inputs, after zero-shot scoring of mutations, delivers variable but consistent improvements across diverse protein sequences and functions. In fact, even if PLMs learn patterns from single protein sequences, their performances in predictions of mutational effects tends to be higher for proteins that have more homologs, due to the presence of a higher number of similar sequences in the pre-training corpus. Therefore, especially in the case of poorly characterized or orphan proteins, transfer learning from other proteins of feature importance in scoring mutations is likely to improve the model performance. Finally, when we consider overall susceptibility of a position to mutations, indicated by average per-position score, we detect an improvement in all models and experimental settings considered, demonstrating that post-processing ranking corrections are beneficial in capturing global cross-protein fitness-related features.

Figure 6.10: Spearman correlations for ESM2-650M zero-shot scores and updated scores in LOGO-CV setting for ProteinGym dataset, by tested function, hydrophobicity class of wild type residue and position in the protein (surface or interior).

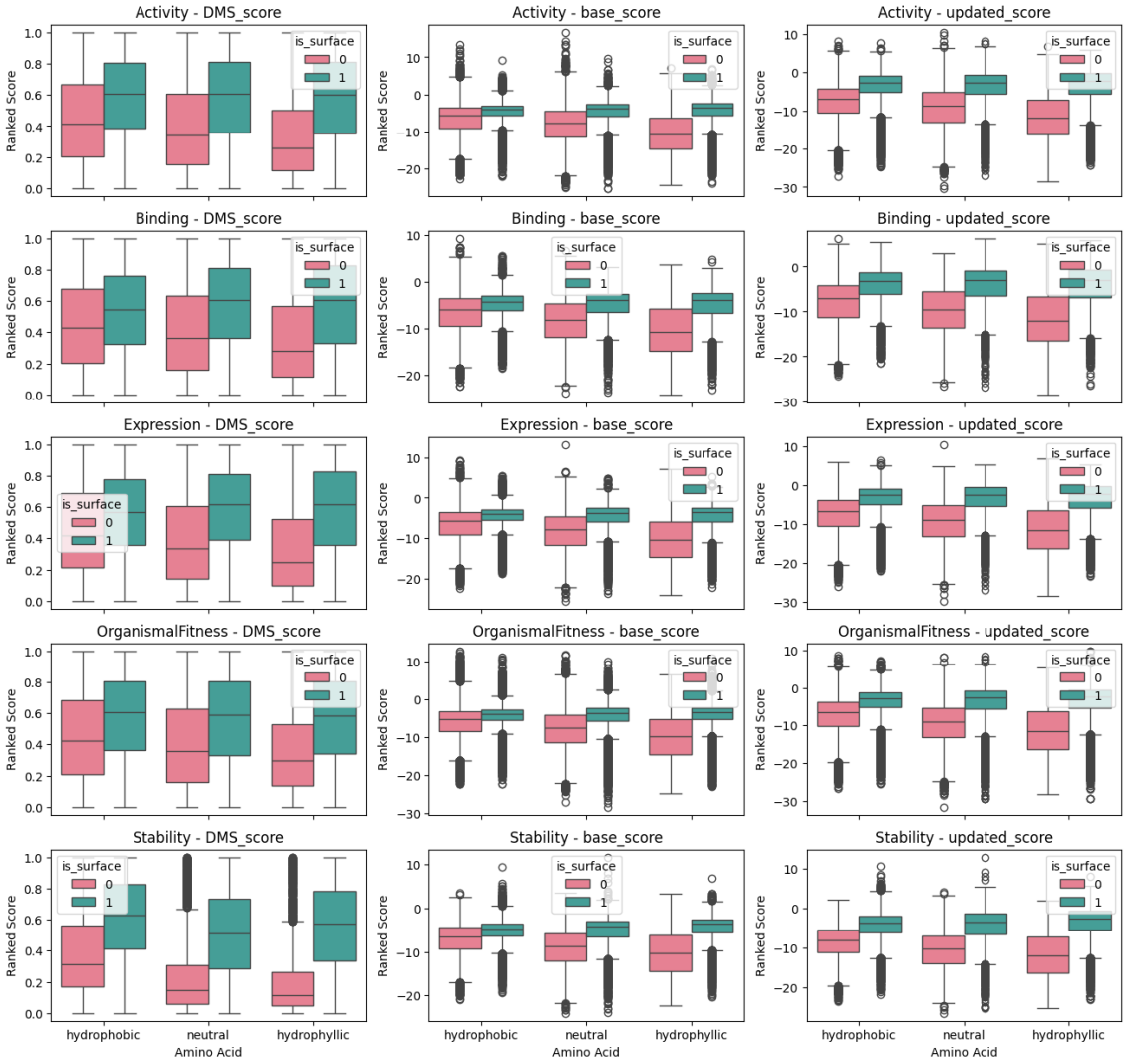

Figure 6.11: Spearman correlations for ProSST-2048 zero-shot scores and updated scores in LOGO-CV setting for ProteinGym dataset, by tested function, hydrophobicity class of wild type residue and position in the protein (surface or interior).

Chapter 7

Conclusions

In this thesis, we tackled different open problems related to the computational prediction of mutational effects in protein sequences and protein-protein interactions. As protein sequence data are increasingly available, the development of methods that can leverage these data to provide precise and biologically-grounded predictions of protein properties has become an exciting and rapidly evolving field. Specifically, foundational PLMs emerged as new tools to encode the complexity of biological molecules like protein sequences, providing high-dimensional representations that capture properties of single residues and relationships between them, that go beyond conservation and coevolution patterns.

These representations can be leveraged and adapted for downstream predictions, either on single proteins or protein-protein interactions. In the first study, we introduced MuLAN, a framework for the prediction of binding affinity changes induced by point mutations in protein complexes, and the identification of interfaces regions and functional sites. We benchmarked its performance on different datasets, and we found that better results for these tasks are obtained with models derived from Ankh PLM. Its relative positional encoding and masking strategy during the pre-training phase, that considers spans of residues instead of single ones, allows the model to better capture relationships between neighbouring residues, that are further enhanced by the light attention module. Additionally, the introduction of amino acid-specific features, like evolutionary scores from iGEMME, improves model's generalization on unseen sequences. Attention scores extracted from the model are also informative for interaction and functional sites.

MuLAN's predictions were tested in different biological systems, where they provided useful insights on the partners' specificity in the considered complexes. The first application was in chloroplast thioredoxins, where MuLAN's mutational scores highlighted electropositive residues around the binding site (also recognized by attention scores) of TRXf2 type, which determine its specific activation towards FBPase redox. The inclusion of such TRXf2 determinants in h type TRX allowed to design a TRXh hybrid able to activate FBPase. Other applications involved self-incompatibility systems in *Petunia* and *Malus* plant families, and olfactory and gustatory receptors in pathogen/host systems. In these studies, MuLAN was able to identify positive selection sites across different protein families, supporting the hypothesis that they were determinant in the mediation of interactions with specific partners. These results were also in agreement with other tools, not based on machine learning, offering a complementary yet diverse perspective that supported the

formulated hypotheses on the biological underlying mechanisms.

Concerning mutational effects on single protein sequences, different methodologies were developed to predict various types of fitness-related scores, from zero-shot scores derived from large foundational models, to protein or family-specific predictors that leverage homologous sequences, limited experimental data, or cross protein transfer to infer properties like thermostability or activity. On this subject, we worked on complementary strategies to augment zero-shot predictors with additional modalities, position-specific biases, or models combination conditioned on the property to predict and the characteristics of the considered protein. All these strategies showcased improved performances with respect to original baselines in well-established benchmarks, and can be combined to increase robustness and generalization.

Finally, we worked on the predictions of disorder from protein sequences. We benchmarked different PLMs on the prediction of intrinsic disordered, using LoRA fine-tuning with careful hyperparameters search on experimentally curated training data. We found that ESM-derived models better capture this notion of disorder with respect to ProtT5 architecture, and can achieve state-of-the-art performance on unseen sequences with limited training data, as demonstrated by CAID3 challenge results. Additionally, we constructed the SoftDis dataset from a non-redundant version of the crystallographic structures in the PDB, that provides labels for soft-disordered, missing, and interacting regions across tens of thousands of protein chains. We also trained classifiers for soft disorder on the new dataset, that offer a broader perspective on protein disorder and, indirectly, on folding models confidence, due to the high correlations between soft disorder predictions and AlphaFold pLDDT.

7.1 Key findings and outlook

The results obtained in this thesis provide useful insights, which open possible research directions that go beyond the specific applications considered here.

We mentioned that representations from PLMs can be successfully applied to downstream tasks involving both single proteins and protein-protein interaction. For single-token predictions, the implementation is straightforward, as even linear layers applied in parallel to each position in the embeddings usually offer good performances. When sufficient data are available, more complex layers or backbone fine-tuning can be introduced. However, one question that arises in these applications is how to extract the fine-grained information contained in PLMs-derived embeddings, while also reducing their dimensionality to get protein-level representations for global predictions. With MuLAN, we demonstrated that learned pooling strategies applied to single sequences, such as light attention, can be beneficial in this sense. One-dimensional convolutions contained in the light attention module learn local patterns in the sequences that, combined with the Siamese structure of the network, are determinant in the identification of hotspots for variations in binding affinity and interface regions. Depending on the specific task and PLM, other learned pooling strategies might provide better results, from recurrent networks to bidirectional attention layers. In addition, the dual applicability of MuLAN can be leveraged for experi-

mental validation and hypothesis generation, especially on less characterized organisms and systems, alone or combined with complementary tools.

Evaluation of many methods in the prediction of mutational effects in single sequences demonstrated that even the most advanced models, incorporating multiple sources of information, are still struggling when dealing with proteins lacking many homologs, with poorly predicted structures, disordered regions and strong epistatic effects. We showed that late fusion strategies, post-processing adaptations, or ensembling of model outputs can mitigate some flaws of individual models. However, a set of fundamental limitations remains, arising from several distinct factors.

First, the experimental data used for both training and evaluation are often noisy, heterogeneous, and condition-dependent, which directly affects the reliability of statistical inference. This issue is reflected, for example, in DMS benchmarks, where the best cross-protein transfer performance is frequently achieved by simple linear baselines rather than complex neural models. Second, both input availability and label distribution in biology are heavily affected by clinical and research biases: some protein families are represented by tens of thousands of homologs, while others are scarcely sampled. This asymmetry makes prediction tasks relatively straightforward in the well-represented families, often producing spuriously high correlations across methods, while leaving predictions unreliable and out-of-distribution for underrepresented families.

Third, biology itself presents intrinsically difficult cases, where the interplay between multiple levels of protein organization, including primary sequence, tertiary structure, surface geometry, conformational dynamics, and cellular environment, leads to context-dependent behaviors. Examples include IDPs, whose functional ensembles are not well captured by static models, or gain-of-function mutations, where pathogenicity emerges from subtle changes in regulatory or interaction dynamics rather than obvious structural disruption.

Future research directions that aim to integrate all these representations (sequence, structure, dynamics) in foundational models are promising, but still limited by data availability, as large databases of structural predictions like AlphaFoldDB inevitably introduces some form of circularity.

The application of PLMs representations to disorder predictions, either intrinsic or soft, currently leads to state-of-the-art performances. Task-specific adaptation of embeddings through LoRA layers can be leveraged to improve accuracy on related prediction tasks, particularly under conditions of data scarcity. An interesting direction for future research could be, for example, the use of soft disorder predictions for the inference of assembly pathways, due to the experimental evidence of their interplay.

7.2 Personal feedback

This last section is dedicated to my personal experience in the field of computational predictions of mutational effects, gained during the development of the thesis.

With the recent proliferation of deep learning models targeting mutation-related prediction tasks, the establishment of rigorous benchmarks for fair evaluation has become critical.

In the biological domain, in particular, the adoption of standardized preprocessing pipelines for widely used databases and the design of data splits that consider not only sequence similarity but also additional sources of data leakage specific to different input modalities would benefit the development of future models. ProteinGym represents a notable example in this direction, having provided a common evaluation framework that stimulated methodological advances, leading to competitive models adopting diverse strategies.

Another important consideration, which is often insufficiently addressed, concerns the balance between accuracy improvements and the computational costs required to achieve them. While larger PLMs tend to produce more informative representations, the distinctive nature of protein sequences, characterized by a relatively small vocabulary encoding highly complex biochemical semantics and physical constraints, affects the applicability of conventional scaling laws. Recent work, such as PoET-2, has shown that smaller multimodal architectures equipped with novel training strategies can outperform billion-parameter sequence-only transformers. This observation is particularly relevant for mutational effect prediction, where model complexity and predictive accuracy are only weakly correlated, and where late fusion of simpler models can often yield superior results compared to training increasingly larger models.

Finally, mechanistic interpretability of model outputs is particularly critical in the biological domain, especially when clinical studies are the ultimate goal. In MuLAN and LoRA-DR, we approached this aspect through indirect proxies of interpretability, such as attention weights and perturbations in attention-derived contact maps. While these representations provide valuable insights into potential residue-residue dependencies and highlight regions of biological relevance, they remain correlational rather than causal, and thus offer only a partial view of the models' internal reasoning. The introduction of novel approaches, such as the disentanglement of biological features through sparse autoencoders and the analysis of neural circuits in transformers, inspired by advances in NLP, represents a promising direction to enhance mechanistic interpretability of foundational PLMs and to improve their reliability and effectiveness in downstream biological tasks.

Addressing these limitations provides a concrete direction for future improvements: developing more efficient, multimodal models that capture the intrinsic complexity of biological objects, while still retaining interpretability of their predictions, would not only enhance the evaluation of mutational effects, but also extend their applicability to real-world clinical contexts, bridging the gap between fundamental and applied research.

Bibliography

[1] Linus Pauling, Robert B Corey, and Herman R Branson. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 37(4):205–211, 1951.

[2] Norma J Greenfield. Using circular dichroism spectra to estimate protein secondary structure. *Nature protocols*, 1(6):2876–2890, 2006.

[3] Georg E Schulz and R Heiner Schirmer. *Principles of protein structure*. Springer Science & Business Media, 2013.

[4] Wesley E Stites. Protein-protein interactions: interface structure, binding thermodynamics, and mutational analysis. *Chemical reviews*, 97(5):1233–1250, 1997.

[5] Errol C Friedberg, Graham C Walker, Wolfram Siede, and Richard D Wood. *DNA repair and mutagenesis*. American Society for Microbiology Press, 2005.

[6] Nimrat Chatterjee and Graham C Walker. Mechanisms of DNA damage, repair, and mutagenesis. *Environmental and molecular mutagenesis*, 58(5):235–263, 2017.

[7] Isabelle Touitou. The spectrum of familial Mediterranean fever (FMF) mutations. *European Journal of Human Genetics*, 9(7):473–483, 2001.

[8] Cyrus Chothia, Julian Gough, Christine Vogel, and Sarah A Teichmann. Evolution of the protein repertoire. *Science*, 300(5626):1701–1703, 2003.

[9] Julian Echave, Stephanie J Spielman, and Claus O Wilke. Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics*, 17(2):109–121, 2016.

[10] Julie D Thompson, Benjamin Linard, Odile Lecompte, and Olivier Poch. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PloS one*, 6(3):e18093, 2011.

[11] Florencio Pazos and Alfonso Valencia. Protein co-evolution, co-adaptation and interactions. *The EMBO journal*, 27(20):2648–2655, 2008.

[12] Michael S Breen, Carsten Kemena, Peter K Vlasov, Cedric Notredame, and Fyodor A Kondrashov. Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421):535–538, 2012.

[**13**] Daniel M Weinreich, Nigel F Delaney, Mark A DePristo, and Daniel L Hartl. Darwinian evolution can follow only very few mutational paths to fitter proteins. *science*, 312(5770):111–114, 2006.

[**14**] Merijn LM Salverda, Eynat Dellus, Florien A Gorter, Alfons JM Debets, John Van Der Oost, Rolf F Hoekstra, Dan S Tawfik, and J Arjan GM de Visser. Initial mutations direct alternative pathways of protein evolution. *PLoS genetics*, 7(3):e1001321, 2011.

[**15**] Clyde A Hutchison, Sandra Phillips, Marshall H Edgell, Shirley Gillam, Patricia Jahnke, and Michael Smith. Mutagenesis at a specific position in a DNA sequence. *Journal of Biological Chemistry*, 253(18):6551–6560, 1978.

[**16**] Randall K Saiki, David H Gelfand, Susanne Stoffel, Stephen J Scharf, Russell Higuchi, Glenn T Horn, Kary B Mullis, and Henry A Erlich. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239(4839):487–491, 1988.

[**17**] Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature methods*, 11(8):801–807, 2014.

[**18**] Lea M Starita, Jonathan N Pruneda, Russell S Lo, Douglas M Fowler, Helen J Kim, Joseph B Hiatt, Jay Shendure, Peter S Brzovic, Stanley Fields, and Rachel E Klevit. Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences*, 110(14):E1263–E1272, 2013.

[**19**] Patrick C Cirino, Kimberly M Mayer, and Daisuke Umeno. Generating mutant libraries using error-prone PCR. In *Directed evolution library creation: Methods and protocols*, pages 3–9. Springer, 2003.

[**20**] Douglas M Fowler, Carlos L Araya, Sarel J Fleishman, Elizabeth H Kellogg, Jason J Stephany, David Baker, and Stanley Fields. High-resolution mapping of protein sequence-function relationships. *Nature methods*, 7(9):741–746, 2010.

[**21**] Jeffrey I Boucher, Daniel NA Bolon, and Dan S Tawfik. Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature. *Protein Science*, 25(7):1219–1226, 2016.

[**22**] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic acids research*, 46(5):2699–2699, 2018.

[**23**] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[**24**] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover,

et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1):D862–D868, 2016.

[**25**] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in medicine*, 17(5):405–423, 2015.

[**26**] Andrew Folta, Adriana E Sedeño Cortés, Pankhuri Gupta, Abbye E McEwen, Eric Y Kao, Martha Horike-Pyne, Jeremy Stone, Brian H Shirts, Marianne E Dubard-Gault, Douglas M Fowler, et al. Imprecision medicine: Systematic gaps in reporting variants of uncertain significance (vus) and their reclassifications. *Genetics in Medicine*, page 101501, 2025.

[**27**] Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pages 16990–17017. PMLR, 2022.

[**28**] Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36:64331–64379, 2023.

[**29**] Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta PI Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2):128–135, 2017.

[**30**] Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, 15(10):816–822, 2018.

[**31**] Kotaro Tsuboyama, Justas Dauparas, Jonathan Chen, Elodie Laine, Yasser Mohseni Behbahani, Jonathan J Weinstein, Niall M Mangan, Sergey Ovchinnikov, and Gabriel J Rocklin. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973):434–444, 2023.

[**32**] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

[**33**] Sean R Eddy. Accelerated profile HMM searches. *PLoS computational biology*, 7(10):e1002195, 2011.

[**34**] Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

[**35**] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[**36**] Daniel Esposito, Jochen Weile, Jay Shendure, Lea M Starita, Anthony T Papenfuss, Frederick P Roth, Douglas M Fowler, and Alan F Rubin. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome biology*, 20(1):223, 2019.

[**37**] Alan F Rubin, Jeremy Stone, Aisha Haley Bianchi, Benjamin J Capodanno, Estelle Y Da, Mafalda Dias, Daniel Esposito, Jonathan Frazer, Yunfan Fu, Sally B Grindstaff, et al. MaveDB 2024: a curated community database with over seven million variant effects from multiplexed functional assays. *Genome biology*, 26(1):13, 2025.

[**38**] Justina Jankauskait, Brian Jiménez-García, Justas Dapk Unas, Juan Fernandez-Recio, and Iain H Moal. SKEMPI 2.0: an updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 07 2018.

[**39**] Emmanuel D Levy. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *Journal of molecular biology*, 403(4):660–670, 2010.

[**40**] H Jane Dyson and Peter E Wright. Intrinsically unstructured proteins and their functions. *Nature reviews Molecular cell biology*, 6(3):197–208, 2005.

[**41**] Peter E Wright and H Jane Dyson. Intrinsically disordered proteins in cellular signalling and regulation. *Nature reviews Molecular cell biology*, 16(1):18–29, 2015.

[**42**] A Keith Dunker, J David Lawson, Celeste J Brown, Ryan M Williams, Pedro Romero, Jeong S Oh, Christopher J Oldfield, Andrew M Campen, Catherine M Ratliff, Kerry W Hipps, et al. Intrinsically disordered protein. *Journal of molecular graphics and modelling*, 19(1):26–59, 2001.

[**43**] Kim Van Roey, Bora Uyar, Robert J Weatheritt, Holger Dinkel, Markus Seiler, Aidan Budd, Toby J Gibson, and Norman E Davey. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chemical reviews*, 114(13):6733–6778, 2014.

[**44**] Victor Neduva and Robert B Russell. Linear motifs: evolutionary interaction switches. *FEBS letters*, 579(15):3342–3345, 2005.

[**45**] Nancy Kedersha and Paul Anderson. Mammalian stress granules and processing bodies. *Methods in enzymology*, 431:61–81, 2007.

[**46**] Salman F Banani, Hyun O Lee, Anthony A Hyman, and Michael K Rosen. Biomolecular condensates: organizers of cellular biochemistry. *Nature reviews Molecular cell biology*, 18(5):285–298, 2017.

[**47**] Steven Boeynaems, Simon Alberti, Nicolas L Fawzi, Tanja Mittag, Magdalini Polymenidou, Frederic Rousseau, Joost Schymkowitz, James Shorter, Benjamin Wolozin, Ludo Van Den Bosch, et al. Protein phase separation: a new phase in cell biology. *Trends in cell biology*, 28(6):420–435, 2018.

[**48**] Anton Abyzov, Martin Blackledge, and Markus Zweckstetter. Conformational dynamics of intrinsically disordered proteins regulate biomolecular condensate chemistry. *Chemical Reviews*, 122(6):6719–6748, 2022.

[**49**] Tobias S Ulmer, Ad Bax, Nelson B Cole, and Robert L Nussbaum. Structure and dynamics of micelle-bound human α-synuclein. *Journal of Biological Chemistry*, 280(10):9595–9603, 2005.

[**50**] Tong Guo, Wendy Noble, and Diane P Hanger. Roles of tau protein in health and disease. *Acta neuropathologica*, 133(5):665–704, 2017.

[**51**] Beatriz Seoane and Alessandra Carbone. The complexity of protein interactions unravelled from structural disorder. *PLOS Computational Biology*, 17(1):e1008546, 2021.

[**52**] Beatriz Seoane and Alessandra Carbone. Soft disorder modulates the assembly path of protein complexes. *PLOS Computational Biology*, 18(11):e1010713, 2022.

[**53**] Damiano Piovesan, Francesco Tabaro, Ivan Mičetić, Marco Necci, Federica Quaglia, Christopher J Oldfield, Maria Cristina Aspromonte, Norman E Davey, Radoslav Davidović, Zsuzsanna Dosztányi, et al. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic acids research*, 45(D1):D219–D227, 2017.

[**54**] Alessio Del Conte, Adel Bouhraoua, Mahta Mehdiabadi, Damiano Clementel, Alexander Miguel Monzon, Silvio CE Tosatto, and Damiano Piovesan. CAID prediction portal: a comprehensive service for predicting intrinsic disorder and binding regions in proteins. *Nucleic Acids Research*, 51(W1):W62–W69, 2023.

[**55**] Damiano Piovesan, Alessio Del Conte, Mahta Mehdiabadi, Maria Cristina Aspromonte, Matthias Blum, Giulio Tesei, Sören von Bülow, Kresten Lindorff-Larsen, and Silvio CE Tosatto. MOBIDB in 2025: integrating ensemble properties and function annotations for intrinsically disordered proteins. *Nucleic Acids Research*, 53(D1):D495–D503, 2025.

[**56**] Satoshi Fukuchi, Shigetaka Sakamoto, Yukiko Nobe, Seiko D Murakami, Takayuki Amemiya, Kazuo Hosoda, Ryotaro Koike, Hidekazu Hiroaki, and Motonori Ota. IDEAL: intrinsically disordered proteins with extensive annotations and literature. *Nucleic acids research*, 40(D1):D507–D511, 2012.

[**57**] Andras Hatos, Alexander Miguel Monzon, Silvio CE Tosatto, Damiano Piovesan, and Monika Fuxreiter. FuzDB: a new phase in understanding fuzzy interactions. *Nucleic acids research*, 50(D1):D509–D517, 2022.

[**58**] Erzsébet Fichó, Rita Pancsa, Csaba Magyar, Zsofia E Kalman, Éva Schád, Bálint Z Németh, István Simon, Laszlo Dobson, and Gábor E Tusnády. MFIB 2.0: a major update of the database of protein complexes formed by mutual folding of the constituting protein chains. *Nucleic acids research*, 53(D1):D487–D494, 2025.

[**59**] Hamidreza Ghafouri, Tamas Lazar, Alessio Del Conte, Luiggi G Tenorio Ku, Peter Tompa, Silvio CE Tosatto, and Alexander Miguel Monzon. PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins. *Nucleic acids research*, 52(D1):D536–D544, 2024.

[**60**] Christian B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[**61**] Nils J Nilsson. *Artificial intelligence: a new synthesis*. Elsevier, 1998.

[**62**] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[**63**] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[**64**] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[**65**] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[**66**] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

[**67**] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[**68**] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[**69**] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[**70**] James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. A primer on deep learning in genomics. *Nature Genetics*, 51(1):12–18, 2019.

[**71**] Jie Hou, Badri K. Adhikari, and Jianlin Cheng. DeepSF: deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics*, 34(8):1295–1303, 2018.

[**72**] Kevin K Yang, Nicolo Fusi, and Alex X Lu. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294, 2024.

[**73**] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649, 2013.

[**74**] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

[**75**] Zhen Shen, Wenzheng Bao, and De-Shuang Huang. Recurrent neural network for predicting transcription factor binding sites. *Scientific reports*, 8(1):15270, 2018.

[**76**] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[**77**] Corentin Tallec and Yann Ollivier. Can recurrent neural networks warp time? *arXiv preprint arXiv:1804.11188*, 2018.

[**78**] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014.

[**79**] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[**80**] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[**81**] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[**82**] Craig W Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. Tokenization is more than compression. *arXiv preprint arXiv:2402.18376*, 2024.

[**83**] Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. Fast WordPiece Tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103, 2021.

[**84**] Guolin Ke, Di He, and Tie-Yan Liu. Rethinking Positional Encoding in Language Pre-training. In *International Conference on Learning Representations*, 2021.

[**85**] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36:24892–24928, 2023.

[**86**] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[**87**] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

[**88**] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[**89**] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[**90**] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[**91**] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[**92**] Mary Phuong and Marcus Hutter. Formal algorithms for transformers. *arXiv preprint arXiv:2207.09238*, 2022.

[**93**] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[**94**] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

[**95**] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

[**96**] Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! What makes rotary positional encodings useful? *arXiv preprint arXiv:2410.06205*, 2024.

[**97**] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

[**98**] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[**99**] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[**100**] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[**101**] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 87(1):012707, 2013.

[**102**] Matteo Figliuzzi, Pierre Barrat-Charlaix, and Martin Weigt. How pairwise coevolutionary models capture the collective residue variability in proteins? *Molecular biology and evolution*, 35(4):1018–1027, 2018.

[**103**] Matteo Figliuzzi, Hervé Jacquier, Alexander Schug, Oliver Tenaillon, and Martin Weigt. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Molecular biology and evolution*, 33(1):268–280, 2016.

[**104**] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

[**105**] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA Transformer. In *International conference on machine learning*, pages 8844–8856. PMLR, 2021.

[**106**] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664):eadg7492, 2023.

[**107**] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

[**108**] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 118(15), April 2021.

[**109**] Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized protein language model unlocks general-purpose modelling. *arXiv preprint arXiv:2301.06568*, 2023.

[**110**] Minh Hoang and Mona Singh. Locality-aware pooling enhances protein language model performance across varied applications. *Bioinformatics*, 41(Supplement_1):i217–i226, 2025.

[**111**] Nicki Skafte Detlefsen, Søren Hauberg, and Wouter Boomsma. Learning meaningful representations of protein sequences. *Nature communications*, 13(1):1914, 2022.

[**112**] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[**113**] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

[**114**] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. ProGen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.

[**115**] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 2025.

[**116**] Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein language modeling with structure-aware vocabulary. *BioRxiv*, pages 2023–10, 2023.

[**117**] Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Foldseek: fast and accurate protein structure search. *Biorxiv*, pages 2022–02, 2022.

[**118**] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4):lqae150, 2024.

[**119**] Ning Sun, Shuxian Zou, Tianhua Tao, Sazan Mahbub, Dian Li, Yonghao Zhuang, Hongyi Wang, Xingyi Cheng, Le Song, and Eric P Xing. Mixture of experts enable efficient and effective protein understanding and design. *bioRxiv*, pages 2024–11, 2024.

[**120**] Pan Li, Xingyi Cheng, Le Song, and Eric Xing. Retrieval Augmented Protein Language Models for Protein Structure Prediction. *bioRxiv*, pages 2024–12, 2024.

[**121**] Mingchen Li, Yang Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Pan Tan, and Liang Hong. ProSST: Protein language modeling with quantized structure and disentangled attention. *Advances in Neural Information Processing Systems*, 37:35700–35726, 2024.

[**122**] Fan Hu, Yishen Hu, Weihong Zhang, Huazhen Huang, Yi Pan, and Peng Yin. A multimodal protein representation framework for quantifying transferability across biochemical downstream tasks. *Advanced Science*, 10(22):2301223, 2023.

[**123**] Yang Tan, Ruilin Wang, Banghao Wu, Liang Hong, and Bingxin Zhou. From high-throughput evaluation to wet-lab studies: advancing mutation effect prediction with a retrieval-enhanced model. *Bioinformatics*, 41(Supplement_1):i401–i409, 2025.

[**124**] Elodie Laine, Yasaman Karami, and Alessandra Carbone. GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects. *Molecular Biology and Evolution*, 36:2604–2619, 11 2019.

[**125**] Stefan Engelen, Ladislas A Trojan, Sophie Sacquin-Mora, Richard Lavery, and Alessandra Carbone. Joint Evolutionary Trees: a large-scale method to pre-

dict protein interfaces based on sequence sampling. *PLoS computational biology*, 5(1):e1000267, 2009.

[**126**] Elodie Laine and Alessandra Carbone. Local geometry and evolutionary conservation of protein surfaces reveal the multiple recognition patches in protein-protein interactions. *PLoS computational biology*, 11(12):e1004580, 2015.

[**127**] Mihaly Mezei. A new method for mapping macromolecular topography. *Journal of Molecular Graphics and Modelling*, 21(5):463–472, 2003.

[**128**] Mustafa Tekpinar, Laurent David, Thomas Henry, and Alessandra Carbone. PRESCOTT: a population aware, epistatic, and structural model accurately predicts missense effects. *Genome Biology*, 26(1):1–42, 2025.

[**129**] Sanna Gudmundsson, Moriel Singer-Berk, Nicholas A Watts, William Phu, Julia K Goodrich, Matthew Solomonson, Genome Aggregation Database Consortium, Heidi L Rehm, Daniel G MacArthur, and Anne O'Donnell-Luria. Variant interpretation using population databases: Lessons from gnomAD. *Human mutation*, 43(8):1012–1030, 2022.

[**130**] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.

[**131**] Steffanie Paul, Aaron Kollasch, Pascal Notin, and Debora Marks. Combining structure and sequence for superior fitness prediction. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

[**132**] Pauline C Ng and Steven Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003.

[**133**] Ivan Adzhubei, Daniel M Jordan, and Shamil R Sunyaev. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics*, 76(1):7–20, 2013.

[**134**] Florian Gnad, Albion Baucom, Kiran Mukhyala, Gerard Manning, and Zemin Zhang. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC genomics*, 14(Suppl 3):S7, 2013.

[**135**] Yongwook Choi and Agnes P Chan. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, 31(16):2745–2747, 2015.

[**136**] Najmeh Alirezaie, Kristin D Kernohan, Taila Hartley, Jacek Majewski, and Toby Dylan Hocking. ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. *The American Journal of Human Genetics*, 103(4):474–483, 2018.

[**137**] Vikas Pejaver, Jorge Urresti, Jose Lugo-Martinez, Kymberleigh A Pagel, Guan Ning Lin, Hyun-Jun Nam, Matthew Mort, David N Cooper, Jonathan Sebat, Lilia M Iakoucheva, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nature communications*, 11(1):5918, 2020.

[**138**] Douglas EV Pires, David B Ascher, and Tom L Blundell. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30(3):335–342, 2014.

[**139**] Nilah M Ioannidis, Joseph H Rothstein, Vikas Pejaver, Sumit Middha, Shannon K McDonnell, Saurabh Baheti, Anthony Musolf, Qing Li, Emily Holzinger, Danielle Karyadi, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics*, 99(4):877–885, 2016.

[**140**] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*, 47(D1):D886–D894, 2019.

[**141**] Laksshman Sundaram, Hong Gao, Samskruthi Reddy Padigepati, Jeremy F McRae, Yanjun Li, Jack A Kosmicki, Nondas Fritzilas, Jörg Hakenberg, Anindita Dutta, John Shon, et al. Predicting the clinical impact of human mutation with deep neural networks. *Nature genetics*, 50(8):1161–1170, 2018.

[**142**] Hong Gao, Tobias Hamp, Jeffrey Ede, Joshua G Schraiber, Jeremy McRae, Moriel Singer-Berk, Yanshen Yang, Anastasia SD Dietrich, Petko P Fiziev, Lukas FK Kuderna, et al. The landscape of tolerated genetic variation in humans and primates. *Science*, 380(6648):eabn8153, 2023.

[**143**] Huali Cao, Jingxue Wang, Liping He, Yifei Qi, and John Z Zhang. DeepDDG: predicting the stability change of protein point mutations using neural networks. *Journal of chemical information and modeling*, 59(4):1508–1514, 2019.

[**144**] Vanessa E Gray, Ronald J Hause, Jens Luebeck, Jay Shendure, and Douglas M Fowler. Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell systems*, 6(1):116–124, 2018.

[**145**] Milind Jagota, Chengzhong Ye, Carlos Albors, Ruchir Rastogi, Antoine Koehl, Nilah Ioannidis, and Yun S Song. Cross-protein transfer learning substantially improves disease variant prediction. *Genome Biology*, 24(1):182, 2023.

[**146**] Pascal Notin, Ruben Weitzman, Debora Marks, and Yarin Gal. ProteinNPT: Improving protein property prediction and design with non-parametric transformers. *Advances in Neural Information Processing Systems*, 36:33529–33563, 2023.

[**147**] Peter Mørch Groth, Mads Kerrn, Lars Olsen, Jesper Salomon, and Wouter Boomsma. Kermut: Composite kernel regression for protein variant effects. *Advances in Neural Information Processing Systems*, 37:29514–29565, 2024.

[**148**] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The FoldX web server: an online force field. *Nucleic acids research*, 33(suppl_2):W382–W388, 2005.

[**149**] Rhiju Das and David Baker. Macromolecular modeling with Rosetta. *Annu. Rev. Biochem.*, 77(1):363–382, 2008.

[**150**] Yves Dehouck, Jean Marc Kwasigroch, Marianne Rooman, and Dimitri Gilis. BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations. *Nucleic acids research*, 41(W1):W333–W339, 2013.

[**151**] Cunliang Geng, Anna Vangone, Gert E. Folkers, Li C. Xue, and Alexandre M.J.J. Bonvin. iSEE: Interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. *Proteins: Structure, Function and Bioinformatics*, 87:110–119, 2 2019.

[**152**] Swagata Pahari, Gen Li, Adithya Krishna Murthy, Siqi Liang, Robert Fragoza, Haiyuan Yu, and Emil Alexov. SAAMBE-3D: predicting effect of mutations on protein–protein interactions. *International journal of molecular sciences*, 21(7):2563, 2020.

[**153**] Gen Li, Swagata Pahari, Adithya Krishna Murthy, Siqi Liang, Robert Fragoza, Haiyuan Yu, and Emil Alexov. SAAMBE-SEQ: a sequence-based method for predicting mutation effect on protein–protein binding affinity. *Bioinformatics*, 37(7):992–999, 2021.

[**154**] Carlos HM Rodrigues, Yoochan Myung, Douglas EV Pires, and David B Ascher. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic acids research*, 47(W1):W338–W344, 2019.

[**155**] Ning Zhang, Yuting Chen, Haoyu Lu, Feiyang Zhao, Roberto Vera Alvarez, Alexander Goncearenco, Anna R Panchenko, and Minghui Li. MutaBind2: predicting the impacts of single and multiple mutations on protein-protein interactions. *Iscience*, 23(3), 2020.

[**156**] Peng Xiong, Chengxin Zhang, Wei Zheng, and Yang Zhang. BindProfX: Assessing Mutation-Induced Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts. *Journal of Molecular Biology*, 429:426–434, 2 2017.

[**157**] Guangyu Zhou, Muhao Chen, Chelsea J.T. Ju, Zheng Wang, Jyun Yu Jiang, and Wei Wang. Mutation effect estimation on protein–protein interactions using deep contextualized representation learning. *NAR Genomics and Bioinformatics*, 2, 6 2020.

[**158**] Menglun Wang, Zixuan Cang, and Guo Wei Wei. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nature Machine Intelligence*, 2:116–123, 2 2020.

[**159**] Yasser Mohseni Behbahani, Elodie Laine, and Alessandra Carbone. Deep Local Analysis deconstructs protein–protein interfaces and accurately estimates binding affinity changes upon mutation. *Bioinformatics*, 39(Supplement_1):i544–i552, 2023.

[**160**] Yunzhuo Zhou, YooChan Myung, Carlos HM Rodrigues, and David B Ascher. DDMut-PPI: predicting effects of mutations on protein–protein interactions using graph-based deep learning. *Nucleic Acids Research*, 52(W1):W207–W214, 2024.

[**161**] Zsuzsanna Dosztányi, Veronika Csizmok, Peter Tompa, and István Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434, 2005.

[**162**] Gábor Erdős and Zsuzsanna Dosztányi. AIUPred: combining energy estimation with deep learning for the enhanced prediction of protein disorder. *Nucleic Acids Research*, 52(W1):W176–W181, 2024.

[**163**] David T Jones and Domenico Cozzetto. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, 31(6):857–863, 2015.

[**164**] Marco Necci, Damiano Piovesan, Zsuzsanna Dosztányi, and Silvio CE Tosatto. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics*, 33(9):1402–1404, 2017.

[**165**] Jack Hanson, Kuldip K Paliwal, Thomas Litfin, and Yaoqi Zhou. SPOT-Disorder2: improved protein intrinsic disorder prediction by ensembled deep learning. *Genomics, proteomics & bioinformatics*, 17(6):645–656, 2019.

[**166**] Gang Hu, Akila Katuwawala, Kui Wang, Zhonghua Wu, Sina Ghadermarzi, Jianzhao Gao, and Lukasz Kurgan. flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nature Communications*, 12(1):4438, Jul 2021.

[**167**] Claudio Mirabello and Björn Wallner. rawMSA: end-to-end deep learning using raw multiple sequence alignments. *PloS one*, 14(8):e0220182, 2019.

[**168**] Dagmar Ilzhöfer, Michael Heinzinger, and Burkhard Rost. SETH predicts nuances of residue disorder from protein embeddings. *Frontiers in Bioinformatics*, 2:1019597, 2022.

[**169**] Elisa Cilia, Rita Pancsa, Peter Tompa, Tom Lenaerts, and Wim F Vranken. From protein sequence to dynamics and disorder with DynaMine. *Nature communications*, 4(1):2741, 2013.

[**170**] Rupashree Dass, Frans AA Mulder, and Jakob Toudahl Nielsen. ODiNPred: comprehensive prediction of protein order and disorder. *Scientific reports*, 10(1):14780, 2020.

[**171**] Yann Vander Meersche, Gabriel Cretin, Alexandre G de Brevern, Jean-Christophe Gelly, and Tatiana Galochkina. MEDUSA: prediction of protein flexibility from sequence. *Journal of molecular biology*, 433(11):166882, 2021.

[**172**] Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

[**173**] Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with TM-score= 0.5? *Bioinformatics*, 26(7):889–895, 2010.

[**174**] Panagiotis L Kastritis and Alexandre MJJ Bonvin. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of The Royal Society Interface*, 10(79):20120835, 2013.

[**175**] Cunliang Geng, Li C Xue, Jorge Roel-Touris, and Alexandre MJJ Bonvin. Finding the ΔΔG spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 9(5):e1410, 2019.

[**176**] Till Siebenmorgen and Martin Zacharias. Computational prediction of protein–protein binding affinities. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 10(3):e1448, 2020.

[**177**] Xianggen Liu, Yunan Luo, Pengyong Li, Sen Song, and Jian Peng. Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLoS computational biology*, 17(8):e1009284, 2021.

[**178**] Hao Xiong, Long Han, Yue Wang, and Pengxin Chai. Evaluating the Reliability of AlphaFold 2 for Unknown Complex Structures with Deep Learning. *bioRxiv*, 2022.

[**179**] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. *bioRxiv*, 2020.

[**180**] Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Henry Chao, Li Chen, Paul A Craig, Gregg V Crichlow, Kenneth Dalenberg, Jose M Duarte, et al. RCSB Protein Data Bank (RCSB. org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic acids research*, 51(D1):D488–D508, 2023.

[**181**] Chloé Dequeker, Elodie Laine, and Alessandra Carbone. INTerface Builder: A fast protein–protein interface reconstruction tool. *Journal of Chemical Information and Modeling*, 57(11):2613–2617, 2017.

[**182**] Hugues Ripoche, Elodie Laine, Nicoletta Ceres, and Alessandra Carbone. JET2 Viewer: a database of predicted multiple, possibly overlapping, protein–protein

interaction sites for PDB structures. *Nucleic acids research*, 45(D1):D236–D242, 2017.

[**183**] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

[**184**] Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Adel Bouharoua, Silvio CE Tosatto, and Damiano Piovesan. DisProt in 2024: improving function annotation of intrinsically disordered proteins. *Nucleic Acids Research*, 52(D1):D434–D441, 2024.

[**185**] Taojunfeng Su, Michael AR Hollas, Ryan T Fellers, and Neil L Kelleher. Identification of splice variants and isoforms in transcriptomics and proteomics. *Annual review of biomedical data science*, 6(1):357–376, 2023.

[**186**] Fabian Birzele, Gergely Csaba, and Ralf Zimmer. Alternative splicing and protein structure evolution. *Nucleic acids research*, 36(2):550–558, 2008.

[**187**] Loredana Lo Conte, Bart Ailey, Tim JP Hubbard, Steven E Brenner, Alexey G Murzin, and Cyrus Chothia. SCOP: a structural classification of proteins database. *Nucleic acids research*, 28(1):257–259, 2000.

[**188**] John-Marc Chandonia, Gary Hon, Nigel S Walker, Loredana Lo Conte, Patrice Koehl, Michael Levitt, and Steven E Brenner. The ASTRAL compendium in 2004. *Nucleic acids research*, 32(suppl_1):D189–D192, 2004.

[**189**] Michael Heyne, Jason Shirian, Itay Cohen, Yoav Peleg, Evette S. Radisky, Niv Papo, and Julia M. Shifman. Climbing Up and Down Binding Landscapes through Deep Mutational Scanning of Three Homologous Protein–Protein Complexes. *Journal of the American Chemical Society*, 143(41):17261–17275, 2021. PMID: 34609866.

[**190**] Jean-Francois Collet and Joris Messens. Structure, Function, and Mechanism of Thioredoxin Proteins. *Antioxidants & Redox Signaling*, 13(8):1205–1216, 2010. PMID: 20136512.

[**191**] Stéphane D Lemaire, Daniele Tedesco, Pierre Crozet, Laure Michelet, Simona Fermani, Mirko Zaffagnini, and Julien Henri. Crystal structure of chloroplastic thioredoxin f2 from *Chlamydomonas reinhardtii* reveals distinct surface properties. *Antioxidants*, 7(12):171, 2018.

[**192**] Laure Michelet, Mirko Zaffagnini, Vincent Massot, Eliane Keryer, Hélène Vanacker, Myroslawa Miginiac-Maslow, Emmanuelle Issakidis-Bourguet, and Stéphane D Lemaire. Thioredoxins, glutaredoxins, and glutathionylation: new crosstalks to explore. *Photosynthesis Research*, 89(2):225–245, 2006.

[**193**] R Vicedomini, JP Bouly, E Laine, A Falciatore, and A Carbone. Multiple profile models extract features from protein sequence data and resolve functional diversity

of very different protein families. *Molecular biology and evolution*, 39(4):msac070, 2022.

[**194**] Stéphane D Lemaire, Gianluca Lombardi, Andrea Mancini, Alessandra Carbone, and Julien Henri. Functional recoding of *Chlamydomonas reinhardtii* thioredoxin type-h into photosynthetic type-f by switching selectivity determinants. *Frontiers in Plant Science*, 16:1554272, 2025.

[**195**] Hannes Stärk, Christian Dallago, Michael Heinzinger, and Burkhard Rost. Light attention predicts protein location from the language of life. *Bioinformatics Advances*, 1, 6 2021.

[**196**] Yasser Mohseni Behbahani, Paul Saighi, Flavia Corsi, Elodie Laine, and Alessandra Carbone. LEVELNET to visualize, explore, and compare protein-protein interaction networks. *Proteomics*, 2023.

[**197**] Sarah Sirin, James R Apgar, Eric M Bennett, and Amy E Keating. AB-Bind: antibody binding mutational database for computational affinity predictions. *Protein Science*, 25(2):393–409, 2016.

[**198**] Sherlyn Jemimah, K Yugandhar, and M Michael Gromiha. PROXiMATE: a database of mutant protein–protein complex thermodynamics and kinetics. *Bioinformatics*, 33(17):2787–2788, 2017.

[**199**] Emmanuel D. Levy. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *Journal of Molecular Biology*, 403(4):660–670, 2010.

[**200**] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, et al. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, 2024.

[**201**] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023.

[**202**] Wei Lu, Jixian Zhang, Jiahua Rao, Zhongyue Zhang, and Shuangjia Zheng. AlphaFold3, a secret sauce for predicting mutational effects on protein-protein interactions. *bioRxiv*, 2024.

[**203**] J. Abramson, J. Adler, J. Dunger, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 2024.

[**204**] Arne Holmgren. Thioredoxin. *Annual review of biochemistry*, 54(1):237–271, 1985.

[**205**] Elias SJ Arnér and Arne Holmgren. Physiological functions of thioredoxin and thioredoxin reductase. *European journal of biochemistry*, 267(20):6102–6109, 2000.

[**206**] Jun Qin, G Marius Clore, and Angela M Gronenborn. The high-resolution three-dimensional solution structures of the oxidized and reduced states of human thioredoxin. *Structure*, 2(6):503–522, 1994.

[**207**] Hans Eklund, Margareta Ingelman, Bengt-Olof Söderberg, Tomas Uhlin, Pär Nordlund, Matti Nikkola, Ulf Sonnerstam, Thorleif Joelson, and Kyriacos Petratos. Structure of oxidized bacteriophage T4 glutaredoxin (thioredoxin): refinement of native and mutant proteins. *Journal of molecular biology*, 228(2):596–618, 1992.

[**208**] Stéphane D Lemaire and Myroslawa Miginiac-Maslow. The thioredoxin superfamily in *Chlamydomonas reinhardtii*. *Photosynthesis Research*, 82(3):203–220, 2004.

[**209**] Stéphane D Lemaire, Laure Michelet, Mirko Zaffagnini, Vincent Massot, and Emmanuelle Issakidis-Bourguet. Thioredoxins in chloroplasts. *Current genetics*, 51(6):343–365, 2007.

[**210**] Jean-Marc Lancelin, Laure Guilhaudis, Isabelle Krimm, Martin J Blackledge, Dominique Marion, and Jean-Pierre Jacquot. NMR structures of thioredoxin m from the green alga *Chlamydomonas reinhardtii*. *Proteins: Structure, Function, and Bioinformatics*, 41(3):334–349, 2000.

[**211**] Théo Le Moigne, Libero Gurrieri, Pierre Crozet, Christophe H Marchand, Mirko Zaffagnini, Francesca Sparla, Stéphane D Lemaire, and Julien Henri. Crystal structure of chloroplastic thioredoxin z defines a type-specific target recognition. *The Plant Journal*, 107(2):434–447, 2021.

[**212**] Victoria Bunik, Günter Raddatz, Stephane Lemaire, Yves Meyer, Jean-Pierre Jacquot, and Hans Bisswanger. Interaction of thioredoxins with target proteins: role of particular structural elements and electrostatic properties of thioredoxins in their interplay with 2-oxoacid dehydrogenase complexes. *Protein Science*, 8(1):65–74, 1999.

[**213**] Jean-Pierre Jacquot, Pierre Gadal, Ann N Nishizawa, Boihon C Yee, Nancy A Crawford, and Bob B Buchanan. Enzyme regulation in C4 photosynthesis: mechanism of activation of NADP-malate dehydrogenase by reduced thioredoxin. *Archives of Biochemistry and Biophysics*, 228(1):170–178, 1984.

[**214**] Paulette Decottignies, JM Schmitter, Myroslawa Miginiac-Maslow, P Le Maréchal, JP Jacquot, and Pierre Gadal. Primary structure of the light-dependent regulatory site of corn NADP-malate dehydrogenase. *Journal of Biological Chemistry*, 263(24):11780–11785, 1988.

[**215**] Peter Schürmann and Bob B Buchanan. Role of ferredoxin in the activation of sedoheptulose diphosphatase in isolated chloroplasts. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 376(1):189–192, 1975.

[**216**] Peter Schürmann and Richardo A Wolosiuk. Studies on the regulatory properties of chloroplast fructose-1, 6-bisphosphatase. *Biochimica et Biophysica Acta (BBA)-Enzymology*, 522(1):130–138, 1978.

[**217**] Michel Droux, J-P Jacquot, M Miginac-Maslow, P Gadal, JC Huet, NA Crawford, BC Yee, and BB Buchanan. Ferredoxin-thioredoxin reductase, an iron-sulfur enzyme linking light to enzyme regulation in oxygenic photosynthesis: purification and properties of the enzyme from C3, C4, and cyanobacterial species. *Archives of biochemistry and biophysics*, 252(2):426–439, 1987.

[**218**] Keisuke Yoshida, Satoshi Hara, and Toru Hisabori. Thioredoxin selectivity for thiol-based redox regulation of target proteins in chloroplasts. *Journal of Biological Chemistry*, 290(23):14278–14288, 2015.

[**219**] Yves Balmer, Antonius Koller, Greg del Val, Peter Schürmann, and Bob B Buchanan. Proteomics uncovers proteins interacting electrostatically with thioredoxin in chloroplasts. *Photosynthesis Research*, 79(3):275–280, 2004.

[**220**] Keisuke Yoshida and Toru Hisabori. Current insights into the redox regulation network in plant chloroplasts. *Plant and Cell Physiology*, 64(7):704–715, 2023.

[**221**] Christophe H Marchand, Simona Fermani, Jacopo Rossi, Libero Gurrieri, Daniele Tedesco, Julien Henri, Francesca Sparla, Paolo Trost, Stéphane D Lemaire, and Mirko Zaffagnini. Structural and Biochemical Insights into the Reactivity of Thioredoxin h1 from *Chlamydomonas reinhardtii*. *Antioxidants*, 8(1):10, 2019.

[**222**] Aurélie Mesnil, Vincent Castric, Gianluca Lombardi, Marie Monniaux, Somia Saidi, Xilong Chen, William Marande, Johann Confais, Anne-Laure Fuchs, Enrique Dapena, et al. Genomic architecture of the self-incompatibility locus in apple provides insights into the evolution of collaborative non-self recognition. *bioRxiv*, pages 2025–10, 2025.

[**223**] Adrian John Richards. *Plant breeding systems*. Garland Science, 1997.

[**224**] Harold LK Whitehouse. Multiple-allelomorph incompatibility of pollen and style in the evolution of the angiosperms. *Annals of Botany*, 14(54):199–216, 1950.

[**225**] Emmanuelle Porcher and Russell Lande. Loss of gametophytic self-incompatibility with evolution of inbreeding depression. *Evolution*, 59(1):46–60, 2005.

[**226**] Hidenori Sassa, Hiroyuki Kakui, and Mai Minamikawa. Pollen-expressed F-box gene family and mechanism of S-RNase-based gametophytic self-incompatibility (GSI) in Rosaceae. *Sexual plant reproduction*, 23(1):39–43, 2010.

[**227**] Michael D Nowak, Aaron P Davis, Francois Anthony, and Anne D Yoder. Expression and trans-specific polymorphism of self-incompatibility RNases in *Coffea* (Rubiaceae). *PloS one*, 6(6):e21019, 2011.

[**228**] Tetsuyuki Entani, Ken-ichi Kubo, Shin Isogai, Yoichiro Fukao, Masahiro Shirakawa, Akira Isogai, and Seiji Takayama. Ubiquitin–proteasome-mediated degradation of S-RNase in a solanaceous cross-compatibility reaction. *The Plant Journal*, 78(6):1014–1021, 2014.

[**229**] Ken-ichi Kubo, Tetsuyuki Entani, Akie Takara, Ning Wang, Allison M Fields, Zhihua Hua, Mamiko Toyoda, Shin-ichi Kawashima, Toshio Ando, Akira Isogai, et al. Collaborative non-self recognition system in S-RNase–based self-incompatibility. *Science*, 330(6005):796–799, 2010.

[**230**] Chi Xu, Maofu Li, Junkai Wu, Han Guo, Qun Li, Yu'e Zhang, Jijie Chai, Tianzhong Li, and Yongbiao Xue. Identification of a canonical SCFSLF complex involved in S-RNase-based self-incompatibility of *Pyrus* (Rosaceae). *Plant molecular biology*, 81(3):245–257, 2013.

[**231**] Seiji Takayama and Akira Isogai. Self-incompatibility in plants. *Annu. Rev. Plant Biol.*, 56(1):467–489, 2005.

[**232**] Xavier Vekemans and Vincent Castric. When the genetic architecture matters. *The New Phytologist*, 231(4):1304–1307, 2021.

[**233**] Ken-ichi Kubo, Timothy Paape, Masaomi Hatakeyama, Tetsuyuki Entani, Akie Takara, Kie Kajihara, Mai Tsukahara, Rie Shimizu-Inatsugi, Kentaro K Shimizu, and Seiji Takayama. Gene duplication and genetic exchange drive the evolution of S-RNase-based self-incompatibility in *Petunia*. *Nature Plants*, 1(1):1–9, 2015.

[**234**] X Ma, Z Cai, W Liu, S Ge, and L Tang. Identification, genealogical structure and population genetics of S-alleles in *Malus sieversii*, the wild ancestor of domesticated apple. *Heredity*, 119(3):185–196, 2017.

[**235**] Ziheng Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591, 2007.

[**236**] Takeshi Ishimizu, Toshinori Endo, Yumi Yamaguchi-Kabata, Kazuo T Nakamura, Fumio Sakiyama, and Shigemi Norioka. Identification of regions in which positive selection may operate in S-RNase of Rosaceae: implication for S-allele-specific recognition sites in S-RNase. *Febs Letters*, 440(3):337–342, 1998.

[**237**] K Ushijima, H Sassa, R Tao, H Yamane, AM Dandekar, TM Gradziel, and H Hirano. Cloning and characterization of cDNAs encoding S-RNases from almond (*Prunus dulcis*): primary structural features and sequence diversity of the S-RNases in Rosaceae. *Molecular and General Genetics MGG*, 260(2):261–268, 1998.

[**238**] Jianke Du, Chunfeng Ge, Tingting Li, Sanhong Wang, Zhihong Gao, Hidenori Sassa, and Yushan Qiao. Molecular characteristics of S-RNase alleles as the determinant of self-incompatibility in the style of *Fragaria viridis*. *Horticulture Research*, 8, 2021.

[**239**] Sergio Gabriel Olvera-Vazquez, Xilong Chen, Aurélie Mesnil, Camille Meslin, Fabricio Almeida-Silva, Johann Confais, Yann Bourgeois, Gianluca Lombardi, Célia Lougmani, Karine Alix, et al. Comprehensive annotation of olfactory and gustatory receptor genes and transposable elements revealed their evolutionary dynamics in aphids. *Molecular Biology and Evolution*, page msaf238, 2025.

[**240**] Pierre Gladieux, Jeanne Ropars, Hélène Badouin, Antoine Branca, Gabriela Aguileta, Damien M De Vienne, Ricardo C Rodríguez De La Vega, Sara Branco, and Tatiana Giraud. Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. *Molecular ecology*, 23(4):753–773, 2014.

[**241**] Jean-Christophe Simon, Emmanuelle d'Alencon, Endrick Guy, Emmanuelle Jacquin-Joly, Julie Jaquiery, Pierre Nouhaud, Jean Peccoud, Akiko Sugio, and Rejane Streiff. Genomics of adaptation to host-plants in herbivorous insects. *Briefings in functional genomics*, 14(6):413–423, 2015.

[**242**] Mareike Möller and Eva H Stukenbrock. Evolution and genome architecture in fungal plant pathogens. *Nature Reviews Microbiology*, 15(12):756–771, 2017.

[**243**] Heather Kirk, Silvia Dorn, and Dominique Mazzi. Molecular genetics and genomics generate new insights into invertebrate pest invasions. *Evolutionary Applications*, 6(5):842–856, 2013.

[**244**] Simone M Cacciò, Rachel M Chalmers, Pierre Dorny, and Lucy J Robertson. Foodborne parasites: Outbreaks and outbreak investigations. A meeting report from the European network for foodborne parasites (Euro-FBP). *Food and Waterborne Parasitology*, 10:1–5, 2018.

[**245**] Nicholas Hamilton Barton. Natural selection and random genetic drift as causes of evolution on islands. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 351(1341):785–795, 1996.

[**246**] Michael Lynch. Mutation pressure, drift, and the pace of molecular coevolution. *Proceedings of the National Academy of Sciences*, 120(27):e2306741120, 2023.

[**247**] JC Fay and PJ Wittkopp. Evaluating the role of natural selection in the evolution of gene regulation. *Heredity*, 100(2):191–199, 2008.

[**248**] Elinor K Karlsson, Dominic P Kwiatkowski, and Pardis C Sabeti. Natural selection and infectious disease in human populations. *Nature Reviews Genetics*, 15(6):379–393, 2014.

[**249**] Sébastien Duplessis, Christina A Cuomo, Yao-Cheng Lin, Andrea Aerts, Emilie Tisserant, Claire Veneault-Fourrey, David L Joly, Stéphane Hacquard, Joëlle Amselem, Brandi L Cantarel, et al. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proceedings of the National Academy of Sciences*, 108(22):9166–9171, 2011.

[**250**] Fanny E Hartmann, Andrea Sánchez-Vallet, Bruce A McDonald, and Daniel Croll. A fungal wheat pathogen evolved host specialization by extensive chromosomal rearrangements. *The ISME journal*, 11(5):1189–1204, 2017.

[**251**] Clémence Plissonneau, Fanny E Hartmann, and Daniel Croll. Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. *BMC biology*, 16(1):5, 2018.

[**252**] Samuel O'donnell, Gabriela Rezende, Jean-Philippe Vernadet, Alodie Snirc, and Jeanne Ropars. Harbouring Starships: the accumulation of large horizontal gene transfers in domesticated and pathogenic fungi. *Genome Biology and Evolution*, page evaf125, 2025.

[**253**] Ioannis Stergiopoulos and Pierre JGM De Wit. Fungal effector proteins. *Annual review of phytopathology*, 47(1):233–263, 2009.

[**254**] David E Torres, Ursula Oggenfuss, Daniel Croll, and Michael F Seidl. Genome evolution in fungal plant pathogens: looking beyond the two-speed genome model. *Fungal Biology Reviews*, 34(3):136–143, 2020.

[**255**] Tyler A Elliott and T Ryan Gregory. Do larger genomes contain more diverse transposable elements? *BMC evolutionary biology*, 15(1):69, 2015.

[**256**] Clément Gilbert, Jean Peccoud, and Richard Cordaux. Transposable elements and the evolution of insects. *Annual Review of Entomology*, 66(1):355–372, 2021.

[**257**] Camille Meslin, Pauline Mainet, Nicolas Montagné, Stéphanie Robin, Fabrice Legeai, Anthony Bretaudeau, J Spencer Johnston, Fotini Koutroumpa, Emma Persyn, Christelle Monsempès, et al. *Spodoptera littoralis* genome mining brings insights on the dynamic of expansion of gustatory receptors in polyphagous noctuidae. *G3*, 12(8):jkac131, 2022.

[**258**] Carole Smadja, Peng Shi, Roger K Butlin, and Hugh M Robertson. Large gene family expansions and adaptive evolution for odorant and gustatory receptors in the pea aphid, *Acyrthosiphon pisum*. *Molecular biology and evolution*, 26(9):2073–2086, 2009.

[**259**] Bill S Hansson and Marcus C Stensmyr. Evolution of insect olfaction. *Neuron*, 72(5):698–711, 2011.

[**260**] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt Von Haeseler, and Robert Lanfear. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, 37(5):1530–1534, 2020.

[**261**] Hugh M Robertson. The insect chemoreceptor superfamily is ancient in animals. *Chemical senses*, 40(9):609–614, 2015.

[**262**] Seong-il Eyun, Ho Young Soh, Marijan Posavi, James B Munro, Daniel ST Hughes, Shwetha C Murali, Jiaxin Qu, Shannon Dugan, Sandra L Lee, Hsu Chao, et al. Evolutionary history of chemosensory-related gene families across the Arthropoda. *Molecular biology and evolution*, 34(8):1838–1862, 2017.

[**263**] Nathan C Sheffield and Christoph Bock. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics*, 32(4):587–589, 2016.

[**264**] Caroline Meguerditchian, Ayse Ergun, Veronique Decroocq, Marie Lefebvre, and Quynh-Trang Bui. Pipeline to detect the positional and directional relationship between transposable elements and adjacent genes in host genome. *Peer Community Journal*, 1, 2021.

[**265**] Motoo Kimura. Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences*, 78(1):454–458, 1981.

[**266**] Franziska Beran and Georg Petschenka. Sequestration of plant defense compounds by insects: from mechanisms to insect–plant coevolution. *Annual Review of Entomology*, 67(1):163–180, 2022.

[**267**] Ryan M Joseph and John R Carlson. Drosophila chemoreceptors: a molecular interface between the chemical world and the brain. *Trends in Genetics*, 31(12):683–695, 2015.

[**268**] Andrea Mancini, Vinh-Son Pho, Alessandro Bianchi, Gianluca Lombardi, Chujun Lyu, and Alessandra Carbone. Scaling the profile of life by function with spin. *Bioinformatics Advances*, 6(1):vbag064, 2026.

[**269**] Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

[**270**] Brandi L Cantarel, Pedro M Coutinho, Corinne Rancurel, Thomas Bernard, Vincent Lombard, and Bernard Henrissat. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic acids research*, 37(suppl_1):D233–D238, 2009.

[**271**] Ivica Letunic and Peer Bork. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic acids research*, 52(W1):W78–W82, 2024.

[**272**] Gianluca Lombardi, Beatriz Seoane, and Alessandra Carbone. LoRA-DR-suite: adapted embeddings predict intrinsic and soft disorder from protein sequences. *Bioinformatics*, 41(Supplement_1):i439–i448, 07 2025.

[**273**] Peter Tompa. Intrinsically unstructured proteins. *Trends in biochemical sciences*, 27(10):527–533, 2002.

[**274**] A Keith Dunker, Pedro Romero, Zoran Obradovic, Ethan C Garner, and Celeste J Brown. Intrinsic protein disorder in complete genomes. *Genome informatics*, 11:161–171, 2000.

[**275**] Bálint Mészáros, István Simon, and Zsuzsanna Dosztányi. Prediction of protein binding regions in disordered proteins. *PLoS computational biology*, 5(5):e1000376, 2009.

[**276**] Gang Hu, Kui Wang, Jiangning Song, Vladimir N Uversky, and Lukasz Kurgan. Taxonomic landscape of the dark proteomes: Whole-proteome scale interplay between structural darkness, intrinsic disorder, and crystallization propensity. *Proteomics*, 18(21-22):1800243, 2018.

[**277**] Zhoutong Sun, Qian Liu, Ge Qu, Yan Feng, and Manfred T Reetz. Utility of B-factors in protein science: interpreting rigidity, flexibility, and internal motion and engineering thermostability. *Chemical reviews*, 119(3):1626–1665, 2019.

[**278**] Vladimir N Uversky. Intrinsic disorder, protein–protein interactions, and disease. *Advances in protein chemistry and structural biology*, 110:85–121, 2018.

[**279**] Ana M Melo, Juliana Coraor, Garrett Alpha-Cobb, Shana Elbaum-Garfinkle, Abhinav Nath, and Elizabeth Rhoades. A functional role for intrinsic disorder in the tau-tubulin complex. *Proceedings of the National Academy of Sciences*, 113(50):14336–14341, 2016.

[**280**] Kumlesh K Dev, Katja Hofele, Samuel Barbieri, Vladimir L Buchman, and Herman van der Putten. Part II: α-synuclein and its molecular pathophysiological role in neurodegenerative disease. *Neuropharmacology*, 45(1):14–44, 2003.

[**281**] Lilia M Iakoucheva, Celeste J Brown, J David Lawson, Zoran Obradović, and A Keith Dunker. Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of molecular biology*, 323(3):573–584, 2002.

[**282**] A Keith Dunker and Vladimir N Uversky. Drugs for 'protein clouds': targeting intrinsically disordered transcription factors. *Current opinion in pharmacology*, 10(6):782–788, 2010.

[**283**] Fanchi Meng, Vladimir N Uversky, and Lukasz Kurgan. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cellular and Molecular Life Sciences*, 74:3069–3090, 2017.

[**284**] Lukasz Kurgan, Gang Hu, Kui Wang, Sina Ghadermarzi, Bi Zhao, Nawar Malhis, Gábor Erdős, Jörg Gsponer, Vladimir N Uversky, and Zsuzsanna Dosztányi. Tutorial: a guide for the selection of fast and accurate computational tools for the prediction of intrinsic disorder in proteins. *Nature Protocols*, 18(11):3157–3172, 2023.

[**285**] Marco Necci, Damiano Piovesan, and Silvio CE Tosatto. Critical assessment of protein intrinsic disorder prediction. *Nature methods*, 18(5):472–481, 2021.

[**286**] Alessio Del Conte, Mahta Mehdiabadi, Adel Bouhraoua, Alexander Miguel Monzon, Silvio CE Tosatto, and Damiano Piovesan. Critical assessment of protein intrinsic disorder prediction (CAID)-Results of round 2. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1925–1934, 2023.

[**287**] Kui Wang, Gang Hu, Sushmita Basu, and Lukasz Kurgan. flDPnn2: Accurate and Fast Predictor of Intrinsic Disorder in Proteins. *Journal of Molecular Biology*, page 168605, 2024.

[**288**] Md Wasi Ul Kabir and Md Tamjidul Hoque. DisPredict3.0: Prediction of intrinsically disordered regions/proteins using protein language model. *Applied Mathematics and Computation*, 472:128630, 2024.

[**289**] Jiangning Song and Lukasz Kurgan. Two decades of advances in sequence-based prediction of MoRFs, disorder-to-order transitioning binding regions. *Expert Review of Proteomics*, 22(1):1–9, 2025.

[**290**] Zhenling Peng, Zixia Li, Qiaozhen Meng, Bi Zhao, and Lukasz Kurgan. CLIP: accurate prediction of disordered linear interacting peptides from protein sequences using co-evolutionary information. *Briefings in Bioinformatics*, 24(1):bbac502, 2023.

[**291**] Kiersten M Ruff and Rohit V Pappu. AlphaFold and implications for intrinsically disordered proteins. *Journal of molecular biology*, 433(20):167208, 2021.

[**292**] András Hatos, Borbála Hajdu-Soltész, Alexander M Monzon, Nicolas Palopoli, Lucía Álvarez, Burcu Aykac-Fas, Claudio Bassot, Guillermo I Benítez, Martina Bevilacqua, Anastasia Chasapi, et al. DisProt: intrinsic protein disorder annotation in 2020. *Nucleic acids research*, 48(D1):D269–D276, 2020.

[**293**] Ying Huang, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682, 2010.

[**294**] Felix Kallenborn, Alejandro Chacon, Christian Hundt, Hassan Sirelkhatim, Kieran Didi, Christian Dallago, Milot Mirdita, Bertil Schmidt, and Martin Steinegger. GPU-accelerated homology search with MMseqs2. *bioRxiv*, pages 2024–11, 2024.

[**295**] Simon J Hubbard and Janet M Thornton. Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London*, 2(1), 1993.

[**296**] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[**297**] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[**298**] Zhenling Peng, Jing Yan, Xiao Fan, Marcin J Mizianty, Bin Xue, Kui Wang, Gang Hu, Vladimir N Uversky, and Lukasz Kurgan. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cellular and Molecular Life Sciences*, 72:137–151, 2015.

[**299**] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pages 8946–8970. PMLR, 2022.

[**300**] Dexiong Chen, Philip Hartout, Paolo Pellizzoni, Carlos Oliver, and Karsten Borgwardt. Endowing protein language models with structural knowledge. *arXiv preprint arXiv:2401.14819*, 2024.

[**301**] Matsvei Tsishyn, Pauline Hermans, Marianne Rooman, and Fabrizio Pucci. Residue conservation and solvent accessibility are (almost) all you need for predicting mutational effects in proteins. *Bioinformatics*, 41(6):btaf322, 2025.

[**302**] Pascal Notin, Lood Van Niekerk, Aaron W Kollasch, Daniel Ritter, Yarin Gal, and Debora S Marks. TranceptEVE: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. *bioRxiv*, pages 2022–12, 2022.

[**303**] Zuobai Zhang, Pascal Notin, Yining Huang, Aurelie C Lozano, Vijil Chenthamarakshan, Debora Marks, Payel Das, and Jian Tang. Multi-scale representation learning for protein fitness prediction. *Advances in Neural Information Processing Systems*, 37:101456–101473, 2024.

[**304**] Timothy Fei Truong Jr and Tristan Bepler. Understanding protein function with a multimodal retrieval-augmented foundation model. *arXiv preprint arXiv:2508.04724*, 2025.

[**305**] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. AutoGluon-Tabular: Robust and accurate AutoML for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

[**306**] Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic acids research*, 52(D1):D368–D375, 2024.

[**307**] Céline Marquet, Julius Schlensok, Marina Abakarova, Burkhard Rost, and Elodie Laine. Expert-guided protein language models enable accurate and blazingly fast fitness prediction. *Bioinformatics*, 40(11):btae621, 2024.

[**308**] Zi-Hang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. ConvBERT: Improving BERT with span-based dynamic convolution. *Advances in Neural Information Processing Systems*, 33:12837–12848, 2020.

[**309**] Timothy Truong Jr and Tristan Bepler. PoET: A generative model of protein families as sequences-of-sequences. *Advances in Neural Information Processing Systems*, 36:77379–77415, 2023.

[**310**] Damiano Sgarbossa, Cyril Malbranke, and Anne-Florence Bitbol. ProtMamba: a homology-aware but alignment-free protein state space model. *Bioinformatics*, 41(6), 2025.

[**311**] Luca Alessandro Silva, Barthelemy Meynard-Piganeau, Carlo Lucibello, and Christoph Feinauer. Fast uncovering of protein sequence diversity from structure. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8):084003, 2025.

[**312**] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[**313**] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

[**314**] Frances MG Pearl, CF Bennett, James E Bray, Andrew P Harrison, Nigel Martin, A Shepherd, Ian Sillitoe, J Thornton, and Christine A Orengo. The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic acids research*, 31(1):452–455, 2003.

[**315**] Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P Waman, Paul Ashford, Harry M Scholes, Camilla SM Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, et al. CATH: increased structural coverage of functional space. *Nucleic acids research*, 49(D1):D266–D273, 2021.

[**316**] Andy M Lau, Nicola Bordin, Shaun M Kandathil, Ian Sillitoe, Vaishali P Waman, Jude Wells, Christine A Orengo, and David T Jones. Exploring structural diversity across the protein universe with The Encyclopedia of Domains. *Science*, 386(6721):eadq4946, 2024.

[**317**] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pages 950–959. PMLR, 2020.

[**318**] Antoni Beltran, Xiang'er Jiang, Yue Shen, and Ben Lehner. Site-saturation mutagenesis of 500 human protein domains. *Nature*, pages 1–10, 2025.

[**319**] Alessandra Carbone, Elodie Laine, and Gianluca Lombardi. Unlocking protein evolution insights: Efficient and interpretable mutational effect predictions with gemme. In *Protein Evolution: Methods and Protocols*, pages 105–122. Springer, 2026.

[**320**] Neil Thomas, David Belanger, Chenling Xu, Hanson Lee, Kathleen Hirano, Kosuke Iwai, Vanja Polic, Kendra D Nyberg, Kevin Hoff, Lucas Frenz, et al. Engineering highly active and diverse nuclease enzymes by combining machine learning and ultra-high-throughput screening. *bioRxiv*, pages 2024–03, 2024.

[**321**] Robert D Finn, Jody Clements, and Sean R Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic acids research*, 39(suppl_2):W29–W37, 2011.

[**322**] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

Appendix A

gemme-utils package

ESCOTT model is a method that combines evolutionary information from sequence alignments with structural features to compute pathogenicity scores for single and multiple-point mutations. While being a fast and effective model, complementary to deep learning approaches, its execution might not be seamless due to its computational implementation, that requires many dependencies.

Additionally, data preprocessing is also a limitation, as the model requires as inputs a multiple sequence alignment in FASTA format and a structure file in PDB format, with the following characteristics:

1. query sequence must be the first in the alignment, and must not contain any gap;
2. all sequences in the alignment must have the same length;
3. importantly, there must be an exact correspondence between the query sequence and the PDB structure.

The first two conditions are required for the execution of JET2 algorithm, that is part of GEMME software. The last one, instead, is referred to ESCOTT implementation.

Indeed, the fastest method to obtain the required inputs for any query sequence is to run Colabfold, that returns an alignemnt in A3M format produced by MMseqs2, and a structure predicted with AlphaFold2. AlphaFold server is a more recent and precise alternative, as it runs AlphaFold3 pipeline with Jackhmmer alignment method. In both cases, the execution is time consuming and the number of requests is limited, making them unfeasible for large scale studies. A possible solution is a local installation of Alphafold, Colabfold or alternative methods, but they require hundreds of gigabytes of disk space for databases storage, and GPUs with enough RAM to run the prediction step.

While in the general case one of the mentioned approaches is required, we could instead have access to experimental or precomputed structures that include or differ very slightly from the query sequence. In that case, one might want to run ESCOTT with such approximated structure, when not many variations are expected with respect to the true one in terms of residues relative positions.

Therefore, to allow for this option, accelerate inputs preprocessing and program execution, and visualize output results, I created different Python scripts and merged them in a

small package, called `gemme-utils`, available at `https://github.com/GianLMB/gemme-utils`, with instructions for the installation. The package was developed for the methodological paper "Unlocking Protein Evolution Insights: Efficient and Interpretable Mutational Effect Predictions with GEMME" [**319**], where we also compared ESCOTT and GEMME results with PLM-based predictions in multiple settings, including predicitons of multiple-points mutations. In the following a list of the package's scripts is provided, as well as a brief description of each of them.

prepare_alignment.py prepares a FASTA-like alignment file for use with GEMME or ESCOTT by:

- Removing lowercase letters from sequences;
- Converting gap-like characters '.' into '-';
- Removing positions containing gaps in the query sequence;
- Shortening headers to a maximum length of 50 characters (an issue in the JET2 package);
- Removing gapped sequences with length different from query.

prepare_pdb.py reformats the PDB file such that it can be used as input for ESCOTT, by removing all residues that are not part of the query sequence.

Query sequence is extracted from the alignment file, in FASTA format. The common subsequence between the query and PDB sequence is extracted, and the PDB file is trimmed to only include these residues. If the query sequence (and thus the alignment) is longer than the PDB sequence, the script produces an additional alignment file whose query corresponds to the common subsequence.

Also, if an exact local match between the query and PDB sequence is not found, due to few amino acids substitutions, the script attempts to find the best local alignment between the two sequences and, if the alignment score is above an acceptance threshold, the corresponding structure is kept.

docker_run.py runs GEMME or ESCOTT in a Docker container. Corresponding images (`elodielaine/gemme:gemme` and `tekpinar/prescott-docker:v1.6.0`, respectively) are pulled from the Docker Hub if not available locally. It is possible to run multiple processes in parallel in different containers, by specifying the different input files.

plot_on_structure.py plots GEMME/ESCOTT average scores on the structure of the protein in an interactive PyMOL session. The script initializes a PyMOL session and colors the structure based on the average per-position scores from the prediction file. The adopted color scale assignes red to positions more sensible to mutations, and blue to less sensible positions.

All scripts are also available as Bash scripts and can be executed from the command line.

On the prediction of multiple mutations with ESCOTT and GEMME

`gemme-utils` package facilitates the usage of GEMME and ESCOTT softwares, complementing their local installation and webserver implementations with an almost automated pipeline. In [**319**], in addition to model's description and usage, we provided some interesting applications to known proteins, like human homeobox protein ARX, in comparsion to other apporaches.

Interestingly, we found that GEMME and ESCOTT performances remain competitive also when evaluated on multiple-point mutations, even if the model does not fully capture the interdependence of residue positions within a sequence. This aspect was demonstrated by analyzing one of the most extensive genotype-phenotype enzyme activity landscape to date, comprising 55 759 diverse variants of the nuclease enzyme NucB (`https://github.com/google-deepmind/nuclease_design`) [**320**]. The dataset was iteratively constructed and used in optimization rounds of NucB, to adapt it for functioning in novel chemical environments, a key objective in synthetic biology.

Specifically, the dataset comprises three rounds of mutational campaigns (G2, G3, G4), in which mutations where selected through a variety of techniques including direct evolution (DE), hit recombination (HR) and machine-learning models fit on evolutionary data and mutational data from previous generations (ML). G2, G3 and G4 campaigns were all built upon the same initial library (G1), obtained by applying error-prone PCR (epPCR) to a single-site saturation mutagenesis library of the wild-type sequence from B. licheniformis. Among more than 211K initial mutations generated by epPCR, employed for the DE campaign, only 9.4K where included in G1, corresponding to those for which catalytic activity was measured (Figure A.1A). Based on their activity, mutated sequences where divided in 4 classes, namely non-functional, activity > 0, activity $>$ wild type (WT) and activity $>$ A73R, the latter being the top-performing variant from the initial epPCR library, with 8-fold higher activity than the WT. GEMME prediction of 2088 single-point mutations in this dataset is reported in Figure A.1B.

Figure A.1C illustrates the superior performance of GEMME compared to a variational auto-encoder trained on homologous sequences, ESM-1b [**108**], pHMM [**321**], BLOSUM62 [**322**]. This comparison is based on two different evaluations: the percentage of functional variants recovered (Figure A.1C, top) and the percentage of sequences that after mutation improve their activity with respect to the WT (Figure A.1C, bottom). Notably, ESCOTT, which incorporates structural information into GEMME, further enhances performance. The analysis was conducted on a subset of the G1 variants: 874 have activity higher than 0, and 103 have an activity higher than WT (Figure A.1A).

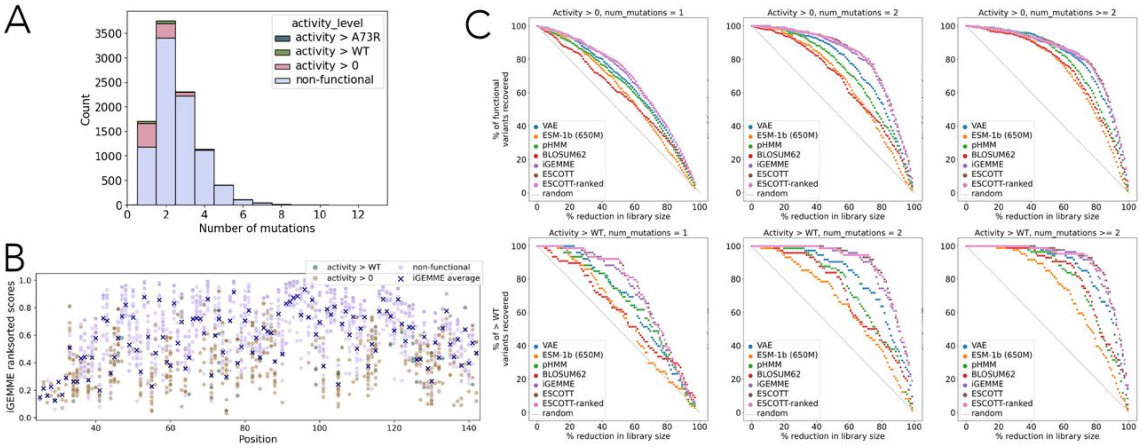

Figure A.1: **Comparative zero-shot analysis of the nuclease activity, NucB, with different tools predicting multiple mutations**. (A) Activity level distribution for the dataset of approximately 9 400 variants in G1 dataset, colored by estimated activity levels. G1 includes a single-site saturation mutagenesis library as well as variants generated by error prone PCR. On average, G1 variants have 2.5 mutations, while functional variants have, on average, 1.6 mutations (from Figure 5f or Supplementary Figure 28 in [**320**]). (B). iGEMME ranked sorted analysis of the 2088 single-point mutations present in the dataset. This analysis includes the approximately 1700 single-point mutations in G1 described in A. The prediction concerns three phenotypic classes of variants: with activity > WT (green), with activity > 0 (brown), non functional (light purple). For each position in the sequence, iGEMME averages are indicated by a cross. Low iGEMME scores indicate no mutational effect, hence functional activity, while high iGEMME scores indicate no functional activity. (C) VAE, ESM-1b, pHMM, BLOSUM62, iGEMME, ESCOTT (raw scores) and ESCOTT (ranked) are compared on two tasks, for sequences with 1, 2 or more mutations: the percentage of functional variants recovered (top); the percentage of sequences that after mutation improve their activity with respect to the WT (bottom). The dataset used for this analysis is a subset of G1, previously used and described in [**320**].

Appendix B

Additional information for MuLAN

Table B1: Benchmark datasets of changes of binding affinity upon mutation

Name	Number of complexes	Number of mutations	Type of point mutations	Source Database
S1102 [**151**]	57	1102	single	SKEMPI1
S4169 [**154**]	319	4169	single	SKEMPI2
S1131 [**156**]	114	1131	single	SKEMPI1
S2003 [**159**]	142	2003	single	SKEMPI2
S1400 [**157**]	113	1400	multiple	SKEMPI1

Table B2: **RMSE of MuLAN predictions for fine-tuned PLMs used as feature extractors**. Model performances are evaluated on the benchmark datasets S1102 and S1400. "Baseline" refers to the pre-trained PLMs.

Pre-trained Model	Baseline		Evo-tuning		Supervised fine-tuning	
	S1102	S1400	S1102	S1400	S1102	S1400
ESM2–35M	1.289	1.188	1.267	1.197	1.242	1.221
ESM2–650M	1.215	1.181	1.202	1.167	1.244	1.245
ProtBert	1.270	1.241	-	-	1.272	1.227

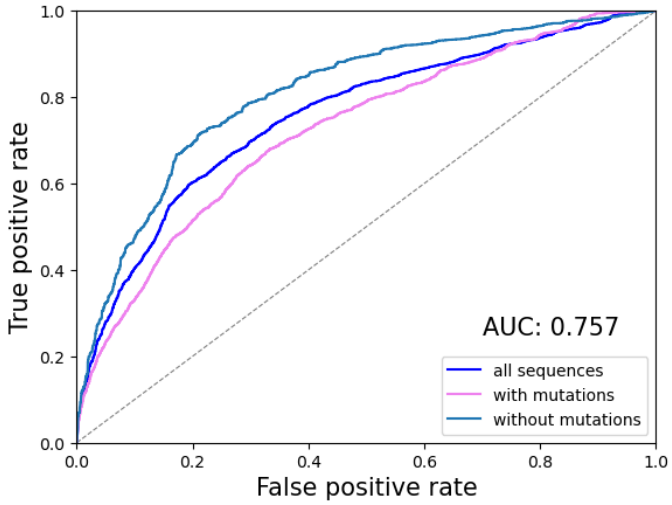

Figure B1: **ROC curve for MuLAN-Ankh-large attention weights over S1102 training set.** Attention weights achieve an average AUC of 0.757 for all sequences, that does not appear to be biased on regions in the dataset where mutations are found (AUC = 0.809 for sequences without mutations, AUC = 0.723 for sequences that present at least a mutation).

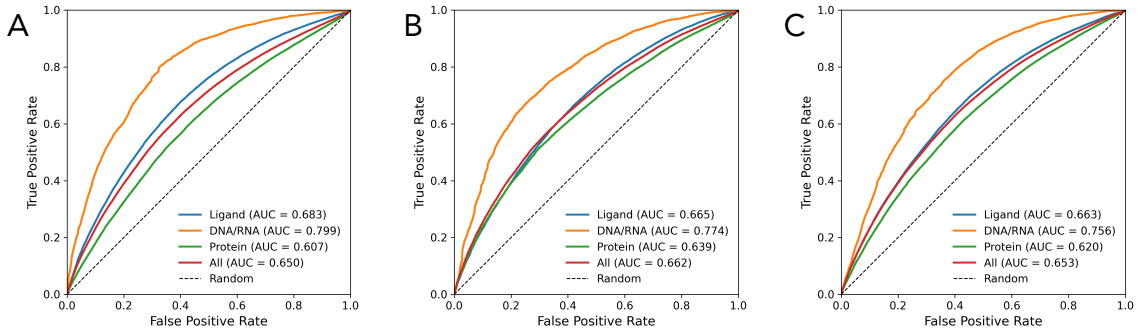

Figure B2: **ROC curves for MuLAN-Ankh-large attention weights over 177,545 interacting proteins from the PDB.** ROC curves are shown for interactions between protein chains and small molecules, DNA/RNA, other proteins, and all interaction types combined. Residue-residue contacts were labelled as positive if the distance between their central atoms was ≤ 8Å. Scores are stratified by DSSP 3-states secondary structure assignment: (A) alpha helices, (B) beta strands, (C) coiled regions.

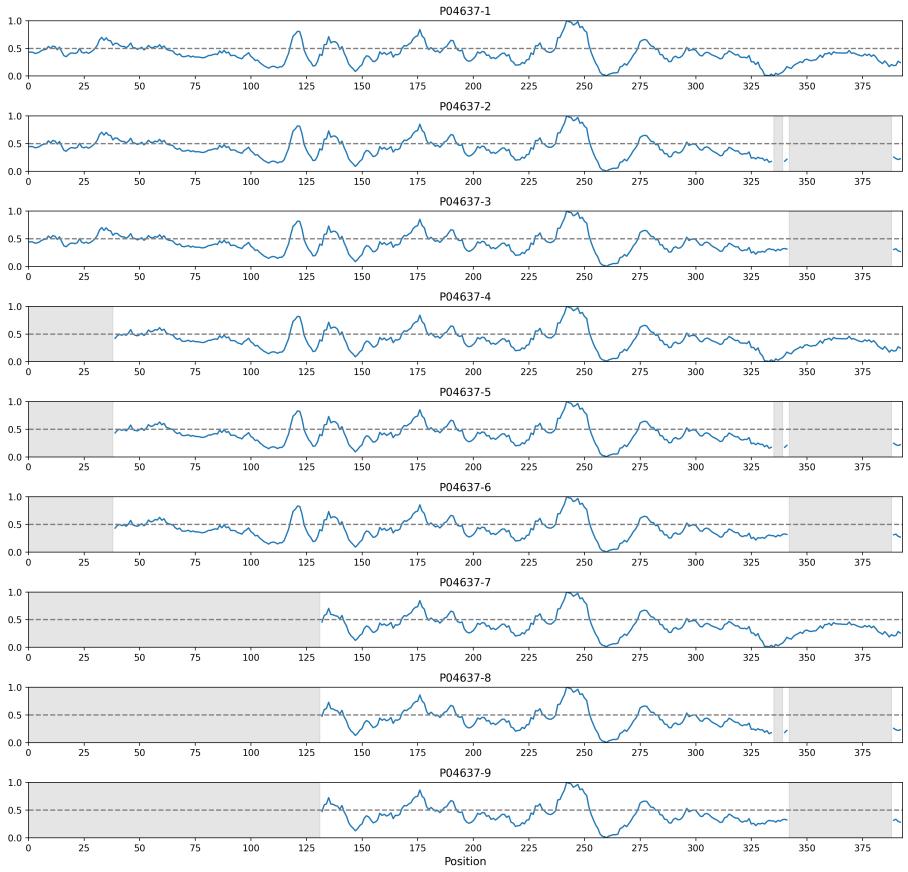

Figure B3: Attention profiles for aligned TP53 isoforms (UniProt ID P04637).

Figure B4: Attention profiles for isoforms (orange) aligned with attention profiles from native sequences (blue).

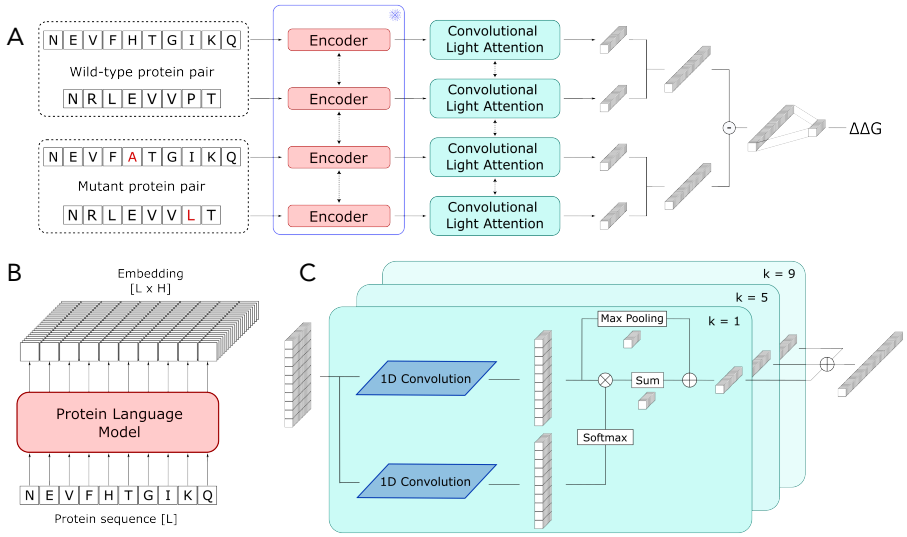

Figure B5: **Representation of the Light Attention model architecture.** (A) MuLAN full architecture: embeddings for both wild-type and mutated proteins are extracted from a pLM encoder, with frozen weights (B); each embedding undergoes processing within a Siamese Light Attention block, which normalizes weights across residues and learns a vector representation (C); these individual representations are subsequently combined pairwise to predict $\Delta\Delta G$.

Appendix C

Additional information for LoRA-DR-suite

Model	ROC AUC	F1	MCC	PR AUC	Max F1
esm2_650M-LoRA-DisProt7	0.847	0.587	0.433	0.675	0.610
esm2_35M-LoRA-DisProt7	0.856	0.629	0.497	0.658	0.647
Ankh-LoRA-DisProt7	0.827	0.590	0.438	0.634	0.590
ProtT5-LoRA-DisProt7	0.823	0.580	0.422	0.582	0.584

Table C1: Results of LoRA-DR-suite models trained on DisProt 7.0 dataset on CAID3_NOX test set.

DisProt ID	Length (in aa)	Fraction of positive labels
DP03745	145	0.241
DP03802	2905	0.149
DP03803	1043	0.015
DP03818	13	1.000
DP03819	13	1.000
DP03824	1136	0.242
DP03826	1138	0.182
DP03927	1084	0.056
DP03930	852	0.147
DP03931	854	0.063
DP04005	3088	0.007
DP04126	736	0.016
DP04128	296	0.145
DP04130	108	0.231
DP04132	230	0.078
DP04133	317	0.262
DP04134	140	0.807
DP04139	2285	0.658
DP04155	1068	0.071
DP04158	1292	0.013
DP04162	1103	0.198
DP04164	1247	0.136
DP04166	955	0.015
DP04178	1132	0.148
DP04187	94	0.426
DP04191	1162	0.057
DP04197	1447	0.345
DP04198	3010	0.013
DP04217	1053	0.153
DP04218	561	0.212
DP04225	1799	0.028
DP04232	1104	0.028
DP04246	218	0.335

Table C2: Proteins from CAID3_NOX dataset that were not included in the reduced test set, as they were not evaluated by at least one of the top 10 models.

Model	ROC AUC	F1	MCC	PR AUC	Max F1
CAID3_PDB (100% residue coverage)					
ESM2_650M-LoRA-ID	0.929	0.780	0.695	0.882	0.783
ESM2_35M-LoRA-ID	0.918	0.775	0.680	0.859	0.777
Ankh-LoRA-ID	0.898	0.741	0.640	0.834	0.752
ProtT5-LoRA-ID	0.911	0.753	0.656	0.851	0.755
AlphaFold3-rsa	**0.949**	**0.852**	**0.786**	**0.906**	**0.853**
LMDisorder	0.941	0.812	0.724	0.897	0.824
PredIDR2-Seq-Art	0.939	0.789	0.689	0.885	0.801
PredIDR2-Prof-Art	0.935	0.777	0.671	0.868	0.785
PredIDR2-Seq-Rnd	0.933	0.768	0.658	0.869	0.788
SETH-0	0.933	0.843	0.777	0.905	0.843
CAID3_PDB (82.7% residue coverage)					
esm2_650M-LoRA-ID	0.940	0.819	0.742	0.906	0.820
esm2_35M-LoRA-ID	0.921	0.798	0.718	0.872	0.799
Ankh-LoRA-ID	0.910	0.781	0.686	0.857	0.787
ProtT5-LoRA-ID	0.925	0.792	0.704	0.875	0.795
PUNCH2	**0.952**	**0.853**	**0.795**	**0.917**	**0.853**
PUNCH2-Light	0.946	0.850	0.791	0.913	0.851
SPOT-Disorder2	0.945	0.827	0.761	0.906	0.827
PredIDR2-Prof-Art	0.943	0.796	0.701	0.880	0.797
AlphaFold-rsa	0.942	0.835	0.762	0.896	0.839

Table C3: Results of LoRA-DR-suite models on the CAID3_PDB test set, on all dataset (considering only models with coverage higher than 95%) and on common proteins scored by top 10 models (206 out of 232 proteins, corresponding coverage 82.7%). See Table C6.

Model	ROC AUC	F1	MCC	PR AUC	Max F1
CAID3_Binding (100% residue coverage)					
esm2_650M-LORA-ID	0.891	0.503	0.424	0.706	**0.618**
esm2_35M-LORA-ID	0.885	0.511	0.441	0.622	0.617
Ankh-LORA-ID	0.875	0.525	0.442	0.614	0.577
ProtT5-LORA-ID	**0.894**	**0.520**	**0.443**	**0.720**	0.617
IPA-All-bind	0.856	0.500	0.424	0.552	0.589
IPA-Nucleotide-bind	0.850	0.522	0.429	0.536	0.530
IPA-Protein-bind	0.850	0.492	0.407	0.518	0.545
AlphaFold3-binding	0.828	0.464	0.384	0.399	0.506
MoRFchibi-light	0.802	0.342	0.289	0.422	0.477
CAID3_Binding (81.2% residue coverage)					
esm2_650M-LoRA-ID	0.885	0.498	0.399	0.729	**0.644**
esm2_35M-LoRA-ID	0.880	0.504	0.416	0.641	0.639
Ankh-LoRA-ID	0.864	0.517	0.415	0.632	0.584
ProtT5-LoRA-ID	**0.893**	0.518	0.424	**0.753**	**0.644**
DisoFLAG-LB	0.867	0.386	**0.450**	0.684	0.591
DisoFLAG-IDR	0.850	**0.529**	0.417	0.614	0.561
IPA-Nucleotide-bind	0.843	0.528	0.423	0.569	0.548
IPA-All-bind	0.841	0.491	0.395	0.562	0.598
DisoFLAG-IB	0.838	0.239	0.338	0.632	0.551

Table C4: Results of LoRA-DR-suite models on the CAID3_Binding test set. Top: on the all dataset (we considered models with coverage higher than 95%). Bottom: on the dataset comprising proteins scored by all the top 10 models (45 out of 51 proteins, corresponding coverage 81.2%). See Table C6 for the discarded proteins.

Model	ROC AUC	F1	MCC	PR AUC	Max F1
CAID3_Linker (100% residue coverage)					
esm2_650M-LORA-ID	0.852	0.370	0.366	0.258	0.390
esm2_35M-LORA-ID	0.838	0.378	**0.376**	0.233	0.430
Ankh-LORA-ID	0.849	0.398	0.387	0.240	0.408
ProtT5-LORA-ID	0.811	0.306	0.271	0.176	0.310
LINKER-Pred2	**0.854**	**0.392**	0.352	**0.359**	**0.485**
DisorderUnetLM	0.851	0.369	0.338	0.258	0.374
LINKER-Pred	0.845	0.405	0.365	0.361	0.478
flDPnn3b	0.840	0.299	0.300	0.211	0.362
flDPnn3a	0.836	0.302	0.305	0.217	0.373
CAID3_Linker (79.8% residue coverage)					
esm2_650M-LoRA-ID	0.853	0.392	0.379	0.279	0.411
esm2_35M-LoRA-ID	0.834	0.388	0.379	0.242	0.436
Ankh-LoRA-ID	0.847	0.407	0.388	0.248	0.416
ProtT5-LoRA-ID	0.810	0.324	0.276	0.195	0.330
IPA-AF2-Linker	**0.883**	**0.450**	0.411	0.389	0.467
LINKER-Pred2	0.882	0.426	0.378	0.413	**0.517**
LINKER-Pred	0.873	0.440	0.391	**0.414**	0.510
DisoFLAG-IDR	0.862	0.448	**0.416**	0.327	0.450
LINKER-Pred-Lite	0.853	0.433	0.383	0.421	0.459

Table C5: Results of LoRA-DR-suite models on the CAID3_Linker test set. Top: on the all dataset (we considered models with coverage higher than 95%). Bottom: on the dataset comprising proteins scored by all the top 10 models (19 out of 20 proteins, corresponding coverage 79.8%). See Table C6 for the discarded proteins.

DisProt ID	Length (in aa)	Fraction of positive labels
CAID3_PDB		
DP02732	1684	0.377
DP03745	145	0.289
DP03802	2905	0.371
DP03803	1043	0.061
DP03818	13	1.000
DP03819	13	1.000
DP03938	473	1.000
DP03972	1735	0.029
DP03980	912	1.000
DP04005	3088	1.000
DP04052	3056	0.010
DP04057	1093	0.140
DP04062	1507	0.012
DP04064	814	1.000
DP04096	1844	0.063
DP04134	140	1.000
DP04139	2285	0.726
DP04163	409	1.000
DP04184	146	0.275
DP04190	88	1.000
DP04197	1447	0.860
DP04198	3010	0.032
DP04208	274	0.184
DP04225	1799	1.000
DP04239	1430	0.093
DP04241	467	1.000
CAID3_Binding		
DP03818	13	1.000
DP03980	912	0.138
DP04134	140	0.271
DP04190	88	0.602
DP04198	3010	0.013
DP04241	467	0.385
CAID3_Linker		
DP04198	3010	0.013

Table C6: Proteins from CAID3_PDB, CAID3_Binding and CAID3_Linker datasets that were not included in the respective test sets, as they were not evaluated by at least one of the top 10 models.

Model	ROC AUC	F1	MCC	PR AUC
Intrinsic disorder				
ESM2_650M-LoRA-ID	**0.881**	**0.625**	**0.506**	**0.748**
ESM2_650M	0.832	0.565	0.416	0.632
Soft disorder				
ESM2_650M-LoRA-SD	**0.817**	**0.626**	**0.439**	**0.694**
ESM2_650M	0.784	0.593	0.390	0.650

Table C7: Performances of logistic regression classifiers trained on model embeddings (either ESM alone or with LoRA) on held-out test sets.

Model	LoRA r	Dropout p	LoRA layers	LR ($\times 10^{-4}$)	Warm up	Weight decay	ROC AUC
ESM2_650M-LoRA-DisProt7	16	0.1	Q, V	4.11	0.1	0.1	0.842
ESM2_35M-LoRA-DisProt7	32	0.3	Q, V	7.09	0.1	0.1	0.836
Ankh-LoRA-DisProt7	16	0.3	Q, V	5.64	0.1	0.001	0.831
ProtT5-LoRA-DisProt7	16	0.3	Q, V	9.74	0.1	0.01	0.853
ESM2_650M-LoRA-ID	32	0.1	Q, V	2.08	0.1	0.001	0.858
ESM2_35M-LoRA-ID	16	0.3	Q, K, V	9.37	0.2	0.001	0.848
Ankh-LoRA-ID	32	0.3	Q, V	2.41	0.2	0.01	0.836
ProtT5-LoRA-ID	8	0.2	Q, V	7.73	0.1	0.1	0.847
ESM2_650M-LoRA-SD	32	0.3	Q, V	5.49	0.2	0.01	0.844
ESM2_35M-LoRA-SD	16	0.1	Q, V	5.80	0.1	0.01	0.815
Ankh-LoRA-SD	8	0.1	Q, K, V	8.09	0.1	0.01	0.837
ProtT5-LoRA-SD	16	0.2	Q, V	9.93	0.1	0.001	0.826

Table C8: Hyperparameters selected for each model, and ROC AUC score obtained on the relative validation set. DisProt7 suffix refers to models evaluated on the first and second edition of CAID, thus trained only on the DisProt 7.0 subset, while models with ID (SD) suffix are trained on the full intrinsic disorder (soft disorder) datasets.

Model	CAID1_Disprot	CAID2_NOX	CAID3_NOX
ESM2_650M-LoRA-ID	0.161 (2.203)	0.248 (4.132)	0.154 (2.151)
ESM2_35M-LoRA-ID	0.025 (0.360)	0.041 (0.776)	0.038 (0.358)
Ankh-LoRA-ID	0.151 (3.436)	0.222 (5.988)	0.262 (3.367)
Ankh-ProtT5-ID	0.272 (4.032)	0.449 (7.143)	0.262 (3.922)
Average L	522	766	492

Table C9: Average execution time per protein sequence (in seconds) for LoRA-DR-suite models on the evaluation sets for intrinsic disorder. Test samples were processed in single batches. For each dataset, we report the average length L of the proteins. Times are reported for GPU (CPU) execution. See text for hardware specifics.

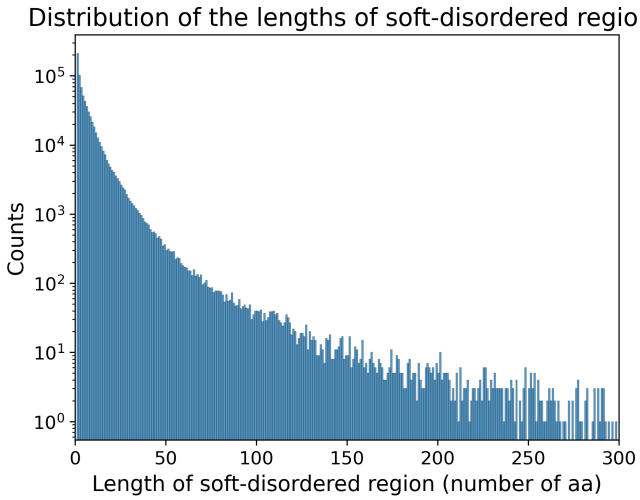

Figure C1: Distribution of the length of soft-disordered regions in SoftDis clustered dataset.

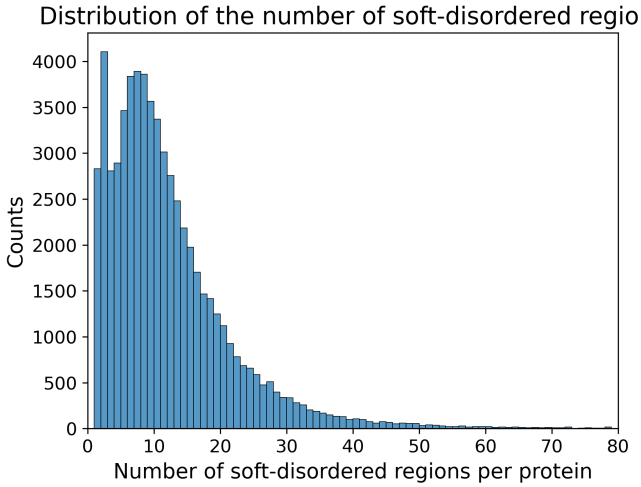

Figure C2: Distribution of the number of soft-disordered regions per protein in SoftDis clustered dataset.

Appendix D

List of articles

G. Lombardi, A. Carbone. "MuLAN: Mutation-driven Light Attention Networks for investigating protein-protein interactions from sequences". *bioRxiv 10.1101/2024.08.24.609515.*
doi: `https://doi.org/10.1101/2024.08.24.609515`

G. Lombardi, B. Seoane, A. Carbone, "LoRA-DR-suite: adapted embeddings predict intrinsic and soft disorder from protein sequences". *Bioinformatics*, Volume 41, Issue Supplement_1, July 2025, Pages i439-i448.
doi: `https://doi.org/10.1093/bioinformatics/btaf185`

S. D. Lemaire, G. Lombardi, A. Mancini, A. Carbone, J. Henri. "Functional recoding of *Chlamydomonas reinhardtii* thioredoxin type-h into photosynthetic type-f by switching selectivity determinants". *Frontiers in Plant Science*, Volume 16, March 2025.
doi: `https://doi.org/10.3389/fpls.2025.1554272`

A. Carbone, E. Laine, G. Lombardi, "Unlocking protein evolution insights: efficient and interpretable mutational effect predictions with GEMME", *Protein Evolution: Methods and Protocols*, in the book series *Methods in Molecular Biology* (MIMB), volume 2979, S.M. Khan, F. Pazos ed., Springer, 2025.
doi: `https://doi.org/10.1101/2024.08.15.608048`

S. G. Olvera-Vazquez, X. Chen, A. Mesnil, C. Meslin, F. Almeida-Silva, J. Confais, Y. Bourgeois, G. Lombardi, C. Lougmani, K. Alix, N. Francillonne, N. Choisne, S. Cauet, J.-C. Simon, C. Buchard, N. Rodde, D. Ogereau, C. Mottet, A. Degrave, E. Segura, A. Carbone, B. Benoît, E. Jacquin-Joly, W. Marande, D. Lavenier, F. Legeai, A. Cornille. "Comprehensive annotation of olfactory and gustatory receptor genes and transposable elements revealed their evolutionary dynamics in aphids". *Molecular Biology and Evolution*, September 2025.
doi: `https://doi.org/10.1093/molbev/msaf238`

A. Mesnil, V. Castric, G. Lombardi, M. Monniaux, S. Saidi, X. Chen, W. Marande, J. Confais, A.-L. Fuchs, E. Dapena, A. Venon, T. Tsuchimatsu, A. Carbone, X. Vekemans, A. Cornille. "Genomic architecture of the self-incompatibility locus in apple provides insights

into the evolution of collaborative non-self recognition". *bioRxiv 10.1101/2025.10.28.683340.*
doi: `https://doi.org/10.1101/2025.10.28.683340`

A. Mancini, V.-S. Pho, A. Bianchi, G. Lombardi, C. Lyu, A. Carbone. "Scaling the Profile of Life by Function with SPIN". *Bioinformatics Advances*, February 2026.
doi: `https://doi.org/10.1093/bioadv/vbag064`