Exploiting Machine Learning in Complex Biological Systems: Insights into Protein Stability Prediction and Mutational Signatures Analysis

UNIVERSITÀ
DI TORINO

University of Turin, Italy
Department of Medical Sciences
PhD Program in Complex Systems for Quantitative Biomedicine
XXXVI cycle
Academic years: 2020/2023

Author: Corrado Pancotti

Advisor: Prof. Piero Fariselli
Advisor: Dr. Tiziana Sanavia
PhD coordinator: Prof. Enzo Medico

Abstract

Since the advent of the next-generation sequencing technologies, it is possible to thoroughly analyze the entire genome and to explore its detailed components with unprecedented accuracy. This advanced exploration have allowed to highlight complex biological mechanisms and relationships that are crucial to understanding genetics at a deeper level. For instance, Whole Exome Sequencing (WES) has allowed to identify non-synonymous mutations, which are responsible for amino acid changes in protein sequences which can influence and modify his three dimensional structure, triggering cascade effects thus consenquently leading to diseases. A deep understanding of the complex mechanisms underlying these mutations and their effects is crucial for the development of new targeted therapies. In addition Whole Genome Sequencing (WGS), provides a complete view of individual's DNA. This comprehensive analysis helps in mapping the human genome of individual patients, identifying mutations associated with specific mutagenic processes, paving the way for innovative targeted treatments.

The present manuscript will present main results of my research project during the PhD period, which generally consists in implementing machine learning solutions in biomedical applications. The present work will be dived in two macro parts: the first one will focus on non-synonymous mutations and on their impact on protein structure and stability. Specifically we will introduce the computational tools developed in the recent year to study and to predict protein stability changes upon mutations, highlighting method evaluations, major caveats and open challenges. Specifically we will describe in details two deep learning predictors we developed in our laboratory to satisfy transitivity and anti-symmetry, two fundamental thermodynamic properties that majority of predictors does not consider. The first method ACDC-NN, uses 3D structural information, while the second, ACDC-NN-Seq, relies on sequence data alone.

The second part of the manuscript will focus on cancer genetics, specifically on mutational signatures, which are unique patterns of mutations associated with mutagenic processes (ie. Tobacco smoking, ultraviolet exposure or DNA repair mechanisms). We will review existing methods for extracting mutational signatures from cancer genomes, highlighting current challenges and limitations. Special attention will be given to two our papers: one which address the issue of similarity of mutational signatures profile present on Catalogue of Somatic Mutation in Cancer (COSMIC), using archetypal analysis. The paper suggests the idea that some profiles, with unknown etiology, could represent overfitted non-biological signals, and could be represented as a linear combination of archetypes profiles, thus claiming the necessity to introduce some constraints in the implementation of future methods. The second study introduces a novel mutational signatures extraction method based on explainable autoencoders (MUSE-XAE), which consist in a nonlinear encoder and a linear decoder with non-negative constraint and a minimum volume regularization, adept at capturing potential nonlinear dependencies while preserving signature interpretability. We evaluated and compared MUSE-XAE with other available tools both on synthetic and real cancer datasets and demonstrated that it achieves superior performance in terms of precision and sensitivity in recovering mutational signature profiles. In addition, MUSE-XAE extracts highly discriminative mutational signature profiles by enhancing the classification of primary tumor types and subtypes in real-world settings.

Preface

I, Corrado Pancotti, confirm that the work presented in this thesis is my own. Where external contributions have been provided, I confirm that this has been indicated in the thesis.

Publications present in this thesis:

- S. Benevenuta*, C. Pancotti*, P. Fariselli, G. Birolo, T. Sanavia: ***An anti-symmetric neural network to predict free energy changes in protein variants*** - Journal of Physics D: Applied Physics (2021)
- C. Pancotti*, S. Benevenuta*, V. Repetto*, G. Birolo, E. Capriotti, T. Sanavia, P. Fariselli: ***A Deep-Learning Sequence-Based Method to Predict Protein Stability Changes upon Genetic Variations*** - Genes (2021)
- C. Pancotti*, S. Benevenuta*, G. Birolo*, V. Alberini, V. Repetto, T. Sanavia, E. Capriotti, P. Fariselli: ***Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset*** - Briefings in bioinformatics (2022)
- L. Montanucci, E. Capriotti, G. Birolo, S. Benevenuta, C. Pancotti, D. Lal, P. Fariselli: ***DDGun: an untrained predictor of protein stability changes upon amino acid variants*** - Nucleic Acids Research (2022)
- C. Pancotti, C. Rollo, G. Birolo, S. Benevenuta, P. Fariselli, T. Sanavia: ***Unravelling the instability of mutational signatures extraction via archetypal analysis*** - Frontiers in Genetics (2023)

Manuscript under revision included in this thesis:

- C.Pancotti, C.Rollo, G. Birolo, P. Fariselli,T. Sanavia, ***MUSE-XAE: MUtational Signature Extraction with eXplainable AutoEncoder enhances tumour type classification*** - Bioinformatics (2023)

Publications not included in this thesis:

- O. De Filippo*, Victoria Camman*, C.Pancotti* et. al ***Machine-learning based prediction of in-hospital death for patients with takotsubo syndrome: the InterTAK-ML model*** - European Journal of Heart Failure (2023)
- C.Pancotti, G. Birolo, C.Rollo, T Sanavia, B. Di Camillo, U. Manera, A. Chiò, P.Fariselli ***Deep learning methods to predict amyotrophic lateral sclerosis disease progression*** - Scientific Reports (2022)

During my PhD, I have collaborated with the University of Copenaghen under the supervision of Professor Anders Krogh .

Acknowledgements

This adventure has also come to an end. First of all, I want to thank my team for having contributed to creating a healthy, serene but also very stimulating environment. A special thanks goes to Piero, my scientific mentor who has been able to guide me over these three years, giving me initiative and trust in various projects, making me appreciate scientific rigour but at the same time showing me how great humility accompanied by profound knowledge is the perfect combination to allow us students to emerge in the best possible way and to have fruitful scientific dialogues.

I would like to thank the members of the lab one by one: Tiziana, a hard worker, an expert in evening submissions, who helped to greatly improve the work I presented, but above all the author of perfect aperol spritzes that cheered up the time spent together with the group; Giovanni, our ”crazy” logician always ready to lend a hand to help and advise, we won together an historic european challenge that will remain in the annals, author of spiced teas at all hours of the day and great host of our lab dinners; Silvia, my programming partner and collaborator on numerous projects, without whose valuable help and natural kindness I would have been lost; Cesare my scientific and life brother who accompanied me from aspiring physics student to aspiring researcher, Francesco with the Bolognese mood, the only real computer scientist in the group, a sincere and smart guy; Flavio the man of the Valley, an expert climber with a contagious laugh; Isabella, a sunny girl defined by someone as the fastest programmer in the west; Valeria, who brought a refreshing taste of the islands to us urban dwellers. I also want to thank Prof. Anders Krogh, who welcomed me into his laboratory in Copenaghen with extreme kindness and availability, giving me the opportunity for some enriching scientific discussions.

Last but not least the most important people in my life: I start by thanking my friends from Senigallia, my life and adventure companions, brothers i can count on at all times. I also want to thank all the friends in Turin, there are too many of you to list one by one, but you have to know that you have taught me a lot.

I thank my life partner Francesca, for all that you are, for never having doubted me and for sharing my own happiness at all my achievements. I thank all the members of my family without whom all this would not have been possible. Your support and continued closeness throughout the years have allowed me not only to get here, but more importantly to be the person i am, all this I owe to you.

Being able to count on all these wonderful people is the constant beacon that gives me serenity.

Ad Maiora!

List of Tables

2.1 Datasets composition . . . 12

3.1 Results on the IvankovDDGun test set . . . 24
3.2 The optimal architectures of the Differential Net before and after transfer learning. . . . 26
3.3 Results on Ssym . . . 28
3.4 Results on myoglobin . . . 29
3.5 Results on p53 . . . 29

4.1 Assessment of the protein stability prediction tools on s669. . . . 34
4.2 Pearson's correlations and root mean square error (RMSE) between the experimental and estimated $\Delta\Delta G$s. . . . 38

6.1 Summary of the *de novo* extraction scenario . . . 49
6.2 *De novo* signature extraction performance . . . 53
6.3 Aetiological information related to each archetype . . . 58

7.1 AUC_{norm} metric for de novo extraction methods . . . 67
7.2 Classification performance for tumour primary types for PCAWG and WGS cohorts . . . 69
7.3 Classification performance for tumour subtypes for PCAWG and WGS cohorts . . . 69

B.1 SBS Signatures used in each extraction scenario . . . 81
B.2 Pairwise cosine similarity (> 0.8) between COSMIC SBS Mutational Signatures . . . 81
B.3 The ten most flat signatures of COSMIC . . . 82
B.4 Cosine similarity and mean absolute error (MAE) between the original catalogues and the reconstructed . . . 82

C.1 Pairwise cosine similarity between matched MUSE-XAE and COSMIC signatures and between MUSE-XAE and Signal ones for the PCAWG cohort . . . 92
C.2 Pairwise cosine similarity between matched MUSE-XAE and databases 93

List of Figures

2.1 Visual representation of the $\Delta\Delta G$ 9
2.2 Distribution of the experimental $\Delta\Delta G$ (kcal/mol) values in the Ssym, S669, S2648, Broom, VariBench, P53 and Myoglobin datasets. 13
2.3 Heatmap showing the percentage of shared variants among the presented datasets 14
2.4 15
2.5 distribution of the amino acids involved in a mutation in each dataset. 16

3.1 Constituent module of ACDC-NN 21
3.2 Constituent modules of ACDC-NN-Seq 22
3.3 Complete ACDC-NN architecture. 23
3.4 Complete ACDC-NN-Seq architecture. 24
3.5 Heatmap of the filter matrices K_T and K_S for ACDC-NN 25
3.6 Heatmap of the filter matrix K_S for ACDC-NN-Seq 25

4.1 Antisymmetry, Pearson correlation and Bias for all predictors 33
4.2 $\Delta\Delta G$ classification accuracy 36
4.3 $\Delta\Delta G$ prediction distribution by stability class 37

5.1 SBS Mutational Signatures example 41
5.2 A summary workflow of de novo extraction and refit methods 43
5.3 An Illustrative example of exposures of mutational signatures on PCAWG dataset 44

6.1 Cluster map of COSMIC SBS Mutational Signatures 51
6.2 Distribution of the COSMIC flatness 52
6.3 Summary of the 29 archetypal profiles 55
6.4 Heatmap highlighting the associations between archetypes and the reconstructed signatures 56

7.1 MUSE-XAE architecture 61
7.2 Sensitivity analysis of data augmentation for the 5 synthetic scenarios 66
7.3 Evaluation metrics for data augmentation 67
7.4 Cosine similarity heatmap between the MUSE-XAE and SigProfilerExtractor 69
7.5 t-SNE representation of PCAWG and WGS cohorts 70

A.1 Comparison of method performance on real and modelled structure on the Ssym+ dataset 76

A.2 Effect on the method performances of the experimental technique: NMR versus X-ray. 77
A.3 Effect on the method performances of the different X-ray resolution . 77
A.4 Assessment of the effects of the relative accessibility of an amino-acid on the prediction of the protein stability 79
A.5 Assessment of the protein stability predictions tools on S669 at different temperature and pH conditions 80

B.1 Cosine similarity distribution for each simulated scenario 83
B.2 Flatness distribution for each simulated scenario 84
B.3 Number of detected signatures for each simulated scenario 85
B.4 Explained variance with respect to the number of archetypes 86
B.5 Cosine similarity between the original COSMIC signatures and the reconstructed ones through the 29 archetypes 87
B.6 Relationship between archetypes and signatures at different cosine similarity thresholds 88
B.7 pairwise cosine similarity distribution of the alpha coefficients profiles for different levels of pairwise cosine similarity between the original mutational signatures 89

C.1 Performance comparison between the top 10 performing methods . . 92
C.2 Performance comparison between the top 10 performing methods for each scenario 94
C.3 22 mutational signatures extracted from PCAWG dataset 95
C.4 23 mutational signatures extracted from the extended cohort dataset 96
C.5 SBS mutation counts for MUSE-SBSP, MUSE-SBSS and MUSE-SBSW for each tumour of the WGS extended cohort 97
C.6 MUSE-XAE confusion matrix for 18 PCAWG tumour primary sites . 98
C.7 MUSE-XAE confusion matrix for 34 PCAWG tumour subtypes . . . 98
C.8 MUSE-XAE confusion matrix for 15 WGS extended tumour primary sites 99
C.9 MUSE-XAE confusion matrix for 23 WGS extended tumour subtypes 99

Contents

1 Introduction 2

1.1 Thesis Outline 4

2 Protein stability 8

2.1 Gibbs free energy and $\Delta\Delta G$ 8

2.1.1 $\Delta\Delta G$'s fundamental properties 9

2.2 Stability prediction: computational methods, datasets and evalutation metrics 10

2.2.1 General overview of prediction algorithms 10

2.2.2 Available Datasets 11

2.2.3 Prediction assessment: performance metrics and open problems 12

2.3 Discussion 17

3 ACDC-NN and ACDC-NN-Seq: antisymmetric neural networks 18

3.1 Introduction 18

3.2 Materials and methods 19

3.2.1 Building ACDC-NN 19

3.3 Results 23

3.3.1 Learning anti-symmetry 23

3.3.2 Fine tuning and prediction of experimental $\Delta\Delta G$ values . . . 26

3.4 Discussion 27

4 Thorough Comparison of the available tools on the 669 dataset: pitfalls and suggestions 30

4.1 Introduction 30

4.2 Materials and methods 31

4.2.1 Evaluated methods 31

4.3 Results 33

4.3.1 Method performance on the new S669 dataset 33

4.3.2 Classification performance 35

4.3.3 The unbalanced predictions among stabilizing, neutral and destabilizing variants 35

4.4 Discussion and Conclusion 39

5 Mutational Signatures: fingerprints of mutational processes in cancer 40

5.1 Introduction 40

5.2 Mathematical formulation 41

5.3 De Novo extraction and refitting 42
5.4 Caveats and open problems 44
5.5 Discussion 45

6 Archetypal analysis reveals the inherent instability of mutational signatures 47
6.1 Introduction 47
6.2 Materials and methods 48
6.2.1 Similarity of COSMIC signatures 48
6.2.2 Flatness of COSMIC signatures 48
6.2.3 *De novo* extraction scenarios 48
6.2.4 Evaluation metrics 49
6.2.5 Archetypal Analysis 50
6.3 Results 50
6.3.1 COSMIC cluster map 50
6.3.2 Flatness analysis 52
6.3.3 *De novo* signature extraction 52
6.3.4 Archetypal Analysis 54
6.4 Discussion 57

7 MUSE-XAE: MUtational Signatures Extraction with an Explainable AutoEncoder enhances tumour types classification 60
7.1 Introduction 60
7.2 Materials and methods 61
7.2.1 Signatures Extraction 62
7.2.2 Signatures Assignment 63
7.2.3 Real World datasets 65
7.3 Results 65
7.3.1 Data augmentation improves robustness and accuracy 65
7.3.2 *De Novo* extraction comparison in synthetic scenarios 66
7.3.3 *De Novo* Extraction in real world datasets 67
7.3.4 MUSE-XAE enhances tumor classification 68
7.4 Discussion 70

8 Conclusion 72

A Protein Stability 75
A.0.1 Method performance on model vs experimental structures . . 75
A.0.2 Effect of the experimental technique on the method performance 75
A.0.3 Surface accessibility, pH and Temperature 78

B Mutational signatures 81
B.1 Archetypal analysis 81

C MUSE-XAE 90
C.1 Evaluation metrics 90

Chapter 1

Introduction

Next-Generation Sequencing (NGS) technologies have revolutionized the landscape of human genomics research. NGS provides the resolution to identify not only large-scale genomic changes but also single nucleotide variants (SNVs), small insertions and deletions (indels), and even more complex mutational events. Such granularity is vital for understanding the genetic basis of diseases and to offer a window into individual variability. Instead of a 'one-size-fits-all' approach, therapies can now be tailored based on individual genomic profiles. By understanding the specific mutations or variants that drive a disease, targeted therapies can be developed to intervene at the molecular level. The objective of this thesis is to highlight the role of genetic variation through a dual lens. On one hand, this work focuses on protein stability prediction due to an aminoacid changes in its sequence. On the other hand, it delves into the genomic dimension by examining mutational signatures and how their etiologies can be used to better understand the mutational landascape of tumour development. Specifically, for both topics, we will begin with a general introduction. Then we will discuss commonly used methodologies and practices, followed by an exploration of open challenges in the field. Finally, we will detail the contributions we have made to advance both protein stability prediction and mutational signature extraction.

Types and Implications of Exomic Variants

Focused on the coding regions of the human genome, Whole Exome Sequencing (WES) is a pivotal method in next-generation sequencing. Although these regions constitute less than 2% of the entire genome, they contain roughly 85% of all disease-related variants [1]. WES is highly adaptable, finding utility in various scientific realms like population genetics and genetic disease research [2]. Exomic variants can be dichotomized into synonymous and non-synonymous types. While synonymous variants do not change the amino acid sequence, they can still be pathogenic by affecting mRNA splicing and thereby altering protein functionality and drug responses [3].

Non-synonymous variants are more direct in their impact, causing amino acid substitutions (missense), introducing premature stop codons (nonsense), or eliminating stop codons altogether (nonstop). Such alterations often result in dysfunctional or non-functional proteins. Repositories such as Human Gene Mutation Database

(HGMD) [4], the Catalogue of Somatic Mutations in Cancer (COSMIC) [5] and others aggregate information of single amino acid variants that cause or are associated with disease and other sequence variations, providing valuable resources for researchers [6]. However, it is crucial to emphasize that not all non-synonymous variations necessarily lead to impactful changes in protein function or structure. Several factors can mitigate the consequences of such variations. For instance, if the replaced amino acid has chemical properties such as charge, size, or hydrophobicity that are highly similar to those of the original amino acid, the protein's overall structure and function may not be significantly affected. Additionally, the location of the amino acid substitution within the protein matters. If the change occurs in a region that is neither involved in the protein's enzymatic active site nor critical for its secondary or tertiary structure, or if it doesn't influence the protein's ability to interact with ligands or other proteins, then the overall functionality of the protein is likely to remain intact.

Disease-associated variants have been cataloged in over 1,000 human genes [4]. These sequence alterations can impact multiple dimensions of protein functionality, including transcription, RNA processing, folding, and stability [6]. In this context, this thesis emphasizes the role of exomic variants in affecting protein stability, a critical aspect often linked to disease development [7, 8, 9]. For example, a majority of monogenic disease-causing variants are known to destabilize the native structure of proteins [10]. Particularly in genes with haploinsufficiency, highly destabilizing variants usually yield non-functional proteins [11] .

Elucidating the mechanisms by which non-synonymous variants affect human diseases is of vital importance [12, 13, 14]. Yet, it is crucial to also consider the complex interplay between synonymous and non-synonymous variants, particularly in the case of complex diseases where multiple factors contribute to pathogenesis.

Challenges and pitfalls in protein stability prediction To accurately predict the impact of non-synonymous variants on protein stability in terms of change in the Gibbs free energy ($\Delta\Delta G$), multiple computational methods have been developed. While these tools are invaluable for understanding disease pathology, they often inadequately capture intrinsic $\Delta\Delta G$ properties like antisymmetry and transitivity. This shortfall is likely a consequence of training on datasets biased towards destabilizing mutations.

Addressing this issue could involve incorporating more stabilizing variants into the training data or leveraging machine learning techniques designed to handle imbalanced datasets. Another limitation lies in the underrepresentation of certain amino acid mutations, such as those involving proline or cysteine residues. This bias restricts the utility of existing tools for a broader range of mutation types.

Additionally, current evaluation protocols for these predictors often overlook testing bias. For a robust assessment, the partition between training and test datasets should account for protein homology, excluding variants from proteins sharing more than 25% sequence identity across the two sets. Unfortunately, this best practice is not universally implemented, resulting in possibly inflated performance metrics in the literature.

Mutational Signatures

The advent of NGS technologies has also enabled the in-depth investigation of mutational landscapes across the entire genome. These mutational patterns, also known as "mutational signatures", can be described as fingerprints of underlying exogenous and endogenous mutagenic factors (i.e tobacco smoking, ultraviolet exposure, DNA mistmatch-repair) or aberrant cellular processes that contribute to cancer development and proliferation. [15, 16]. Understanding these signatures provides valuable insights into disease etiology and can potentially guide therapeutic interventions [17, 18].

There are various types of mutational signatures that capture different kinds of mutational events, such as Single Base Substitution signatures (SBS), Insertions and Deletions (Indels, ID), Double Base Substitutions (DBS), and Copy Number Variations (CNV). Over the years, several methodologies have been developed, leading to the discovery of around 80 distinct mutational signatures, cataloged within the COSMIC database.

Open challenges in Mutational Signatures analysis

However, several open challenges remained to be addressed. Among these, of our interest we can include the COSMIC signatures redundancy, i.e the fact that a high cosine similarity between some mutational signatures can be found in the COSMIC catalogue and this could suggests the presence of mathematical artifacts due to overfitting [19]. Mathematical artifacts become a concern, particularly when signature extraction is highly dependent on the number of samples available. Moreover, several signatures in COSMIC databases still lack experimentally-validated etiology, supporting the hypothesis of possible overlapping signatures that describe the same mutagenic process. These challenges underscore the need for enhanced methods and frameworks for accurately interpreting mutational signatures and to ensure their correct application in both clinical and research settings.

In addition, current methods for extracting mutational signatures from tumor catalogs predominantly rely on Non-Negative Matrix Factorization (NMF) techniques. While NMF's linear approach offers the advantage of explainability in mutagenic processes, it may oversimplify the complex mechanisms underlying cancer development. This limitation could potentially result in overlooking non-linear relationships that better capture the intricacies of tumorigenesis.

1.1 Thesis Outline

In the first part of the present manuscript (Chapters 2-4) we introduce the concept of protein stability and present two new computational methods we have developed in our laboratory to predict the effects on protein stability due to a point variants on the amino acid sequence. In additionwe provide a comparison of available tools on a new dataset we have manually curated to assess methods performance in an unbiased manner. In the second part, after an introduction of mutational signatures, we present an archetypal analysis we performed on COSMIC signatures to prove that redundancy in the catalogues could be eliminated while mantaining the proper amount of information. Finally we describe a Mutational Signature extraction method we have developed in our laboratory.

Chapter 2: Protein stability, general concepts and properties In this chapter, we define the concept of protein stability and how its changing can be described with the difference in the Gibbs free energy, $\Delta\Delta G$.

we explore the impact of protein stability on various disorders, focusing particularly on monogenic conditions and diseases associated with haploinsufficient genes. we outline the $\Delta\Delta G$ characteristics and review the evolution of $\Delta\Delta G$ prediction computational methods. we detail the principal datasets employed for training and validation in machine learning-based prediction methods, the metrics used for performance assessment, and key challenges in the current landscape of stability prediction tools and their evaluations.

The content of this chapter can be found in the following publications belong to our laboratory:

- S. Benevenuta, P. Fariselli: ***On the Upper Bounds of the Real-Valued Predictions*** - Bioinformatics and Biology Insights (2019)
- G. Birolo, S. Benevenuta, P. Fariselli, E. Capriotti, E. Giorgio, T. Sanavia: ***Protein stability perturbation contributes to the loss of function in haploinsufficient genes*** - Frontiers in Molecular Biosciences (2021)
- S. Benevenuta, G. Birolo, T. Sanavia, E. Capriotti, P. Fariselli: ***Challenges in predicting stabilizing variations: an exploration*** - Frontiers in Molecular Biosciences

Chapter 3: ACDC-NN: A Convolutional Antisymmetric neural network In this Chapter we present ACDC-NN, a Convolutional Antisymmetric Differential Concatenated Neural Network, designed for respect $\Delta\Delta G$ thermodynamic properties. In particolar we describe both ACDC-NN 3D, which uses both proteins structure and sequence information, and its sequence based version ACDC-NN-Seq which does not require tertiary protein structure.

The content of this chapter originally appears in :

- S. Benevenuta*, C. Pancotti*, P. Fariselli, G. Birolo, T. Sanavia: ***An antisymmetric neural network to predict free energy changes in protein variants*** - Journal of Physics D: Applied Physics (2021)
- C. Pancotti*, S. Benevenuta*, V. Repetto*, G. Birolo, E. Capriotti, T. Sanavia, P. Fariselli: ***A Deep-Learning Sequence-Based Method to Predict Protein Stability Changes upon Genetic Variations*** - Genes (2021)

Chapter 4: Thorough comparison of available $\Delta\Delta G$ prediction tools on s669 dataset From the comprehensive database ThermoMutDB (v1.3) we manually collected and cleaned a new dataset, named s669, of point variants. The variants in s669 belong to proteins with a sequence identity lower than 25% with proteins found in commonly used dataset for $\Delta\Delta G$ prediction. Using s669 dataset, we conducted the most extensive and comprehensive comparison by including 20 tools, highlighting the weaknesses and strengths of these methods in $\Delta\Delta G$ prediction. Specifically, we compared their performance in predicting direct and reverse variants, and testing their antisymmetry. Additionally, we sought to identify common issues affecting all methods and potential improvements, pinpointing which variant classes (destabilizing, neutral, and stabilizing) are the most challenging to predict and trying to understand the reason behind that.

The content of this chapter originally appears in:

- C. Pancotti*, S. Benevenuta*, G. Birolo*, V. Alberini, V. Repetto, T. Sanavia, E. Capriotti, P. Fariselli: ***Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset*** - Briefings in bioinformatics (2022)
- S. Benevenuta, G. Birolo, T. Sanavia, E. Capriotti, P. Fariselli: ***Challenges in predicting stabilizing variations: an exploration*** - Frontiers in Molecular Biosciences

Chapter 5: Mutational Signatures: fingerprints of mutational processes in cancer In the present Chapter we mathematically introduce the concept of mutational signature and its importance to understand cancer heterogeneity and etiology, thus providing an invaluable resource to better implement target therapies. We discuss computational methods developed so far to extract mutational signatures from cancer genomes, highlighting their weaknesses and strengths, underlying their biological assumptions and stressing common pitfalls and caveats in the field.

Chapter 6: Archetypal Analysis reveals the inherent instability of mutational signatures In the present Chapter we present in detail an archetypal analysis we performed on COSMIC catalogue, suggesting the idea that numerous mutational signatures with unknown etiology could be the result of overfitting due to lack of statistical power. In fact, our study show that the amount of information in the COSMIC catalogue can be totally preerved with the half of the component. Interestingly the 'archetypal' signatures we found can be well explained and seems to have a biological meaning. This analysis highlight the need to develop computational methods without redundancy, by incorporating some mathematical constraints in order to extract realistic and biologically plausible signals.

The content of this chapter originally appeared in the following paper:

- C. Pancotti, C. Rollo, G. Birolo, S. Benevenuta, P. Fariselli, T. Sanavia: ***Unravelling the instability of mutational signatures extraction via archetypal analysis*** - Frontiers in Genetics (2023)

Chapter 7: MUSE-XAE: MUtational Signatures Extraction with an Explainable Auto Encoder In the present chaprter we describe a novel method we implemented in our lab, called MUSE-XAE, MUtational Signatures Extraction with an Explainable Auto Encoder, that consists in a non linear encoder and linear decoder with minimum volume regularization and non negativity constraint. Similar architectures has already been successfully applied in the context of single cell RNA seq and transcriptomic data. However this is the first application in mutational signature analysis. Our method show a very accurate extraction capabilities and enhances tumour types classification.

The content of this chapter originally appeared in the following paper:

- C.Pancotti, C.Rollo, G. Birolo, P. Fariselli,T. Sanavia, ***MUSE-XAE: MUtational Signature Extraction with eXplainable AutoEncoder enhances tumour type classification*** - Bioinformatics (2023)

Chapter 2

Protein stability

2.1 Gibbs free energy and $\Delta\Delta G$

The stability of a protein is quantified by the change in Gibbs free energy ΔG, which represents the difference in free energy between the native (folded, F) and denatured (unfolded, U) states of the protein. A negative ΔG indicates that the folded state is energetically more favorable, while a positive ΔG suggests that the unfolded state is favored. Thermodynamically, ΔG is directly related to the equilibrium constant (K_{eq}) for the folding-unfolding transition, according to the equation:

$$\Delta G = -RTlog(K_{eq}) \tag{2.1}$$

where R is the universal gas constant and T is the temperature in Kelvin. The equilibrium constant K_{eq} is defined as the ratio of the concentrations of the unfolded to folded states at equilibrium, $K_{eq} = \frac{[F]}{[U]}$.

A stable configuration of a protein corresponds to a minimum in the ΔG; in particular during its functional cycle a protein experiences different free energy local minima which correspond to stable three dimensional structures that are needed to correctly perform biological activities. The Gibbs free energy includes different contribution term such as hydrophobic effects, conformational configurations (entropic contributions) and interaction terms such as electrostatic and hydrogen bonds and Van der Waals forces. An aminoacid change in the protein sequence (which represents non-synonimous DNA variation) can alter the Gibbs free energy and consequently the protein stability, thus changing the protein structure and its ability to perform functions, leading to cascade effects that consequently can cause diseases. Protein stability changes can be described as difference of the free energy of unfolding between the wild type (W) and mutated state (M) of a protein:

$$\Delta\Delta G_{WM} = \Delta G_W - \Delta G_M. \tag{2.2}$$

A graphical representation of the $\Delta\Delta G$ is represented in Fig. 2.1.

Depending on the sign of $\Delta\Delta G$, variants can be classified in destabilizing (negative sign) and stabilizing (positive sign). Mutations that leads to a $\Delta\Delta G$ values close to zero, given the experimental uncertainties ([20, 21]) make their $\Delta\Delta G$ signs less

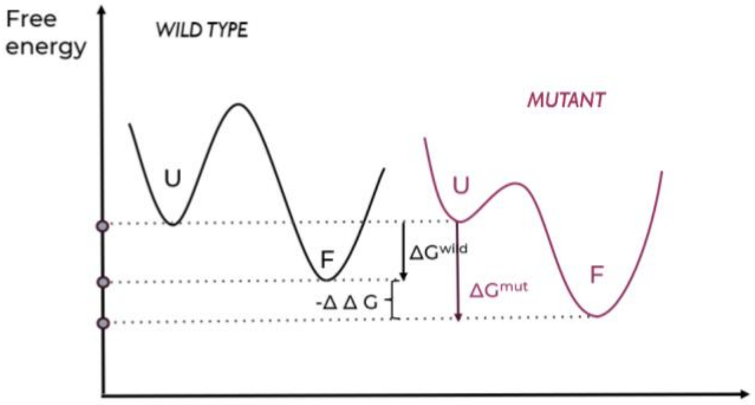

Figure 2.1: Schematic representation of the $\Delta\Delta G$. The black curve represent the energy profile of the wild type protein while the red one the mutant profile after an amminoacid change in the sequence.

reliable. To account for this issue, it is possible to consider $\Delta\Delta G$ values in the range between -0.5 and 0.5 kcal/mol as 'neutral'. The choice of 0.5 kcal/mol is based on the average experimental error, as reported in ([22]).

According to the classical view, if a mutation destabilizes, then this can lead to pathological conditions. On the other hand, if it tends to stabilize, there is no substantial change. However, recent studies are highlighting the possibility that it is not the sign of $\Delta\Delta G$ that is correlated with pathological conditions, but rather its absolute value $|\Delta\Delta G|$. Therefore, it is not the direction (destabilizing-stabilizing) of the mutation that is important, but rather its magnitude. From what has been said so far, it become clear the importance of being able to accurate predict the change in stability of a protein following one or more amminoacid substitutions.

2.1.1 $\Delta\Delta G$'s fundamental properties

Antisymmetry Given a wild-type protein (W) and its mutated version (M) that differ from each other by a single amino acid at position X, we can calculate the change in stability due to the substitution $X_W \rightarrow X_M$ as $\Delta\Delta G_{WM} = \Delta G_W - \Delta G_M$. Similarly, for the reverse variation $X_M \rightarrow X_W$, the magnitude of the change in Gibbs free energy is the same, with the opposite sign. Thus, the $\Delta\Delta G$ antisymmetry property can be summarized as:

$$\Delta\Delta G_{WM} = -\Delta\Delta G_{MW}. \tag{2.3}$$

This property came from the thermodynamics and can be derived by considering unfolding (U) and folding states (F) of a protein at the equilibrium and by calculating the Gibbs free energy through the equation 2.1 :

Hence, by writing the formula both for the direct for the reverse variation and exploiting the log properties we obtain:

$$\Delta G_{FU} = -RTlog\left(\frac{[F]}{[U]}\right) = RTlog\left(\frac{[U]}{[F]}\right) = -\Delta G_{UF}. \tag{2.4}$$

Finally , given the wild type protein (W) and the mutated one (M), we can obtain the antisymmetry property:

$$\Delta\Delta G_{WM} = RTlog\left(\frac{[U]}{[F]}\right)_W - RTlog\left(\frac{[F]}{[U]}\right)_M = -\Delta\Delta G_{MW}. \tag{2.5}$$

Transitivity Another fundamental $\Delta\Delta G$ property is the transitivity [23]. Let A, B, and C be three protein structures that differ by a single amino acid. $\Delta\Delta G_{A\to C}$ represents the change in stability for the mutation A → C . This free energy change can be expressed as the sum of the $\Delta\Delta G_{A\to B}$ due to the A → B mutation and the $\Delta\Delta G_{B\to C}$ due to the B → C mutation. Mathematically, this can be written as

$$\begin{aligned}\Delta\Delta G_{A\to C} &= \Delta G_C - \Delta G_A = (\Delta G_C - \Delta G_B) + (\Delta G_B - \Delta G_A) = \\ &= \Delta\Delta G_{B\to C} + \Delta\Delta G_{A\to B}.\end{aligned} \tag{2.6}$$

Naturally this property can be extended to N different proteins, forming chains of different lengths. The only requirement is that all the quantity are measured under the same experimental conditions.

2.2 Stability prediction: computational methods, datasets and evalutation metrics

2.2.1 General overview of prediction algorithms

Various computational methods have been developed so far to predict protein stability changes, spanning a range of approaches from force field-based methods to machine learning models. These methods employ features that can be broadly categorized into four types:

- **Structural**: Includes contacts between residues and distances between primary atoms.
- **Sequential**: Incorporates evolutionary information such as sequence similarity and homologies, along with the position of amino acids in the sequence.
- **Energetic**: Considers electrostatic interactions, Van der Waals forces, hydrogen bonds, and solvation energy.
- **Molecular**: Includes solvent accessibility and identifies hydrophobic and hydrophilic regions.

Initially, force field-based methods like FoldX and Monte Carlo based approach such as Rosetta were predominant, relying on physical free energy functions, by modelling the three dimensional structure of the protein. However those kind of methods suffer for intensive computational costs which represent a limit when analysing a high number of mutations. Consequently, less time-consuming methods such as PoPMuSiC and CUPSAT were designed. Those kind of methods apply Maxwell-Boltzmann statistics to estimate atomic interaction propensities from known protein structures. In particular they include linear combination of functions, often referred to as statistical potential, in order to model different kind of interactions based on the aminoacid type, torsion angles and solvent accessibility.

In contrast to knowledge-based predictors, thanks to the increasing amount of available data, machine learning-based methods have gained prominence. Thus, many machine learning based methods were implemented, such as MAESTRO, DeepDDG, PremPS, our method ACDC-NN and many others. These predictors exploit different kind of techniques ranging from Random Forest to Gradient Boosting and Neural Networks. The main peculiarity of ML based approaches consists in modeling the biophysical principles starting from structures and sequence based features, during the learning process with less a priori assumptions, revealing unrecognized patterns and dependencies in the data. However, the accuracy of these tools is highly dependent on the availability of extensive and diverse experimental training data, which can also make them prone to overfitting and sometimes not easy to interpret in physical therms.

Despite the progress made in the field of stability prediction as we will see later in more detail, most predictors suffer from lack of antisimmetry and transitivity. In addition all predictors have more difficulties to predict stabilizing variants than destabilizing. In the next section we will introduce the commonly used datasets to better explain the open challenges and discuss possible solutions.

2.2.2 Available Datasets

In recent years, the number of datasets available with experimental measurements of protein stability changes has increased considerably. Especially starting from resources such as ProthermDB and ThermoMUTDB, several datasets of variants with associated DDG have been cleaned and collected. Especially the most used to train or test machine learning methods are as follows.

- **S2648**, which contains 2648 manually curated variants with experimentally measured $\Delta\Delta G$ values [24];
- **VariBench**, which contains 1420 manually curated variants with experimentally measured $\Delta\Delta G$ values [25] extracted from ProTherm;
- **Broom**, which contains 599 variants with experimentally measured $\Delta\Delta G$ values[26] ;
- **Ssym**, which is a data set with an equal number of stabilizing and destabilizing mutations. It contains 684 variations in total, 342 direct variations and 342 reverse variations with both structures available[27]. This feature makes the dataset particularly useful for testing antisymmetry.
- **S669**, a manually-cleaned dataset extracted from ThermoMutDB ([28]) curated by our group. We recently extracted 900 variants from ThermoMutDB belong-

ing to proteins having less than 25% sequence identity with those of S2648 [24] and VariBench [29], whose union includes almost all variants available. From the 900 variants we excluded about ~24% of them due inconsistencies (e.g. free energies measured in terms of transition state kinetics, affinity binding, multiple variants, etc.) and we corrected some variants with wrong sign in the $\Delta\Delta G$. This datasets is particularly suitable for testing methods that were trained on the others.

- **P53 and Myoglobin**: two small datasets variations of p53 protein [30] and myoglobin [31] respectively of 42 and 134 mutations

The composition of all the datasets, their intersection and the distribution of their $\Delta\Delta G$s are reported in Tab.2.1, Fig.2.3 and Fig.2.2, respectively.

	Destabilizing	Neutral	Stabilizing
S2648	1597 (60%)	755 (29%)	295 (11%)
S669	387 (58%)	195 (29%)	85 (13%)
Ssym	225 (33%)	234 (34%)	225 (33%)
VariBench	800 (56%)	426 (30%)	194 (14%)
Broom	357 (60%)	171 (28%)	71 (12%)
P53	21 (50%)	19 (45%)	2 (5%)
Myoglobin	64 (48%)	55 (41%)	15 (11%)

Table 2.1: **Datasets composition.** The variants are grouped according to their $\Delta\Delta G$ values into three classes: destabilizing ($\Delta\Delta G \leq -0.5$ kcal/mol), neutral ($|\Delta\Delta G| < 0.5$ kcal/mol) and stabilizing ($\Delta\Delta G \geq 0.5$ kcal/mol). The corresponding percentages are reported into brackets.

2.2.3 Prediction assessment: performance metrics and open problems

Performance evaluation The most common metrics used to evaluate the performance of the computational methods are the Pearson correlation coefficient (indicated by r), the root mean square error (RMSE) and the mean absolute error (MAE).These metrics measures the accordance between the predicted and observed $\Delta\Delta G$ values. They are defined as:

$$r = \frac{Cov(\Delta\Delta G^{exp}, \Delta\Delta G^{pred})}{\sigma_{\Delta\Delta G^{exp}}\,\sigma_{\Delta\Delta G^{pred}}} \tag{2.7}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\Delta\Delta G_i^{exp} - \Delta\Delta G_i^{pred})^2}{N}}. \tag{2.8}$$

$$MAE = \frac{\sum_{i=1}^{N}|\Delta\Delta G_i^{exp} - \Delta\Delta G_i^{pred}|}{N}. \tag{2.9}$$

where Cov is the covariance matrix, σ represents the standard deviation and N is the number of total variants.

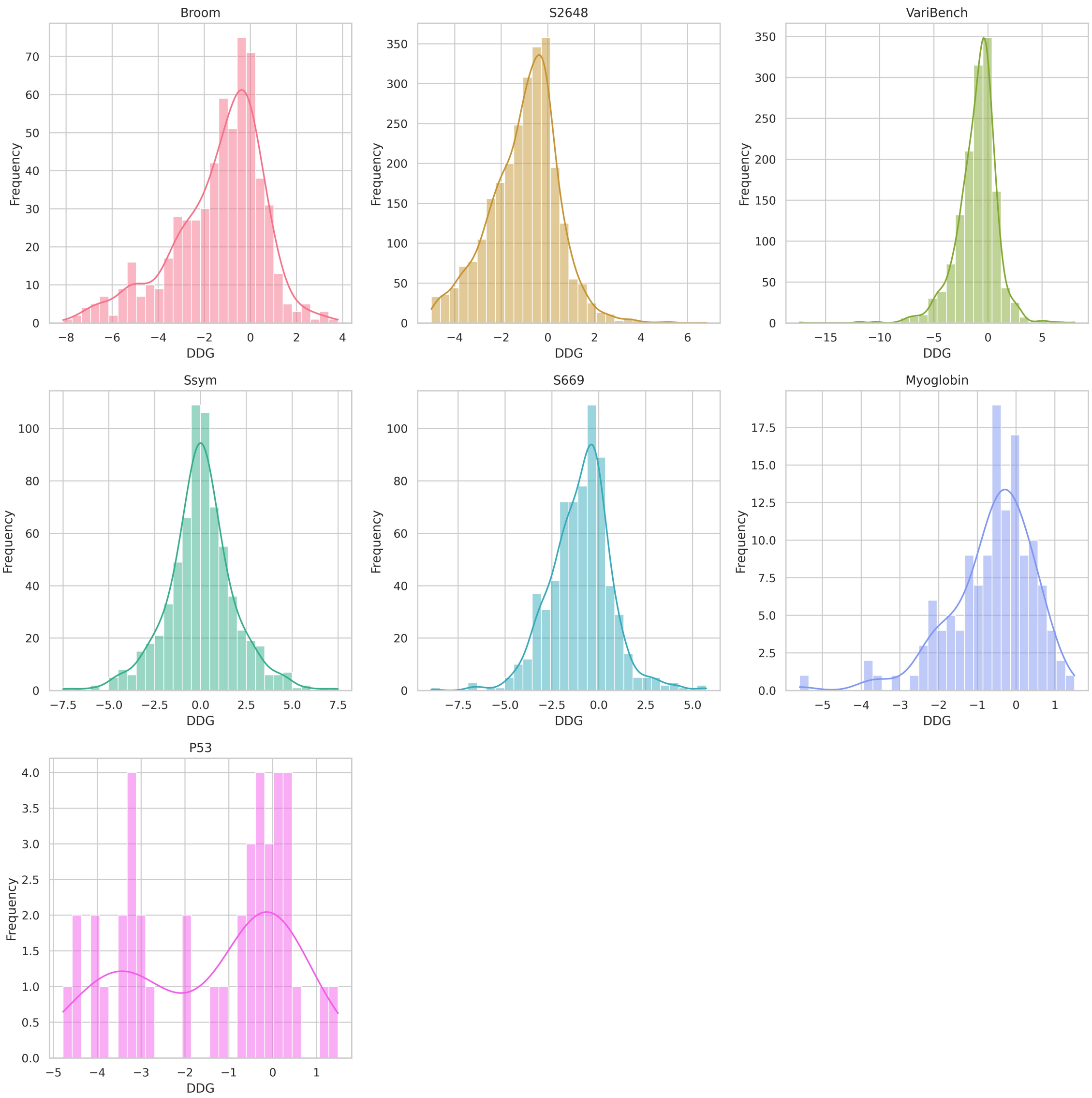

Figure 2.2: Distribution of the experimental $\Delta\Delta G$ (kcal/mol) values in the Ssym, S669, S2648, Broom, VariBench, P53 and Myoglobin datasets.

However, as already pointed out, predictors should not only be accurate based on the above metrics, but they should also respect the thermodynamics property of antisymmetry. To assess this property two scoring indices were generally adopted: r_{d-r} and $\langle\delta\rangle$. r_{d-r} is the Pearson correlation coefficient between the direct and the corresponding reverse variations:

$$r_{d-r} = \frac{Cov(\Delta\Delta G^{dir}, \Delta\Delta G^{rev})}{\sigma_{dir}\sigma_{rev}} \tag{2.10}$$

To measure the average bias towards a specific class, the **bias** score $\langle\delta\rangle$ is adopted, defined as:

$$\langle\delta\rangle = \frac{\sum_{i=1}^{N}(\Delta\Delta G_i^{dir} + \Delta\Delta G_i^{rev})}{2N}. \tag{2.11}$$

As shown in figure 2.2, all the datasets are strongly unbalanced towards the

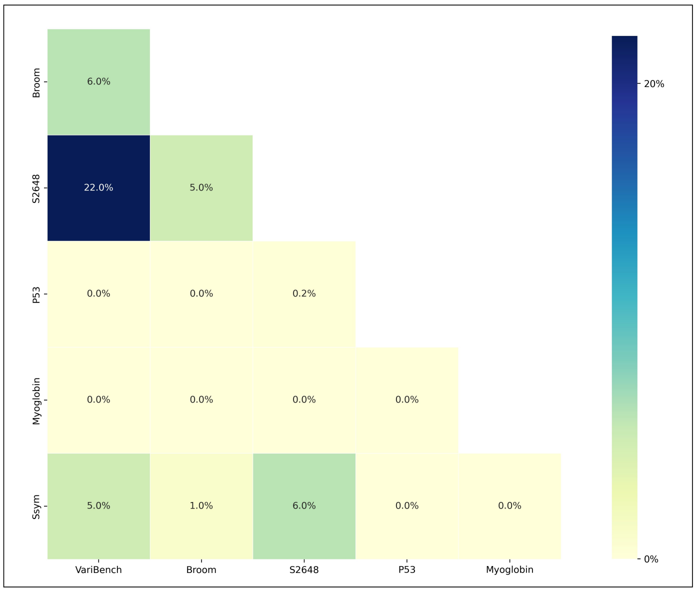

Figure 2.3: Heatmap showing the percentage of shared variants among the presented datasets

destabilizing class; consequently most of predictors suffers of high bias and very low antisimmetry. A perfectly antisymmetric method should have r_{d-r} equal to -1 and $\langle\delta\rangle$ equal to 0.

Upper and lower bounds for performance metrics Given the distribution of the experimental $\Delta\Delta G$s and the range of experimental errors, it is possible to compute upper and lower bound for performance metrics. In particular in [20, 21], authors show that these bounds depend on the average variance of the data $\overline{\sigma^2} = \frac{1}{N}\sum_i \sigma_i^2$, where σ_i^2 is the uncertainty associated with each $\Delta\Delta G$ point and on the variance of the distribution of the real $\Delta\Delta G$ values, $\sigma_{DB}^2 = \frac{1}{N}\sum_i(\mu_i - \overline{\mu})^2$, which depend only on the $\Delta\Delta G$ values in a specific dataset. Hence, the theoretical estimation for the pearson correlation coefficient (r) is lower than 1. The formula is:

$$\langle r\rangle \cong \frac{\sigma_{DB}^2}{\overline{\sigma^2} + \sigma_{DB}^2}, \tag{2.12}$$

This means that the theoretical value critically depends on both the average uncertainty of the data ($\overline{\sigma}^2$) and the spread of the dataset used (σ_{DB}^2).

Authors also derived a lower bound for the Root Mean Square Error (RMSE):

$$\langle RMSE\rangle \cong \sqrt{2\overline{\sigma}}. \tag{2.13}$$

Figure 2.4 shows the Pearson correlation coefficient r vs ($\overline{\sigma}^2$) for different (σ_{DB}^2) values: :

All the presented datasets have a $\sigma_{DB} < 2$, thus leading to an upper bound in the range of 0.70-0.85 to the Pearson correlation and a lower bound of about 1 kcal/mol for the root mean square error.

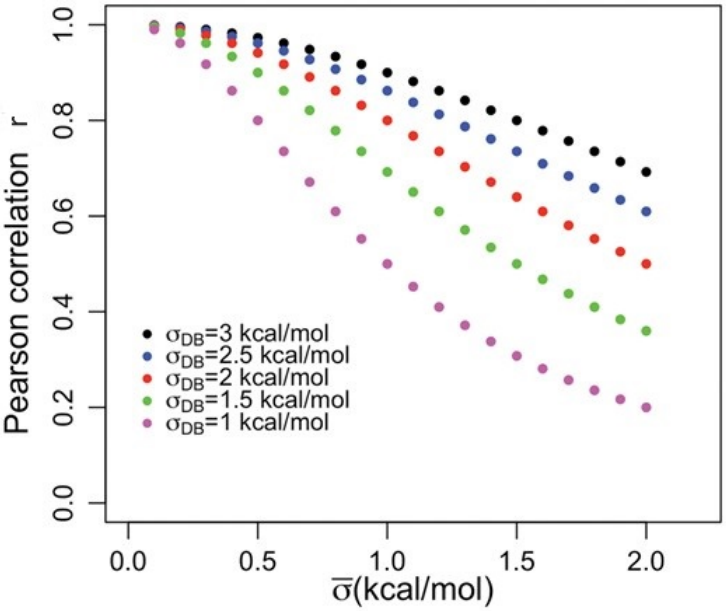

Figure 2.4

The burden of sequence identity When assessing performance of protein stability predictor, it is essential to consider the evolutionary links between proteins in the training and test sets. While proteins may differ in sequence, they can still be functionally or evolutionarily related, termed as homologs. Thus, to take this effect into account we use a 25% sequence identity cut-off to ensure that no two proteins across the data splits share more than a quarter of their amino acid sequences. Unfortunately, many existing tools neglect this step during the evaluation, potentially leading to overly optimistic results. In fact, it is crucial to consider sequence similarity also during cross-validation to obtain a more realistic error estimate. By ignoring it can result in a improper model training and can mask overfitting because the test set is not truly independent. Therefore, applying the 25% sequence identity threshold or a similar metric during the cross-validation process can help in generating a more robust and generalizable model.

Datasets biases As already shown in Table 2.1 and Fig.2.2, the most common datasets used for training are highly unbalanced towards the destabilizing variants. This bias in the data can propagate into the predictive models, making them more inclined to classify mutations as destabilizing. For instance, a model trained on a dataset where 80% of mutations are destabilizing is likely to incorrectly label most mutations in that manner. This affects both classification (sign of $\Delta\Delta G$) and regression (value of $\Delta\Delta G$) tasks. To mitigate this effect, recent computational approaches either inherently account for this imbalance with an appropriate loss function or artificially balance the dataset by introducing reverse variants. The antisimmetry 2.3 and bias 2.11 measures are good metrics to assess if predictors are correctly balanced.

Another source of bias in predictive models comes from the specific amino acids involved in mutations [32]. Alanine-related mutations are over-represented because many studies employ alanine scanning to assess individual residue contributions to protein stability. On the other hand, mutations involving amino acids like tryptophan,

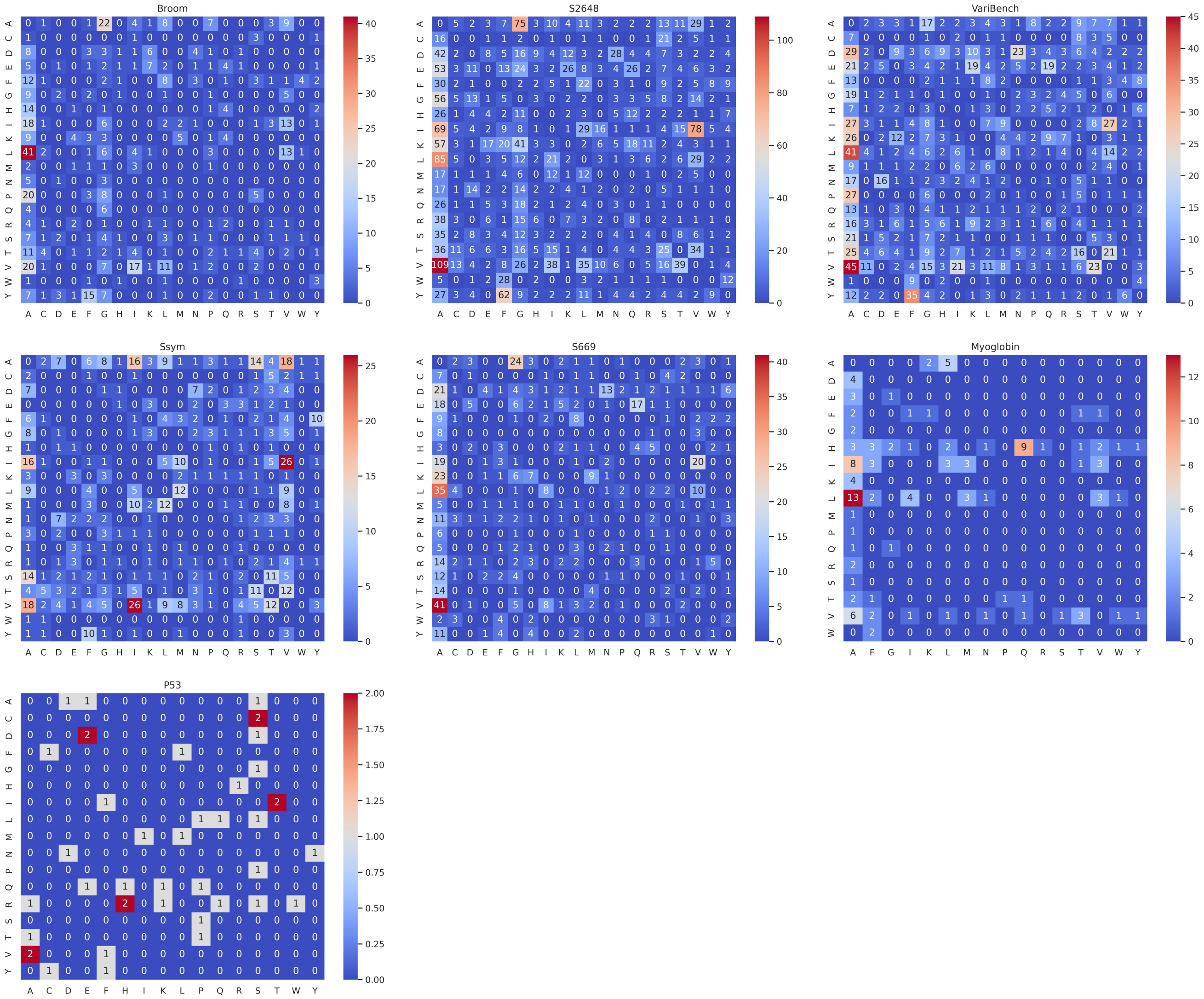

Figure 2.5: **Distribution of the amino acids involved in a mutation in each dataset** In all datasets, there is a an over-representation of mutation involving Alanine.

proline, or cysteine are less frequent due to their potential to disrupt the native protein structure. In Fig 2.5 we shown an heatmap representing all type of mutations for each presented datasets. In all datasets, except for P53 that is a very small one, there is a strong representation of mutation involving Alanine, while proline or cysteine are under represented.

This bias could have a large impact on the application of the methods to a broader context of mutations [32].

2.3 Discussion

In this chapter, we outlined fundamental $\Delta\Delta G$ properties that must be respected and we highlighted key challenges that must be overcome to enhance the accuracy of stability predictors. In particular it would important not only to increase dataset size but also its quality trying to balance it and to include a variety of amino acids. In addition new methods should respect both antisimmetry and transitivity properties. Finally during parameters optimization and metrics evaluation, sequence identity should be taken into account in order to fairly assess methods performance. In the next chapter we present our answer to the problem of antisymmetry by describing our new methods ACDC-NN, Antisymmetric Convolutional Differential Concatenated Neural Network and ACDC-NN-Seq, its sequence based counterpart, testing their performance on difference datasets and compare them with existing tools.

Chapter 3

ACDC-NN and ACDC-NN-Seq: antisymmetric neural networks

3.1 Introduction

As outlined in Chapter 2, various methods have been developed over the years for predicting protein stability. However, most existing methods do not preserve the property of antisymmetry. To address this issue, our lab has developed a method based on a siamese neural network, that uses a specifically designed loss function to ensure perfect antisymmetry. We have created two versions of this method: one that requires tertiary structural information of the protein (ACDC-NN) and a sequence-based version (ACDC-NN-Seq). The latter was developed because predicting Gibbs free energy changes solely from sequence information has the advantage of not requiring the more costly 3D structural data. Given that the latest release of the UniProtKB/TrEMBL protein database contains over 214 million sequence entries, including around 176,000 human proteins, and the Protein Data Bank has about 178,000 entries, with 52,485 of them being human, there is a pressing need for computational methods that can predict the impact of genetic variations on protein stability using only sequence information.

While deep neural networks are powerful, they usually require a large dataset for effective training. To face the challenge of limited availability of experimental $\Delta\Delta G$ data, we initially pre-trained ACDC-NN using another predictor, DDGun3D [33], as a teacher model. DDGun3D is primarily based on statistical potentials and is available in both sequence (DDGun) and structure-based (DDGun3D) versions. It has comparable performance with other state-of-the-art predictors and it respects the antisymmetric properties. After this pre-training phase, we employed a transfer learning strategy to fine-tune ACDC-NN on our smaller set of experimental data.

The work presented in this chapter is published in [34] and [35].

3.2 Materials and methods

3.2.1 Building ACDC-NN

Datasets During the ACDC-NN development we used both real and artificial datasets to train and test the method. We considered the already described S2648, Ssym, VariBench, p53 and Myoglobin datasets. In addition we used Ivankov 2000, which contains 2000 single-point variants with available structures, 1000 in a given direction and 1000 in the opposite one [36]. This dataset does not report experimental $\Delta\Delta G$ values because it has been specifically designed to score the anti-symmetric property of the available predictors; starting from the latter dataset we artificially generated IvankovDDGun. In particular we realised all the possible direct and reverse variations in every sequence position and assigning the corresponding DDGun3D predictions as $\Delta\Delta G$ values. This dataset was only used to pre-train the neural network.

Input ACDC-NN employs a convolutional architecture that accepts two distinct inputs: one for direct variations and another for inverse variations. These inputs undergo convolutional operations to extract relevant features. These extracted features are then fed into a pair of Siamese neural networks that operate with shared weights.

Each of the two ACDC-NN-Seq inputs consists of 160 elements to encode variation and sequence evolutionary information, while ACDC-NN's ones consists of 620 elements to also encode structure information:

- **Variation (V), common to both networks**: 20 features (one for each amino acid) coding for the variation by setting all the entries to 0 with the exception of the wild-type and the variant residue positions set to -1 and 1, respectively. This input corresponds to a one-dimensional matrix $V \in \mathbb{R}^{20\times 1}$;

- **Sequence (S or 1D-input), for ACDC-NN**: 100 features representing protein profile information of the sequence neighbourhood. Considering i as the variant position in the sequence, we used a window of 2 residues, i.e. $[i-2, i-1, i, i+1, i+2]$, so to obtain 20×5 elements, with the profile information of these 5 positions. This input corresponds to a matrix $S \in \mathbb{R}^{5\times 20}$;

- **Sequence (S or 1D-input), for ACDC-NN-Seq**: 140 features representing protein profile information of the variation neighbourhood. Considering i as the variant position in the sequence, we used a window of 3 residues, i.e., $[i-3; i+3]$, so to obtain 20×7 elements, with the profile information of these 7 positions. This input then corresponds to a sequence of 7 vectors taken from the protein profile. We experimented with some other input window sizes, but there was not much difference up to 11 residues, where we found a performance decrease;

- **Structure (T or 3D-input), for ACDC-NN**: 500 features representing protein profile information of the 3D-structure neighbourhood. We considered the residues up to 5 Å from the variation, taking a maximum of 25 residues sorted according to their distance in Å from the amino acid of interest. This input corresponds to a matrix $T \in \mathbb{R}^{25\times 20}$.

ACDC-NN and ACDC-NN-Seq modules A 2D Convolutional operation was applied on both S and T for ACDC-NN and only on the S matrix for ACDC-NN-Seq using a kernel equal to $(1, 20)$ and stride $(1, 1)$. This 2D Convolution generates a 20×20 filter matrix for each original matrix: K_S for S and K_T for T. The filter matrices K_S and K_T were learnt during the training phase.

$$\begin{aligned} \text{2D-Conv}(S, K_S) &= S' \qquad \text{where } S' \in \mathbb{R}^{5\times20} \\ \text{2D-Conv}(T, K_T) &= T' \qquad \text{where } T'' \in \mathbb{R}^{25\times20}. \end{aligned} \tag{3.1}$$

To make the ACDC-NN model invariant to the order of the 3D neighbours, 2D global average pooling operation (2D-GAP) was applied on T' (which encodes the structure information) :

$$\text{2D-GAP}(T') = T'' \qquad \text{where } T'' \in \mathbb{R}^{1\times20}. \tag{3.2}$$

Then the Dot product was calculated between the variant vector V and both T'' and S':

$$\begin{aligned} D &= T'' \cdot V, \qquad \text{where } D \in \mathbb{R}^{1\times1} \\ E &= S' \cdot V, \qquad \text{where } E \in \mathbb{R}^{5\times1}. \end{aligned} \tag{3.3}$$

As displayed in Figures 3.1 and 3.2, after convolution operations we concatenated the obtained features with the variant vector (V). In particular, for ACDC-NN we concatenated the variant vector V with 6 features for structure information (D) and with 5 features encoding the sequence information (E), while for ACDC-NN-Seq we only concatenated (V) with the 7 sequence encoding features (E). The concatenated features are used as input to a Differential Siamese Network [37, 38], both in ACDC-NN and ACDC-NN-Seq architectures. The Siamese Network is constituted by two neural network with share weights. The two outputs O_D for the direct variations and O_I for the reverse variations are combined in two final outputs,one wich represents the difference in $\Delta\Delta G$ between O_D and O_I while the other represents the average between the two. The meaning of the two outputs will be clarified in the next paragraph where we introduce the custom loss function we used to train the model.

The loss function In order to realised a perfectly antisymmetric predictor we designed a specific loss function that constrains the neural network to learn the property by considering the following equations:

$$\begin{aligned} &\Delta\Delta G_{WM} = \frac{1}{2}(\Delta\Delta G_{WM} + \Delta\Delta G_{WM}) = \frac{1}{2}(\Delta\Delta G_{WM} - \Delta\Delta G_{MW}) \\ &\Delta\Delta G_{WM} = -\Delta\Delta G_{MW} \Rightarrow \Delta\Delta G_{WM} + \Delta\Delta G_{MW} = 0. \end{aligned} \tag{3.4}$$

We implemented these constraints through the following customized loss function:

$$J = log(cosh(D - y)) + abs(S), \text{with } D = (O_D - O_I)/2,\ S = (O_D + O_I)/2, \tag{3.5}$$

where y is the experimental $\Delta\Delta G$ value, while O_D and O_I are the outputs of the direct and reverse modules of the Differential Siamese Network, respectively. The

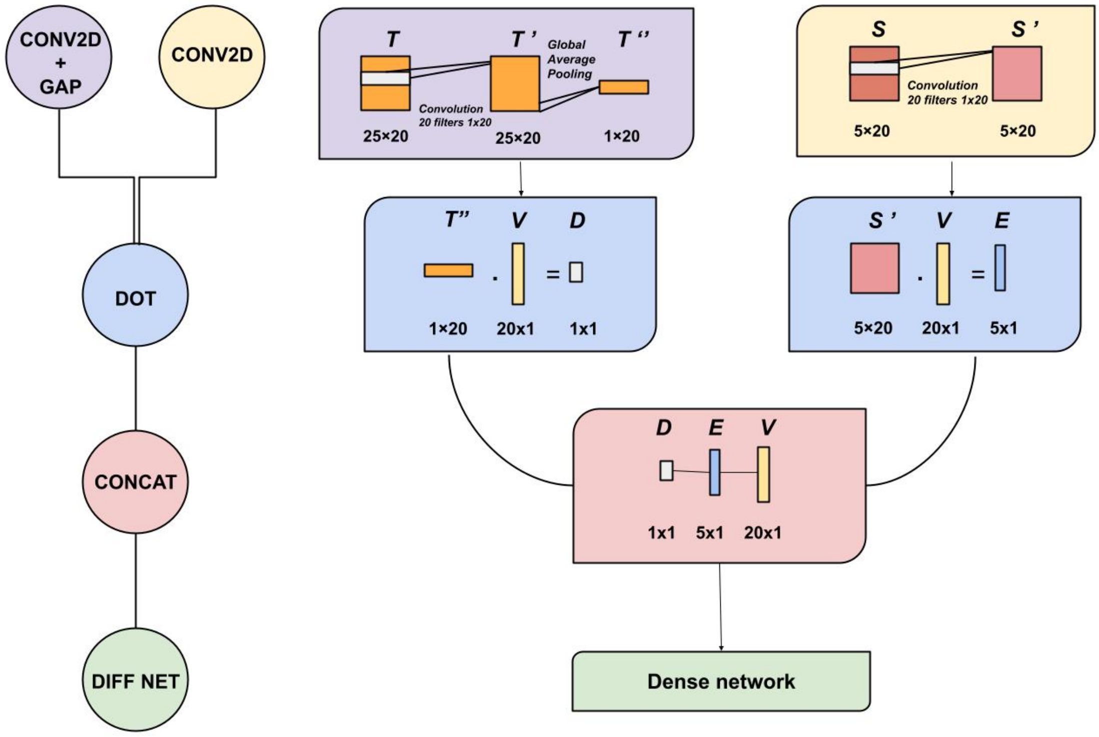

Figure 3.1: **Constituent module of ACDC-NN**. The module consists in a 2D-convolution operation applied to both T (3D inputs) and S (1D inputs) with 20 filters with kernel (1,20) and stride (1,1); 3D information undergoes 2D-Global Average Pooling (GAP). Dot product is then applied to both T" and S' processed information with the variation encoding vector V. Finally, all the 26 features are concatenated and used as input in a Dense Network. The notation used in the figure is the same as in the text.

$log(cosh(x))$ function was chosen since it is less sensitive to outliers like the mean absolute error, but it is still differentiable in 0 as the mean squared error. From the equation 3.5 it is now clear why the two Siamese Network outputs are combined in difference (D) and average (S). Given that O_D and O_I should have opposite sign, the difference (D) should be as close as possible to the real $\Delta\Delta G$ values (y), while the average S should be equal to 0, which means perfect antisimmetry. The complete ACDC-NN and ACDC-NN-Seq architectures are shown in Figures 3.3 and 3.4

Pre training phase The pre-training of the network was conducted on a subset of $\Delta\Delta G$ values artificially generated from the IvankovDDGun dataset. Here, we generated all permutations of sequence positions, including both direct and reverse variations, and we used the $\Delta\Delta G$ predictions from DDGun3D as the target output. The idea was to try to inherit the fundamental property of antisimmetry from DDGun3D using an architecture that can naturally learn the property with the specifically designed loss function. The ACDC-NN was trained on 400,000 simulated variations, divided between direct and inverse changes (200,000 each). Hyperparameter optimization was performed on a validation set of 100,000 examples, avoiding sequence overlap with the training data. The model was finally evaluated on an independent test set of the same size (100,000).

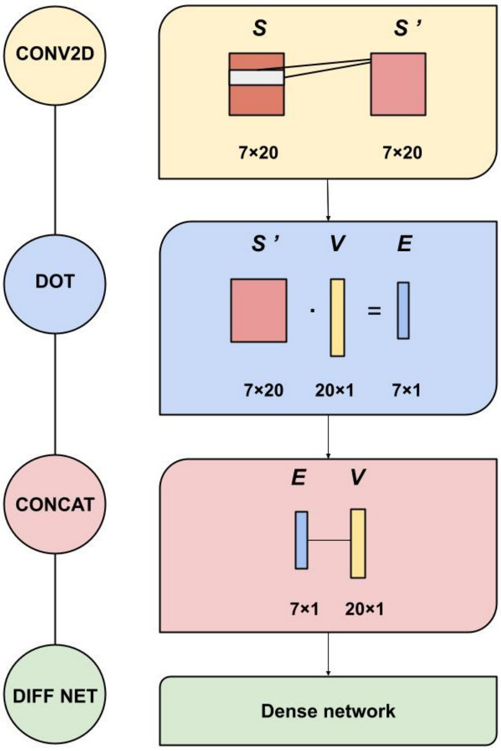

Figure 3.2: **Constituent modules of ACDC-NN-Seq.** CONV2D: a 2D-convolution operation applied to S (1D inputs) with 20 filters with kernel $(1, 20)$ and stride $(1, 1)$; DOT: dot product is applied to S' processed information with the variation encoding vector V; CONCAT: all the 27 features are concatenated and used as input in a Dense Network (DIFF NET).

Transfer Learning Starting with the pre-trained networks, we fine-tuned ACDC-NN using the s2648 dataset, which is comprised of experimental $\Delta\Delta G$ values. We employed a cross-validation strategy that eliminated sequence similarity across the training, validation, and test sets. The weights of the convolutional layers, responsible for capturing structural and sequence features, were frozen. Only the Differential Siamese network components were retrained. To augment the training data, we incorporated experimentally-determined structures from the Ivankov2000 dataset, assigning their corresponding DDGun3D predictions as target outputs.

ACDC-NN predictions The ACDC-NN model was built so that it can be used to make predictions in two different cases:

- when both the wild-type and variant structure are available, these are respectively used as direct and reverse inputs so that the network can provide a prediction that, by construction, is perfectly antisymmetric;
- when only the wild-type structure is available, as usual, the reverse input is created starting from the direct one by inverting the variation encoding, but preserving the same structure.

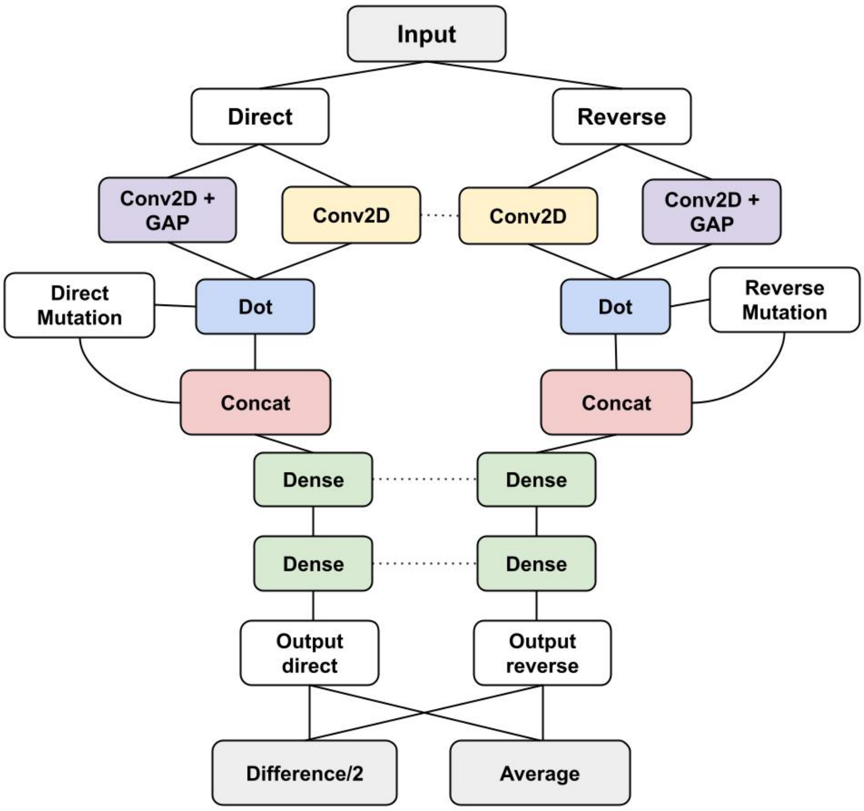

Figure 3.3: **Complete ACDC-NN architecture.** The module displayed in Figure 3.1 is used for both direct and reverse variations. A final layer takes the average and the difference between the two outputs (representing the $\Delta\Delta G$ predictions for direct and reverse variations, respectively). The two Siamese networks have shared weights, both in the convolutional and dense parts of the network, represented by the dashed lines.

3.3 Results

3.3.1 Learning anti-symmetry

During the pre-training phase on the synthetic IvankovDDGun dataset, we aimed to enable the network to inherit the antisymmetry property from DDGun3D. We assessed the network's performance using several metrics. The ability to reconstruct the predictions is measured with the Pearson correlation coefficient r (Eq.2.7) and root mean squared error (RMSE) (Eq.2.8); the $\Delta\Delta G$ antisymmetry is measured with r_{d-i} (equation 2.10) and the bias $\langle\delta\rangle$.

Table3.1 shows that on the external test set extracted from the IvankovDDGun both ACDC-NN and ACDC-NN-Seq were capable of learnDDGun3D predictions $(r = 0.97 - 0.98)$ while achieving perfect antisymmetry with r_{d-i} close to -1 and a bias δ close to 0.

The obtained results showed that using only sequence (ACDC-NN-Seq) and structure information (ACDC-NN) we were able to encode the information that DDGun3D calculates using the statistical potentials by Bastolla-Vendruscolo [39] and Skolnick [40], the BLOSUM62 substitution matrix [41], the Kyte-Doolittle hydrophobicity [42] and the residue solvent accessibility. In a more detailed analysis of the network, we found that the convolutional filters encode both structural and sequence information

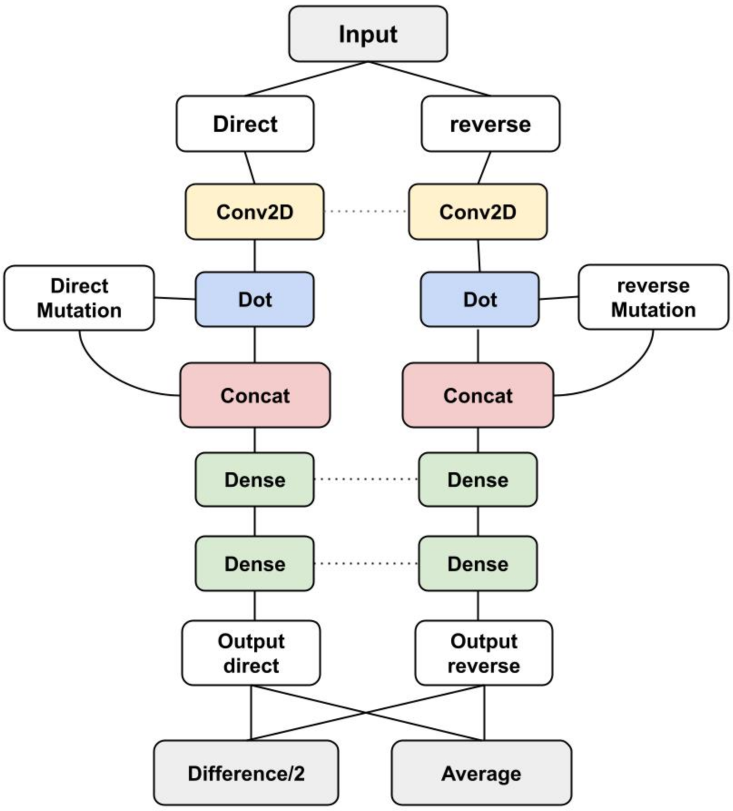

Figure 3.4: **Complete ACDC-NN-Seq architecture.** The module displayed in Figure 3.2 is used for both direct and reverse variations. Given a variation, we provide to the network its coding to the left, and the coding to the reverse variation to the right. A final layer takes the average and the difference between the two outputs. The difference computes $(\Delta\Delta G_{direct} - \Delta\Delta G_{inverse})/2$, which in case of perfect antisymmetry is exactly equal to $\Delta\Delta G$. The average computes $(\Delta\Delta G_{direct} + \Delta\Delta G_{inverse})/2$, which in case of perfect antisymmetry is equal to 0. The ACDC-NN-Seq outputs are estimations of the target $\Delta\Delta Gs$ learned during the training phase. The two Siamese networks have shared weights, both in the convolutional and the dense parts of the network.

Dataset	**Pearson/RMSE**	**Antisymmetry**	
	Direct	r_{d-i}	$\langle\delta\rangle$
ACDC-NN	0.98/0.26	−1.0	0.01
ACDC-NN-Seq	0.97/0.06	−1.0	0.00

Table 3.1: **Results on the IvankovDDGun test set:** The performance of ACDC-NN-Seq in learning DDGun3D was measured in terms of Pearson correlation coefficient (r) and root mean square error (RMSE). The antisymmetry property was assessed in terms of Pearson correlation coefficient (r_{d-i}) and the bias ($\langle\delta\rangle$) between the predicted values. RMSE and $\langle\delta\rangle$ are expressed in kcal/mol. IvankovDDGun (Test) is the test set extracted from the IvankovDDGun artificial dataset (never seen during training).

in a manner consistent with statistical potentials. Specifically, the network comprises two convolutional layers: one spanning the 3D residue contacts (K_T) and the other spanning the sequence nearest neighbors (K_S). The K_S filters cover an input that includes the central residue—the one undergoing the residue substitution—while the

K_T filters focus on an input containing only the 3D neighbors. The two types of filters, K_S and K_T, are illustrated in Figures 3.5 and 3.6.

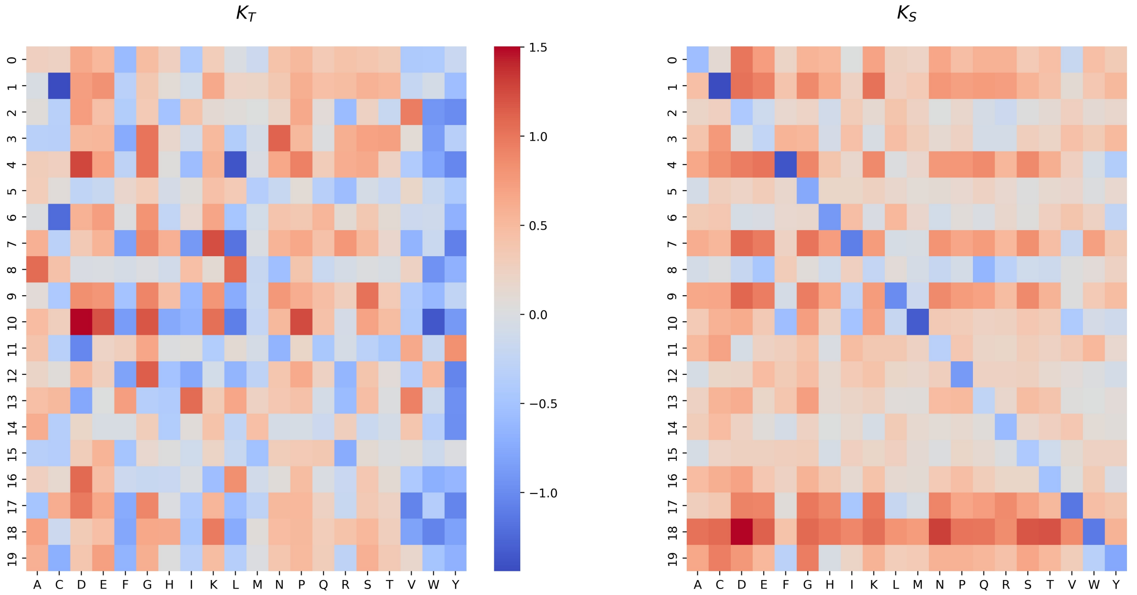

Figure 3.5: Heatmap of the filter matrix K_T and K_S spanning over the tridimensional and sequence nearest neighbour of the variant residue for ACDC-NN. The rows indicate the filters while the columns represent the residue position in the input matrix.

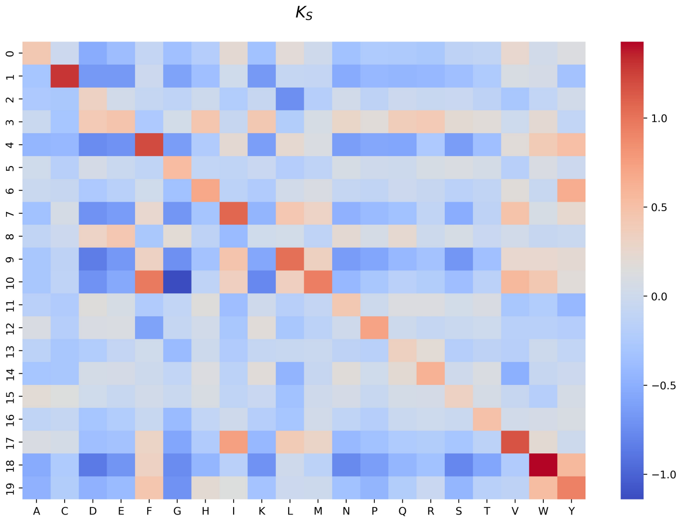

Figure 3.6: Heatmap of the filter matrix K_S for ACDC-NN-Seq spanning over the sequence nearest neighbour of the variant residue.

Both matrices correlate with statistical potentials and substitution matrices, indicating that the network is able to capture physical relationships within the data. Specifically, for ACDC-NN, the K_T matrix correlates with a Pearson coefficient of 0.51 with the three-dimensional statistical potential of Bastolla-Vendruscolo, while the K_S matrix correlates with the Blosum62 substitution matrix with a Pearson of

-0.75 and with Miyazawa's contact statistical potential with a Pearson of 0.60. The same can be said for ACDC-NN-Seq. Interestingly, the convolutional filter K_S not only correlates with 0.76 with Blosum62 and with Miyazawa with a Pearson of 0.65, but it also seems to encode structural information using only sequence data, thanks to labeled pre-training with DDGun3D; it correlates with Bastolla-Vendruscolo's potential at -0.45.

3.3.2 Fine tuning and prediction of experimental $\Delta\Delta G$ values

After the pre-training phase, the convolutional part of the network was frozen and last layers were retrained on the experimentally $\Delta\Delta G$ values from S2648 (both) and VariBench (only Seq) through a 10-fold cross-validation without sequence similarity. The optimal architectures resulted before and after transfer learning are presented in Table 3.2.

NN Parameters	**Before Transfer Learning**		**After Transfer Learning**	
	ACDC-NN	ACDC-NN-Seq	ACDC-NN	ACDC-NN-Seq
Hidden units	32,16	128,64	32,16	128,64
Dropout	0.2	0.05	0.2	0.35
Epochs	40	45	150	30
Batch-size	100	500	150	150
Optimizer	Adam	Adam	Adam	Adam
Loss	logcosh + abs	logcosh + abs	logcosh + abs	logcosh + abs

Table 3.2: **The optimal architectures of the Differential Net before and after transfer learning.** The optimal parameters were selected on a validation set without intersections or homologies with the training set.

Evaluation and comparison on the Ssym dataset To assess antisymmetry and prediction accuracy on experimental data, we utilized the Ssym dataset [27]. This dataset is designed to evaluate the predictive capability for both direct and reverse variants. For training, we used proteins with less than 25% sequence identity to those in the Ssym dataset. The obtained results were reported in Table 3.3 in terms of Pearson correlation coefficient (r) and root mean square error (RMSE) between the predicted values and the experimental $\Delta\Delta G$s. The antisymmetry was assessed through r_{d-i} and the bias $\langle\delta\rangle$. RMSE and $\langle\delta\rangle$ are expressed in kcal/mol.

The data in Table 3.3 indicate that ACDC-NN outperformed other structure-based methods in predicting reverse variants, closely followed by DDGun3D, which was used for pre-training. Except for ThermoNet and INPS, all methods that performed well on direct variants showed poorer results for reverse variants, revealing a lack of thermodynamic antisymmetry, as already highlighted by Pucci et al. [27, 43].

Regarding sequence-based predictors, ACDC-NN-Seq predicts equally well both direct and reverse variants with nearly perfect antisymmetry (-0.99). ACDC-NN-Seq performance is higher than the one obtained by INPS-Seq, which is the only machine-learning method proven to be antisymmetric [43, 44]. It is worth highlighting that, among the methods reported in Table 3.3, only Inps-NoSeqId, ThermoNet and ACDN-NN* were trained in cross-validation removing the sequence identity (i.e. sequence similarity $< 25\%$).

In addition, ACDC-NN showed very good results in antisymmetry (-0.98) when only one structure was available (ACDC-NN) and it achieved perfect antisymmetry (-1.00) with both structures (ACDC-NN+).

Evaluation and comparison on p53 and Myoglobin datasets Finally we evaluated and compare both ACDC-NN and ACDC-NN-Seq on the smaller p53 transcription factor and the myoglobin datasets. Also in this case, since both myoglobin and p53 are present in the S2648 training dataset, we predicted the variants in cross-validation, removing sequence similarity with the two proteins. In order to predict the reverse variants we used MODELLER [61] to generate the mutated structures. The performance obtained by ACDC-NN outperforms ThermoNet, the other available deep-learning approach both on Myoglobin and P53 in all the metrics. Regarding the sequence based method ACDC-NN-Seq is second only to INPS.

3.4 Discussion

Very few methods in the literature satisfy the principle of antisymmetry imposed by thermodynamics, either due to the unbalance of training datasets or their implementation. ACDC-NN incorporates the principle of thermodynamic antisymmetry into its learning process, achieved through a custom-designed loss function. This ensures minimal discrepancy between direct and reverse predictions. When 3D structures for both direct and reverse variants are available, ACDC-NN achieves perfect antisymmetry, while in cases where only one structure is known, the model generates a reverse variant, maintaining high-quality predictions and near-perfect antisymmetry, as shown in Table 3.3. As demonstrated through an analysis of the convolutional filters 3.5, both ACDC-NN and ACDC-NN-Seq exhibit physical consistency and are capable of internally encoding evolutionary information and statistical potentials solely based on the target $\Delta\Delta G$.

Finally, both ACDC-NN and ACDC-NN-Seq were tested on various datasets on protein variants dissimilar to those used in the training set and they show comparable or better performance to existing methods while maintaining thermodynamic antisymmetry.

In conclusion, the method we have developed proves to be a good predictor that can be used for predicting stability changes, both when structural information is available or only with sequence information mantaining its core property and without losing too much accuracy. The code for ACDC-NN and ACDC-NN-Seq is publicly available at `https://github.com/compbiomed-unito/acdc-nn`.

Method	Pearson/RMSE		Antisymmetry	
	Direct	Reverse	r_{d-i}	$\langle\delta\rangle$
Structure-based				
ACDC-NN+	**0.57/1.45**	**0.57/1.45**	**−1.00**	**0.00**
ACDC-NN	**0.58/1.42**	**0.55/1.47**	**−0.99**	**−0.01**
INPS-NoSeqId [43]	0.48/1.42	0.47/1.45	−0.99	−0.06
ThermoNet [45]	0.47/1.56	0.47/1.55	−0.96	−0.01
DDGun3D [33]	0.56/1.42	0.53/1.46	−0.99	−0.02
INPS3D [46]	0.59/1.29	0.44/1.64	−0.86	−0.55
PopMusicSym [27]	0.48/1.58	0.48/1.62	−0.77	0.03
SDM [47]	0.51/1.74	0.32/2.28	−0.75	−0.32
CUPSAT [48]	0.39/1.71	0.05/2.88	−0.54	−0.72
Rosetta [49]	0.69/2.31	0.43/2.61	−0.41	−0.69
FoldX [50]	0.63/1.56	0.39/2.13	−0.38	−0.47
Maestro [51]	0.52/1.36	0.32/2.09	−0.34	−0.58
PoPMuSiC [24] v2.1	0.63/1.21	0.25/2.18	−0.29	−0.71
mCSM [30]	0.61/1.23	0.14/2.43	−0.26	−0.91
DUET [52]	0.63/1.20	0.13/2.38	−0.21	−0.84
AUTOMUTE [53]	0.73/1.07	−0.01/2.61	−0.06	−0.99
iSTABLE [54]	0.72/1.10	−0.08/2.28	−0.05	−0.60
I-Mutant [55] v3.0	0.62/1.23	−0.04/2.32	0.02	−0.68
NeEMO [56]	0.72/1.08	0.02/2.35	0.09	−0.60
STRUM [57]	0.75/1.05	−0.15/2.51	0.34	−0.87
Sequence-based				
ACDC-NN-Seq	**0.55/1.44**	**0.55/1.44**	**−0.99**	**−0.01**
INPS [58]	0.51/1.42	0.50/1.44	−0.99	−0.04
I-Mutant2.0 [55]	0.7/1.12	0.05/2.54	−0.17	−1.01
MUpro [59]	0.79/0.94	0.07/2.51	−0.02	−0.97
SAAFEC-SEQ [60]	0.71/1.09	−0.39/2.71	0.58	−1.84

Table 3.3: **Results on Ssym:** The performance on direct and reverse variants was measured in terms of Pearson correlation coefficient (r) and root mean square error (RMSE). The antisymmetry was assessed using the correlation coefficient r_{d-i} (Eq,2.10) and the bias $\langle\delta\rangle$ (Eq.2.11). RMSE and $\langle\delta\rangle$ are expressed in kcal/mol. All the predictions of the methods were taken from Pucci et al. [27], except for INPS-NoSeqId, INPS, INPS3D, DDGun3D and ThermoNet. The results of SAAFEC-SEQ and I-mutant2.0 were obtained using their stand-alone code, those of MUpro were obtained using the webserver available. The results of DDGun3D, INPS and ThermoNet were taken from Montanucci et al. [33], Fariselli et al. [58] and from [45], respectively. ACDC-NN+ reports the cross-validation performance using in input both structures, wild-type and mutated (as experimentally defined). Only Inps-NoSeqId, ThermoNet, ACDN-NN and ACDC-NN+ were trained in cross-validation addressing the sequence identity issue (sequence similarity $< 25\%$).

Method	**Pearson/RMSE**		**Antisymmetry**	
	Direct	**Reverse**	r_{d-i}	$\langle\delta\rangle$
Structure-based				
ACDC-NN+	0.58/0.89	0.58/0.89	−1.00	0.00
ACDC-NN	0.58/0.89	0.57/0.89	−0.99	−0.01
ThermoNet	0.38/1.16	0.37/1.18	−0.97	−0.02
Sequence-based				
ACDC-NN-Seq	0.56/0.97	0.56/0.97	−1.00	0.00
INPS	0.60/0.99	0.61/0.98	−1.00	0.01
SAAFEC-SEQ	0.63/0.89	0.30/1.63	−0.21	−1.50
I-Mutant2.0	0.56/1.12	0.39/1.71	−0.45	−0.88
MUpro	0.51/0.99	0.35/1.75	−0.17	−0.79

Table 3.4: **Results on myoglobin.** Comparison on structure-based methods and sequence-based methods on myoglobin. We compared ACDC-NN with Thermonet, the other available deep-learning approach. ACDC-NN+ values were obtained when both direct and reverse structures are used as input. The results of ThermoNet were taken from [45]. We compare ACDC-NN-Seq with INPS, SAAFEC-SEQ, I-mutant2.0 and MUpro.The INPS, SAAFEC-SEQ and I-mutant2.0 results were obtained using their stand alone code, those of MUpro were obtained using the webserver available.

Method	**Pearson/RMSE**		**Antisymmetry**	
	Direct	**Reverse**	r_{d-i}	$\langle\delta\rangle$
Structure-based				
ACDC-NN+	0.61/1.69	0.61/1.69	−1.00	0.00
ACDC-NN	0.62/1.67	0.61/1.72	−0.99	−0.01
ThermoNet	0.45/2.01	0.56/1.92	−0.93	−0.04
Sequence-based				
ACDC-NN-Seq	0.62/1.62	0.62/1.62	−1.00	0.00
INPS	0.72/1.49	0.70/1.54	−0.99	−0.01
SAAFEC-SEQ	0.52/1.64	−0.18/2.97	0.06	−1.79
I-Mutant2.0	0.35/1.75	0.22/2.81	−0.24	−1.02
MUpro	0.23/1.78	0.04/2.87	0.12	−0.98

Table 3.5: **Results on p53.** Comparison on structure-based methods and sequence-based methods on p53. We compared ACDC-NN with Thermonet, the other available deep-learning approach. ACDC-NN+ values were obtained when both direct and reverse structures are used as input. The results of ThermoNet were taken from [45]. We compare ACDC-NN-Seq with INPS, SAAFEC-SEQ, I-mutant2.0 and MUpro.The INPS, SAAFEC-SEQ and I-mutant2.0 results were obtained using their stand alone code, those of MUpro were obtained using the webserver available.

Chapter 4

Thorough Comparison of the available tools on the 669 dataset: pitfalls and suggestions

4.1 Introduction

Most available $\Delta\Delta G$ predictors use the same data for the training phase (in the case of ML predictors) or to extract biological scores and statistical potentials (in the case of energy-based predictors and others). Thus, considering there is a significant overlap between datasets, it has always been challenging to appropriately and fairly evaluate the performance of these tools. To address this issue, we have manually curated a new dataset from a recent source, ThermoMutDB [28]. From version 1.3 of ThermoMutDB we identified 669 novel variants with less than 25% sequence identity with those present in the most commonly used datasets such as S2648 [24] and VariBench [29].

To thoroughly assess prediction performance, we artificially balanced the dataset by creating the reverse mutations and we evaluated and compared the commonly used computational tools, described in section 4.2.1. Results clearly show a group of methods that outperformed the others, achieving a Pearson correlations around 0.5 and performing equally well on both direct and reverse variations, thus respecting the antisimmetry property.

Our analysis also highlighted that, among various methods, the ability to classify stabilizing, neutral, and destabilizing variants was generally more precise for destabilizing variants compared to stabilizing ones, particularly when focusing solely on direct variants. This observation was true even for the most recent tools, that tried to artificially balance or encode the anti-symmetry property in the method.

By respecting the anti-symmetry property and/or by balancing its training dataset, a method should theoretically be able to predict the stabilizing mutations with the same accuracy as the destabilizing mutations. While this statement was true when considering both direct and reverse variants, the performances were highly unbalanced when considering only the direct ones.

A recent study by [62] highlighted how some of the commonly used features are relevant only to score the destabilizing variants and not helpful for the stabilizing ones. This study demonstrate an intrinsic difference between the type of variants (stabilizing vs destabilizing) and suggest the importance of finding and using different properties, in order to correctly described the stabilizing mutations.

4.2 Materials and methods

4.2.1 Evaluated methods

We predicted the $\Delta\Delta G$s on the S669 dataset with 21 different tools. Either web server (when available) or stand-alone versions were used with default parameters, as indicated in the following:

- **ACDC-NN** [34] and its sequence-based version **ACDC-NN-Seq** [35] (stand-alone tool): neural network-based methods whose architectures satisfy the antisymmetry properties by construction. They both take as input the local information from the amino acids in the neighbourhood of the mutation and they both use multiple sequence alignments considering the two amino acids involved in the mutation.
- **DDGun3D** and **DDGun** [33] (stand-alone tool): untrained methods that combine evolutionary information and statistical potentials to predict the $\Delta\Delta G$. Compared to the sequence-based DDGun, DDGun3D includes the structural information scored by the Bastolla-Vendruscolo statistical potential [39] and weights the linear combination through the accessibility of the mutated amino acid. They both include antisymmetric features and provide an easy extension to the prediction of multiple variations.
- **mCSM** [30] (web server): considers graph-based structural signatures, encoding for the distance patterns between atoms and used to represent the protein residue environment, to study and predict the impact of single-point mutations on the protein stability.
- **SDM** [47] (web server): statistical potential energy function that uses environment-specific amino-acid substitution frequencies within homologous protein families to calculate a stability score as proxy of the free energy difference between the wild-type and mutant protein.
- **DUET** [52] (web server): web server implementing a meta-classifier based on the combined results from mCSM and SDM using Support Vector Machines (SVM).
- **Dynamut and Dynamut2** [63, 64] (web server): machine learning methods implementing a consensus prediction. They combine the effects of mutations on protein stability and dynamics calculated by DUET, Bio3D and ENCoM to generate an optimized and more robust predictor.
- **FoldX** [50] (stand-alone tool): empirical force field-based method predicting the effect of a single-point variation through a linear combination of empirical free energy terms, including entropy contribution, Van der Walls forces, hydrogen

bonds and electrostatic interactions.

- **SAAFEC-SEQ** [60] (web server): gradient boosting decision-tree machine learning method that uses physico-chemical properties, sequence features and evolutionary information to predict the $\Delta\Delta G$ values.

- **MUpro** [59] (web server): sequence-based SVM-based approach which considers the local mutation environment encoding the residues in a window centered on the target residue. The input corresponding to the deleted residue is set to -1 and the newly introduced residue to 1; all other inputs are set to 0.

- **Rosetta** [49] (stand-alone tool): a method based on structural modeling that computes the difference in Rosetta energy between the simulated wild-type versus the mutated structures.

- **ThermoNet** [45] (stand-alone tool): deep 3D-convolutional neural network designed for structure-based prediction of the $\Delta\Delta G$ values. Input protein structures are treated as if they were multi-channel 3D images, therefore by using multi-channel voxel grids based on biophysical properties derived from raw atom coordinates.

- **PremPS** [65] (web server): random forest regression-based method that uses evolutionary and structure-based features to make $\Delta\Delta G$ predictions. It has been trained on a balanced dataset with an equal number of stabilizing and destabilizing mutations to obtain unbiased predictions.

- **PoPMuSiC** [24] (web server): energy function-based method providing a linear combination of 13 statistical potentials, two volume-dependent terms of the wild-type and mutant amino acids, and an independent term. The coefficients depend on the solvent accessibility of the mutated residue, based on a sigmoid function whose parameters are optimized through a neural network.

- **MAESTRO** [51] (stand-alone tool): multi-agent prediction method based on statistical scoring functions (SSFs) and exploiting an ensemble of neural networks, support vector and multiple linear regressors, combined into a consensus model.

- **INPS3D** [66] and its sequence-based version **INPS** [44] (stand-alone tool): SVM-based methods using radial basis function kernel. Specifically, INPS uses the substitution score derived from the BLOSUM62 matrix, the difference in the alignment score between the native and variant sequences, hydrophobicity, evolutionary information and others; INPS3D also considers the relative solvent accessibility of the native residue and the difference between wild-type and mutated structures, scored by the Bastolla-Vendruscolo statistical potential [39].

- **I-Mutant** and its sequence-based version **I-Mutant-Seq** [55] (web server): SVM-based methods using radial basis function kernel with 42 features as input, including temperature, Ph, 20 features encoding for the mutations and 20 features encoding for the spatial residue environment when the protein structure is available or the nearest sequence neighbours when only the protein sequence is available.

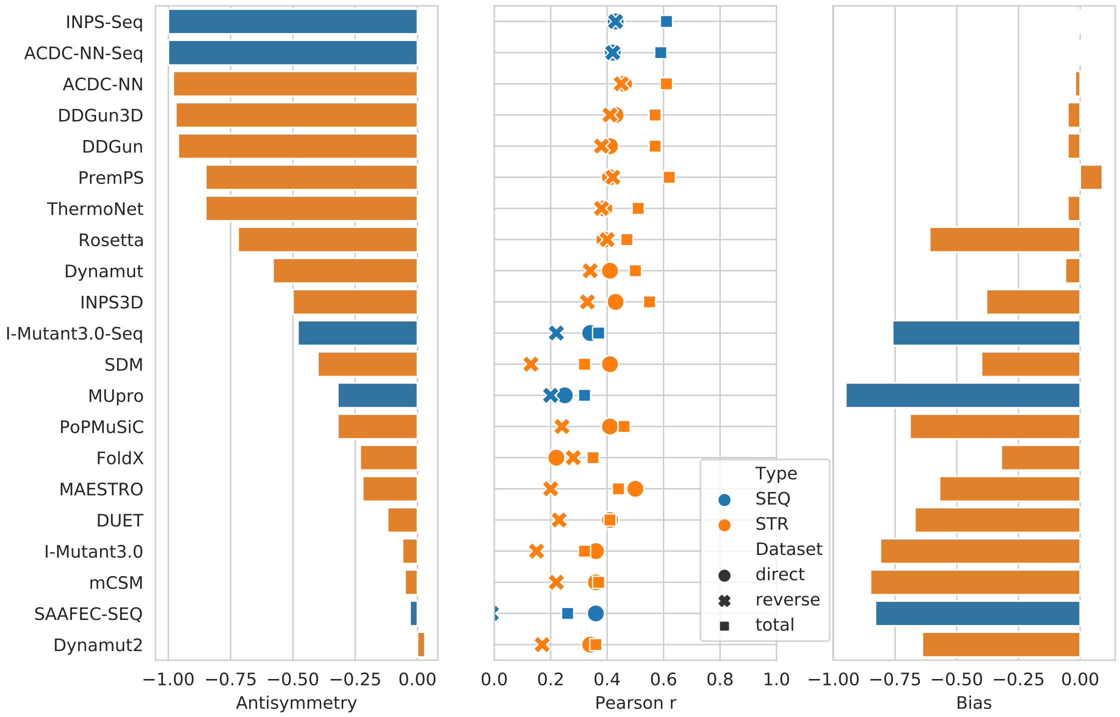

Figure 4.1: **Antisymmetry, Pearson correlation and Bias for all predictors**. From the left: antisymmetry expressed as the Pearson correlation between direct and reverse $\Delta\Delta G$ predictions, where perfect antisymmetry corresponds to -1; Pearson correlation of predicted with experimental $\Delta\Delta G$ values for the sets of direct, reverse and total (both direct and reverse) variants; bias expressed as the average of the predicted $\Delta\Delta G$ on the total (direct and reverse) dataset: since the average experimental $\Delta\Delta G$ on the total dataset is zero, unbiased predictors have also a bias of 0, while predictors biased towards destabilization have negative values. Colour show which predictors need structural (3D) data in orange and which use only sequence data in blue. Predictors are sorted from the most antisymmetric (top) to the least (bottom).

4.3 Results

4.3.1 Method performance on the new S669 dataset

To assess the generalization capability of different prediction methods of protein stability changes, we performed the analysis on S669, a dataset, as previously mentoned, of never seen variants with less than 25% of sequence identity to previously studied proteins in most commonly used datasets (S2648 and Varibench).

Fig.4.1 and Table 4.1 show the performance metrics, such as Pearson correlation, RMSE and MAE, the bias and the antisymmetry of each method. The highest correlations observed across all the methods are in the range 0.4-0.6 depending on the group of considered variants.

When direct variants are considered, the Pearson correlation of all the methods ranges from 0.2 to 0.5 (Fig. 4.1, circles in the central plot). These values are lower than those reported in the original papers but close to the expected performance for methods developed avoiding proteins with high sequence similarity in the training and testing sets to avoid overfitting. It is worth noticing that S669 can be considered

	Total			**Direct**			**Reverse**			**Antisimmetry/Bias**	
Method	**r**	**RMSE**	**MAE**	**r**	**RMSE**	**MAE**	**r**	**RMSE**	**MAE**	r_{d-r}	$\langle\delta\rangle$
Structure-based											
ACDC-NN	0.61	1.5	1.05	0.46	1.49	1.05	0.45	1.5	1.06	-0.98	-0.02
DDGun3D	0.57	1.61	1.13	0.43	1.6	1.11	0.41	1.62	1.14	-0.97	-0.05
PremPS	0.62	1.49	1.07	0.41	1.5	1.08	0.42	1.49	1.05	-0.85	0.09
ThermoNet	0.51	1.64	1.2	0.39	1.62	1.17	0.38	1.66	1.23	-0.85	-0.05
Rosetta	0.47	2.69	2.05	0.39	2.7	2.08	0.4	2.68	2.02	-0.72	-0.61
Dynamut	0.5	1.65	1.21	0.41	1.6	1.19	0.34	1.69	1.24	-0.58	-0.06
INPS3D	0.55	1.64	1.19	0.43	1.5	1.07	0.33	1.77	1.31	-0.5	-0.38
SDM	0.32	1.93	1.45	0.41	1.67	1.26	0.13	2.16	1.64	-0.4	-0.4
PoPMuSiC	0.46	1.82	1.37	0.41	1.51	1.09	0.24	2.09	1.64	-0.32	-0.69
MAESTRO	0.44	1.8	1.3	0.5	1.44	1.06	0.2	2.1	1.66	-0.22	-0.57
FoldX	0.31	2.39	1.53	0.22	2.3	1.56	0.22	2.48	1.5	-0.2	-0.34
DUET	0.41	1.86	1.39	0.41	1.52	1.1	0.23	2.14	1.68	-0.12	-0.67
I-Mutant3.0	0.32	1.96	1.49	0.36	1.52	1.12	0.15	2.32	1.87	-0.06	-0.81
mCSM	0.37	1.96	1.49	0.36	1.54	1.13	0.22	2.3	1.86	-0.05	-0.85
Dynamut2	0.36	1.9	1.42	0.34	1.58	1.15	0.17	2.16	1.69	0.03	-0.64
Sequence-based											
INPS-Seq	0.61	1.52	1.1	0.43	1.52	1.09	0.43	1.53	1.1	-1	0
ACDC-NN-Seq	0.59	1.53	1.08	0.42	1.53	1.08	0.42	1.53	1.08	-1	0
DDGun	0.57	1.74	1.25	0.41	1.72	1.25	0.38	1.75	1.25	-0.96	-0.05
I-Mutant3.0-Seq	0.37	1.91	1.47	0.34	1.54	1.15	0.22	2.22	1.79	-0.48	-0.76
MUpro	0.32	2.03	1.58	0.25	1.61	1.21	0.2	2.38	1.96	-0.32	-0.95
SAAFEC-SEQ	0.26	2.02	1.54	0.36	1.54	1.13	-0.01	2.4	1.94	-0.03	-0.83

Table 4.1: **Assessment of the protein stability prediction tools on s669.** Performance reported in terms of Pearson correlation coefficient (r), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The antisymmetry property was assessed in terms of Pearson correlation coefficient(r_{d-r}) and bias ($\langle\delta\rangle$), as described in Section 2.2.3. RMSE, MAE, $\langle\delta\rangle$, and $\langle\gamma\rangle$ are expressed in kcal/mol. The methods are ordered by r_{d-r}.

an external validation set for the tested methods; thus, a performance drop can be expected.

When the reverse variants are considered (Fig. 4.1, crosses in the central plot), there is a first group of methods built to be antisymmetric (INPS-Seq, ACDC-NN-Seq, ACDC-NN, DDGun3D, DDGun, ThermoNet, PremPS), which perform significantly better, followed by Rosetta, Dynamut and INPS3D. On the other hand, the not-antisymmetric predictors (I-Mutant3.0-Seq, SDM, MUpro, PoPMuSiC, MAESTRO, FoldX, DUET, I-Mutant3.0, mCSM, Dynamut2, SAAFEC-SEQ) performed remarkably worse for the reverse variants, showing a strong bias towards the destabilizing class (negative values), as highlighted by the values reported in the last columns of Tables 1 and in right-most bar plot of Fig. 4.1.

The majority of the methods improve when we consider the complete and balanced dataset (Fig. 4.1, squares in the central plot). This improvement is partially due to the increase of the $\Delta\Delta G$ distribution variance [20, 21] and most likely due to the learnt thermodynamic property, that allows the antisymmetric methods to increase the performance.

In our study we also evaluated the dependence of the method performance in the choice of the protein structure. We found that most methods are quite insensitive to experimental strategy, but we observed a general trend of slightly increased performance for NMR-derived structures. The overall performance of the methods seems mostly unaffected also by the X-ray resolution, at least in the range from 1.2 to 3.2 Angstrom seen in S669. This analysis is better described in Appendix A.0.2. Simlarly we also investigated the effect of the environmental conditions, i.e pH and temperature. We conducted the analysis splitting variants whose temperature and pH are in physiological conditions (T in range $20 - 40C$ and pH $6 - 8$) and those that aren't. Results suggest that there are no clear evidence to conclude that non-physiological conditions lead to more prediction errors. Additionally, residue accessibility appeared to have no significant effect on the performance of the methods. Further details on both these analyses can be found in Appendix A.0.3, where they are explored more extensively.

4.3.2 Classification performance

In many applications, the identification of destabilizing and stabilizing variations is more relevant than the prediction of the exact $\Delta\Delta G$ value. In Fig. 4.2 we evaluated the classification accuracy of the different methods. The figure shows three broad groups with similar accuracy in the various stability classes and variant subsets. The first group is represented by the most antisymmetric and unbiased predictors: PremPS, ACDC-NN, ACDC-NN-Seq, DDGun3D, DDGun, Dynamut, ThermoNet, INPS-Seq, INPS3D and FoldX. They showed good performance in both stabilizing and destabilizing classes, especially PremPS, ACDC-NN, DDGun and INPS-Seq. However, all these predictors showed a lower accuracy in the under-represented direct-stabilizing variants and their reverse class, i.e. the reverse-destabilizing variants. This is especially true for the best performing PremPS, ACDC-NN and INPS-Seq, suggesting a trade-off where they "sacrifice" accuracy in the smaller classes for greater scores in the whole dataset. A second group includes I-Mutant3.0, I-Mutant3.0-Seq, mCSM, MUpro and the newer SAAFEC-SEQ. They are heavily biased towards destabilizing predictions, therefore their accuracies on stabilizing variants are extremely low in all the datasets. The remaining group includes Rosetta, SDM, Maestro, DUET, Dynamut2 and PoPMuSiC, which still showed a bias towards destabilization but to a lower extent, with Rosetta and SDM being quite balanced across different classes. Among the tested sequence-based methods, the more balanced and the best performing tools are: INPS-Seq, ACDC-NN-Seq and DDGun.

4.3.3 The unbalanced predictions among stabilizing, neutral and destabilizing variants

In Fig 4.3, we can see the prediction distributions of the various methods on both direct and reverse variants. The variants are split by their experimental $\Delta\Delta G$ into destabilizing ($\Delta\Delta G \leq -0.5$), neutral ($|\Delta\Delta G| < 0.5$) and stabilizing variants ($\Delta\Delta G \geq 0.5$). Compared to the experimental distributions, all the methods tended to compress their predictions towards zero (neutral), generating a significant overlap among the three distributions (Rosetta is the only exception here). However, many of them maintained the relation of order among the three classes except for SAAFEC-SEQ (Fig. 4.3). The sequence-based DDGun appeared to be the only one that consistently keeps a minimum of difference among the three types of prediction distributions separating the means and the box quantile borders (stabilizing, neutral and destabilizing) for all direct and reverse sets. As shown in Fig 4.3, when considering both direct and reverse variants, the methods that do not respect the antisymmetry property have a more unbalanced performance.

Examining only direct variant predictions (Table 4.2), a greater discrepancy is

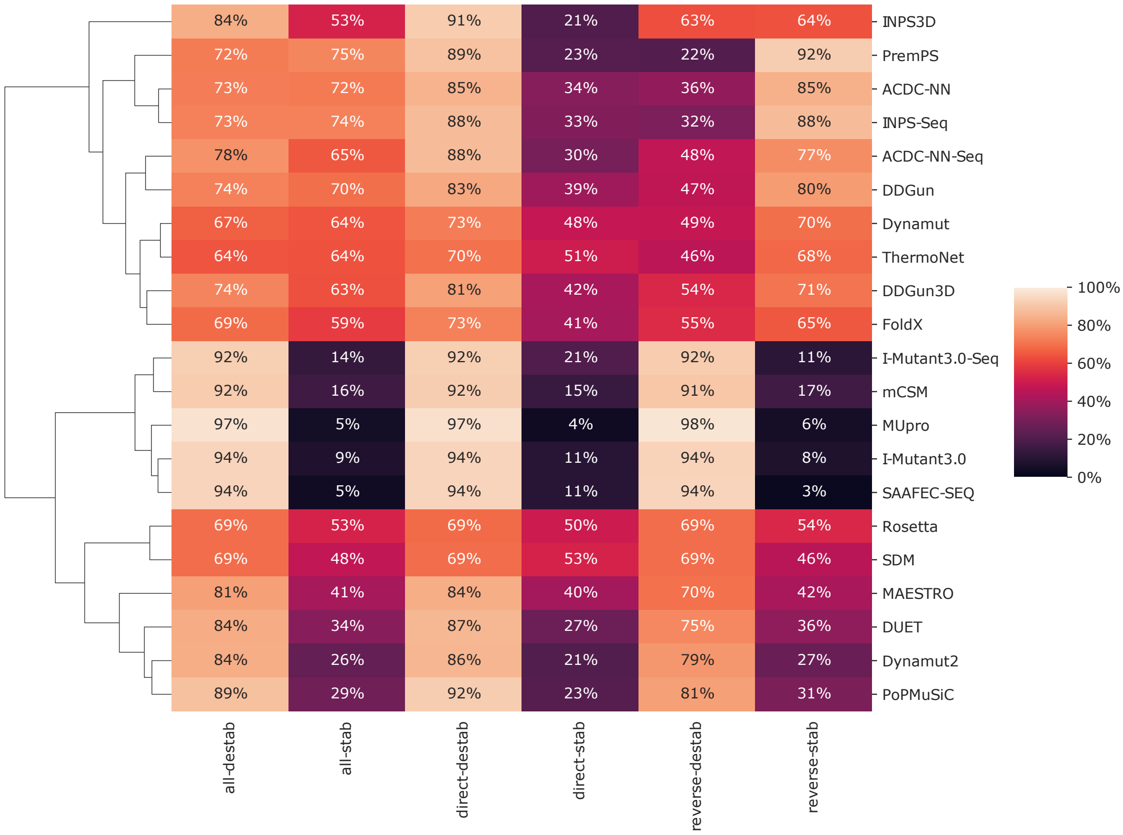

Figure 4.2: $\Delta\Delta G$ **classification accuracy**. Here we explore the classification accuracy when predicting the stability change direction. Predicted and experimental $\Delta\Delta G$ values were split in two classes: stabilizing ($\Delta\Delta G \geq 0$) and destabilizing ($\Delta\Delta G < 0$). For each subset (direct, reverse and both direct and reverse together) and for each experimental $\Delta\Delta G$ class (columns), the heatmap shows the ratio of variants predicted to be in the correct $\Delta\Delta G$ class for each predictor (rows).

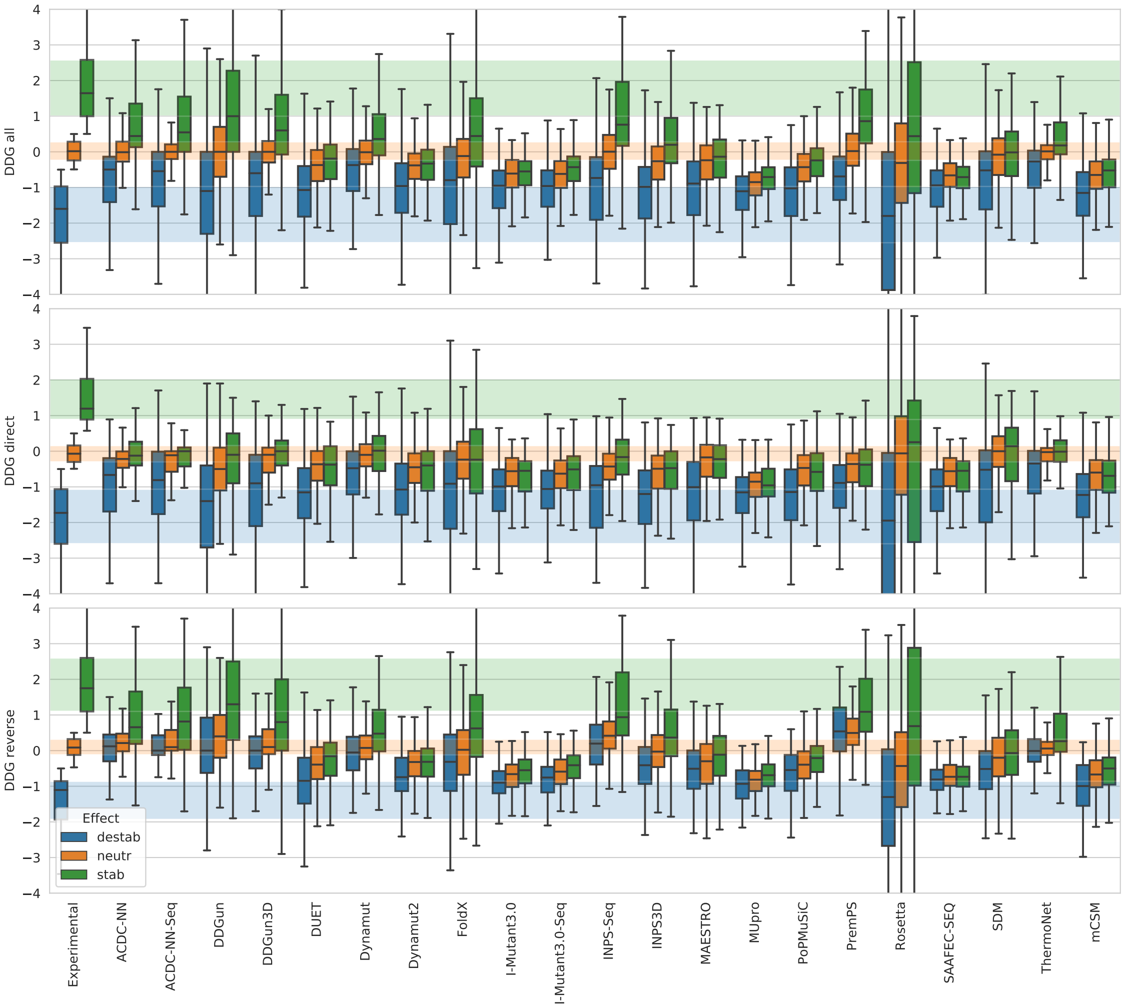

Figure 4.3: $\Delta\Delta G$ **prediction distribution by stability class**. These boxplots show the distribution of the predicted $\Delta\Delta G$ values. Variants were split into three classes by their experimental $\Delta\Delta G$: destabilizing ($\Delta\Delta G \leq 0.5$), neutral ($-0.5 < \Delta\Delta G \leq 0.5$), stabilizing ($0.5 < \Delta\Delta G$). The experimental $\Delta\Delta G$ values were plotted (to the left) and their boxes extended as transparent horizontal bands as a reference. The plot was repeated for all (top), direct (center) and reverse variants only (bottom).

	Pearson/RMSE				
Dataset	Total	Destabilizing	Neutral	Stabilizing	Non neutral
MAESTRO	0.50/1.44	0.42/1.46	-0.01/0.84	0.28/2.26	0.48/1.63
ACDC-NN	0.46/1.49	0.34/1.60	0.09/0.69	0.09/2.14	0.44/1.71
INPS3D	0.43/1.50	0.35/1.40	0.03/1.00	0.02/2.55	0.42/1.67
DDGun3D	0.43/1.60	0.32/1.69	0.13/0.94	0.13/2.22	0.41/1.80
INPS-Seq	0.43/1.52	0.26/1.56	0.10/0.92	0.13/2.25	0.42/1.70
ACDC-NN-Seq	0.42/1.53	0.28/1.64	0.08/0.76	0.07/2.18	0.40/1.75
PremPS	0.41/1.51	0.43/1.48	-0.02/0.84	-0.08/2.53	0.04/1.72
PopMusic	0.41/1.51	0.37/1.40	0.11/0.96	0.09/2.67	0.39/1.69
DUET	0.41/1.52	0.34/1.48	0.02/0.89	0.10/2.54	0.38/1.72
Dynamut	0.41/1.60	0.32/1.81	0.16/0.66	0.29/2.00	0.40/1.85
SDM	0.41/1.67	0.33/1.81	0.16/1.01	0.09/2.14	0.40/1.88
DDGun	0.40/1.75	0.25/1.75	0.12/1.29	0.11/2.46	0.39/1.90
SAAFEC-Seq	0.36/1.54	0.31/1.48	0.03/0.87	0.07/2.60	0.34/1.74
mCSM	0.36/1.54	0.30/1.42	-0.01/0.96	0.06/2.73	0.33/1.73
I-Mutant3.0	0.36/1.54	0.31/1.48	0.03/0.87	0.07/2.60	0.34/1.74
I-Mutant3.0-Seq	0.34/1.56	0.23/1.53	0.05/0.92	0.21/2.53	0.33/1.75
MuPro	0.25/1.61	0.19/1.45	0.08/1.08	-0.01/2.84	0.24/1.78
FoldX	0.21/2.32	0.20/2.25	0.01/2.28	0.17/2.66	0.24/2.33

Table 4.2: **Pearson's correlations and root mean square error (RMSE) between the experimental and estimated $\Delta\Delta G$s.** The $\Delta\Delta G$s are predicted by 18 state-of-the-art protein stability prediction tools on the S669 dataset. The correlations and RMSE are calculated on each class separately (”Destabilizing”-”Neutral”-”Stabilizing”), on the whole dataset (”Total”) and only on the destabilizing and stabilizing variants, excluding the neutral (”Non neutral”).

seen in non-antisymmetric methods between destabilizing and stabilizing variants, further accentuated without reverse variants. However also anti-symmetric methods show substantial imbalance in direct variants. Most methods align reasonably well with experimental $\Delta\Delta G$s trends, achieving $r \geq 0.4$, and predict destabilizing variants with a fair accuracy, evidenced by $r \geq 0.3$. However, stabilizing variants are poorly predicted by all methods, with the highest Pearson's correlation at ≤ 0.3 and the lowest RMSE at ≥ 2, highlighting a significant shortcoming in recognizing direct-stabilizing variants.

4.4 Discussion and Conclusion

The Pearson correlation of the methods tested on S669 is lower than those reported in the original papers. However, this is expected since S669 can be considered as external validation. Nonetheless, the Pearson correlations are not too far from those reported for the methods evaluated by limiting sequence identity between training and test sets. The more antisymmetric methods tend to perform better, and those built to be antisymmetric perform better in the regression task (prediction of value), in particular PremPS, ACDC-NN and INPS-Seq. It is also worth noticing that the methods perform equally well on NMR or X-ray structure and are relatively insensitive to pH and temperature outside the physiological conditions, making them useful also when these types of information are not available.

Overall, our assessment highlighted that the predictors satisfying the antisymmetry property can perform better than the other tools in regression or when the test set is balanced. For some of them, as in the case of ACDC-NN and DDGun, their sequence-based version showed similar results compared with their structure-based counterpart. This indicates that both evolutionary information and antisymmetry are important features for narrowing down the gap in the performance between sequence- and structure-based methods. Most methods, especially the non-antisymmetric, show a bias toward the destabilizing class. This makes them unsuitable for variant classification because they tend to predict every variant as destabilizing, misclassifying most stabilizing variants.

When only stabilization/destabilization information is considered, the antisymmetric methods tend to predict better on the whole datasets. However, it appears that the direct stabilizing variants in the datasets are more challenging to assign correctly. In particular, SDM, Rosetta, ThermoNet and Dynamut are the most balanced. All the tested methods tend to compress the predictions toward neutrality, generating a significant overlap between stabilizing, neutral and destabilizing variations. The compression of the predictions indicates that a possible improvement for future methods is to work on the calibration of the prediction distributions. The destabilizing variants show a stronger signal in terms of $\Delta\Delta G$ which are easier to detect on average. Indeed, the antisymmetric predictors showed to capture very well the reverse variations, as stabilizing. These contrasting results may open a future direction of study, improving our understanding of these types of variants and possibly increasing the method performance.

Chapter 5

Mutational Signatures: fingerprints of mutational processes in cancer

5.1 Introduction

Thanks to the advent of next-generation technologies and their rapid diffusion due to the reduction of costs, sequencing the human genome to identify possible deleterious DNA variations, has become a standard procedure not only in research laboratories, but also in many hospitals where precision medicine is applied. Indeed, WGS and WES techniques combined with the use of open source tools have allowed us to understand the underlying mechanisms of many diseases and to guide drug therapies on the basis of mutations involved in the patient's genome as reported in [67], [68],[69] [70] and other studies. Cancer is linked to the accumulation of somatic mutations, caused by a wide range of endogenous (i.e genome instability or deficiency in a DNA repair mechanism) or exogenous (environmental exposure such as UV light or tobacco smoking) mutagenic agents, which stratify over time. Genome sequencing captures snapshots that reflect the accumulation of various mutations over time, resulting from diverse biological processes. Therefore, it has been hypothesized [15] that a genome mutational pattern can be deconvoluted considering different generative processes that shape the mutational landascape, which are called *mutational signatures.* From a mathematical point of view a mutational signature is represented by a frequency vector over mutational classes. Somatic mutations are categorized into various types, including single base substitutions (SBS), doublet-base substitutions (DBS), and structural variants like insertions and deletions (IDs) and others. For SBS, six distinct classes can be identified ($C > A$, $C > G$, $C > T$, $T > A$, $T > C$, $T > G$) based on the mutated pyrimidine base (C or T). The use of Single Base Substitutions (SBS) as an example is primarily because they represent the most frequent mutations and are extensively employed in analyzing mutational signatures. The classification becomes more granular depending on the sequence context. In some studies, the two adjacent bases, $5'$ and $3'$ to the mutation, are analyzed, leading to 1,536 classes. In others, only one adjacent base is considered, resulting in 96 classes. This classification approach can be also applied to others mutational events. For a better understanding, in figure 5.1 we show how mutational classes for SBS mutations are composed and an example of three related mutational signatures.

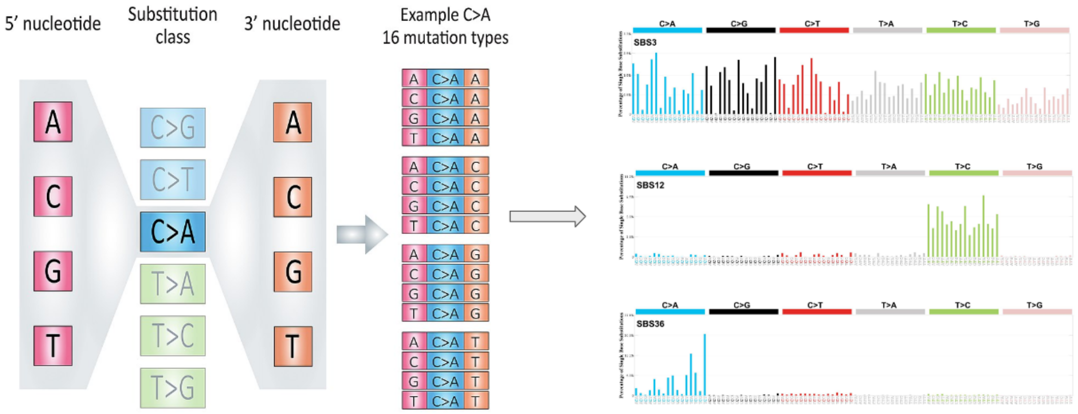

Figure 5.1: **SBS Mutational Signatures** On the left the figure shows how the 96 mutational classes were built. On the right side an example of 3 mutational signatures profile

5.2 Mathematical formulation

The notation generally used in the field of mutational signatures consists in defining a set of mutational catalogues as a matrix of counts $M \in R^{K \times G}$, where G is the number of available genomes into the catalogues and K represents the number of possible mutational classes based on the particular somatic mutation considered (i.e 96 for the three nucleotide context SBS as described above). The first approach to decompose the matrix M was proposed by Alexandrov et al. [15] in his pioneering work and it consisted in using non negative matrix factorization (NNMF) to obtain a signatures matrix $P \in R^{K \times N}$ of N (a priori chosen) mutational signatures generated by N generative processes, shaping the exposure matrix $E \in R^{N \times G}$.

From a mathematical point of view, a genome mutational profile g can be represented as a vector of mutation $\mathbf{m}_g = (m_{g1}, \ldots, m_{gK})^T$ over the K mutational classes. Consequently, a mutational signature, generated by the i-th process, can be represented as a probability vector $\mathbf{p}_i = (p_{i1}, \ldots, p_{iK})^T$, where each p_{ik} is the probability that the mutational process i (with $i = 1, \ldots, N$) will induce a mutation of type k over K classes. Being a probability distribution, a signature is subject to the normalization condition $\sum_{k=1}^{K} p_{ik} = 1$ with $0 \leq p_{ik} \leq 1$.

The intensity of exposure to the i-th mutational process for a genome g, is then represented by a coefficient e_{gi}.

Hence each element of the genome mutational count m_{gk} can be represented as:

$$m_{gk} = \sum_{i=1}^{N} e_{gi} p_{ik} \tag{5.1}$$

Usually, when performing the extraction of mutational signatures, one has to deal with multiple genomes G and therefore it is convenient to use matrix notation. The collection of G catalogues is generally represented by a $K \times G$ matrix:

$$M = \begin{pmatrix} m_1^1 & m_2^1 & \cdots & m_G^1 \\ \vdots & \vdots & & \vdots \\ m_1^K & m_2^K & \cdots & m_G^K \end{pmatrix} \tag{5.2}$$

The extracted signatures are then represented by a $K \times N$ matrix (with the signatures in columns):

$$P = \begin{pmatrix} p_1^1 & p_2^1 & \cdots & p_N^1 \\ \vdots & \vdots & & \vdots \\ p_1^K & p_2^K & \cdots & p_N^K \end{pmatrix} \tag{5.3}$$

And the exposure matrix by an $N \times G$ matrix:

$$E = \begin{pmatrix} e_1^1 & e_2^1 & \cdots & e_G^1 \\ \vdots & \vdots & & \vdots \\ e_1^N & e_2^N & \cdots & e_G^N \end{pmatrix} \tag{5.4}$$

thus leading to the equation 5.1 in matrix notation:

$$M \sim P \cdot E \tag{5.5}$$

The formulation described above represents the original deconvolution approach proposed by Alexandrov in the pioneering work on mutational signatures. The key assumption in this data modeling approach is the independence of mutational processes, which is reflected in their additive nature. This leads to the use of non-negative matrix factorization, an intrinsically linear method. Following Alexandrov's work, various methods have been developed, mainly based on non-negative matrix factorization and its probabilistic formulation [71, 72, 73]. Various authors have tried to include the tumor-specific opportunity for different mutation types, trying to model different genomic regions where the mutation opportunity varies across samples [74, 75]. More recently a new method [76] based on tensor matrix factorization incorporates some important genomic properties, such as the transcriptional orientation, the replication direction and epigenetic states that can influence mutation rates and spectra, thus leading to a more complete description of mutational landscape. However, all aforementioned methods rely on the assumption of additive processes. This could be represent an oversimplification and may not fully capture the complex biological mechanisms involved in cancer development; however these methods nonetheless provide a highly interpretable tools. In addition, these methods has largely been in agreement with experimental data, despite there still being many aspects that need further investigation.

5.3 De Novo extraction and refitting

The research community has focused on addressing two major types of computational challenges associated with mutational signatures: *de novo extraction* and *refitting*. The first aims to infer the mutational signatures and calculate their frequencies

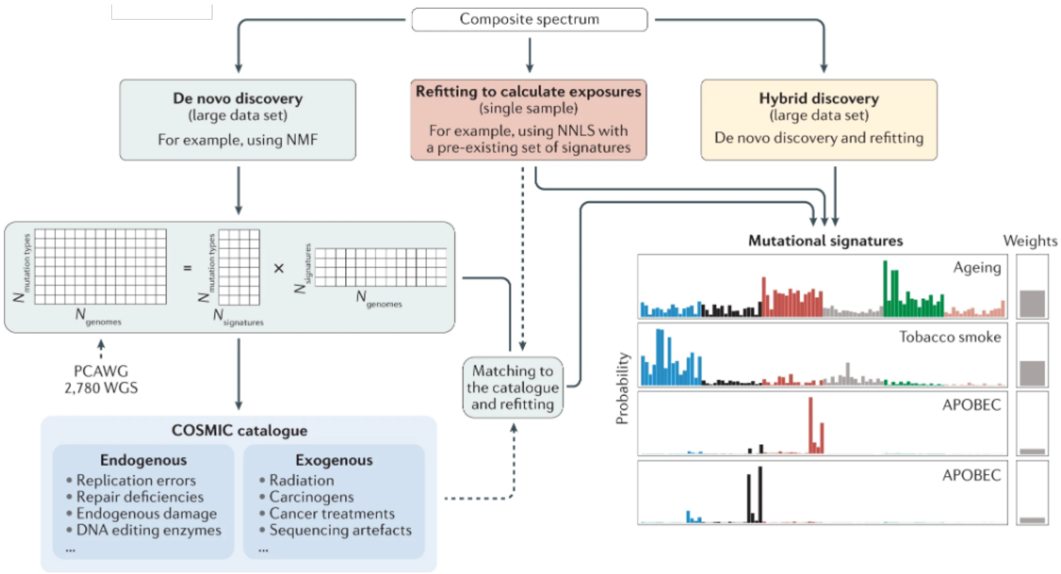

Figure 5.2: A summary workflow of de novo extraction and refit methods for mutational signatures extraction. Figure taken from Cortés-Ciriano et al., [84]

within the cohort. This challenge was the primary focus of the pioneering studies in this field, and efforts to resolve it are ongoing. Among the most widely used tool we mentioned SigProfilerExtractor [77], MutationalPatterns[78], CancerSign[79], Mutspec[80] and many others. The second involves estimating the occurrences of a pre-defined set of signatures in a group of individuals, utilizing techniques such as quadratic programming, non-negative least square, linear combination decomposition and simulated annealing [78, 81, 82, 83]. Both de novo extraction and refitting procedures are summarized in Figure 5.2

This distinction between de novo extraction and refitting methods arose essentially for two main reasons, the first is mathematical. In fact since the solutions of the NMF decomposition are not unique, to increase robustness during the de novo phase, most methods perform the extraction M times for a fixed number N of mutational signatures, thus obtaining M repetitions for each signature.

Using the matrix notation, the whole set of extracted signatures matrix can be represented by $\{\mathbf{P}^1 ... \mathbf{P}^M\}$. Consequently a clustering analysis is performed, thus obtaining a consensus signature matrix $\hat{P}$. However the final set of exposures $\hat{E}$ cannot be simply obtain by taking the average of the exposures because:

$$\frac{1}{M}\sum_{m=1}^{M}\mathbf{P}^{(m)}\mathbf{E}^{(m)} \neq \left(\frac{1}{M}\sum_{m=1}^{M}\mathbf{P}^{(m)}\right)\left(\frac{1}{M}\sum_{m=1}^{M}\mathbf{E}^{(m)}\right) \tag{5.6}$$

Therefore, computational tools usualy re-attribute (refit) exposures to the entire dataset by fixing the P matrix to the consensus set of signatures and only calculate E.

The second reason of the arising of refitting tools is that with the increasing amount of available data, many signatures have been extracted and experimentally validated. Thus, instead of performing de novo extraction, especially in small cohorts, the reference set of signatures used is the one of COSMIC which is attributed to the

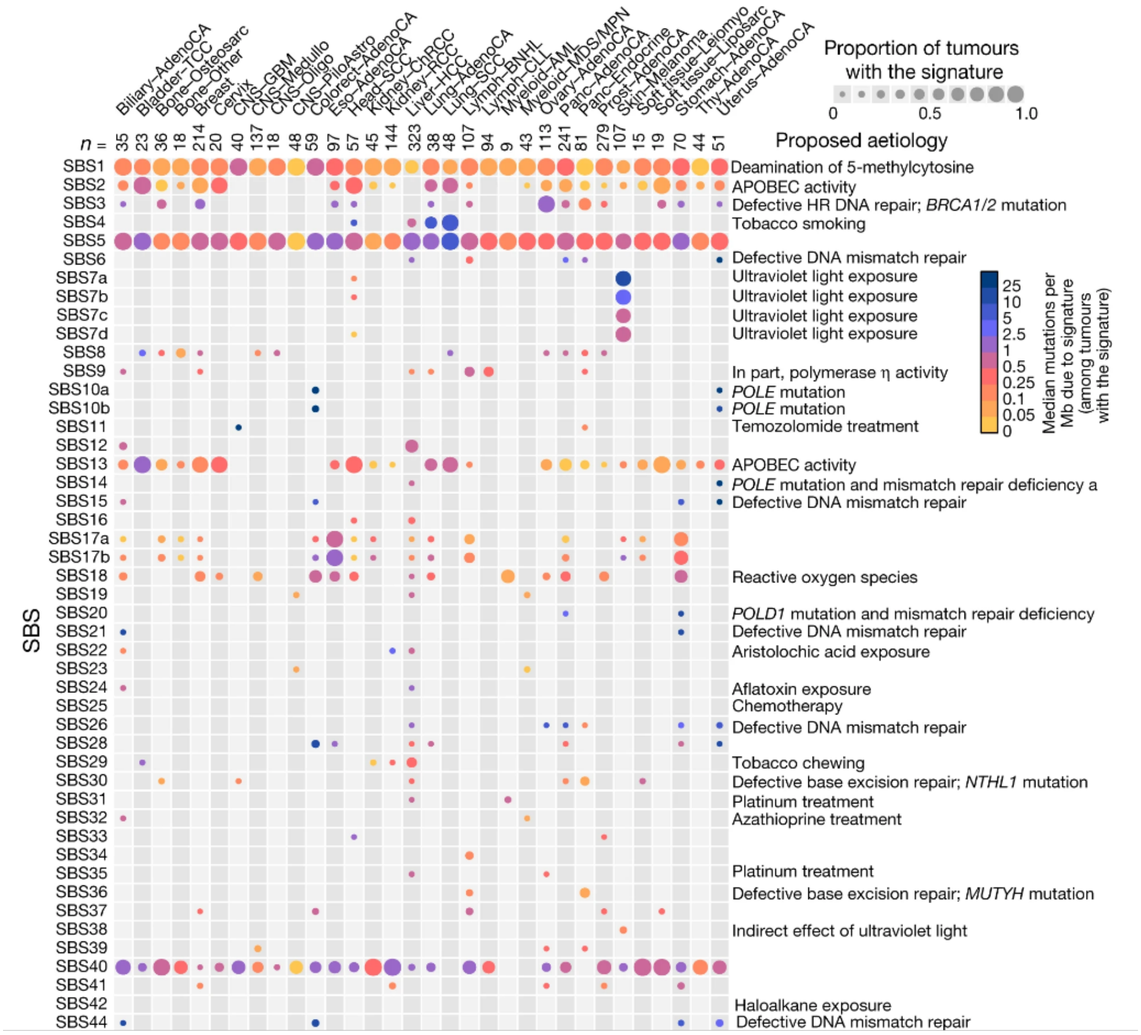

Figure 5.3: An Illustrative example of exposures of mutational signatures on PCAWG dataset. This figure is taken from Alexandrov et. al [86]

samples. In particular, the latest version (3.4) of the Catalogue Of Somatic Mutations In Cancer (COSMIC) [85] hosts 86 single-base substitution (SBS) signatures extracted from the Pan Cancer Analysis of Whole Genomes (PCAWG) and other WGS samples, using SigProfilerExtractor [77].

In summary, the overall objective of these methodologies is to extract the mutational signatures active in cancer genomes, to understand which mutagenic processes have operated, thereby enhancing our understanding of its origins and aiding in the development of targeted treatments. Figure 5.3 illustrates a map of the assignments for each signature within the PCAWG cohort.

5.4 Caveats and open problems

Despite the undeniable successes of mutational signature analysis, there are several open problems, and challenges that remain to address. For example among the 67 non artefactual signatures (signatures associated to contamination of reagents or error during the sequencing analysis) in COSMIC, 24 profiles neither have a direct association with an experimentally validated mutagenic process nor are they supported by statistical association with a specific aetiology. This observation, combined with the high similarity among many signatures in the catalog, suggests that unexplained signals might actually be mathematical artifacts arising from the employed methodology and thus representing a possible overfitting. Lal et al. [72] pointed out that state-of-the-art NMF based methods aim to minimise the residual error after fitting the data with the discovered signatures to fit the data perfectly, which may generate overfitting issues by including stochastic noise in the data as part of the signatures, or multiple similar signatures for the same underlying process.

These issues become more critical when the signature extraction is highly dependent on the number of samples available, which complicates the correct identification of the true components and the stability of the results. In addition, the studies of Maura et al. [87] highlighted that the presence of flat signatures, showing similar frequencies across all the 96 mutational classes, could represent a source of background noise and collinearities, making the de novo signature extraction task difficult and ambiguous. Particularly, weaker signatures that occur less frequently are prone to being misidentified or confused with other signatures. Signatures that display prominent characteristics, like pronounced peaks in specific trinucleotide sequence contexts, are more easily identified as distinct [17].

Besides that, traditional methods for inferring mutational signatures, which are assumed to be additive, are increasingly being challenged by the understanding that mutagenic processes are complex and interactive [88, 89]. The cancer genome's mutational landscape is shaped by a combination of factors: the type of DNA damage, the vulnerability of specific sites to damage, and the effectiveness of repair mechanisms. DNA repair processes, in particular, modify the effects of other mutagens, suggesting that different combinations of DNA damage and repair deficiencies could result in numerous, distinct signatures, complicating their interpretability. Recent studies have shown that some of these are composite signatures, resulting from different types of DNA damage combined with MMR deficiency. This finding raises the possibility that many signatures, especially those associated with DNA repair deficiencies, might be similarly composite, leading to questions about decomposing complex signatures into their individual contributing factors.

5.5 Discussion

Over the past decade, significant progress has been made in understanding mutational signatures in cancer genomes, shedding light on the complex relationship between signatures and mutagenic processes. A key insight is the connection between specific mutational patterns and the malfunction of DNA repair mechanisms. For instance, mutations in key repair genes like BRCA1 and BRCA2 result in distinct mutational signatures, exemplified by SBS signature 3 in cancer genomes [15, 90, 91]. This pattern is indicative of disrupted homologous recombination repair pathways. Similarly, the APOBEC family's role in mutating cytosine bases in DNA is linked to distinct mutational signatures, reflecting broader immune response activities in cancer [92, 18]. On the other hand some signatures are the results of concomitant effects of processes, as in the case of SBS signatures 14 and 20, that arise from a concurrent loss of proofreading (POLE or POLD1) and mismatch repair (MMR) [88]

However as highlighted in section 5.4 there are several open challenges which need further analysis and investigation.

During my doctoral research, we contributed to the field of mutational signatures in two key aspects, which I will detail in the following two chapters:

- Firstly, we conducted a comprehensive analysis of extraction scenarios to understand the situations in which mutational signatures are more challenging to identify. Additionally, through archetypal analysis, we observed that many of COSMIC signatures are well-represented by a subset of archetypal signatures with biological significance. This partially explains why it can be difficult to extract and discern the true signal in certain contexts.
- Secondly, we developed a framework based on autoencoders for the de novo extraction of mutational signatures. This method is the first explainable autoencoder capable of extracting mutational signatures. It can be seen as a foundational architecture, offering extreme flexibility. This flexibility allows for the potential development of more complex architectures that could naturally integrate other biological data.

Chapter 6

Archetypal analysis reveals the inherent instability of mutational signatures

6.1 Introduction

In the previous chapter, we have delved into the numerous ongoing challenges in extracting mutational signatures. Our paper entitled *Unravelling the instability of mutational signatures via Archetypal analysis* [19] seeks to quantitatively explain the difficulties in accurately identifying the true number of these signatures in certain contexts. We have conducted an in-depth analysis, which we will discuss here, focusing on the COSMIC version 3.3 signatures. This version lists 79 mutational signatures, but we have identified that 19 of them are likely artifacts from experimental methods or reagent contaminations. So, our analysis concentrated on the remaining 60.

Our study both provides a systematic assessment of the de novo extraction task through simulation scenarios based on the latest version of the COSMIC signatures and highlights, through a novel approach using archetypal analysis, which COSMIC signatures are redundant and more likely to be considered as mathematical artefacts. 29 archetypes were able to reconstruct the profile of all the COSMIC signatures with cosine similarity > 0.8. Interestingly, these archetypes tend to group similar original signatures sharing either the same aetiology or similar biological processes. In particular the proposed study have focused on two main goals:

- to provide a systematic approach to assess to which extent the extraction of the newest version of COSMIC signatures can be affected by the high similarity among the signatures in the same catalogue, the presence of flat signatures and the number of available samples
- to provide a compact representation of the current catalogue by prioritising the identification of those profiles representing extreme patterns in the data so that all the observations can be reproduced as mixtures of their extremes. To this aim, Archetypal Analysis [93] was applied to represent how the information from COSMIC can be projected into a reduced number of dimensions and to explain potential instability issues in specific extraction scenarios.

We believe that these findings will be useful to encourage the development of new de novo extraction methods avoiding the redundancy of information among the signatures while preserving the biological interpretation.

6.2 Materials and methods

6.2.1 Similarity of COSMIC signatures

Analyses were performed on COSMIC catalogue v3.3 considering the 79 SBS mutational signatures profiles on the reference genome GRCh37 identified by SigProfilerExtractor [77]. Among these, we removed 19 signatures classified by the catalogue as sequencing artefacts. Of the remaining 60 signatures considered, 19 neither have a direct association with an experimentally validated mutagenic process nor are they supported by statistical association with a specific process.

We first quantified the pairwise level of similarity in the signature catalogue. Therefore, for each pair of signatures $\mathbf{s_1}$ and $\mathbf{s_2}$, we calculated the cosine distance:

$$cos(\mathbf{s_1}, \mathbf{s_2}) = \frac{\mathbf{s_1} \cdot \mathbf{s_2}}{\|\mathbf{s_1}\| \|\mathbf{s_2}\|} \tag{6.1}$$

obtaining a pairwise cosine distance matrix $D \in \mathbb{R}^{d \times d}$, where $d = 60$ is the number of signatures considered. We then built a cluster map to simulate different *de novo* extraction scenarios by applying to the D distance matrix a Hierarchical Clustering [94] with average linkage.

6.2.2 Flatness of COSMIC signatures

Signatures SBS3, SBS5, SBS40 and SBS8 are often referred to as flat signatures given their relative featureless profile, almost uniformly distributed across the 96 mutational classes. However, to the best of our knowledge, no quantitative definition of flatness has ever been provided. To fill this gap we formulate a simple definition of signature flatness by calculating the cosine similarity between the signature and the uniform distribution. Therefore the flatness of a signature $\mathbf{s_1}$ can be defined as:

$$flatness(\mathbf{s_1}) = cos(\mathbf{s_1}, \mathbf{s_u}) \tag{6.2}$$

where $\mathbf{s_u}$, in the case of SBS mutational signatures, consists of a signal uniformly distributed over the 96 mutational classes. Hence, the flatness is a score ranging from 0 to 1, where 1 represents a perfectly flat profile. Since the presence of flat signatures in a catalogue can complicate the extraction task, a quantitative definition of flatness can be useful to build robust *de novo* extraction methods and to test their capabilities to correctly extract multiple signatures with different levels of flatness. In this regard, we constructed scenarios with different levels of similarity and flatness.

6.2.3 *De novo* extraction scenarios

To assess both the reliability and the feasibility of the *de novo* extraction procedure, several synthetic catalogues were generated considering COSMIC mutational signatures as underlying generative processes. The SigsPack R package was used to generate 10 synthetic mutational catalogues for each extraction scenario to take into account statistical fluctuations [95]. Different ranges between a minimum of 200 and a maximum of 10,000 samples were set according to the chosen scenario and the number of signatures involved. The number of mutations in each tumour sample was set to 5,000 for each scenario, based on the PCAWG median number of sample mutations. All the simulated scenarios are summarised in Table 6.1 and the generated catalogues are available at the Github repository `https://github.com/compbiomed-unito/archetypal-analysis-cosmic`.

De novo extraction analysis was applied to each scenario using the gold-standard approach SigProfilerExtractor, with the aim of evaluating the extraction performance from catalogues with groups of similar latent signatures, varying in number and flatness score.

Scenario	N° Signatures	Median Similarity	Median Flatness	N° of Samples
1	6	0.73	0.76	200:500
2	5	0.83	0.34	200:500
3	11	0.50	0.64	200:10000
4	11	0.22	0.44	200:5000
5	20	0.22	0.45	1000:10000

Table 6.1:
Summary of the *de novo* extraction scenario. For each simulated scenario, the number of active signatures, the cosine similarity level. the flatness level and the n° of samples.

6.2.4 Evaluation metrics

Four metrics were considered for the performance evaluations:

- Frequency (F) of simulation runs where all the signatures are correctly identified:

$$F = \frac{\text{N° of successful runs}}{\text{N° of total runs}} \tag{6.3}$$

- Mean square error (MSE) between simulated and reconstructed catalogues:

$$MSE = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{i,j} - \hat{x}_{i,j})^2}{n \cdot m} \tag{6.4}$$

  where $x_{i,j}$ and $\hat{x}_{i,j}$ are the matrix elements of the original $X \in \mathbb{R}^{n\times m}$ and the reconstructed $\hat{X} \in \mathbb{R}^{n\times m}$ catalogues, respectively.

- Average stability measured by mean silhouette coefficient score of the signature clusters generated by SigProfilerExtractor:

$$C_{mean} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{N} c_{ik}}{K \cdot N} = \frac{\sum_{k=1}^{K} C_k}{K} \tag{6.5}$$

  c_{ik} is the silhouette coefficient of the $i-th$ sample which belongs to cluster k. N is the number of NMF runs performed by SigProfilerExtractor, K is the number of cluster labels, and C_k is the mean silhouette score of the $k-th$ cluster

c_{ik} is given by:

$$c_{ik} = \frac{b_{ik} - a_{ik}}{max(a_{ik}, b_{ik})} \tag{6.6}$$

where a_{ik} is the mean intra-cluster distance and b_{ik} is the mean nearest-cluster distance of the i-th sample which belongs to cluster k.

- Minimum stability represented by the minimum silhouette coefficient score of the signature clusters generated by SigProfilerExtractor:

$$C_{min} = min\{C_k\}, \quad with \quad k = 1, ..., K \tag{6.7}$$

6.2.5 Archetypal Analysis

Archetypal analysis (AA) is an unsupervised learning method that aims to represent data points as sparse convex combinations of extreme elements of a given dataset. More formally, let $X \in \mathbb{R}^{p \times q}$ be a matrix whose row vectors are $\mathbf{x_i} \in \mathbb{R}^q$ and let $Z \in \mathbb{R}^{r \times q}$ be another matrix, whose column vectors $\mathbf{z_k} \in \mathbf{R}^q$ represent the archetypes. AA reconstructs X through a linear combination of archetypes $\mathbf{z_k}$, which are themselves convex combinations of the X rows $\mathbf{x_i}$. Therefore, AA solves the following constrained equation:

$$minimize \quad \sum_{i=1}^{p} \|\mathbf{x_i} - \sum_{k=1}^{r} \alpha_{ik} \mathbf{z_k}\|^2 \quad with \quad \mathbf{z_k} = \sum_{j=1}^{q} \beta_{kj} \mathbf{x_j} \tag{6.8}$$

where $\alpha_{ik} \geq 0$; $\sum_{k=1}^{r} \alpha_{ik} = 1$ and $\beta_{kj} \geq 0$; $\sum_{j=1}^{q} \beta_{kj} = 1$ ensure that the archetypes fall on the convex hull of X.

In this study, AA was applied directly to the COSMIC signatures matrix $M \in \mathbb{R}^{60 \times 96}$ where, as reported above, 60 is the number of COSMIC SBS mutational signatures and 96 are the mutation contexts. Since AA is a matrix decomposition technique, the number of basis elements has to be chosen, which in our case corresponds to the number of archetypes. This number was set in order to correspond to the 95% of the explained variance. AA was performed using the Python-based *Archetypal Analysis Package* freely available at `https://data.csiro.au/collection/csiro:40600v1` [96].

6.3 Results

6.3.1 COSMIC cluster map

The cluster map on COSMIC v3.3 catalogue revealed that there are several groups of signatures showing pairwise cosine similarity > 0.8 (Fig 6.1). The first group in the top-left corner consists of 6 signatures and it includes those signatures which are commonly referred to as *flat*, i.e. presenting a relatively featureless profile distributed over all the 96 mutational classes, as the well-known SBS3 experimentally associated with defective homologous recombination-based DNA damage repair and the clock-like SBS5, statistically associated with age. The median pairwise similarity is 0.73, with a maximum equal to 0.88 for the pairs SBS3-SBS40 and SBS5-SBS92, while the median flatness, calculated through 6.2, is 0.76. These 6 signatures were used to build up the first extraction scenario presented in Table 6.1.

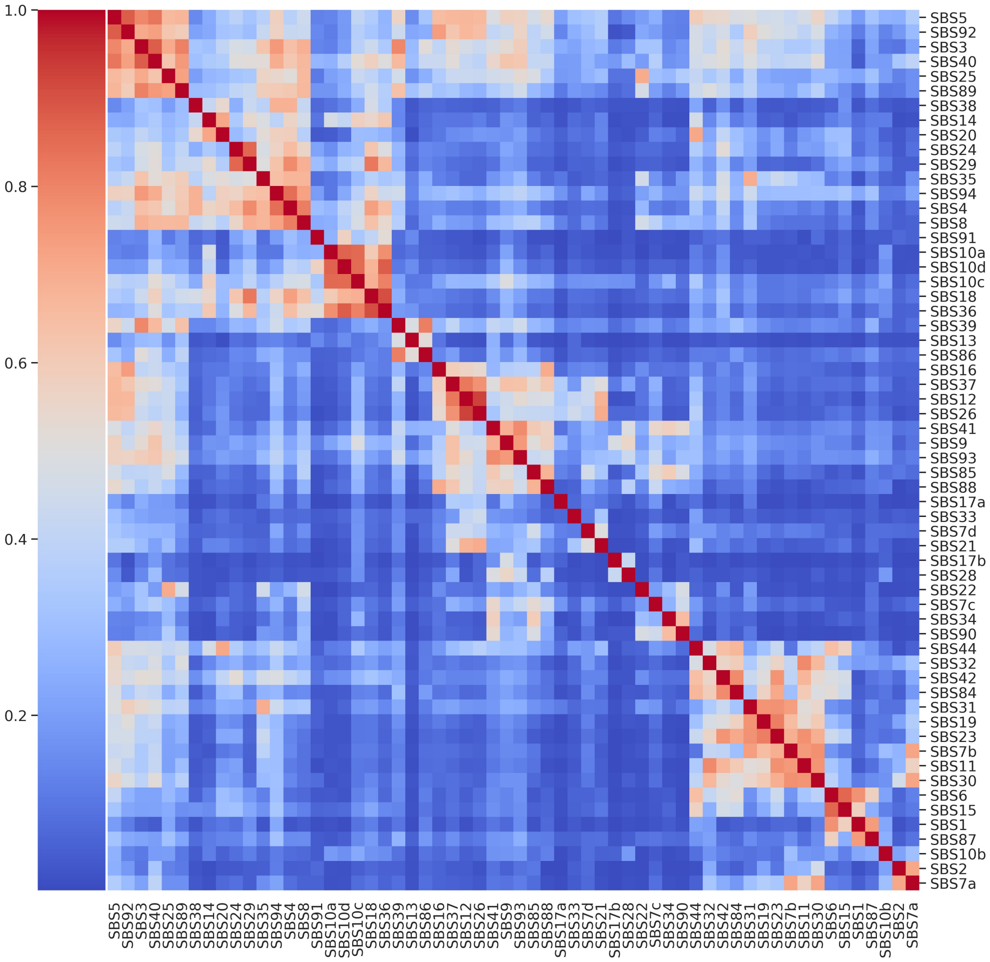

Figure 6.1: **Cluster map of COSMIC SBS Mutational Signatures.** Pairwise cosine similarity displayed for the 60 SBS signatures from COSMIC catalogue.

A second notable group is characterised by a high pairwise cosine similarity among signatures but with a low level of flatness and it includes SBS36, SBS18 and the three signatures SBS10a, SBS10c and SBS10d associated with an altered activity of polymerase (polymerase epsilon exonuclease domain mutations and defective POLD1 proofreading), which were considered for the second extraction scenario. The median pairwise similarity is 0.83 with a maximum equal to 0.91 between SBS36 and SBS18 while the median flatness is 0.34. In the third extraction scenario the synthetic catalogues were generated from the signatures used in the first and the second scenario together (11 signatures). Finally, in the fourth and fifth extraction scenarios, 11 and 20 signatures with a low flatness score were considered, respectively, where each signature has at least another similar one (cosine similarity > 0.8). The complete list of signatures used in each scenario and the list of those signatures with a cosine similarity > 0.8 were reported in Tables B1 and B2 (Appendix B), respectively. In Fig B1 we reported the pairwise similarity distribution for each scenario compared with the full set of non-artefactual COSMIC signatures. The first and second scenarios have a very high similarity level as they were built by taking the two largest clusters of the similarity-based clustermap. The others are gradually less similar since the considered number of signatures increases but the number of similar signatures in each cluster decreases.

6.3.2 Flatness analysis

To overcome the qualitative description of flatness, in Equation 6.2 we defined a simple way to quantitatively assess the flatness of the signatures, being in line with the qualitative description. Indeed, as shown in Table B3, the known flat signatures SBS3, SBS40 and SBS5 show the highest degree of flatness, but a similar level to SBS5 can be found for SBS25 and SBS89. In addition, this definition of flatness appears to be well distributed within COSMIC from a minimum of 0.15 (SBS1) to a maximum 0.87 (SBS3), showing that this metric can emphasise the differences in shape between the various signatures in COSMIC, as shown in Fig 6.2. As mentioned in the previous section, the extraction scenarios built to highlight possible issues in the *de novo* extraction process differ in the number of signatures involved, the pairwise similarity between profiles, and the level of flatness. In Fig B2 (Appendix B), the flatness distribution for each scenario is shown.

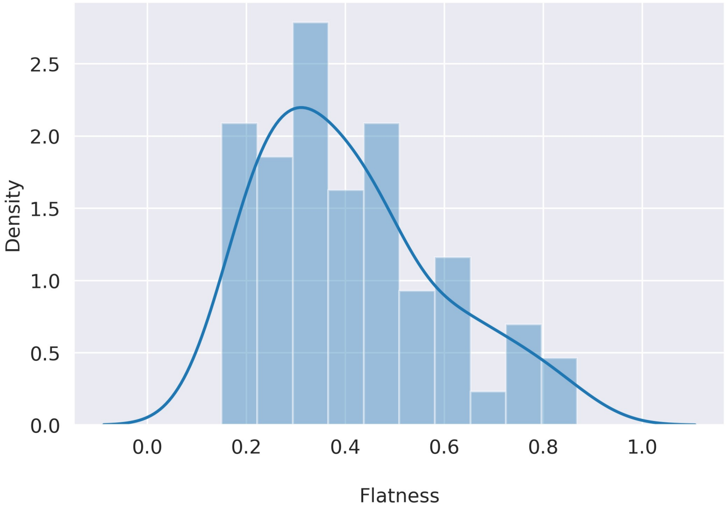

Figure 6.2: **Distribution of the COSMIC flatness.** On the x axis the flatness defined in 6.2, on the y axis the density for each flatness level.

6.3.3 *De novo* signature extraction

The SigProfilerExtractor performance for each scenario is shown in Table 6.2. MSE, C_{mean} and C_{min} are reported as their corresponding median values across 10 repetitions, together with their inter-quartile range. When considered separately, signatures involved in scenarios 1 and 2 were almost always correctly extracted at 200 samples (F=0.9), regardless of the high level of similarity in each group.

However, when the extraction was performed by combining these two types (scenario 3), SigProfilerExtractor was never able to identify the correct number of signatures up to a high number of samples (5,000) and only at 10,000 samples it succeeded 80% of the times (F=0.8). In this scenario, it is worth noticing that, as the number of available samples increases, while the MSE decreases and the average stability decreases but remaining relatively high, the minimum stability decreases considerably reaching a very low level. As expected, by further increasing the number of samples, both the mean and minimum stability rise again. However, given that obtaining 10,000 tumour samples is often unfeasible in practice, this scenario highlights well a limitation of the NMF-based extraction process. Indeed, this scenario is particularly complex since it considers of 2 main subgroups, highly similar internally (median pairwise cosine similarity 0.73 and 0.83, respectively) but one at high and the other one at low flatness score (0.76 and 0.34, respectively, as shown in Table 6.1). Therefore, this difference in the flatness levels makes the extraction process much more difficult if there is not a very large number of samples. As shown in Fig B3, the algorithm starts to differentiate similar signatures inside each of these two groups at 3,000 samples, still failing at differentiating them well even at 5,000 samples.

On the other hand, considering again 11 signatures but with a lower level of similarity and flatness (scenario 4), the algorithm required at least 1,000 samples to identify the signatures with F=0.9 (Fig S3). Finally, when 20 signatures were considered (scenario 5), the algorithm always failed even at 5,000 samples, and it only succeeded 10% of the times at 10,000 samples. It is worth highlighting that the maximum number considered is significantly higher compared to the 2,780 genomes from PCAWG used to built the gold-standard mutational signatures catalogue available in COSMIC.

Scenario	**Samples**	**F**	**MSE**	$\mathbf{C}_{mean}$	$\mathbf{C}_{min}$
1	200	0.9	45.59 (45.45, 46.34)	0.88 (0.86-0.89)	0.78 (0.74-0.82)
	500	1	46.64 (46.41, 46.89)	0.93 (0.9-0.94)	0.86 (0.75-0.88)
2	200	0.9	27.75 (27.62, 28.05)	0.84 (0.83-0.86)	0.67 (0.54-0.69)
	500	1	28.11 (27.85, 28.45)	0.87 (0.86-0.89)	0.69 0.65-0.72
3	200	0.0	124.36 (120.66, 127.48)	0.98 (0.97-0.98)	0.96 (0.94-0.97)
	500	0.0	129.08 (126.64, 130.17)	0.99 (0.99-0.99)	0.98 (0.97-0.99)
	1000	0.0	90.49 (41.68, 127.92)	0.90 (0.81-0.99)	0.80 (0.58-0.98)
	3000	0.0	41.05 (40.89, 41.28)	0.81 (0.8-0.82)	0.47 (0.41-0.5)
	5000	0.0	40.18 (38.99, 41.38)	0.81 (0.80-0.82)	0.35 (0.26-0.49)
	10000	0.8	35.58 (35.55, 35.72)	0.81 (0.80-0.81)	0.36 (0.33-0.47)
4	200	0.0	101.93 (85.45, 103.0)	0.88 (0.84, 0.9)	0.76 (0.68-0.81)
	500	0.1	40.38 (38.24, 41.01)	0.82 (0.82-0.83)	0.38 (0.3-0.45)
	1000	0.9	33.20 (32.81, 33.42)	0.83 (0.81-0.85)	0.52 (0.47-0.58)
	3000	1	32.58 (32.56, 32.69)	0.86 (0.85-0.87)	0.58 (0.53-0.62)
	5000	1	32.54 (32.36, 32.69)	0.89 (0.87-0.89)	0.67 (0.6-0.7)
5	1000	0.0	68.98 (67.01, 70.04)	0.83 (0.81-0.86)	0.49 (0.33-0.59)
	3000	0.0	37.42 (37.16, 37.68)	0.81 (0.81-0.82)	0.33 (0.26-0.39)
	5000	0.0	37.01 (36.94, 37.12)	0.81 (0.80-0.82)	0.34 (0.29-0.39)
	10000	0.1	36.63 (36.44, 36.77)	0.81 (0.81-0.82)	0.38 (0.32-0.4)

Table 6.2: ***De novo* signature extraction performance** For each simulated scenario, the frequency of runs with all the signatures correctly identified (F), the mean square error (MSE) between simulated and reconstructed catalogues, the average C_{mean} and minimum C_{min} stability scores of signature clusters are displayed.

6.3.4 Archetypal Analysis

The application of AA to the COSMIC SBS mutational signatures matrix $M \in \mathbb{R}^{60 \times 96}$, revealed that 29 archetypes $\mathbf{z_k} \in \mathbb{R}^{96}$ $(k = 1, .., 29)$ were able to explain the 95% of the variance (Fig S4) and that through a combination of them it is possible to reconstruct each COSMIC signature profile with cosine similarity >0.8 (Fig B5).

The archetypal profiles are summarised in Figure 6.3. Most of the archetypal profiles coincide almost perfectly with COSMIC signatures. Specifically, 26 out of 29 archetypes correspond to at least one COSMIC signature with cosine similarity of at least 0.97 (Fig B6). These results suggest that a subset of signature profiles represent extreme patterns of the catalogue and that a combination of them are capable of reconstructing the entire catalogue with a high level of accuracy. The relationship between signatures and archetypes can be better understood considering the α coefficients of the equation 6.8. In particular, the coefficients a_{ik} represent the weights that each archetype $\mathbf{z_k}$ has in the reconstruction of the $i - th$ signature $\mathbf{x_i}$.

Fig 6.4 shows the association between the COSMIC signatures and the archetypes through the α coefficients. The heatmap was consequently clustered to find those signatures which share a common reconstruction pattern through the archetypes. It can be seen that 19 archetypes reconstruct only one signature, indicating a one-to-one relationship between them. Others were found to contribute in more than one signature at different weights, as well as there are groups of reconstructed signatures that are mainly represented by the same archetype, highlighted by different colours in Fig 6.4.

Interestingly, AA tends to group similar profiles together fairly well, since the signatures belonging to the same group usually share either the same aetiology or similar biological processes. In Fig B7 we further explored the relationship between the mutational signatures and the archetypes by plotting the pairwise cosine similarity distribution of the alpha coefficients profiles for different categories of pairwise cosine similarity between the original signatures. It is possible to clearly observe that, as the pairwise cosine similarity between the signatures increases, the cosine similarity between alpha coefficient profiles increases. This confirms that, while providing a more compact representation of the COSMIC signatures, the archetypal analysis is able to maintain a good consistency with the original profiles.

Table 6.3 summarises some of the qualitative information that can be extracted from the heatmap, showing the relationships between the reconstructed signatures and the archetype that contributed most to it. Each signature was reported with its aetiology, and whether it had been validated experimentally or by statistical association (i.e. unclear evidence for real signature, as reported in COSMIC).

It is possible to observe that seven signatures (SBS6, SBS14, SBS15, SBS20, SBS21, SBS26, and SBS44), experimentally associated with mismatch repair (MMR) deficiency, are divided into three groups: Blue, Silver Blue and Pink. The Blue group includes two signatures associated with the concurrent effect of MMR deficiency and DNA polymerase (POLD1 and POLE), showing a profile mainly polarised on C>A mutations, whereas the Silver Blue and Pink groups have mutational peaks at C>T and T>C, respectively. Thus, although all these 7 signatures are involved in MMR deficiency, they probably refer to different types of deficiency in MMR

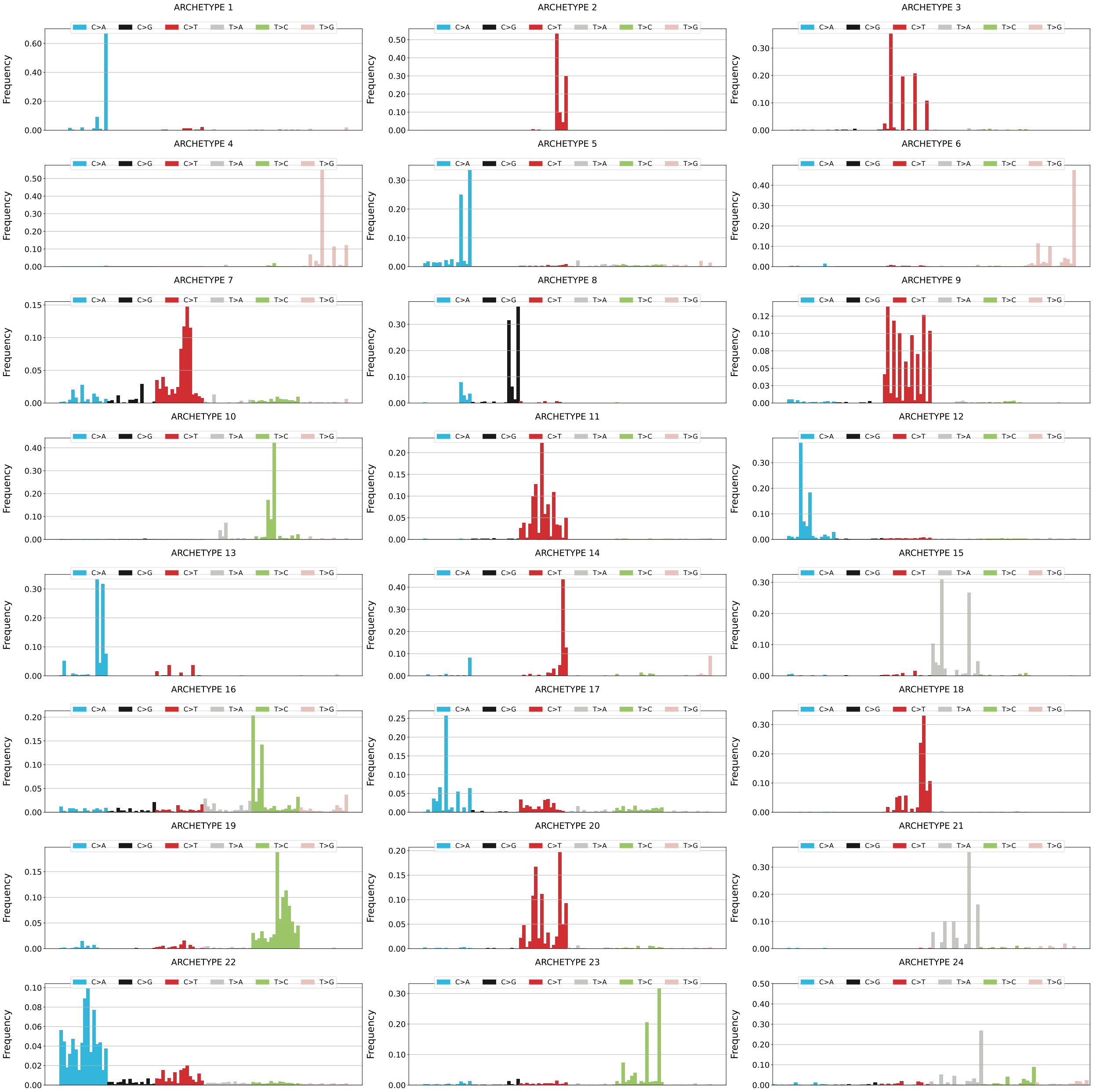

Figure 6.3: **Summary of the 29 archetypal profiles**

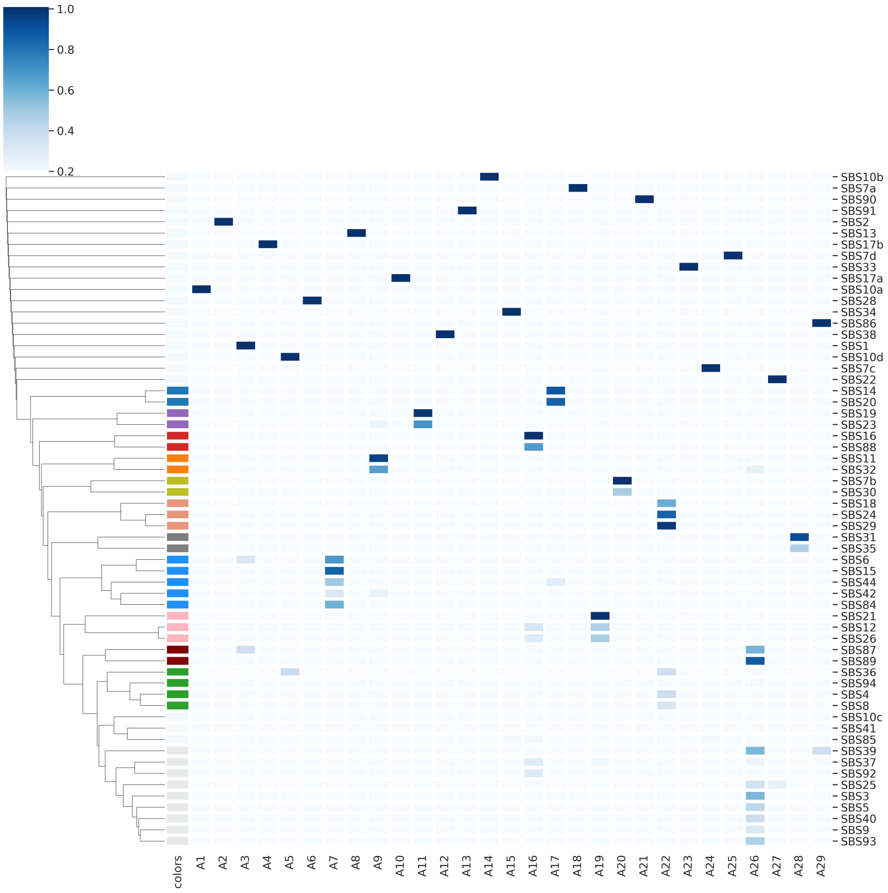

Figure 6.4: **Heatmap highlighting the associations between archetypes and the reconstructed signatures**.Different colors highlight groups of reconstructed signatures that are mainly represented by the same archetype. For a better visualisation, α coefficients < 0.2, even if used for clustering, were not displayed.

genes. The Silver Blue group, in addition to MMR deficiency associated signatures, also includes SBS84, which is statistically associated with AID activity, found in the immunoglobulin genes and other regions in lymphoid cancers. Although the MMR pathway is generally involved in repairing errors that can occur in DNA during the replication and the recombination,it was found to cooperate with the AID enzyme to generate DNA mutations as part of the antibody diversification process [97, 98, 99]. Thus, MMR pathway and AID mechanism are closely related. In this context, it is very interesting to note that SBS6, although it is implicated in several tumour types, was mainly found in B-Cell Non-Hodgkin Lymphoma samples `https://cancer.sanger.ac.uk/signatures/sbs/sbs6/` and that SBS84 is associated with the AID activity in B-Cell Non-Hodgkin Lymphoma. This might suggest that SBS6 relates to an MMR deficiency for genes involved in the development of antibody specificity that cooperates with AID activity. On the other hand, in the Pink group, in addition to the two signatures SBS21 and SBS26 associated with MMR deficiency, there is also SBS12 that was mainly found in liver cancer-related tissue; the high similarity (0.93, Table B2) between the signatures SBS26 and SBS12 and the unknown aetiology could suggest that either SBS12 is also related to MMR deficiency or that these two could actually correspond to the same signature. A cluster formed by these two signatures was also highlighted by [71] by performing an organ-wise mutational signature extraction. Another interesting group is the Grey one, including SBS31 and SBS35, both referring to platinum chemotherapy. The Orange group (SBS11 and SBS32) refers to two treatments as well: Temozolomide, an alkylating agent used as treatment for high-grade brain tumors and melanoma, and Azathioprine, an immunosuppressant. Both these treatments were shown to induce myelosuppression [100, 101]. The associated signatures are mainly represented by archetype A9, whose profile is characterised by high frequency of C>T mutations. The Yellow group includes SBS7b and SBS30, which are both characterised by C>T mutations and linked with ultraviolet (UV) radiation and base excision repair (BER) deficiency. Recently, it was found that BER increases cellular tolerance to UV independently of nucleic excision repair [102]. The Salmon group consisted in three signatures (SBS18, SBS24 and SBS29) associated with reactive oxygen species (ROS) damage, aflatoxin exposure and tobacco chewing, respectively. These three signatures are linked together by oxidative stress processes, since it is known that aflatoxin biosynthesis is linked with ROS [103, 104, 105, 106], as well as the cytotoxic effects from tobacco chewing are mediated by ROS production [107, 108]. The Green group is composed of SBS4, SBS8, SBS36 and SBS94. These signatures are mainly represented by the archetype A22 as the Salmon group. Indeed SBS29 (Salmon group) associated with tobacco chewing seems to be ”complementary” to SBS4 since it was found in some liver and lung cancers where SBS4, related to tobacco smoking, has not been detected. In addition, also SBS36 is associated with BER deficiency including DNA damage due to ROS, as SBS18 in the Salmon group. SBS8 is statistically associated with HR/NER deficiency. Purple, Red and Brown groups include mainly signatures with unknown aetiology and therefore it was not possible to establish qualitative associations between their signatures. Finally, the Light Grey group includes mainly signatures with a high level of flatness profile (SBS3, SBS5, SBS25 and SBS40) or at least with mutations distributed over all 96 mutational classes. Indeed, these signatures are mostly represented by A26, whose profile is in turn homogeneously distributed over all the classes.

6.4 Discussion

The study presented in this chapter investigates the extraction stability issues among the SBS mutational signatures of the most recent version of COSMIC catalogue (3.3). Through a series of simulations considering different scenarios, we showed that high levels of similarity combined with some peculiar (e.g. showing high level of flatness) signatures profiles considerably complicate the *de novo* extraction. Most of the previous studies evaluated stability issues on COSMIC signatures version 2, which includes 30 signatures. However, here we showed that these issues are becoming more critical in the newest version by evaluating 60 non-artefactual signatures. Although SigProfilerExtractor has been proven to be a robust method for signature extraction, even when the number of samples was high (i.e. up to 5,000), it failed in identifying the correct number of signatures and it succeeded 80% of times with 10,000 samples when we simulated a combined set of 6 similar signatures at high level of flatness and 5 similar signatures at a lower level of flatness(scenario 3). Similarly, in scenario 5, considering a higher number of latent signatures (20) with at least each signature

Color	**Main Archetype**	**Signature**	**Aetiology**	**Validation**
Blue	A17	SBS14	MMR deficiency + POLE	Experimental
		SBS20	MMR deficiency + POLD1	Experimental
Purple	A11	SBS19	Unknown	-
		SBS23	Unknown	-
Red	A16	SBS16	Unknown	-
		SBS88	Colibactin Exposure	Experimental
Orange	A9	SBS11	Temozolomide treatment	Experimental
		SBS32	Azathioprine treatment	Statistical
Yellow	A20	SBS7b	UV exposure	Experimental
		SBS30	BER deficiency	Experimental
Salmon	A22	SBS18	Damage by ROS	Experimental
		SBS24	Aflatoxin exposure	Experimental
		SBS29	Tobacco chewing	Statistical
Grey	A28	SBS31	Platinum chemiotherapy	Experimental
		SBS35	Platinum chemiotherapy	Experimental
Silver Blue	A7	SBS6	MMR deficiency	Experimental
		SBS15	MMR deficiency	Experimental
		SBS44	MMR deficiency	Experimental
		SBS84	AID activity	Statistical
Pink	A19	SBS12	Unknown	-
		SBS21	MMR deficiency	Experimental
		SBS26	MMR deficiency	Experimental
Brown	A26	SBS87	Thiopurine treatment	Experimental
		SBS89	Unknown	-
Green	A22	SBS4	Tobacco smoking	Experimental
		SBS8	HR/NER deficiency	Statistical
		SBS36	BER deficiency (ROS damage)	Experimental
		SBS94	Unknown	-
Light grey	A26	SBS3	HR deficiency	Experimental
		SBS5	Aging/Tobacco/NER deficiency	Statistical
		SBS9	Polymerase eta hypermutation	Statistical
		SBS25	Unknown Chemiotherapy	Statistical
		SBS39	Unknown	-
		SBS40	Unknown	-
		SBS93	Unknown	-

Table 6.3: **Aetiological information related to each archetype.** For each archetype, the corresponding reconstructed signatures and, when available, their associated aetiologies are reported, indicating the validation studies supporting the biological interpretation.

highly similar (i.e. pairwise cosine similarity > 0.8) to another one, it always failed up to 5,000 samples and with 10,000 samples it succeeded only 10% of the times.

Although the mutational signatures are not orthogonal by definition, the presence of highly similar signatures, together with the fact that some have a very high level of flatness and there is a lack of an aetiology for many of them, cast some doubts on the real existence of some of these, suggesting that they may be the result of overfitting and hence a mathematical artefact. Several studies already pointed to this issue [17, 87]. However, to the best of our knowledge, the most recent assessment of the signature stability observed among COSMIC signatures was performed by [95], where they considered the second version of this database, therefore working on half the number of signatures compared to our study and without exploring different scenarios in terms of number of samples, cosine similarity and flat vs. non-flat signatures. A limitation of the catalogues used in this work, realised with SigsPack functions, is represented by the random exposure assigned to each latent signature to create the count matrices, subsequently extracted by SigProfilerExtractor. Hence, simulated catalogues may not represent realistic cancer samples. However, this does not affect the technical evaluations of the limitations in the extraction process highlighted by our simulations.

A novelty introduced by this study was the application of AA to investigate whether the information contained in the COSMIC catalogue could be represented more compactly. AA was shown to be an intuitive and straightforward approach to interpret the data like the clustering, but including the flexibility of the matrix factorization [109, 110]. In contrast to the common distance-based approaches, archetypes characterise extremal rather than average properties of the given data and therefore lead to a more compact representation [111]. AA is a type of decomposition where convex combinations of extremal points lie on the convex hull of the data and are themselves restricted to being convex combinations of individual observations [110, 93]. In our study, by applying AA to the COSMIC catalogue, it was possible to identify 29 archetypes able to explain 95% of the variance. Interestingly, it emerged that most of the archetypes correspond almost perfectly (similarity>0.97) to some signatures and that, through a combination of them, it is possible to reconstruct with a certain degree of accuracy the other signatures of the COSMIC catalogue. As further validation of the reconstruction process, Table S4 shows the refitting performance for the simulated catalogues in Scenario 1 with 500 samples, comparing the archetypes to the original COSMIC signatures using MutationalPatterns. As expected, since the simulated catalogues are generated based on the original signatures, these latter showed high cosine similarity and low mean absolute error (MAE), i.e. 0.967 and 6.2 on average, respectively. However, keeping in mind that the archetypes explain the 95% of the variance in the simulated catalogue, they were able to perform well by achieving average cosine similarity and MAE equal to 0.958 and 9.91, respectively.

However, it is worth highlighting that archetypes do not substitute the COSMIC signatures, but emphasise the importance of considering alternative approaches able to reduce redundant information. These observations, together with the lack of known aetiology and experimental validation for many signatures, suggest the need to reformulate the COSMIC catalogue using representations including sparsity constraints in latent vectors during the extraction procedure without loss of information. In the future, archetypal analysis can be also considered to evaluate sparse representations of signatures not only in the context of single base substitutions but also for other types of variants like copy number variations and structural variants [112, 113].

Chapter 7

MUSE-XAE: MUtational Signatures Extraction with an Explainable AutoEncoder enhances tumour types classification

7.1 Introduction

Chapter 6 focused on a quantitative examination of the challenges in extracting mutational signatures under specific conditions. This chapter presents a new method for the de novo extraction of mutationa signatures developed in our lab. As mentioned in Chapter 5, Section 5.4, emerging studies suggest that mutational processes may not be strictly additive; they might also be interactive. To overcome the linear constraints of non-negative matrix factorization, we created MUSE-XAE. This neural network-based tool offers a model with enhanced flexibility. MUSE-XAE, MUtational Signatures Extraction with eXplainable AutoEncoder, includes a nonlinear encoder and a linear decoder with a non-negative constraint and a minimum volume regularization [114] to detect potential nonlinear dependencies while preserving signature interpretability.

Autoencoders have been successfully implemented across various domains, including genomics, to obtain compact and informative data representations. In particular, autoencoders employing a hybrid architecture with a nonlinear encoder and a linear decoder has been applied in the context of single-cell RNA-seq and trascriptomic data [115, 116], achieving great success due to their explainability while preserving powerful performance capabilities. However, from the best of our knowledge, this is the first application of such architecture in the context of mutational signatures analysis.

To fill this gap, in this paper we present MUSE-XAE and demonstrate its effectiveness on various cancer datasets by comparing it to existing state-of-the-art approaches in both synthetic scenarios and real-world applications. Specifically, we assessed our approach on the PCAWG dataset [117], which includes 2,780 cancer samples, and another whole-genome sequencing (WGS) cohort of 1,865 samples [77].

In a comprehensive comparison with 10 other *de novo* extraction tools considering realistic synthetic scenarios, MUSE-XAE resulted the best performing model with both high sensitivity and high precision in recovering the true signatures profiles. When applied to a real-world setting, MUSE-XAE has been shown to extract highly discriminative signatures profiles that can significantly improve the classification of tumor types and subtypes in the analyzed real datasets.

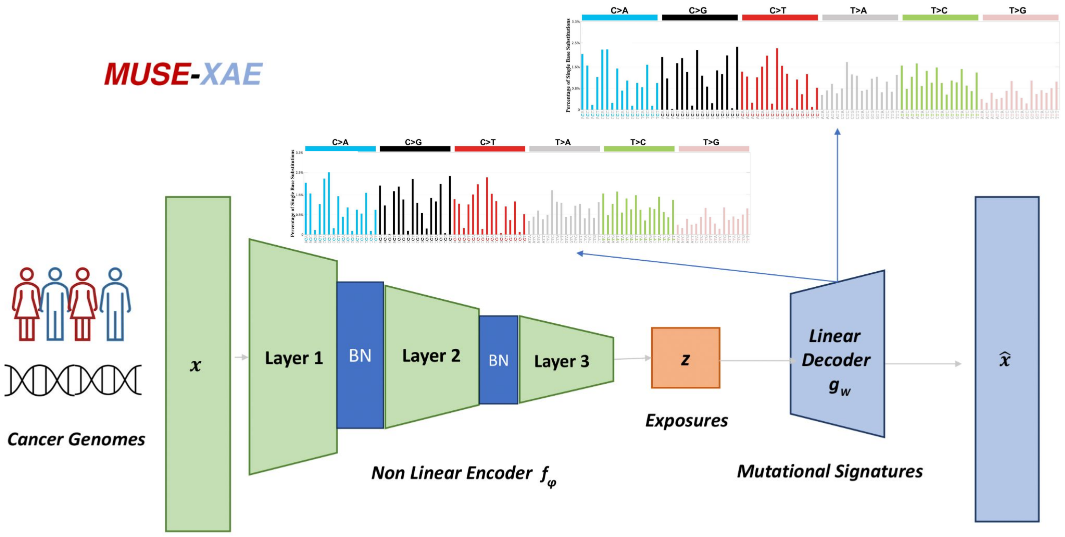

Figure 7.1: **MUSE-XAE schematic architecture.** MUSE-XAE features a nonlinear encoder, made up of from three layers that leverage a softplus activation function with batch normalization. The decoder is designed to be linear in order to enhance the interpretability.

7.2 Materials and methods

MUSE-XAE architecture

An Autoencoder is a type of neural network able to learn a lower dimensional representation of the data. Given an input space X, it consists of an encoder network f, represented by one or more layers, that maps the input data to a lower-dimensional latent space Z, and a decoder network g that reconstructs the input space from the latent representation. The goal of an autoencoder is to minimize the reconstruction error $\mathcal{L}(x, \hat{x})$ between the original input x and the reconstructed output $\hat{x}$. In particular, the general equations that define an autoencoder are:

$$z = f(x) = Encoder \tag{7.1}$$

$$\hat{x} = g(z) = g(f(x)) = Decoder \tag{7.2}$$

Usually, both f and g represent nonlinear activation functions.

MUSE-XAE implements a hybrid architecture with a nonlinear encoder for learning a latent representation z of cancer samples and a linear decoder, with non negative constraint and minimum volume regularization to reconstruct the original input, such as $\hat{x} = zW^T$. More specifically, MUSE-XAE encoder f is composed by three hidden layers with batch normalization and a softplus activation function. The softplus activation function offers continuous differentiability and a smoother transition from negative to positive values compared to ReLU, reducing the risk of neuron inactivation, thus providing a better stability [118]. The decoder g is constituted by a weight matrix W with non negativity constraint and linear activation function that ensures the interpretability, and a minimum volume regularization that helps the model to find a more disentangled representation.

In addition MUSE-XAE exploits a non negative Poisson likelihood function to better take into consideration the count nature of the input data and an early stopping criteria to avoid overfitting. Considering all the contributions, the total loss function $\mathcal{L}(x, \hat{x})$ can be written as:

$$\begin{aligned}\mathcal{L}(x, \hat{x}) = -x\log(\hat{x}) + \hat{x} + \beta \log\left(\det\left(WW^T + I\right)\right) \\ \text{subjected to} \quad W \geq 0\end{aligned} \tag{7.3}$$

Where the first two terms refers to the Poisson likelihood function, while the third term represents the logarithm of the minimum volume constraint. The β coefficient regulates the strength of the regularization. Referring to the mutational signatures terminology, the latent representation z represents the cancer genome's exposures, while the decoder weight matrix W represents the mutational signatures. MUSE-XAE architecture is displayed in Fig.7.1.

7.2.1 Signatures Extraction

Training a neural network requires a substantial amount of data to exploit its capacity and fit the parameters effectively. In MUSE-XAE implementation we used a data augmentation strategy to overcome this challenge. Specifically, given a tumour catalogue matrix $C \in R^{m \times 96}$, where m is the number of samples and 96 is the number of mutational channels, for each cancer genome with N total number of mutations, we determined the relative mutation frequency p for each of the 96 mutational classes. Then, we generated new data points by bootstrapping cancer genomes t times through a multinomial distribution $M(N, p)$, obtaining the augmented count matrix C_{aug}. This bootstrap approach has been already used by other tools with the aim to ensure the stability of a consensus signature [15, 77, 86]. In our case, we repeated this process t times to increase the dataset size.

Then, in order to select the optimal number of active signatures K, we used a revised version of the NMFk approach, originally described by [119] and also adopted by SigProfilerExtractor. Specifically, for each number $k = 1, ..., K$ candidate signatures, MUSE-XAE was trained n times with different weights' initialization to guarantee a better stability. Subsequently, a custom K-Means clustering with matching and based on cosine similarity distance was performed on the set of the decoder weight matrices $\{W_{1k}...W_{nk}\}$ to find a consensus signatures matrix S_k. This custom clustering approach exploits the Jonker-Volgenant algorithm [120] to solve the linear assignment problem (i.e the matching) and to find K clusters of equal size n.

Once obtained K clusters, whose centroids represent the signature matrices $S_k, k = 1...K$, we considered only the solutions with a mean and a minimum silhouette scores above a fixed threshold, choosing as the best solution that one allowing the minimum reconstruction error of the original matrix of cancer samples. MUSE-XAE *de novo* extraction procedure is summarized in Algorithm 1.

All the parameters mentioned above can be specified by the user, including the size factor t for data augmentation, the number of repetitions n for each candidate signature k and the thresholds for the mean and minimum silhouette scores, th_{mean} and th_{min}. An open source version of the code is available at https://github.com/compbiomed-unito/MUSE-XAE.

Algorithm 1: **MUSE-XAE De Novo Extraction procedure**

```
Given $C \in R^{m \times 96}$    ▷ tumour catalogue matrix
Step 1: Data Bootstrapping
Obtain the augmented count matrix $C_{\text{aug}}$
bootstrapping each genome $t$ times from $M(N, p)$

Step 2: Training
for $k$ in 1..$K$ signatures do
    for $i$ in 1..$n$ iterations do
        MUSE-XAE.train($C_{aug}$)
        $C_{pred}$=MUSE-XAE.predict($C$)
        Collects $E_{rec} = \|C - C_{pred}\|_F$
        Collects $W_{ik}$

Step 3: Clustering
for each $k$ in 1..$K$ do
    K-Means clustering with matching on $\{W_{1k}, ..., W_{nk}\}$
    Obtain consensus Signatures matrix $S_k$

Step 4: Filtering
Given threshold $th_{\text{mean}}$ and $th_{\text{min}}$
Filter solutions with
$silhouette_{\text{mean}} > th_{\text{mean}}$ and $silhouette_{\text{min}} > th_{\text{min}}$

Step 5: Optimal Solution Selection
Sort filtered solutions based on $E_{rec}$,
The optimal solution is the one with the lowest $E_{rec}$
```

7.2.2 Signatures Assignment

Once the profiles of active signatures within a set of cancerous genomes have been identified, it is necessary to understand which signatures are causing mutations within a genome and in which quantity, i.e. we need to assign the contribution of each extracted signature to each genome. To accomplish this, we utilized a slightly modified version of MUSE-XAE used for the signatures extraction. Specifically, we normalized the computed consensus matrix S_k into $S_{k_{norm}}$, which is used to initialize the weights of the decoder and then freeze them, meaning that the decoder is no longer trainable. Therefore, we only allowed the weights of the encoder to be trained.

In order to obtain a sparse representation and to avoid over-assignments of mutational signatures, we used an L1 penalty for both the weights of the last layer of the encoder and for the output of the encoder after a ReLU activation, training the network until convergence. Our new latent representation z represents the exposure of the signatures within the genomes, i.e., the number of mutations of a certain mutational class that a signature causes within a genome. We summarized the signature assignment procedure in the Algorithm 2.

Algorithm 2: **Signature Assignment Procedure**

```
Given $C \in R^{m \times 96}$    ▷ SBS catalogue matrix
Step 1: Signature Profiles Normalization
Normalize the consensus matrix $S_k$ from Alghoritm 1.
Get $S_{k_{norm}}$

Step 2: Initialization and Freezing of Weights
Initialize and freeze decoder weights with $S_{k_{norm}}$

Step 3: Improve Sparsity
Add an L1 penalty on the weights of the last encoder layer
Add an L1 penalty on the output after activation

Step 4: Train the network and obtain Exposures
MUSE-XAE.train($C$)
Exposures=MUSE-XAE.z
```

De Novo Extraction Scenarios

To evaluate the performance of MUSE-XAE in the context of mutational signature extraction, we utilized 5 publicly available realistic synthetic scenarios (`ftp://alexandrovlab-ftp.ucsd.edu/pub/publications/Islam_et_al_SigProfilerExtractor/`), also used in the thorough benchmarking conducted in the SigProfilerExtractor paper [77] for the currently available mutational signature tools. Specifically:

- **Scenario 1:** 1,000 synthetic samples, modeling a subset of the pancreatic adenocarcinoma PCAWG dataset. The 11 ground-truth signatures are based on COSMIC.
- **Scenario 2:** 1,000 synthetic tumors from flat, relatively featureless mutational signatures, including a mix of 500 synthetic renal cell carcinomas (high prevalence and mutation load from SBS5 and SBS40) and 500 synthetic ovarian adenocarcinomas (high prevalence and mutation load from SBS3), with 11 COSMIC-based signatures.
- **Scenario 3:** 1,000 synthetic tumors from signatures with overlapping and potentially interfering profiles, mostly SBS2, SBS7a, and SBS7b. The mutational load distributions were drawn from bladder transitional cell carcinoma (SBS2) and skin melanoma (SBS7a, SBS7b), with 11 COSMIC-based signatures.
- **Scenario 4:** 1,000 synthetic tumors emulating a mix of 500 synthetic renal cell carcinomas (high prevalence and mutation load from SBS5 and SBS40) and 500 synthetic ovarian adenocarcinomas (high prevalence and mutation load

from SBS3). In this scenario, only 3 COSMIC-based signatures (SBS3, SBS5, SBS40) are present.

- **Scenario 5:** 2,700 synthetic samples with mutational spectra matching the ones observed in PCAWG, including 300 spectra from each of 9 different cancer types: bladder transitional cell carcinoma, esophageal adenocarcinoma, breast adenocarcinoma, lung squamous cell carcinoma, renal cell carcinoma, ovarian adenocarcinoma, osteosarcoma, cervical adenocarcinoma, and stomach adenocarcinoma. The ground-truth signatures are 21 signatures based on COSMIC.

Specifically, we extracted mutational signatures from each of these scenarios using MUSE-XAE and applied the same performance metrics as in [77]. We used the Hungarian algorithm [120] to match the predicted and the known signatures based on the cosine-similarity scores. Since the signatures in each scenario are known, an extracted signature was considered correctly identified, or a True Positive (TP), if the cosine similarity between extracted and real signatures was $\geq$ *threshold.* If the profile of a signature is missing or the cosine similarity was $<$ *threshold*, it was considered as a False Negative (FN) or Positive (FP), respectively.

For each scenario, precision, sensitivity, and F1 Score were calculated from the corresponding confusion matrices at different cosine similarity thresholds, ranging between 0.8 and 1. A description of the evaluation metrics was reported in supplementary (section *Evaluation metrics*).

7.2.3 Real World datasets

In order to evaluate the performance of MUSE-XAE also in real-world scenarios, we applied our method to both the Pancancer Analysis of Whole Genomes (PCAWG) dataset, including 2,780 tumor samples, and a WGS cohort of 1,865 genomes collected from various studies and including the International Cancer Genome Consortium (ICGC), as compiled in [77]. For both datasets, we performed: 1) a *de Novo* extraction of mutational signatures and comparison of the profiles with those of SigProfilerExtractor and with the known signatures from COSMIC [121] and Signal [122] databases; 2) an evaluation of how the signatures and consequently the exposures are discriminative, performing a multi-class classification of the cancer types. Specifically, we used the exposures as new features which were fed into a Random Forest to classify both the primary sites and the cancer subtypes. Finally, we evaluated the performance in terms of balanced accuracy, Matthews Correlation Coefficient (MCC) and Kohen Kappa score in a 5-fold cross-validation setting. A description of the metrics was reported supplementary (section *Evaluation metrics*).

7.3 Results

7.3.1 Data augmentation improves robustness and accuracy

In our initial analysis, we investigated the influence of data augmentation on the extraction of mutational signatures across each of the the five synthetic scenarios. Specifically, we performed *de novo* extraction with MUSE-XAE for each of the five datasets, varying the data augmentation level from 1 to 100 times the original dataset size. We repeated the extraction five times at each augmentation level to evaluate stability and accuracy. As depicted in Fig. 7.2, for all the five datasets there is a trend where an increase in data augmentation not only enhances run-to-run stability, but also improves the correct estimation of the real number of profiles.

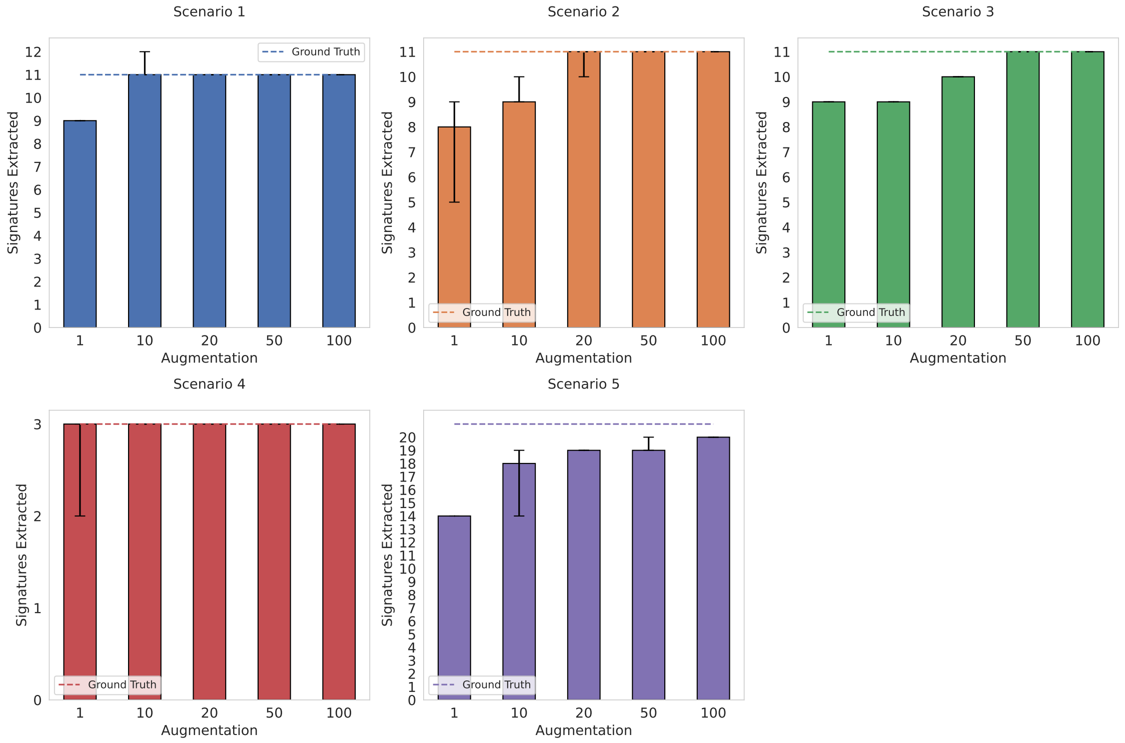

Figure 7.2: Sensitivity analysis of data augmentation for each of the five synthetic scenario. Each bar represents the average number of extracted signatures over 5 repetitions. The dashed line represents the ground truth, while the black error bar represents the minimum and maximum.

To further assess the effects of data augmentation, we computed the average precision, sensitivity and F1 scores across the five scenarios at different thresholds of the cosine similarity between the extracted and the real profile, ranging between 0.8 and 1. Fig 7.3 shows that the overall performance, notably the sensitivity in the signature profile detection, improves with the size of the data augmentation. This confirms that employing data augmentation is a strategy that improves the detection of signature profiles and it can be used as an effective technique to further enhance the extraction performance.

7.3.2 *De Novo* extraction comparison in synthetic scenarios

To assess the performance of our approach with the current state-of-the-art tools, we compared MUSE-XAE (using 100 data augmentation) with 10 state-of-the-art *de novo* signature extraction tools, considering the results reported in `ftp://alexandrovlab-ftp.ucsd.edu/pub/publications/Islam_et_al_SigProfilerExtractor/`. Precision, sensitivity and F1-score were computed in each scenario at different thresholds of the cosine similarity between the extracted and the real profile, ranging between 0.8 and 1 for each method.
Supplementary Fig. C1 (Appendix C) shows precision, sensitivity and F1 score of the top 10 performing methods, averaged across the five scenarios, while Supplementary

Fig. C2 reports, for each scenario, the F1 scores at different thresholds of the cosine similarity. Both Fig. C1 and Table 7.1 reveal that MUSE-XAE is, on average, the best performing method in all metrics, followed by SigProfilerExtractor and SigProfilerPCAWG.

Table 7.1: AUC_{norm} for precision, sensitivity and F1 score curves for each method, averaged across the five synthetic scenarios. Methods are ordered according to the AUC of the F1 score

Method	**AUC Precision**	**AUC Sensitivity**	**AUC F1-score**
MUSE-XAE	$\mathbf{0.92 \pm 0.05}$	$\mathbf{0.93 \pm 0.04}$	$\mathbf{0.92 \pm 0.05}$
SigProfilerExtractor	0.89 ± 0.07	0.91 ± 0.04	0.90 ± 0.06
SigProfilerPCAWG	0.89 ± 0.07	0.91 ± 0.05	0.90 ± 0.06
SigneR	0.87 ± 0.09	0.91 ± 0.12	0.89 ± 0.09
SignatureAnalayzer	0.85 ± 0.09	0.90 ± 0.03	0.88 ± 0.05
MutationPatterns	0.80 ± 0.11	0.92 ± 0.03	0.86 ± 0.07
SignaturesToolsLib	0.84 ± 0.08	0.87 ± 0.07	0.85 ± 0.07
MutSpec	0.76 ± 0.14	0.92 ± 0.03	0.83 ± 0.09
SomaticSignatures	0.68 ± 0.19	0.86 ± 0.08	0.75 ± 0.14
Maftools	0.64 ± 0.27	0.81 ± 0.13	0.69 ± 0.22
SigMiner	0.54 ± 0.20	0.85 ± 0.12	0.65 ± 0.19

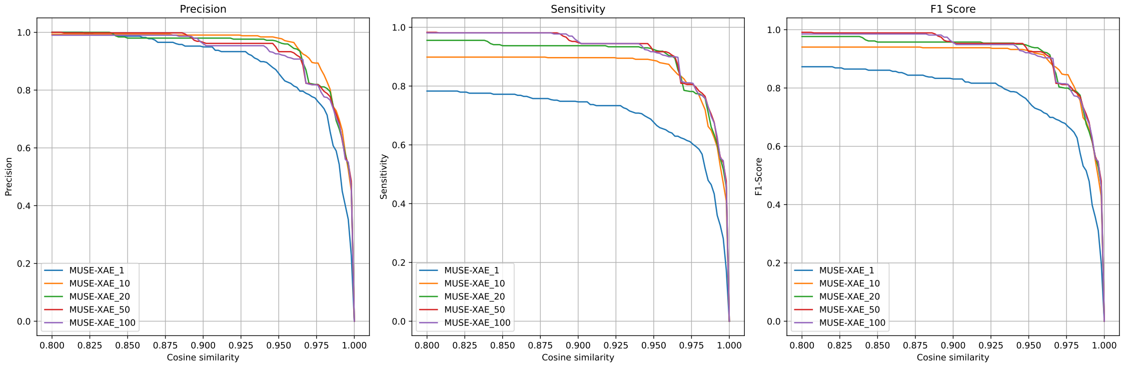

Figure 7.3: Average precision, sensitivity and F1 scores across five scenarios for MUSE-XAE at different level of data augmentation and at varying thresholds of the cosine similarity, ranging between 0.8 and 1.

7.3.3 *De Novo* Extraction in real world datasets

We applied MUSE-XAE for *de novo* extraction of mutational signatures to both the PCAWG cohort, including 2,780 samples from 18 cancer primary sites and 37 cancer subtypes, and an additional extended WGS cohort, collecting genomes from 1,865 samples across 15 cancer primary sites and 23 cancer subtypes. Specifically, we used MUSE-XAE with the data augmentation strategy (i.e. 100 times the original dataset size for 100 iterations) to find stable consensus signatures. MUSE-XAE found 22 and 23 mutational signatures profiles in the PCAWG and the extended

WGS cohort, respectively. Their profiles are presented in Appendix C, Fig. C3 and Fig. C4. By matching the 22 profiles identified by MUSE-XAE with the 21 found by SigProfilerExtractor in the PCAWG cohort, it resulted that the two methods extracted 21 highly similar profiles, showing a mean cosine similarity of 0.98, with a minimum of 0.92.

On the other hand, in the extended WGS cohort MUSE-XAE found 23 signatures while SigProfilerExtractor 21, with a mean cosine similarity of 0.90 but with a minimum of 0.36 between the 21 most similar signatures. Fig. 7.4 shows an heatmap representing the cosine similarity between MUSE-XAE and SigProfilerExtractor 21 most similar signatures on both PCAWG and the extended WGS cohort. It is possible to observe that in the PCAWG cohort, the 21 profiles from the two methods are highly aligned. However, in the extended WGS cohort, although there are 19 out of 21 profiles with a cosine similarity greater than 0.8, the distribution along the diagonal is lower than that one observed in PCAWG. Moreover, there are two pairs of signatures with a notably low cosine similarity, specifically 0.69 and 0.36, meaning that the two methodologies extract different signature profiles. Therefore, in general, although the two methods are fairly in agreement, MUSE-XAE seems to identify more and different profiles compared to SigProfilerExtractor.

Given that two random 96-component vectors have a cosine similarity of 0.75, and 0.80 is commonly used as a threshold to determine if two signatures represent the same profile, we can observe from Appendix Table C1 that almost all MUSE-XAE profiles from the PCAWG cohort are in agreement with the known signatures from COSMIC and Signal databases. An exception is MUSE-SBSV, which shows a cosine similarity of 0.78 in both the COSMIC and Signal databases, potentially indicating an incomplete extraction of the original signature. On the other hand, in the WGS extended cohort, despite most extracted profiles align well with those in COSMIC and Signal databases (Appendix Table C2), there are three signatures (MUSE-SBSP, MUSE-SBSS, and MUSE-SBSW) with a cosine similarity below 0.75 with the matched signatures in both databases.

Hence, we further investigated the exposures of these three signatures in the WGS extended cohort. Notably, as depicted in Appendix Figure C5, MUSE-SBSW is predominantly observed in Eye-Melanoma samples (32 out of 46), indicating that it might be a tumor-specific signature. To validate this hypothesis, we carried out a *de novo* extraction exclusively for the Eye-Melanoma samples, which revealed a strikingly similar profile, showing a pairwise cosine similarity of 0.94 with the one extracted from the pancancer analysis. Given the limited number of samples, this finding reinforces the need for a comprehensive examination of this profile, focusing on its origin and validation in an external cohort. Such an in-depth investigation, however, exceeds the objectives of our study and it will be a focus of our future research.

7.3.4 MUSE-XAE enhances tumor classification

Considering *de novo* extraction of mutational signatures on real cancer datasets, although COSMIC and Signal databases can be used as references for the extracted profiles, there is no actual ground truth to calculate the evaluation metrics. Therefore, to thoroughly evaluate the performance of MUSE-XAE, we examined the exposures of mutational signatures, i.e the latent representation z of tumor samples, both

Model	Dataset	Matthews score	Cohen score	Balance Accuracy
MUSE-XAE	**PCAWG**	$\mathbf{0.75 \pm 0.02}$	$\mathbf{0.75 \pm 0.02}$	$\mathbf{0.65 \pm 0.02}$
SigProfiler	PCAWG	0.71 ± 0.01	0.71 ± 0.01	0.61 ± 0.02
MUSE-XAE	**Extended**	$\mathbf{0.73 \pm 0.01}$	$\mathbf{0.72 \pm 0.01}$	$\mathbf{0.65 \pm 0.01}$
SigProfiler	Extended	0.70 ± 0.02	0.69 ± 0.02	0.61 ± 0.02

Table 7.2: Matthews correlation coefficient, Kappa score, and Balanced accuracy metrics calculated in 5 fold cross-validation for PCAWG and Extended WGS cohort for primary tumour type classification

Model	Dataset	Matthews score	Cohen score	Balance Accuracy
MUSE-XAE	**PCAWG**	$\mathbf{0.73 \pm 0.01}$	$\mathbf{0.73 \pm 0.01}$	$\mathbf{0.58 \pm 0.01}$
SigProfiler	PCAWG	0.67 ± 0.02	0.67 ± 0.02	0.52 ± 0.02
MUSE-XAE	**Extended**	$\mathbf{0.68 \pm 0.03}$	$\mathbf{0.67 \pm 0.03}$	$\mathbf{0.50 \pm 0.03}$
SigProfiler	Extended	0.66 ± 0.02	0.66 ± 0.02	0.50 ± 0.03

Table 7.3: Matthews correlation coefficient, Kappa score, and Balanced accuracy metrics calculated in 5 fold cross-validation for PCAWG and Extended WGS cohort for tumour subtype classification.

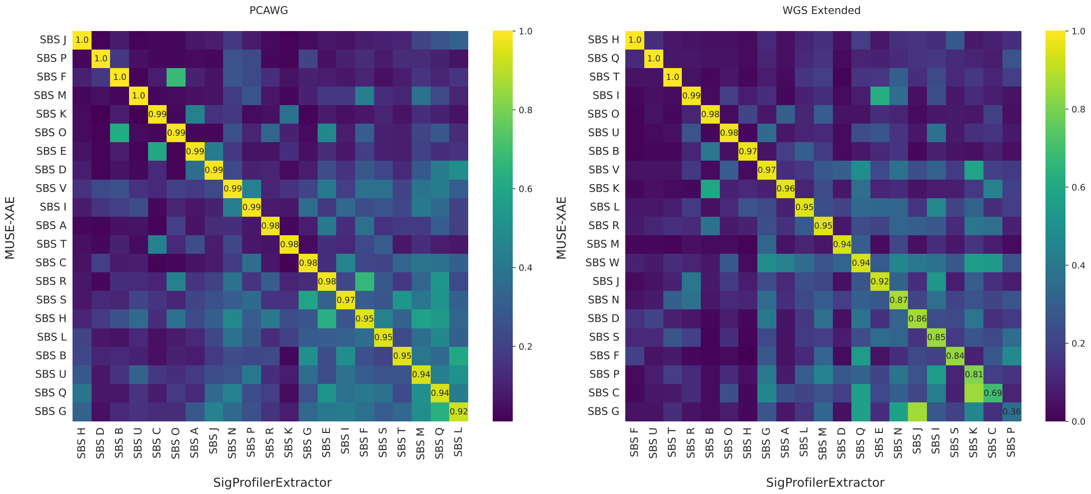

Figure 7.4: Cosine similarity heatmap between the MUSE-XAE and SigProfilerExtractor most similar extracted signatures for PCAWG and WGS extended cohort.

qualitatively and quantitatively. While acknowledging that tumors of the same type may demonstrate a degree of heterogeneity, we assumed that these exposures, representing the mutations caused by a signature within a particular sample, could serve as a key discriminant between different tumor types and subtypes.

Fig. 7.5 shows the t-distributed stochastic neighbor embedding (t-SNE) of the latent representations (exposures), coloured by the primary tumour types both for the PCAWG and the extended WGS cohort. The t-SNE of exposures displays a clear grouping pattern in both datasets, which provides compelling evidence in support of this hypothesis and indicating a coherent relationship between signatures exposures and tumor types.

To quantitatively assess this hypothesis, we implemented a Random Forest Classifier which considers the signature exposures as input features to classify both primary

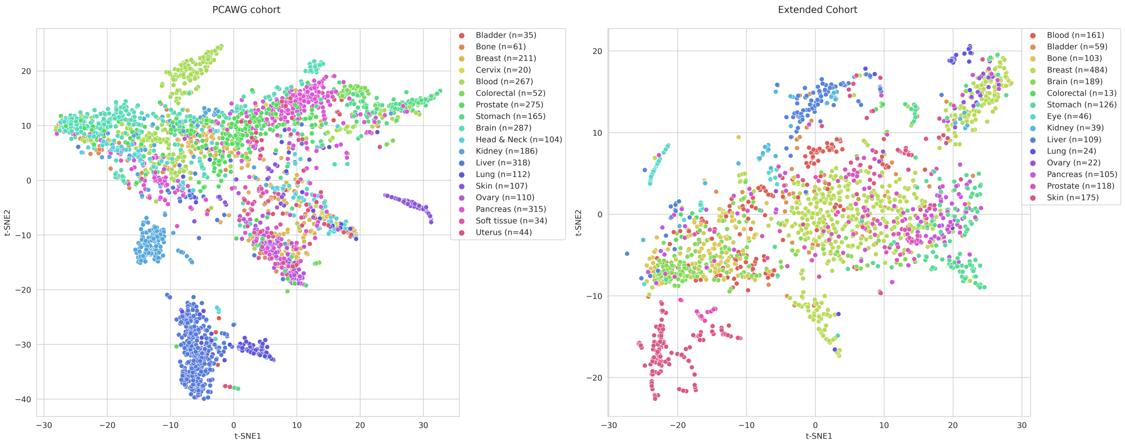

Figure 7.5: t-SNE representation of the latent representation of PCAWG and the extended WGS cohorts, post-hoc coloured by primary tumour sites

types and tumour subtypes. This classifier was applied both to MUSE-XAE and SigProfilerExtractor exposures using a balanced 5-fold cross-validation approach. To properly train the Random Forest in both datasets, we removed tumor types with less than 10 counts, i.e. the tumor subtypes Myeloid-MDS (n=4), Breast-DCIS (n=4) and Cervix-AdenoCA (n=2) in the PCAWG dataset, while in the extended WGS dataset, we excluded Blood-CMDI (n=9), Sarcoma (n=3), and Bone-cancer (n=2). Classification performance metrics for primary tumour types and tumour subtypes in PCAWG and in the extended WGS cohort are reported in Tables 7.2 and 7.3, respectively.
In both classification tasks, MUSE-XAE outperformed SigProfilerExtractor across all metrics, suggesting that the exposures and the corresponding signature profiles generated by MUSE-XAE are more discriminative and able to accurately identify the tumor types. MUSE-XAE particularly outperformed than SigProfilerExtractor in primary types classification (Fig. 7.2) and it discriminates tumor subtypes much better, notably in the PCAWG cohort (Table 7.3). Appendix Figures C6-C9 display the complete confusion matrices of MUSE-XAE for both primary tumour types and subtypes in PCAWG and the extended WGS cohort.

7.4 Discussion

The study presented in this chapter introduces MUSE-XAE, a novel method for mutational signatures extraction based on an explainable autoencoder. MUSE-XAE combines a nonlinear encoder with a linear decoder by adding a non-negative constraint and a minimum volume regularization. Our method demonstrated high accuracy in the *de novo* extraction of mutational signatures, proven through a sensitivity analysis and a comprehensive comparison with 10 other available tools. In particular, MUSE-XAE resulted as the best-performing and the most robust method in different realistic synthetic scenarios, with an average F1-AUC of 0.92. In addition, MUSE-XAE identified 22 mutational signature profiles in the PCAWG cohort and 23 mutational signatures in the WGS extended cohort with a high agreement with the known signatures from both COSMIC v3.4 and Signal databases. Notably, in the extended WGS cohort we found a candidate novel signature specific to Eye-Melanoma.

This finding will need to be further investigated and validated in an independent cohort.

A detailed investigation of the mutational signature exposures revealed that MUSE-XAE profiles are very informative and capable of enhancing primary tumour type and subtype classifications. Indeed, the classification performance based on the signature exposures showed MCC around 0.70 in predicting primary types and tumour subtypes in both PCAWG and the extended WGS cohort.

MUSE-XAE opens up new possibilities for the development of interpretable neural network-based models for mutational signature extraction, which can leverage the increasing amount of available data and their scalability for larger datasets. Our architecture, given its extreme flexibility, can be used to build more sophisticated models which could integrate the profile of somatic mutations with other clinical and genomic information, potentially improving and refining the extraction of mutational signatures.

Chapter 8

Conclusion

The objective of this thesis is to highlight the role of genetic variations through a dual lens. On one hand, this work focuses on missense variants and in particular on protein stability prediction due to an aminoacid changes in the protein sequence. On the other hand, it delves into the genomic dimension by examining mutational signatures and their relationship with tumorigenesis.

Accurate prediction of protein stability due to variations is pivotal for advancing precision medicine and protein engineering. As outlined in Chapter 2, over half of the variants responsible for monogenic diseases are those that disrupt the stable state of proteins [10]. Moreover, missense variants significantly affecting protein stability are key factors in disease progression, particularly in cases of gene haploinsufficiency [11].

Hence, computational methods capable of reliably predicting variants effects on protein stability are essential for accurately determining pathogenicity. In Chapter 2, we discussed several challenges in developing a new $\Delta\Delta G$ predictor such as the small size of the training datasets and the unbalanceness towards the destabilizing variants. In addition we presented two fundamental thermodynamic properties of $\Delta\Delta G$ such as antisymmetry and transitivity that most of the state of the art predictor does not consider thus performing worst on reverse mutations.

In Chapter 3, we presented an Antisymmetric Convolutional Differential Concatenated Neural Network (ACDC-NN) and its sequence-based version ACDC-NN-Seq, two intrinsically antisymmetric predictors we developed in our lab that naturally incorporate antisimmetry. Both methods used transfer learning to face the challenge of small datasets size and consequentyly fine tuned their hyperparameters on real world datasets (S2648 and Varibench). In addition even if both ACDC-NN and ACDC-NN- Seq were not specifically built to satisfy the transitivity property, yet they do, as shown in [23], thanks to the incorporation of the antisymmetry through a specifically designed loss function.

In chapter 4, we presented a new dataset we manually cleaned from ThermoMutDB, named S669. On this dataset we performed a comprehensive comparison among over 20 widely-used predictors. This dataset can be used as fair benchmarking because includes variants found in proteins with less than 25% sequence identity compared to those in the S2648 and VariBench datasets. Therefore, S669 serves as blind set to test the real performance for all the methods.

Our analysis indicated that the accurate prediction of stabilizing variants is still an open challenge and existing methods are only effective at predicting stabilizing ones, with a tendency of predictions towards neutrality. It was also observed that, in a balanced test set, antisymmetric predictors outperformed other methods. Notably, for some tools like ACDC-NN and DDGun, the sequence-based versions were as effective as their structure-based counterparts.

In conclusion, future predictors should be trained and tested on datasets with low sequence similarity. Moreover, their performance should be reported distinctly for stabilizing, neutral, and destabilizing classes, rather than just overall.

In Chapter 5 we described what mutational signatures are and their relevance in clinical settings to predict prognosis, to understand the stage of cancer and to develop target therapies. In the same chapter we also presented the open challenges in the field, in particular we focused on the problem of *de novo* extraction in the presence of small datasets size, different tissue types, concomitant similar signatures and flat profiles. In addition we suggested the idea that many COSMIC signatures without a known aetiology, given their high similarity with other signatures they could represent mathematical artefacts.

In Chapter 6 we presented in detail our Archetypal Analysis of COSMIC (v 3.2) signatures that quantitatively explain the instability of the extraction of mutational signatures in the presence of similar signatures and with a high flatness. This study support the hypothesis of signatures redundancy in the database. In particular, it was possible to identify 29 archetypes able to explain 95% of the variance of the COSMIC catalogue that contains 60 non artefactual mutational signatures. Interestingly, it emerged that most of the archetypes correspond almost perfectly (similarity>0.97) to some signatures and that, through a combination of them, it is possible to reconstruct with a high degree of accuracy the other signatures of the COSMIC catalogue. However, it is worth highlighting that archetypes do not substitute the COSMIC signatures, but emphasise the importance of considering alternative approaches able to reduce redundant information. These observations, together with the lack of known aetiology and experimental validation for many signatures, suggest the need to reformulate the COSMIC catalogue using representations including sparsity constraints in latent vectors during the extraction procedure without loss of information.

In Chapter 7 we introduced a novel method for mutational signatures extraction based on an explainable autoencoder, named MUSE-XAE. MUSE-XAE combines a nonlinear encoder with a linear decoder by adding a non-negative constraint and a minimum volume regularization. Our method demonstrated high accuracy in the *de novo* extraction of mutational signatures, proven through a sensitivity analysis and a comprehensive comparison with 10 other available state of the art tools. In particular, MUSE-XAE resulted as the best-performing and the most robust method in different realistic synthetic scenarios, with an average F1-AUC of 0.92. In addition, MUSE-XAE identified 22 mutational signature profiles in the PCAWG cohort and 23 mutational signatures in the WGS extended cohort with a high agreement with the known signatures from both COSMIC v3.4 and Signal databases. Notably, in the extended WGS cohort we found a candidate novel signature specific to Eye-Melanoma. This finding will need to be further investigated and validated in an independent cohort.

A detailed investigation of the mutational signature exposures revealed that MUSE-XAE profiles are very informative and capable of enhancing primary tumour type and subtype classifications. MUSE-XAE opens up new possibilities for the development of interpretable neural network-based models for mutational signature extraction, which can leverage the increasing amount of available data and their scalability for larger datasets. Our architecture, given its extreme flexibility, can be used to build more sophisticated models which could integrate the profile of somatic mutations with other clinical and genomic information, potentially improving and refining the extraction of mutational signatures.

Appendix A

Protein Stability

A.0.1 Method performance on model vs experimental structures

The current data repositories and the derived datasets are skewed towards the destabilizing variants. Using the thermodynamic property of antisymmetry, we can double the data and perfectly balance the distribution by adding the reverse variants. This procedure works smoothly for sequence-based methods, but structure-based methods require the atomic coordinates, and unfortunately, very few pairs of wild-type and mutated protein structures with experimental $\Delta\Delta G$s are available.
The most significant effort in this direction has produced the Ssym dataset, including 684 variants (342 direct and 342 reverse) with 19 experimental structures for the direct variants and 342 experimental structures for each of the reverse variants [27]. From ThermoMutDB [28] we extracted 10 more reverse structures, which slightly increased the Ssym dataset (Ssym+ consists of 704 variants).
Another way to generate the reverse structure when the experimental one is not available is through comparative modelling. However, it is not clear if using a predicted model can hamper the predictive performance of the methods. To test the possibility of using single-point mutation models as reverse structures, we generated 704 protein models for each Ssym+ structure. Thus, a model of the direct protein is obtained from the corresponding reverse PDB structure (and vice versa). To compute the model structures we used Rosetta/Robetta server.
To assess the generated models regarding the PDB structures, we performed 1,408 predictions (704 for the experimental structures and 704 for the modelled ones) for each structure-based method. The comparison of the performance obtained for each method in the two scenarios (experimental versus modelled structures) is reported in Fig.A.1 and in [123]. The results indicate that there is no performance degradation using the models as a proxy for the experimental structure. This finding supports the idea of balancing datasets by adding the reverse of all variants and modelling the missing mutated structures with Rosetta/Robetta.

A.0.2 Effect of the experimental technique on the method performance

One interesting point that has been recently studied is the possible dependence of the method performance in the choice of the protein structure [124]. Caldararu et al.

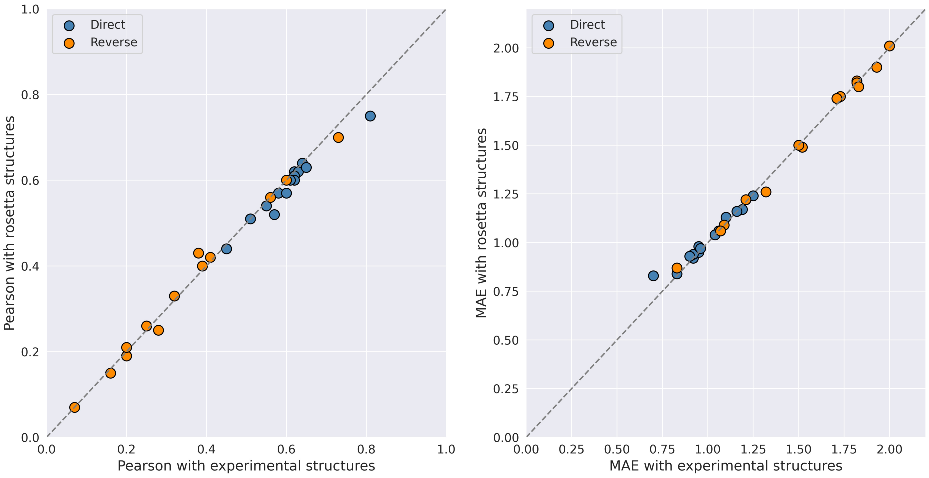

Figure A.1: **Comparison of method performance on real and modelled structure on the Ssym+ dataset**. Pearson correlation coefficients (r) and mean absolute error (MAE) are displayed in the left and in the right figure respectively. The prediction performances obtained from the experimental structures (x-axis) is plotted against those from the Rosetta-simulated structures (y-axis). Performances calculated with real or modeled structures are very consistent, with correlations of 0.995 (p-value$< 10^{-13}$) and 0.993 (p-value $< 10^{-12}$) for the Pearson and MAE, respectively.

showed that some methods, such as FoldX, are more sensitive to the change of the three-dimensional protein structures [124]. To test whether different experimental strategies have an impact on the performance of the structure-based $\Delta\Delta G$ predictors, we divided the S669 into variants from structures that were obtained by Nuclear Magnetic Resonance spectroscopy (NMR) and structures obtained by X-Ray diffraction. The protein structure solved using the NMR technique usually presents several models in the corresponding PDB file that are all compatible with the experimental constraints. As usually done, we selected the first model as representative. Fig. A.2 displays the method performance on the two subsets of variants. Largely overlapping error bars show that most methods are quite insensitive to experimental strategy, even though a general trend of slightly increased performance for NMR-derived structures can be observed. Only FoldX and PremPS showed a clear preference for X-Ray- and NMR-derived structures, respectively. However, the observed differences are probably due to the variations in the NMR and X-Ray sets rather than to the specific experimental technique.
Furthermore, the overall performance of the methods seems mostly unaffected by the X-ray resolution, at least in the range from 1.2 to 3.2 Angstrom seen in S669. Fig. A.3 displays the results obtained by splitting the X-Ray structures in those that have been crystallized at a resolution above or below Ångstrom. The only methods that seem sensitive to the resolution are PremPS and Dynamut2 for the Pearson correlation and Rosetta for the MAE.

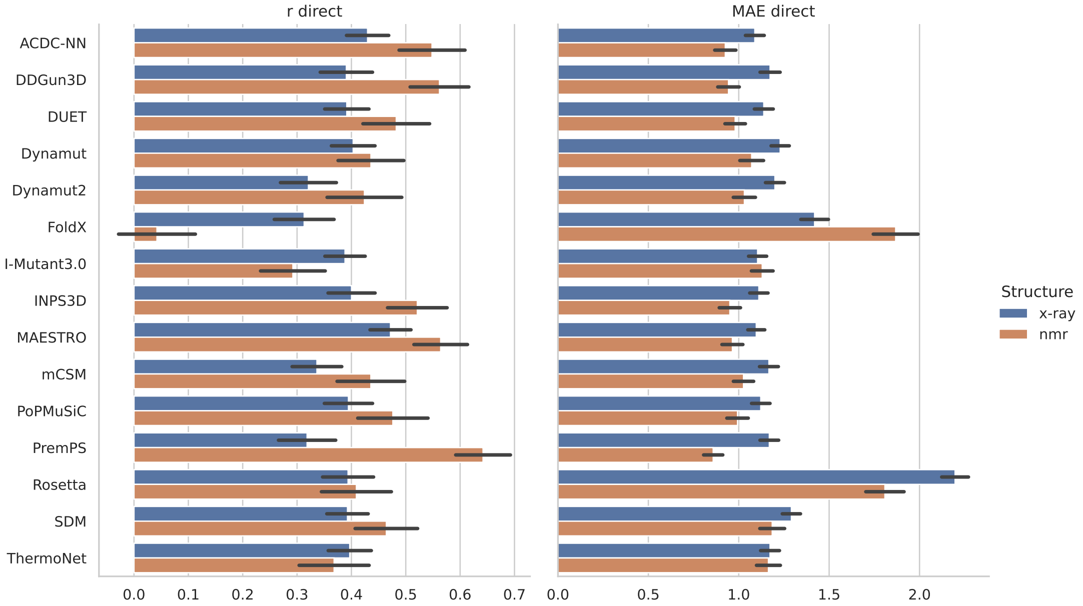

Figure A.2: **Effect on the method performances of the experimental technique: NMR versus X-ray.** After splitting the S669 dataset into NMR and X-ray derived structures (with 196 and 473 variants in 23 and 71 proteins, respectively), Pearson correlation coefficients (r direct, on the left) and mean absolute error (MAE direct, on the right) for the direct variants are shown for all structure-based methods. The black error bars represent the bootstrap estimated standard error.

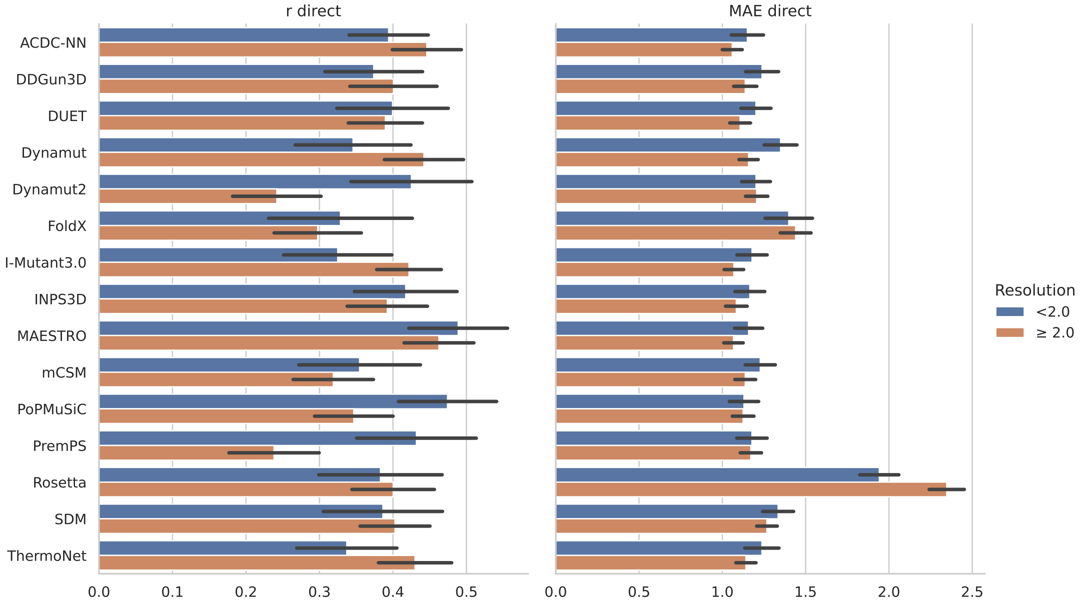

Figure A.3: **Effect on the method performances of the different X-ray resolution**. Evaluation is made on the direct variants in S669 whose structures were obtained by X-Ray diffraction. The dataset is split in two classes using 2.0 Årmstrong as a threshold for the resolution, with 177 variants in 34 proteins with resolution < 2.0 and 296 variants in 37 proteins with a resolution ≥ 2.0. Pearson correlation coefficients (r direct, on the left) and mean absolute error (MAE direct, on the right) are shown for all the structure-based methods. The black error bars represent the bootstrap estimated standard error.

A.0.3 Surface accessibility, pH and Temperature

As already observed in previous studies [125, 24, 32], the residue accessibility impacts the method performance. Fig.A.4 shows the results for the variants classified by their relative accessibility median value (Buried=[0-24%), Superficial=[24-100%]). Most predictors (even sequence-based) show much lower Pearson correlations on surface residues, with the exception of FoldX, and to a lower extent PremPS and INPS3D. However, the MAE, which measures the distance between the predicted and observed $\Delta\Delta G$ values, are lower (better) on the surface residues. This means that the methods are able to recognize that the surface residues have a lower impact on stability and coherently predict $\Delta\Delta G$ values closer to zero. However, when values are close to 0, the noise is higher, reducing the Pearson correlation performance.

Another very relevant point is to which extent the methods are affected by $\Delta\Delta G$ measures obtained outside physiological conditions. A recent paper [126] showed that there are some predictors in some extreme ranges of pH and temperature that decreases the performance. S669 dataset was divided into two parts: the former group containing variants whose temperature and pH are in physiological ranges $[293.15, 313.15]$ K (20-40 $°C$) and $[6.0, 8.0]$, respectively. This physiological group consists of 443 variants, while the non-physiological one of 226 variants. The results reported in Fig. A.5 show that there is not a clear indication of the fact that non-physiological conditions induce more errors in the predictions. The Pearson correlation is slightly better for variants in the group of physiological conditions; however, the MAE has an opposite trend (Fig. A.5). In the future, when a far larger set of clean data will hopefully be available, a more thorough study should be carried out.

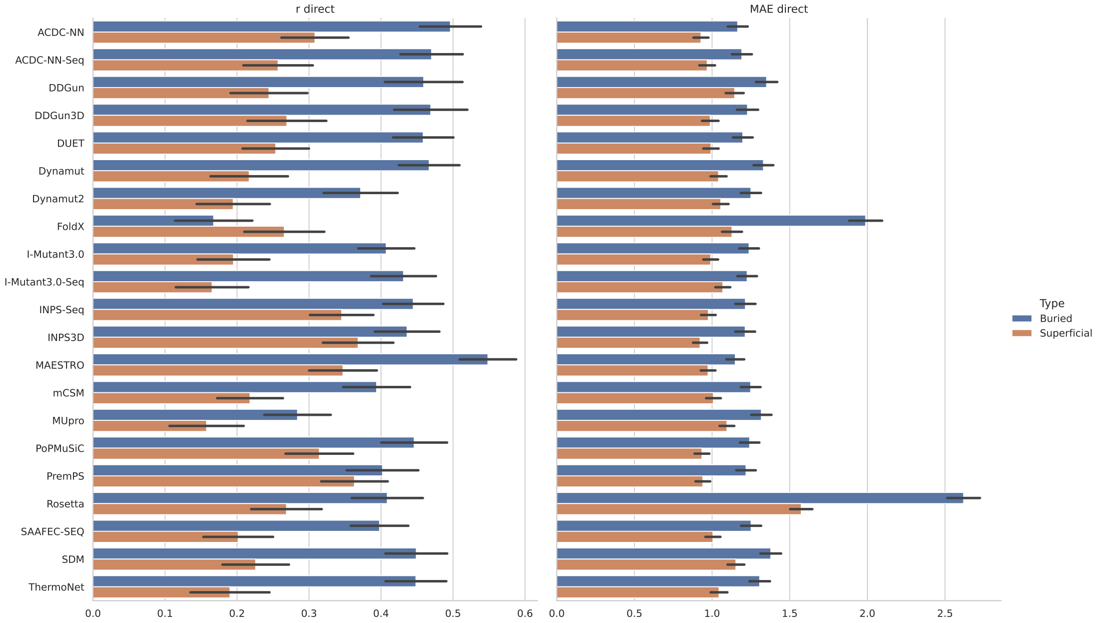

Figure A.4: **Assessment of the effects of the relative accessibility of an amino-acid on the prediction of the protein stability**. The effects of the relative accessibility (RA) are estimated by splitting the direct variants in the s669 dataset with respect to the RA median value (24%). Pearson correlation coefficients (r) and mean absolute errors (MAE direct) are displayed in the left and in the right plot, respectively. RA ranges from 0 to 1, with 0 representing a completely buried residue and 1 representing a residue on the surface. The black error bars represent the bootstrap estimated standard error.

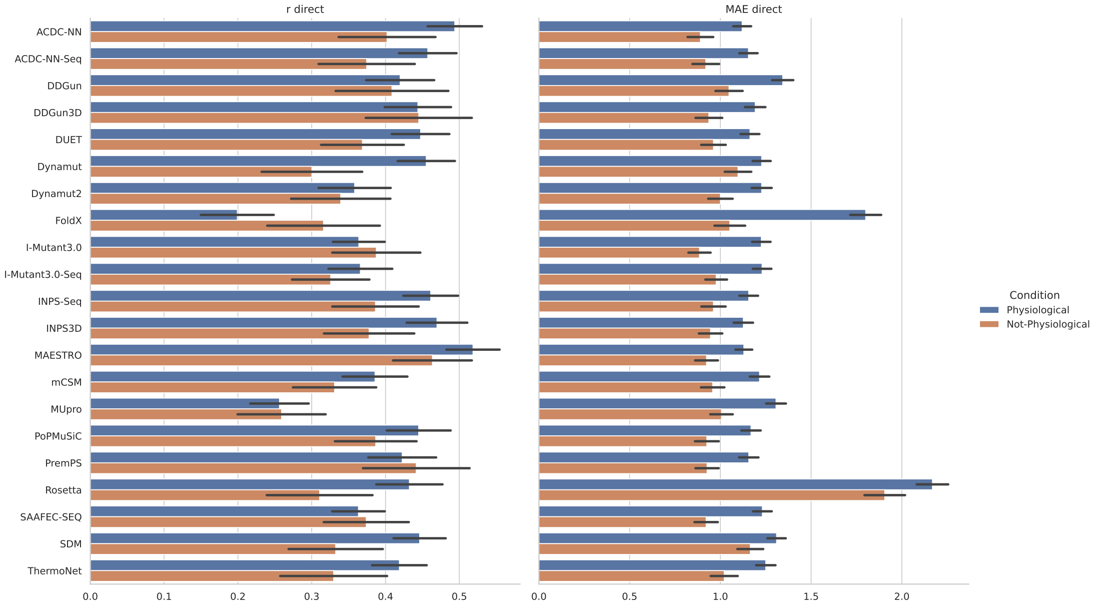

Figure A.5: **Assessment of the protein stability predictions tools on S669 at different temperature and pH conditions**. We compared all the prediction tools at physiological ($T \in [293.15, 313.15]$ K,$pH \in [6.0, 8.0]$, 443 variations) and not-physiological temperature and pH conditions (226 variations). After dividing the S669 dataset accordingly, the effects of different temperature and pH conditions were estimated by calculating the Pearson correlation coefficients (r) and mean absolute error (MAE) between predicted and real values on the two classes. These two measures are displayed in the left and in the right figure respectively. The black error bars represent the bootstrap estimated standard error.

Appendix B

Mutational signatures

B.1 Archetypal analysis

Scenario	SBS Signatures
1	3-5-25-40-89-92
2	10a-10c-10d-18-36
3	3-5-25-40-89-92-10a-10c-10d-18-36
4	6-10a-10c-10d-12-15-18-24-26-29-36
5	4-6-8-10a-10c-10d-12-15-18-19-23-24-26-29-36-37-39-86-92-94

Table B.1: **SBS Signatures used in each extraction scenario**

Signature 1	Signature 2	Cosine Similarity
SBS26	SBS12	0.93
SBS36	SBS18	0.91
SBS92	SBS5	0.88
SBS40	SBS3	0.88
SBS36	SBS10d	0.88
SBS10d	SBS10a	0.87
SBS15	SBS6	0.86
SBS29	SBS24	0.86
SBS10c	SBS10a	0.86
SBS10d	SBS10c	0.86
SBS94	SBS4	0.85
SBS40	SBS5	0.83
SBS29	SBS18	0.83
SBS37	SBS12	0.82
SBS8	SBS4	0.82
SBS23	SBS19	0.81
SBS86	SBS39	0.81
SBS36	SBS10c	0.81
SBS89	SBS3	0.81

Table B.2: **Pairwise cosine similarity (> 0.8) between COSMIC SBS Mutational Signatures**

Signature	**Flatness**
SBS3	0.87
SBS40	0.83
SBS5	0.78
SBS89	0.75
SBS25	0.74
SBS39	0.72
SBS94	0.64
SBS9	0.64
SBS92	0.64
SBS8	0.64

Table B.3: **The ten most flat signatures of COSMIC**

Profiles	**Cosine Similarity**	**MAE**
Archetypes	0.958 (0.958-0.959)	9.91 (9.89-9.92)
COSMIC SBS	0.967 (0.967-0.968)	6.20 (6.18-6.24)

Table B.4: **Cosine similarity and mean absolute error (MAE) between the original catalogues and the reconstructed ones, using MutationalPatterns as a refitting tool. The catalogues are those used in Scenario 1 with 500 samples. Performance is reported in terms of median and interquartile range over 10 runs.**

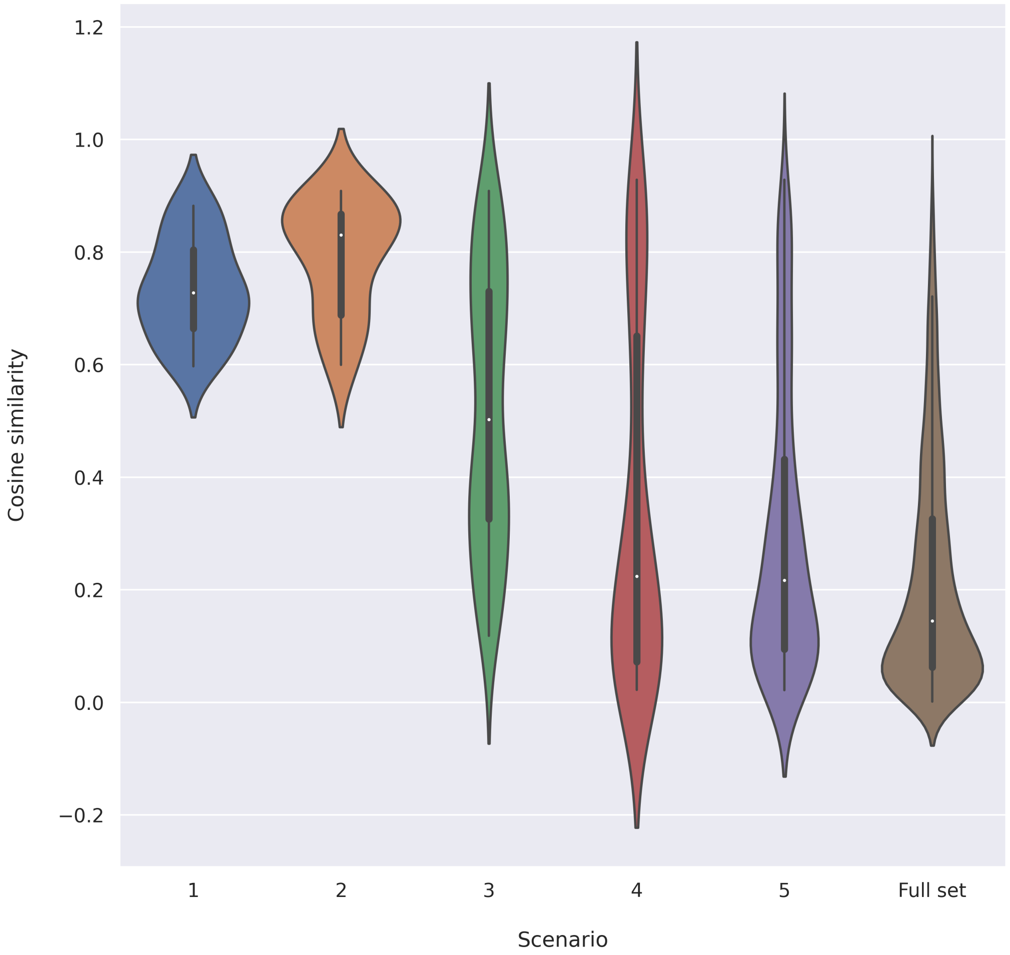

Figure B.1: **Cosine similarity distribution for each simulated scenario compared with the full set of non-artefactual COSMIC signatures.**

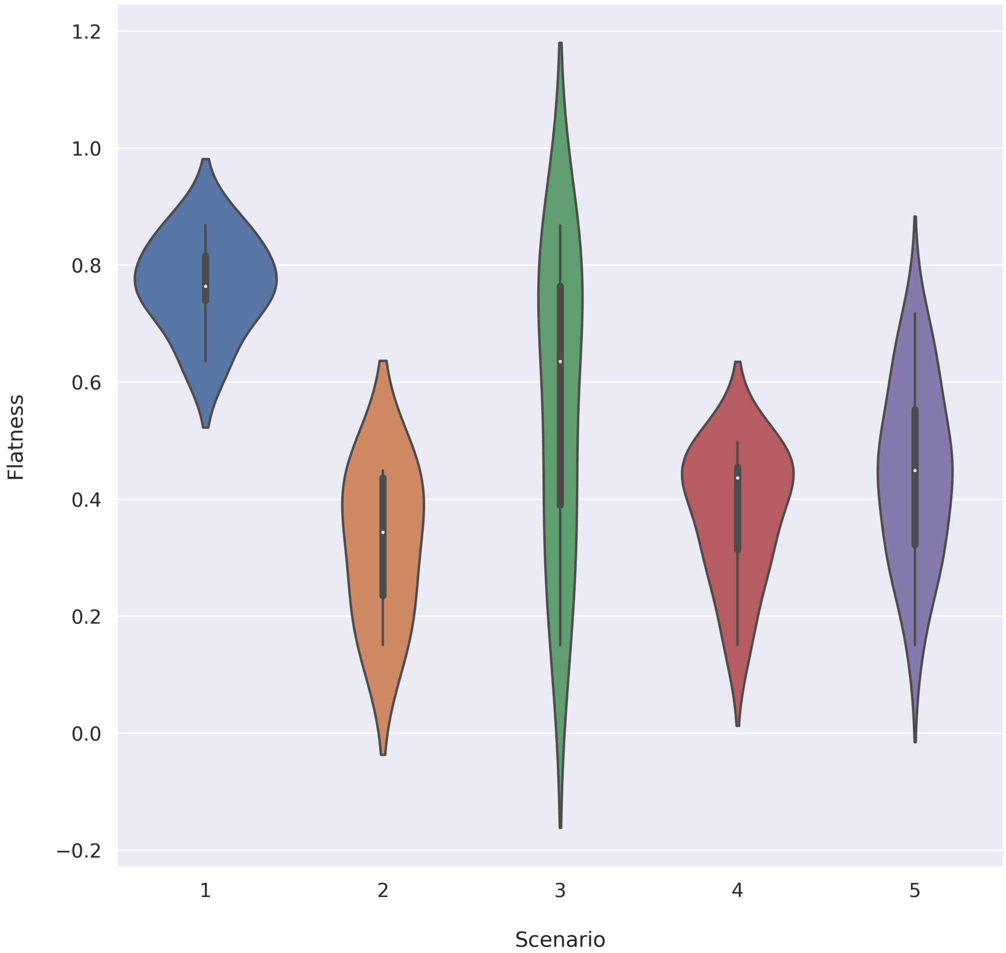

Figure B.2: **Flatness distribution for each simulated scenario.**

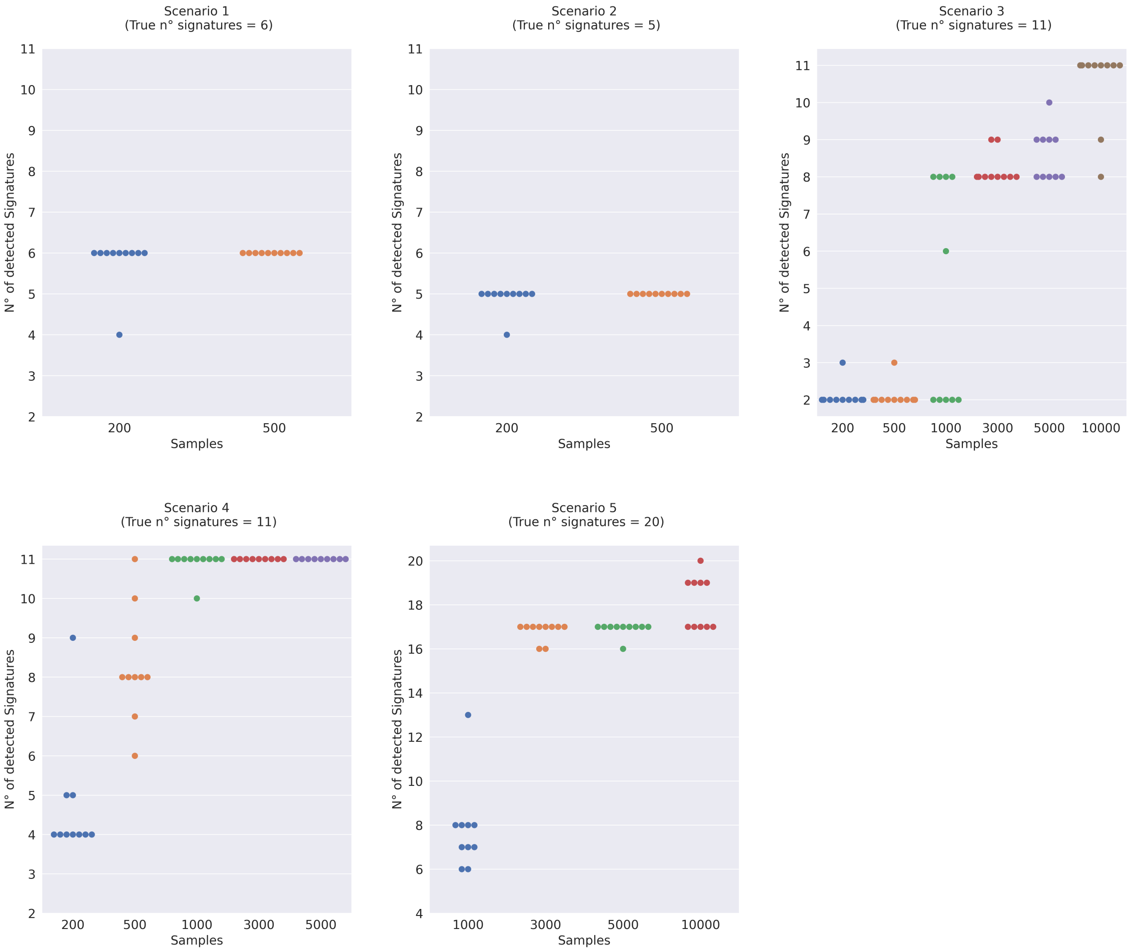

Figure B.3: **Number of detected signatures for each simulated scenario.** Each point at a fixed number of samples represents an individual run.

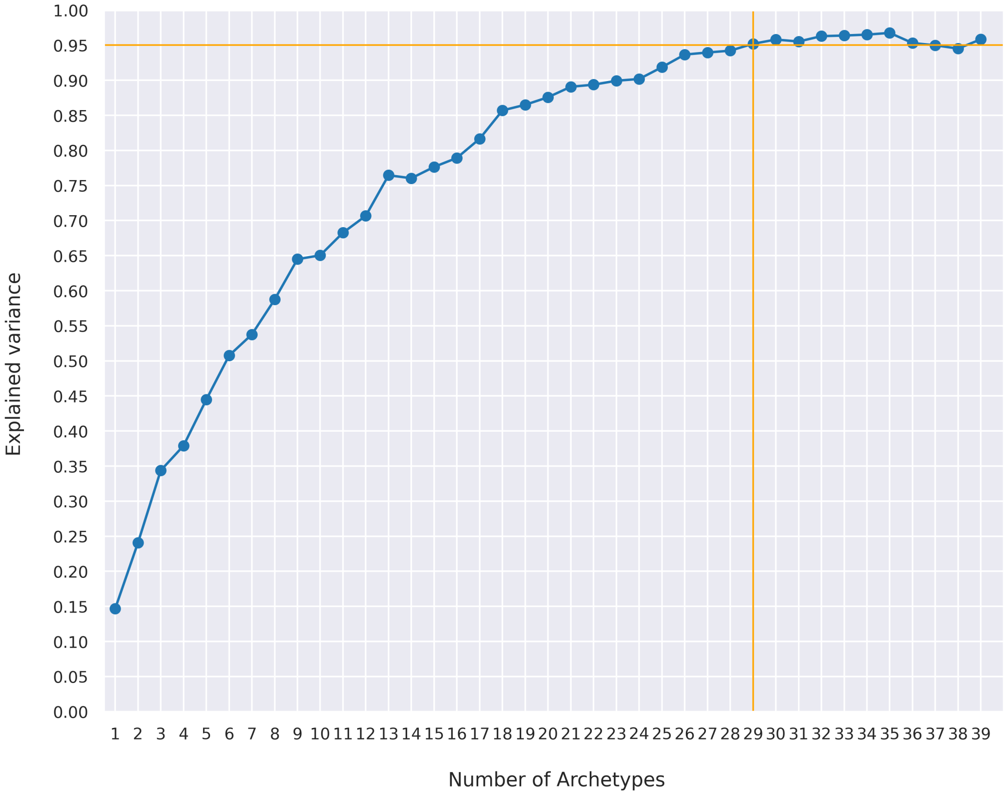

Figure B.4: **Explained variance with respect to the number of archetypes.**

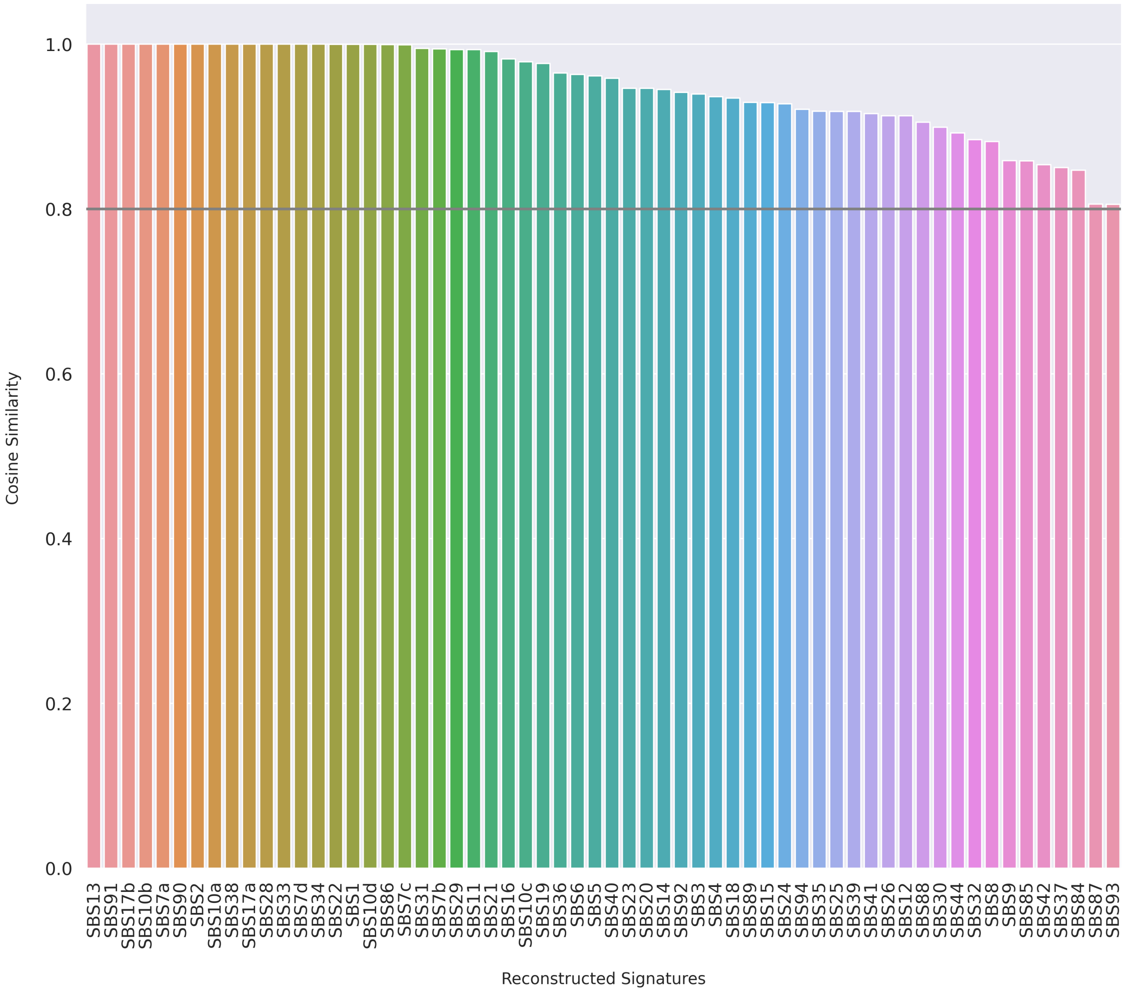

Figure B.5: **Cosine similarity between the original COSMIC signatures and the reconstructed ones through the 29 archetypes.**

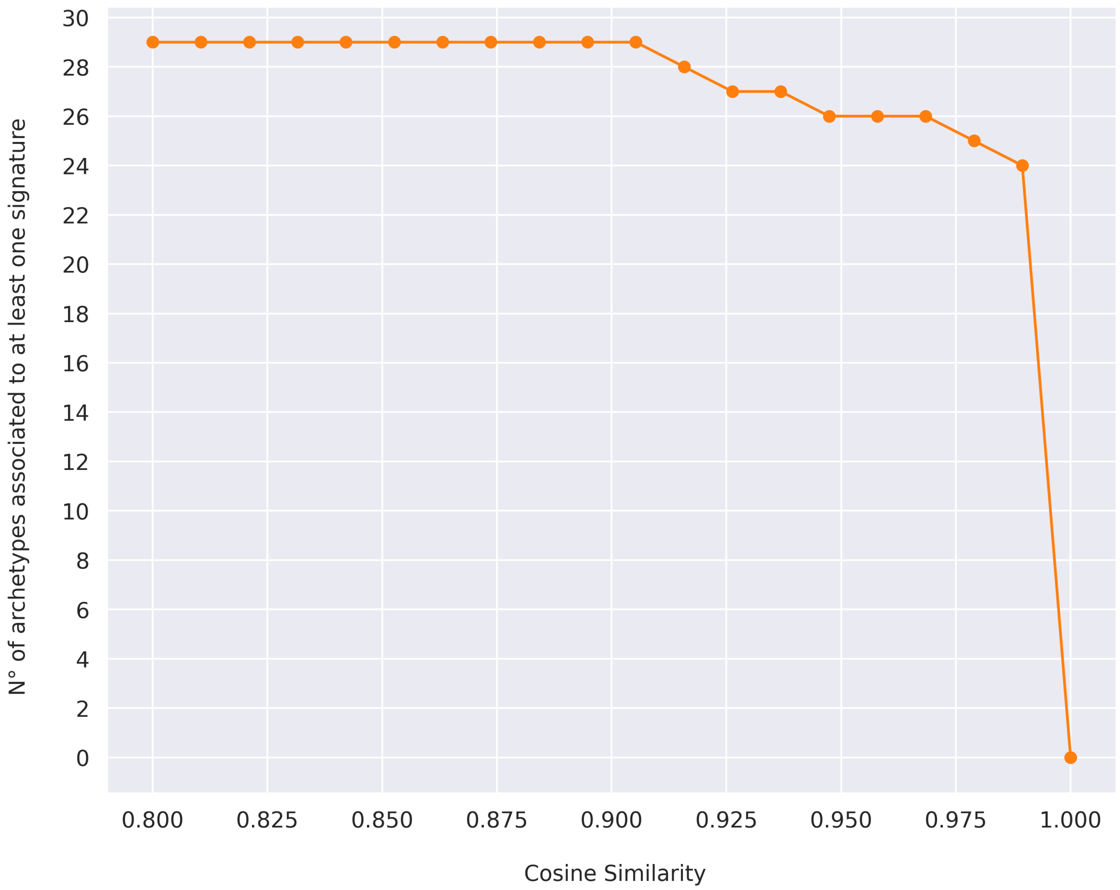

Figure B.6: **Relationship between archetypes and signatures at different cosine similarity thresholds.**

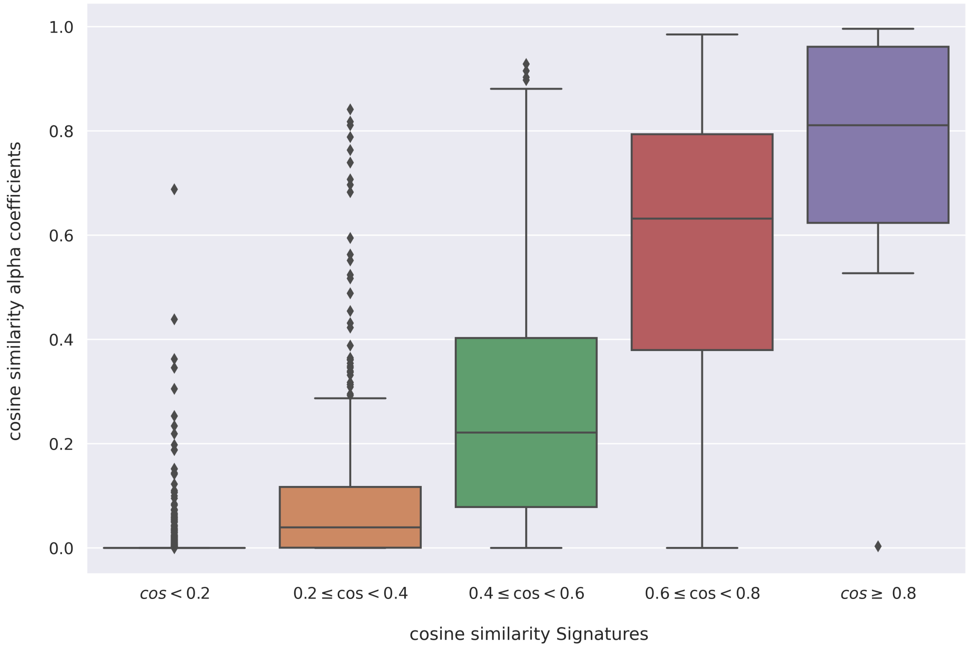

Figure B.7: **pairwise cosine similarity distribution of the alpha coefficients profiles for different levels of pairwise cosine similarity between the original mutational signatures**

Appendix C

MUSE-XAE

C.1 Evaluation metrics

Synthetic scenarios

For each synthetic scenario, we calculated metrics for precision, sensitivity, and the F1 Score. Specifically, we computed each metric by varying the cosine similarity threshold from 0.8 to 1.

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{Sensitivity} = \frac{TP}{TP + FN}$$

$$\textbf{F1-score} = 2 \cdot \frac{Precision \cdot Sensitivity}{Precision + Sensitivity}$$

For each of these metrics, we evaluated the score as a function of the cosine similarity threshold. We then computed their Area Under the Curve (AUC). Since the cosine similarity threshold ranges from 0.8 to 1 and the maximum AUC value is 0.2 (AUC_{max}), we normalized the AUC score as follows:

$$\textbf{AUC}_{norm} = \frac{AUC}{AUC_{max}}$$

This normalization allowed us to provide a more comparable and interpretable measure of performance across different methods and scenarios.

Real World datasets

In PCAWG and WGS extended cohort we used signatures exposures as new features which were fed into a Random Forest to classify both the primary sites and the cancer subtypes. We evaluated the performance in terms of balanced accuracy, Matthews Correlation Coefficient (MCC) and Kohen Kappa score in a 5-fold cross-validation setting.

- **Balanced Accuracy**, i.e. the average sensitivity obtained on each class:

$$\text{Balanced Accuracy} = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c}$$

where TP_c and FN_c are the number of true positives and false negatives for the c-th class, and C is the total number of classes.

- **Multi-class Matthews Correlation Coefficient**, defined as:

$$\text{MCC} = \frac{c \times s - \sum_{k=1}^{K}(p_k \times t_k)}{\sqrt{(s^2 - (\sum_{k=1}^{K} p_k)^2) \times (s^2 - (\sum_{k=1}^{K} t_k)^2)}}$$

where $t_k = \sum_{i=1}^{K} C_{ik}$ is the number of times the class k truly occurred, $p_k = \sum_{i=1}^{K} C_{ki}$ is the number of times the class k was predicted, $c = \sum_{k=1}^{K} C_{kk}$ is the total number of samples correctly predicted, and $s = \sum_{i=1}^{K} \sum_{j=1}^{K} C_{ij}$ is the total number of samples.

- **Cohen's Kappa**, which measures the agreement between two raters who each classifies N items into C mutually exclusive categories. The formula for the Kappa score is:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the relative observed agreement among the raters and p_e is the hypothetical probability of chance agreement.

SBS MUSE-XAE	COSMIC	Cosmic Similarity	SIGNAL	Signal Similarity
SBS A	SBS38	0.95	SBS33	0.92
SBS B	SBS54	0.81	SBS26	0.84
SBS C	SBS12	0.85	SBS13	0.88
SBS D	SBS23	0.93	SBS24	0.92
SBS E	SBS7a	0.97	SBS6	0.99
SBS F	SBS10a	0.90	SBS10	1.00
SBS G	SBS44	0.78	SBS34	0.80
SBS H	SBS8	0.88	SBS8	0.94
SBS I	SBS9	0.85	SBS9	0.96
SBS J	SBS1	1.00	SBS46	0.99
SBS K	SBS2	0.99	SBS1	0.99
SBS L	SBS39	0.89	SBS76	0.77
SBS M	SBS22a	0.99	SBS23	1.00
SBS N	SBS43	0.90	SBS49	0.60
SBS O	SBS36	0.97	SBS20	0.89
SBS P	SBS17b	0.93	SBS18	0.99
SBS Q	SBS5	0.65	SBS53	0.86
SBS R	SBS29	0.79	SBS3	0.83
SBS S	SBS92	0.90	SBS17	0.95
SBS T	SBS13	0.90	SBS14	0.99
SBS U	SBS34	0.78	SBS77	0.78
SBS V	SBS40b	0.86	SBS75	0.74

Table C.1: Pairwise cosine similarity between matched MUSE-XAE and COSMIC signatures and between MUSE-XAE and Signal ones for the PCAWG cohort

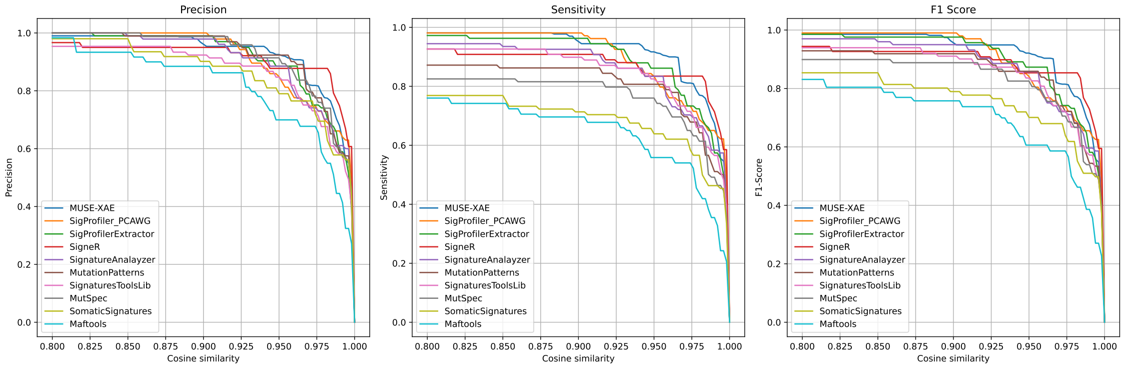

Figure C.1: Performance comparison between the top 10 performing methods. On the x-axis cosine similarity thresholds, while on y axes mean Precision, Sensitivity and F1 score across the five synthetic scenarios. Methods are ordered by the F1 score AUC.

SBS MUSE-XAE	**COSMIC**	**Cosmic Similarity**	**SIGNAL**	**Signal Similarity**
SBS A	SBS9	0.75	SBS9	0.90
SBS B	SBS13	0.92	SBS14	0.99
SBS C	SBS19	0.86	SBS24	0.92
SBS D	SBS26	0.89	SBS73	0.89
SBS E	SBS1	0.85	SBS46	0.86
SBS F	SBS57	0.91	SBS37	0.91
SBS G	SBS12	0.84	SBS17	0.82
SBS H	SBS17b	0.94	SBS18	0.99
SBS I	SBS36	0.91	SBS20	0.86
SBS J	SBS24	0.82	SBS25	0.77
SBS K	SBS7a	0.98	SBS6	0.99
SBS L	SBS39	0.91	SBS2	0.75
SBS M	SBS6	0.83	SBS0	0.90
SBS N	SBS34	0.84	SBS93	0.76
SBS O	SBS2	0.98	SBS1	0.97
SBS P	SBS32	0.67	SBS30	0.73
SBS Q	SBS43	0.97	SBS49	0.49
SBS R	SBS58	0.97	SBS111	0.57
SBS S	SBS8	0.75	SBS8	0.74
SBS T	SBS22a	0.99	SBS23	0.99
SBS U	SBS38	0.95	SBS33	0.92
SBS V	SBS44	0.96	SBS34	0.88
SBS W	SBS5	0.72	SBS4	0.75

Table C.2: Pairwise cosine similarity between matched MUSE-XAE and COSMIC signatures and between MUSE-XAE and Signal ones for the WGS extended cohort

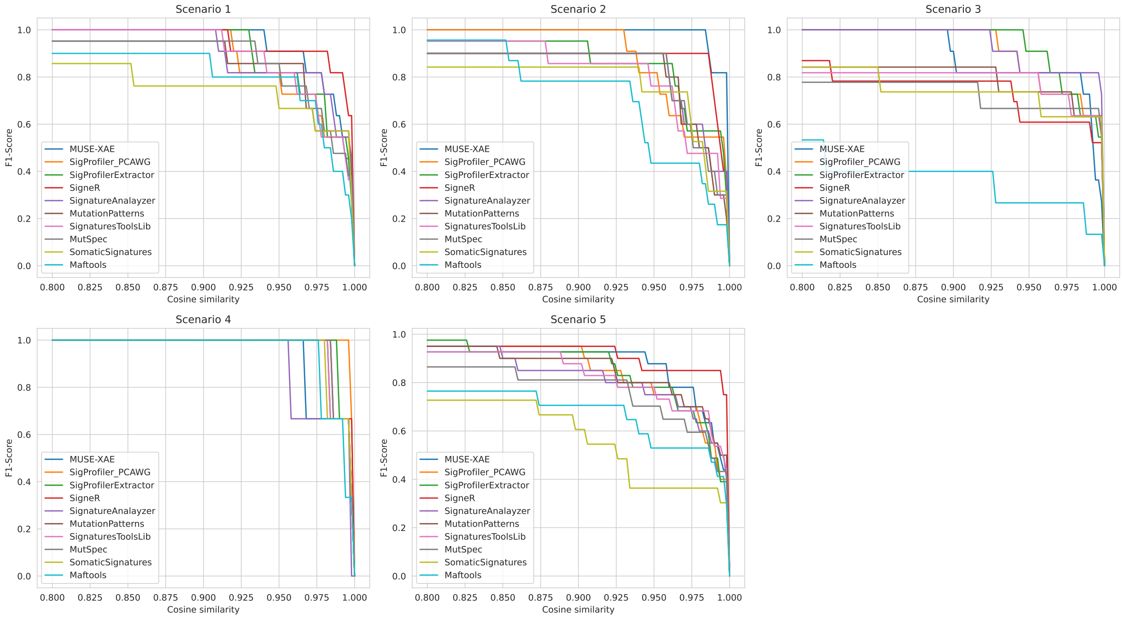

Figure C.2: Performance comparison between the top 10 performing methods. On the x-axis cosine similarity thresholds, while on y axis F1 score for each synthetic scenario. Methods are ordered by the F1 score AUC.

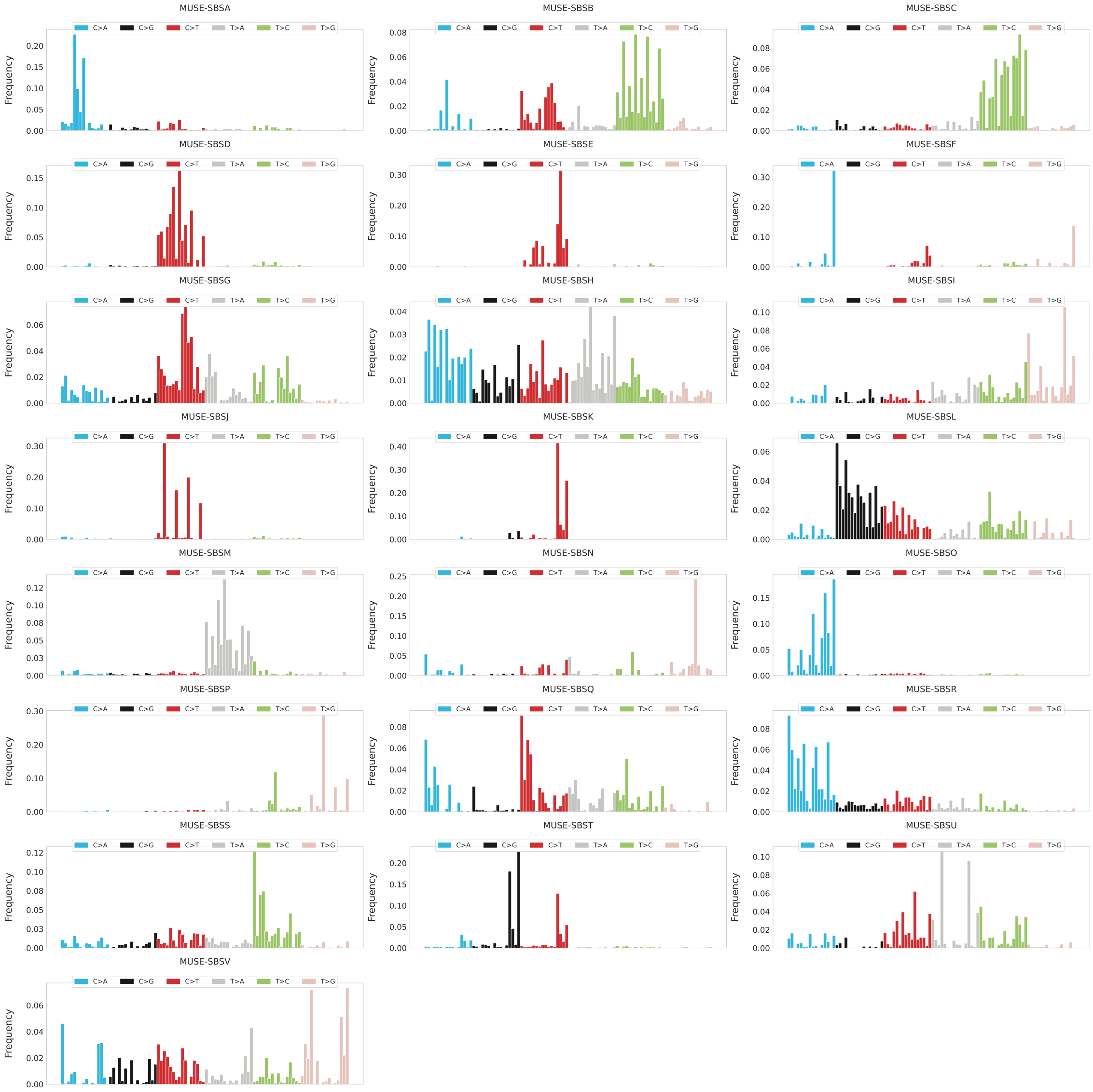

Figure C.3: 22 mutational signatures extracted from PCAWG dataset

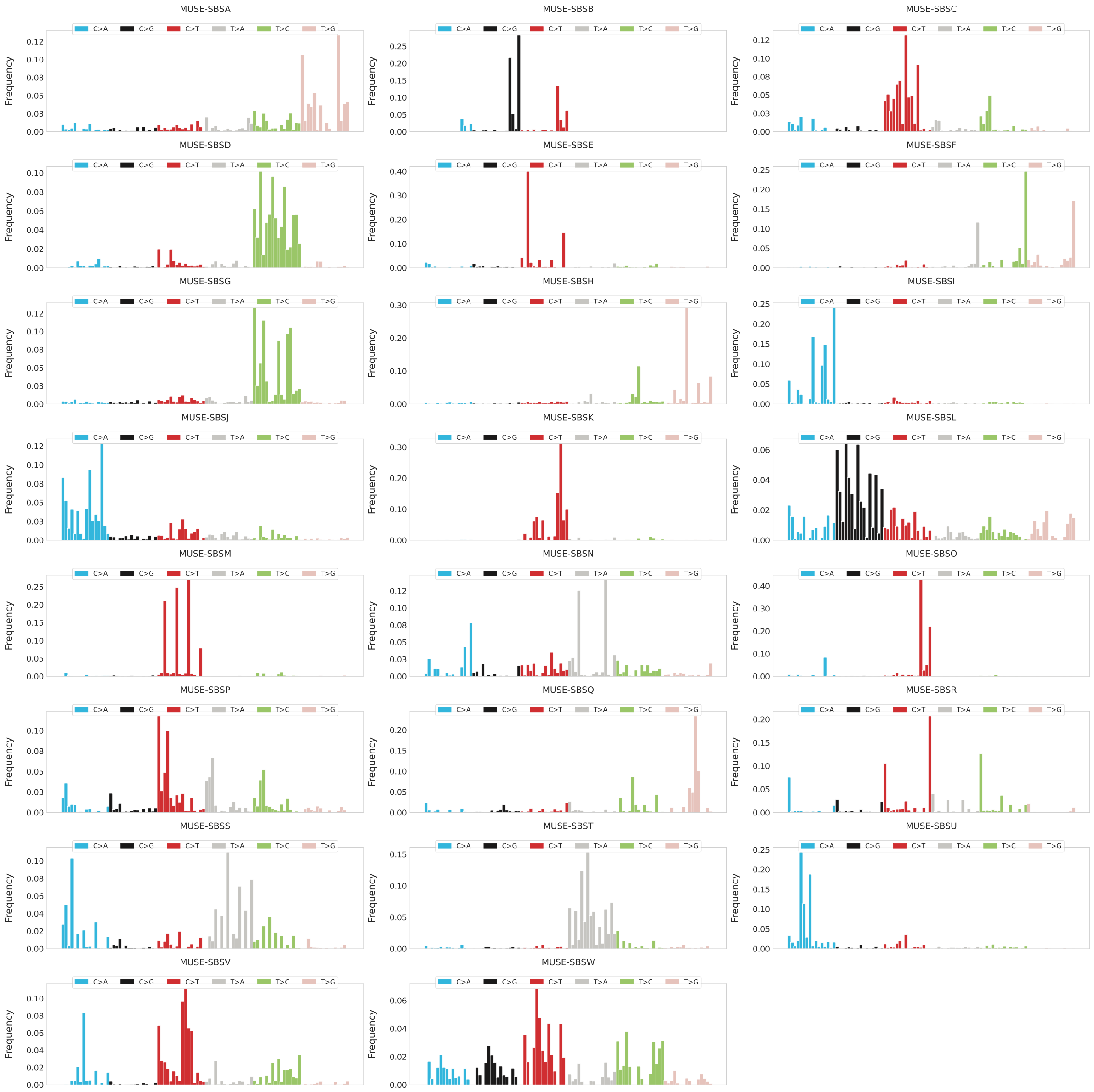

Figure C.4: 23 mutational signatures extracted from the extended cohort dataset

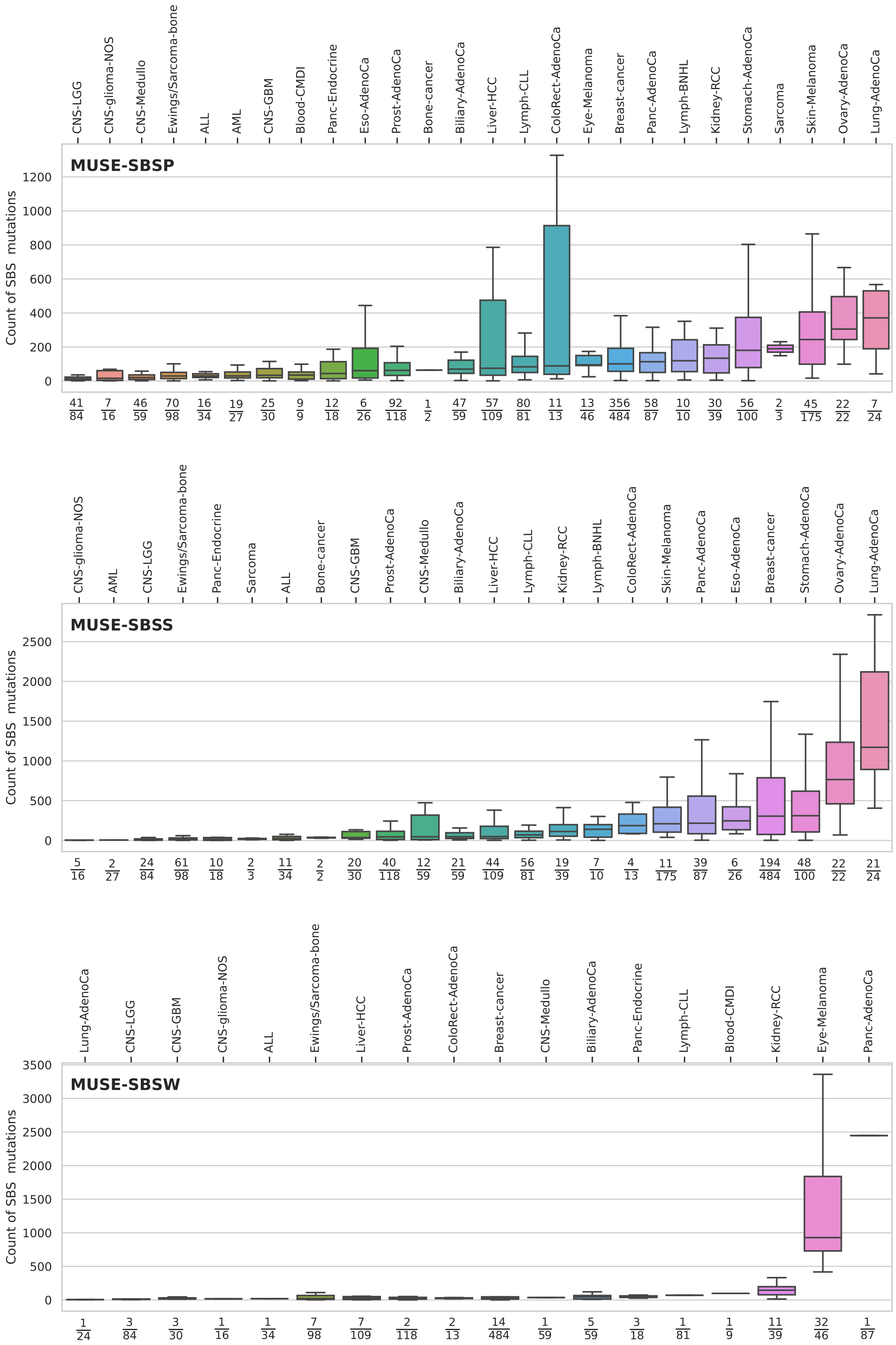

Figure C.5: SBS mutation counts for MUSE-SBSP, MUSE-SBSS and MUSE-SBSW for each tumour of the WGS extended cohort

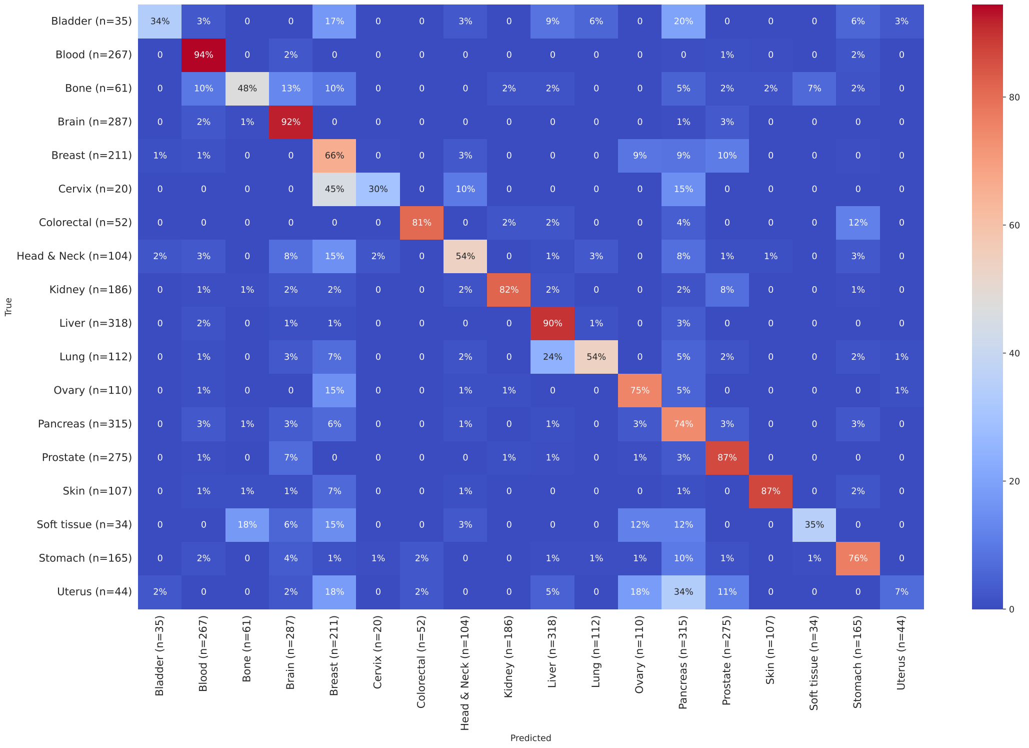

Figure C.6: MUSE-XAE confusion matrix for 18 PCAWG tumour primary sites

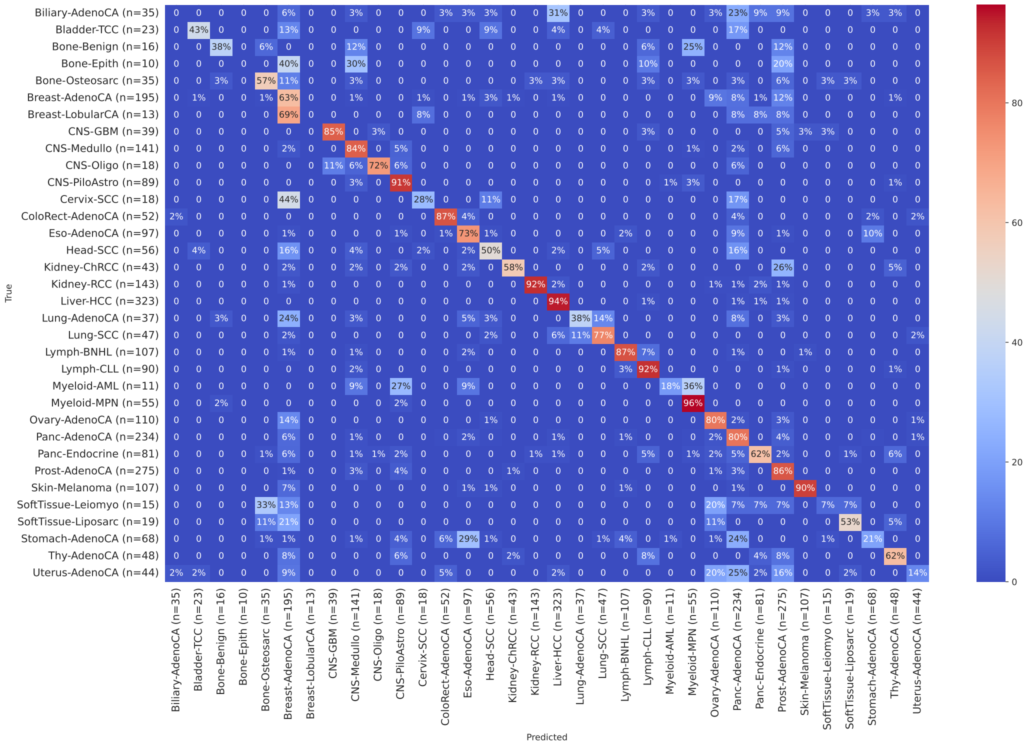

Figure C.7: MUSE-XAE confusion matrix for 34 PCAWG tumour subtypes

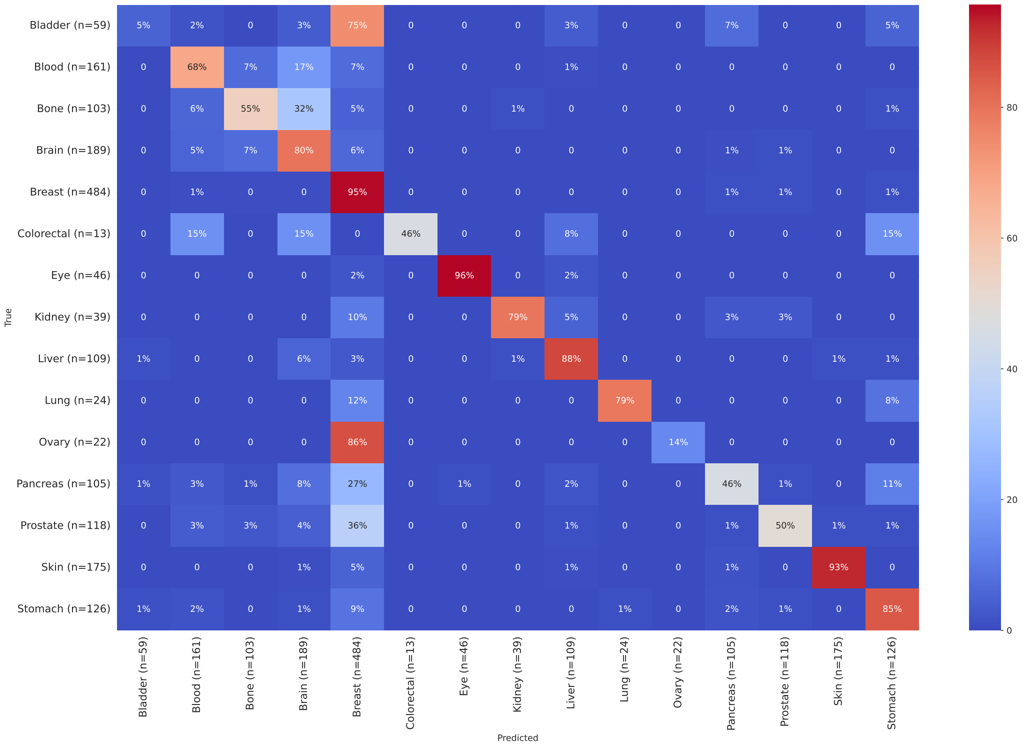

Figure C.8: MUSE-XAE confusion matrix for 15 WGS extended tumour primary sites

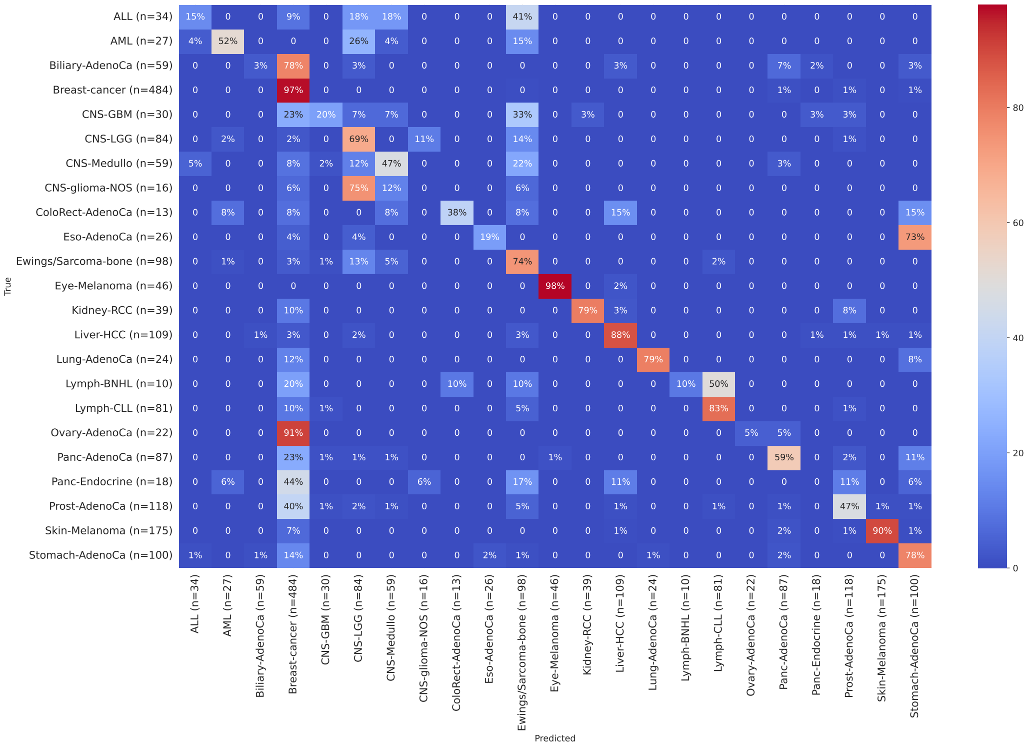

Figure C.9: MUSE-XAE confusion matrix for 23 WGS extended tumour subtypes

Bibliography

[1] Erwin L Van Dijk et al. “Ten years of next-generation sequencing technology”. In: *Trends in genetics* 30.9 (2014), pp. 418–426.

[2] Gerald Goh and Murim Choi. “Application of whole exome sequencing to identify disease-causing variants in inherited human diseases”. In: *Genomics & informatics* 10.4 (2012), p. 214.

[3] Ryan Hunt et al. “Silent (synonymous) SNPs: should we care about them?” In: *Single nucleotide polymorphisms* (2009), pp. 23–39.

[4] Peter D Stenson et al. “Human gene mutation database (HGMD®): 2003 update”. In: *Human mutation* 21.6 (2003), pp. 577–581.

[5] John G Tate et al. “COSMIC: the Catalogue Of Somatic Mutations In Cancer”. In: *Nucleic Acids Research* 47.D1 (Oct. 2018), pp. D941–D947. ISSN: 0305-1048. DOI: `10.1093/nar/gky1015`. eprint: `https://academic.oup.com/nar/article-pdf/47/D1/D941/27441712/gky1015.pdf`. URL: `https://doi.org/10.1093/nar/gky1015`.

[6] Tiziana Sanavia et al. “Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine”. In: *Computational and structural biotechnology journal* (2020).

[7] Tammy MK Cheng et al. “Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms”. In: *PLoS Comput Biol* 4.7 (2008), e1000135.

[8] M. N. Reza et al. “Pathogenic genetic variants from highly connected cancer susceptibility genes confer the loss of structural stability”. In: *Sci Rep* 11.1 (2021), p. 19264.

[9] L. Cheng et al. “Functional alterations caused by mutations reflect evolutionary trends of SARS-CoV-2”. In: *Brief Bioinform* 22.2 (Mar. 2021), pp. 1442–1450.

[10] Peng Yue, Zhaolong Li, and John Moult. “Loss of protein structure stability as a major causative factor in monogenic disease”. In: *Journal of molecular biology* 353.2 (2005), pp. 459–473.

[11] Giovanni Birolo et al. “Protein Stability Perturbation Contributes to the Loss of Function in Haploinsufficient Genes”. In: *Frontiers in Molecular Biosciences* 8 (2021), p. 10. ISSN: 2296-889X. DOI: `10.3389/fmolb.2021.620793`. URL: `https://www.frontiersin.org/article/10.3389/fmolb.2021.620793`.

[12] F Ulrich Hartl. “Protein misfolding diseases”. In: *Annual review of biochemistry* 86 (2017), pp. 21–26.

[13] Pier Luigi Martelli et al. “Large scale analysis of protein stability in OMIM disease related human protein variants”. In: *BMC genomics* 17.2 (2016), pp. 239–247.

[14] Mario Compiani and Emidio Capriotti. “Computational and theoretical methods for protein folding”. In: *Biochemistry* 52.48 (2013), pp. 8601–8624.

[15] Ludmil B Alexandrov et al. “Signatures of mutational processes in human cancer”. In: *Nature* 500.7463 (2013), pp. 415–421.

[16] Thomas Helleday, Saeed Eshtad, and Serena Nik-Zainal. “Mechanisms underlying mutational signatures in human cancers”. In: *Nature reviews genetics* 15.9 (2014), pp. 585–598.

[17] Gene Koh et al. “Mutational signatures: emerging concepts, caveats and clinical applications”. In: *Nature reviews cancer* 21.10 (2021), pp. 619–637.

[18] Mia Petljak and John Maciejowski. “Molecular origins of APOBEC-associated mutations in cancer”. In: *DNA repair* 94 (2020), p. 102905.

[19] Corrado Pancotti et al. “Unravelling the instability of mutational signatures extraction via archetypal analysis”. In: *Frontiers in Genetics* 13 (2023), p. 1049501.

[20] L. Montanucci et al. “A natural upper bound to the accuracy of predicting protein stability changes upon mutations”. In: *Bioinformatics* 35.9 (May 2019), pp. 1513–1517.

[21] S. Benevenuta and P. Fariselli. “On the Upper Bounds of the Real-Valued Predictions”. In: *Bioinform Biol Insights* 13 (2019), p. 1177932219871263.

[22] Emidio Capriotti et al. “A three-state prediction of single point mutations on protein stability changes”. In: *BMC bioinformatics* 9.2 (2008), pp. 1–9.

[23] Yashas BL Samaga, Shampa Raghunathan, and U Deva Priyakumar. “SCONES: Self-consistent neural network for protein stability prediction upon mutation”. In: *The Journal of Physical Chemistry B* 125.38 (2021), pp. 10657–10671.

[24] Yves Dehouck et al. “PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality”. In: *BMC bioinformatics* 12.1 (2011), pp. 1–12.

[25] Preethy Sasidharan Nair and Mauno Vihinen. “V ari B ench: a benchmark database for variations”. In: *Human mutation* 34.1 (2013), pp. 42–49.

[26] Aron Broom et al. “Computational tools help improve protein stability but with a solubility tradeoff”. In: *Journal of Biological Chemistry* 292.35 (2017), pp. 14349–14361.

[27] Fabrizio Pucci et al. “Quantification of biases in predictions of protein stability changes upon mutations”. In: *Bioinformatics* 34.21 (2018), pp. 3659–3665.

[28] J. S. Xavier et al. “ThermoMutDB: a thermodynamic database for missense mutations”. In: *Nucleic Acids Res* 49.D1 (Jan. 2021), pp. D475–D479.

[29] P. Sasidharan Nair and M. Vihinen. “VariBench: a benchmark database for variations”. In: *Hum Mutat* 34.1 (2013), pp. 42–49.

[30] Douglas EV Pires, David B Ascher, and Tom L Blundell. “mCSM: predicting the effects of mutations in proteins using graph-based signatures”. In: *Bioinformatics* 30.3 (2014), pp. 335–342.

[31] Kasper P Kepp. “Towards a “Golden Standard” for computing globin stability: Stability and structure sensitivity of myoglobin mutants”. In: *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1854.10 (2015), pp. 1239–1248.

[32] Octav Caldararu et al. “Systematic investigation of the data set dependency of protein stability predictors”. In: *Journal of Chemical Information and Modeling* 60.10 (2020), pp. 4772–4784.

[33] Ludovica Montanucci et al. “DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations”. In: *BMC bioinformatics* 20.14 (2019), p. 335.

[34] S Benevenuta et al. “An antisymmetric neural network to predict free energy changes in protein variants”. In: *Journal of Physics D: Applied Physics* 54.24 (2021), p. 245403.

[35] Corrado Pancotti et al. “A Deep-Learning Sequence-Based Method to Predict Protein Stability Changes upon Genetic Variations”. In: *Genes* 12.6 (2021), p. 911.

[36] Dinara R Usmanova et al. “Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation”. In: *Bioinformatics* 34.21 (2018), pp. 3653–3658.

[37] Jane Bromley et al. “Signature verification using a” siamese” time delay neural network”. In: *Advances in neural information processing systems* 6 (1993), pp. 737–744.

[38] Sumit Chopra, Raia Hadsell, and Yann LeCun. “Learning a similarity metric discriminatively, with application to face verification”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05).* Vol. 1. IEEE. 2005, pp. 539–546.

[39] Ugo Bastolla et al. “How to guarantee optimal stability for most representative structures in the protein data bank”. In: *Proteins: Structure, Function, and Bioinformatics* 44.2 (2001), pp. 79–96.

[40] Jeffrey Skolnick et al. “Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct?” In: *Protein science* 6.3 (1997), pp. 676–688.

[41] Steven Henikoff and Jorja G Henikoff. “Amino acid substitution matrices from protein blocks”. In: *Proceedings of the National Academy of Sciences* 89.22 (1992), pp. 10915–10919.

[42] Jack Kyte and Russell F Doolittle. “A simple method for displaying the hydropathic character of a protein”. In: *Journal of molecular biology* 157.1 (1982), pp. 105–132.

[43] L. Montanucci et al. “On the biases in predictions of protein stability changes upon variations: the INPS test case”. In: *Bioinformatics* 35.14 (July 2019), pp. 2525–2527.

[44] C. Savojardo et al. “On the critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation”. In: *Brief Bioinform* (2019).

[45] Bian Li et al. “Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks”. In: *PLOS Computational Biology* 16.11 (Nov. 2020). Ed. by Piero Fariselli, e1008291. DOI: `10.1371/journal.pcbi.1008291`. URL: `https://doi.org/10.1371/journal.pcbi.1008291`.

[46] C. Savojardo et al. “INPS-MD: a web server to predict stability of protein variants from sequence and structure”. In: *Bioinformatics* 32.16 (Aug. 2016), pp. 2542–2544.

[47] Catherine L Worth, Robert Preissner, and Tom L Blundell. “SDM—a server for predicting effects of mutations on protein stability and malfunction”. In: *Nucleic acids research* 39.suppl_2 (2011), W215–W222.

[48] Vijaya Parthiban, M Michael Gromiha, and Dietmar Schomburg. “CUPSAT: prediction of protein stability upon point mutations”. In: *Nucleic acids research* 34.suppl_2 (2006), W239–W242.

[49] Elizabeth H Kellogg, Andrew Leaver-Fay, and David Baker. “Role of conformational sampling in computing mutation-induced changes in protein structure and stability”. In: *Proteins: Structure, Function, and Bioinformatics* 79.3 (2011), pp. 830–838.

[50] Joost Schymkowitz et al. “The FoldX web server: an online force field”. In: *Nucleic acids research* 33.suppl_2 (2005), W382–W388.

[51] Josef Laimer et al. “MAESTROweb: a web server for structure-based protein stability prediction”. In: *Bioinformatics* 32.9 (2016), pp. 1414–1416.

[52] Douglas EV Pires, David B Ascher, and Tom L Blundell. “DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach”. In: *Nucleic acids research* 42.W1 (2014), W314–W319.

[53] Majid Masso and Iosif I Vaisman. “AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements”. In: *Protein Engineering, Design & Selection* 23.8 (2010), pp. 683–687.

[54] Chi-Wei Chen, Jerome Lin, and Yen-Wei Chu. “iStable: off-the-shelf predictor integration for predicting protein stability changes”. In: *BMC bioinformatics*. Vol. 14. S2. Springer. 2013, S5.

[55] Emidio Capriotti, Piero Fariselli, and Rita Casadio. “I-Mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure”. In: *Nucleic acids research* 33.suppl_2 (2005), W306–W310.

[56] Manuel Giollo et al. “NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation”. In: *BMC genomics* 15.4 (2014), pp. 1–11.

[57] Lijun Quan, Qiang Lv, and Yang Zhang. “STRUM: structure-based prediction of protein stability changes upon single-point mutation”. In: *Bioinformatics* 32.19 (2016), pp. 2936–2946.

[58] Piero Fariselli et al. “INPS: predicting the impact of non-synonymous variations on protein stability from sequence”. In: *Bioinformatics* 31.17 (2015), pp. 2816–2821.

[59] Jianlin Cheng, Arlo Randall, and Pierre Baldi. “Prediction of protein stability changes for single-site mutations using support vector machines”. In: *Proteins: Structure, Function, and Bioinformatics* 62.4 (2006), pp. 1125–1132.

[60] Gen Li, Shailesh Kumar Panday, and Emil Alexov. “SAAFEC-SEQ: A Sequence-Based Method for Predicting the Effect of Single Point Mutations on Protein Thermodynamic Stability”. In: *International Journal of Molecular Sciences* 22.2 (2021), p. 606.

[61] A. Sali and T. L. Blundell. “Comparative protein modelling by satisfaction of spatial restraints”. In: *J Mol Biol* 234.3 (1993), pp. 779–815.

[62] Silvia Benevenuta et al. “Challenges in predicting stabilizing variations: An exploration”. In: *Frontiers in Molecular Biosciences* 9 (2023), p. 1075570.

[63] Carlos HM Rodrigues, Douglas EV Pires, and David B Ascher. “DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability”. In: *Nucleic acids research* 46.W1 (2018), W350–W355.

[64] C. H. M. Rodrigues, D. E. V. Pires, and D. B. Ascher. “DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations”. In: *Protein Sci* 30.1 (Jan. 2021), pp. 60–69.

[65] Yuting Chen et al. “PremPS: Predicting the impact of missense mutations on protein stability”. In: *PLoS computational biology* 16.12 (2020), e1008543.

[66] Castrense Savojardo et al. “INPS-MD: a web server to predict stability of protein variants from sequence and structure”. In: *Bioinformatics* 32.16 (2016), pp. 2542–2544.

[67] Jeffrey Gagan and Eliezer M Van Allen. “Next-generation sequencing to guide cancer therapy”. In: *Genome medicine* 7.1 (2015), pp. 1–10.

[68] Kornelius Schulze et al. “Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets”. In: *Nature genetics* 47.5 (2015), pp. 505–511.

[69] Maria Secrier et al. “Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance”. In: *Nature genetics* 48.10 (2016), pp. 1131–1141.

[70] Gianmarco Contino et al. “Whole-genome sequencing of nine esophageal adenocarcinoma cell lines”. In: *F1000Research* 5 (2016).

[71] Andrea Degasperi et al. “A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies”. In: *Nature cancer* 1.2 (2020), pp. 249–263.

[72] Avantika Lal et al. “De novo mutational signature discovery in tumor genomes using SparseSignatures”. In: *PLoS computational biology* 17.6 (2021), e1009119.

[73] Rafael A Rosales et al. “signeR: an empirical Bayesian approach to mutational signature discovery”. In: *Bioinformatics* 33.1 (2017), pp. 8–16.

[74] Andrej Fischer et al. “EMu: probabilistic inference of mutational processes and their localization in the cancer genome”. In: *Genome biology* 14.4 (2013), pp. 1–10.

[75] Yuichi Shiraishi et al. “A simple model-based approach to inferring and visualizing cancer mutation signatures”. In: *PLoS genetics* 11.12 (2015), e1005657.

[76] Harald Vöhringer et al. “Learning mutational signatures and their multidimensional genomic properties with TensorSignatures”. In: *Nature communications* 12.1 (2021), p. 3628.

[77] SM Ashiqul Islam et al. “Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor”. In: *Cell Genomics* 2.11 (2022).

[78] Francis Blokzijl et al. “MutationalPatterns: comprehensive genome-wide analysis of mutational processes”. In: *Genome medicine* 10 (2018), pp. 1–11.

[79] Masroor Bayati et al. “CANCERSIGN: a user-friendly and robust tool for identification and classification of mutational signatures and patterns in cancer genomes”. In: *Scientific reports* 10.1 (2020), p. 1286.

[80] Maude Ardin et al. “MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes”. In: *BMC bioinformatics* 17 (2016), pp. 1–10.

[81] Andy G Lynch. "Decomposition of mutational context signatures using quadratic programming methods". In: *F1000Research* 5 (2016), p. 1253.

[82] Xiaoqing Huang, Damian Wojtowicz, and Teresa M Przytycka. "Detecting presence of mutational signatures in cancer with confidence". In: *Bioinformatics* 34.2 (2018), pp. 330–337.

[83] Sandra Krüger and Rosario M Piro. *Identification of mutational signatures active in individual tumors*. Tech. rep. PeerJ Preprints, 2017.

[84] Isidro Cortés-Ciriano et al. "Computational analysis of cancer genome sequencing data". In: *Nature Reviews Genetics* 23.5 (2022), pp. 298–314.

[85] Sally Bamford et al. "The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website". In: *British journal of cancer* 91.2 (2004), pp. 355–358.

[86] Alexandrov LB, Kim J, Haradhvala NJ, et al. "The repertoire of mutational signatures in human cancer". In: *Nature* 578.7793 (2020), pp. 94–101.

[87] Francesco Maura et al. "A practical guide for mutational signature analysis in hematological malignancies". In: *Nature communications* 10.1 (2019), p. 2969.

[88] NJ Haradhvala et al. "Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair". In: *Nature communications* 9.1 (2018), p. 1746.

[89] Nadezda V Volkova et al. "Mutational signatures are jointly shaped by DNA damage and repair". In: *Nature communications* 11.1 (2020), p. 2169.

[90] Paz Polak et al. "A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer". In: *Nature genetics* 49.10 (2017), pp. 1476–1486.

[91] Yoo-Ah Kim et al. "Network-based approaches elucidate differences within APOBEC and clock-like signatures in breast cancer". In: *Genome medicine* 12 (2020), pp. 1–12.

[92] Charles Swanton et al. "APOBEC enzymes: mutagenic fuel for cancer evolution and heterogeneity". In: *Cancer discovery* 5.7 (2015), pp. 704–712.

[93] Adele Cutler and Leo Breiman. "Archetypal analysis". In: *Technometrics* 36.4 (1994), pp. 338–347.

[94] Stephen C Johnson. "Hierarchical clustering schemes". In: *Psychometrika* 32.3 (1967), pp. 241–254.

[95] Franziska Schumann et al. "SigsPack, a package for cancer mutational signatures". In: *BMC bioinformatics* 20.1 (2019), pp. 1–9.

[96] Benyamin Motevalli Soumehsaraei and Amanda; Barnard. "Archetypal Analysis Package." In: *Commonwealth Scientific and Industrial Research Organisation (CSIRO).* (2019). DOI: 10.25919/5d3958889f7ff.

[97] Alberto Martin and Matthew D Scharff. "AID and mismatch repair in antibody diversification". In: *Nature Reviews Immunology* 2.8 (2002), pp. 605–614.

[98] Kimberly J Zanotti and Patricia J Gearhart. "Antibody diversification caused by disrupted mismatch repair and promiscuous DNA polymerases". In: *DNA repair* 38 (2016), pp. 110–116.

[99] Man Liu and David G Schatz. "Balancing AID and DNA repair during somatic hypermutation". In: *Trends in immunology* 30.4 (2009), pp. 173–181.

[100] Robert K Sylvester et al. "Temozolomide-induced severe myelosuppression: analysis of clinically associated polymorphisms in two patients". In: *Anticancer Drugs* 22.1 (2011), pp. 104–110.

[101] W R Connell et al. “Bone marrow toxicity caused by azathioprine in inflammatory bowel disease: 27 years of experience”. In: *Gut* 34.8 (1993), pp. 1081–1085.

[102] Liton Kumar Saha et al. “Topoisomerase I-driven repair of UV-induced damage in NER-deficient cells”. In: *Proc Natl Acad Sci U S A* 117.25 (2020), pp. 14412–14420.

[103] Han-Ming Shen et al. “Detection of elevated reactive oxygen species level in cultured rat hepatocytes treated with aflatoxin B1”. In: *Free Radical Biology and Medicine* 21.2 (1996), pp. 139–146.

[104] Yanan An et al. “Aflatoxin B1 induces reactive oxygen species-mediated autophagy and extracellular trap formation in macrophages”. In: *Frontiers in cellular and infection microbiology* 7 (2017), p. 53.

[105] Boyan Huang et al. “Aflatoxin B1 induces neurotoxicity through reactive oxygen species generation, DNA damage, apoptosis, and S-phase cell cycle arrest”. In: *International journal of molecular sciences* 21.18 (2020), p. 6517.

[106] Debasish Kumar Dey and Sun Chul Kang. “Aflatoxin B1 induces reactive oxygen species-dependent caspase-mediated apoptosis in normal human cells, inhibits Allium cepa root cell division, and triggers inflammatory response in zebrafish larvae”. In: *Science of the Total Environment* 737 (2020), p. 139704.

[107] Hans F. Stich and Fritz Anders. “The involvement of reactive oxygen species in oral cancers of betel quid/tobacco chewers”. In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 214.1 (1989). Special Issue Effects of Active Oxygen Species on the Genetic Apparatus, pp. 47–61. ISSN: 0027-5107. DOI: `https://doi.org/10.1016/0027-5107(89)90197-8`. URL: `https://www.sciencedirect.com/science/article/pii/0027510789901978`.

[108] Manashi Bagchi et al. “Role of reactive oxygen species in the development of cytotoxicity with various forms of chewing tobacco and pan masala”. In: *Toxicology* 179.3 (2002), pp. 247–255.

[109] Yuansi Chen, Julien Mairal, and Zaid Harchaoui. “Fast and robust archetypal analysis for representation learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1478–1485.

[110] Morten Mørup and Lars Kai Hansen. “Archetypal analysis for machine learning and data mining”. In: *Heurocomputing* 80 (2012), pp. 54–63.

[111] Vinayak Abrol and Pulkit Sharma. “A Geometric Approach to Archetypal Analysis via Sparse Projections”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 42–51.

[112] Christopher D Steele et al. “Signatures of copy number alterations in human cancer”. In: *Nature* 606.7916 (2022), pp. 984–991.

[113] David Heller et al. “SDip: A novel graph-based approach to haplotype-aware assembly based structural variant calling in targeted segmental duplications sequencing”. In: *BioRxiv* (2020). DOI: `https://doi.org/10.1101/2020.02.25.964445`.

[114] Miao L and Qi H. “Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization”. In: *IEEE Trans Geosci Remote Sens* 45.3 (2007), pp. 765–777.

[115] Valentine Svensson et al. “Interpretable factor models of single-cell RNA-seq via variational autoencoders”. In: *Bioinformatics* 36.11 (2020), pp. 3418–3421.

[116] Seninge L, Anastopoulos I, Ding H, et al. “VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics”. In: *Nature Commun* 12.1 (2021), p. 5684.

[117] “Pan-cancer analysis of whole genomes”. In: *Nature* 578.7793 (2020), pp. 82–93.

[118] Zheng H, Yang Z, Liu W, et al. “Improving deep neural networks using softplus units”. In: *2015 International joint conference on neural networks (IJCNN).* IEEE. 2015, pp. 1–4.

[119] Nebgen BT, Vangara R, Hombrados-Herrera M, et al. “A neural network for determination of latent dimensionality in non-negative matrix factorization”. In: *Machine Learning: Science and Technology* 2.2 (2021), p. 025012.

[120] Jonker R and Volgenant T. “A shortest augmenting path algorithm for dense and sparse linear assignment problems”. In: vol. 38. Springer, 1987, 325–340.

[121] Tate JG, Bamford S, Jubb HC, et al. “COSMIC: the catalogue of somatic mutations in cancer”. In: *Nucleic Acids Res* 47.D1 (2019), pp. D941–D947.

[122] Degasperi A, X Zou, Amarante TD, et al. “Substitution mutational signatures in whole-genome–sequenced cancers in the UK population”. In: *Science* 376.6591 (2022), abl9283.

[123] Corrado Pancotti et al. “Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset”. In: *Briefings in Bioinformatics* (Jan. 2022). bbab555. ISSN: 1477-4054. DOI: 10.1093/bib/bbab555. eprint: https://academic.oup.com/bib/advance-article-pdf/doi/10.1093/bib/bbab555/42244654/bbab555.pdf. URL: https://doi.org/10.1093/bib/bbab555.

[124] Octav Caldararu, Tom L Blundell, and Kasper P Kepp. “A base measure of precision for protein stability predictors: structural sensitivity”. In: *BMC bioinformatics* 22.1 (2021), pp. 1–14.

[125] Vladimir Potapov, Mati Cohen, and Gideon Schreiber. “Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details”. In: *Protein engineering, design & selection* 22.9 (2009), pp. 553–560.

[126] Shahid Iqbal et al. “Assessing the performance of computational predictors for estimating protein stability changes upon missense mutations”. In: *Briefings in Bioinformatics* (2021).