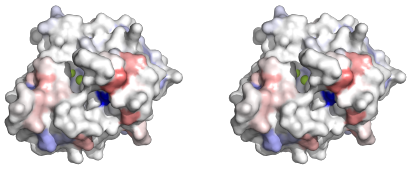
OLIVIER MAILHOT

PREDICTING BIOMOLECULAR FUNCTION FROM 3D DYNAMICS: SEQUENCE-SENSITIVE COARSE-GRAINED ELASTIC NETWORK MODEL COUPLED TO MACHINE LEARNING

PREDICTING BIOMOLECULAR FUNCTION FROM 3D DYNAMICS: SEQUENCE-SENSITIVE COARSE-GRAINED ELASTIC NETWORK MODEL COUPLED TO MACHINE LEARNING

OLIVIER MAILHOT



A framework for the efficient study of three-dimensional motions of RNA, proteins and drug-target complexes

Département de biochimie et médecine moléculaire
Faculté de médecine
Université de Montréal

August 2022 – classicthesis v4.6

Olivier Mailhot: *Predicting biomolecular function from 3D dynamics: sequence-sensitive coarse-grained elastic network model coupled to machine learning*, A framework for the efficient study of three-dimensional motions of RNA, proteins and drug-target complexes, © August 2022

SUPERVISORS:
Rafael Najmanovich
François Major

LOCATION:
Montréal

TIME FRAME:
August 2022

Identification des membres du jury

PRÉSIDENT:
Nazzareno D'Avanzo

DIRECTEUR:
François Major

CODIRECTEUR:
Rafael Najmanovich

MEMBRE:
Michelle Scott

EXAMINATEUR EXTERNE:
Giovanni Bussi

REPRÉSENTANTE DU DOYEN:
Pascale Legault

À Pierrot, qui m’a appris à quel point la vie est belle.

RÉSUMÉ

La dynamique structurelle des biomolécules est intimement liée à leur fonction, mais très coûteuse à étudier expériementalement. Pour cette raison, de nombreuses méthodologies computationnelles ont été développées afin de simuler la dynamique structurelle biomoléculaire. Toutefois, lorsque l'on s'intéresse à la modélisation des effects de milliers de mutations, les méthodes de simulations classiques comme la dynamique moléculaire, que ce soit à l'échelle atomique ou gros-grain, sont trop coûteuses pour la majorité des applications. D'autre part, les méthodes d'analyse de modes normaux de modèles de réseaux élastiques gros-grain (ENM pour "elastic network model") sont très rapides et procurent des solutions analytiques comprenant toutes les échelles de temps. Par contre, la majorité des ENMs considèrent seulement la géométrie du squelette biomoléculaire, ce qui en fait de mauvais choix pour étudier les effets de mutations qui ne changeraient pas cette géométrie. Le "Elastic Network Contact Model" (ENCoM) est le premier ENM sensible à la séquence de la biomolécule à l'étude, ce qui rend possible son utilisation pour l'exploration efficace d'espaces conformationnels complets de variants de séquence. La présente thèse introduit le pipeline computationel ENCoM-DynaSig-ML, qui réduit les espaces conformationnels prédits par ENCoM à des Signatures Dynamiques qui sont ensuite utilisées pour entraîner des modèles d'apprentissage machine simples. ENCoM-DynaSig-ML est capable de prédire la fonction de variants de séquence avec une précision significative, est complémentaire à toutes les méthodes existantes, et peut générer de nouvelles hypothèses à propos des éléments importants de dynamique structurelle pour une fonction moléculaire donnée. Nous présentons trois exemples d'étude de relations séquence-dynamique-fonction: la maturation des microARN, le potentiel d'activation de ligands du récepteur mu-opioïde et l'efficacité enzymatique de l'enzyme VIM-2 lactamase. Cette application novatrice de l'analyse des modes normaux est rapide, demandant seulement quelques secondes de temps de calcul par variant de séquence, et est généralisable à toute biomolécule pour laquelle des données expérimentale de mutagénèse sont disponibles.

Mots-clés: Dynamique structurelle · Analyse des modes normaux · Effet des mutations · Dynamique de l'ARN · Dynamique ligand-récepteur · Dynamique des protéines · Prédiction à haut débit de l'effet des variants

ABSTRACT

The dynamics of biomolecules are intimately tied to their functions but experimentally elusive, making their computational study attractive. When modelling the effects of thousands of mutations, time-stepping methods such as classical or enhanced sampling molecular dynamics are too costly for most applications. On the other hand, normal mode analysis of coarse-grained elastic network models (ENMs) provides fast analytical dynamics spanning all timescales. However, the vast majority of ENMs consider backbone geometry alone, making them a poor choice to study point mutations which do not affect the equilibrium structure. The Elastic Network Contact Model (ENCoM) is the first sequence-sensitive ENM, enabling its use for the efficient exploration of full conformational spaces from sequence variants. The present work introduces the ENCoM-DynaSig-ML computational pipeline, in which the ENCoM conformational spaces are reduced to Dynamical Signatures and coupled to simple machine learning algorithms. ENCoM-DynaSig-ML predicts the function of sequence variants with significant accuracy, is complementary to all existing methods, and can generate new hypotheses about which dynamical features are important for the studied biomolecule's function. Examples given are the maturation efficiency of microRNA variants, the activation potential of mu-opioid receptor ligands and the effect of point mutations on VIM-2 lactamase's enzymatic efficiency. This novel application of normal mode analysis is very fast, taking a few seconds CPU time per variant, and is generalizable to any biomolecule on which experimental mutagenesis data exist.

Key words: Structural dynamics · Normal mode analysis · Effect of mutations · RNA dynamics · Ligand-receptor dynamics · Protein dynamics · High-throughput variant effect prediction

No, no, you're not thinking; you're just being logical.

— Niels Bohr

ACKNOWLEDGMENTS

It is a difficult task to identify all people who positively influenced me while I was undertaking my doctorate studies. I strongly believe in chaos as a driving force of our lives, such that events or encounters seemingly unimportant can have a great impact on one's path. Nonetheless, I will try to acknowledge here the more direct contibutions to my scientific coming of age and reserve the use of French for friends and family.

François and Rafael, my two advisors, have had profound and complementary roles in turning me into the scientist I am today. Their common passion for computational structural dynamics shaped the way I think, and I feel blessed by the mix of freedom and guidance they both gave me.

I am thankful to all the colleagues with whom I have exchanged ideas over the years. Vincent, Francis and the other NRG members in Sherbrooke hold a special place as they were the ones who introduced me to structural bioinformatics. Similarly Paul, Mathieu, Nicolas and other Major lab members were great influences on my programming practice. Helping out newer members of both labs has also been a great source of joy and learning. Guillaume, Gabriel, Natalia, Thomas and others have all contributed to making my time at UdeM worthwhile.

A thesis would not be complete without a thesis jury. I am grateful to Giovanni, Michelle, Nazzareno and Pascale for being a part of the adventure.

Lastly, I want to thank Brian, John and the members of their labs at UCSF, who are my current colleagues to my great delight. Preparing my interview seminar around a year ago is what started me down the path to the main ideas presented in this thesis.

J'ai mis mon grain d'sel dans un grain d'sable.
J'ai mis le grain d'sable dans un château fort.
Mon château y'était fort mais y'a pas d'pont-levis.
Le royaume y vaut rien si t'as pas d'bons amis.

— Robert Nelson

REMERCIEMENTS

Je tiens d'abord à remercier mes parents, Marie et Frédéric, tant pour leur éternel support que pour l'inspiration qu'ils ont été et seront toujours pour moi. Dès mon plus jeune âge, mon père m'a transmis sa passion pour les sciences et ma mère sa fascination pour les mystères du vivant.

Florence est ma source de bonheur et de bien-être au quotidien. Elle a toléré avec amour mes nuits de travail et mes périodes de procrastination intensive à travers les années, sans mentionner mes nombreuses obsessions loufoques et passagères. Elle a mon amour et mon admiration les plus profonds.

Je remercie mon frère Antoine pour tous les mauvais plans que nous avons accomplis ensemble au fil des ans. Ceux qui sont surpris qu'on soit frères ne nous connaissent pas vraiment.

Sans Marlène, je n'aurais pas pu terminer l'écriture de ma thèse à temps pour mon départ à UCSF. Elle m'a prêté son chalet, nourri, écouté et supporté pendant le dernier droit et je lui en serai toujours reconnaissant.

On ne choisit pas sa famille, mais je choisirais la mienne si c'était à refaire. Mes oncles, mes tantes, mes cousins et cousines, mes grands-parents, ont tous été sources d'inspiration, d'amour et de fous rires.

Un merci bien spécial à Albert, sans qui je n'aurais pas la chance de passer autant de temps avec Florence tout en habitant à San Francisco.

L'amitié n'a pas de prix. Merci à Etienne, Justin, Jessica, Antoine, Audrey, Xavier, Maude, Vincent, Mathieu, Hélène, Richard, Gabriel, Benjamin, Alexis, Marie-Noëlle et bien d'autres de faire partie de ma vie.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

ANM — Anisotropic Network Model
ASA — Solvent accessible surface area
AU-PR — Area under the precision-recall curve
AU-ROC — Area under the receiver operating characteristic curve
CO — Cumulative overlap
CPU — Central processing unit
cryo-EM — Cryogenic electron microscopy
DMS — Deep mutational scan
DNA — Deoxyribonucleic acid
DOF — Degree of freedom
DynaSig — Dynamical Signature
ENCoM — Elastic Network Contact Model
ENM — Elastic Network Model
EntroSig — Entropic Signature
GA — Genetic algorithm
GNM — Gaussian Network Model
GPCR — G protein-coupled receptor
LASSO — Least absolute shrinkage and selection operator LOO — Leave-one-out
LOOCV — Leave-one-out cross-validation
MBL — Metallo-β-lactamase
MC — Monte Carlo
MD — Molecular dynamics
MFE — Minimum free energy

ML — Machine learning
MLP — Multilayer perceptron
MOR — μ-opioid receptor
mRNA — Messenger RNA
MSF — Mean square fluctuation
NA — Not applicable
NCO — Normalized cumulative overlap
ncRNA — Non-coding RNA
NMA — Normal mode analysis
NMR — Nuclear magnetic resonance
NRGTEN — Najmanovich research group toolkit for elastic networks
NRMSE — Negative root mean square error
nrt-PCA — Non-rotational-translational PCA
PCA — Principal component analysis
PC — Principal component
PDB — Protein Data Bank
PR — Precision-recall
QM/MM — Quantum mechanics/molecular mechanics
RISC — RNA-induced silencing complex
RMSD — Root mean square displacement
RMSE — Root mean square error
RNA — Ribonucleic acid
ROC — Receiver operating characteristic
RZSS — Rescaled Z-score sum
SNP — Single nucleotide polymorphism
STeM — Generalized spring tensor model
SVD — Singular value decomposition
TM — Transmembrane helix
VEP — Variant effect predictor
WT — Wild-type

Part I

BACKGROUND AND METHODS

1 INTRODUCTION

[1]

1.1 SEQUENCE, STRUCTURE, DYNAMICS, FUNCTION

Biomolecules are the fundamental building blocks of all living organisms. Like all matter subjected to high enough temperature, they *are in constant motion*. However, biomolecular motion is unique with regards to the large scale at which it happens and its importance for biological function. Thus, one necessary step in order to fully understand biomolecular function will be to completely characterize biomolecular motion, or structural dynamics. For now, that crucial step lies in the distant future. Structural dynamics are experimentally elusive, computationally expensive, and even the storage needed to represent the full conformational dynamics of a relatively large molecule is vaster than what is currently accessible. Furthermore, a given biomolecule may have tremendous numbers of theoretical mutations which do not affect its structure but *change its structural dynamics*. Predicting how mutations affect dynamics-function relationships is of interest to biomolecular engineering, molecular biology, virology, drug design, as well as many other life science disciplines. How can we predict the effect of vast numbers of theoretical mutations on biomolecular function through their influence on structural dynamics? To answer this question, the present work introduces the novel ENCoM-DynaSig-ML computational pipeline, composed of a sequence-sensitive coarse-grained elastic network model coupled to simple machine learning algorithms.

ENCoM-DynaSig-ML captures dynamics-function relationships, a capacity which we will illustrate in three diverse case studies: microRNA maturation, G protein-coupled receptor activation, and enzymatic activity of the VIM-2 lactamase. ENCoM-DynaSig-ML is fast enough to be trained on datasets of tens of thousands sequence variants and has been applied to predict theoretical maturation efficiencies of 30 million microRNA sequence variants. All parts of the pipeline are open-source and distributed as part of the user-friendly and extensively documented NRGTEN

1 This version of the thesis has had some PDF figures converted to bitmaps in order to respect Université de Montréal's guidelines. You can find an unpixelated version here: https://github.com/gregorpatof/omailhot_phd_thesis

and DynaSig-ML Python packages, thus lowering the barriers for the scientific community to use ENCoM-DynaSig-ML to study dynamics-function relationships.

1.1.1 *The central dogma*

The central dogma of molecular biology describes the transfer of information from a gene to its intended function in an organism: a DNA gene is transcribed to messenger RNA (mRNA); the mRNA is in turn translated to a protein; most proteins fold in distinct three-dimensional shapes determined by their sequence; the structure of a protein dictates its biological function based on its molecular interactions [1]. Thus, according to the dogma, there is a direct correspondence between the DNA sequence of a gene and its function. In addition to proteins, non-coding RNA (ncRNA) also functions in a manner that is dependent on its structure. Figure 1.1 shows this central dogma, with the inclusion of ncRNA.

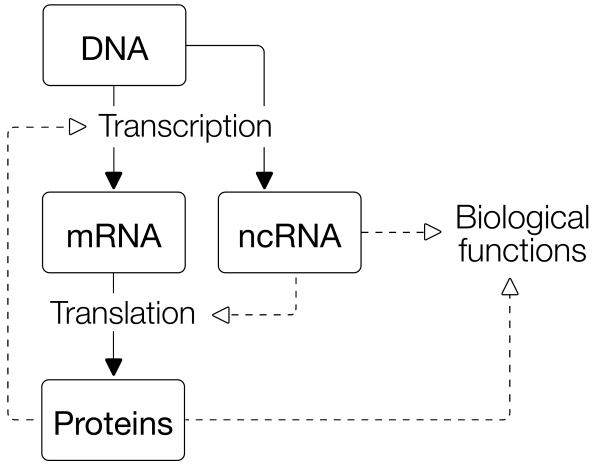


Figure 1.1: **Central dogma of molecular biology.** DNA is transcribed to RNA, which can be mRNA destined for translation into protein or non-coding RNA (ncRNA) exerting its own functions. The transcription is carried on by proteins (RNA polymerases) while ribosomes (of which ribosomal RNA is the catalytic element) are responsible for translation. Full arrows represent the transfer of information (transcription and translation) and dashed arrows represent the execution of a biomolecular function.

1.1.2 *Structure*

A biomolecule's structure is the three-dimensional (3D) configuration of its constituent atoms. The Thermodynamic Hypothesis in molecular biology, also called Anfinsen's dogma in honor of Nobel Prize laureate Christian Anfinsen, states that a protein's equilibrium structure in its physiological environment is simply the conformation minimizing the total Gibbs free energy of the system [2]. While this

statement may seem somewhat evident, when one considers the fact that even a modestly-sized protein has uncountable numbers of conformations, it becomes puzzling how a newly synthesized or denatured protein can "find" the proper conformation which minimizes Gibbs energy in the short timescales that are relevant to living cells. However, when considering the cellular environment (which contains chaperones that help the folding of some biomolecules), the Thermodynamic Hypothesis holds for most functional proteins, as they evolved to have defined structures in order to accomplish their functions, and thus also evolved funnel-shaped energy landscapes in order to guide their folding to the minimum energy conformation [3]. Figure 1.2 illustrates idealized and rugged protein folding funnels. The funnel-shaped energy landscape extends to structured RNA molecules, however the energy barrier required to break a base pair means that the landscape is more rugged and has deeper kinetic traps [4]. Some biomolecules, whether proteins or nucleic acids, do not possess this funnel-shaped energy landscape and thus the very concept of a defined structure does not apply to them. Around 30% of the human proteome is believed to be composed of intrinsically disordered proteins, which are defined as having more than 30% of their residues without defined structure [5]. Similarly, numerous functional RNA molecules exert their functions as result of their sequences and do not adopt defined 3D structures, for instance the coding regions of messenger RNA [6]. Since the present thesis is concerned with the computational study of structural dynamics happening around an equilibrium structure, disordered biomolecules will not be further discussed.

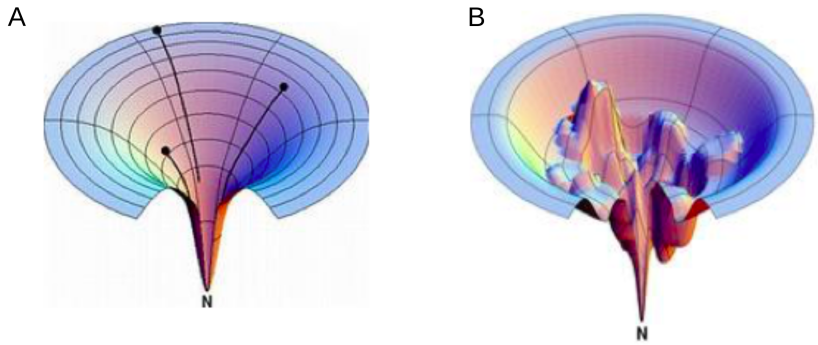


Figure 1.2: **Funnel-shaped energy landscapes explain protein folding.** A) An idealized funnel-shaped folding landscape. The three points represent suboptimal conformations in terms of Gibbs free energy, and the funnel shape guides them to the native conformation (N). B) A more realistic, slightly rugged folding funnel, with kinetic traps and energy barriers. The adoption of a single conformation of minimum free energy is still favored by this rugged funnel. The two images are authored by Ken A. Dill [3] and licensed under CC BY 4.0. They were not modified.

Biomolecular structure can be determined experimentally by techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) or cryo-electron mi-

croscopy (cryo-EM). Molecules with similar structures often have similar functions; structure can be used to classify protein and RNA families [7, 8]. For enzymes, the precise positioning of catalytic residues is crucial in determining what the activated state will be. The electronic surface of a receptor's binding site determines what ligands can interact with it. The primary transcripts of microRNAs have to form hairpin loops of approximately 35 base pairs to be recognized and processed by the Microprocessor complex. These are a few examples of how structure can relate to function and the list obviously goes on to include almost any functional biomolecule. In fact, structure also relates with function at the macroscopic scale: the shape of the human hand is ideal for tool use, the wings of an eagle for flight, etc. However, it is hard to properly describe function, both at macroscopic and molecular scales, if one ignores the contribution of dynamics.

1.1.3 *Structural dynamics*

In many cases, structure alone is not sufficient to understand function. Most molecular functions necessitate motion to be fully understood: the transformation of a substrate [9], the change of a conformation [10], the elongation of a polymer [11], etc. The whole landscape of possible conformations that a biomolecule occupies, along with their probabilities of occurring and the relationships between the different conformations, is referred to as structural dynamics. Structural dynamics arise from the combination of sequence, structure and context (pH, ionic strength, interaction partners, etc.). Since the equilibrium structure is described by structural dynamics, we can say that structure is encompassed by structural dynamics. A broader way of describing the sequence-function relationship is thus: sequence dictates structure; context and structure dictate structural dynamics; structural dynamics determine function.

Figure 1.3 gives two simplified examples of molecular functions that need dynamics to be explained. The first one is an enzyme of which the active site is not in the proper conformation to support substrate binding in the equilibrium structure. This activated conformation only exists as a tiny fraction of the ensemble population, but is favored upon substrate binding [12], a typical instance of conformational selection. The second example illustrates the processing of primary transcripts of microRNAs (pri-miRs). A specific position at the eighth base pair from the start of the hairpin has been identified as more flexible in pri-miRs that are efficiently recognized and cleaved by the Microprocessor. However, flexibility in that region does not affect the equilibrium structure of the pri-miR, but rather makes the opening of that specific base pair more probable [13, 14]. The Microprocessor recognizes the flexibility at the eighth base pair, which enhances the cleavage rate of the pri-miR. The cleavage happens about three base pairs higher, after the eleventh base pair from the 3'-end (2 nucleotides higher from the 5'-end) [15].

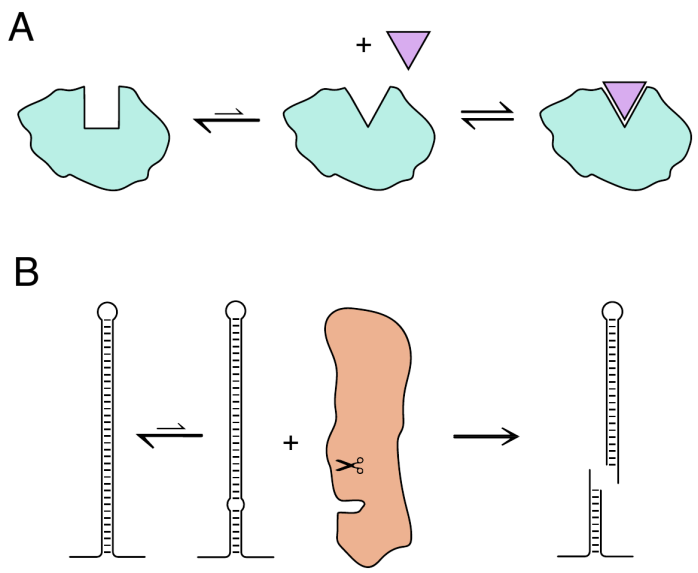


Figure 1.3: **Dynamics-function relationships in biomolecules.** A) At equilibrium, an enzyme (colored in teal) has its active site in a conformation incompatible with substrate binding The necessary conformational change happens with low probability but leads to substrate binding in the presence of the substrate (colored in purple). B) A microRNA primary transcript forms an equilibrium stem-loop structure of 35 base pairs. The Microprocessor enzyme (colored in orange) recognizes the flexibility at position 8, which is part of the signal that triggers the cleavage of the primary transcript 11 base pairs from its basal end.

Biomolecular functions are incredibly diverse, and novel functions are constantly being discovered as our understanding of biology grows. Since any measurable function of a biomolecule can be used as input to the ENCoM-DynaSig-ML pipeline, we do not define additional properties of function beyond its measurability.

1.1.4 *Biomolecular engineering*

Biomolecular engineering is the process of designing biomolecules with specific properties, from enhanced enzymatic activity to novel material properties. The first applications of biomolecular engineering used directed evolution to optimize the function of proteins, and it is still one of the most successful tools to design new biomolecular functions. In addition to directed evolution, rational and semi-rational design philosophies are now routinely employed to design DNA, RNA, as well as protein molecules with a wide range of applications [16]. Machine learning-aided protein design is also becoming very popular due to great successes in enzyme engineering. Usually, the machine learning (ML) algorithms are trained on either

sequence or structure data in relation to protein stability, solubility or catalytic efficiency [17].

1.1.5 *The problem*

In the last decades, sequence data has been generated at an exponentially growing rate thanks to next-generation sequencing technologies. Our capacity to solve experimental structures has also hugely grown, but at a much slower pace. The advent of high precision tools for the computational prediction of protein stuctures such as AlphaFold holds the promise to fill the gap between sequence and stucture. How about structural dynamics? From an experimental point of view, they require more work to capture than structure alone and are often impossible to study under physiological conditions. Most importantly, there are no actual or foreseeable experimental techniques capable of studying the structural dynamics from vast numbers of sequence variants of a given biomolecule. Thus, if the solution to this problem exists, it must be computational in nature. However, the best 3D structure prediction tools do not consider dynamics, instead outputting one or a few high-confidence equilibrium structure(s). Computational tools to simulate 3D structural dynamics can be split in three broad categories:

1. Time-stepping methods, whether all-atom or coarse-grained, commonly called molecular dynamics (MD) simulations [18].

2. Sampling techniques such as Monte Carlo (MC) methods [19].

3. Normal mode analysis (NMA) methods, which give analytical solutions spanning all timescales but make strong assumptions about the energy landscape [20].

MD simulations represent powerful tools for the study of detailed structural dynamics. However, even the most simple of these methods require computational time proportional to the desired timescale at which structural dynamics are studied [21]. This linear dependency on time makes all time-stepping methods, whether all-atom or coarse-grained, too costly for the study of slow-timescales dynamics happening in vast numbers of sequence variants.

Monte Carlo (MC) methods are based on randomly perturbing a biomolecule's conformation to obtain new conformations, assigning energies to them, keeping conformations with low enough energy as probable conformations and repeating the procedure [22]. MC methods do not explicitly consider time as MD methods do, hence they are more likely to sample conformational changes that need to cross large energy barriers to happen. One issue with both MC and MD methods is that they are ill-conditioned, meaning that different runs on the same input may give rise to vastly different solutions. It is thus commonplace to perform replicate

simulations in order to ensure convergence has been reached, further increasing the computational cost of both techniques.

NMA methods, on the other hand, provide information on biomolecular dynamics spanning all timescales at a fixed computational cost. This decoupling between computational cost and simulation timescale is made possible by assuming a harmonic energy well for the energy landscape of the biomolecule, with the input conformation at the bottom (Figure 1.4). With NMA, the molecule's conformational space is described with a set of normal modes, each one representing a harmonic, oscillatory motion around the equilibrium structure. The normal modes are ordered according to their associated frequency, from lowest to highest. The slowest modes correspond to the largest deformations of the molecule, which are also the most collective motions and have been shown to capture known biomolecular motions surprisingly well, for example the opening and closing of the citrate synthase active site [23]. Moreover, the computational cost can be further reduced by coarse-graining the studied molecule, for example with the usage of a single bead per amino acid, and such coarse-grained NMA models still capture well the slow biologically relevant motions [24]. Coarse-grained NMA thus presents an appealing alternative to time-stepping and Monte Carlo methods for the study of structural dynamics. However, popular coarse-grained NMA models consider backbone geometry alone in their potential function and are consequently insensitive to the effect of mutations which do not alter the backbone geometry [25, 26]. To overcome this, the Elastic Network Contact Model (ENCoM) has been introduced by Frappier and Najmanovich in 2014 as the first sequence-sensitive coarse-grained NMA model [27].

From the set of normal modes and their associated frequencies, thermodynamical properties such as vibrational entropy can be efficiently obtained. The difference in vibrational entropy upon mutation (ΔS_{vib}) computed with ENCoM was shown to correlate with experimentally measured $\Delta\Delta G$ of folding. Moreover, it can be used to classify homologous proteins from thermophile and mesophile organisms [28]. However, ΔS_{vib} is a single value measuring the molecule's overall rigidity and mutations can affect structural dynamics locally without changing this global property. We hypothesize that considering the flexibility change at every position in the molecule has the potential to capture finer dynamics-function relationships. Our study of these flexibility changes in the context of mutations on the Spike protein from SARS-CoV-2 can be seen as the first step towards validating that hypothesis [29].

1.1.6 *The solution*

This thesis proposes to bridge the knowledge gap at the interface between sequence, structure and structural dynamics with the introduction of the ENCoM-DynaSig-

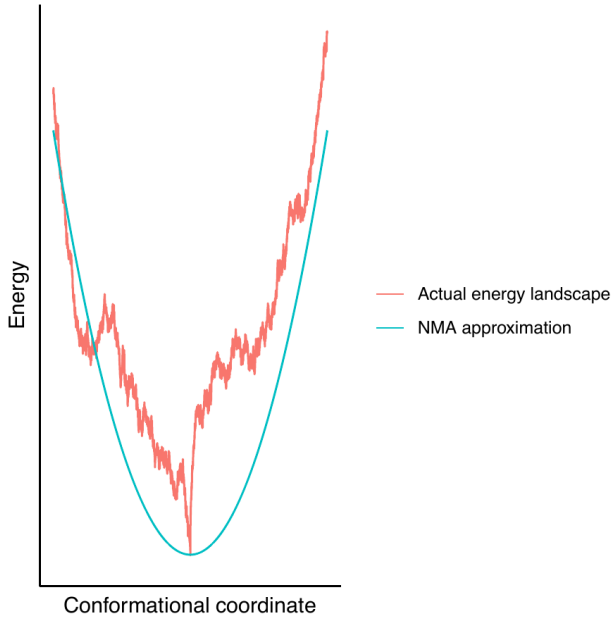


Figure 1.4: **The NMA energy landscape.** The energy of a fictional biomolecule is plotted against a varying arbitrary conformational coordinate. The real energy landscape is rugged, with many local minima, while the normal mode analysis (NMA) approximation has the input conformation at the bottom of a quadratic energy well.

ML computational pipeline. The pipeline starts with the sequence-sensitive elastic network ENCoM, which is used to compute the whole conformational space of the studied sequence variants. These conformational spaces are then reduced to Dynamical Signatures, which are vectors of flexibility at every position. The Dynamical Signatures are used to train simple machine learning algorithms with experimentally measured functional data. Our answer to the problem has the following characteristics:

1. It is fast enough to predict structural dynamics for sequence variants in an ultra-high-throughput manner in reasonable computational time.

2. It gives dynamical information about all timescales in a single analytical solution.

3. It is applicable to RNA, protein, DNA, small molecules and their complexes.

4. The model learns fine dynamical characteristics necessary for biomolecular function from experimental data. These can be mapped back to the sequence and structure, enabling both the generation of new hypotheses about the biology of the studied molecules and the prediction of novel sequences with desired properties.

1.2 THE ENCOM-DYNASIG-ML PIPELINE

1.2.1 *Coarse-grained elastic network models*

The study of protein dynamics using normal mode analysis (NMA) of coarse-grained elastic network models (ENMs) dates back to the publication of Monique Tirion's seminal 1996 article [24]. Tirion showed that slow vibrational modes of globular proteins can be captured with normal mode analysis of a simplified, single-parameter potential as opposed to the complex semi-empirical potentials that were previously the norm. Coarse-grained ENMs allow the fast analytical study of biomolecular dynamics spanning all timescales and even very simple models such as ANM and GNM show good correlations with experimental data [25, 26]. The perhaps surprising performance of these simple models is explained by the fact that geometry has a large role to play in dynamics [30].

1.2.2 *The Elastic Network Contact Model*

Coarse-grained elastic network models (ENMs) capture large-scale dynamics well at low computational cost, however they are based on the studied biomolecule's backbone geometry alone. This property prevents their use to predict the effect of point mutations on dynamics if such mutations do not change the backbone geometry and consequently the position of the beads in the system. In order to capture the effects of such backbone geometry preserving mutations within a coarse-grained ENM, Frappier and Najmanovich introduced the Elastic Network Contact Model (ENCoM) in 2014 [27]. ENCoM is the first and only coarse-grained ENM that is sensitive to the sequence of biomolecules through the inclusion of non-bonded interaction terms between all atoms in its potential function. This inclusion still allows for coarse-graining, so the computational cost of ENCoM is comparable to that of the simplest coarse-grained ENMs, enabling high-throughput *in silico* prediction of the effect of mutations on structural dynamics. For example, computing the entire set of normal modes for a 250 amino acid protein takes around 3 seconds on a single modern CPU (single core from AMD Rome 7532 @ 2.40 GHz). The whole set of normal modes in turn allows the generation of any conformation as a linear combination of the modes. As part of the present work, ENCoM has been extended to RNA, has had new metrics added including the Entropic Signature described below, has been re-parameterized on a diverse benchmark of experimentally measured biomolecular dynamics and is now distributed through the user-friendly NRGTEN Python package [31].

1.2.3 *Dynamics-function relationships with DynaSig-ML*

ENMs provide fast computational access to structural dynamics, however the predicted conformational space represents a huge linear space. For example, the entire solution for a 250 amino acid protein comprises 744 normal modes, each a vector of length 750, for a total of 558 000 scalar values. In order to perform statistical analyses and/or machine learning using ENM-predicted dynamics, these huge linear spaces need to be reduced. One concept which is explored extensively in this thesis is the Dynamical Signature (DS). Simply put, a DS is a vector of the same length as the number of beads in the ENM which describes the flexibility of each bead (Figure 1.5). The standard way to obtain such flexibility at every residue in the molecule is to compute the mean-square fluctuations (MSF), which are a vector of the same length as the number of residues in the studied molecule. MSF arise directly from the normal modes and their associated frequencies and have been shown to correlate well with experimental temperature factors (B-factors) from X-ray crystallography experiments as will be defined in Section 2.1.1 [26]. Temperature has no impact on the relative MSF as it only rescales them linearly. However, the relative contribution of each normal mode to the total vibrational entropy of the molecule does vary with temperature. Moreover ENCoM, like all coarse-grained ENMs, is a pseudo-physical model in which physical quantities such as temperature have no definite units. To model the impact of varying the pseudo-temperature on the contribution of each normal mode to the Dynamical Signature, we introduced the Entropic Signature (EntroSig). As its name implies, the EntroSig captures the entropy at every residue by scaling the square fluctuations by the vibrational entropy of the normal mode, which depends on the temperature as mentioned earlier (see Section 4.3.3).

Dynamical Signatures computed with ENCoM, whether MSF or EntroSig, capture the impact of a point mutation on the whole flexibility profile of the biomolecule. Figure 1.6 illustrates how a mutation can have both localized, as well as global effects. These predicted effects arise from the change in local surface complementarity of the atoms and their respective atom types [27], without change in the backbone positions of the beads in the system. Thus, the observed changes are a case of pure dynamical effects, without any impact on the structure. The most important flexibility change apparent in Figure 1.6B is at the mutated position, however significant differences in flexibility are present far from that position, for instance in the loop region around positions 40-45. This phenomenon is made clearer in Figure 1.6C, where we can see a white region (representing a zero difference in Dynamical Signature) between the mutated position and the distal loop. Thus, the changes in flexibility upon mutation happen as the result of both local and global perturbations to predicted structural dynamics.

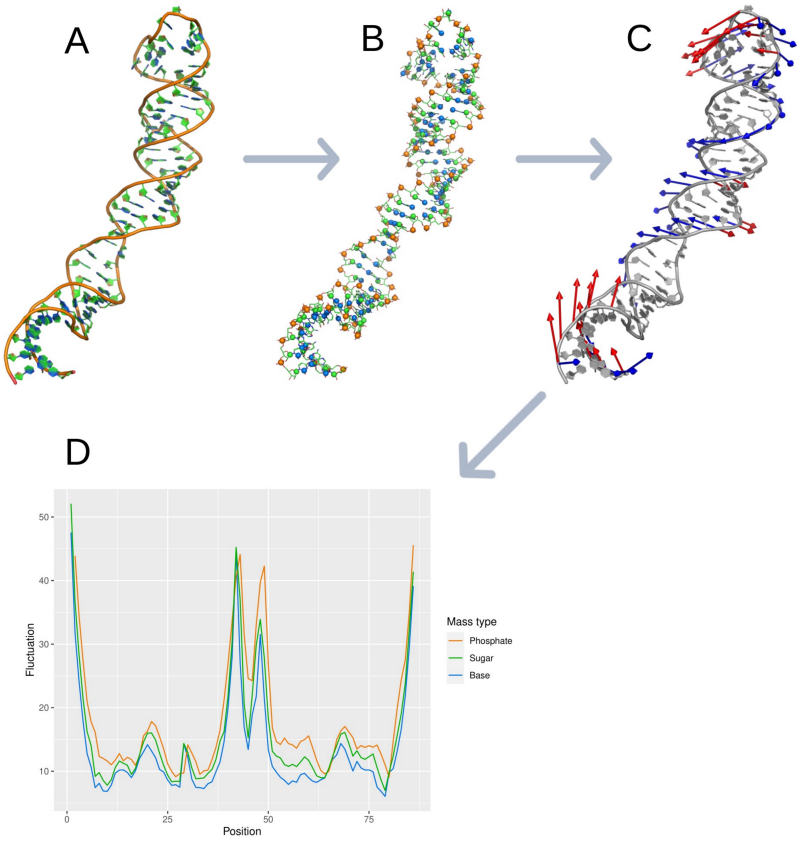


Figure 1.5: **Dynamical Signature from NMA.** A) An input structure is needed (pri-miR-125a is used as example). B) Coarse-graining represents groups of atoms with individual beads. C) The analytical solution gives 765 nontrivial normal modes, of which the first two are shown with red and blue arrows respectively. D) The whole set of normal modes can be reduced to a Dynamical Signature representing the flexibility of every bead in the system. Notice that if more than one bead is used per sequence unit (a nucleotide in this case), the contribution of different parts, here phosphate, sugar and base, can be individually analyzed.

When mutagenesis data are available, one can use the ENCoM Dynamical Signatures to train machine learning (ML) models. This combination of methods is the main contribution of the present thesis, which we call DynaSig-ML. We used two different ML back-ends: LASSO regression, which is a type of regularized multivariate linear regression, and multilayer perceptrons (MLPs), a type of artificial neural network. The idea behind the use of LASSO regression is to enable mapping the learned coefficients on the studied biomolecule, driving hypotheses about important features. This mapping is possible since all input variables are independent in such a model, hence the coefficient associated with every position in the DS can be interpreted at that position in the biomolecule. MLPs, on the other hand, are very powerful models which can capture complex relationships within

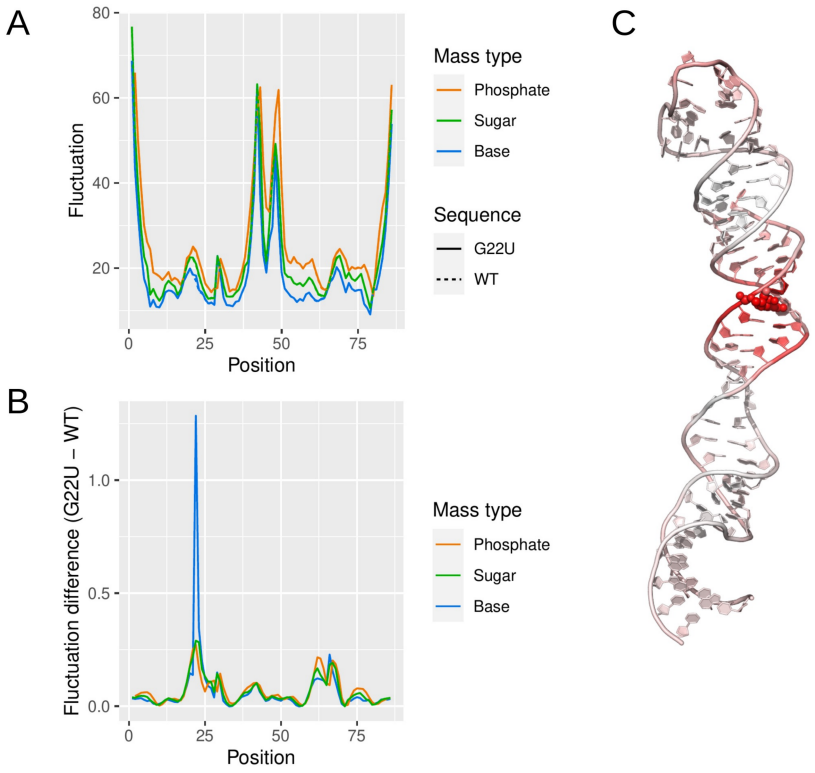


Figure 1.6: **Dynamical Signatures capture the effect of mutations.** A) The Dynamical Signatures from WT pri-miR-125a and the G22U mutation. B) The G22U - WT Dynamical Signature difference illustrates the localized and global effects the mutation has on structural dynamics. C) The Dynamical Signature difference of the backbone beads is mapped to the 3D structure of miR-125a, with a color gradient from white to red. No differences are lower than zero, so white represents zero and red represents the maximal value observed (0.26). The G22 mutated position is shown as spheres.

the input variables. Their improved performance comes at the cost of being hard to interpret once trained, thus having limited power to generate new hypotheses.

In order to investigate dynamics-function relationships with ENCoM, one starts with a dataset of experimental measures of a given function (e.g. pri-miR catalysis by the Microprocessor) for a number of sequence variants. The equilibrium 3D structure of the biomolecule has to be known, or modeled with acceptable confidence. The complete set of normal modes from every sequence variant are computed and reduced to an Entropic Signature (EntroSig), which represents the flexibility at each residue. By feeding these signatures in either a linear regression model or a simple neural network, the model learns which dynamical features correlate with function. It is then possible to predict in an ultra-high-throughput fashion the function from theoretical sequence variants.

ENCoM has been used in the past to predict the effect of mutations on thermal stability via the differences in vibrational entropy, but neither the whole Dynamical Signature nor prediction of function were considered. In this respect, the explicit use of the ENCoM Dynamical Signature coupled to supervised machine learning makes the methodology presented here novel. Indeed, we believe ENCoM-DynaSig-ML is the first entropy-based tool considering flexibility changes at every position and allowing ultra-high-throughput computational predictions. It has obvious applications in protein engineering and through the all-atom sensitivity of the ENCoM potential, it can predict allosteric effects from the binding of different ligands to the same receptor. In addition to the NRGTEN package [31], the DynaSig-ML package [32] contains all the necessary tools to perform the analyses presented here in a streamlined fashion and both packages are extensively documented online with simple examples and tutorials.

Our work opens up new avenues for biomolecular engineering, bringing together sequence, structure and structural dynamics in order to better understand and predict the function of biomolecular systems. The next section will introduce the topics of microRNA maturation, GPCR activation and VIM-2 lactamase catalytic efficiency, which are the object of the case studies presented in Chapter 6, Chapter 7 and Chapter 8. Chapter 2 reviews existing experimental and computational techniques to study structural dynamics and tools to predict the effects of mutations. Chapter 3 outlines the theory of normal mode analysis and the ENCoM model, LASSO regression and multilayer perceptrons. Chapter 4 covers the global methodology of the present work, from the modifications made to the ENCoM model to the diverse benchmarks assembled for its re-parameterization. Chapter 5 will give the detailed results from this extensive parameter search. Chapters 6 to 8, as mentioned, are dedicated to dynamics-function case studies of human microRNA miR-125a, the μ-opioid receptor and VIM-2 lactamase.

The three application chapters will outline the methodology specific to their respective experimental datasets, while the biological background of each biomolecular system will be covered here, in the next section. This division enables better appreciation of both the biological diversity behind these three applications and the similar methodology in the application of the ENCoM-DynaSig-ML pipeline, highlighting the ENCoM-DynaSig-ML pipeline as a general tool applicable to any biomolecule for which dynamics-function relationships are suspected.

Of the three application chapters, Chapter 6 is the first and also the longest. There are four reasons for this added length:

1. The dataset of miR-125a maturation efficiency contains a great deal of sequence redundancy, so care has to be taken in ensuring the machine learning models are not merely learning sequence.

2. Because of this opportunity to distinguish sequence and dynamics, we use Chapter 6 to compare both LASSO regression and multilayer perceptrons, the two machine learning backends explored in the thesis. Our findings lead us to use LASSO regression exclusively in the subsequent chapters.

3. The 3D structure of miR-125a has to be modeled as no experimental structure is available. There is thus the added step of 3D model selection.

4. Although outside the scope of the present thesis, we have the experimental setup to test theoretical miR-125a variants. Ultra-high-throughput predictions are thus presented exclusively in that chapter.

The four results chapters, namely the parameter search chapter and the three application chapters, will each have a dedicated discussion section which will focus exclusively on findings from that chapter. The general findings from these four results chapters will be discussed together in Chapter 9 along with our ideas for future work and Chapter 10 will list our conclusions.

1.3 Case studies of dynamics-function relationships

The present thesis introduces the ENCoM-DynaSig-ML computational pipeline for the study of dynamics-function relationships in biomolecules. In order to illustrate the wide applicability of the pipeline, we have performed dynamics-function case studies for three diverse biomolecules:

1. The cleavage efficiency of human microRNA miR-125a by the Microprocessor, with mutagenesis data about pri-miR-125a sequence variants.

2. μ-opioid receptor activation, where different ligands act as pseudo-sequence variants.

3. VIM-2 lactamase catalytic efficiency, using deep mutational scan data.

The following subsections will introduce the biological background for these three case studies.

1.3.1 *microRNA maturation: miR-125a mutagenesis*

Mature microRNAs (miRs) are small non-coding single-stranded RNAs of approximately 22 nucleotides. Their main function in the cell is to regulate gene expression by guiding RNA-Induced Silencing Complexes (RISCs) to complementary regions within messenger RNA (mRNA) molecules, triggering silencing of these targets [33]. There are upwards of 2000 miRs in the human genome [34], which collectively can target more than 60% of human genes at the mRNA level [35]. miRs

play important roles in cell differentiation as they enable the fine-tuning of gene expression [36]. They have recognized roles in numerous physiological processes and diseases, including cardiovascular disease [37], neurodegenerative diseases [38] and cancer [39].

miR biogenesis can happen through the so-called canonical pathway, illustrated in Figure 1.7, as well as through various noncanonical pathways which include the production of miRs from introns in mRNA transcripts (mirtrons) [40]. Our case study focuses on the first step of canonical miR biogenesis, hence we will not further discuss the noncanonical miR biogenesis pathways.

Canonical miR production starts with a DNA miR gene, transcribed to RNA by the enzyme RNA polymerase II. The generated transcript is called a primary miR transcript (pri-miR) and adopts a stem-loop structure comprising approximately 35 base pairs, single-stranded basal segments and an apical loop [41]. One miR gene can give rise to a single pri-miR stem-loop, or include several pri-miRs linked together in what is referred to as a polycistronic miR cluster [42].

After transcription, the first step in miR biogenesis is the cleavage of the pri-miR by the Drosha/DGCR8 heterotrimer, also called Microprocessor. This enzymatic complex acts as a molecular ruler and measures the length of the stem-loop structure as one of the conditions for processing, with the cleavage happening around the 11th base pair from the single-stranded basal part [15]. Since the Microprocessor is located in the nucleus, the mechanism for miR recognition needs to be very stringent to avoid the cleavage of other functional ncRNAs, such as ribosomal RNAs which dominate the transcriptional landscape. Beyond structural geometry, several features are important for pri-miR recognition by the Microprocessor, some of which have already been identified as dynamical in nature [13, 43].

The cleavage of the pri-miR by the Microprocessor generates a miR precursor (pre-miR), which has a shorter stem-loop structure and is characterized by a 2 nucleotide overhang at its 3'-end. The pre-miR is exported out of the nucleus to the cytoplasm, where the 3'-end overhang is recognized by the Dicer enzyme, which cleaves the stem-loop approximately 22 nucleotides from this overhang to generate a duplex of mature miR strands [15].

The precise mechanism through which one of the two mature miR strands is loaded in the RNA-Induced Silencing Complex (RISC) remains debated but there is compelling evidence to believe the intact duplex is loaded in the Argonaute protein, which then rejects one of the two based on physicochemical properties [44]. In any case, one of the strands from the duplex ends up loaded in the RISC and then serves as the template for recognizing mRNA transcripts to regulate while the other strand is degraded.

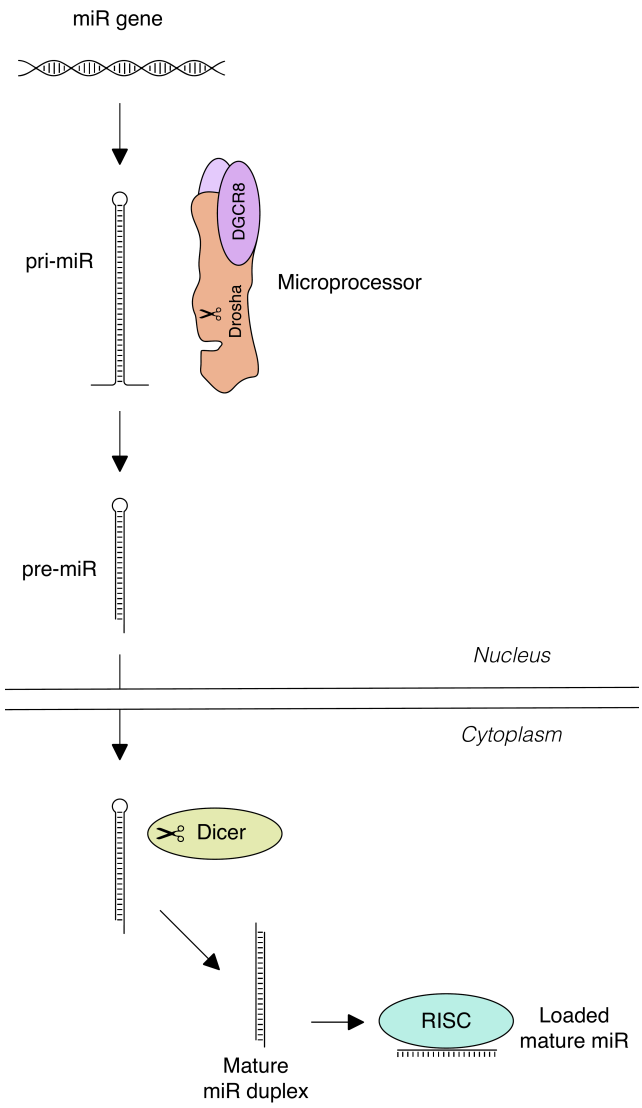


Figure 1.7: **microRNA biogenesis.** The production of microRNAs (miRs) starts in the nucleus with a miR DNA gene. That gene is transcribed to a primary microRNA transcript (pri-miR) forming a stem-loop structure of approximately 35 base pairs. The Microprocessor complex, formed of Drosha and DGCR8, cleaves the pri-miR to form a precursor microRNA (pre-miR). The pre-miR is then exported out of the nucleus to the cytoplasm, where it is cleaved by the Dicer enzyme, generating a duplex of mature miR strands. One of them is loaded in the RISC complex, which can then interact with complementary sites within mRNA transcripts to regulate their expression.

miR-125a is a human miR in which a single nucleotide polymorphism (SNP) predisposes to aggressive types of breast cancer. The SNP (G22U) is located in the sequence of the mature miR-125a and it was initially hypothesized that loss of

complementarity of miR-125a to its target genes could explain the predisposition to breast cancer. However, the main effect of G22U is to prevent cleavage by the Microprocessor, leading to the almost complete loss of mature miR-125a from the minor allele [45]. The SNP does not affect the global 2D structure of the minimum free energy (MFE) state and simply introduces a noncanonical base pair in the stem. The lost signal for Microprocessor cleavage seems to be dynamical in nature, as it was shown that changes in 2D structural dynamics of the 16 possible base pairs at the SNP position correlate well with their maturation efficiency measured from cellular luciferase assays [14].

David Bartel's group has generated high-throughput data about the maturation efficiency of upwards of 50 000 pri-miR-125a sequence variants using an enzymatic assay with purified Microprocessor [13]. In Chapter 6, miR-125a dynamics-function relationships will be investigated by applying ENCoM-DynaSig-ML on this mutagenesis dataset.

1.3.2 *μ-opioid receptor activation*

G protein-coupled receptors (GPCRs) represent the largest family of human receptors. GPCRs are of tremendous pharmacological importance, with over 450 FDA-approved drugs targeting a total of 108 different GPCRs as of 2017 [46]. GPCRs are membrane-bound proteins and can recognize a plethora of extracellular signals, from small molecules to mechanical forces [47, 48]. These extracellular signals are transmitted inside the cell via interaction between the receptor and G proteins. GPCRs are divided in 6 classes, of which class A is the most studied, the most pharmacologically targeted and the one for which the activation mechanism is understood the best. The case study presented in Chapter 7 focuses on the activation of the μ-opioid receptor, which is part of class A. Thus, other classes of GPCRs will not be further discussed.

Class A GPCRs have seven transmembrane helices, an extracellular binding pocket and intracellular loops which can interact with G proteins. The activation of class A GPCRs is perhaps one of the better known examples of dynamics-function relationships. Two distinct states of the receptor can be indentified from its conformational landscape, called the active and inactive states. Figure 1.8 shows these two states for the μ-opioid receptor, as captured by X-ray crystallography [49, 50]. In the active state, the receptor can interact with the intracellular G protein and activate it as the first step in signal transduction. In the inactive state, this interaction is highly unprobable and no signal is transduced. It is thought that both states exist in the absence of a ligand and that class A GPCR activation is thus a case of conformational selection [51].

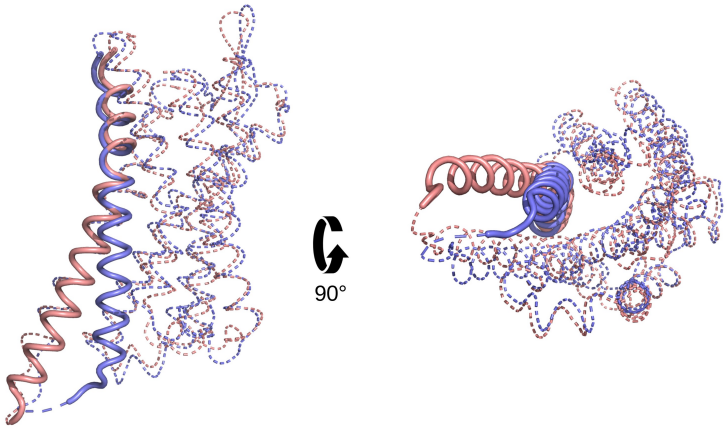


Figure 1.8: **Active and inactive states of the mu-opioid receptor.** The active state (PDB ID 5C1M) is shown in salmon and the inactive state (PDB ID 4DKL) in slate. Transmembrane helix 6 (TM6), which undergoes the largest conformational change between the two states, is shown in bold. All other parts of the receptor are dashed.

The largest structural change observed between the inactive and active states of class A GPCRs is the outward position of transmembrane helix 6 (TM6) in the active state (Figure 1.8). This change, along with other rearrangements, generally favors the binding of different types of G proteins or other intracellular partners such as arrestins [52, 53]. GPCR signal transduction can be biased towards the activation of G proteins, arrestins, or balanced between the activation of both [54]. Moreover, different types of G proteins exist, with the potential for different cellular effects dependent on their type [55, 56]. To encompass all possibilities, let us consider that the binding of an agonist to a GPCR increases the probability of the receptor occupying an active conformation, in which it favorably interacts with some intracellular partners in order to transduce the signal from the extracellular ligand inside the cellular environment. This unified, simple model is illustrated in Figure 1.9.

As its name suggests, the μ-opioid receptor (MOR) is the primary target of the most effective class of analgesics, opioids. As such, it is a widely studied receptor: over 5000 manually curated ligands of MOR with associated biological assays are listed in the ChEMBL database as of June 2022 [57]. Like other class A GPCRs, MOR can exhibit balanced G protein and arrestin signalling or bias towards either G proteins or arrestins, depending on the agonist [58]. This ability of different agonists to recruit divergent intracellular partners illustrates the receptor's conformational flexibility and the coupling of ligand binding with the intracellular receptor conformation through allosteric effects [59].

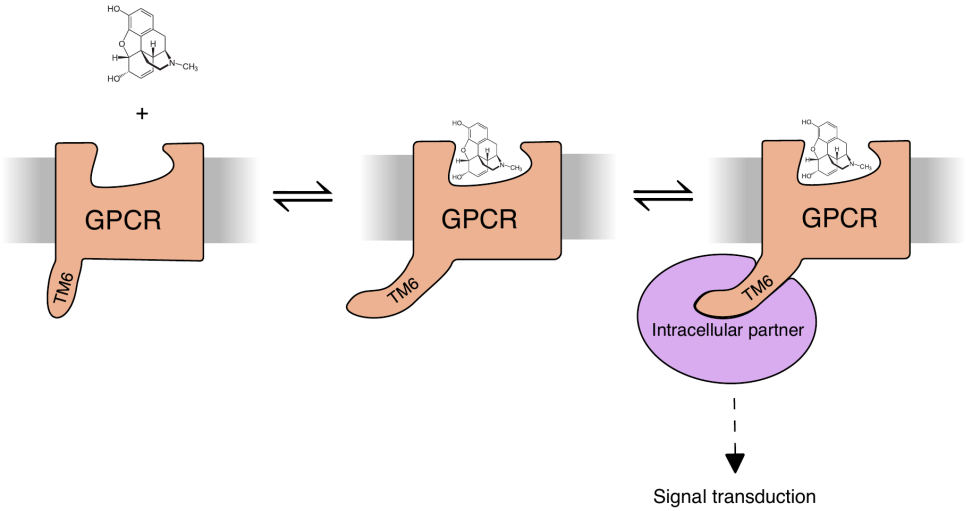


Figure 1.9: **Simplified GPCR signal transduction.** The binding of an agonist favors the active state, in which structural changes are apparent at the GPCR's intracellular moiety. The most prominent of these changes is the outward position of transmembrane helix 6 (TM6). The active conformation allows the interaction and activation of an intracellular partner, which further transduces the signal once activated.

Beyond signalling bias, the cellular context plays a role in the effect of GPCR activation, as the concentration and idendity of downstream effectors affect how the activation signal ultimately propagates. Indeed, a study from 2016 identified hundreds of thousands distinct connections between human GPCRs and downstream pathways, which are dependent on tissue type [60]. [^{35}S]GTPγS binding assays measure G protein activation after the binding of a ligand to a GPCR, offering a way to assess GPCR activation almost directly, without the added noise of the downstream effectors [61].

In pharmacology, the maximal biological response induced by a drug is termed the maximal efficacy or Emax [62]. When a measurement of maximal efficacy relative to a known full agonist is made, the term Emax is also frequently used, and in that case represents a percentage of maximal efficacy. For our dataset of MOR ligands with experimentally measured activation potential, we select ligands with Emax relative to DAMGO, a strong and potent MOR agonist [63], measured by [^{35}S]GTPγS binding assays.

These criteria allow us to obtain 198 MOR ligands with uniform experimental measures of MOR activation potential, which constitute the starting dataset for the case study presented in Chapter 7. The change of ligand does not constitute a variation of sequence *per se*, however it can be considered as such within ENCoM:

a single bead is used in the model to represent every ligand, and the change in all-atom surface complementarity leads to a change in dynamics in a manner similar to an *in silico* generated sequence variant.

1.3.3 *VIM-2 lactamase catalytic efficiency*

β-lactamases are bacterial enzymes which degrade β-lactam antibiotics, thus providing the bacteria expressing them with antibiotic resistance. Since β-lactams are the most important class of antibiotic by use, the evolution of more efficient β-lactamases is a real threat to health care systems worldwide [64]. β-lactamases can be divided in four classes: classes A, C and D share many similarities and are thought to have evolved from a common ancestor, while class B, also called metallo-β-lactamases (MBLs), depends on metal ions for its activity and has distinct evolutionary origins [65]. The standard treatment for β-lactam resistant infection is the co-administration of β-lactam antibiotics with β-lactamase inhibitors. Until recently, no MBL inhibitors had made it to the clinic [66], however there is now a standard combination of aztreonam with avibactam that is routinely used against MBLs [67].

Verona integron-encoded metallo-β-lactamase 2 (VIM-2) is one of the most widespread MBLs and thus represents a major source of antibiotic resistance [68]. Its 3D crystal structure is shown in Figure 1.10 as solved by Brem and coworkers (PDB ID 4bz3) [69].

A deep mutational scan (DMS) of VIM-2 lactamase was recently conducted, with the experimental measurement being bacterial fitness under various concentrations of β-lactam antibiotics [70]. The measurement thus encompasses the protein's expression, stability and catalytic efficiency at once. β-lactamases are considered relatively static enzymes, however there is evidence for an important role of dynamics in their ability to hydrolyze β-lactam antibiotics [71]. The application of the ENCoM-DynaSig-ML pipeline to the VIM-2 evolutionary fitness DMS data can thus help shed light on whether dynamics play a role in the enzyme's activity. This question will constitute the basis for the last of our dynamics-function case studies, presented in Chapter 8.

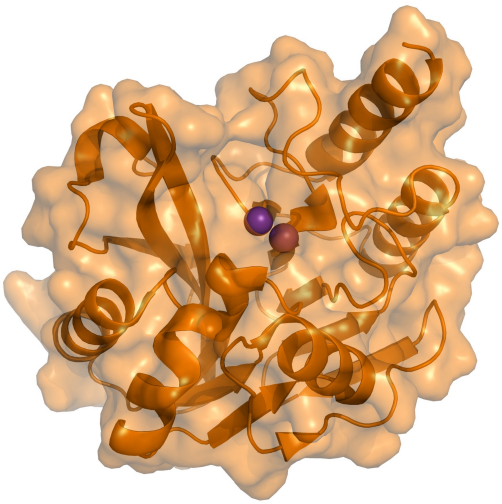


Figure 1.10: **Crystal structure of VIM-2 lactamase.** Cartoon and semi-transparent surface representations of the VIM-2 lactamase are shown in orange, with the active site zinc ions represented as purple spheres. The first of two biological units submitted to the PDB by the authors was used (PDB ID 4bz3) [69].

2

LITERATURE REVIEW

The present thesis introduces the ENCoM-DynaSig-ML computational pipeline, which enables the fast prediction of biomolecular function from perturbations to simulated structural dynamics. The concept of studying dynamics to elucidate function is well established, as well as the study of perturbations to structural dynamics as a result of changes in biomolecular sequence (mutations) or ligand binding. Two broad approaches to study relationships between perturbations to structural dynamics and biomolecular function can be distinguished: experimental techniques and computer simulations.

The focus of this thesis is on computer simulations, which need experimental structural dynamics in order to exist; experiments are crucial in order to develop and validate simulation methodologies. The main techniques for obtaining such information, namely X-ray crystallography, nuclear magnetic resonance and cryo-electron microscopy, will be discussed in the first section of the present literature review.

The second section will present the main techniques used for computer simulations of biomolecular structural dynamics. These fall in three broad categories: molecular dynamics methods, sampling (or Monte Carlo) methods and normal mode analysis methods.

All experimental and most computational techniques discussed in the first two sections can in theory be used to study changes in structural dynamics as a result of perturbations (the exception being sequence agnostic normal mode analysis methods which do not capture, for example, the effect of point mutations that do not change the backbone conformation). However, studying vast numbers (from hundreds of thousands and upwards) of potential perturbations is out of reach for all structure determination experimental methods at the moment, and has a very high computational cost in the case of most computer simulation approaches. For biomolecular engineering, such vast numbers of predictions are routinely needed in order to aid in the design of optimized biomolecules. Two categories of computational tools can be used to accomplish high-throughput predictions in the context of biomolecular engineering: stability prediction tools and variant effect predictors. We will review these tools in Section 2.3 and support our claim

that ENCoM-DynaSig-ML is the first variant effect predictor based on detailed computational structural dynamics.

2.1 EXPERIMENTAL STRUCTURAL DYNAMICS

Structural dynamics can be studied experimentally via two main approaches: the direct measurement of dynamical properties, and the elucidation of static structures of the same biomolecule occupying different conformations, either as a result of varying experimental conditions or as the direct capture of distinct conformational states (achievable with cryo-EM).

Table 2.1: **Number of entries in the PDB by experimental method.** The number of entries in the Protein Data Bank (PDB) [72] when restricting the search to entries resolved by X-ray diffraction, solution NMR, electron microscopy or none of these ("other"), with the percentage of total shown in parentheses, is given for all entries, for unique biomolecular entities (no sequence redundancy), and for groups of biomolecular entities sharing at leat 95% sequence identity. The data were retrieved as of July 7, 2022.

Method	Number of entries	Unique entities	Unique groups at 95% identity
X-ray diffraction	257 294 (67.7%)	88 187 (76.2%)	59 362 (71.1%)
Solution NMR	15 499 (4.1%)	9 958 (8.6%)	9 229 (11.1%)
Cryo-electron microscopy	106 779 (28.1%)	17 373 (15.0%)	14 645 (17.5%)
Other	535 (0.1%)	236 (0.2%)	216 (0.3%)

Table 2.1 presents the number of entries in the protein data bank (PDB) [72], the the hub of all publicly available experimentally resolved biomolecular structures, per experimental technique used. As is evident, most of our structural knowledge emerges from three experimental techniques: X-ray crystallography, solution NMR and cryo-electron microscopy. The next subsections will present these experimental techniques, focusing on how they can provide information about biomolecular structural dynamics.

2.1.1 *X-ray crystallography*

X-ray crystallography is first and foremost a technique to determine biomolecular structure, as it elucidates the atomic configuration of a biomolecule in crystal form, permitting very little movement [73]. Fluctuations of atomic positions around the equilibrium structure, commonly called temperature factors or B-factors, are also

captured as part of the experiment. However, these do not necessarily represent biologically relevant motions and have been shown to arise in part due to rotations and translations of individual biomolecules in the crystal lattice, which are completely irrelevant to functional dynamics in a cellular context, and their relevance as a benchmark for ENMs has been questioned [74]. Nonetheless, we will use crystallographic B-factors for 2 of the 7 diverse benchmarks presented in Chapter 5 as they are a common standard in the field, and represent a readily accessible source of experimental structural dynamics.

As X-ray crystallography is the experimental method behind the majority of known biomolecular structures (see Table 2.1), indirect structural dynamics can be gathered from these experiments. One example of curated indirect structural dynamics from crystal structures is the PSCDB database of protein structural change upon ligand binding [75]. The database contains 839 structural changes that were identified from pairs of high resolution crystal structures, each pair representing the same protein in bound and unbound conformations. In a similar fashion, it is possible to search the PDB for pairs of entries representing the same biomolecule in different conformations, as we will do for RNA structures in Chapter 5.

The principle behind X-ray crystallography is X-ray diffraction: a beam of coherent light in the X-ray spectrum is shined on a crystal, and the photons are diffracted by interacting with the electrons in the crystal. The periodic nature of the crystal lattice gives rise to a diffraction pattern which is unique to the geometry of the asymmetric unit in the lattice. The mathematical framework to deduce atomic geometry from the diffraction patterns was first laid out by Max von Laue [76], for which he won the Nobel Prize in physics in 1914.

The first protein structure to be solved by X-ray crystallography was that of sperm whale myoglobin, a 152 amino acid protein, in 1958 by John Kendrew [77]. Kendrew later shared the 1962 Nobel Prize in Chemistry with Max Perutz for this accomplishment. Since then, X-ray crystallography has grown to be the most prolific experimental method for biomolecular structure elucidation. However, significant limitations exist as to the interpretation of the structural information obtained by this method. The most obvious is the fact that the conditions in which the structure is elucidated are far from physiological conditions: most biomolecules exist in solution or in membranes, and with the exception of very rare instances (such as the crystalline δ-endotoxins from *Bac. thurigiensis* [78]), do not form crystals under physiological conditions. This unphysiological nature of crystallized biomolecules means that the elucidated conformation might be slightly different from the equilibrium conformation in a living organism, especially at the interfaces between two individual biomolecules in the crystal. These intermolecular interactions, known as crystal contacts, happen as a result of the tight packing of the crystal units, slightly influence the equilibrium conformation and have different physicochemical properties than physiological biomolecular interfaces

[79]. Nevertheless, the overall fold and intramolecular interaction details (such as ligand-binding site interactions) determined by X-ray crystallography can be considered correct. Notorious X-ray crystallography successes include the structure of the whole ribosome [80] and the crystallization of GPCRs [81]. Crystallography can be considered an essential step in most drug design campaigns [82].

Protein and RNA structures from X-ray crystallography experiments are used extensively for the benchmarks in Chapter 5 and as the input structures in Chapter 7 and Chapter 8.

2.1.2 *Solution NMR spectroscopy*

When atomic-level information about a biomolecule's structural dynamics is desired, solution nuclear magnetic resonance spectroscopy (NMR) is widely regarded as the go-to experimental method, with modern equipment allowing the study of biomolecular motions happening on timescales from picoseconds to several seconds and even days, depending on the specific technique employed [83]. Since the experiments are performed in solution, physiologically relevant dynamic processes can be captured, such as enzymatic catalysis [84], the effects of ligand binding on the conformational landscape [85] or transient protein-protein interactions [86]. Beyond specific information about known or suspected dynamical processes, solution NMR allows the resolution of complete atomic-level structures. Moreover, the distance restraints obtained from the experiments usually allow the generation of an ensemble of structural models which collectively explain the observables [87].

As the name implies, nuclear magnetic resonance spectroscopy is based on the fact that atomic nuclei will interact with magnetic fields at characteristic frequencies which depend on their surroundings. This phenomenon relies on nuclear magnetic spin, which in turn depends on the ratio of protons to neutrons [88]. This dependence of nuclear resonance on nucleon ratios allows the usage of isotopically labelled biomolecules or deuterated solvent in order to maximize the signal-to-noise ratio in solution NMR experiments [89]. The theory behind NMR spectroscopy is beyond the scope of the present thesis and has been extensively reviewed elsewhere [88, 90].

Without isotope labelling, the size of the molecules which can be studied at atomic resolution by solution NMR is severely limited [89]. Isotope labelling allows the study of bigger biomolecules, however the production of the labelled biomolecule is experimentally demanding. Another limitation of solution NMR is the high cost of the necessary magnetic apparatus. The systems capable of producing the strongest magnetic fields (higher frequency) and thus study the biggest biomolecules cost millions of dollars. Thus, solution NMR can be seen as the gold standard for the experimental study of biomolecular structural dynamics, however it is not

applicable to all biomolecules and is relatively costly both in terms of experimental time and actual equipment cost. This is reflected by the still relatively low amount of solution NMR entries in the PDB, at 8.6% of unique entities (see Table 2.1). Because of these high costs in time and equipment, it is not currently possible to use solution NMR to study large quantities of sequence variants from the same biomolecule.

Solution NMR ensembles of RNA molecules and RNA-protein ensembles will be benchmarked against for the parameter search presented in Chapter 5.

2.1.3 *Cryo-electron microscopy*

Of the three experimental techniques presented here, cryo-electron microscopy is the only one in which the experimenters directly "look" at the biomolecule: a frozen sample containing the dissolved molecule of interest is imaged by an electron microscope, and a density map is constructed algorithmically by processing vast amounts of particles in different orientations [91]. Since the biomolecules are flash-frozen, different conformational states are also sampled, and with high enough resolution the density maps corresponding to each state can be obtained [92].

Since the advent of better electron detectors in the early 2010s, cryo-EM is leading what has been termed a revolution in structural biology [93]. Combined with advances in image processing, modern electron detectors allow the determination of cryo-EM electron density maps which rival the resolution of X-ray crystallography [94]. In 2020, the first atomic-level resolution protein structures solved by cryo-EM were reported [95, 96], and there are now 139 distinct biomolecular entities in the PDB that have been solve by cryo-EM at a resolution below 2 Å.

Since the advent of atomic-resolution cryo-EM structures is fairly recent and most of them are of very big biomolecules/biomolecular complexes, we do not employ cryo-EM structures in the present thesis. However, the field is developing at such a high rate that we expect cryo-EM conformational ensembles to become a central part of experimental structural dynamics in the coming years.

2.2 COMPUTATIONAL STRUCTURAL DYNAMICS

The present section is dedicated to methods for the computational simulation of biomolecular structural dynamics, which we split in three broad categories: time-stepping, or molecular dynamics (MD) methods; sampling methods, such as Monte Carlo techniques; and normal mode analysis methods, of which the present thesis makes use. Since coarse-grained normal mode analysis is detailed in Chapter 3, we will spend more time on the first two categories in what follows. Let us mention that hybrid quantum mechanics/molecular mechanics (QM/MM) methods offer

the greatest level of simulation detail, especially when chemical changes are to be simulated. However, these methods are very costly to use and are thus beyond the scope of our work [97].

2.2.1 *Molecular dynamics methods*

As all molecules, biomolecules are composed of a number of atoms linked together by covalent bonds and interacting both within the molecule and with their environments through different forces. When the approximation is made that these forces can be described using classical newtonian interactions, the problem of simulating biomolecular dynamics is reduced to a variation of the n-body problem.

Molecular dynamics (MD) simulations of biomolecules were first reported in 1975 following the seminal work of Michael Levitt and Arieh Warshel, in which the authors report the simulated folding of bovine pancreatic trypsin inhibitor, a 58 amino acid protein, from a completely denatured state to a near-native state [98]. To accomplish this, they used a simple interaction potential based on simulations of all 400 possible dipeptides, which was then the basis of a coarse-graining scheme leaving two interaction centers per amino acid: the C_α atom and the centroid position of the side-chain. In the years following that work, which can be considered the first coarse-grained MD protocol applied to biomolecules, there was a great emergence of MD force fields, coarse-graining and enhanced sampling methodologies [99]. Levitt, Warshel and Karplus share the 2013 Nobel Prize in Chemistry for pioneering the field of biomolecular simulation. [100]. Today, MD simulations are routinely employed for the study of biomolecular dynamics and can be considered an indispensable tool of modern structural biology [18].

The next subsection will discuss all-atom MD methods, which provide the greatest level of details and also exhibit the greatest uniformity in their mathematical definitions. Then, coarse-grained MD methods will be reviewed.

2.2.1.1 *All-atom MD methods*

All-atom molecular dynamics methods, as their name implies, explicitly simulate all atoms of the studied biomolecule and of its immediate environment, which includes solvent molecules, ions, lipid molecules in the case of transmembrane or membrane-bound biomolecules, along with small molecules of interest if they are relevant for the studied system. In order to perform the simulation, every atom has to have xyz coordinates assigned to it, a velocity and a force acting on it. To compute the force acting on every atom, a force field (also called potential) is needed. The force field describes the instantaneous force on all atoms in the system from their cartesian coordinates. Most all-atom MD force fields have the following form:

$$U = \sum_{\text{bonds}} V_{\text{bonds}} + \sum_{\text{angles}} V_{\text{angles}} + \sum_{\text{dihedrals}} V_{\text{dihedrals}} + \sum_{\text{improper}} V_{\text{improper}} + \sum_{\text{Lennard-Jones}} V_{\text{Lennard-Jones}} + \sum_{\text{electrostatic}} V_{\text{electrostatic}} \tag{2.1}$$

The first four terms represent intramolecular interactions: covalent bond stretching, angle bending in trios of covalently connected atoms, torsional movements from quartets of connected atoms, and so-called improper dihedrals which happen between unconnected atoms but play a big role in fine-tuning the potentials in order to match experiments and quantum mechanical theory [101]. The last two terms describe long-range interactions, which can be both inter- and intra-molecular. They represent the Lennard-Jones potential [102] and the electrostatic potential.

Once the size of the timestep is set, the simulation is carried on by the numerical integration of the potential, starting from the initial positions and velocities for all atoms in the system. The initial atomic coordinates for the studied biomolecule are usually those of the experimentally resolved structure when the goal is to study dynamics near the equilibrium, native state. However, it is possible in theory to start from a completely denatured molecule, and long enough simulations should allow it to fold into its native conformations if the potential is accurate enough. Such complete folding simulations can now be carried for fast-folding proteins [103]. The state-of-the-art machine for all-atom MD simulations is the Anton 3 supercomputer, a purpose-built machine with circuit architecture optimized for MD simulations [104]. The 64-node version achieves 212.2 μs/day on the classic 24 000 atoms DHFR benchmark. For comparison, a performance of 160 ns/day was reported of the same DHFR benchmark using one 8-core Intel Core i7 5960X CPU @ 3.5GHz [105]. Anton 3 thus represents a speedup of over 1000 times on this classic benchmark, but its advantage lies in its ability to carry that speedup over to very large systems, such as the whole ribosome and even entire satellite viruses [104].

Large biomolecular motions tend to happen on large timescales; for most proteins, domain motions happen on the microsecond-to-millisecond timescale and large collective motions happen on the millisecond-to-second timescale [106]. For RNA, the presence of deeper kinetic traps due to the strength of the base pairs and the limited number of nucleotides (which leads to many possible suboptimal pairings) results in the equivalent collective motions happening on timescales that can reach hours [107]. In the case of proteins, the recent advancements in hardware mean that most biologically relevant motions are within the grasp of state-of-the-art all-atom MD simulations. However, these require tremendous resources and it is still beyond our grasp to use these techniques for the study of high-throughput dynamics. For example, the DHFR enzyme used as a classical MD benchmark comprises 159 amino acids [108]. A virtual deep mutational scan would thus generate $159 * 19 = 3021$ point mutations. Large collective motions happen on

the millisecond-to-second timescale [106], but let's assume the goal would be to accumulate just one millisecond of simulation time for every variant. Even using the Anton 3 supercomputer, it takes aournd 5 days per DHFR variant to get to a millisecond (212 μs/day) [104]. Thus, in order to finish this study within a year, one would need more than 41 instances of Anton 3. By comparison, the simplifications made by coarse-grained normal mode analysis mean that the same timescales can be reached with ENCoM in less than an hour on a quad-core, modern laptop.

2.2.1.2 *Coarse-grained MD methods*

The same basic principle of time-stepping numerical integration used for all-atom MD applies to coarse-grained MD methods. However, the available potentials exhibit much more variation since they depend on the coarse-graining strategy used. These strategies vary from potentials that make the solvent molecules implicit but otherwise model all atoms [109] to simulations of huge biomolecular complexes in which blocks of many amino acids are represented as single rigid elements interacting between themselves [21]. On its own, removing the solvent allows for considerable speedup, as the majority of atoms in an all-atom MD simulation are solvent atoms. When representing groups of atoms by interacting beads, one can define a coarse-graining factor as the average number of atoms represented by each bead. For instance, since the average number of atoms in an amino acid is around 19, a CG MD model with one bead per amino acid would have a coarse-graining factor of 19. The maximal theoretical speedup achievable by coarse-graining is proportional to the inverse square of the coarse-graining factor [21], thus such a model can hope to achieve a maximal speedup of $19^2 = 361$ compared to atomistic simulation. The significant speedups achievable make CG MD models very attractive for the study of structural dynamics in large systems.

However, even the remarkable speedups achieved by the CG MD cannot compete with the speed of coarse-grained normal mode analysis for the study of slow timescales dynamics. For instance, let's go back to the DHFR example used in the last section. It was recently reported that a single CPU core from an Intel Gold 6148 @ 2.40GHz can achieve around 10 ns/day simulation time [110]. The DHFR protein atoms constitute around 10% of the atoms in the system, thus a solvent-implicit model could hope to achieve a maximal speedup of 100 times. Combined with coarse-graining at the 1-bead per amino acid level, a total maximal speedup of 36 100 could be achieved, for a total of 361 μs/day. To get to 1 ms for the 3021 point mutations in a virtual DMS, one would thus need approximately 23 core-years even with this maximal-speedup, fictional CG MD model. Again, computing the ENCoM Dynamical Signatures for 3021 variants of the DHFR enzyme takes less than four core-hours. Thus, while coarse-grained molecular dynamics simulation play vital roles in our understanding of dynamical biomolecular phenomenon happening over long timescales or in very large systems, we do not believe that

these techniques can be used to predict the dynamical effects of sequence variants in an ultra-high-throughput fashion within reasonable computational cost.

2.2.2 *Monte Carlo methods*

Monte Carlo algorithms applied to the simulation of molecular systems were first described in 1953 in Nicholas Metropolis' seminal work [111]. The principle behind MC sampling is to perturb a starting system with a slight random change, then compute the energy associated with the deformed system, and accept the change with a probability that depends on the difference in energy between the current state and the deformed state, with lower energies being favored.

There is a long-standing debate about whether MD or MC methods are the most efficient for conformational sampling, and the answer seems to depend on the properties of the studied system and the strategies employed to accelerate sampling [112, 113]. However, in the recent decades MD methods have gained more widespread use after successes such as the reversible folding of fast-folding proteins using purpose-built supercomputers [103]. For our purposes, we will consider the computational cost of MC methods and MD methods to be roughly equivalent, and as we have demonstrated in the last subsection, even the most optimistic speedup achievable by coarse-graining is not comparable to the speedup afforded by coarse-grained normal mode analysis in the study of collective, slow-timescales dynamics.

2.2.3 *Normal mode analysis methods*

Coarse-grained normal mode analysis (NMA), as already stated, allows the fast and analytical simulation of biomolecular dynamics around an equilibrium structure. Because of the quadratic energy well assumed around the input structure, the computational cost scales only with the size of the system, not with the desired timescale of study [114].

ENCoM, the central part of the ENCoM-DynaSig-ML pipeline which is the subject of the present thesis, is a sequence-sensitive coarse-grained normal mode analysis model. Since most of Chapter 3 is dedicated to the presentation of the ENCoM coarse-grained normal mode analysis (NMA) model, we will not further discuss NMA in the present literature review.

2.3 BIOMOLECULAR ENGINEERING

To the best of our knowledge, the ENCoM-DynaSig-ML pipeline is the first computational tool allowing the fast and systematic learning of biomolecular dynamics-function relationships from experimental datasets followed by high-throughput pre-

diction of new sequence variants, with obvious application potential for biomolecular engineering. Let us clarify this claim: while numerous studies can be found in which simulated or experimental structural dynamics of some sort guide the design of sequence variants with engineered properties, these approaches are often time-consuming, require expert knowledge and are not generalizable to any biomolecule. In contrast, the ENCoM-DynaSig-ML pipeline is very fast and thus has the potential to be used for high-throughput "virtual screening" of sequence variants, based on their structural dynamical properties. This high-throughput applicability in the context of biomolecular engineering is what constitutes, again to the best of our knowledge, a novel contribution.

To support our claim, we conducted an extensive literature search in order to build a repertoire of computational biomolecular engineering tools routinely used. The next subsections will present the most widely used tools, divided in two categories: stability change prediction tools and variant effect tools.

2.3.1 *Stability prediction tools*

In their recent work, Gerasimavicius et al. benchmarked the performance of several computational tools designed for the prediction of ΔΔG upon mutation to see if they could predict the observed pathogenicity of missense mutations observed in the human population [115]. Interestingly, they included the ENCoM ΔS_{vib} predictions in their set of 13 tested stability prediction tools. They found that contrary to the popular assumption that pathogenic mutations lead to a loss in stability, all the tools tested performed better when the absolute value of the predicted ΔΔG was used. This observation could mean two things: either the tools are good at predicting the magnitude of change but bad at predicting the direction, or some pathogenic mutations actually make the protein more stable, which disrupts its function through other effects like the interaction with molecular partners. The authors also compared the stability prediction tools with tools specifically developed for the prediction of pathogenic missense mutations. Interestingly, while the best predictors overall are from this category, some are outperformed by the stability prediction tools. Out of the 12 tools specifically developed for pathogenic missense mutation calling, the ENCoM ΔS_{vib} outperformed 6 and was outperformed by the 6 others. FoldX, the best stability prediction tool according to their benchmark, outperformed an additional 2.

Among the 13 stability prediction tools tested by Gerasimavicius and coworkers, the ENCoM ΔS_{vib} ranked 5^{th} overall. We think this study represents a good blind test of the different methods, as the usual datasets of experimentally measured ΔΔG like ProTherm [116] are widely adopted, leading to the issue of the consensus methods possibly being over-specialized for the represented protein families. Thus, we will focus the discussion on the four stability prediction tools which outperform

the ENCoM ΔS_{vib} in the prediction of pathogenicity: FoldX [117], INPS3D [118], Rosetta [119] and PoPMuSiC [120]. In addition, we will discuss DynaMut [121] as it is a consensus predictor which includes the ENCoM ΔS_{vib} predictions as a component of the prediction. Each of these five stability-based predictors will be dedicated a subsection in the text that follows.

2.3.1.1 *DynaMut*

From the list of stability prediction tools tested by Gerasimavicius *et al.*, DynaMut and ENCoM stand out as the only tools considering protein dynamics in their predictions of stability change upon mutation [27, 115, 121]. The DynaMut approach is based on the integration of different stability change prediction algorithms within a random forest predictor. The integrated predictors include the authors' graph-based signatures as part of DUET [122], which is already a consensus of two other methods, SDM [123] and mCSM [124]. In addition, the dynamics component of DynaMut is the inclusion of the predicted change in vibrational entropy (ΔS_{vib}) using ENCoM. It was already shown by Frappier & Najmanovich that ENCoM exhibits good complementarity to enthalpy-based tools like FoldX and statistical consensus tools like PoPMuSiC [27], so it is unsurprising that the authors of DynaMut find good complementarity between their set of predictors. In addition to predicting $\Delta\Delta G$, their webserver also allows for NMA of the WT protein. However, it is unclear how this can add information to the analysis, and it is not performed by ENCoM but by one of the NMA models implemented by the Bio3D R package according to the user's choice [125]. It is slightly puzzling to us that they do not use ENCoM for the NMA analysis, as it is the only NMA model used for their $\Delta\Delta G$ predictions.

Interestingly, Gerasimavicius *et al.* tested both ENCoM and DynaMut as part of their set of stability prediction tools. Both tools have very similar performance profiles and led to the same surprising behaviour of their ROC curves, which have the false positive rate become higher than the true positive rate for higher values. Surprisingly, ENCoM outperforms DynaMut according to this analysis, which makes us question the soundness of the methodology used in the training of DynaMut. Indeed, since DynaMut is a consensus method that includes ENCoM predictions, it should not be outperformed by one of its singled-out components if the training led to true generalizability.

Beyond these caveats for the DynaMut method, let us remind that their integration of dynamics in the prediction of the effects of mutations is limited to changes in vibrational entropy, which represent a global property of the studied biomolecule. As the next chapters will outline, the ENCoM-DynaSig-ML pipeline considers a complete vector of fluctuations, the Dynamical Signature, for the predictions. Both approaches have merits and caveats; using the full DynaSig for prediction requires the availability of an experimental dataset probing sequence-function relationships

to train the ENCoM-DynaSig-ML pipeline but allows the learning of fine dynamical patterns that do not necessarily depend upon global changes in entropy; the use of changes in S_{vib} alone has the advantage of being directly applicable to any biomolecule of which the 3D structure is known, without the need for experimental datasets on its function.

2.3.1.2 *FoldX*

FoldX is a popular tool for protein engineering that can both model mutated protein structures and predict ΔΔG upon mutation, with the most recent version, FoldX 5.0, now also including parameters for RNA and small molecules [117]. At its core, FoldX is an empirical potential that was developed with the specific goal of predicting ΔG of folding from a high-resolution structure of a protein [126]. It contains 10 terms covering van der Walls interactions, solvation, H bonds, inter-molecular interactions, electrostatics and others. Beyond the specifics of the empirical potential, let us note that FoldX is entirely structure-based, and was developed as such: the potential evaluates a single minimum energy conformation for both the WT and mutant sequences in the evaluation of ΔΔG. Among the stability predictors tested by Gerasimavicius and colleagues, FoldX had the highest performance when taking the absolute value of the prediction [115]. In the past, the linear combination of FoldX with ENCoM ΔS_{vib} has led to complementary improvement over both methods alone for the prediction of experimental ΔΔG [27].

FoldX is routinely used in protein engineering and like all computational tools, its accuracy is far from perfect. However, according to a recent meta-analysis, it is able to reliably enrich for stabilizing mutations and thus reduce the amount of experimental work needed for protein engineering campaigns [127].

Methods such as FoldX and Rosetta, which are based on empirical force fields that recapitulate the enthalpy component of the Gibbs free energy, strike us as potentially complementary to the ENCoM Dynamical Signatures, especially when using the Entropic Signatures we introduce in Section 4.3.3. Indeed, these signatures describe the vibrational entropy at every position in the studied biomolecule and thus offer a way to capture both enthalpic and entropic properties in the same machine learning model.

2.3.1.3 *INPS3D*

INPS3D is a tool integrating information derived from a protein's 3D structure to the sequence-based stability predictor INPS [128] in order to improve the predictions [118]. The INPS predictor is a support vector machine integrating information about amino acid substitution through the Blosum62 matrix [129], along with 6 other sequence-based features describing change in molecular weight, hydrophobicity, and alignment-based features [128]. The added 3D descriptors correspond

to solvent accesibility of the native residue and a simple energy change computed according to the neighbouring residues of the mutated residue and a pairwise interaction potential [130].

INPS3D appears to be a simpler model than FoldX or Rosetta (described next), yet performs in-between the two in the Gerasimavicius study. This highlights how more detailed potential functions can sometimes hinder performance through the introduction of artifacts, for instance by making the resulting tools highly sensitive to slight incaccuracies in the input structures. INPS3D is also a purely static-based predictor and does not explicitly consider dynamics for its predictions, though one could argue that they are implicitly considered by the sequence-based features.

2.3.1.4 *Rosetta*

Rosetta is an all-atom empirical potential developed for both protein structure prediction and *de novo* protein design. Its potential is a sum of empirical and statistical energy terms, for a total of 19 specific terms of which the detailed description is beyond the scope of the present literature review [119]. Rosetta is routinely used to perform *in silico* mutagenesis and evaluate the impact of the mutation on the folding free energy. However, as mentioned in the above subsection, having such a detailed potential can lead to artifacts, which can explain the fact that Rosetta was outperformed by FoldX and INPS3D in the Gerasimavicius study.

2.3.1.5 *PoPMuSiC*

PoPMuSiC is a statistical structure-based potential consisting of a total of 16 terms [120]. Its only application is the prediction of $\Delta\Delta G$ upon mutation, and its set of 64 parameters was specifically fitted for that task. It has the advantage of being very fast since the statistical potential is simply a sum of the terms and depends on geometric properties such as distance and torsional angles between amino acids and the solvent accessible area of the mutated residue. Indeed, it can process many thousands mutations per minute. However, it seems that the webserver is no longer accessible under the name PoPMuSiC and the URL links to the "Dezyme" company, which appears to be selling protein design services.

As all other tools presented in this section except DynaMut, PoPMuSiC bases its prediction on structure alone. In the original ENCoM work, the ENCoM + PoPMuSiC combination demonstrated the highest overall performance on par with the FoldX3 + ENCoM combination for experimental $\Delta\Delta G$ prediction [27].

2.3.2 *Variant effect predictors*

In recent years, deep learning tools such as AlphaFold [131] and protein language models [132] have become much more prominent in the structural bioinformatics

field. Unsurprisingly, we can observe an associated rise in the use of such models to predict the effect of sequence variants on protein function. In a recent review, published in April 2022, Horne & Shukla detail the progresses and challenges associated with machine-learning based variant effect predictors (VEPs) [133]. They divide these predictors in four categories: fixed feature, unsupervised, supervised and metapredictors. The authors also state that despite great potential for the systematic integration of dynamical information within VEPs, to the best of their knowledge there exists no such tool, further reinforcing our previous statement about the novelty of the ENCoM-DynaSig-ML pipeline. In their own words [133]:

> *One foreseeable path forward in expanding the types of data models used is to include protein dynamics in predicting functional effects*. [...] *Mutations often shift the relative populations in the conformational ensemble, and acquiring and introducing this data into the ML pipeline would likely enhance VEP accuracy.*

Thus, we confirm the novelty of ENCoM-DynaSig-ML and we note that its integration within metapredictors will probably yield improvements for all existing metapredictors, as it adds dynamical information no other VEP considers. This high potential complementarity to all existing VEPs renders a detailed review unnecessary for the present thesis, and we instead refer the interested reader to the excellent review by Horne and Shukla [133].

3

THEORETICAL FRAMEWORK

3.1 COARSE-GRAINED NMA WITH ENCOM

The field of coarse-grained normal mode analysis emerged in 1996 with the seminal work of Monique Tirion, which championed the perhaps counterintuitive idea that global motions could be captured almost as well with a very simplified potential as with all-atom normal mode analysis [24]. Follow-up studies on different levels of coarse-graining for proteins found that a single bead per amino acid led to a good tradeoff between accuracy and performance [26].

In parallel, studies using more realistic coarse-grained potentials started to emerge. In fact, it is striking that the use of the Anisotropic Network Model (ANM) is still as widespread in the field nowadays [134], considering that the single feature the model takes into account is the cartesian coordinates of the coarse-grained beads. More realistic schemes of coarse-graining that can capture the underlying connectivity of the studied biomolecule have been developed. One of these models is the generalized Spring Tensor Model (STeM), on which ENCoM is based [135].

Both the Elastic Network Contact Model (ENCoM) [27] and the generalized Spring Tensor Model (STeM) [135] maintain the popular C_α coarse-graining, thus having computational costs comparable with the much simpler ANM. Early studies on coarse-grained normal mode analysis have put forward the hypothesis that the suprisingly high performance of simple cutoff-based models is explained by the fact that large-scale collective motions depend mostly on the geometry of the biomolecule [136]. Thus, including more information about such geometry in the model should lead to increased performance. In order to do so, STeM uses a four-term Gō-like potential, akin to classical molecular dynamics potentials minus the electrostatic term. However, let us remind the reader that this potential is at the level of entire amino acids. For instance, "bonds" represent the connectivity between consecutive residues, "angles" the angle between three sequential amino acids, etc. Both STeM and ENCoM share this potential function containing four terms: covalent bond stretching, angle bending, dihedral angle torsion and non-bonded (or long-range) interactions [27, 135]. Equation 3.1 gives the four terms of the ENCoM potential and Figure 3.1 illustrates them conceptually on the villin headpiece protein (PDB code 2RJY). Here is the ENCoM potential function:

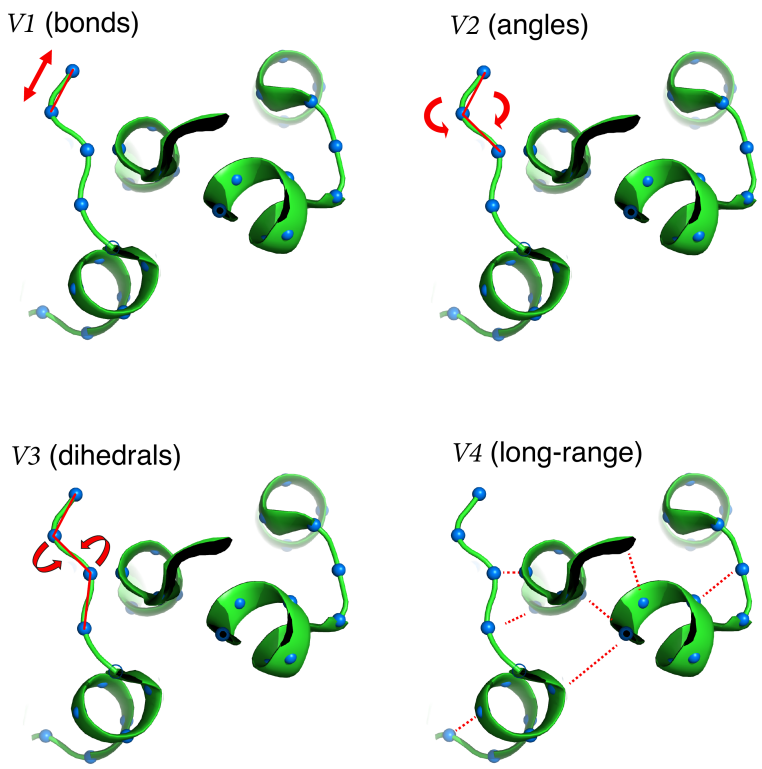


Figure 3.1: **The four terms of the ENCoM potential.** The structure of the villin headpiece (PDB code 2RJY) is used as an example. The beads are located on the C_α atom for every amino acid. V_1 restricts bond stretching, the elongation or contraction of the distance between two connected residues. V_2 restricts angle bending, the change in the angle formed by the beads located on three consecutive residues. V_3 restricts dihedrals (torsion angles), rotations of residues 1 and 4 about the axis passing through residues 2 and 3 from a quartet of connected residues. V_4 is the long-range interaction potential and restricts changes in distance between pairs of residues not covered by the other V_1-V_3 potentials. In the case of ENCoM, V_4 depends on both distance between the residues and the atomic surface area in contact complementarity term (details in Figure 3.2)

$$
\begin{aligned}
V_{\text{ENCoM}}(\vec{R}, \vec{R_0}) &= \sum_{\text{bonds}} V_1(r, r_0) + \sum_{\text{angles}} V_2(\theta, \theta_0) \\
&+ \sum_{\text{dihedrals}} V_3(\phi, \phi_0) + \sum_{i<j-3} V_4(r_{ij}, r_{ij_0}) \\
&= \sum_{\text{bonds}} \alpha_1 (r - r_0)^2 + \sum_{\text{angles}} \alpha_2 (\theta - \theta_0)^2 \\
&+ \sum_{\text{dihedrals}} \left[\alpha_3 (1 - \cos(\phi - \phi_0)) + \frac{\alpha_3}{2} (1 - \cos 3(\phi - \phi_0)) \right] \\
&+ \sum_{i<j-3} (\beta_{ij} + \alpha_4) \left[5 \left(\frac{r_{ij_0}}{r_{ij}} \right)^{12} - 6 \left(\frac{r_{ij_0}}{r_{ij}} \right)^{10} \right]
\end{aligned}
\tag{3.1}
$$

$\vec{R}$ and $\vec{R}_0$ are vectors of length $3N$ for a system with N beads, corresponding to the three-dimensional coordinates of every bead. $\vec{R}_0$ represents the input, or equilibrium, conformation, while $\vec{R}$ represents an arbitrary conformation. In the ENCoM potential, the non-bonded interaction term is modulated according to the surface area in contact between residues by the β_{ij} term:

$$\beta_{ij} = \sum_{k}^{N_i} \sum_{l}^{N_j} \epsilon_{T(k)T(l)} S_{kl} \tag{3.2}$$

$$\epsilon_{T(k)T(l)} = \begin{cases} \sigma_+ & \text{for favorable interactions} \\ \sigma_- & \text{for unfavorable interactions} \end{cases} \tag{3.3}$$

$\epsilon_{T(k)T(l)}$ represents the interaction between atoms of types $T(k)$ and $T(l)$ while S_{kl} is the surface area in contact between the two atoms, calculated using a constrained Voronoi procedure, as described by McConkey and co-workers [137]. This Voronoi procedure is very fast, yet takes into account all neighbouring atoms when evaluating the surface area in contact between a given pair of atoms. Moreover, the van der Waals radii are extended by the approximate radius of a water molecule (1.4 Å) to form what is termed the extended contact radius. It allows the implicit consideration of solvent-mediated contacts as two atoms can have a surface area in contact up to a distance given by the sum of their extended radii. For example, an oxygen atom forming a double bond (no hydrogen partner) has a van der Waals radius of 1.42 Å and thus and extended contact radius of 2.82 Å. To be considered in contact with another such atom, the pair could be as far apart as just under two times that distance, or 5.6 Å. However, the surface area at such a distance would be very small, or inexistent if there are neighbouring atoms blocking out this possibility.

Figure 3.2 illustrates the Voronoi procedure applied to a subset of atoms from a pair of interacting amino acids, at the 2D level for the sake of simplicity. In reality, Voronoi polyhedra are constructed instead of Voronoi polygons and the radical planes of intersection between spheres are used in place of lines of intersection. The final surface area is given by the projection of the contact planes on the surface of the sphere representing each atom. The interested reader can find algorithmic details of the procedure in the work of McConkey *el al.* [137].

The surface area in contact between two atoms is modulated according the atoms' types, for which the eight classes developed by Sobolev and co-workers [138] are used. The eight atom types are: hydrophilic, acceptor, donor, hydrophobic, aromatic, neutral, neutral-donor and neutral-acceptor. Only two types of interactions are considered in the model: favorable and unfavorable, associated with the respective

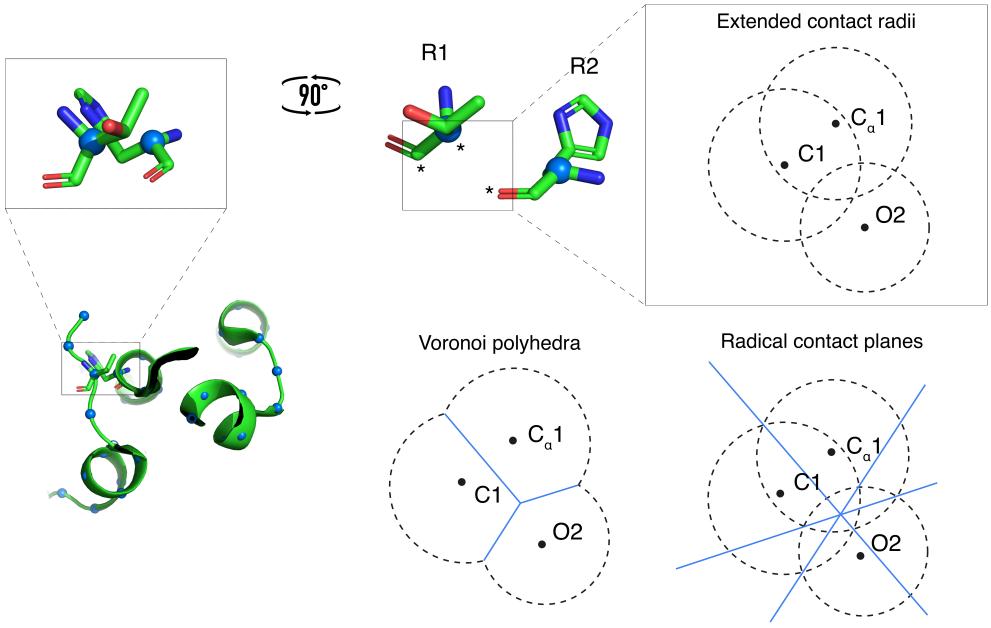


Figure 3.2: **ENCoM's surface complementarity term.** The structure of the villin headpiece (PDB code 2RJY) is used as an example. All heavy atoms are considered for the computation of ENCoM's surface complementarity (β_{ij}) term, but here we show them only for threonine 15 (labelled R1) and histidine 41 (labelled R2), for the sake of simplicity. For every heavy atom, the extended contact radius is the sum of its van der Waals radius and the radius of a water molecule (1.4 Å). Again for the sake of example, the radii for three atoms only are shown at the 2D level (discs instead of sphere surfaces). The three atoms shown are the backbone carbon and alpha carbon for R1 (C1 and $C_\alpha 1$) and the backbone oxygen for R2 (O2). From the extended contact radii, radical contact planes are computed at the intersection between each pair of spheres. From these contact planes, Voronoi cells can be built aroud every atom. The polygon at the intersection between two atoms (here line segment) represents the surface area in contact between these two atoms. It is projected onto the sphere surface to compute the surface in contact.

weight parameters σ_+ and σ_-. The interaction types for every pair of atom types are listed in Table 3.1.

3.1.1 *Solving the Hessian matrix*

The quadratic approximation is used in normal mode analysis, in which we consider only the quadratic terms of the Taylor expansion from the potential function. Thus, the normal modes and their associated squared frequencies are obtained as the eigenvectors and eigenvalues of the Hessian matrix of the potential function, as described elsewhere [114].

Table 3.1: **Interaction between atom types in the ENCoM potential.** The atom types are the ones from Sobolev and coworkers, divided in eight classes. The model considers an interaction to be either favorable (+) or unfavorable (-).

	Atom type	I.	II.	III.	IV.	V.	VI.	VII.	VIII.
I.	Hydrophilic	+	+	+	-	+	+	+	+
II.	Acceptor	+	-	+	-	+	+	+	-
III.	Donor	+	+	-	-	+	+	-	+
IV.	Hydrophobic	-	-	-	+	+	+	+	+
V.	Aromatic	+	+	+	+	+	+	+	+
VI.	Neutral	+	+	+	+	+	+	+	+
VII.	Neutral-donor	+	+	-	+	+	+	-	+
VIII.	Neutral-acceptor	+	-	+	+	+	+	+	-

3.2 MACHINE LEARNING MODELS

Since the ENCoM-DynaSig-ML occupies a novel niche as a variant effect predictor considering the effect of dynamics, we chose to compare two simple machine learning backends in the present thesis. The first is LASSO regression, arguably the most widespread and easily interpretable form of regularized linear regression. The second ML backend explored is multilayer perceptrons (MLPs), which can capture complex relationships between the input variables.

3.2.1 *LASSO regression*

Multivariate linear regression models the relationship between a series of input variables, or predictors X_i, and an outcome variable:

$$Y = \beta_0 + \beta_1 * X_1 + ... + \beta_i * X_n + \epsilon \tag{3.4}$$

where the model will learn the β coefficients that maximize the fit to the observed values of the outcome variable Y, and ϵ is the error. The usual procedure for the fit is to minimize the squared error of the prediction [139]. The problem with linear regression is that it can lead to poor generalizability due to high prediction variance, since the model can learn very high coefficients for input variables which do not vary much in the training set.

One of the simplest ways to deal with this problem is to introduce regularization to the model, which will minimize the sum of both squared error and a term depending on the size of the coefficients. The term LASSO was coined in 1996 by Robert Tibshirani and stands for "least absolute shrinkage and selection operator" [140].

Since LASSO is an acronym, we prefer to use of capital letters in the present work. As the name indicates, the regularization term introduced by LASSO penalizes the absolute sum of coefficients, so the objective function becomes:

$$\text{Minimize} : \lambda \sum |\beta_i| + \sum \epsilon^2 \tag{3.5}$$

where λ is the regularization strength, and standard multivariate linear regression is a special case at $\lambda = 0$.

Beyond its simplicity and elegance, LASSO regression has the property to drive coefficients to zero at high regularization strengths. This is the reason for our preference of LASSO over other types of regularized linear regression for the present thesis, as this feature selection makes the biological interpretation of the coefficients easier.

3.2.2 *Multilayer perceptron*

One caveat of linear regression models is in fact their linearity. Indeed, they assume linear independence between input variables, and in many cases this assumption is not true. This does not generate problems *per se*, but can lead to a loss in performance if there are complex relationships between the input variables and the outcome variable.

Artificial neural networks are machine learning models inspired by biological neural networks which are capable of learning arbitrarily complex relationships between input and outcome variables, provided they have sufficient numbers of neurons and enough training iterations are performed [141].

For the present thesis, we will investigate the use of multilayer perceptons (MLPs), a type of feedforward neural network with interconnected hidden layers of neurons [142]. MLPs have been applied to a large variety of problems, from diagnosis using genomic data [143] to weather forecast [144], demonstrating their ability to learn intricate nonlinear relationships happening between the input variables themselves and between these and the outcome variable(s).

Since the datasets studied in the present thesis are of relatively modest size, we will restrict the MLP architectures tested to having hidden layers of homogeneous size. The number of free parameters in an MLP with hidden layers of homogeneous size is given by:

$$F_{\text{params}} = 2 + S * P_{\text{input}} + 2 * S + \sum_{i=2}^{N_{\text{hidden}}} \left[S^2 + S\right] \tag{3.6}$$

where P_{input} is the number of input predictor variables (the size of the input layer), S is the size of every hidden layer and N_{hidden} is the number of hidden layers. Where possible, we will try to limit the number of free parameters to below the number of training set degrees of freedom, to prevent overfitting.

4

METHODOLOGY

The ENCoM-DynaSig-ML computational pipeline is composed of three parts: the ENCoM model, the generation of Dynamical Signatures, and the application of machine learning algorithms to predict biological properties from the DynaSigs. The ENCoM model was detailed in the last chapter (Section 3.1), while Section 4.1 will describe our adaptation of it for RNA molecules. The next sections of the present chapter will roughly follow the computational steps of a typical application of the pipeline: Section 4.2 details our approaches for modeling mutations to RNA and protein molecules; Section 4.3 gives the definitions of the two different types of Dynamical Signatures used in the present thesis, namely mean-square fluctuations and Entropic Signatures, and defines vibrational entropy; Section 4.4 will outline the general methodology for training ML algorithms on Dynamical Signatures and the specifics of the two models used in the present thesis, LASSO regression and multilayer perceptrons. These five sections present all methodology that is shared among the three case studies of Chapters 7-9, while a section at the beginning of each case study gives the specific methodology for that application.

As part of this thesis, we use a variety of elastic network model (ENM) performance metrics for the parameter sweep presented in Chapter 5. These include metrics widely used in the field and some original contributions, the most notable of which is our non-rotational-translational principal component analysis (nrt-PCA) correction (Section 4.5.3.1). Section 4.5 gives the definition of each metric. Since the parameters resulting from the re-parameterization presented in Chapter 5 are used throughout the remainder of the thesis, the methodology used is outlined in the present chapter, in Section 4.6.

Finally, an important contribution of the present thesis is the high usability of the complete ENCoM-DynaSig-ML pipeline, distributed as Python packages with online documentation and tutorials. Section 4.7 will briefly present these packages, NRGTEN [31] and DynaSig-ML [32], which are used throughout the results chapters that will follow.

ENCoM was originally developed for applications on proteins and notably the prediction of changes in protein thermal stability as a result of mutations [27, 28]. For the present work, we were interested in extending the applicability of ENCoM to study the dynamics of RNA molecules. In the case of proteins, it is generally believed that ENMs using a single bead per residue, situated on the C_α atom, still capture the essential low-frequency motions of the molecule [30]. This accepted property of coarse-grained ENMs is the reason why more beads per amino acid were never tested in the context of ENCoM. However, since RNA molecules are intrinsically more flexible per residue than proteins, using more beads per residue leads to an increase in the predictive power of the model [145]. Pinamonti and colleagues thoroughly inverstigated the relationship between the number of beads per nucleotide, the cutoff distance and the performance of an ANM model applied to four different RNA structures with varied topologies [146]. The authors used agreement with all-atom MD trajectories performed on these four molecules as a performance metric for the different models tested. In addition to coarse-grained models with all combinations of one, two or three beads located at the phosphate, sugar or base groups (seven total combinations), they tested a model with one bead per heavy atom. They found that the model with three beads per nucleotide, positioned at the C1′, C2 and P atoms for the sugar, base and phosphate groups respectively, offered a good tradeoff between performance and computational cost. Moreover, this coarse-graining scheme led to an optimal interaction cutoff of 9 Å, which is in close proximity to the 8 Å optimal interaction cutoff of found by Fuglebakk *et al.*, who screened interaction cutoffs from 8 to 23 Å for their ability to reproduce covariance matrices from MD trajectories of diverse proteins [147]. This proximity of the optimal ANM interaction cutoffs for the three-beads-per-nucleotide model and the standard C_α protein model makes the three-beads-per-nucleotide coarse-graining scheme attractive for the prospective study of RNA-protein complexes.

We adapted ENCoM to work on RNA molecules with the three-beads-per-nucleotide coarse-graining scheme of Pinamonti *et al.* [146]. We report similar findings with increased performance of ENCoM across all RNA benchmarks when using three beads per residue instead of one (data not shown). Since this was a pre-established fact, we use three beads per nucleotide for all applications of ENCoM to RNA tested thereafter in the present work. Figure 4.1 shows the positioning of the beads in the four standard nucleotides. The atom type assignation for A, U, C and G nucleotides is given in Table A.1.

The adaptation to RNA is fairly simple in the case of an ANM, as in the study by Pinamonti *et al.*; an additional bead is positioned at the cartesian coordinates of the selected atom, and the interaction cutoff can be varied to account for the changing

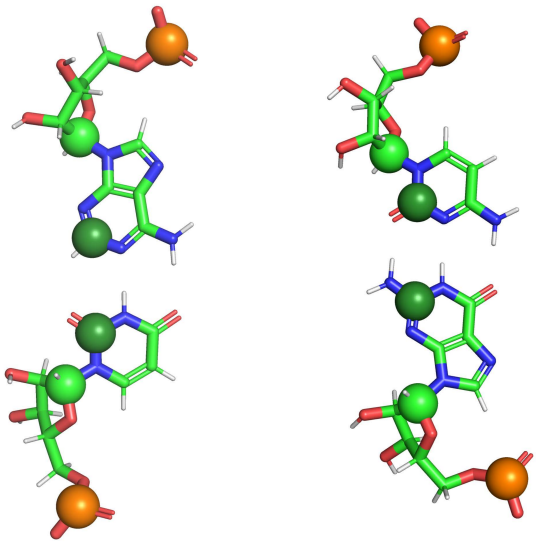


Figure 4.1: **Assignment of beads on the four standard nucleotides.** The nucleotides are arranged in two base pairs extracted A-form RNA helix generated using the MC-Fold and MC-Sym pipeline [148]. Phosphate atoms are in gold, C1′ carbons from the sugar group in light green and C2 carbons from the nucleobase in dark green. All three beads are included in RNA adaptation of ENCoM introduced in the present thesis.

density of beads [146]. However, the use of three beads per nucleotide in the context of ENCoM introduces ramifications. Since the connectivity between the beads is considered by ENCoM, these ramifications significantly change the implementation logic. Indeed, proteins are linear when coarse-grained at the C_α level and thus the V_2 and V_3 terms of the potential, representing angle bending and dihedral torsions, can be computed by taking all groups of respectively three or four consecutive beads that are part of the same protein chain. Figure 4.2 illustrates the ramifications in the connectivity of the beads introduced by the RNA adaptation, along with all types of angles and dihedrals introduced. As part of the present thesis, we have extended ENCoM so that beyond the proper treatment of these RNA-induced ramifications, the computation of the V_2 and V_3 terms is now generalized. As such, any coarse-graining approach will be properly treated, and it is straightforward to test novel coarse-graining approaches for any biomolecule, with the necessary input files described in the online NRGTEN documentation (Section 4.7.1).

4.2 MODELING MUTATIONS

In order to study dynamics-function relationships with the ENCoM-DynaSig-ML pipeline, one needs a large enough dataset of experimental measures for sequence variants of the studied biomolecule. The sequence variants then have to

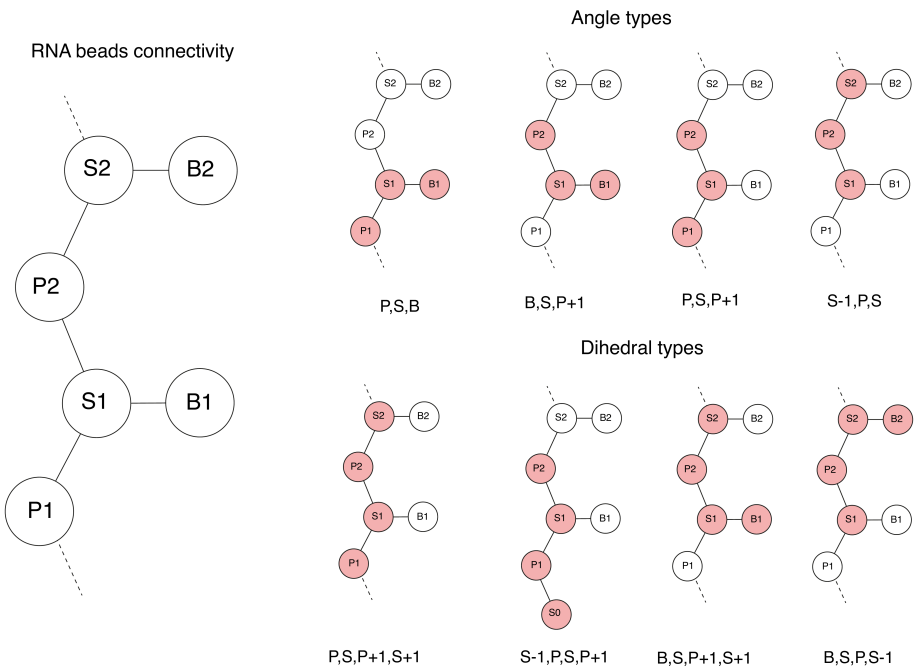


Figure 4.2: **RNA beads connectivity, types of angles and types of dihedrals.** In contrast to proteins, which are completely linear when coarse-grained at the C_α level, the use of three beads per nucleotide in the case of RNA introduces a ramification at the sugar bead. This leads to four different types of both angles and dihedrals. The phosphate, sugar and base beads are respectively represented by the letters P, S and B. The angle and dihedral types are named with these letters and relative numbering. The numbering corresponds to the position of a nucleotide in the RNA chain.

be modeled on the experimentally resolved structure of the WT biomolecule, or on a predicted 3D model if no experimental structure is available. Keeping in mind that the ultimate application of ENCoM-DynaSig-ML is the prediction of functional properties for theoretical variants, we prioritized fast modeling procedures in the present work. Indeed, a single DynaSig computation on a sequence variant takes between 3 to 5 seconds CPU time for all three case studies presented in Chapters 6-8. In order to take advantage of this speed, we restricted ourselves to modeling procedures which are either faster or equivalent to that time.

The next two subsections will present our approaches for RNA mutation and for protein mutation. These two approaches are used throughout the thesis to perform all *in silico* mutagenesis.

4.2.1 *RNA mutations*

ModeRNA is a tool developed for the template-based modeling of 3D RNA structures [149]. The problem of modeling a sequence variant on a given 3D structure is a special case of template-based modeling where the alignment between target and template has no gaps and no insertions, only mismatches. Thus, ModeRNA also allows the easy and fast modeling of nucleotide substitutions in RNA structures. We model all RNA sequence variants with ModeRNA by simply constructing a one-to-one alignment of the variant sequence with the WT sequence and using the WT 3D structure as template. Since there are no insertions or deletions in the alignment, ModeRNA simply substitutes the bases in place by restricting the rings to be in the same plane. We do not perform energy minimization for two reasons:

1. As mentioned, we want to optimize for speed in order to allow for ultra-high-throughput predictions on theoretical sequence variants.

2. Keeping the backbone positions the same between all variants makes the ENCoM DynaSigs the result of base substitution alone (V_4 potential). Changing backbone positions would affect all terms of the potential and could thus introduce some unwanted noise in the DynaSigs.

Moreover, our application to RNA in the present thesis is the study of pri-miR-125a presented in Chapter 6. As outlined in that chapter, we restrict the analysis to sequence variants having the same predicted 2D structure as the WT 2D structure, which means all base pairing patterns are constant across all variants. Since the ENCoM surface complementarity term does not depend on precise distances between atoms (see Section 3.1), we make the hypothesis that it can recapitulate the base pairing and stacking interactions well enough on the substituted nucleobases, without the need for energy minimization.

4.2.2 *Protein mutations*

We use another template-based modeling tool, MODELLER, for the generation of protein sequence variants [150]. However, the size of amino acid side chains varies significantly across all 20 types, from 0 to 10 heavy atoms, which means that steric clashes or unnatural interactions could be introduced by in-place substitutions. For this reason, we use the `mutate_model.py` script from MODELLER (salilab.org/modeller/wiki/Mutate model), which optimizes the mutated residue's energy. The script mutates one position at a time, but the dataset of VIM-2 lactamase sequence variants we present in Chapter 8 was generated using deep mutational scanning so it does not contain variants with more than one mutated position. For the search of high fitness VIM-2 variants containing numerous mutated positions,

we proceed by sequential mutation in an evolution-mimicking search procedure, so again the `mutate_model.py` MODELLER script can be used.

4.3 DYNAMICAL SIGNATURES

We have defined the complete structural dynamics of a biomolecule as the set of all its possible conformations associated with their respective probabilities of occuring. Exhaustively enumerating these complete dynamics, which include states with very low probabilities, is currently beyond our computational reach for all but the smallest biomolecules (very small peptides or ligands). However, normal mode analysis (NMA) allows for the complete description of a subset of harmonic, oscillatory dynamics (normal modes). For a biomolecule represented by N beads (coarse-grained or all-atom), $3N$ normal modes exist, each of which is a vector of length $3N$ with one associated eigenvalue. Thus, under the NMA approximation, structural dynamics are fully captured by around $3N * 3N = 9N^2$ values for a biomolecule of length N ($3N$ normal modes, each of length $3N$). While these NMA structural dynamics represent a tremendous gain in space and time compared to those arising from techniques such as molecular dynamics, their size is still very significant. For instance, a 250 amino acid protein gives rise to a linear space characterized by 563 250 scalars.

Thankfully, these huge linear spaces can be reduced to more tractable size. Perhaps the simplest way to do so is the computation of a Dynamical Signature: a vector of fluctuation at every bead in the system. We prefer the term Dynamical Signature for two reasons:

1. It encompasses the general concept of reducing structural dynamics to a vector of the same length as the studied biomolecule, without emphasizing a specific way of doing so. We use two different types of Dynamical Signatures in the present thesis, but the concept extends beyond these two.

2. It also puts forward the notion that these vectors are characteristic of certain molecules or certain functions.

Dynamical Signatures are the central part of the ENCoM-DynaSig-ML pipeline. For the present thesis, we use the classical mean-square fluctuations (MSF) and also introduce Entropic Signatures (EntroSigs). Since the EntroSig is based on vibrational entropy, let us introduce it before defining both MSF and EntroSig in the sections that will follow.

4.3.1 *Vibrational entropy*

Under the harmonic oscillator approximation, the vibrational entropy of a biomolecule can be computed from the eigenvalues associated with the nontrivial normal modes. In the present thesis, we use the classical vibrational entropy formula, which arises from the vibrational partition function of a harmonic oscillator [151]:

$$S_{\text{vib}} = \sum_{n=7}^{3N} S_{\text{vib}n} \tag{4.1}$$

$$S_{\text{vib}n} = \frac{\beta\nu_n}{e^{\beta\nu_n} - 1} - \ln\left(1 - e^{-\beta\nu_n}\right) \tag{4.2}$$

$$\nu_n = \frac{1}{2\pi}\sqrt{\lambda_n} \tag{4.3}$$

$$\beta = \frac{h}{kT} \tag{4.4}$$

where ν_n is vibrational frequency of the n^{th} normal mode, computed from its associated eigenvalue λ_n. β is a thermodynamic scaling factor, h is Planck's constant, T is the temperature and k is Boltzmann's constant.

In the original ENCoM version, the $h\nu \ll kT$ approximation was made, which leads to a simpler S_{vib} formula that linearly depends on temperature. However, since we are dealing with a pseudo-physical system the h, k and T quantities are hard to estimate. This is the reason for our preference of the classical formula in the present thesis. Furthermore, the β scaling factor is necessary to our Entropic Signature, which we will define below, after the introduction of the classical mean-square fluctuations.

4.3.2 *Mean-square fluctuations*

From the eigenvectors and eigenvalues of the nontrivial normal modes, the mean-square fluctuations of individual residues can be computed [152]:

$$\text{MSF}_i = \sum_{n=7}^{3N} \frac{E_{n,i,x}^2 + E_{n,i,y}^2 + E_{n,i,z}^2}{\lambda_n} \tag{4.5}$$

where $E_{n,i}$ represents the xyz diplacement of bead i in the n^{th} eigenvector and λ_n the associated eigenvalue of that eigenvector. These mean-square fluctuations (MSF)

of beads in the system have been widely used in the normal mode analysis field as so-called "predicted B-factors", and correlated with experimentally measured temperature factors (B-factors) from X-ray crystallography [25, 26, 153]. The MSF directly arise from the energy potential of the system as formally proved in 1990 by Nobuhiro Gō [152], and temperature has no impact on these fluctuations except to scale them linearly.

4.3.3 *Entropic Signatures*

Despite the invariance of relative MSF with regards to temperature, the contribution of each normal mode to vibrational entropy does depend on temperature, with the first nontrivial mode contributing most of the entropy at temperatures approaching zero as is apparent from Equation 4.2. Indeed, β tends to infinity as the temperature drops near zero and thus the contribution of the lowest frequency becomes the dominant driver of vibrational entropy. For this reason, we hypothesized that scaling the square fluctuations at every position by the entropic contribution of that normal mode would give the model the power to better capture fluctuations happening in biological contexts. Indeed, since ENCoM and other coarse-grained ENMs have pseudo-physical potentials, they do not incorporate a temperature with physical units and thus the possibility to scale the contribution of the normal modes to the Dynamical Signature according to a Boltzmann distribution is attractive. Moreover, from a functional point of view, the entropy at each residue can be of great significance and exhibit high complementarity to enthalpy-based properties computed by other models, such as the MC-Fold enthalpy of folding used in Chapter 6 or the predicted $\Delta\Delta G$ of folding from Rosetta [154] used in Chapter 8. In order to satisfy these desired properties, we introduce the Entropic Signature (EntroSig), which scales the square displacements at every bead in the system by the contribution of the corresponding nontrivial normal mode to the vibrational entropy of the molecule (defined in Section 4.3.1):

$$\text{EntroSig}_i = \sum_{n=7}^{3N} S_{\text{vib}n} \left(E_{n,i,x}^2 + E_{n,i,y}^2 + E_{n,i,z}^2 \right) \tag{4.6}$$

where $E_{n,i}$ represents the xyz diplacement of bead i in the n^{th} eigenvector. $S_{\text{vib}n}$ is defined in Equation 4.2 and importantly depends on a thermodynamic scaling factor β.

This β scaling factor allows for varying the relative contributions of high- and low-frequency normal modes to the fluctuations of the individual beads, with higher values of β leading to contributions from the lowest-frequency modes dominating and lower values leading to a more nuanced distribution across all modes. For all molecules for which we have studied EntroSigs, we observe that it is always

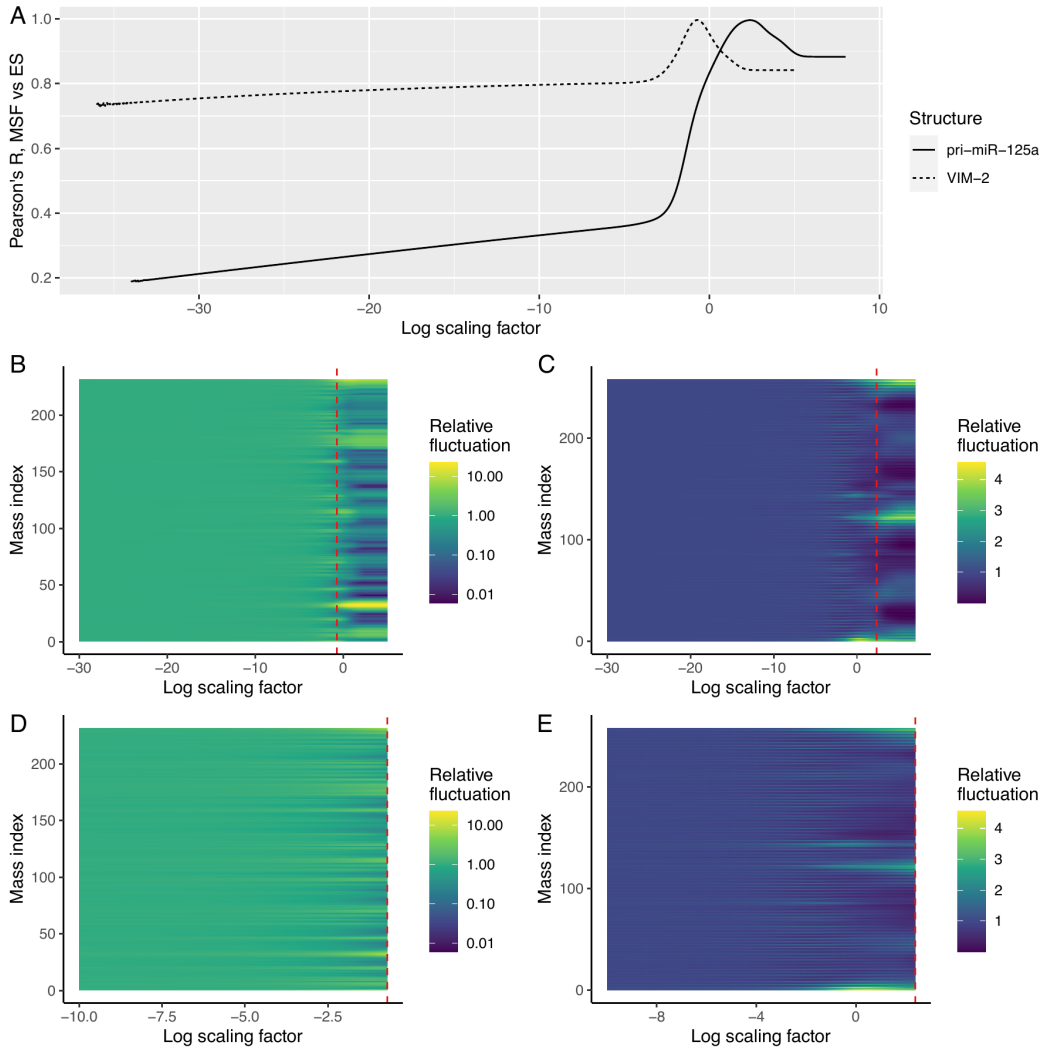


Figure 4.3: **Comparison between MSF and EntroSigs.** The Entropic Signatures were computed for the predicted 3D structure of pri-miR-125a selected in Chapter 6 (MC-Sym model 61) and the crystal structure of VIM-2 lactamase (PDB code 4BZ3). The widest possible range of scaling factors was tested, in log increments of 0.01. The incrementation was stopped in both directions when the output became constant. A) The Pearson correlation between the Entropic Signature and the mean-square fluctuations as a function of the β thermodynamic scaling factor for the EntroSig. Almost perfect agreement is reached for VIM-2 at $\beta = e^{-0.71}$ and for pri-miR-125a at $\beta = e^{2.39}$. B)-C) VIM-2 and pri-miR-125a EntroSigs, respectively, as a function of β. The red dashed line shows the position of almost-perfect agreement (Pearson's R > 0.995) with MSF. C)-D) Same as B)-C), zoomed in on β values at and lower than the almost-perfect agreement with MSF.

possible to find a value of β which gives almost exactly the same relative coefficients as the MSF.

For example, Figure 4.3 shows the correlation between the MSF and the EntroSig for different scaling factors, for two very different selected structures, the predicted pri-miR-125a 3D model which we selected as the result of our analyses in Chapter 6 and VIM-2 beta-lactamase experimental structure which is the starting structure in Chapter 8. pri-miR-125a is a rod-shaped RNA hairpin structure while VIM-2 is very globular. The β values at which their EntroSig perfectly agrees with their MSF are far apart, with pri-miR-125a reaching agreement at $\beta = e^{2.39}$ and VIM-2 at $\beta = e^{-0.71}$. Looking at Equation 4.2 it Equation 4.3, it is apparent that agreement happens for the scaling factor which leads to:

$$S_{\text{vib}n} = \frac{1}{\lambda_n} \tag{4.7}$$

4.4 SIGNATURES INSIDE MACHINE LEARNING ALGORITHMS

4.4.1 *DynaSig standardization*

The two machine learning models which we explore in the present thesis, LASSO regression and multilayer perceptrons (MLPs), were detailed in Section 3.2.1 and Section 3.2.2. When training these models on DynaSigs, we first standardize each position in the DynaSig using all values encountered in the dataset. This is to decouple both the average fluctuation value of the positions and their variance across the tested sequences from their importance in the model. The coefficients of the LASSO model are readily interpretable, as they now directly relate the predicted outcome to the standard deviation of the fluctuation at each position. For example, a coefficient of 1 at a specific position means that the predicted outcome will be 1 more for every standard deviation in fluctuation at that position.

The usage of the whole dataset to standardize the DynaSigs does not introduce information about the testing set to the model, since only average properties are obtained from it. The reason for including both training and testing DynaSig values in the standardization is that we often construct these sets so that specific positions are only mutated in the testing set. Standardizing on training set alone would introduce some very high standardized values for these positions in the testing set, since the biggest effect a mutation has is most often local.

4.4.2 *Binary classification performance*

In all three case studies chapters, we analyze the performance of ENCoM-DynaSig-ML both in terms of the goodness of fit of the predictions to the observed experimental data, and by reducing the data to a binary classification problem. The two performance metrics that we use, the area under the receiver operating charac-

teristic curve, and the area under the precision-recall curve, will be introduced here.

4.4.2.1 *Receiver operating characteristic curves*

The receiver operating characteristic (ROC) curve plots the relationship between false positive rate and true positive rate for the classification of binary outcomes, at varying decision thresholds. For a classification problem with equal proportions of both classes, the expected area under the ROC curve (AU-ROC) for a random classifier is 0.5; a perfect classifier has an AU-ROC of 1 [155]. However, when class imbalance is present, AU-ROC can give the impression of good classification when the underlying model has just learned the class proportion but is not making meaningful predictions. Nonetheless, we include AU-ROC as it is a popular metric in the machine learning field, and we take care in using statistical simulations to sample the performance of random classifiers in order to assess the significance of our predictors when there is class imbalance in our prediction problems.

4.4.2.2 *Precision-recall curves*

Another method for evaluating classification performance is the area under the precision-recall curve (AU-PR). The precision-recall curve, as its name implies, plots precision as a function of recall. Precision is the proportion of positive guesses that are true, and recall is the proportion of relevant elements found. A random classifier is expected to have an AU-PR equal to the proportion of positive elements in the dataset [156].

4.4.3 *Softening/rigidifying biases of LASSO models*

When looking at the coefficients learned by LASSO models for the Entropic Signature input variables, we can readily interpret them in terms of structural dynamics: negative coefficients mean that sequence variants in which these positions are more rigid than the average lead to a higher predicted functional property, while the reverse is true for positive coefficients: variants leading to these positions being more flexible will have higher predicted functional properties. We define the rigidifying/softening bias of the LASSO model as the sum of Entropic Signature coefficients divided by the absolute sum of coefficients, expressed in percentage. If the sum is negative, we reverse the sign and call it a rigidifying bias; if it is positive, we call it a softening bias. These biases identify general trends for the whole biomolecule in terms of the predicted property.

4.5 ENM PERFORMANCE METRICS

In order to assess the performance of ENMs, various metrics have been developed as the field matured. In the present thesis, four metrics will be used as part of the parameter search presented in Chapter 5:

1. Pearson correlation between experimental B-factors and Dynamical Signatures (MSF and EntroSig).

2. Cumulative overlap between a set of low-frequency normal modes and a sampled conformational change.

3. Normalized cumulative overlap (NCO) between a set of low-frequency normal modes and motions apparent from solution NMR ensembles.

4. Root-mean-square error from a linear fit of predicted change in vibrational entropy to experimental $\Delta\Delta G$ of folding as the result of mutation.

The next subsections will describe each metric and outline the contributions of the present thesis, which are centered around the selection of the slow-frequency normal modes, the use of a more accurate vibrational entropy computation, the introduction of the Entropic Signature and the non-rotational-translational correction to principal component analysis (nrt-PCA). nrt-PCA is a correction to standard PCA that is needed in order to allow the relevant comparison of nontrivial normal modes with motions from an NMR ensemble, especially in the case of NCO which weights the cumulative overlaps by the percentage of variance explained by the principal components. The nrt-PCA correction was first introduced in our work presently published as a preprint; some parts of the following sections are taken directly from that work [157].

4.5.1 *Pearson correlation with experimental B-factors*

Experimental B-factors capture fluctuations of atomic positions in the crystallized biomolecule (Section 2.1.1). As outlined in Section 4.3.3, it is commonplace in the coarse-grained normal mode analysis field to assess the models' performance by correlating the computed mean-square fluctuations (MSF) at every residue with experimental B-factors from X-ray crystallography (Equation 4.5) [25]. We also introduced the Entropic Signature (EntroSig), in which the square fluctuations at every residue are scaled by the vibrational entropy of every nontrivial normal mode (Section 4.3.3). The EntroSig thus allows to vary the relative contribution of the slow-frequency normal modes compared with the high-frequency normal modes with a Boltzmann scaling factor as part of the vibrational entropy computation (Section 4.3.1).

Whichever Dynamical Signature type is used, MSF or EntroSig, the Pearson correlation is then calculated between the experimental B-factors and the computed Dynamical Signature. Crystallographic B-factors are reported for every resolved atom, whereas ENMs represent many atoms with a single bead (one bead per amino acid or three beads per nucleotide in the case of ENCoM). Thus, a mapping has to be made to reduce the experimental B-factors to the same lenght as the Dynamical Signature's length. In the present thesis, the chosen method is to average the B-factors for all atoms considered as part of the same bead in the system. In ENCoM, this is an unambiguous mapping as the potential already considers all atoms in the computation of the surface area complementarity term.

4.5.2 *Individual conformational changes: overlap and cumulative overlap*

At it simplest expression, a biologically relevant conformational change can be described using two conformations of a biomolecule, called the input and target conformation. For instance, the activation of GPCRs can be described in such a simplified way: the input conformation is the inactive state, and the target conformation is the active state [51]. The overlap metric is a measure of the similarity between an eigenvector $\vec{E_n}$ predicted using the start conformation and the displacement vector $\vec{R}$ calculated between the coordinates of the target and input conformations, after both have been superimposed [23, 158].

$$\text{overlap}(\vec{E_n}, \vec{R}) = \frac{|\vec{E_n} \cdot \vec{R}|}{\|\vec{E_n}\| \|\vec{R}\|} \tag{4.8}$$

For each eigenvector, it has a value between 0 and 1, which describes how well we can reproduce the target conformation by deforming the start conformation along the eigenvector. A common practice is to measure the maximum overlap between the experimental conformational change and the first N slowest normal modes. However, this technique fails to capture a difference between a set of normal modes which collectively capture the change very well and another set in which only one mode has significant overlap with the observed change. For this reason, we prefer to use of cumulative overlap (CO) in the present thesis as it corrects this artifact. The CO is also a value between 0 and 1 and describes how well a set of orthogonal motions (eigenvectors) can collectively reproduce the target conformation from the start conformation.

$$\text{CO} = \sqrt{\sum_n \text{overlap}(\vec{E_n}, \vec{R})^2} \tag{4.9}$$

4.5.2.1 *Linear proportion of normal modes*

To the best of our knowledge, all articles investigating overlaps between normal modes predicted by ENMs and experimentally sampled conformational changes do so using a fixed number of the slowest normal modes [26, 146, 159]. This has the clear advantage of representing a linear space of the same number of dimensions for every tested biomolecule, and would give rise to the same number of conformations is they were enumerated from the start conformation using a given RMSD step for each normal mode (grows exponentially with the number of modes). However, the number of normal modes grows linearly with the number of beads in the system ($3N - 6$ nontrivial modes for N beads). This means that using a fixed number, say 10 or 20 normal modes as is commonplace in the field [159], artificially favors benchmarks with smaller structures as a bigger proportion of the orthogonal space spanned by all normal modes gets used in the comparison of the predicted motions with experimentally validated conformational changes. We first proposed to circumvent this problem by using a linear proportion of nontrivial normal modes. This correction does not favor the use of smaller biomolecules in benchmarking ENM performance and does not come with significant added computational cost for computing the cumulative overlap. Depending on the proportion chosen and the size of the studied biomolecules however, it may give rise to a large number of low-frequency normal modes being selected. In a typical use case, it may not be feasible to exhaustively enumerate conformations from such a high number of normal modes, however neither is it for even the 10 slowest modes at 5 deformation steps per mode (which give rise to 5^{10} exhaustively enumerated conformations, or over 9 million). As such, we maintain that the use of a low linear proportion of the total nontrivial normal modes, such as 5% as we use in the present thesis, should be used for ENM benchmarking purposes. This practice should remove the intrinsic bias towards smaller structures of the traditional fixed number of slow-frequency modes.

4.5.3 *Conformational ensembles: normalized cumulative overlap*

Solution NMR experiments are perhaps the most reliable source of experimental biomolecular structural dynamics. For small enough molecules, the obtained restraints allow the construction of an ensemble of conformations that represent the motions of the studied biomolecule in solution. One way of describing the motions apparent within the ensemble is to apply principal component analysis (PCA) [160] to the cartesian coordinates of the structures in the ensemble. However, as discussed earlier, PCA used in this way introduces rotational and translational motions in the components obtained. In normal mode analysis, these types of motions are captured by the first 6 trivial modes, which are discarded as they have zero-valued eigenfrequencies. Furthermore, these motions are irrelevant in the context of biomolecular structural dynamics happening in solution, where they

happen randomly as a result of motion within the solvent. Therefore, we proposed the non-rotational-translational PCA correction (nrt-PCA) as part of our recent work adapting ENCoM for RNA. Let us introduce PCA along with the the nrt-PCA correction before moving on to explain normalized cumulative overlap.

4.5.3.1 *Non-rotational-translational principal component analysis*

Principal component analysis (PCA) is a statistical technique that is used to transform observations of variables that may be correlated into linearly uncorrelated variables which are called principal components (PCs) [160]. PCA is commonly used in the normal mode analysis field to extract dominant motions apparent within the conformational ensembles obtained from solution NMR experiments [159]. Each conformation is represented as a vector of length 3N where N is the number of beads in the system. In the present work, PCA is computed on these vectors using a singular value decomposition (SVD) algorithm [161]. The PCs obtained are analogous to normal modes in that they are the eigenvectors of the covariance matrix of the 3N coordinates from the ensemble of structures. The first PC describes the largest proportion of the variance in the ensemble, and each subsequent component captures the largest proportion of the remaining variance and is orthogonal to the preceding components. However, when there are more than two conformations in the ensemble, rotational and translational motions can be present in the principal components. To our knowledge, the published workarounds in the context of biomolecular 3D dynamics introduce a change of coordinates, which is undesirable in the context of cartesian normal mode analysis [162]. We thus introduce a correction to obtain cartesian principal components without rotational and translational degrees of freedom. First, standard PCs are computed from the ensemble, with all conformations superimposed to the first model. Then, the first six rotational and translational normal modes from that first model are used as the starting basis for Gram-Schmidt orthonormalization [163] of the PCs. This transformation ensures all rotational-translational motions from the PCs are removed and all relevant internal motions are maintained. However, the proportion of variance explained needs to be corrected for each PC according to the amount of rotational-translational variance initially present:

$$c_i = \frac{v_i \left(1 - \sqrt{\sum_{j=1}^{6} (\vec{\mathrm{PC}}_i \cdot \vec{\mathrm{RT}}_j)^2}\right)}{\sum_n c_n} \tag{4.10}$$

where v_i is the initial proportion of variance explained by $\vec{\mathrm{PC}}_i$, c_i is the corrected proportion of variance explained, and $\vec{\mathrm{RT}}_j$ are the six rotational-translational normal modes of the first model in the structural ensemble. The PCs are then reordered in decreasing order of corrected proportion of variance explained. We call this method nrt-PCA (non-rotational-translational PCA) and whenever we refer to PCA

or PCs, it is implied that nrt-PCA is used. It ensures the accurate comparison of non-trivial normal modes computed from a representative structure with strictly internal motions apparent from the ensemble. In most cases, nrt-PCA will not lead to drastically different PCs and will just remove a small proportion of rotational-translational motions from each PC.

4.5.3.2 *Normalized cumulative overlap*

When an ensemble of experimentally validated conformations is available, such as what typical NMR solution experiments provide, nrt-PCA is computed on the cartesian coordinates of the ensemble as described in the last section. The same atoms as the ones used in the coarse-grained ENM representation tested are used in order to have the same dimensionality. The normal modes are computed on the minimum energy conformation, which by convention is the first model in the ensemble submitted by the NMR spectroscopists to the PDB [164]. The normalized cumulative overlap (NCO) between the first N normal modes and the first M PCs is given by:

$$\mathrm{NCO} = \sum_{j=1}^{M} \left[v_j \sqrt{\sum_{i=1}^{N} \mathrm{overlap}(\vec{E}_i, \vec{\mathrm{PC}}_j)^2} \right] \tag{4.11}$$

where $\vec{E}_i$ is normal mode i, $\vec{\mathrm{PC}}_j$ is principal component component j and v_j is the proportion of variance explained by component j. NCO ensures a value between 0 and 1 as both the normal modes and the components are orthogonal with respect to themselves, and v_j sums to 1 over all the components.

4.5.4 *Root-mean-square error from ΔΔG of folding predictions*

ENCoM is unique among coarse-grained ENMs in its ability to predict the effect of mutations, even when such mutations do not affect the backbone geometry of the studied biomolecule. The first application of this sequence sensitivity was to predict the effect of mutations on the vibrational entropy change, ΔS_{vib}. It was shown that ΔS_{vib} can be used as a linear approximation of the experimentally measured ΔΔG of folding induced by the mutation [27]. In order to use the most stringent metric possible, the authors used a simple linear fit between ΔS_{vib} and experimental ΔΔG of folding, forcing the intercept to be zero. The root-mean-square error (RMSE) between the linear fit and experimental values was then computed, with lower values representing better performance.

We use the same approach in the present work, however since we changed the formula for the computation of vibrational entropy for a more realistic one (Section 4.3.1), it now includes a thermodynamic scaling factor. In a similar manner as

for Pearson correlation of Entropic Signatures with B-factors (Section 4.5.1), this thermodynamic scaling factor β is varied across a range of log spaced values, and the mean RMSE across the whole dataset of mutations is computed once for every β value, preventing the use of one overfitted β value for every example in the dataset.

4.6 DIVERSE BENCHMARK: ENCOM RE-PARAMETERIZATION

ENCoM exhibits robust performance across a wide range of parameter combinations [27]. However, in the present work we have extended its applicability beyond proteins to nucleic acids, small molecules and all complexes of these biomolecules. This prompted us to ask whether there exists a significant tradeoff in parameter space between applications to different biomolecular types, in particular between proteins and RNA. Conversely, it could also be that a single set of parameters can reasonably cover most applications. Finally, as part of the initial development of ENCoM, it was stated that ENCoM has 4 parameters, while in truth the model has 6 parameters. This fact alone is enough to prompt re-parameterization and ensure good fit to experimental data.

To answer the question of whether a tradeoff exists between RNA and protein applications, we decided to run three benchmarks on proteins, three benchmarks on RNA and one on RNA-protein complexes, for a total of seven diverse benchmarks. Each of these benchmarks contains between 38 and 313 structures, for a total of 940 structures, on which ENCoM was run for a total of 221 255 parameter combinations. The three protein benchmarks come from the original ENCoM publication [27] and measure correlation with experimental B-factors, conformational change prediction, and the prediction of folding $\Delta\Delta G$ upon mutation. The three RNA benchmarks come from our work extending ENCoM to RNA [157] and measure correlation with experimental B-factors, conformational change prediction, and NMR ensemble structural variance prediction. The RNA-protein benchmark was constructed as part of the present re-parameterization and measures NMR ensemble structural variance prediction. The next subsections will list the details of how each individual benchmark was constructed and how the metrics were computed. The list of PDB codes for the experimental structures and ensembles in each benchmark can be found in the Appendix.

4.6.1 *Protein experimental B-factors*

For constructing the protein experimental dataset, we started from 113 non-redundant, high-resolution protein structures selected by Kundu *et al.* [165], which were used in the original ENCoM parameterization [27]. We downloaded each structure directly from the PDB, and removed all non-protein residues (ions, water molecules, ligands). Two PDB codes, 3B5C and 4PTP, were listed as obsolete by

the PDB and were replaced by their listed superseded entries, respectively 1CYO and 5PTP. In order to reduce computational cost, the structures with more than 300 amino acids were removed, leaving us with 80 high-resolution protein structures. The PDB codes for these 80 structures are given in Table A.2. No sequence clustering was performed as the entries are non-redundant.

The performance metric used is the Pearson correlation between the experimental B-factors, averaged by amino acid, and the ENCoM Dynamical Signatures. As outlined in Section 4.3.3, these can be of two types:

1. Mean-square fluctuations (MSF), which are temperature-independent when normalized and arise as a unique vector for a given set of eigenvectors and eigenvalues.

2. Entropic Signatures, which depend on as thermodynamic scaling factor and are introduced as part of the present thesis (Section 4.3.3).

For each set of parameters, ENCoM is run on every of the 80 selected structures. Then, the computed eigenvectors and eigenvalues are used to compute the MSF and a set of Entropic Signatures, with scaling factors ranging from e^{-5} to e^{5} in log increments of 0.25. For each scaling factor, the mean Pearson correlation across the whole set of structures is the value we keep, because using the best correlation for each structure could lead to overfitting specific features of individual structures. The best of these mean correlations, whether from MSF or EntroSig, is the final value representing the performance of a parameter set on this first benchmark.

4.6.2 *Protein conformational change*

Besides crystallographic B-factors, another source of information on biomolecular structural dynamics is the crystallization of the same biomolecule under different conditions, therefore changing its energy landscape and favoring a different conformation as the minimum energy conformation. In the case of proteins, the PSCDB (Protein Structural Change upon ligand binding DataBase) has repertoried 714 pairs of X-ray crystallography experiments where the same protein was crystallized both with and without a ligand [75]. Of these 714 pairs, 403 include significant conformational changes between the ligand-bound (holo) and ligand-unbound (apo) structures. These 403 pairs were used as part of the original ENCoM parameterization.

Of the original 403 PSCDB apo-holo pairs exhibiting significant conformational change, we keep the 37 pairs that are single-chain, complete from the crystallographers' submission (no missing residues) and exactly the same length. We use these stringent filters to minimize potential artifacts from missing residues or changes in the geometry of the complex in the case of multi-chain structures. Moreover, since

many of these proteins are quite large, we cannot afford the computational cost of running the full 403 pairs with the 221 255 parameter combinations. The 37 kept pairs (74 structures total) are listed in Table A.3. We computed the cumulative overlap at 5% normal modes both ways for every pair and report the mean cumulative overlap for every parameter set.

4.6.3 *Protein ΔΔG of folding*

In the original ENCoM article, the effect of mutations was predicted for 303 mutations selected from the ProTherm database [116] as part of the validation for the PoPMuSiC-2.0 software [166], a machine-learning based tool for predicting protein stability changes upon mutation to which ENCoM was compared [27]. This set of mutations contains 45 stabilizing mutations, 84 neutral mutations and 174 destabilizing mutations. The mutations come from 66 distinct wild-type proteins. We use this exact set to measure the performance of the different parameter sets in the prediction of folding ΔΔG, with the exception of four cases: first, the single mutation reported on a structure with PDB code reported as 1AON in the supplementary data from the ENCoM article. PDB code 1AON is a complex with over 8000 residues, so this must be an error. The other three cases removed are the mutations on yeast phosphoglycerate kinase (PDB code 3PGK). We modeled all mutations with MODELLER [150] and this starting structure generated an error. While investigating this, we noticed that the PDB quality metrics for this structure are very poor, with more than 33% sidechain outliers and over 20% Ramachandran outliers. This led us to simply exlude this case. The list of included mutations is given in Table A.4.

As mentioned, the remaining 299 mutations were performed *in silico* with the MODELLER software. We use the same protocol throughout the present thesis, which is outlined in Section 4.2.2. For each of the mutants, we computed ΔS_{vib} as follows:

$$\Delta S_{\text{vib}} = S_{\text{vib}_{\text{WT}}} - S_{\text{vib}_{\text{mutant}}} \tag{4.12}$$

with the calculation for S_{vib} given in Equation 4.1. According to this definition, a positive ΔS_{vib} means the mutation is rigidifying (loss of vibrational entropy), while a negative ΔS_{vib} corresponds to softening mutations (gain of vibrational entropy). Since the metric used is RMSE after a linear fit without intercept (see Section 4.5.4), the sign of the prediction does not matter for the measure of performance.

4.6.4 *Dataset of RNA structures*

For the three RNA benchmarks, we start from the same sets of structures that we used to benchmark our adaptation of ENCoM to RNA [157], which were constructed as follows. Note that sections of the text for the description of these three benchmarks come from that work. In the PDB as of 2022-01-08, there were 982 structures solved by X-ray crystallography and 417 NMR ensembles when restricting the search to entries containing only RNA. We used these data in the RNA-only benchmarks after applying appropriate filters, described at the end of the current section. We kept all chains present in the biological assembly for X-ray structures and all chains in the structural ensemble for NMR experiments.

ENCoM requires the manual assignment of atom types for every residue we want to consider so that the surface in contact term can be adequately computed. It is thus possible in theory to include every modified residue. In order to simplify the conducted benchmarks, we chose to restrict the model to the four standard ribonucleotides, replacing all modified ribonucleotides with their unmodified analog using the ModeRNA software [149], which contains information about the 170 modifications present in the MODOMICS database [167]. The atom type assignations for these four standard ribonucleotides are given in Table A.1.

We also used ModeRNA to add missing atoms where it was the case, for example terminal phosphate groups, so that each residue was complete. This addition ensured that the assignment of the beads to the residues was standard for all residues. The structures for which ModeRNA produced an error or which produced an error for any of the models tested subsequently were removed from the analysis. Moreover, the size of the X-ray resolved RNA structures was restricted to below 300 nucleotides (900 beads) for the B-factors benchmark and below 100 nucleotides (300 beads) for the conformational change benchmark, in order to minimize computational cost. In the case of the solution NMR resolved RNAs, no size threshold was applied as no molecules bigger than 155 residues have been elucidated this way. Since we were interested in conformational ensembles solved by NMR, we restricted our analysis to submissions containing at least two models. The lists of PDB codes for the structures kept for the three benchmarks (38 structures for the B-factors benchmark, 116 structures for the conformational change benchmark, 313 NMR ensembles for the structural variance benchmark) are given in Table A.5, Table A.6 and Table A.7.

4.6.4.1 *Sequence clustering*

Since we assembled the RNA structures automatically from the PDB, some sequences are sampled more than once, or have very close homologs sampled. In fact, we rely on this property for the RNA conformational change benchmark. However, we need to prevent RNA families which have more members present in our

database from driving the performance metrics. We first compute a Needleman-Wunsch global alignment [168] between all pairs of two sequences, with the following scoring scheme: gap penalty: -1, mismatch penalty: -1, match score: 1. We compute the alignment distance as follows:

$$\text{distance}(s,t) = \frac{\text{score}(s,t)}{\min\left[\text{length}(s), \text{length}(t)\right]} \tag{4.13}$$

We then perform complete linkage clustering [169] on the sequences from our database of structures with a distance threshold of 0.1, ensuring that all pairs of sequences within a given cluster are at least 90% similar in terms of their Needleman-Wunsch global alignment. Because of the scoring scheme used, this is equivalent to 90% sequence identity, with differences in length counting as mismatches.

For all reported RNA-only benchmark metrics, we first normalize by sequence cluster before reporting the mean metric. This ensures that molecules which have a lot of conformations sampled in the PDB do not drive the score and instead each RNA or family of RNAs has an equal weight.

4.6.5 *RNA experimental B-factors*

In addition to a resolution filter of 2.5 Å or better, we used only the structures for which no modified nucleobases were initially present and in which all residues were complete as the set of X-ray crystallography RNA structures from which to predict experimental B-factors. These restrictions were applied to ensure that every atom used in the prediction had a corresponding experimental B-factor, and not an extrapolated value from the rebuilding of the modified nucleobases and missing atoms by ModeRNA. The 38 remaining structures were clustered according to their sequence as described, giving us 34 clusters. We computed the mean Pearson correlation normalized by cluster.

4.6.6 *RNA conformational change*

We started by listing all pairs of X-ray structures from the same sequence cluster as potential conformational changes. We then rejected pairs of conformations for which the root mean squared deviation (RMSD) of the center atoms of each bead is lower than 2 Å. Clustering was performed as described, leaving us with 116 unique structures forming 227 pairs of conformations with RMSD greater than or equal to 2 Å, divided in 25 sequence clusters. We calculated the cumulative overlap for 5% of normal modes both ways for every pair and report the mean cumulative overlap normalized by cluster.

4.6.7 *RNA NMR ensemble variance*

For the 313 RNA-only NMR ensembles, we compute the normalized cumulative overlap (NCO) between the principal components (PCs) representing 99% of the structural variance in the ensemble and the 5% slowest normal modes computed from the minimum energy conformation. As already mentioned, we use the nrt-PCA correction introduced in Section 4.5.3.1 to compute the PCs. We report the mean NCO for every parameter set.

4.6.8 *RNA-protein NMR ensemble variance*

We assembled a set of solution NMR ensembles of RNA-protein complexes by using the advanced search feature of the PDB (as of 2022-05-31), limiting results to entries containing both protein and nucleic acids, rejecting entries with more than 70% sequence identity. We then sorted the entries by date, with the newest entries first, and manually selected the first 20 entries that contained only RNA and protein residues (no DNA). The 20 PDB codes selected are listed in Table A.8.

4.6.9 *Parameter search*

As outlined in Section 4.1, ENCoM has been adapted to work on RNA molecules as part of the present thesis and its performance has been extensively compared to other coarse-grained elastic network models (ENMs) as part of our work currently submitted for publication [157]. However, the ENCoM parameters were not explored as part of that work, as we found that the parameters already optimized for proteins [27] were leading to good performance. The ENCoM potential, presented in more details in Section 3.1, is the following:

$$
\begin{aligned}
V_{\mathrm{ENCoM}}(\vec{R}, \vec{R}_0) &= \sum_{\text{bonds}} V_1(r, r_0) + \sum_{\text{angles}} V_2(\theta, \theta_0) \\
&+ \sum_{\text{dihedrals}} V_3(\phi, \phi_0) + \sum_{i<j-3} V_4(r_{ij}, r_{ij_0}) \\
&= \sum_{\text{bonds}} \alpha_1 (r - r_0)^2 + \sum_{\text{angles}} \alpha_2 (\theta - \theta_0)^2 \\
&+ \sum_{\text{dihedrals}} \left[\alpha_3 (1 - \cos(\phi - \phi_0)) + \frac{\alpha_3}{2} (1 - \cos 3(\phi - \phi_0)) \right] \\
&+ \sum_{i<j-3} (\beta_{ij} + \alpha_4) \left[5 \left(\frac{r_{ij_0}}{r_{ij}} \right)^{12} - 6 \left(\frac{r_{ij_0}}{r_{ij}} \right)^{10} \right]
\end{aligned}
\tag{4.14}
$$

The four apparent parameters in the potential are the scaling weights for each term, α_1 to α_4. However, the β_{ij} term is the surface area in contact between atoms of

different residues, multiplied by their complementarity coefficient in the interaction matrix (Section 3.1). The simplified atom typing system from Sobolev *et al.* [138] has two interaction types: favorable and unfavorable. These constitute an additional two parameters, which we call σ_+ and σ_- for favorable and unfavorable interactions respectively. Surprisingly, these additional parameters were not explored as part of the initial ENCoM parameter grid search, instead being set to 3 and 1 for favorable and unfavorable interactions respectively. Another version of ENCoM, ENCoM_{ns} (for ENCoM non-specific), was tested as part of the original benchmarking of ENCoM, which had both σ parameters set to 1. It seems that the authors were under the impression that the varying of α_4 was enough to cover the different possibilities. This is not the case however, as the long-range term in the potential is modulated by the sum of α_4 and β_{ij}. α_4 is responsible for background, non-specific long-range interactions which only depend on the xyz coordinates of the interacting residues. β_{ij} is the term responsible for ENCoM's sequence sensitivity (detailed in Equation 3.1), and thus the σ parameters need to be optimized as well.

The introduction of these two additional parameters greatly expands the size of the search space. For instance, the $\log_{10}$ parameter grid search performed by Frappier and Najmanovich spanned 13 orders of magnitude per parameter and thus led to 28 561 distinct parameter sets. Performing such a wide search on 6 parameters would lead to over 4 million combinations, which is beyond the computational resources we have access to. Nonetheless, the original parameter search led to the following set of parameters:

$$\alpha_1 = 10^3, \quad \alpha_2 = 10^4, \quad \alpha_3 = 10^4, \quad \alpha_4 = 10^{-2} \tag{4.15}$$

As mentioned, the σ parameters were set to:

$$\sigma_+ = 3, \quad \sigma_- = 1 \tag{4.16}$$

4.6.9.1 *First round of parameter search*

Thus, all original published parameters are within 7 orders of $\log_{10}$ magnitude despite a search space covering 13 orders of magnitude, suggesting that combinations with differences of more than 7 orders of magnitude between pairs of parameters led to suboptimal performance. Furthermore, since the normal modes are obtained as the eigenvectors of the Hessian matrix (Section 3.1.1), two parameter sets which have exactly the same ratios between them will lead to exactly the same set of normal modes, and the eigenvalues will simply be rescaled linearly. Since this rescaling does not affect any of the performance metrics investigated, it is possible to speed up the computation of the benchmarks by considering only one parameter set per given ratio, instead of performing a truly exhaustive search.

We thus decided to perform the parameter search in successive rounds: in the first round, all combinations of 7 orders of $\log_{10}$ magnitude were tested, with the added possibility of a null value for α_4 and σ_-. In the case of α_4, it was to test whether the non-specific long range interaction is necessary for performance, since the surface area in contact term β_{ij} might be sufficient on its own. σ_- represents interactions which are unfavorable and even repulsive in some cases. However, repulsive terms cannot be part of a normal mode analysis potential since the input conformation is assumed to represent the minimum energy of the system. The closest one can get to repulsive behaviour is thus to assign a zero value to the unfavorable interaction term, hence why we tested this possibility.

The main reason for the parameter search across seven diverse benchmarks is to ask whether a single set of parameters can lead to good performance across many scenarios. In order to answer this question, we standardized the performance of every benchmark, taking the Z-score of the negative RMSE (NRMSE) in the case of predicted $\Delta\Delta G$ of folding since a lower score is better on that metric. For all other metrics, a higher score represents better performance. We consider the seven benchmarks to be balanced between protein and RNA molecules: three for RNA-only structures, three for protein-only structures, and one for RNA-protein complexes. Since the performance metrics across parameter sets are not guaranteed to be normally distributed, we rescaled the Z-scores linearly, dividing by the maximal Z-score for each benchmark. This rescaling is necessary to ensure an equal contribution of every benchmark to the Z-score sum across the seven benchmarks, which is the combined performance metric we use in order to select the parameter set with the best overall performance:

$$P_i = \sum_{\text{Benchmarks}} \frac{Z_i}{\max(Z)} \tag{4.17}$$

where P_i denotes the global performance of the i^{th} parameter set.

4.6.9.2 *Second and last rounds of parameter search*

The best parameter set from the first round according to the P metric defined in Equation 4.17 was used as the seed for the second round, in which multipliers spanning 7 $\log_2$ orders of magnitude were applied combinatorially to each parameter from the seed set. The idea behind the reduction in logarithmic base from 10 to 2 was to search the vicinity of the best parameter set from round 1 for potential improvements. The null values tested for the α_4 and σ_- parameters in round 1 did not lead to good performance, hence we did not include these further.

Both the largest and smallest multipliers were applied to the seed parameter set when examining the best parameter set from round 2, prompting us to perform a last round of parameter search, again applying multipliers spanning 7 $\log_2$

orders of magnitude. As outlined in Section 5.3, this last round did not lead to improved performance and the same set of parameters as in round 2 is found as the optimal set. The absence of further performance gains is the reason for stopping the parameter search after this final round.

4.6.10 *Combined performance across benchmarks*

The seven performance metrics evaluated are: protein B-factors prediction, protein conformational change prediction, protein mutations $\Delta\Delta G$ prediction, RNA B-factors prediction, RNA conformational change prediction, RNA ensemble variance prediction, RNA-protein complexes ensemble variance prediction. Each performance metric is standardized across the whole range of parameter sets and then rescaled so that the maximum Z-score is always 1, in order to give equal weight to every category. The average rescaled Z-score is then used to rank the parameter sets. To investigate covariance between the performance across pairs of benchmarks, the covariance matrix of the Z-score before rescaling is reported. This ensures a possible range of values from -1 to 1 for covariance.

4.7 DEVELOPED TOOLS AS PYTHON PACKAGES

Importantly, all parts of the EMCoM-DynaSig-ML pipeline are distributed as open-source, extensively documented Python packages that were produced as part of the present thesis. This availability and ease of use will help lower the barriers to the usage of the pipeline and to its integration as part of consensus variant effect predictors.

4.7.1 *The NRGTEN Python package*

The Najmanovich Research Group Toolkit for Elastic Networks (NRGTEN) contains the implementation of ENCoM and was published recently [31]. Since it is extensively documented online, we refer the interested reader to the online documentation: nrgten.readthedocs.io. Figure 4.4 shows the NRGTEN documentation page for the computation of changes in ΔS_{vib} upon mutation.

4.7.2 *The DynaSig-ML Python package*

The DynaSig-ML Python package implements the necessary tools for the streamlined execution of the ENCoM-DynaSig-ML pipeline, building upon the NRGTEN package for the execution of ENCoM. It is currently submitted for publication and is available as a preprint [32]. Online documentation is available at dynasigml.readthedocs.io, along with a tutorial allowing the replication of our results on VIM-2 lactamase catalytic efficiency presented in Chapter 8, albeit with the

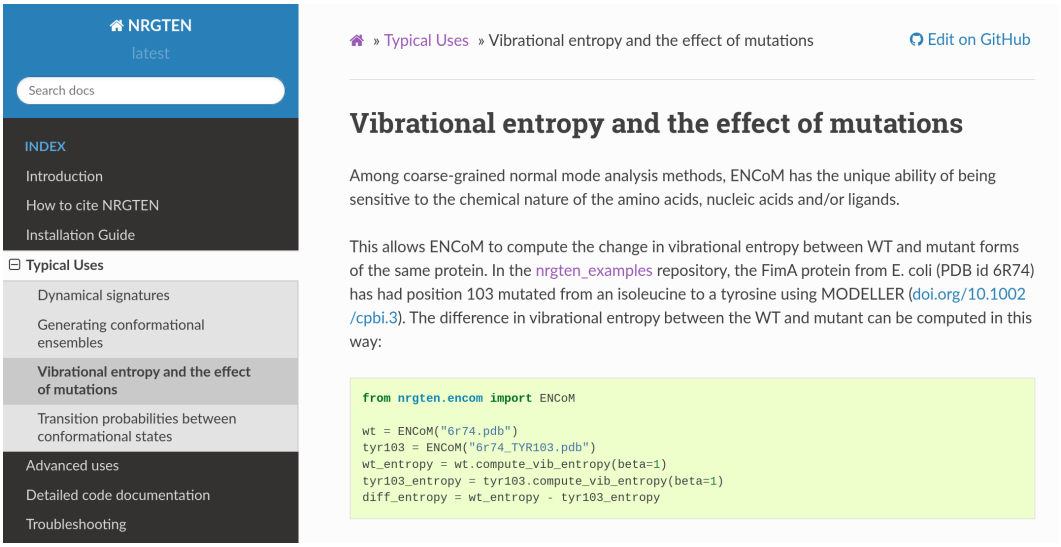


Figure 4.4: **Web page from the NRGTEN online documentation.**

original ENCoM parameters, coarser search of thermodynamic scaling factors and simplified statistical analyses.

Part II

RESULTS AND DISCUSSION

5

PARAMETER SEARCH: DIVERSE BENCHMARKS

The ENCoM-DynaSig-ML pipeline is the present thesis' central piece. ENCoM is used as the pipeline's first step to generate Dynamical Signatures, so its performance should be maximized in order to maximize the pipeline's performance. Moreover, ENCoM was adapted to work on RNA molecules and the case studies presented in Chapter 6, Chapter 7 and Chapter 8 respectively study RNA molecules, protein-small molecule complexes and an enzyme (protein). The original ENCoM parameters were not changed as part of our study assessing ENCoM's performance on RNA, as we wished to maintain optimal performance on protein as part of a single parameter set and the performance on RNA from the original parameters already outperformed two common ENMs [157]. However, the question remained of whether a single set of parameters can attain high performance on both RNA and protein molecules. In order to answer this question, the current chapter presents the results of a wide parameter search across seven diverse benchmarks in which RNA and protein molecules are equally represented.

The parameter search was performed in three distinct rounds, as outlined in Section 4.6.9. Table 5.1 lists the performances of the best parameter sets found in each round across the seven benchmarks, in addition to the performances of the original parameter set. The best parameter set for each round was defined as the one leading to the highest rescaled Z-score sum across the seven benchmarks. Table 5.2 lists the individual rescaled Z-scores per benchmark, for the three rounds of parameter search. Table 5.3 lists the different optimal parameter sets found after each searching round. The same parameter set is listed for the combination of rounds 2 and 3 because it was the optimal parameter set for both rounds. As the next sections will outline, this optimal parameter set from rounds 2 and 3 was chosen as the new set of ENCoM parameters.

5.1 ROUND 1 PARAMETER SEARCH

The set of values used for the first round parameter sweep was the following: [0.01, 0.1, 1, 10, 100, 1000, 10 000]. As mentioned in Section 4.6.9, α_4 and σ_- were also allowed to take null values. With the inclusion of these zero values, there are 153 664 exhaustively enumerated parameter sets for the first round of parameter

Table 5.1: **Results by round for the parameter search.** For each benchmark, the best performance attained is highlighted in bold.

Metric	Original	Round 1	Round 2	Round 3
Protein B-factors	0.572	0.583	**0.585**	**0.585**
Protein overlaps	0.791	0.798	**0.805**	**0.805**
Protein ΔΔG RMSE	1.549	1.542	**1.539**	**1.539**
RNA-protein NCO	0.752	0.759	**0.762**	**0.762**
RNA B-factors	0.446	**0.458**	0.456	0.455
RNA overlaps	0.690	**0.718**	0.715	0.715
RNA NCO	0.774	0.775	**0.779**	**0.779**

Table 5.2: **Rescaled Z-scores by round for the parameter search.** The rescaled Z-scores (as defined in Equation 4.17) are given for the best parameter set found in each round of parameter search. The best parameter set is defined as the one leading to the highest rescaled Z-score sum across the seven benchmarks.

Metric	Round 1	Round 2	Round 3
Protein B-factors	0.833	0.777	0.772
Protein overlaps	0.975	0.843	0.772
Protein ΔΔG RMSE	0.928	0.646	0.703
RNA-protein NCO	0.973	0.704	0.685
RNA B-factors	0.323	0.311	0.289
RNA overlaps	0.905	0.242	0.377
RNA NCO	0.947	0.745	0.711
Rescaled Z-score sum	5.884	4.268	4.310

search. After removing sets with duplicate parameter ratios, we are left with 96 244 parameter sets to test. Running the seven benchmarks for one of these parameter sets takes 67 minutes using one CPU from a 32-core AMD Rome 7532 @ 2.40 GHz, which represents a total cost of 107 472 core-hours, or 12.3 core-years for the complete first round.

The parameters leading to the highest rescaled Z-score sum (RZSS) across all seven benchmarks are given in Table 5.3, with an RZSS of 5.9. Figure 5.1 shows the performances attained for the individual benchmarks as histograms, with the performances of the optimal parameter set from round 1 indicated by dashed lines and the performance of the original ENCoM parameters by a full line.

Table 5.3: **Parameter sets by round of parameter search.** In addition to the best parameter set found after each round of parameter search, the original ENCoM parameter set is given. The optimal parameter set for round 2 is identical to the optimal parameter set for round 3. This set represents the new ENCoM parameters and is shown in bold.

Parameter	Original	Round 1	Round 2	Round 3
α_1	1000	1000	**4000**	**4000**
α_2	10 000	10 000	**80 000**	**80 000**
α_3	10 000	10 000	**20 000**	**20 000**
α_4	0.01	0.01	**0.08**	**0.08**
σ_-	1	0.1	**0.0125**	**0.0125**
σ_+	3	10	**80**	**80**

Interestingly, the parameter set selected exhibits high performance across all benchmarks, with the exception of the B-factors prediction benchmarks where it performs well but other combinations perform better. This is in accordance with the tradeoff already observed a part of ENCoM's original parameterization between B-factors prediction performance on the one hand and overlaps and effects of mutations on the other hand [27].

Strikingly, the α_1-α_4 parameters are the same as in the original ENCoM parameter set, while σ_+ and σ_- respectively get higher and lower values. This is in accordance with the hypothesis that ENCoM's performance edge over other models comes in part from the surface complementarity term. Indeed, we expected the exploration of the σ parameters to yield such a pattern of higher weight for favorable interactions and lower weight for unfavorable interactions.

However, this set of parameters, as the original set, spans 7 orders of $\log_{10}$ magnitude, which is the exact breadth of the parameters explored (with the exception of the null values for α_4 and σ_-). This points to a better combination potentially existing outside of these bounds, and is the reason for performing another round of optimization.

Figure 5.2 shows the covariance of the Z-scores between all possible combinations of two benchmarks. A positive covariance means that performance on these two benchmarks tends to go in the same direction across parameter sets, whereas negative covariance means that performance from one benchmark is inversely correlated to performance from the other benchmark across the whole set of tested parameters. However, such negative covariance does not necessarily imply that a tradeoff between the two benchmarks has to be made, as outlier parameter sets could give rise to good performance for both benchmarks even if this is not the general trend.

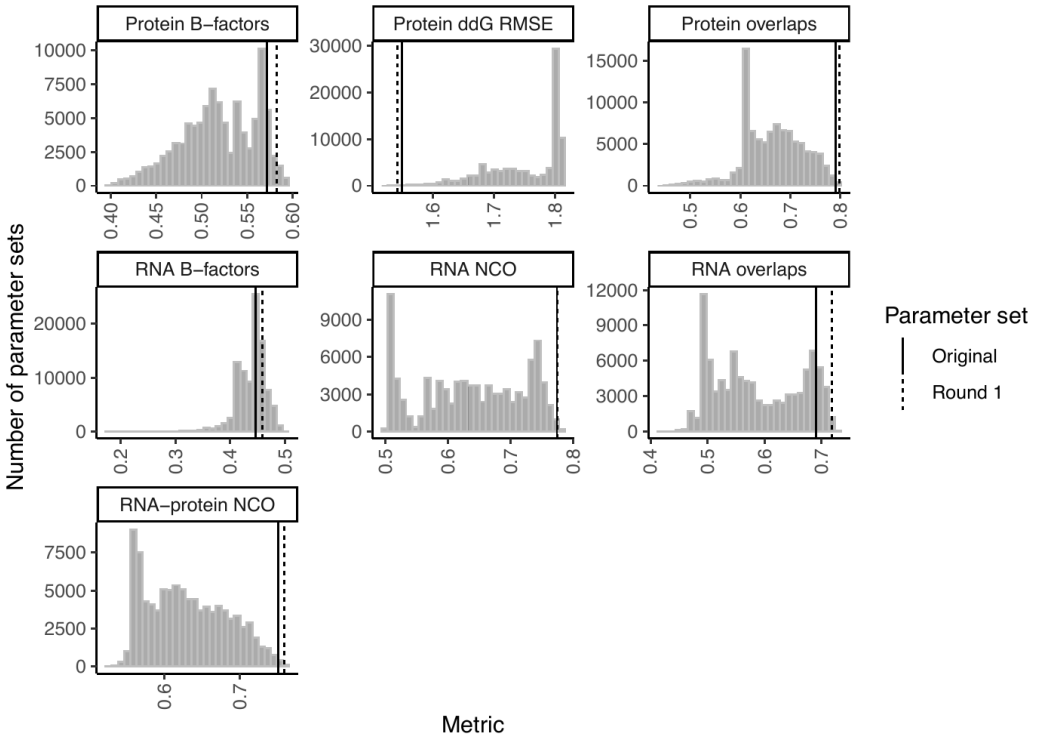


Figure 5.1: **Round 1 parameter sets performance on the diverse benchmark.** The performances on each individual benchmark are shown as histograms. In order from left to right and top to bottom, the x axis represents the Pearson correlation between protein experimental B-factors and ENCoM Entropic Signature, the root-mean-square error (RMSE) on predicted $\Delta\Delta G$ of folding upon mutation, cumulative overlap at 5% non-trivial normal modes for protein conformational change prediction, Pearson correlation between RNA experimental B-factors and ENCoM Entropic Signature, normalized cumulative overlap (NCO) at 5% non-trivial normal modes for RNA NMR ensemble structural variance prediction, cumulative overlap at 5% non-trivial normal modes for RNA conformational change, and NCO at 5% non-trivial normal modes for RNA-protein complexes NMR ensemble structural variance prediction.

Figure 5.3 gives a more detailed view of how the performance varies across each pair of benchmarks. This detailed view reveals that despite negative covariance between some pairs of benchmarks, such as the protein B-factors and the three RNA-only benchmarks, a combination of parameters leading to high performance across both benchmarks can still be found in all cases. The performance of the optimal parameter set is indicated by red dots, highlighting the fact that a rescaled Z-score close to 1 is attained for all benchmarks with the exception of the RNA B-factors prediction benchmark. For example, despite a general inverse correlation between performance across the RNA-only and protein B-factors benchmarks, parameter sets exist that give rise to rescaled Z-scores close to 1 for both metrics.

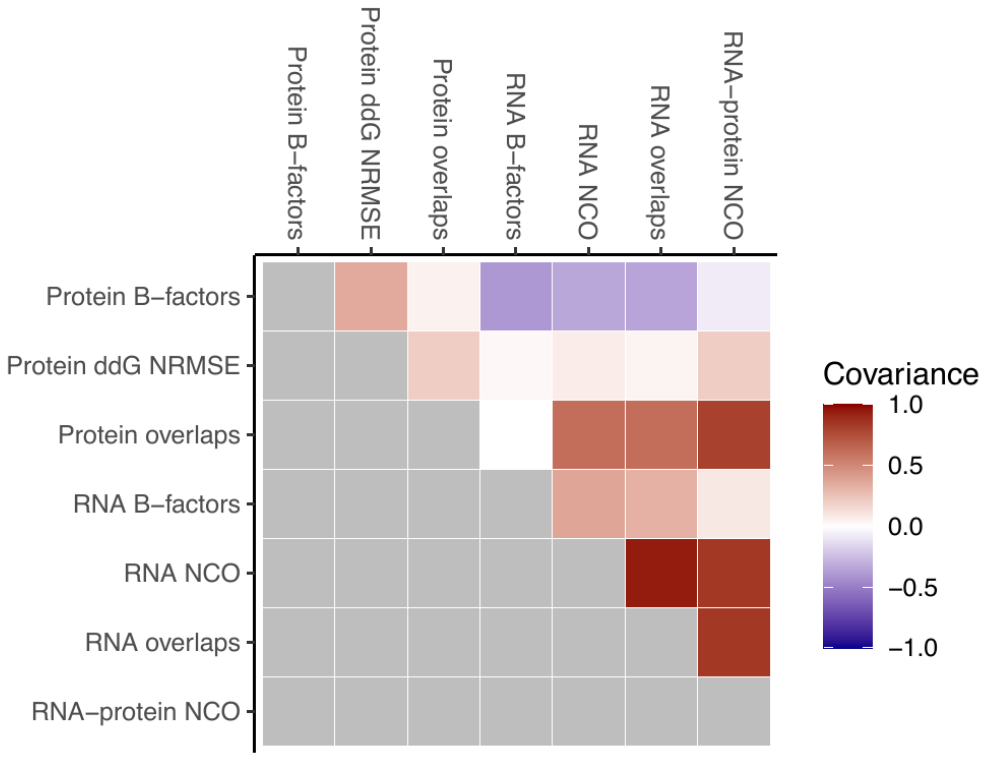


Figure 5.2: **Round 1 covariance between benchmarks.** The Z-score covariance between each pair of benchmarks is shown as a half-matrix. For the protein ΔΔG of folding benchmark, the Z-score of the negative RMSE (NRMSE) is used in order for positive Z-score to represent better performance. To avoid redundancy, the bottom half and the diagonal of the matrix are colored in gray.

Again, the exception is the RNA B-factors benchmark which have a rescaled Z-score of 0.32 using the best parameter set.

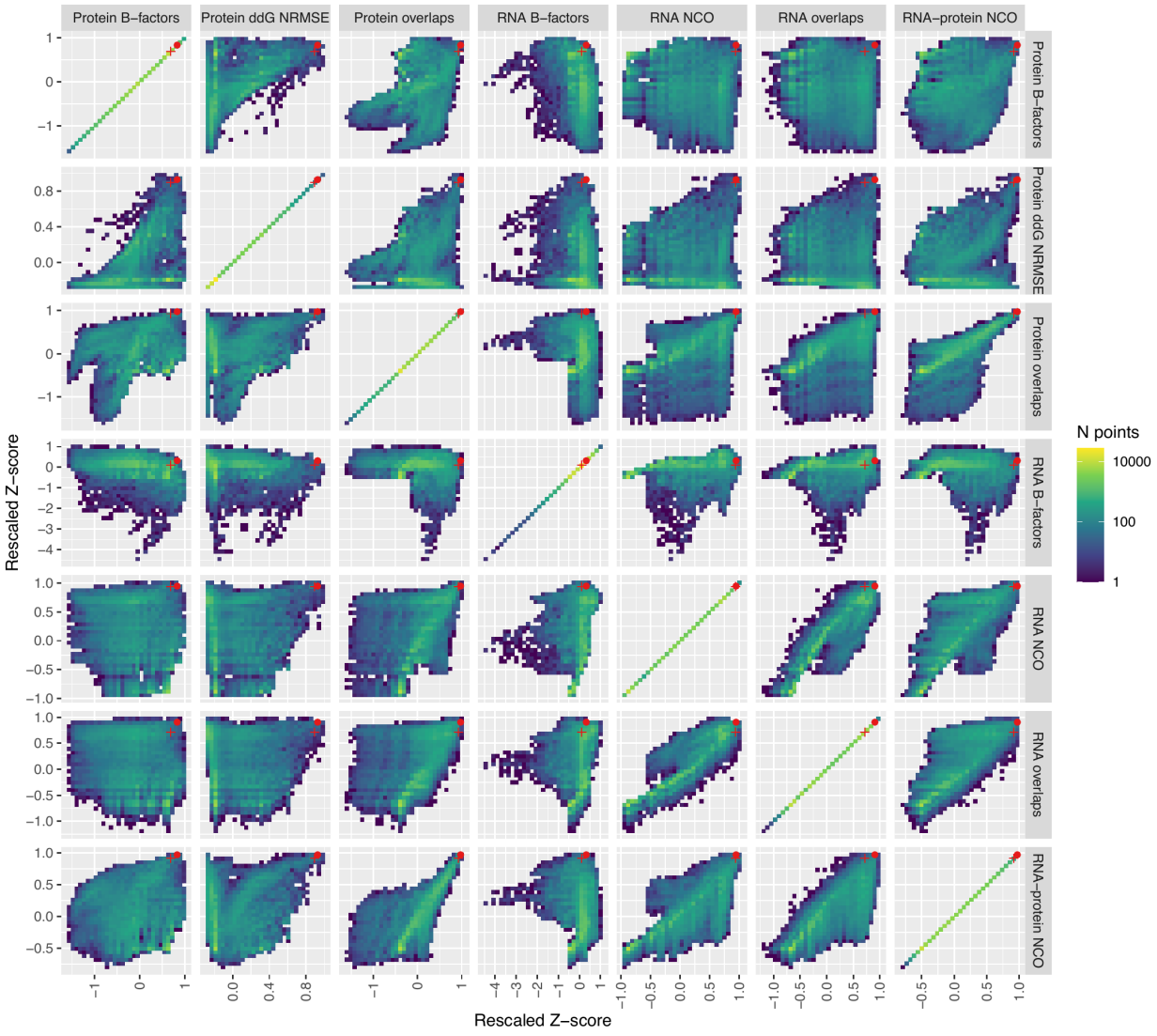


Figure 5.3: **Round 1 detailed performance across benchmark pairs.** For each pair of benchmarks, the rescaled Z-scores across all parameter sets are plotted against each other. In the case of the protein ΔΔG prediction benchmark, the rescaled Z-score of the negative RMSE (NRMSE) is plotted, as a lower RMSE value represents higher performance. The performance of the optimal parameter set found is denoted with red dots and the performance of the original ENCoM parameter set is denoted with red crosses.

5.2 ROUND 2 PARAMETER SEARCH

For the second round of parameter search, we started with the optimal set of parameters from round 1, which is shown in Table 5.3. In order to both perform a finer search in the vicinity of these parameters and allow for the potential

widening of the gap between the smallest and largest values, the second ensemble of parameter sets was generated by multiplying each parameter from the round 1 set by the following powers of 2: [0.125, 0.25, 0.5, 1, 2, 4, 8]. Since there are 7 combinations for every parameter, the exhaustive enumeration of the parameter sets gives 117 649 sets. After removal of duplicate ratios, we are left with 70 992 parameter sets to test, for a total computational cost of 79 274 core-hours, or 9.1 core-years.

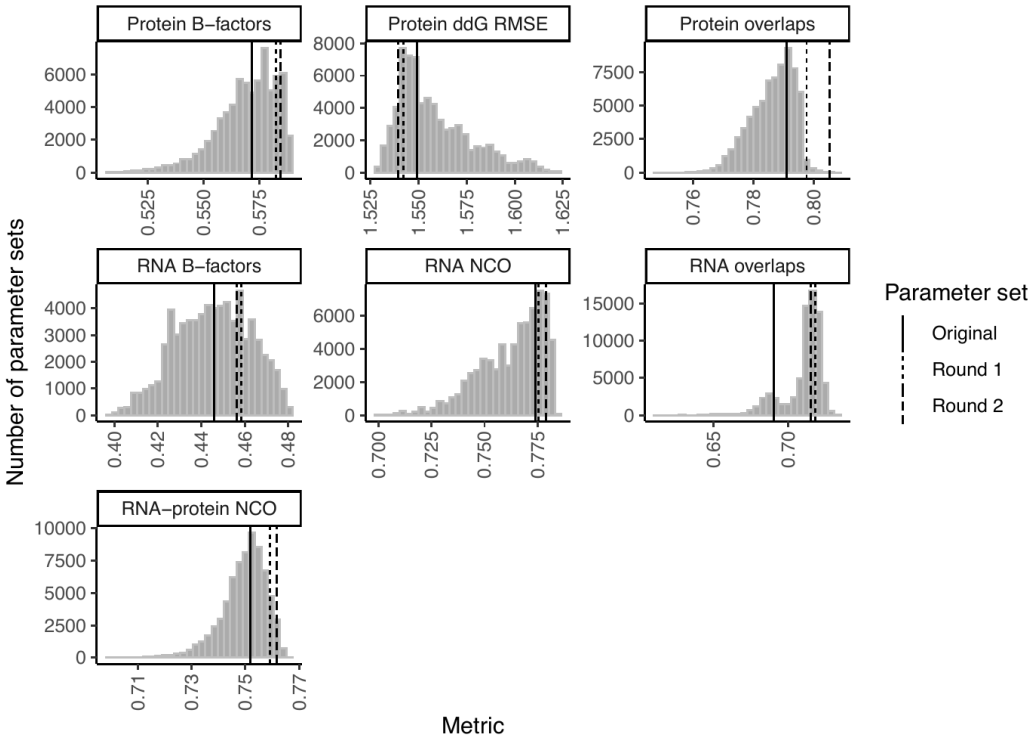


Figure 5.4: **Round 2 parameter sets performance on the diverse benchmark.** The performances on each individual benchmark for the second round of parameter search are shown as histograms, as in Figure 5.1 for the first round.

The optimal parameter set for round 2 is given in Table 5.3. Both the maximum and minimum multipliers are part of the transformation from round 1 optimal parameters to round 2 optimal parameters. Indeed, α_2 went from 10 000 to 80 000 while σ_- went from 0.1 to 0.0125. However, the smallest parameter from round 1 (α_4) was multiplied by 8, meaning that the gap between smallest and biggest parameter value was not maximally widened. These results prompted us to run a last round of parameter search, again with multipliers spanning 7 $\log_2$ orders of magnitude. We hypothesized that the parameter values should change less as part of this last round and that the gap should not widen significantly between minimal and maximal parameter.

Figure 5.4 illustrates the performance across the seven benchmarks for all round 2 parameter sets, with lines indicating the performance from the original parameters, the optimal round 1 parameters and the optimal round 2 parameters.

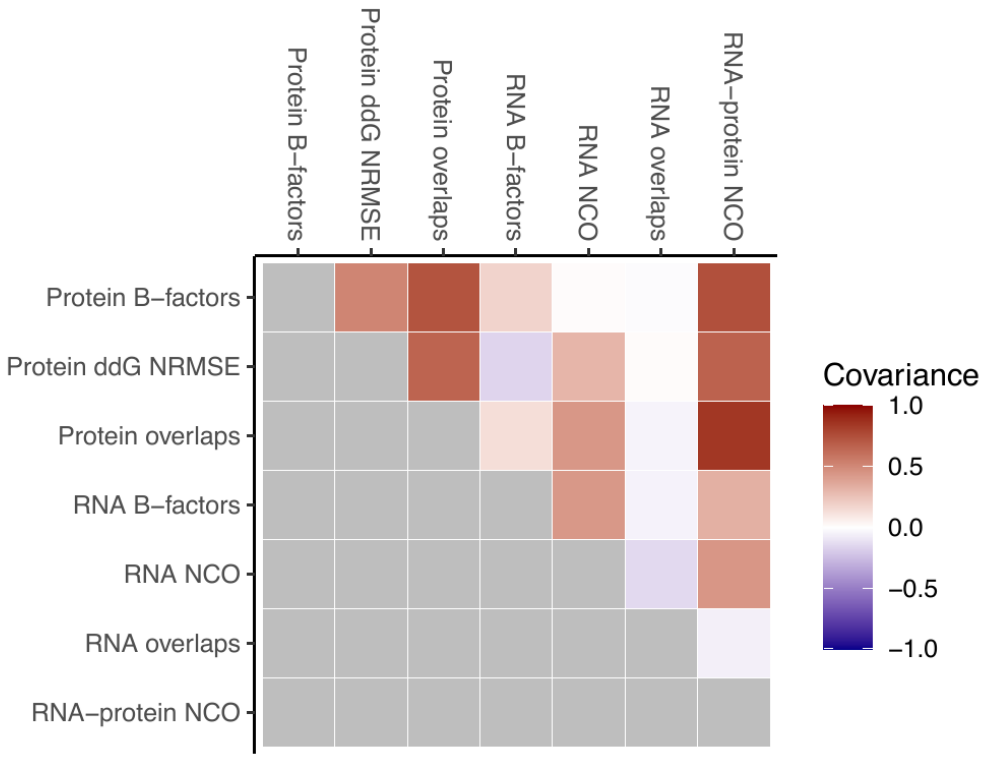


Figure 5.5: **Round 2 covariance between benchmarks.** The Z-score covariance between each pair of benchmarks for the second round of parameter search is shown as a half-matrix, as in Figure 5.2 for the first round.

Interestingly, σ_+ and σ_- got respectively multiplied by 8 and 0.125 compared to the starting parameter set, which corresponds to the largest and smallest multipliers tested. This reinforces the notion that the capturing of localized, all-atom interactions between complementary atom types is part of what constitutes ENCoM's performance edge.

Figure 5.5 illustrates the covariance matrix for the seven benchmarks across the round 2 parameter sets. A striking feature from this matrix is the disappearance of strong negative covariance between the protein B-factors prediction benchmark and the three RNA-only benchmarks. This might be explained by the overall higher performance of the parameter sets tested in round 2, as they are generated across a much smaller range of values around the optimal set from round 1.

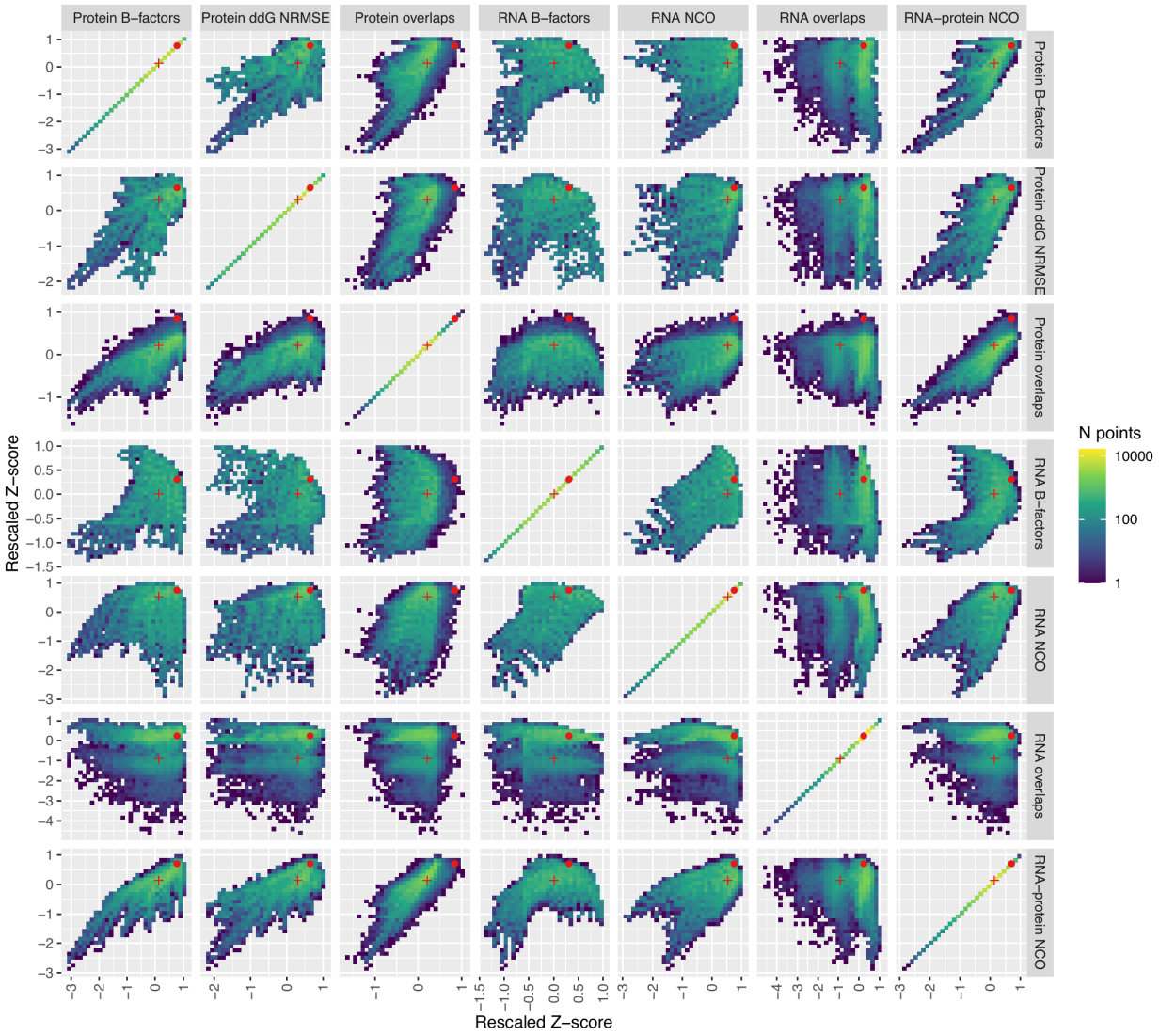


Figure 5.6: **Round 2 detailed performance across benchmark pairs.** As in Figure 5.3, the rescaled Z-scores across all parameter sets are plotted against each other, the performance of the optimal parameter set found is denoted with red dots and the performance of the original ENCoM parameters with red crosses.

The detailed performance across benchmark pairs for round 2 is outlined in Figure 5.6. As in round 1, the RNA B-factors benchmark suffers a big gap between the maximal performance and the performance of the optimal parameter set, with a rescaled Z-score of 0.311. In addition, the RNA overlaps benchmark has the lowest rescaled Z-score at 0.242. However, this is probably due to the fact that almost all parameter sets tested in round 2 have excellent performance on this benchmark, with mean cumulative overlap of 0.71 at 5% nontrivial normal modes. For reference, the cumulative overlap is a value between 0 and 1 which describes how well a

set of normal modes can capture a conformational change (see Section 4.5 for more details). A value of 0.71 means that the ensemble of normal modes can deform the starting conformation towards the target conformation and reduce the RMSD between the two by 71%. Almost all parameter sets tested in round 2 have a cumulative overlap above 0.65 for the RNA overlaps benchmark, as is apparent in Figure 5.4. Thus, the rescaled Z-score for this benchmark is to be taken with a grain of salt, as the variation in performance is very slight. Moreover, the optimal parameter set still performs better than the average set tested, despite a tiny loss in performance compared to the optimal round 1 set (0.715 vs 0.718 for round 1).

5.3 ROUND 3 PARAMETER SEARCH

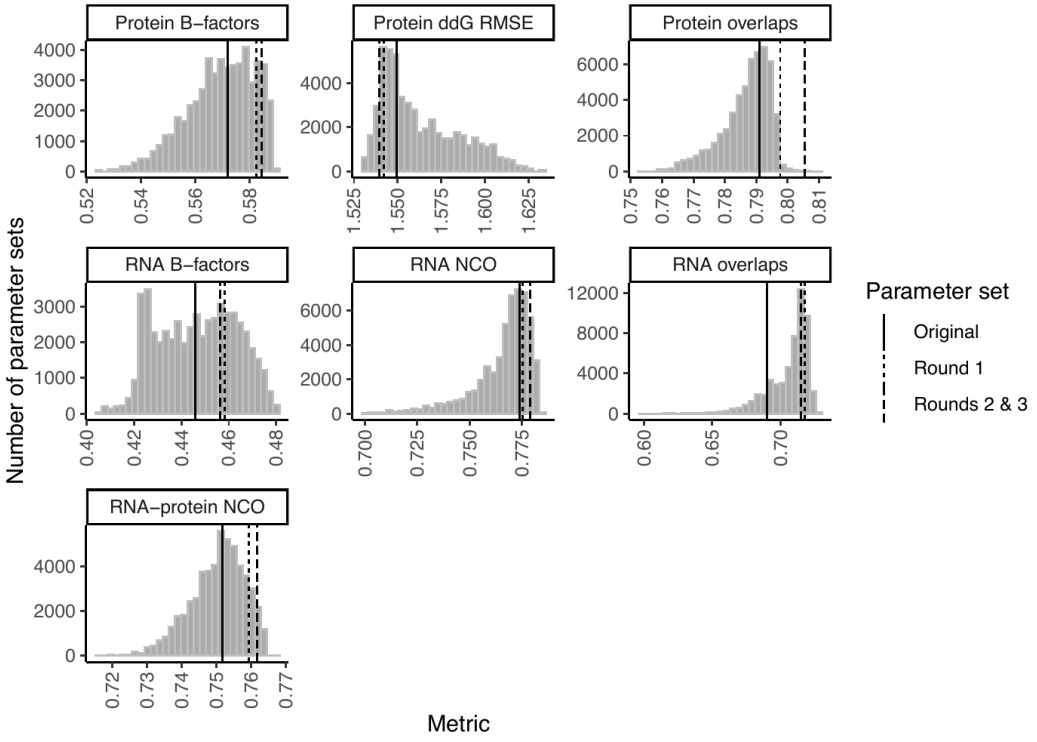


Figure 5.7: **Round 3 parameter sets performance on the diverse benchmark.** The performances on each individual benchmark for the third round of parameter search are shown as histograms, as in Figure 5.1 for the first round.

The third and final round of parameter search was performed in the same fashion as the second round, starting from the optimal round 2 parameters and applying the following multipliers: [0.125, 0.25, 0.5, 1, 2, 4, 8]. As for round 2, the exhaustive enumeration gives 117 649 parameter sets. However, since both round 2 and round 3 parameter sets were generated with multipliers that are powers of 2, there are more redundant parameter ratios in the round 3 ensemble of parameter sets because some were already tested in round 2. After removal of all duplicate ratios, we are

left with 54 019 parameter sets to test, for a total computational cost of 60 321 core-hours (6.9 core-years).

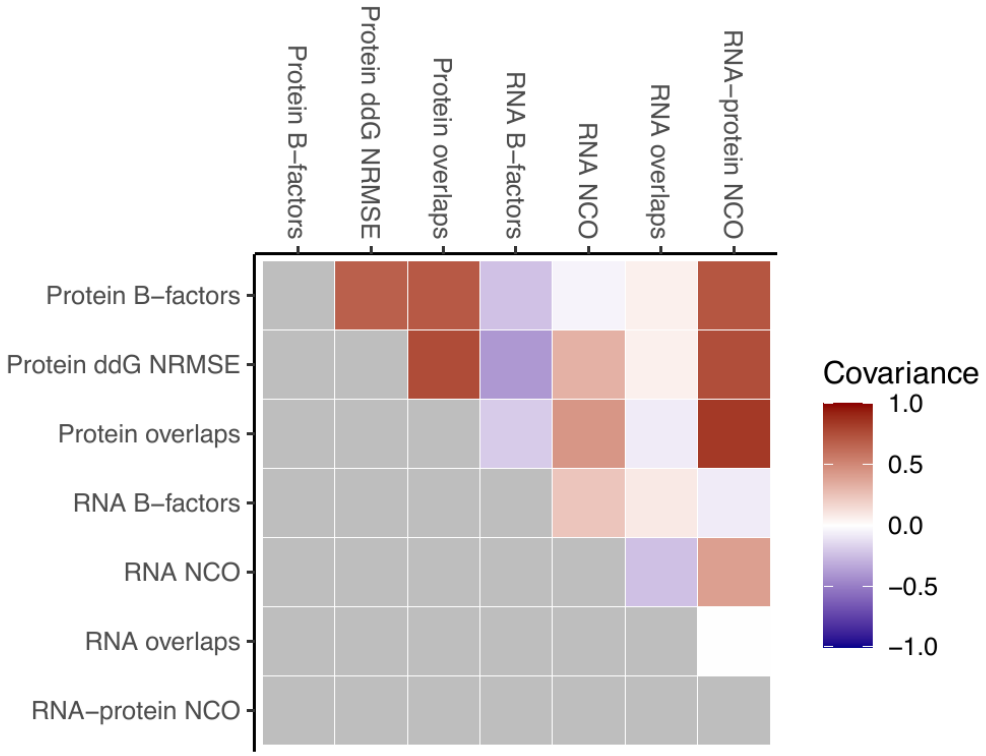


Figure 5.8: **Round 3 covariance between benchmarks.** The Z-score covariance between each pair of benchmarks for the third round of parameter search is shown as a half-matrix, as in Figure 5.2 for the first round.

Strikingly, the optimal parameter set emerging from round 3 is the same as the optimal round 2 set, meaning that no other combination tested leads to a higher rescaled Z-score sum than the 4.3 value it obtains. Figure 5.7 illustrates the performance of the round 3 parameter sets across the benchmarks, Figure 5.8 shows the Z-score covariance matrix for round 3 and Figure 5.9 gives the detailed performances for benchmark pairs.

Since no change was observed in the optimal parameter set, we decided to stop the parameter search with this third round and use the optimal parameter set from rounds 2 and 3 as the new ENCoM parameters.

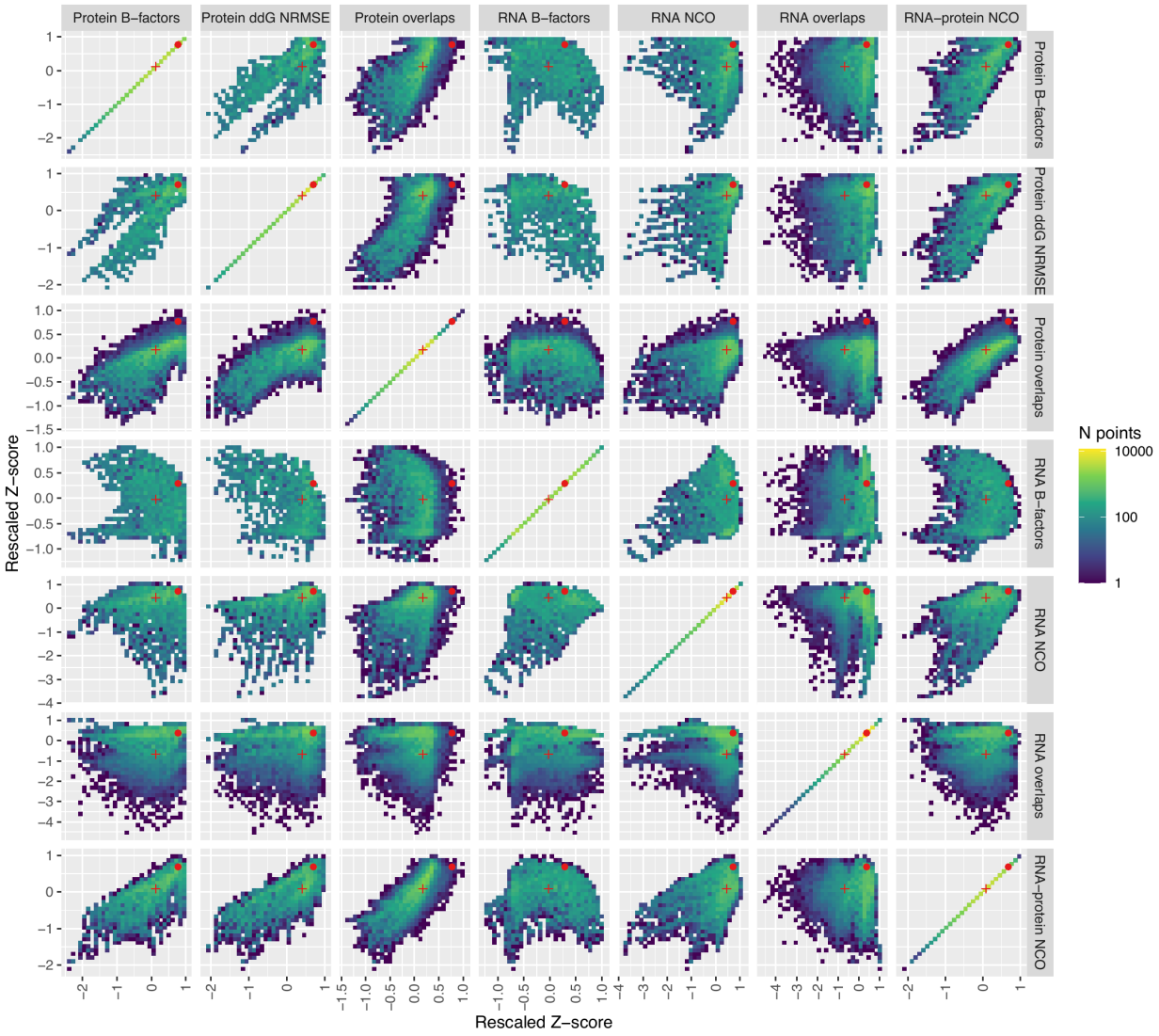


Figure 5.9: **Round 3 detailed performance across benchmark pairs.** As in Figure 5.3, the rescaled Z-scores across all parameter sets are plotted against each other, the performance of the optimal parameter set found is denoted with red dots and the performance of the original ENCoM parameters with red crosses.

5.4 PERFORMANCE ACROSS ALL SEARCH ROUNDS

In order to further investigate our initial question of whether a single parameter set can lead to good performance across all benchmarks, we decided to combine all parameter sets tested across the three search rounds and analyze how performance covaries between benchmark pairs. This combination of all results means that a wider range of values is obtained for every benchmark, as the wide search of the first round led to poorer performances on average. However, the presence of

more parameter sets leading to good performance also prevents the driving of the covariance by these poorer parameter sets from round 1.

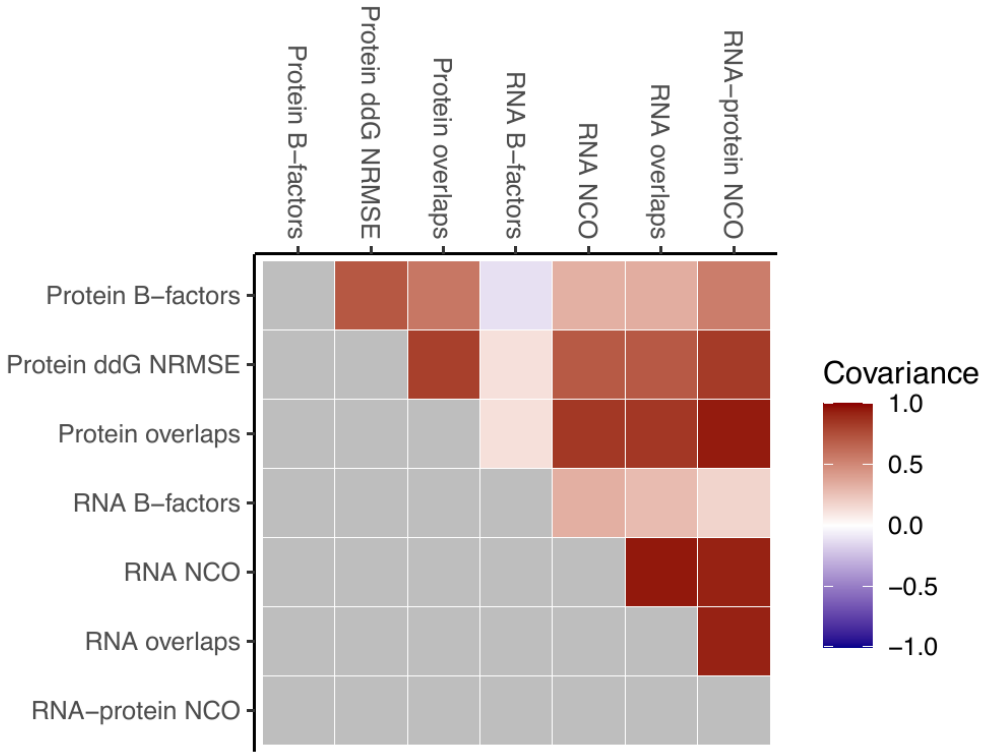


Figure 5.10: **Combined covariance between benchmarks.** The Z-score covariance between each pair of benchmarks across all rounds of parameter search is shown as a half-matrix, as in Figure 5.2 for the first round.

Figure 5.10 gives the covariance matrix for the combined results from all search rounds. Strikingly, almost all benchmark pairs exhibit high covariance, with the exception of the RNA B-factors benchmark with all other benchmarks. Moreover, the protein B-factors benchmark exhibits the second lowest set of covariances, and the only negative covariance happens between RNA and protein B-factors benchmarks. These observations further reinforce the notion that a tradeoff has to be made in parameter space between the prediction of experimental B-factors and the prediction of both large conformational changes and the effects of mutations.

Figure 5.11 illustrates the detailed performance across the benchmark pairs, combining results from the three search rounds. Since round 1 led to the largest performance variance, the shapes of the distributions are closely related to those observed in Figure 5.3. However, the red dots now represent the performance of the new ENCoM parameters (optimal from rounds 2 and 3) and more density is

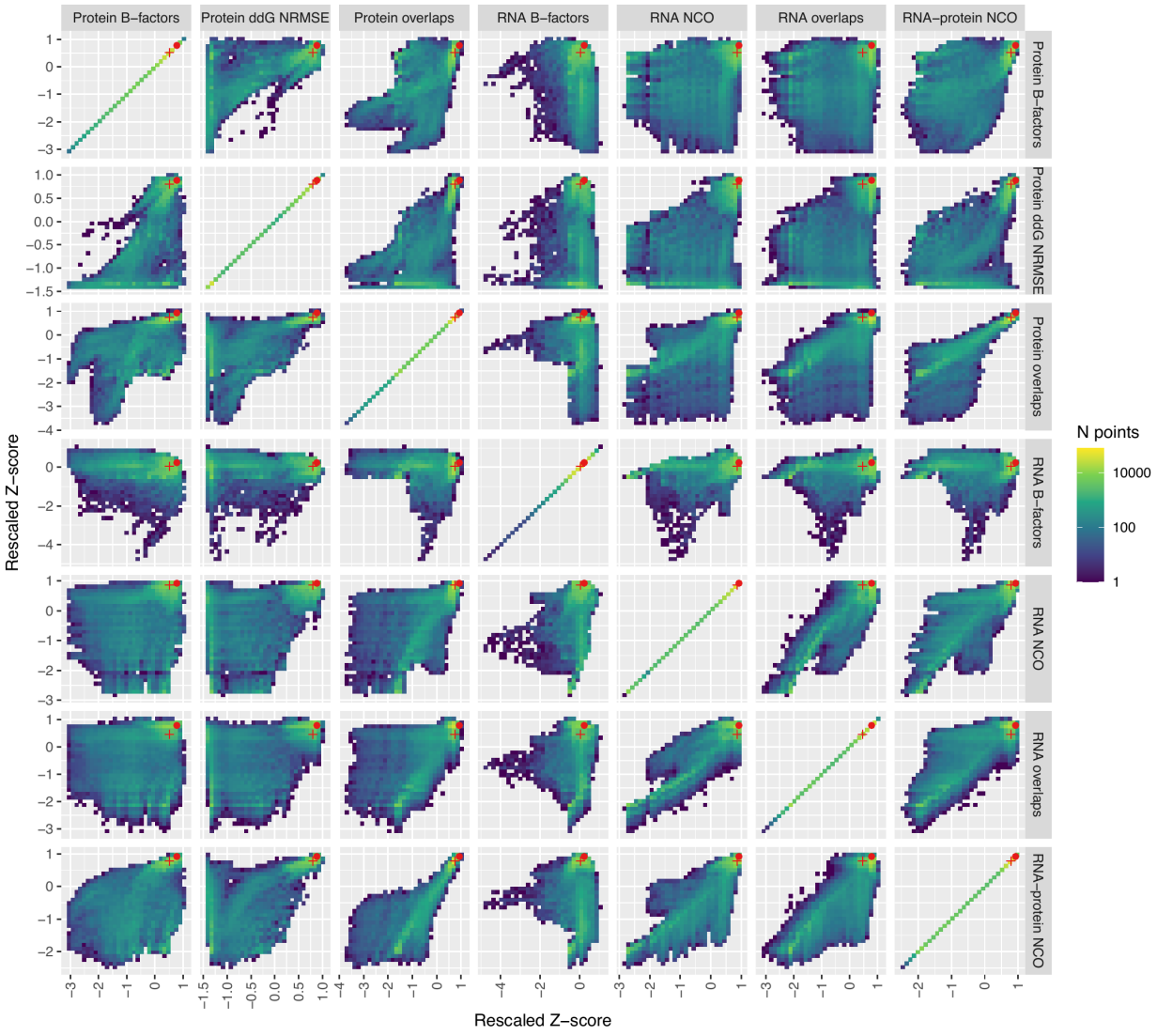


Figure 5.11: **Combined detailed performance across benchmark pairs.** As in Figure 5.3, the rescaled Z-scores across all parameter sets are plotted against each other, the performance of the optimal parameter set round rounds 2 and 3 is denoted with red dots and the performance of the original ENCoM parameters with red crosses.

observed in the higher range of performance, due to the addition of the higher performance round 2 and round 3 results.

The parameter search across the seven diverse benchmarks presented here has allowed the identification of a parameter set leading to superior performance across all benchmarks when compared to the original ENCoM parameters. We have thus redefined the ENCoM parameters in light of this and will be using these improved parameters for the rest of the present thesis. More importantly, the present chapter highlights that a single set of parameters leads to high performance for conformational change prediction of RNA, protein and RNA-protein complexes. This was an important question we initially raised as part of the adaptation of ENCoM for RNA [157], as we hypothesized that a tradeoff could exist between performance on these two types of biomolecules. Indeed, the use of three masses per nucleotide introduces ramifications in the connectivity of the masses in the system, which do not exist in the case of proteins. However, it seems that such a tradeoff is not necessary after all and that in the case of ENCoM parameter optimization, one can have their cake and eat it too.

There are nonetheless some sacrifices to be made, particularly when it comes to the prediction of experimental B-factors. This comes as no surprise and was reported as part of the original ENCoM publication [27]. Moreover, the biggest tradeoff comes from the RNA B-factors prediction benchmark. We already observed that some very low, sometimes negative correlations were present in our dataset of high-resolution RNA crystal structures used for this benchmark [157]. These were present not only in the case of ENCoM, but also for the other two models tested in that work, a power-dependent ANM and a cutoff-based ANM. This observation could mean that some artifacts are present in the dataset used for this benchmark. However, the chosen parameter set still achieves performance in the upper range for the RNA B-factors benchmark. This reinforces the idea that ENCoM, like all coarse-grained ENMs, is a very robust model. The artifacts are averaged out and both the original parameters and the new optimal set perform strikingly well across all pairs of benchmarks, as is apparent in Figure 5.11.

It would have been interesting to include a dataset of dynamics-function relationships in the set of diverse benchmarks, for example a training and testing set pair from one of the three chapters that will follow. However, the computational cost of doing this across the more than 200 000 parameter sets test would have been very high, at least doubling the amount of computation required just to compute the Dynamical Signatures. For example, the miR-125a training set from the hard benchmark presented in Chapter 6 contains 1849 sequence variants, and it takes around 3 seconds CPU time to compute the Dynamical Signature from one variant. The inclusion of this dataset as an additional benchmark would thus have more than doubled the time needed to compute the benchmarks for one parameter set, adding approximately 92 minutes to the 67 taken by the seven actual benchmarks.

More importantly, the testing of so many parameter sets necesarily introduces noise, and thus the machine learning algorithms trained on the Dynamical Signatures could experience performance boosts when the noise patterns aligns with the relevant data by chance. The benchmarks presented here are shielded from this effect because of the large amount of diverse structures that constitute every benchmark. Nonetheless, it will be very interesting to see how the new parameters affect the performance of ENCoM-DynaSig-ML on the miR-125a hard benchmark, as we already reported the performance of the original parameters on this exact dataset [157].

The goal of the parameter search was mainly to investigate the performance covariation across the different benchmarks, and the finding of significantly better ENCoM parameters as a result comes as a welcome surprise. The assumption for the whole parameter search was that the performance landscape is relatively smooth, however this may not be the case. In the future, it would be interesting to investigate this question by using more sophisticated optimization methods, such as genetic algorithms [170] or Monte Carlo optimization [171].

6

MIR-125A MATURATION EFFICIENCY

The present chapter is dedicated to our first dynamics-function case study, in which we investigate dynamical features of microRNA (miR) biogenesis from a dataset of experimental maturation efficiencies of over 26 000 sequence variants of pri-miR-125a. As mentioned in Chapter 1, we refer the reader to Section 1.3.1 for the detailed biological background concerning miR biogenesis.

It is also in the present chapter that we investigate the potential performance increase that can be obtained by using multilayer perceptrons (MLPs) instead of LASSO regression as the machine learning backend of the ENCoM-DynaSig-ML pipeline. As detailed in Section 3.2, LASSO regression allows relatively agressive feature selection and assumes linear independence between the predictor variables, both desired properties for the biological interpretation of the trained model. On the other hand, MLPs can model complex relationships between input variables, hence if such relationships exist they should lead to a gain in performance over LASSO.

As a result of our analyses, we end up selecting LASSO regression as the desired ML backend for the ENCoM-DynaSig-ML pipeline for Chapter 7 and Chapter 8. However, as will be discussed, we do not reject the possibility to use other ML backends, including neural networks. We simply have higher confidence in the generalizability of the pipeline with LASSO backend to sequence variants with higher numbers of mutations than what was seen in training, and this high-throughput prediction capability is one of the advantages of ENCoM-DynaSig-ML.

Ultra-high-throughput predictions of 30 million pri-miR-125a theoretical sequence variants will be shown towards the end of the results section, along with sequence variants specifically optimized to have given properties with the use of a simple asexual genetic algorithm. The LASSO models resulting from our analyses of μ-opioid receptor activation in Chapter 7 and VIM-2 lactamase catalytic efficiency in Chapter 8 could be applied in the same way to virtual screening experiments and the prediction of optimized enzyme variants, respectively. However, we only show predictions in the present chapter for two reasons. First, since the validation of these predictions is beyond the scope of the thesis, it would not add much to show more unvalidated predictions. Second, we have special interest in miR biogenesis and

Mélanie Lemaire, a fellow PhD student from the Major group, has developed an experimental setup to measure the maturation efficiency of pri-miR-125a variants from cellular assays. She is working on predictions shown in the current chapter and we expect the results of these experiments to be submitted to publication during 2023.

The chapter is divided in three sections: first the methodology pertaining to the specific dataset, modeling of pri-miR-125a in 3D, exploration of different ML backends and ultra-high-throughput will be outiled; then the results and discussion specific to microRNA maturation will follow.

6.1 METHODOLOGY

The next subsections will first outline the methodology associated with the experimental miR-125a maturation dataset: the selection of sequence variants, the construction of the different benchmarks for evaluating the machine learning models' performance and the reduction of maturation efficiency measurements to a class prediction problem. We will then detail our approach to modeling pri-miR-125a in 3D, including the selection of the most relevant MC-Sym 3D model, followed by the exploration of thermodynamic scaling factors for the Entropic Signatures and the optimization of MLP architectures. Finally, the ultra-high-throughput prediction methodology will be described and the results will follow.

6.1.1 *Dataset of miR-125a mutations*

The Fang *et al.* dataset of high-throughput miR-125a maturation efficiency contains all possible mutations for thirteen 6-nucleotide boxes plus all fifteen possibilities at the 2-nucleotide bulge, which in addition to the major allele sequence (WT) amount to 53 251 sequence variants (Figure 6.1A). We submitted all sequence variants to 2D structure prediction using the accelerated MC-Flashfold implementation [172] of the MC-Fold software [148]. We found that 29 478 of them adopt WT 2D minimum free energy (MFE) structure (Figure 6.1A). Because NMA assumes the input structure is at equilibrium and thus all variants tested need to share a close 3D structure, we restricted the analysis to these 29 478 sequence variants. Furthermore, we realized that the first eight mutated boxes account for the vast majority (over 90%) of these variants that adopt the WT MFE structure. The proportion of mutations at each position leading to the WT MFE structure is plotted in Figure 6.1B, showing a very clear drop in the proportion of sequence variants adopting the WT MFE structure beyond box 8. We thus further restricted our analysis to the set of 26 960 variants with WT MFE structure from boxes 1-8 to have higher confidence in a common 3D structure for all variants.

6.1.2 *Sequence redundancy: hard benchmark*

Since the mutations were performed exhaustively for each 6-nucleotide box, there is a great deal of sequence redundancy in the dataset. For example, every of the three possibilities of mutation at a given position appears in as much as 1024 sequence variants in the full dataset. To prevent the learning of sequence features by the ML models, which could be guessed as a result of ENCoM's sequence-sensitive potential function (Section 3.1), we contructed a hard benchmark which ensures all mutated positions in the testing set were never mutated in the training set. In that hard benchmark, the middle base pair of mutated boxes 1-8 was reserved for the testing set while all sequence variants affecting only the bottom and/or top base pairs of every box were selected for the training set (illustrated in Figure 6.1C). This left us with 1849 sequence variants in the training set and 116 in the testing set.

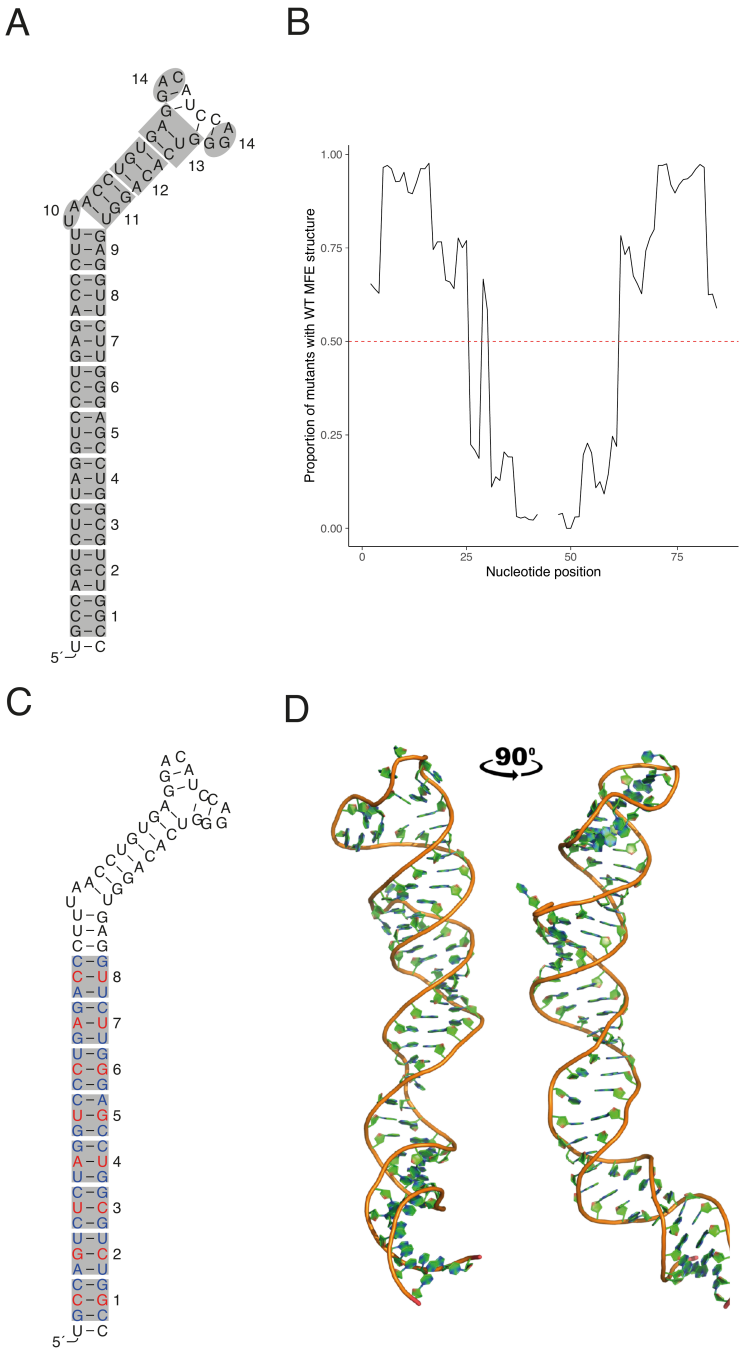


Figure 6.1: **miR-125a 2D MFE structure, mutation boxes, hard benchmark sets and 3D structure.** A) The 14 boxes that were each exhaustively mutated in the Bartel study, shown on the MC-Fold predicted WT MFE structure. B) Proportions of sequence variants containing each position which adopt the WT MFE structure. C) Positions in the hard benchmark testing set shown in red, and positions in the training set in blue. D) The medoid 3D miR-125a structure from the 67 structures predicted by MC-Sym. Taken from our own work [157].

6.1.3 *Class prediction problem*

In some respects, the catalytic efficiency of pri-miR cleavage by the Microprocessor can be seen as a binary classification problem. Indeed, the set of transcripts which have the possibility of being cleaved to become mature miR is very well defined. Thus, the Microprocessor has to somehow "decide" whether to cleave or not when it encounters RNA hairpin structures in the nucleus, which are extremely abundant and among which microRNAs are relatively rare [40]. Moreover, the high-throughput sequencing approach used in the generation of the Fang & Bartel dataset can lead to noise regarding the precise maturation efficiency recorded. Reducing the problem to classification can help mitigate some of that noise.

We decided to reduce the data to binary classes using the following scheme: variants with maturation efficiencies less than 0.5 (half the WT efficiency) were deemed unproductive, and variants with maturation efficiencies higher than 0.8 were deemed productive. In the case of the hard benchmark, the testing set contains 47 productive and 29 unproductive variants.

Since we are interested in predicting theoretical variants with very high and very low maturation efficiencies, we still train all the machine learning models tested in this chapter to predict a scalar maturation efficiency. We then use these predicted efficiencies to generate receiver operating characteristic (ROC) curves and precision-recall (PR) curves for the binary class prediction problem and compute the area under the curve as the performance metric in both cases (AU-ROC and AU-PR, defined in Section 4.4.2). This reduction to classification is useful in the case of the hard benchmark described above and the mutated boxes benchmarks described below, because they represent hard tasks where the positions which are mutated in the testing set are never mutated in the training set. For the benchmarks which allow sequence redundancy, we prefer the classical predictive coefficient of determination (predictive R^2) as it is directly interpretable as the proportion of variance explained by the model.

6.1.4 *Mutated boxes benchmarks*

Another possibility to remove sequence redundancy between training and testing sets is to take all variants from a single mutated box as the testing set, and all variants affecting the other 7 mutated boxes in our filtered dataset as the training set. Not only does these 8 additional benchmarks contain no sequence redundancy, they also test the models' abilities to capture longer-range effects. Since the most apparent effects of a mutation tend to be local, the most important features learned by the model will tend to be located on the 7 boxes which are unaffected in the testing set. Thus, these 8 benchmarks can be used to investigate the effect of

different β values for the EntroSigs and regularization strengths on the models' abilities to capture long-range effects.

Table 6.1: **Mutated boxes, hard benchmark and whole dataset statistics.** The number of sequence variants in the training set, testing set, testing set restricted to binary classification (test classes), processed variants (maturation efficiency above 0.8) and unprocessed variants (maturation efficiency below 0.5) is given. The numbers differ from exhaustive enumeration of sequence variants since the variants are filtered to adopt the WT 2D MFE structure. For the two classes, the percentage of the binary classification test set they represent is given in parentheses. For the whole dataset, the testing set size column is used and the training set column is left empty.

Dataset	Train size	Test size	Classes size	Processed variants	Unprocessed variants
Box 1	24 418	2543	2079	183 (9%)	1896 (91%)
Box 2	23 001	3960	3204	580 (18%)	2624 (82%)
Box 3	23 132	3829	3629	38 (1%)	3591 (99%)
Box 4	23 259	3702	3208	624 (19%)	2584 (81%)
Box 5	23 003	3958	2925	1311 (45%)	1614 (55%)
Box 6	23 825	3136	2803	66 (2%)	2737 (98%)
Box 7	24 261	2700	2396	294 (12%)	2102 (88%)
Box 8	23 829	3132	1885	1197 (64%)	688 (36%)
Whole dataset	—	26 960	22 129	4293 (19%)	17 836 (81%)
Hard benchmark	1849	116	76	47 (62%)	29 (38%)

Table 6.1 gives the number of productive and unproductive variants for each of the 8 boxes benchmarks, as well as for the hard benchmark and for the whole dataset of 26 960 variants from the first 8 mutated boxes with WT 2D MFE. As was already observed by Fang and Bartel [13], some boxes are much more tolerant of mutations than others. For instance, box 3 corresponds to the mismatched GHG motif and is very intolerant of mutations, hinting at the importance of that specific motif.

6.1.5 *5-fold cross-validation*

In order to explore the performance gains to be made when there is sequence redundancy between training and testing sets, we performed 5-fold cross-validation of the entire dataset of variants from the first 8 boxes with WT 2D MFE. To ensure uniform sampling of the mutated boxes, all variants affecting every individual box were split randomly in 5 sets, and the combination of the n^{th} set from every box constituted the n^{th} testing set for this cross-validation.

6.1.5.1 *Sequence vectors*

Since there is now sequence information shared between training and testing sets, we test the performance of sequence vectors for this 5-fold cross-validation. We simply construct the vectors as 1-hot encodings of the variant sequence: each position is represented by four values, of which three are set to -1 and one is set to 1. The first is set to 1 if the nucleotide is A, the 2nd if it is C, the 3rd for a G and the last for a U. We then standardize these sequence vectors in the same fashion as we do for all predictors used to train ML models. Since the length of our modeled pri-miR-125a is 86 nucleotides, the sequence vectors are of length 344.

6.1.5.2 *Inverted dataset*

Our observations of MLP performance patterns on the 5-fold cross-validation led us to hypothesize that the models could somehow memorize some important sequence patterns. In order to test this hypothesis, we constituted another training-testing pair with sequence redundancy, which we call the inverted dataset because the training set is much smaller than the testing set. This inversion was done on purpose as we wanted to see whether the models could classify the variants containing multiple mutations after being trained on variants containing few mutations. Precisely, the training set contains all variants with at most 2 mutations, while the testing set contains all variants with 3 or more mutations, again restricting the analysis to the subset of variants already filtered for sharing the WT 2D MFE and affecting only the first 8 boxes. This selection scheme leads to a training set of size 1094 and a testing set of size 25 866.

6.1.6 *MLP architectures optimization*

When trained on P input predictors, a multilayer perceptron (MLP) with N hidden layers all of size S will have a number of free parameters given by (already defined in Equation 3.6):

$$F_{\text{params}} = 2 + S * P_{\text{input}} + 2 * S + \sum_{i=2}^{N_{\text{hidden}}} \left[S^2 + S\right] \tag{6.1}$$

For the 5-fold cross-validation, the training sets all contain 21 558 observations, thus 21 557 degrees of freedom (DOFs). We restrict MLP architectures for this test to architectures with a number of free parameters at most 75% of the training set DOFs, in order to prevent the learning of noise patterns. An observation that many studies have made is that provided a sufficient number of neurons in the hidden layer(s), MLPs rarely perform better with more than one or two hidden layers [173]. Thus, we tested architectures with 1 or 2 hidden layers, with equal sizes of 2, 5, 10,

20, 30, 40, 50 or 60 hidden neurons. We rejected the combinations which generated more free parameters than our cutoff of 75% the number of training DOFs.

For the inverted dataset, the small training set size of 1094 restricts the number of hidden neurons we can use in the first layer without the free parameters vastly exceeding the training DOFs. We this reason, we made two modifications to the architectures tested: we allowed the free parameters to go up to 200% the training DOFs, and we tested the inclusion of more hidden layers since we had to restrict the layer size to a maximum of 8 neurons. We thus tested all combinations of 1 to 5 hidden layers, each comprising 1 to 8 neurons. Again, we did not test the combinations leading to more free parameters than our cutoff of 200% the training DOFs.

6.1.7 *pri-miR-125a structure prediction*

Until recently, there were no solved 3D structures of pri-miRs despite great interest from the community to understand how their structure varies across miR families and what features are necessary for their cleavage by the Microprocessor [40, 43, 174]. The first glimpse into pri-miR structure came in 2020 when Jin *et al.* published the cryo-EM structure of pri-miR-16-1 in complex with Drosha [175]. Their work illustrates the detailed interaction of Drosha with the pri-miR, with the cleavage site positioned close to the catalytic residues but the pri-miR still intact through the use of an inactivated Drosha variant. However, a 26 nucleotide region comprising the apical loop is missing from the structure. Nonetheless, this structure would represent a great starting point for *is silico* mutagenesis in the same manner as outlined in the present chapter in the case of pri-miR-125a. We decided to use predicted 3D structures of pri-miR-125a for the following reasons:

1. We have special interest in miR-125a because of the population SNP in its sequence leading to higher risks of breast cancer. Previous work from our group uncovered the correlation between pri-miR-125a 2D structural dynamics and the processing efficiency of 15 sequence variants at the SNP position [14].

2. Because of this interest, we currently have the experimental setup to measure maturation efficiency of miR-125a variants using northern blot [176]. The predictions made at the end of the present chapter are currently being tested in the lab and while outside the scope of the present thesis, we expect to have the results in 2023.

3. pri-miRs adopt simple stem-loop structures, and in the case of pri-miR-125a it is striking that over half sequence variants from the Fang and Bartel experimental dataset have the same predicted 2D MFE structure. We thus have confidence in 3D RNA structure prediction programs to generate models

that are close to reality. Moreover, the most complex part of pri-miRs is the apical loop, which would have to be modeled in any case as it is missing from the cryo-EM structure.

4. Lastly, to the best of our knowledge, pri-miR-16-1 is the only pri-miR for which a partial 3D structure is known. As part of future work, we are interested in applying the methods developed in the present chapter to study the structural dynamics across human miR families. In order to do so, their 3D structures will have to be modeled and thus the present chapter can be seen as validation that our modeling protocol with the MC-Fold | MC-Sym pipeline [148] is sufficient for such a study.

To generate a 3D model of pri-miR-125a, MC-Sym [148] was run using the WT sequence and 2D MFE structure as input, with default parameters. This generated 67 predicted 3D structures, of which the medoid structure is shown in Figure 6.1D. In the MC-Sym algorithm, the free energy of folding is set by the preliminary prediction of nucleotide cyclic motifs, which happens at the 2D level in the MC-Fold algorithm. The order of the 67 models thus does not reflect any predicted biophysical property and simply corresponds to the order in which they are generated by the search procedure [148].

6.1.8 *MC-Sym model selection*

Among the 67 pri-miR-125a models predicted with MC-Sym, some are almost guaranteed to be closer to the physiological equilibrium structure than others. Under the hypothesis that structural dynamics play a role in pri-miR cleavage by the Microprocessor, the hard maturation benchmark can be used to select the MC-Sym model which leads to the best predictive capabilities as the closest one to a biologically relevant state. The training set size of the hard benchmark is relatively modest at 1849 sequence variants, so we restrict the ML backend of the ENCoM-DynaSig-ML pipeline to LASSO regression for this 3D model selection step.

6.1.9 *Entropic Signatures scaling factors*

In order to identify an interesting range of values for the β thermodynamic scaling factor of the Entropic Signatures, we computed them using the medoid MC-Sym model (model 41) for the widest possible range of values, using a logarithmic search in either direction until the output became constant.

Figure 6.2 illustrates the result of this search of scaling factors. As already discussed in Section 4.3.3, lower scaling factor values correspond to higher temperatures and thus more even distribution of vibrational entropy across all normal modes. We

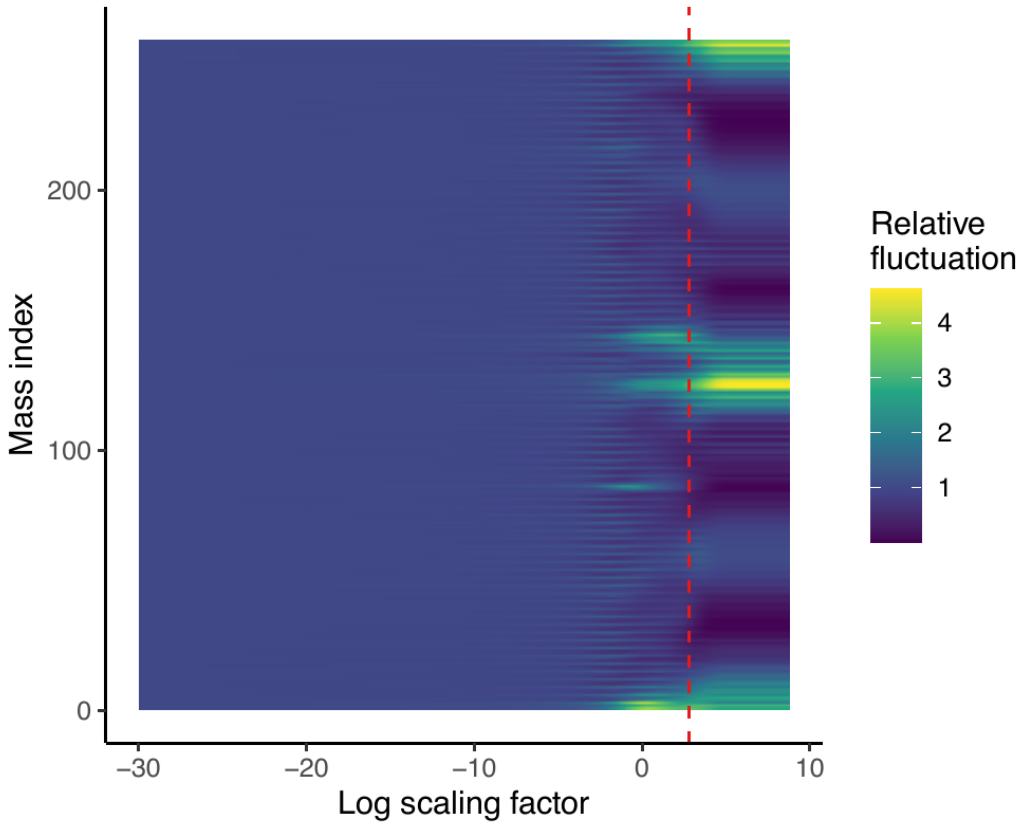


Figure 6.2: **Entropic Signatures for the medoid pri-miR-125a 3D structure across a wide range of scaling factors.** The relative fluctuation is shown for every bead in the system (y axis) across a logarithmic search of thermodynamic scaling factors. The Entropic Signatures are linearly rescaled so that the average fluctuation is always 1. The scaling factor leading to the highest correlation with the mean-square fluctuations ($e^{2.8}$) is shown with a red dashed line.

identify the range of scaling factors from e^{-10} to e^{5} as capturing all essential features, from an almost constant entropic signature at the lower end to a dominance of the first normal mode at the higher end.

Figure 6.3 shows the contributions to vibrational entropy for all normal modes, across the selected range of scaling factors. Indeed, we can see that the highest values lead to a dominant entropic contribution from the slowest mode, while the lowest values lead to an almost constant contribution across the whole range of normal modes.

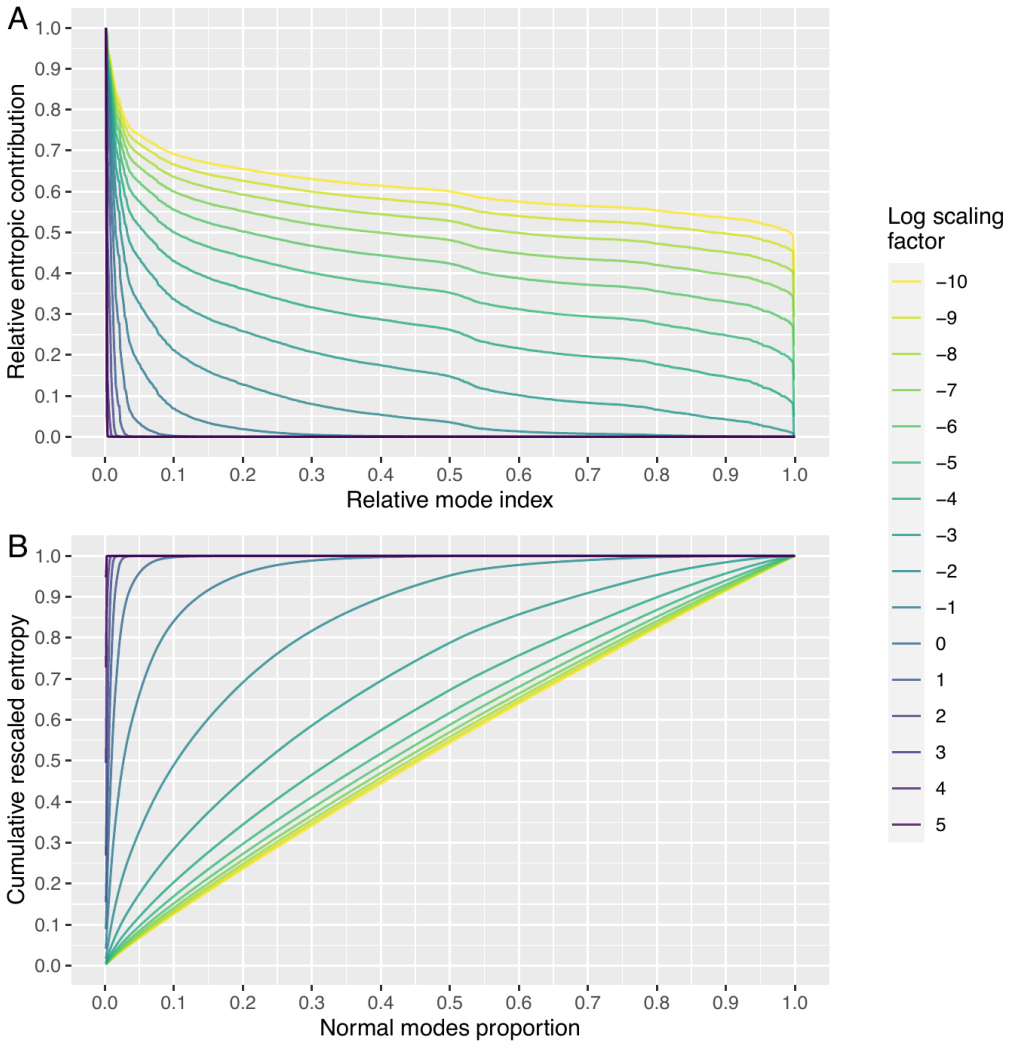


Figure 6.3: **Vibrational entropy proportions for pri-miR-125a across selected scaling factors.** A) The relative entropic contribution of each normal mode is given, rescaled so that the first nontrivial mode has a contribution of 1. The mode indices are also rescaled from 0 to 1 so that the proportion of total internal space they represent is directly represented on the axis. B) Same as A), but cumulative entropy is shown, rexcaled this time to sum to 1.

6.1.10 *Ultra-high-throughput maturation efficiency prediction for pri-miR-125a variants*

After the analysis across all benchmarks, we selected the LASSO model trained on MC-Sym model 61, $\beta = e^{-1.5}$, $\lambda = 2^{-7}$ as the most promising and most generalizable model. This model was used to generate ultra-high-throughput predictions of maturation efficiency for pri-miR-125a sequence variants.

6.1.10.1 *Random sequence variants*

We first interrogated the model on randomly generated pri-miR-125a sequence variants spanning the whole sequence. These random variants were generated with a uniform distribution of 3-12 affected base pairs. For each affected base pair, one of the 15 mutated possibilities was randomly selected. In order to maintain the ability to model the random variants as base substitutions on the MC-Sym 3D model, we pre-filtered the random variants through MC-Flashfold and kept those which adopted the WT 2D MFE structure. We binned the found sequences according to their folding energy in three classes: low folding energy, medium folding energy and high folding energy. We kept 10 million random sequences for each of these bins, for which we ran the LASSO model outlined above on the combination of ENCoM EntroSigs at $\beta = e^{-1.5}$ and MC-Fold enthalpy of folding, for a total of 30 million evalutations of random sequence variants.

6.1.10.2 *Asexual genetic algorithm*

In order to find theoretical variants with either very high or very low maturation efficiencies, we came up with a simple asexual genetic algorithm (GA) outlined in Listing 6.1.

Listing 6.1: Single iteration of the asexual genetic algorithm for maturation efficiency optimization

```
enthalpies = []
for variant in n_variants_folded:
        n_mutations = uniform(min_mutations, max_mutations)
        variant = mutate_random_positions(starting_sequence, n_mutations)
        enthalpies.append(MC-Flashfold(variant))
keep Y variants with top/bottom/specific enthalpy
efficiencies = []
for variant in top_variants_enthalpy:
        efficiencies.append(evaluate_efficiency(variant))
keep X variants with top/bottom efficiencies
```

The parameters of the GA are thus the range of random number of mutations introduced, the number of random variants folded, the number of variants fully evaluated by computing the ENCoM EntroSig and computing the predicted maturation efficiency from the LASSO model selected, and the number of top or bottom variants, according to maturation efficiency, to keep for the next generation. Note that this simple algorithm will tend to continuously accumulate mutations along a specific path, so we run it repeatedly in order to obtain diversified sequence variants.

6.2 RESULTS

The following subsections will first present the selection of the most relevant MC-Sym pri-miR-125a 3D model using the hard benchmark, the performance on the 8 boxes benchmarks and the features learned by the best LASSO models. Then, the 5-fold cross-validation will be presented along with the different MLP architectures tested and the preformance of the sequence vectors, followed by the performance on the inverted dataset. Finally, ultra-high-throughput maturation efficiency predictions of upwards of 30 million theoretical sequence variants will be made using the most generalizable model found and will be followed by the presentation of variants optimized through our simple asexual genetic algorithm.

6.2.1 *MC-Sym model selection*

While our computational pipeline is fast enough to be applied to tens of millions of pri-miR-125a variants, as presented later in Section 6.2.4, we decided to first select an MC-Sym model using performance on the hard benchmark as our criteria. The reason for selecting a single 3D model is to prevent the finding of good statistical models in the later analyses by chance, due to the high number of combinations that would be tested if we kept all MC-Sym models. Moreover, it is expected that a small subset of the predicted MC-Sym models better represent the real equilibrium conformation of pri-miR-125a, and so we can take advantage of the experimental data to look for the optimal 3D model. Finally, we want to find parameter combinations that generalize beyond sequence, hence the use of the hard benchmark for this selection step. Had we used a training/testing split containing redundant sequence information, it could have led to good performance simply due to the statistical models inferring sequence from the Dynamical Signatures (as ENCoM is sensitive to sequence). However, the training set size is relatively modest for this hard benchmark, so we restricted ourselves to LASSO regression for the model selection step. At this size, an MLP with just one hidden layer of 8 neurons would have around the same number of free parameters as the degrees of freedom in the training set, so it could pick up noise patterns and overfit (the input layer would be around $1/8$ the training set size, with 257 Dynamical Signature positions and the MC-Fold enthalpy).

On its own, the MC-Fold enthalpy of folding already gives statistically significant classification of the variants in the hard benchmark. Figure 6.4 shows the classification performance when training a simple linear regression model to predict maturation efficiency from the enthalpy of folding alone. Since there is class imbalance in our classification problem and the testing set size is modest, we performed random simulations to assess the significance of the MC-Fold-based classifier. MC-Fold performs better than the best random classifier from 1000 iterations, thus its performance is statistically significant at the $p < 0.001$ level.

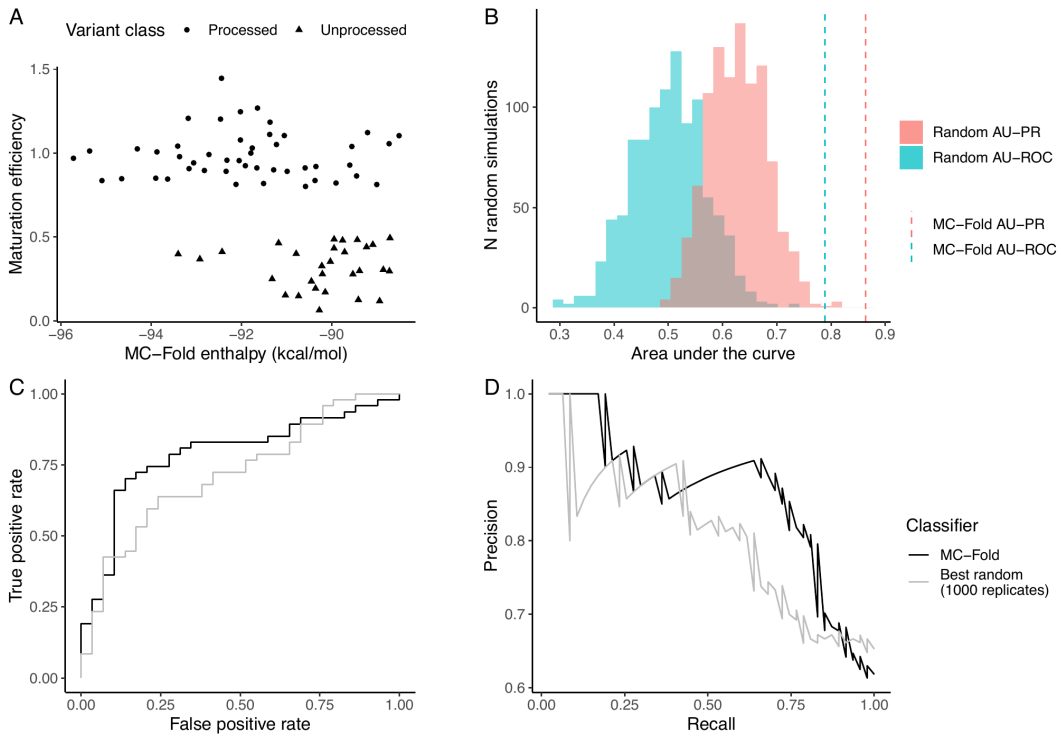


Figure 6.4: **Classification performance using the MC-Fold enthalpy of folding alone.** A) The maturation efficiency for the 76 variants in the class prediction hard benchmark is shown as a function of the enthalpy of folding predicted by MC-Fold. There are 47 variants in the "processed" class (measured efficiency above 0.8) and 29 variants in the "unprocessed" class (measured efficiency below 0.5). B) Simulated distributions showing area under the receiver operating characteristic (ROC) and precision-recall (PR) curves. The distributions result from 1000 replicates of fitting a linear regression model to the maturation efficiency values in the training set using normally distributed noise as the predictor variable. The AUCs are computed after predicting 76 values from each random regression model and using these to predict the classes. The MC-Fold performance is shown with lines. In both cases, MC-Fold outperforms the best random model from the 1000 replicates, which means its performance is statistically significant at the $p < 0.001$ level.

For every of the 67 MC-Sym pri-miR-125a structures, we generate Entropic Signatures with 65 different values for the β thermodynamic scaling factor, from e^{-10} to e^{6} in log increments of 0.25. For each of these 4355 combinations, we test 16 values for the regularization strength of the LASSO model, from 2^{-15} to 1 in $\log_2$ increments. The predictor variables for every LASSO model are the 257 positions of the Entropic Signatures plus the MC-Fold folding enthalpy. Since 4355 combinations of 3D model and β scaling factor are tested, improvement in performance relative to the MC-Fold enthalpy alone could due to chance. Thus, in order to assess whether the addition of the Entropic Signatures leads to a significant gain in

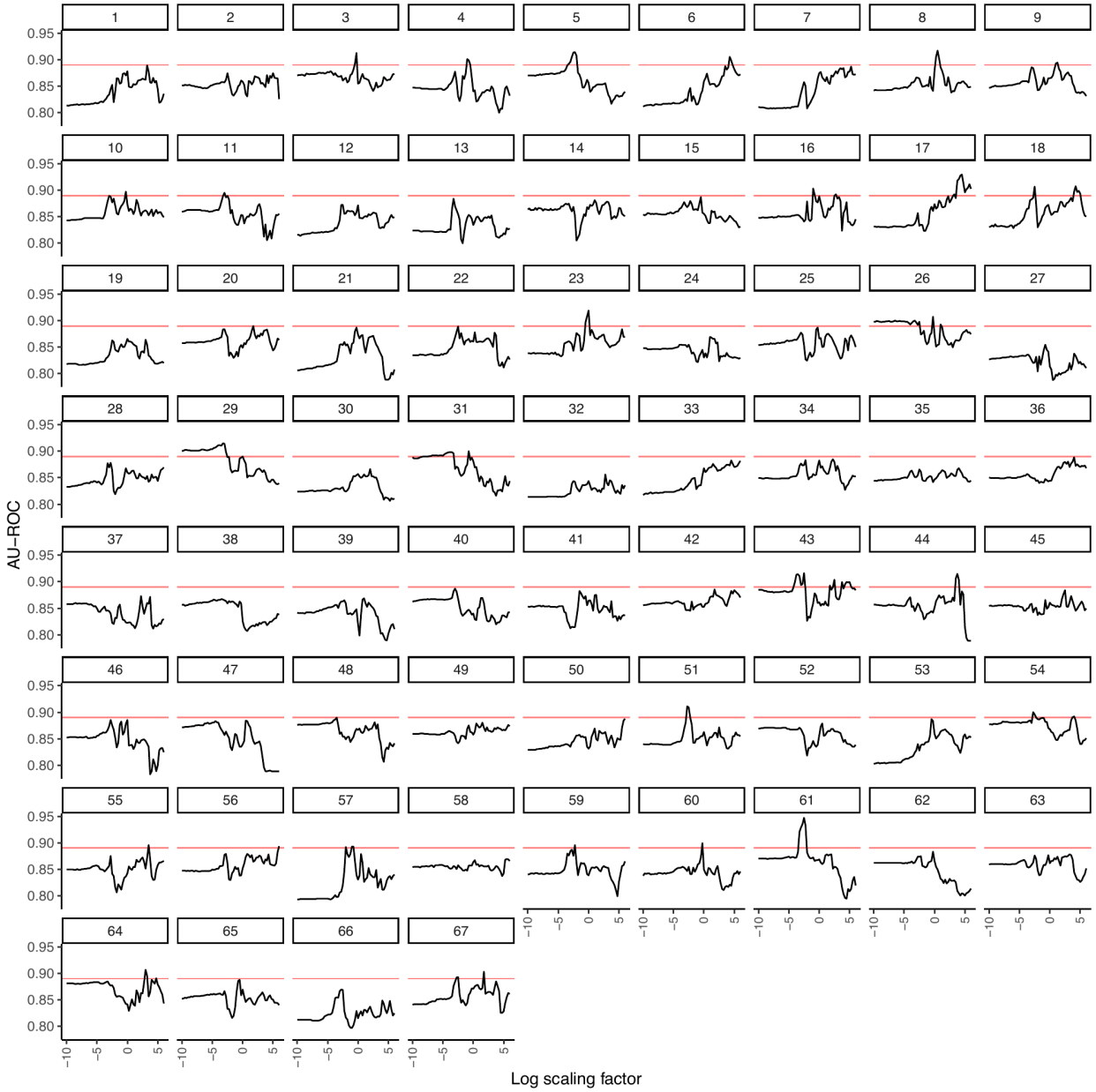


Figure 6.5: **Best AU-ROC across 67 MC-Sym models for the hard benchmark.** The area under the receiver operating characteristic curve (AU-ROC) is shown for every of the 67 models as a function of the β scaling factor. The red line denotes the threshold for statistical significance at the $p < 0.01$ level, according to our simulations. For every combination of MC-Sym model and β value, the best AU-ROC from the 16 regularization strengths is shown. The detailed performances are given in Figure 6.6. The best performance of 0.947 AU-ROC is achieved by MC-Sym model 61 at $\beta = e^{-2.5}$ and regularization strength of 2^{-10}. This performance is significant at the $p < 0.001$ level.

performance when combined with MC-Fold, we simulated for 1000 replicates the training of 4355 sets of 16 LASSO models (one for each regularization strength), with 257 predictors taken from a random normal distribution and 1 predictor being the true predicted MC-Fold energy. For every replicate, we select the best AUC-ROC and AUC-PR obtained. The values obtained impose a stringent threshold for significant improvement relative to MC-Fold alone, equivalent to the Bonferroni correction [177]. In truth, this is probably a very conservative threshold, as the 65 scaling factors tested for every MC-Sym model give rise to related EntroSigs (see Figure 6.2).

Figure 6.5 gives the AU-ROC obtained at the optimal regularization strength from the range tested for every combination of MC-Sym model and β scaling factor. A red line shows the stringent threshold for significant improvement over MC-Fold enthalpy alone at the $p < 0.01$ level. Model 61 exhibits a striking performance peak of 0.947 AU-ROC at $\beta = e^{-2.5}$ and $\lambda = 2^{-10}$. This 0.947 has a significance value of $p < 0.001$ for improvement over MC-Fold enthalpy, which is the lowest value our simulations can detect.

Figure 6.6 shows the detailed AU-ROC values obtained for every MC-Sym model and every combination of λ regularization strength and β scaling factor. Values under the stringent threshold for significant improvement at $p < 0.01$ relative to the MC-Fold enthalpy alone are shown in gray. Let us remind that according to the Bonferroni correction, even if a single combination of parameters led to a value above this threshold, the improvement could be considered statistically significant as we have corrected for the number of tests performed. We can observe that numerous such combinations lead to significant improvement. Moreover, they seem to cluster on specific MC-Sym models and specific ranges of β scaling factor. The highest values are reached for model 61, as described above. Furthermore, for model 61, the parameter combinations leading to significant improvement cluster together across a range of 5 adjacent β values, from $e^{-2.25}$ to $e^{-3.25}$. Similar clustering patterns of significant improvements can be observed for other MC-Sym models, for instance models 5, 17, 26, 29 and 31. These clustering patterns of good performance add to our confidence that the EntroSigs do contibute significant information to the predictions, even in the case of this hard benchmark. Indeed, if the improvements were due to chance alone, one would expect a random pattern of improvement across the MC-Sym models and parameter combinations.

We also investigated the area under the precision-recall curve (AU-PR) as an additional performance metric for the classification of the hard benchmark test set. As defined in Section 4.4.2, ROC curves only depend on positive predictions, as they plot the true positive rate against the false positive rate, and PR curves can be seen as complementary to the analysis since they also consider false negatives (part of the recall definition, see Section 4.4.2.2). Figure 6.7 presents the best AU-PR obtained for every combination of MC-Sym model and β value, again with the

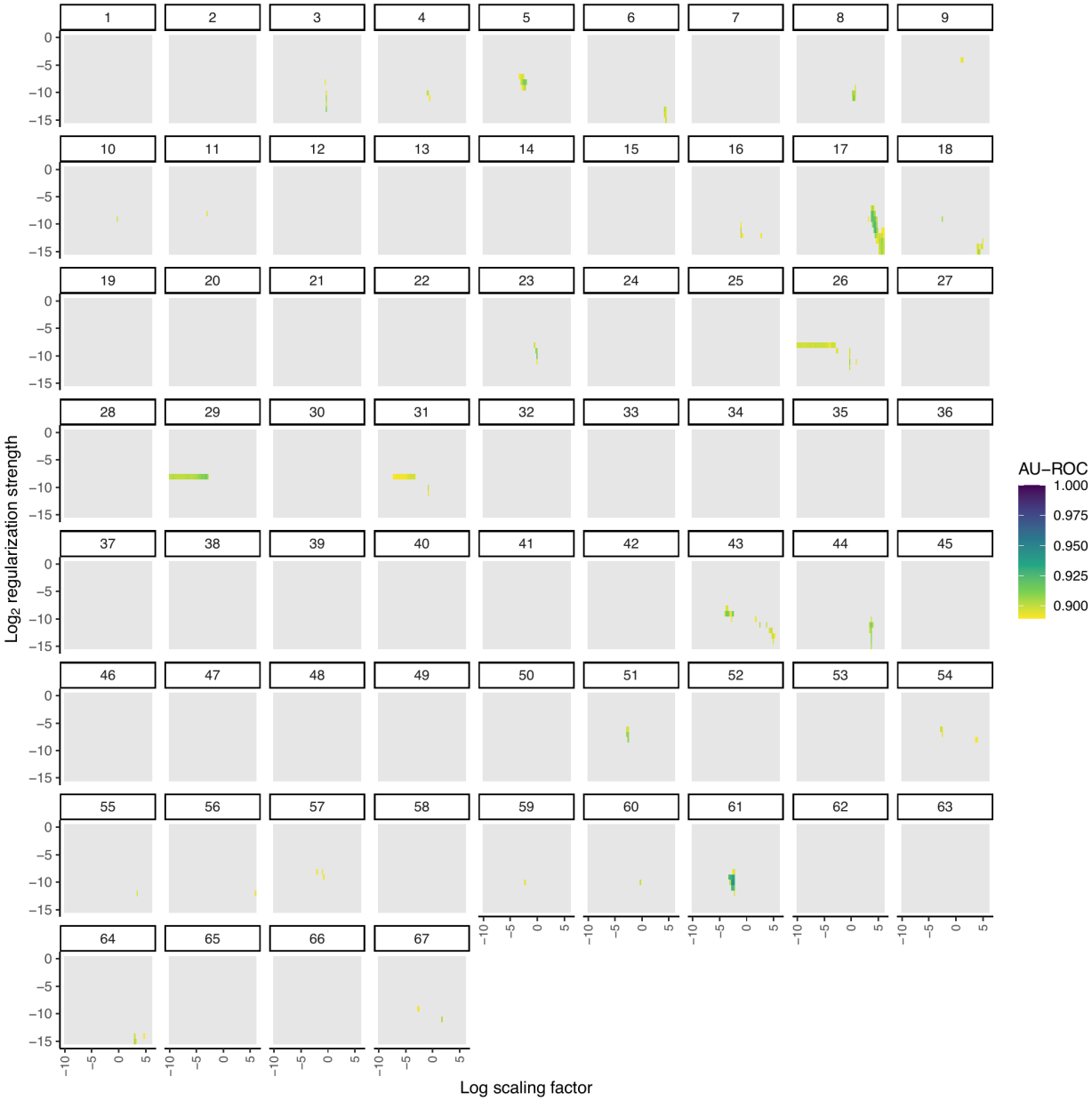


Figure 6.6: **Detailed AU-ROC across 67 MC-Sym models for the hard benchmark.** The area under the receiver operating characteristic curve (AU-ROC) is shown for every of the 67 models as a function of the β scaling factor and λ regularization strength, only when it reaches statistical significance compared to MC-Fold alone at the $p < 0.01$ level (otherwise shown in gray). The best performance of 0.947 AU-ROC is reached for MC-Sym model 61, $\beta = e^{-2.5}$ and $\lambda = 2^{-10}$.

threshold for statistically significant improvement compared to MC-Fold enthalpy alone at the $p < 0.01$ significance level shown as a red line.

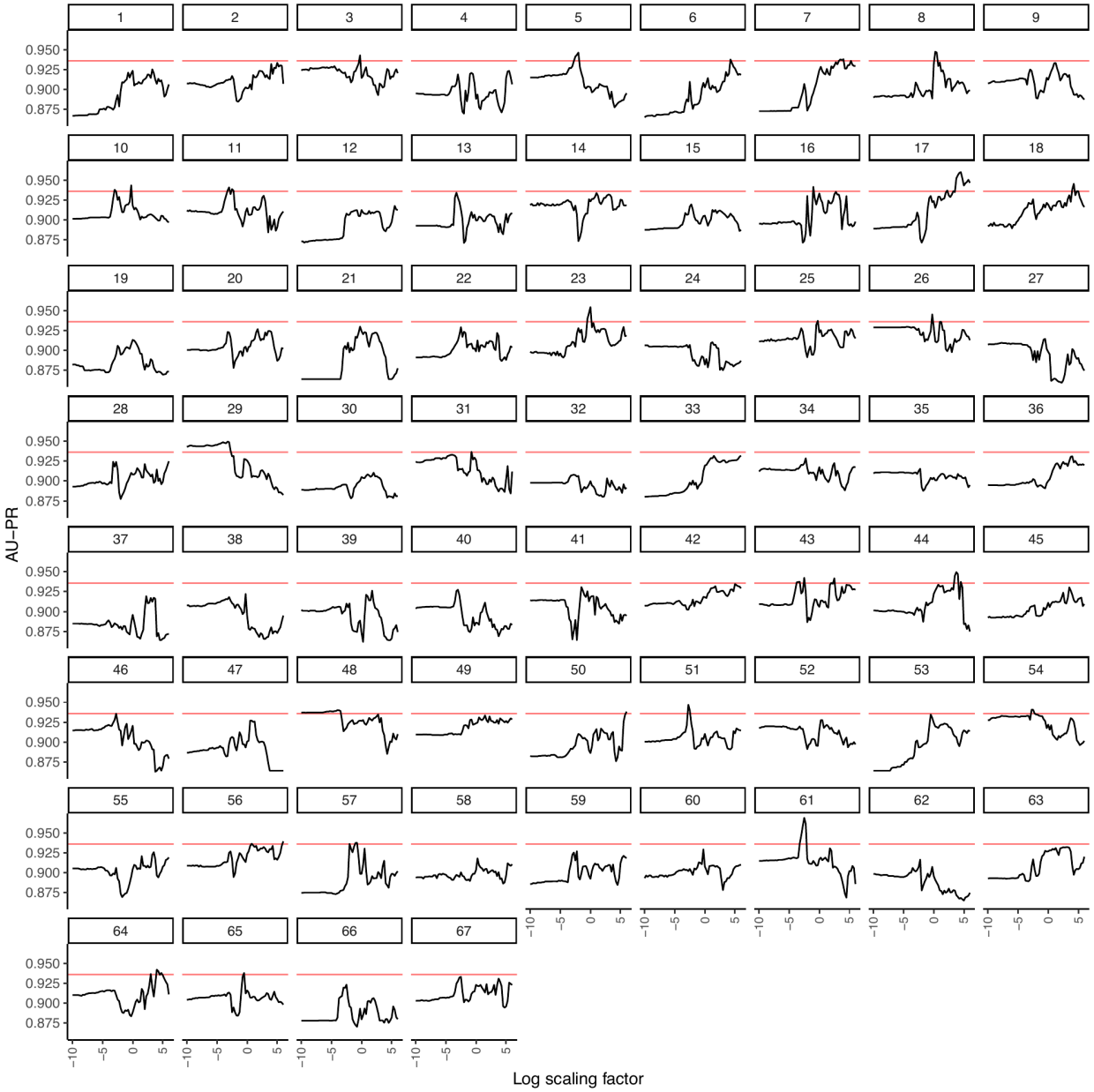


Figure 6.7: **Best AU-PR across 67 MC-Sym models for the hard benchmark.** The area under the precision-recall curve (AU-PR) is shown for every of the 67 models as a function of the β scaling factor, in a manner similar to Figure 6.5. The best performance of 0.969 AU-PR is achieved by the same parameter combinations as for AU-ROC: MC-Sym model 61, $\beta = e^{-2.5}$, $\lambda = 2^{-10}$.

The detailed AU-PR across all parameter combinations are shown in Figure 6.8, similarly as for the AU-ROC in Figure 6.6. Again, clusters of significant performance appear for specific MC-Sym models, and the highest AU-PR of 0.969 is obtained

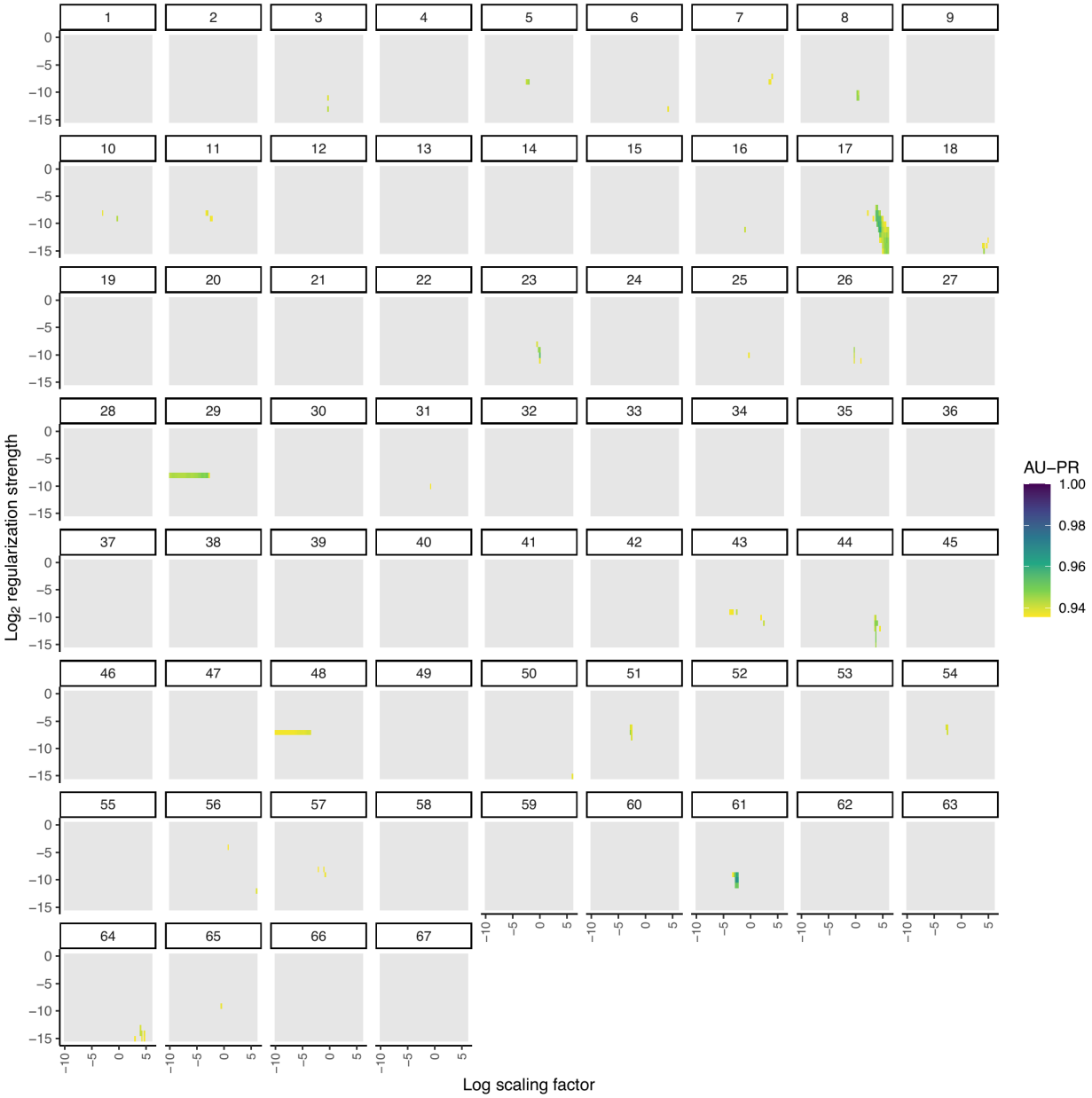


Figure 6.8: **Detailed AU-PR across 67 MC-Sym models for the hard benchmark.** The area under the precision-recall curve (AU-PR) is shown for every of the 67 models as a function of the β scaling factor and λ regularization strength, only when it reaches statistical significance compared to MC-Fold alone at the $p < 0.01$ level (otherwise shown in gray). The best performance of 0.969 AU-ROC is reached for MC-Sym model 61, $\beta = e^{-2.5}$ and $\lambda = 2^{-10}$ (same combinations of parameters as the best AU-ROC).

with model 61, $\beta = e^{-2.5}$ and $\lambda = 2^{-10}$. These are the same parameters that lead to the highest AU-ROC. Thus, we selected model 61 as the most biologically

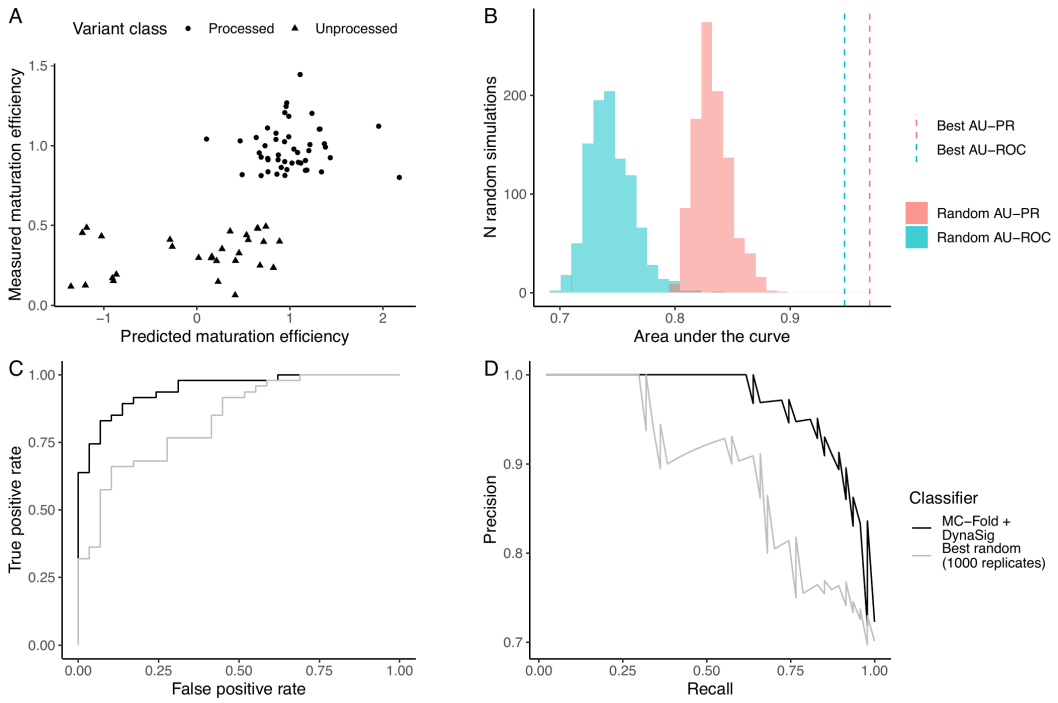


Figure 6.9: **Classification performance using the MC-Fold + DynaSig combination with model 61.** A) The measured maturation efficiency for the 76 variants in the class prediction hard benchmark is shown as a function of the predicted maturation efficiency using the LASSO model trained on MC-Sym model 61 Entropic Signatures with $\beta = e^{-2.5}$, combined with the MC-Fold enthalpy, and with regularization strength $\lambda = 2^{-10}$. There are 47 variants in the "processed" class (measured efficiency above 0.8) and 29 variants in the "unprocessed" class (measured efficiency below 0.5). B) Simulated distributions showing area under the receiver operating characteristic (ROC) and precision-recall (PR) curves. The distributions result from 1000 replicates of fitting 4355 linear regression models to the maturation efficiency values in the training set using normally distributed noise as the predictor variable and keeping the best performance. 4355 is the number of combinations of MC-Sym models and scaling factors, so these simulations are equivalent to the Bonferroni correction. The performance of the combination is shown with lines, significant a the $p < 0.001$ level. C) ROC curves for the best model and the best random simulation. D) PR curves for the best model and the best random simulation.

significant model from the 67 MC-Sym models. Figure 6.9 shows the classification performance of the LASSO model trained with these aforementioned parameters. We observe a more striking improvement over random when comparing to the MC-Fold enthalpy alone (Figure 6.4) despite correcting for the number of tests performed, highlighting the high complementarity of the EntroSigs and MC-Fold enthalpy.

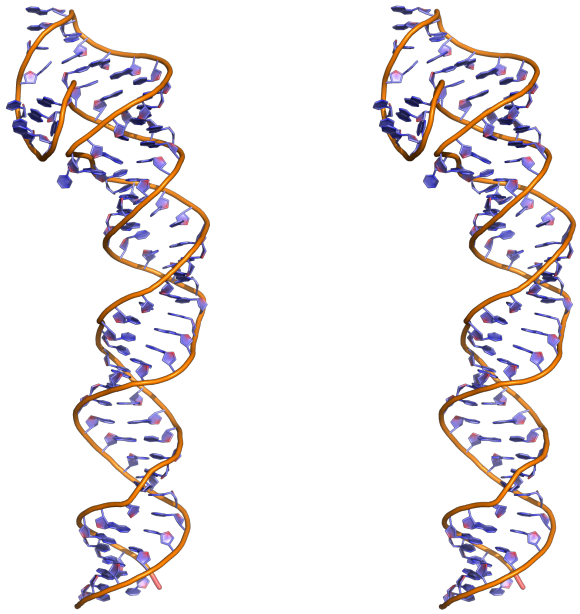


Figure 6.10: **MC-Sym model 61 for pri-miR-125a.** The model is presented in crosseye setero 3D view. Model 61 led to the highest classification performance for the hard benchmark according to both AU-PR and AU-ROC. We select this model for pri-miR-125a for all further analyses presented in this chapter.

Figure 6.10 shows MC-Sym model 61, which we select as the most biologically relevant model according to our results on the hard benchmark. Strikingly, the 5′ end of the Y-shaped loop is in close proximity with the 2-nt bulge, hinting at this interaction potentially playing a role in the structural dynamics of pri-miR-125a. While the backbone might be a bit too close in this region and lead to energetic frustration, let us remind the reader that ENCoM's all-atom sensitivity comes from a constrained Voronoi procedure which is robust to slight inaccuracies in the input structure. Indeed, the same procedure is used in the FlexAID docking software, of which one advantage is robust performance in the case of non-native docking [178]. Thus, while further relaxation of model 61 might prove interesting, it is not required for our analyses performed with ENCoM.

We find striking that the optimal β value is relatively low, corresponding to a high contribution from higher-frequency normal modes and thus greater importance of local dynamics in the EntroSigs. Since the hard benchmark was built by withholding the central base pair from every mutated box, it makes sense that local dynamics are favored: the model can approximate the effect of central mutations by how they affect the dynamics of the top and bottom base pair of the same box. Since we are

also interested in capturing long-range effects of pri-miR-125a variants, we decided to further explore how performance responds to varying β values for the mutated boxes benchmarks presented next.

6.2.2 *Mutated boxes benchmarks*

As described in Section 6.1.4, we constituted 8 train-test sets pairs by withholding all mutations performed on one of the first 8 boxes as the testing set and using the rest as the training set.

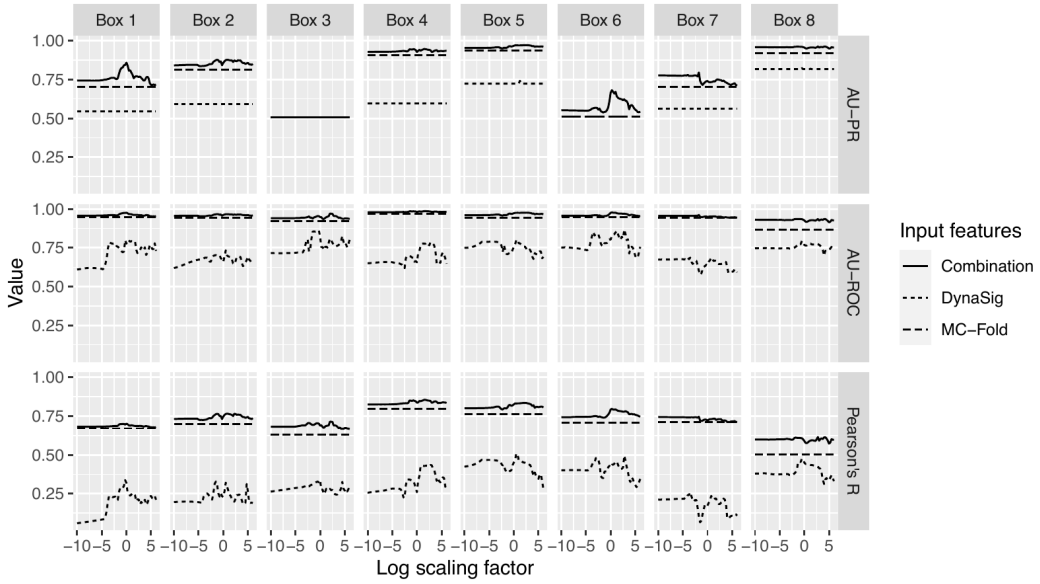


Figure 6.11: **Performance of LASSO models on the 8 boxes benchmarks.** Three performance metrics are detailed across all 8 boxes benchmarks: AU-PR, AU-ROC and Pearson's R linear correlation. The AU-PR and AU-ROC are computed on the classification problem while Pearson's R is computed on the full predictions, amibuous variants (maturation efficiencies between 0.5 and 0.8) included. The performances are shown for MC-Fold alone, for the DynaSigs alone and for the combination of both. The best performance across all regularization strengths tested is given.

Figure 6.11 gives three performance metrics, namely AU-PR, AU-ROC and the Pearson correlation, for the 8 boxes benchmarks. The performance from the best λ regularization strength is shown across the whole range of β values for MC-Fold alone, the DynaSigs alone and the combination of the two. Strikingly, MC-Fold alone performs better than or on par with the DynaSigs across the whole 8 boxes and on the three metrics. However, the combination is also always at least as good as MC-Fold, which is unsurprising since the LASSO model will drive the predictors which are uncorrelated to the outcome to zero at high enough regularization strenght. Nonetheless, significant improvements are made, especially for β values

in the vicinity of 1. Moreover, the improvements seem more consistent across the 8 boxes in the case of the Pearson correlation.

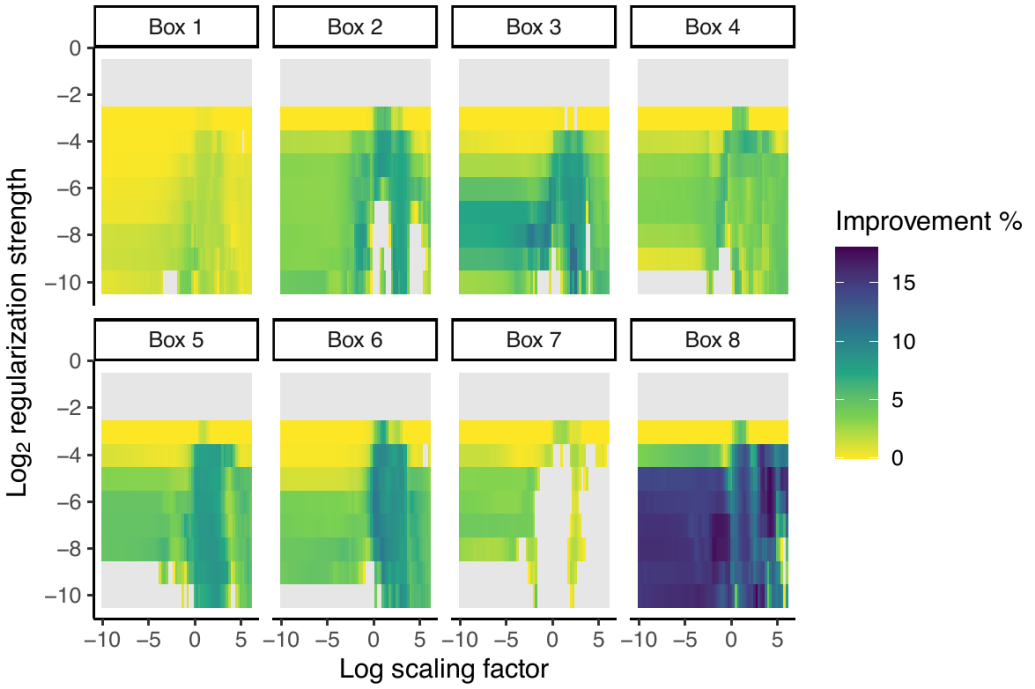


Figure 6.12: **Pearson correlation improvement by the MC-Fold + DynaSig combination on the 8 boxes benchmarks.** The percentage of improvement in Pearson's R correlation by adding the DynaSig to MC-Fold alone is shown for each box and for every combination of λ regularization strength and β thermodynamic scaling factor. Combinations of parameters leading to improvements under 0% are shown in gray.

We were interested in finding β scaling factors which led to optimal performance across the 8 boxes benchmarks since these benchmarks test the ability of the DynaSigs to capture fluctuations happening away from the site of mutation. Indeed, since the model is trained on mutations happening only in the 7 other boxes, it will mostly evaluate the impact of a variant in the testing set through its effect on distant positions. We restricted further analysis to the Pearson correlation since it is computed on the whole set of predictions and leads to improvements which are correlated with the AU-PR improvements, the more stringent of the two classification metrics and also the least sensitive to class imbalances. Figure 6.12 shows the percentage of Pearson correlation improvement across the whole range of λ regularization strengths and β scaling factors tested. The biggest improvement happens for box 8, which also corresponds to the box on which MC-Fold enthalpy alone performs the worse (see Figure 6.11) This finding further illustrates the complementarity of the two methods, with the DynaSig rescuing performance

in the cases where the enthalpy of folding alone cannot explain the difference in maturation efficiency.

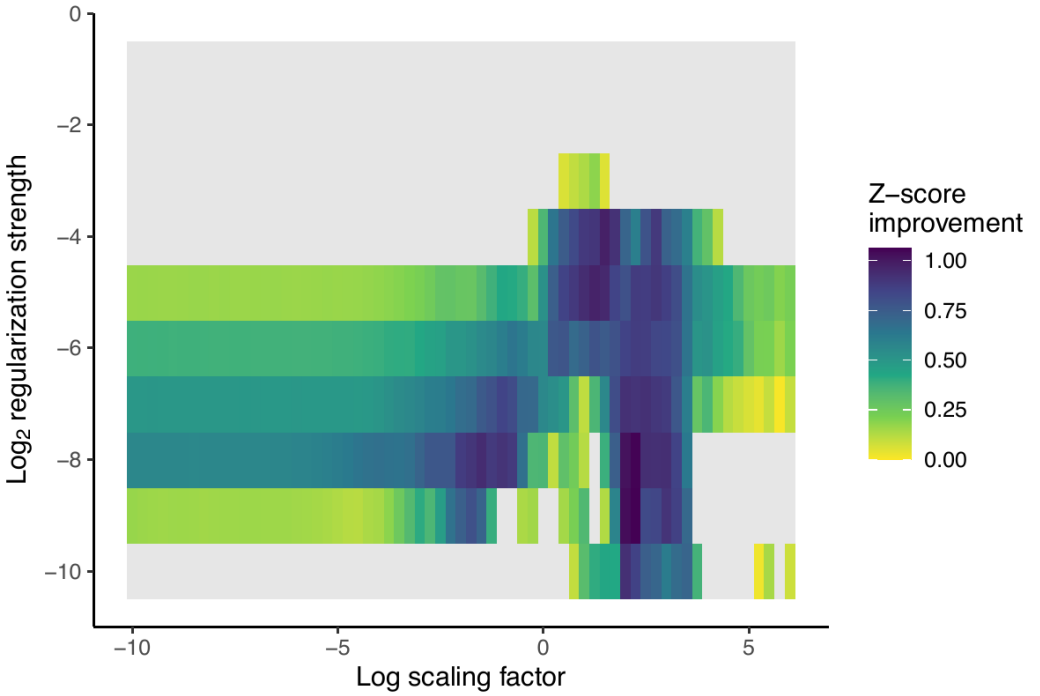


Figure 6.13: **Average Z-score Pearson correlation improvement by the MC-Fold + DynaSig combination on the 8 boxes benchmarks.** The average Z-score of the improvement in Pearson's R correlation by adding the DynaSig to MC-Fold alone is shown for every combination of λ regularization strength and β thermodynamic scaling factor. Combinations of parameters leading to average Z-scores under 0 are shown in gray.

In order to investigate the trends across the 8 boxes benchmarks, we computed the Pearson correlation improvement Z-score for the addition of the DynaSig to MC-Fold alone, for each individual box. Figure 6.13 shows the average improvement Z-score across all 8 boxes for all the combinations of β and λ. Three clusters of good performance strike us as very interesting:

1. The two highest average Z-scores of 1.06, at $\beta = e^{2.25}$ and $\lambda = 2^{-9}$ or 2^{-8}.

2. The cluster of good performance at high regularization strength, centered on $\beta = e^{1.25}$ and $\lambda = 2^{-5}$. The best average improvement Z-score from this cluster is 0.98.

3. The cluster of good performance at low scaling factor, centered of $\beta = e^{-1.5}$ and $\lambda = 2^{-8}$. The best average improvement Z-score from this cluster is 0.94.

In order to investigate what features the LASSO models that lead to these good performances, we trained a LASSO model on all pri-miR-125a variants from our dataset for each of the three clusters. We took the parameters leading to the best Z-score sum in the case of the last cluster, and the two best parameter sets for the first and second clusters in order to investigate how the different regularization strengths affect the LASSO coefficients. Table 6.2 lists these 5 parameter sets.

Table 6.2: **Optimal parameters for the 8 boxes benchmark.**

Cluster	$\log\beta$	$\log_2\lambda$	Average improvement Z-score
1A	2.25	-9	1.061
1B	2.25	-8	1.056
2A	1.50	-4	0.978
2B	1.25	-5	0.972
3	-1.50	-8	0.939

Figure 6.14 illustrates the LASSO coefficients obtained for every DynaSig position when training the model on the whole dataset of pri-miR125a sequence variants, with every of the interesting β and λ parameter combinations outlined avobe. The most prominent feature is the obvious shrinkage of the coefficients at higher regularization strengths in the case of cluster 2, which corresponds to β values around $e^{1.5}$. Another striking feature is the learning of coefficients dominantly on nucleobase beads in the case of cluster 1, corresponding to the highest scaling factor of $e^{2.25}$, while in the case of cluster 3, which corresponds to the lowest scaling factor of $e^{-1.5}$, the bead types selected are predominantly phosphate and sugar. This observation might lead to the conclusion that the parameter combinations from cluster 3 result in better generalization beyond sequence, since the mutations happen at the nucleobase and thus the model trained with these parameters is capturing fluctuation patterns further away from the mutations. Moreover, cluster 3 is also the only combination of parameters in which non-zero coefficients appear far from the 8 mutated boxes, in the apical loop. So while the motions at this lower scaling factor tend to be slightly more localized in nature, it seems that the model is nonetheless learning the long-range impact of these motions. Looking back at Figure 6.3B, this scaling factor of $e^{-1.5}$ leads to around 25% of the slowest modes contributing 90% of the entropy. Very local motions are thus not yet a significant part of the Entropic Signatures, as they happen mostly in the last third of the modes.

6.2.3 *5-fold cross-validation*

So far, we have focused on pairs of testing and training sets without sequence redundance between them, and we have restricted our analysis to LASSO regression

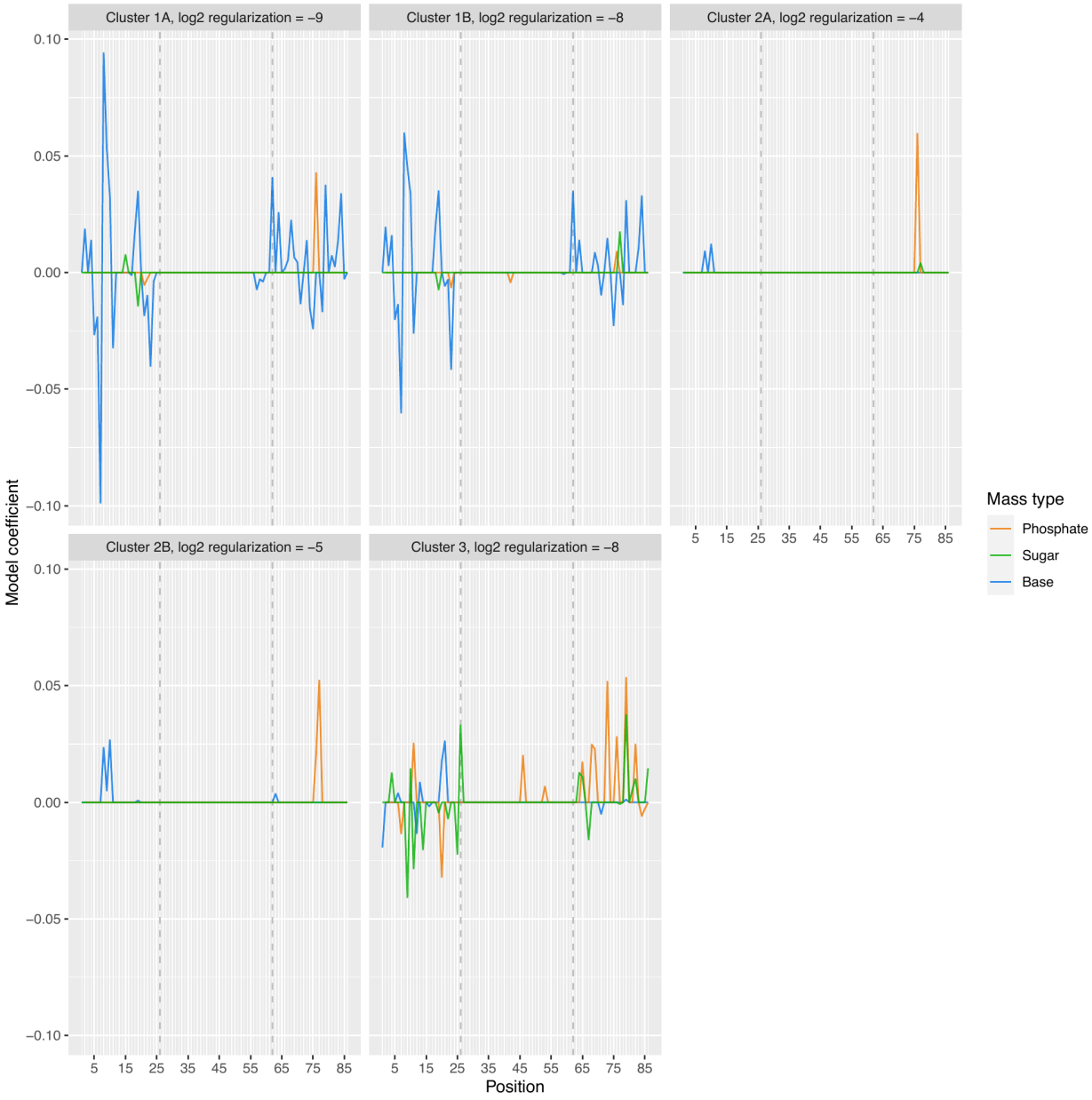


Figure 6.14: **LASSO coefficients for models trained on the whole dataset.** The LASSO coefficients for the DynaSig are given for five parameter combinations corresponding to the three highest improvement parameter clusters (see Table 6.2 for the detailed values). Cluster 1 corresponds to high scaling factor, cluster 2 to mid scaling factor and cluster 3 to low scaling factor. The models are trained on the whole dataset of 26 960 pri-miR-125a sequence variants.

in the search of both the most biologically releval MC-Sym model and the optimal values for the β scaling factor. However, an advantage of ENCoM is its sequence sensitivity, which means it can capture both sequence and dynamical patterns at

the same time. Indeed, the sequence variations that are captured by ENCoM lead to DynaSig patterns, which the ML models can learn. In order to test the gain of DynaSigs over sequence alone, and to explore if performance gains are achievable with multilayer perceptrons (MLP), we perform a carefully constructed 5-fold cross-validation across the whole dataset of 29 960 sequence variants. Each of the 5 test sets is constructed by sampling 20% of the variants present at each mutated box, without sampling the same variant in two different test sets. Sampling by mutated box ensures homogeneity in the proportion of processed and unprocessed variants, as the different boxes have wildly different tolerances for mutations (see Table 6.1).

To answer the question of whether the EntroSigs offer a gain of performance over sequence alone when sequence information can be useful in the predictions, we contruct sequence vectors as detailed in Section 6.1.5.1.

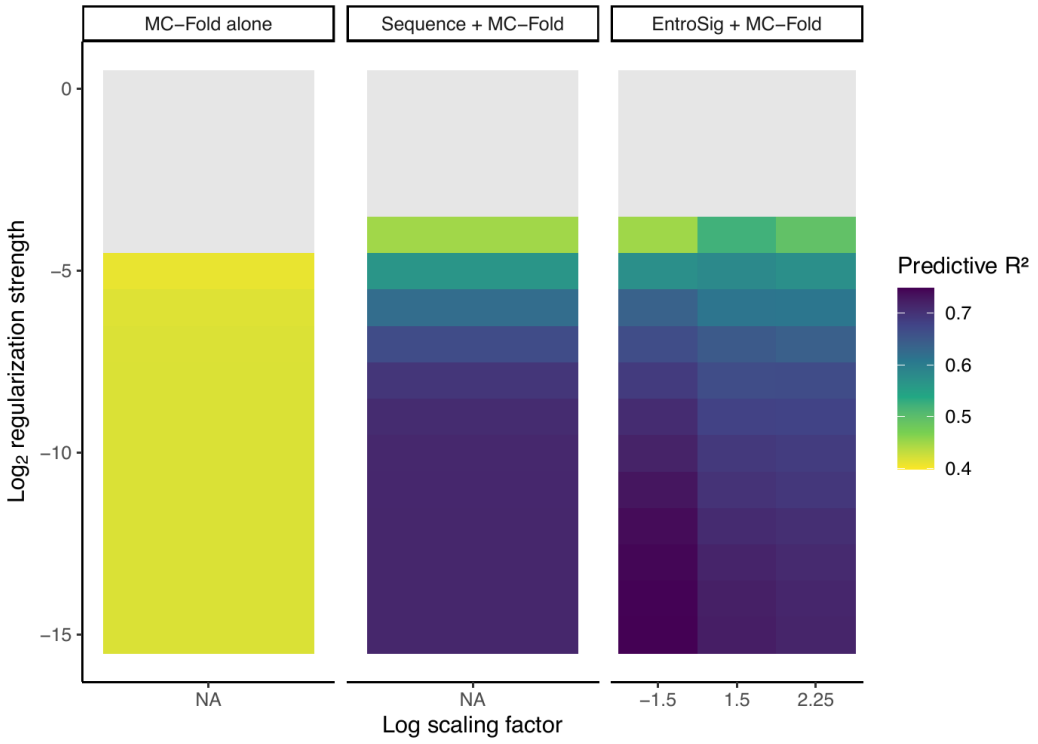


Figure 6.15: **Predictive R^2 for LASSO models on the 5-fold cross-validation.** The average predictive coefficient of determination (R^2) is given for LASSO models trained with the MC-Fold enthalpy alone, and its combination with either a sequence vector or the EntroSigs at the three identified optimal scaling factors. The models reach their highest performance at the lowest regularization strength of 2^{-15}: 0.42 for MC-Fold enthalpy, 0.71 for the combination with sequence vector, 0.72 for the EntroSig combinations with $\beta = e^{2.25}$ and $beta = e^{1.5}$, and 0.75 for the EntroSig + MC-Fold combination with $beta = e^{-1.5}$.

Figure 6.15 illustrates the performance of the different LASSO models in terms of average predictive coefficient of determination (R^2) over the 5-fold cross-validation. There is high complementarity between the MC-Fold enthalpy and either the EntroSigs or sequence vectors. Strikingly, the MC-Fold + EntroSig combination with the lowest thermodynamic scaling factor outperforms other models for five lowest regularization strengths tested.

6.2.3.1 *MLP architecture*

Another question we want to investigate is whether the linear independence between the EntroSig positions that the LASSO model assumes is hurting the predictions. For instance, one could imagine that complex relationships exist between the flexibility of different positions in the signal that the Microprocessor recognizes to cleave pri-miRs. To investigate this, we trained multilayer perceptrons (MLPs) on the same 5-fold cross-validation dataset. As outlined in Section 3.2.2, MLPs can capture complex relationships between the input variables and are thus a good choice to answer the question of how much interplay between flexibility at different positions affects the maturation efficiency of pri-miR-125a variants.

As discussed in Section 6.1.6, we tested MLP architectures having one or two hidden layers and up to 60 neurons per layer, resctricting the combinations which led to a number of free parameters exceeding 75% of the training set degrees of freedom.

Figure 6.16 gives the performance of MLPs trained with architectures of 1 or 2 hidden layers, with equal sizes of 2, 5, 10, 20, 30, 40, 50 or 60 hidden neurons. Surprisingly, the best performance is attained with the combination of MC-Fold and the sequence vector, for a big range of architectures. However, when looking at the performance as a function of the number of free parameters in Figure 6.16B, we notice that the EntroSig combination with the two lowest thermodynamic scaling factors achieve good complementarity with the MC-Fold enthalpy even at the lowest numbers of free parameters.

This fact led us to hypothesize that the extensive mutagenesis behind the pri-miR-125a maturation dataset might generate enough sequence redundancy so that the MLP can learn "by heart" some sequence features: for instance, all given quartet of nucleotides occur 16 times in the full dataset, corresponding to the other 16 possibilities out of a given 6 nucleotide box. To assert this hypothesis and still keep sequence redundancy between the testing and training sets in order to answer our initial questions of whether the ENCoM EntroSigs capture something beyond sequence, we constructed a final train-test pair by taking all variants affecting at most 2 positions as the training set, and all other variants affecting 3 positions or more as the testing set. This ensures that there can be no "memorizing" of specific quartets and triads of mutations by the MLP with the sequence vectors as input. We

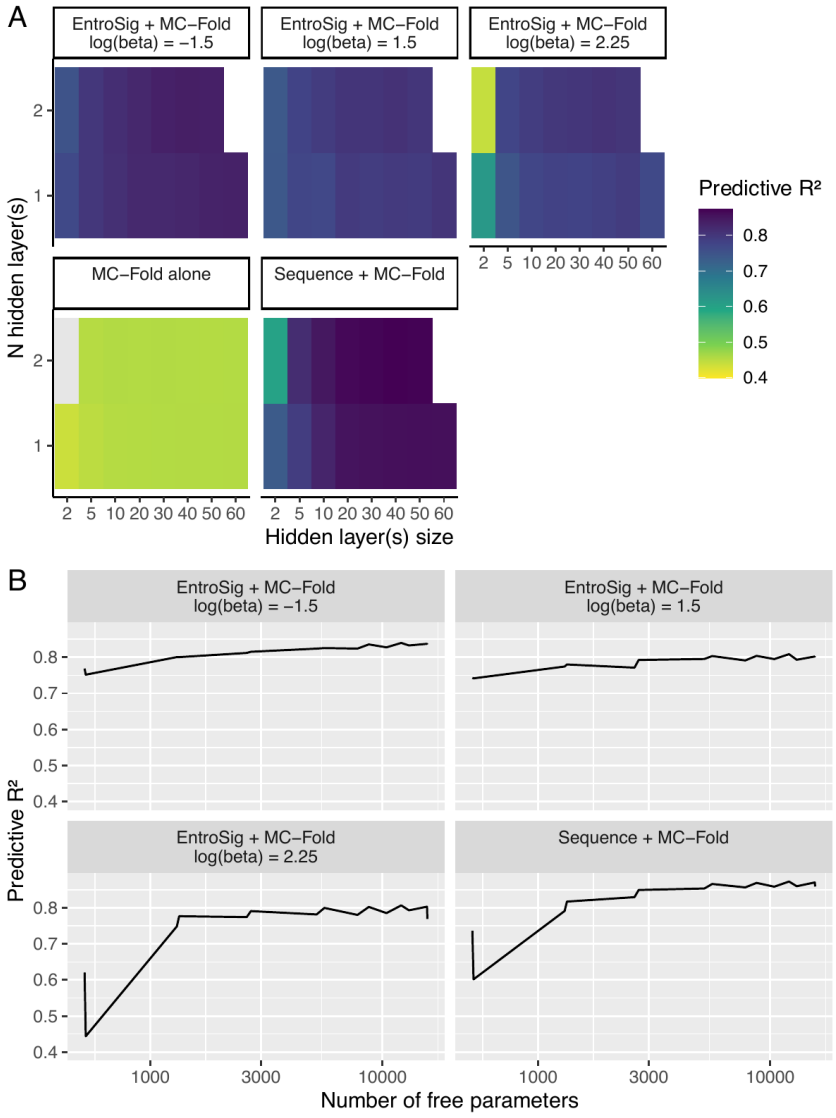


Figure 6.16: **Predictive R^2 for MLP models on the 5-fold cross-validation.** A) The average predictive coefficient of determination (R^2) is given for MLP models trained with same combinations of input variables as in Figure 6.15. The highest performances reached are: 0.87 for the MC-Fold + sequence combination, 0.84 for the MC-Fold + EntroSig at logβ -1.5, 0.81 at logβ 1.5 and 2.25 and 0.46 for the MC-Fold enthalpy alone. B) The performance is given for each combination as a function of the number of free parameters in the MLP.

call this final train-test pair the inverted dataset, because in contains 1094 variants in the training set and 25 866 variants in the testing set.

Figure 6.17 illustrates the performance of the reduced MLP architectures tested on the inverted dataset. These results seem to confirm our hypothesis as the

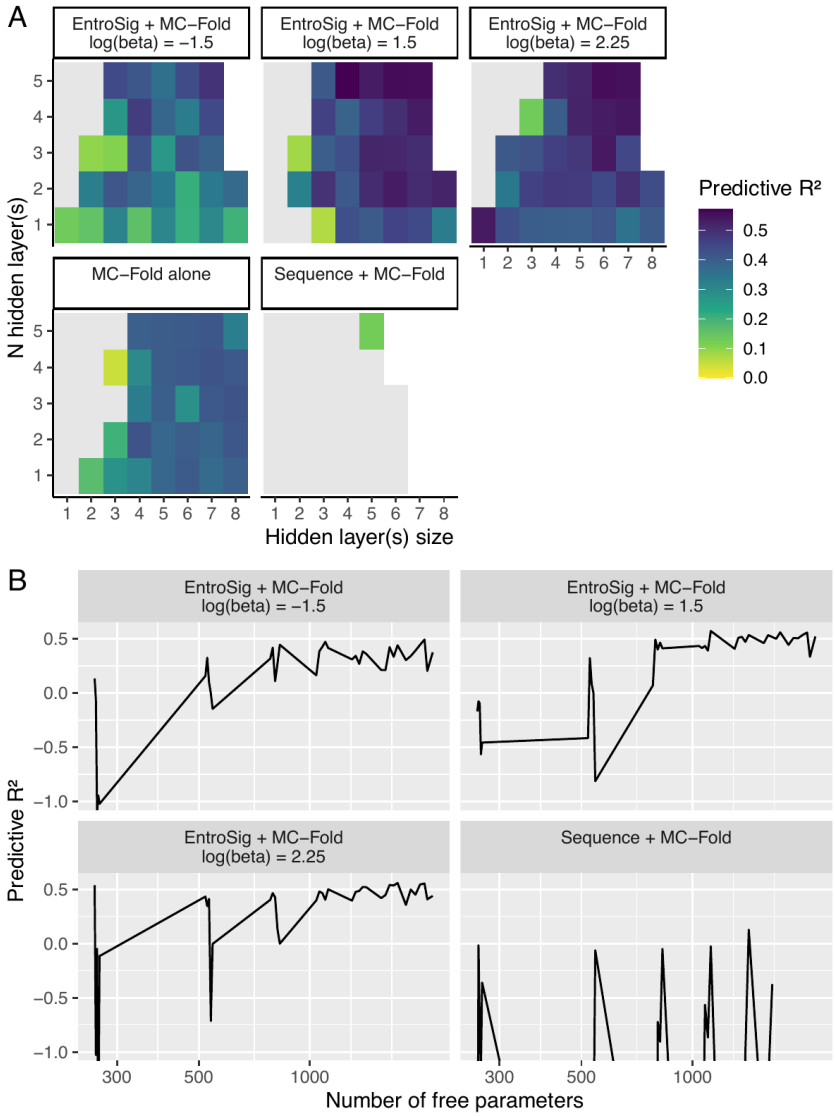


Figure 6.17: **Predictive R^2 for MLP models on the inverted dataset.** A)-B) Same as in Figure 6.16, on the inverted dataset. Since the models experience high variability on this dataset, we show the average performance from 5 reploicates. The highest performances reached are: 0.13 for the MC-Fold + sequence combination, 0.57 for the MC-Fold + EntroSig at logβ 1.5, 0.56 at logβ 2.25, 0.49 at logβ -1.5 and 0.42 for the MC-Fold enthalpy alone.

combination of MC-Fold enthalpy with the sequence vector performs very poorly, leading to no improvement compared to MC-Fold alone. Interestingly, the best scaling factor for this benchmark is $\beta = e^{1.5}$ followed closely by $\beta = e^{2.25}$, with $\beta = e^{-1.5}$ leading to poorer performance, which is contrary to what we observed on the 5-fold cross-validation but is in line with our observations in the 8 boxes benchmark.

Finally, we wanted to come back to the question of whether the linear independence assumption of LASSO regression is sufficient to capture dynamical patterns of pri-miR-125a processing. We thus again trained LASSO models using either sequence vectors + MC-Fold, MC-Fold enthalpy alone or each of the three retained scaling factors for the EntroSigs + MC-Fold combinations.

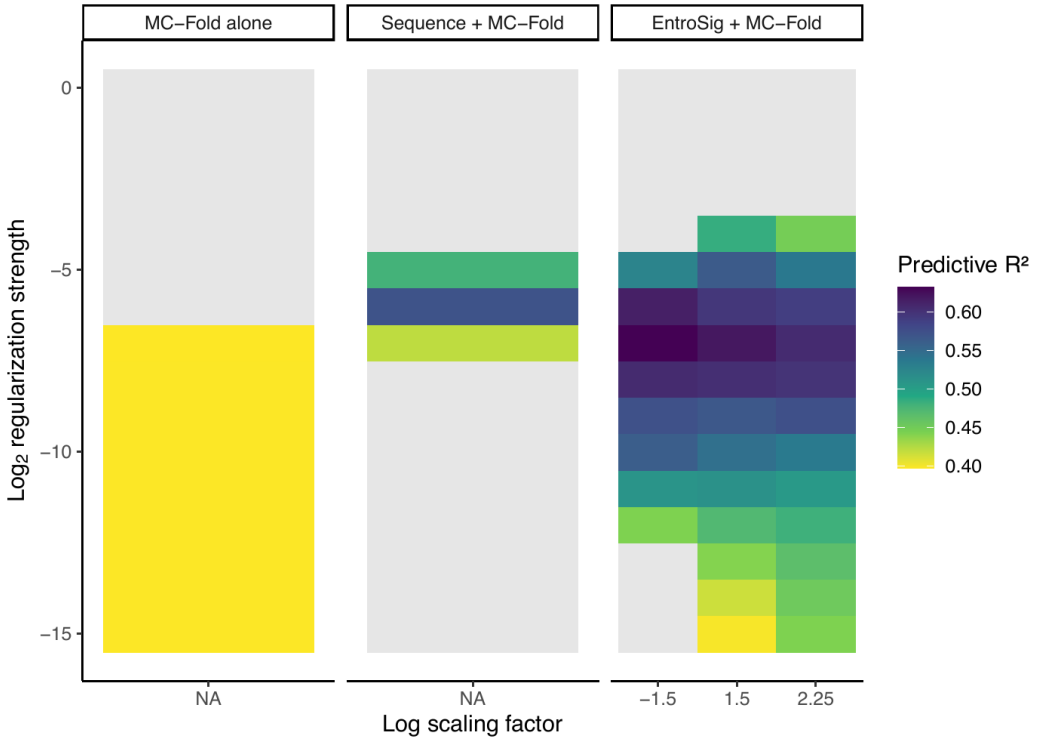


Figure 6.18: **Predictive R^2 for LASSO models on the inverted dataset.** The average predictive coefficient of determination (R^2) is given for LASSO models trained with the MC-Fold enthalpy alone, and its combination with either a sequence vector or the EntroSigs at the three identified optimal scaling factors. The MC-Fold + EntroSigs models all reach their highest performance at regularization strength 2^{-7}: 0.63 for logβ -1.5, 0.62 for logβ 1.5 and 0.61 for logβ 2.25. The combination with the sequence vector reaches 0.57 at $\lambda = 2^{-6}$ and MC-Fold enthalpy alone reaches 0.40 at all regularization strengths lower than 2^{-5}.

Figure 6.18 shows the LASSO performance on the inverted dataset. Very interestingly, the performance exceeds that of MLPs in all cases except for the MC-Fold enthalpy alone. Moreover, the best performance of all is attained by the EntroSigs + MC-Fold combination at $\beta = e^{-1.5}$, reversing the trend observed using MLPs with this same dataset. Compared to the combination with sequence vector, this scaling factor leads to a gain in performance of 0.06 R^2. While seemingly modest, this 0.06 gain still represents an additional 6% of variance explained, which is significant in the context of ultra-high-throughput *in silico* predictions. Moreover, the fact that the EntroSig model is learning both sequence and structural dynamics features at once

through the patterns apparent from the Entropic Signatures gives us confidence in its ability to generalize better in the context of sequence variants combining mutated positions across the whole structure.

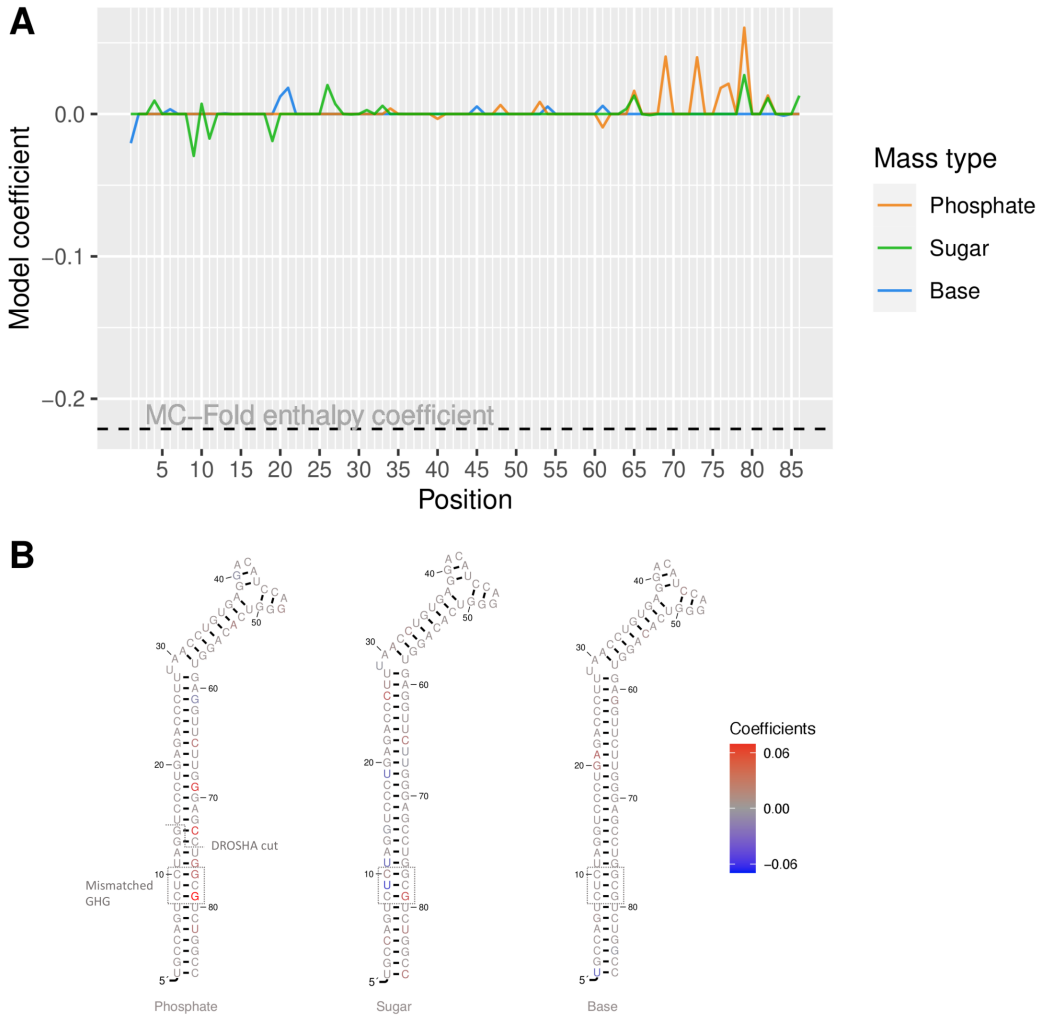


Figure 6.19: **Coefficients for the final LASSO model trained on all available data.** The model was trained on all 29 478 pri-miR-125a sequence variants adopting the WT 2D MFE as predicted by MC-Fold. $\beta = e^{-1.5}$ is the scaling factor used for the EntroSigs and $\lambda = 2^{-7}$ is the regularization strength used. A) The coefficients are shown graphically, with a dashed line for the MC-Fold enthalpy of folding coefficient, the highest in absolute value at -0.22. The EntroSig coefficients have a sum of 0.30 and an absolute sum of 0.50, for a softening bias of 60%. B) Same as A), mapped on the 2D MFE structure of priomiR-125a.

For the reasons outlined above, we train a final LASSO model on the complete dataset of pri-miR-125a sequence variants which adopt the WT 2D MFE, including the ones with mutations beyond the 8$^{\text{th}}$ box. We use parameters $\beta = e^{-1.5}$ and $\lambda = 2^{-7}$, as this combination led to the highest performance on the inverted dataset.

We view this dataset as the ultimate test of generalizability for the model, hence the selection of these parameters. Figure 6.19 illustrates the coefficients learned by this final model. As in all other models with sufficiently low regularization strength, the MC-Fold enthalpy coefficient is -0.22. It represents the largest coefficient by absolute value and thus, since all predictors are standardized, is the single most explanatory variable in the model. However, the sum of all Entropic Signature coefficients together is 0.30, and the sum of absolute EntroSig coefficients is 0.50. The sum of all positive coefficients is 0.40 and the sum of negative coefficients is -0.10. These values lead us to two interesting observations: first, the positive sum and predominance of positive coefficients mean that higher vibrational entropy is favored by the model. Moreover, vibrational entropy is slightly more favored than low folding enthalpy: we can directly compare the coefficients since, again, the predictors are standardized. However, there is an obvious tradeoff happening: lower folding enthalpy is achieved by the introduction of more rigid base pairs by either swapping a noncanonical base pair for a canonical, or an AU base pair for a GC. This rigidification is captured by ENCoM's all-atom sensitivity [157] and leads to lower vibrational entropy, which on average is detrimental to predicted maturation efficiency in the model.

On panel B, the coefficients are mapped on the pri-miR-125a MFE 2D structure, by bead type. The mismatched GHG motif and the DROSH cut site are shown on the structures. Interestingly, the highest LASSO coefficients happen on the phosphate bead 79, part of the mismatched GHG. However, it seems that more flexibility at the GC base pair is learned as favorable by the model, whereas the model proposed by Fang & Bartel states that the two GC base pairs above and below the noncanonical UC base pair are important for proper maturation [13].

6.2.4 *Ultra-high-throughput maturation efficiency predictions*

We submitted 540 million random pri-miR-125a sequence variants to 2D structure prediction with MC-Flashfold, of which almost 70 million adopted the WT 2D MFE. As described in Section 6.1.10.1, each random variant affected between 3 and 12 base pairs compared to the WT sequence, with an average of 7.5 affected base pairs. Thus, these sequences are more mutated that the Fang & Bartel sequences, which have at most 3 affected base pairs and at most 6 mutations. Figure 6.20 shows the distribution of folding enthalpy for both sets of variants. The two distributions are very similar, with the random variants shifted slightly towards higher folding enthalpy but also spanning a wider range, thus sampling bigger proportions of both high and low folding enthalpy variants. This was expected as the average number of affected base pairs is high for the random variants. Since most base pairs are already canonical base pairs in the WT sequence, changing more of them on average leads to higher folding enthalpy. However, having more affected base

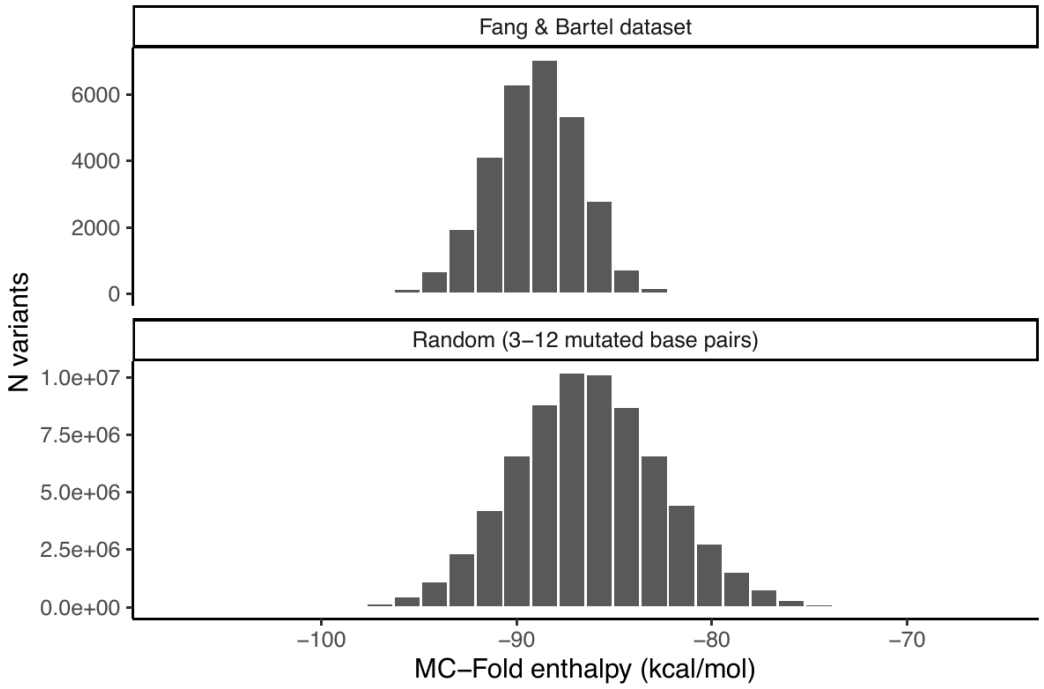


Figure 6.20: **MC-Fold enthalpy for experimental dataset and random variants with WT 2D MFE.** The Fang & Bartel variants used are the 29 478 with WT 2D MFE. A set of 540 million random variants was generated, with a uniform random number of affected basepairs between 3 and 12. Of these, the folding enthalpy is shown for the 69 817 456 which adopt the WT 2D MFE.

pairs also increases the potential to sample very low energies and very high folding energies, thus the wider range.

Since the folding enthalpy coefficient is the largest by absolute value in the selected LASSO model, we wanted to investigate the model's predictive behavior across three ranges of folding enthalpies: low, medium and high. The selected cutoffs for the three categories are given in Table 6.3. They were selected so that all 10 million lowest folding energy random variants contitute the low category, the 10 million medium folding enthalpy are centered on the mean enthalpy from the Fang & Bartel dataset (-89 kcal/mol) and also contain 10 million variants, and the high folding enthalpy contains the 10 million variants with highest folding enthalpy, while staying at or below -80kcal/mol. At higher energies, the amount of noncanonical base pairs introduced mean that the probability of the variant forming a defined structure becomes very low.

Figure 6.21 illustrates the distribution of predicted maturation efficiencies for the 30 million random sequence variants, along with the experimentally measured maturation efficiencies from the same folding enthalpy categories. We can observe

Table 6.3: **Cutoffs for the three folding enthalpy categories.**

Category	Low cutoff (kcal/mol)	High cutoff (kcal/mol)	N random variants
Low ΔG	$-\infty$	-90.31	10 015 077
Mid ΔG	-89.87	-88.16	10 002 397
High ΔG	-83.08	-80.00	10 012 741

a clear linear trend between maturation efficiency and folding free energy in both cases, which is expected in both cases. Interestingly, the random sequence variants do not lead to much higher predicted maturation efficiencies than the highest experimentally measured efficiencies, reinforcing the notion that our LASSO model is capturing a specific dynamical pattern necessary for pri-miR maturation, which random sequences have low probability of finding.

Figure 6.22 shows the superimposition of density distributions for the experimentally measured and predicted random sequence variants, across the three folding enthalpy categories. We can observe a very good fit of the predicted and measured values in the case of the two lowest folding enthalpy categories. For the high folding enthalpies, the LASSO model predicts a lot of negative maturation efficiencies, which cannot be observed experimentally. However, the experimental data is very sparse in that region and most variants falling in that region had very poor maturation efficiencies.

6.2.5 *Optimized pri-miR-125a variants*

Using the simple asexual genetic algorithm described in Section 6.1.10.2, we searched for three types of pri-miR-125a variants:

1. Variants with the same folding enthalpy as the WT sequence (-92kcal/mol) but high predicted maturation efficiency (> 1.5)

2. Variants with WT folding enthalpy but low predicted maturation efficiency (< 0.5)

3. Variants with extremely high predicted maturation efficiency (> 2.0), restricted to folding enthalpy higher no lower than -105 kcal/mol (minimum observed from random variants).

Table 6.4 lists three variants obtained with our asexual GA, one for each of the categories we listed. We find fascinating that the two first variants presented, which have approximately the same folding enthalpy, have such divergent predicted maturation efficiencies. Moreover, it seems that it is much easier to find mutations

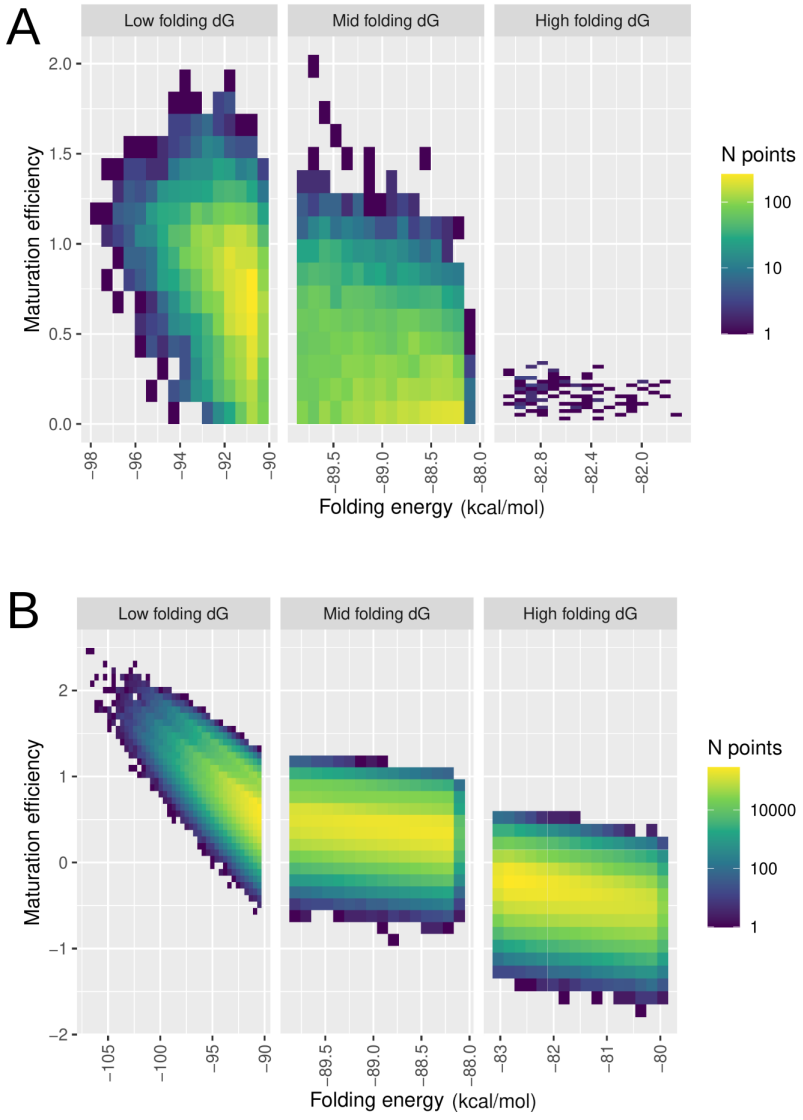


Figure 6.21: **Relationship between maturation efficiency and folding enthalpy for measured and predicted sequence variants.** A) The maturation efficiency measured for the Fang & Bartel variants, split in the three folding enthalpy categories defined in Table 6.3. B) Predicted maturation efficiencies for the same folding enthalpy categories, each containing over 10 million random sequence variants.

which kill maturation when starting from the WT sequence than to find ones that favor it further, as illustrated by the much higher number of mutations the algorithm has accumulated in the case of the positive example. This makes sense from an evolutionary perspective as the WT pri-miR-125a is under selective pressure to

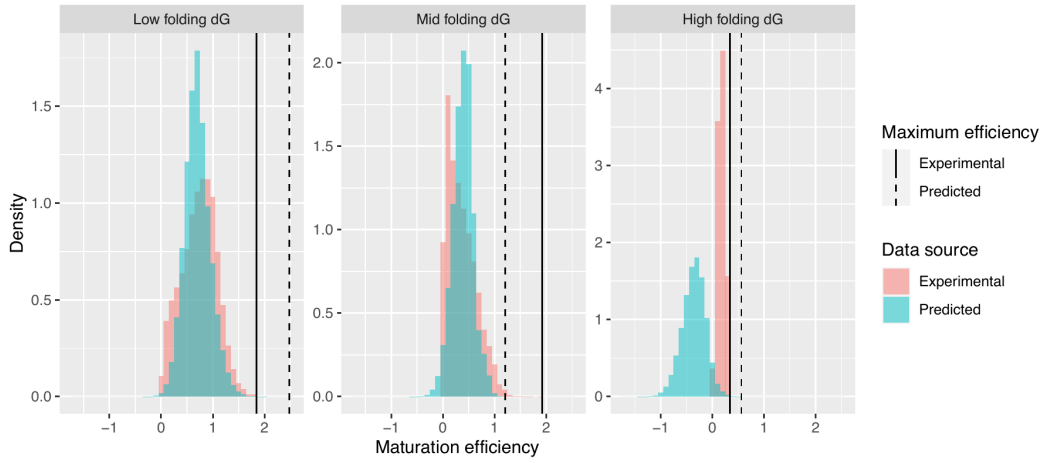


Figure 6.22: **Maturation efficiency distributions for measured and predicted sequence variants.** The distributions from measured maturation efficiencies and predicted efficiencies from random sequence variants are superimposed for the three folding enthalpy categories defined in Table 6.3.

Table 6.4: **Examples of optimized sequences for three categories.**

Category	Enthalpy (kcal/mol)	Predicted efficiency	Mutations
WT enthalpy, positive	-91.9	1.56	C3A, C4G, A5U, G6C, C10U, A23U, C24U, U28C, C42A, U44C, G48A, A60C, G61C, U64A, U67C, G68A, A71C, G72A, G76C, G77A, C81G, G83C, G84U, C86G
WT enthalpy, negative	-92.4	-0.13	C4G, U7A, C8G, U9G, U11A, U67G, G68A, C74A, G83C
Low enthalpy, positive	-104.9	2.56	C3G, C4G, A5U, G6C, C10U, U11C, A12U, G59A, A60U, U63G, U67C, G68A, A71G, U75A, G77A, C78U, C81G, U82A, G83C, G84C, C86A

be processed by the Microprocessor, thus its maturation efficiency is presumably already optimized to some extent.

6.3 DISCUSSION

Throughout this chapter, we have shown the ability of LASSO regression models to capture patterns of structural dynamics necessary for pri-miR-125a maturation as apparent from the high-throughput experimental dataset of Fang & Bartel [13].

We have carefully assembled benchmarks without sequence redundancy between the testing and training sets and shown that the ENCoM Entropic Signatures are not only highly complementary to the MC-Fold enthalpy of folding for these benchmarks, but are able to capture signal beyond sequence since the benchmarks do not contain any signal from sequence.

The hard benchmark has allowed us to select the most relevant 3D model of pri-miR-125a from the 67 generated by MC-Sym. When compared to the other models, the selected model (#61) striked us by the higher quality of its stem and the close proximity of the 2-nt bulge and the 5' end of the Y-shaped loop. We hypothesized that this proximity might play a role in miR biogenesis, and we note that the apical loop structure is missing in the only experimental structure of a pri-miR that we know of, the cryo-EM structure of miR-16-1 [175].

The MC-Sym model selection step highlighted the crucial importance of the Entropic Signature which we introduced as part of the present thesis. Indeed, statistically significant performance in terms of AU-ROC and AU-PR was found for scaling factors far from the one leading to equivalent DynaSig as the MSF ($e^{2.77}$). Moreover, it seems that performance on the hard benchmark is often happening at scaling factors lower than the MSF factor, hinting at an important role of normal modes which are basically excluded from the MSF DynaSig. Specifically, model #61 would not have been selected if it was not for the range of values tested, however it led to good performance on all subsequent tests performed.

We also explored the performance of both LASSO regression and multilayer perceptrons on 5-fold cross-validation on the whole dataset, this time allowing the models to learn both sequence and structural dynamics features by including sequence redundancy between training and testing sets. To our surprise, MLP models using sequence vectors combined with MC-Fold enthalpy outperformed MLPs using EntroSigs + MC-Fold. However, further analysis with the inverted dataset consisting of a minimal training set still covering all mutation possibilities revealed that this high performance was due to the MLPs' abilities to capture sequence patterns which appear across many variants.

The question we wanted to answer by training MLPs was whether the ability to model complex relationships between the EntroSigs positions would improve performance. While performance did improve in the 5-fold cross-validation, LASSO models outperformed MLPs in the inverted dataset. Perhaps after all, the number of relationships between many DynaSig positions is so vast that an MLP with sufficient power will mostly fail to generalize and overfit the relationships that are apparent within the dataset. Thus, we conclude that LASSO regression models are sufficient for our purposes as of now, and will not further explore MLPs in the following chapters.

The final LASSO model we trained, at $\beta = e^{-1.5}$ and $\lambda = 2^{-7}$ on the whole dataset of 29 960 pri-miR-125a sequence variants, captures flexibility at the mismatched GHG motif as the most important feature from the Entropic Signature. However, instead of being captured at the noncanonical UC base pair, the highest coefficient (favoring flexibility) is captured at the phosphate bead of G79. This pattern confirms that the model captures biologically relevant features, nevertheless it seems surprising that the necessary flexibility is not learned at the UC noncanonical base pair. One possible explanation is that the GC base pairs are needed for precise chemical recognition by the Microprocessor, while the role of the noncanonical base pair is to allow greater flexibility of these rigid base bairs in order for the recognition to happen. Indeed, it has been proposed that the mismatched GHG is the defining feature allowing the determination of the DROSHA cleavage site, with the two guanines being recognized by key residues of DROSHA [179]. Since the canonical base pairs are favored in our model by the MC-Fold enthalpy, perhaps it is enough to capture the detrimental effects of losing them, and the flexibility patterns learned reflect the effects of the noncanonical UC and longer-range effects making this region more flexible. It would be interesting to further investigate the effects of specific variants on the Entropic Signature to see whether such long-range effects are indeed captured by the model.

Our selected LASSO model allowed the exploration of predicted maturation efficiencies for over 30 million random pri-miR-125a sequence variants. The associated computational cost of these predictions is fairly low at around 3 seconds CPU per sequence, for a total cost of 2.85 core-years. This low computational cost is one of the main advantages of the ENCoM-DynaSig-ML pipeline, opening the door to ultra-high-throughput predictions of dynamics-function relationships in biomolecules. At such speed, some inaccuracies in the model can be tolerated, as the goal of such predictions is to enrich for some property before testing the top variants experimentally. In fact, while beyond the scope of the present thesis which focuses on the computational tools, we are currently testing some predictions of pri-miR-125a variants in the lab and expect to have the results in 2023. We are also interested in applying the same methodology outlined in this chapter to the other pri-miRs extensively mutated as part of the Fang & Bartel study, miR-16 and miR-30, to try and generalize our model further. Specifically, we are curious to see how alike the LASSO coefficients would be if we were to compare between the three pri-miRs.

Dallaire *et al.* had found that 2D structural dynamics play an important role in the miR biogenesis pathway by correlating distance between 2D conformational landscapes of 15 sequence variants and the WT sequence [14]. In this chapter, we confirm the important of structural dynamics for miR biogenesis at the three-dimensional level, on a dataset of over 26 thousand sequence variants. We have also found that the linear independence of the input variables made by LASSO

regression does not hurt the performance when it comes to models trained on the Entropic Signatures. While counterintuitive, this finding leads us to conclude that LASSO is enough for our purposes and is more generalizable to combinations of more mutated positions than what the model has seen during training, which is what we are most interested in predicting. Thus, subsequent chapters will focus on LASSO regression.

7

μ-OPIOID RECEPTOR ACTIVATION

The present chapter is dedicated to our second dynamics-function case study, the prediction of activation potential for μ-opioid receptor ligands. In this study, the variation in "sequence" is actually the binding of different ligands. Since ENCoM is sensitive to all-atom context, changes in the surface complementarity term β_{ij} result from different ligands or different poses of the same ligand, which are then captured by the Dynamical Signature. The main idea is then that agonists and antagonists lead to different patterns that the LASSO model can learn, since G protein-coupled receptors (GPCRs) occupy diverse conformational states which dictate their activity. As for the previous chapter, we refer the reader to Section 1.3.2 for the biological background concerning GPCR activation.

The chapter's organization will be the same as for the other two case studies chapters: we will start with methodology and follow with results and discussion. However, we will first outline the contributions made by Gabriel Tiago Galdino, as the present chapter is the only one in the thesis for which some of the data shown were not directly generated by the author.

7.1 CONTRIBUTIONS FROM GABRIEL TIAGO GALDINO

As mentioned, this chapter contains joint work between myself (Olivier) and my colleague Gabriel Tiago Galdino, a fellow PhD student from the Najmanovich Research Group. Gabriel's PhD project is dedicated to the study of how ligand binding affects GPCR dynamics. As I was developing the ENCoM-DynaSig-ML pipeline, we collaborated closely on applying the pipeline to the μ-opioid receptor, which is the subject of the present chapter. Gabriel graciously gave me permission to use data he generated, and I am thankful for fruitful discussions with him. The work presented here will soon be submitted to a scientific journal, in a slightly different form, with Gabriel and me sharing first authorship since our contributions are of roughly equal importance. Here are the specific contributions Gabriel made to what will be presented next:

1. Selection of the ligands from the ChEMBL database.
2. FlexAID docking experiments for all selected ligands.

3. Mapping of SYBYL atom types to ENCoM atom types and generation of the ENCoM configuration files for the ligands.

Gabriel and myself designed the docking experiments together. I computed the ENCoM DynaSigs, designed the validation experiments, trained all LASSO models, performed analyses and statistical simulations, generated all figures and wrote the text.

7.2 METHODOLOGY

The next subsections will detail the dataset generation, the docking experiments, the assignment of ligand atom types, the selection of thermodynamic scaling factors and the leave-one-out and 5-fold cross-validations performed.

7.2.1 *Selection of MOR ligands*

In order to select ligands with high-confidence experimental measures of μ-opioid receptor (MOR) activation, we first obtained the list of all compounds in ChEMBL [57] with some experimental measure of Emax for MOR (target ID CHEMBL233). This list contained 840 compounds with measures from 148 different ChEMBL biological assay IDs. While no single assay ID covered more than 21 ligands, a large portion of reported Emax were measured by [^{35}S]GTPγS assays, a reliable and almost direct measure of G protein activation [61], as discussed in Section 1.3.2. A baseline of activation has to be established for Emax measurements, and the most common baseline used was DAMGO, a very potent and selective MOR agonist [63, 180].

Thus, in order to maximize the number of ligands in our dataset while maintaining uniformity in the experimental measures, we selected all listed MOR ligands from ChEMBL which had a measurement of Emax relative to DAMGO from experimental assays using [^{35}S]GTPγS. This filtering yielded 198 unique ligands. We rejected ligands which generated errors during FlexAID preprocessing and ligands which did not lead to at least 10 poses with a negative docking energy, as outlined in the next subsection. This left us with 89 MOR ligands with both high-confidence experimental measurements and high-confidence docking results. Each ligand's interaction with MOR is captured by 10 unique docking poses, as outlined below, for a total of 890 data points.

7.2.2 *Docking experiments with FlexAID*

The structure we used for the MOR is the cystal structure solved by Huang *et al.* [50] (PDB code 5C1M), from which we removed all heteroatoms and kept only the MOR chain, removing the G protein. The reason for the removal of the G protein

is that we want to capture the process of receptor activation by the ligand, which happens before interaction with the G protein [53]. We used the FlexAID molecular docking software for the docking experiments, with published parameters [178]. For each ligand, 10 replicated docking experiments were conducted with 4000 generations of 4000 individuals for the genetic algorithm, keeping the top 50 poses from every experiment. After discarding duplicate poses, we kept the 10 poses with lowest docking energy for every ligand. As mentioned above, we discarded ligands which either caused preprocessing errors or did not yield 10 docking poses with negative energy. This restriction on negative docking energy was to ensure reliability of the docking poses. In addition, energetically frustrated docking poses could break the NMA assumption that the input structure is at equilibrium if they do not represent a local minimum in the energy landscape of the ligand-receptor interaction.

7.2.3 *Assignation of ENCoM atom types*

As outlined in Section 3.1, ENCoM uses a simplified atom typing system first described by Sobolev *et al.* [138]. As part of the NRGTEN package [31], it is now straightforward to extend ENCoM to include new residue types, including small molecules, provided an atom type is assigned to every atom in the new residue. The details of this procedure are part of the online NRGTEN guide and can be found at: nrgten.readthedocs.io/en/latest/custom_atypes.html. FlexAID uses 40 atom types corresponding to a subset of Sybyl atom types [181], assigned with the Open Babel software [182]. We thus started from the Sybyl atom types of the ligands to assign one of the 8 Sobolev atom types used by ENCoM to every atom automatically. There are ambiguities in the case of carbon, nitrogen and oxygen atom Sybyl classes, but we were able to resolve these by looking at an atom's neighbours. The mapping between Sybyl and Sobolev atom types for all elements observed in the 89 MOR ligands is given in the appendix, in Table A.9.

7.2.3.1 *Position of the beads*

We assigned a single bead for every ligand, regardless of its number of atoms, in order to maintain the same number of beads in the system for all ligands. The bead was located on the medoid atom of the ligand for each docking pose, which is the atom closest to the centroid coordinates. Since FlexAID considers ligand flexibility as part of the docking simulations, this center atom was not necessarily the same between poses of the same ligand, but was always closest to the ligand's center of mass, thus making the interactions with receptor beads more physically realistic.

7.2.4 *EntroSigs scaling factors*

Similarly as in Chapter 6, we wanted to identify an interesting range of thermodynamic scaling factors for the Entropic Signatures. We computed the EntroSigs on the complex of MOR with docked morphine, using the lowest energy docking pose.

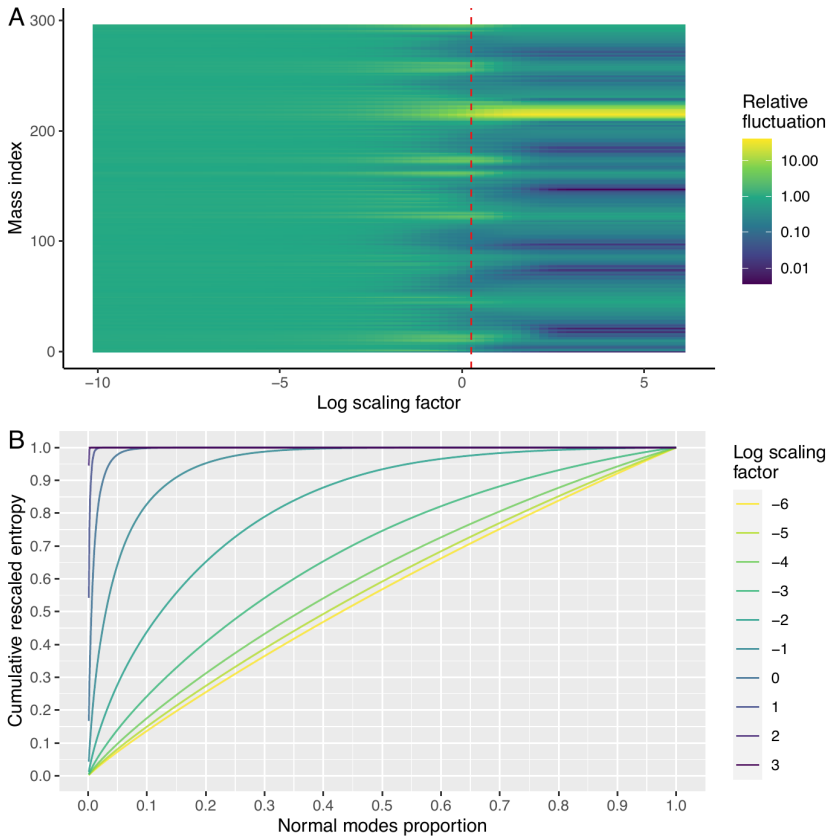


Figure 7.1: **Entropic Signatures and entropy proportions for the MOR-morphine complex across selected scaling factors.** A) Entropic Signatures across scaling factors ranging from e^{-10} to e^{6} in log increments of 0.25. B) The cumulative entropy across all normal modes is given for scaling factors across the selected range, rescaled to sum to 1.

Figure 7.1A illustrates the EntroSigs across a wide range of scaling factors, showing a stabilization of contributions from the first mode past $e^{2.5}$ and a dominance of high-frequency modes below e^{-5}. We thus chose to test values from e^{-6} to e^{3} in log increments of 0.25 for the present investigation of μ-opioid ligand-induced dynamics. Figure 7.1B shows the cumulative vibrational entropy proportions resulting from scaling factors across the selected range.

7.2.5 *Classification problem*

We classify ligands with measured Emax relative to DAMGO >= 50% as agonists, and ligands with Emax < 50% as antagonists. While this binary classification can appear too simple for capturing the intricacies of GPCR activation at different levels, let us remind that we have a limited number of 89 ligands in our dataset. Moreover, with these thresholds, there are 30 "antagonists" and 59 "agonists", therefore lowering the threshold would further imbalance the classes. Making the threshold higher does not make biological sense to us, as DAMGO illicits a strong activation response at the MOR. In fact, many ligands we classify as "antagonists" are actually partial agonists. However, we are satisfied with this classification as it can be used to test for the model's ability to enrich agonists which can generate strong responses. This ability would be useful in a virtual screening context to prioritize ligands with strong predicted Emax for experimental validation. In addition, we also measure the Pearson correlation between the predictions and real Emax values, so the simplified classification problem is not the only performance metric used to assess the models.

7.2.6 *Leave-one-out cross-validation*

In order to assess the capacity of LASSO models trained on MOR-ligand complexes to classify ligands never before seen by the model, we performed leave-one-out cross-validation. For each combination of thermodynamic scaling factor and regularization strength, we trained 89 LASSO models, removing the 10 docking poses for each ligand and training on the remaining 880 complexes. We then predicted the Emax for the 10 left out poses and combined predictions from 89 separate models in order to assess performance. This represents a very stringent test as it does not average out the predictions from every pose before assessing performance.

For the leave-one-out cross-validation, we test 16 regularization strengths across 37 scaling factors, for a total of 592 combinations. To assert the statistical significance of our classification results, we compute AU-PR and AU-ROC for 592 predictions from random noise, for 1000 replicates. Since the FlexAID docking score had no classification ability, we only use the Dynamical Signatures as predictor variables and thus do not test for improvement over a given baseline as we did in Chapter 6.

7.2.7 *5-fold cross-validation*

To assess the performance the model in a best-case scenario in which the training set samples a diversity of chemotypes, we performed 5-fold cross-validation using the scaling factor identified as leading to the best performance from the leave-on-out cross-validation. We split the dataset in 5 by randomly sampling 2 poses from each ligand for every testing set, without replacement. Sampling by ligand ensured

that all ligands are seen in equal proportions for each of the 5 training-testing set pairs. We train one LASSO model for each of the 16 regularization strengths, for each of the 5 train-test pairs. We report both testing performance, and pooled performance from combining the prediction of all models for a given regularization strength.

7.3 RESULTS

7.3.1 *Docking scores*

In the past chapter, we observed high complementarity between the ENCoM Entropic Signatures and the MC-Fold enthalpy of folding for the prediction of pri-miR-125a maturation efficiency. For the present study, the FlexAID docking score, which is outputted for every pose and corresponds to arbitrary energy units, with lower values representing more favorable poses. We thus investigated the prediction ability of the docking score for the class prediction problem across the 89 MOR ligands.

Figure 7.2 shows the docking score of the 10 poses for each of the 89 ligands, as individual boxplots sorted by their mean value. No apparent enrichment of either class at either exteme of the docking scores is apparent. In fact, the main apparent feature is that antagonists tend to be enriched at both ends.

Figure 7.3 shows the detailed predictive ability of the docking score alone. Panel B gives the classification in terms of AU-ROC and AU-PR compared to 1000 replicates of random predictions, confirming that the classification is not statistically better than random ($p = 0.41$ and $p = 0.94$ for AU-ROC and AU-PR, respectively). Thus, we will not include docking score in the predictor variables for the following analyses. We note that this absence of classification ability for the docking scores is not necessarily related to a lack of accuracy in the scoring function. The docking score can be interpreted as an approximation of the complex's free energy, which dictates the ligand's affinity for the receptor. However, affinity is not what determines the activation potential of a ligand, as an antagonist can have higher affinity for the receptor than an agonist and *vice versa*.

7.3.2 *Leave-one-out cross-validation*

For the leave-one-out cross-validation, we pool the predictions from all 89 LASSO models with the same β scaling factor for the EntroSigs and λ regularization strengths. We predict separate Emax values for each of the 10 poses of the withheld ligand.

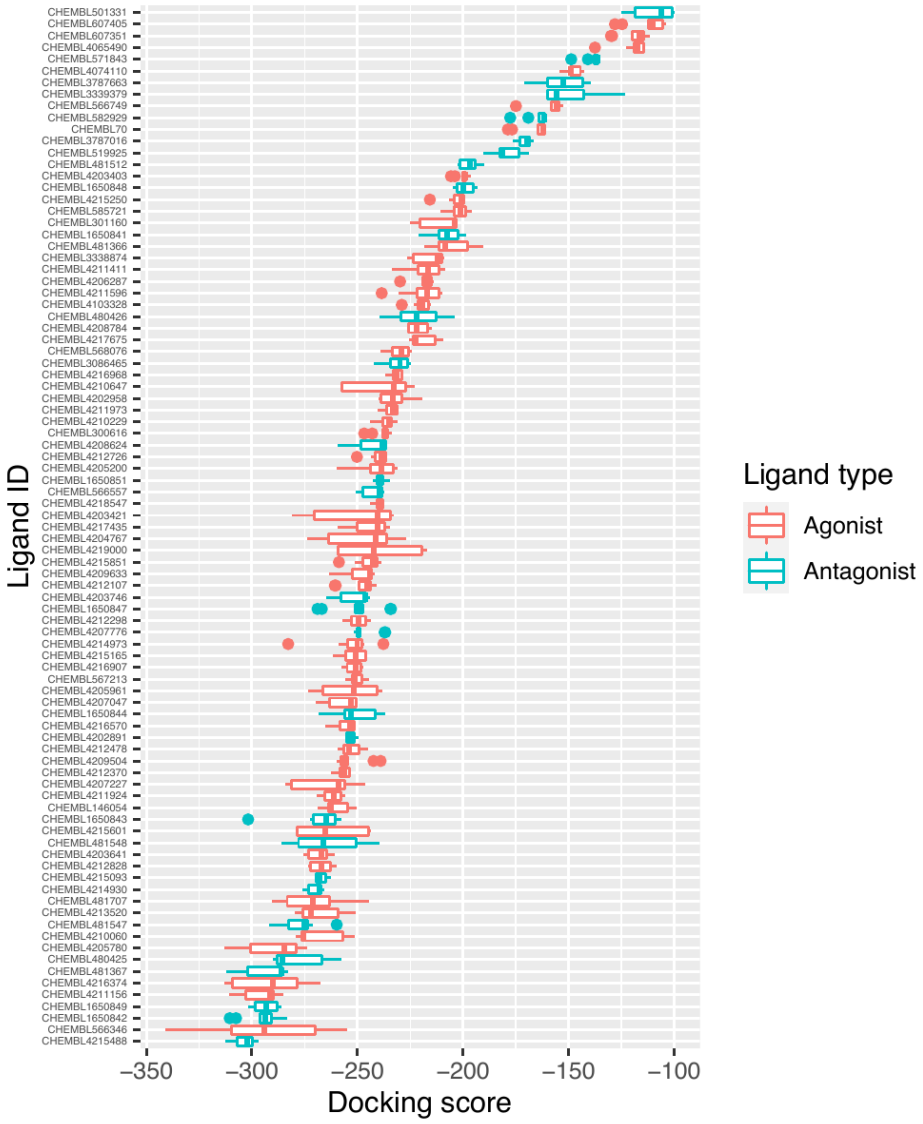


Figure 7.2: **FlexAID docking scores for the 10 selected poses from each MOR ligand.** The FlexAID docking scores for each of the 89 MOR ligands in the dataset are shown for the top 10 unique poses as boxplots, ordered by increasing mean docking score. The middle line represents the mean. The boxplots are colored blue for antagonists (Emax < 50%) and red for agonists (Emax >= 50%).

Figure 7.4 illustrates the pooled performances across three performance metrics: AU-ROC, AU-PR and Pearson correlation. The p-value for the Pearson correlation is also given, after correcting for the number of tests performed. For every scaling factor tested, we report the best performance attained from the 16 regularization strengths tested. All three metrics reach their highest values for $\beta = e^{-0.5}$ or $\beta = e^{-0.75}$.

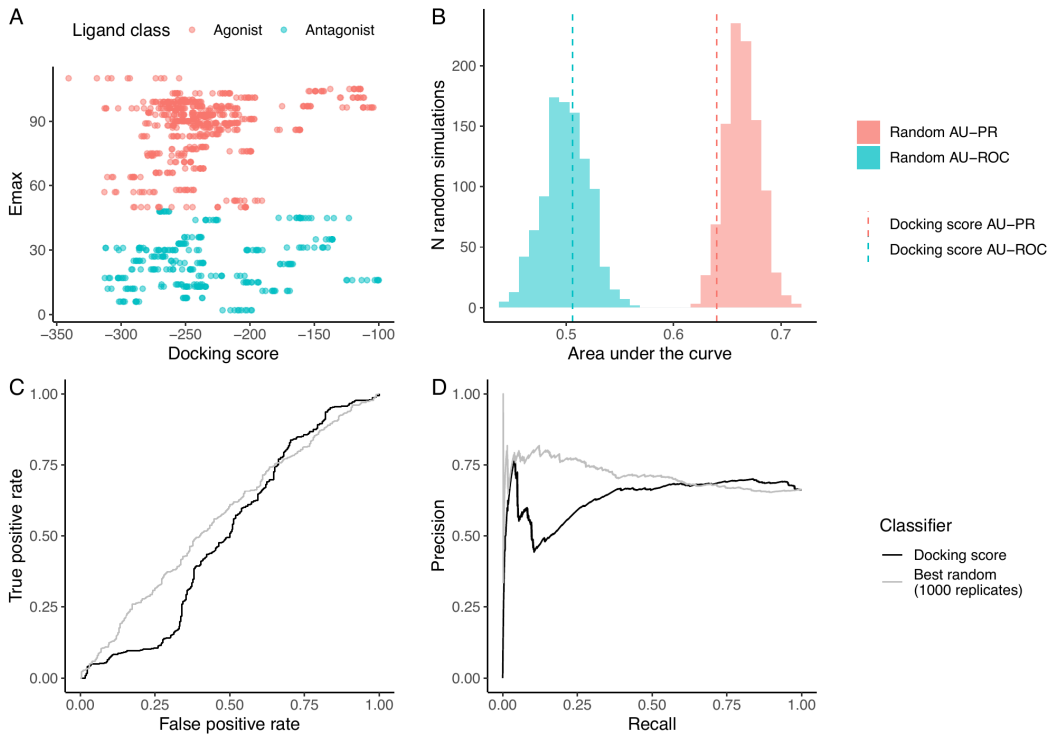


Figure 7.3: **Classification performance using the FlexAID docking score alone.** A) The Emax value for the 89 MOR ligands is shown as a function of the FlexAID docking score. Agonists are defined as having Emax >= 50%, and antagonists as having Emax < 50%. B) Simulated distributions showing area under the receiver operating characteristic (ROC) and precision-recall (PR) curves. The distributions result from 1000 replicates of using normally distributed noise as the predictor variable. The FlexAID docking score performance is shown with lines. In both cases, it does not lead to statistically significant classification of the ligands over random classification, with $p = 0.412$ for AU-ROC and $p = 0.938$ for AU-PR. C) ROC curve, with the best random curve in gray and the curve from the docking score in black. D) Same as C) for PR curve.

For the AU-PR, we observe an interesting upward trend towards higher scaling factors. In fact, one might be tempted to think that we should have tested higher values as the trend might continue. However, at the highest value tested, the proportion of total vibrational entropy contributed by the first normal mode is 0.9998, up from 0.9987 for the second highest value tested. Put another way, the proportional contribution to EntroSigs from modes other than the first is $2 * 10^{-4}$ for the highest β value tested, and $1.3 * 10^{-3}$ for the second highest. Thus, we are confident that the highest AU-PR obtained of 0.768 would not surpass the best AU-PR obtained, 0.775 at $\beta = e^{-0.5}$. Furthermore, the Pearson correlation suffers an abrupt drop in performance for scaling factors higher than $e^{1.5}$, confirming that these values are not beneficial.

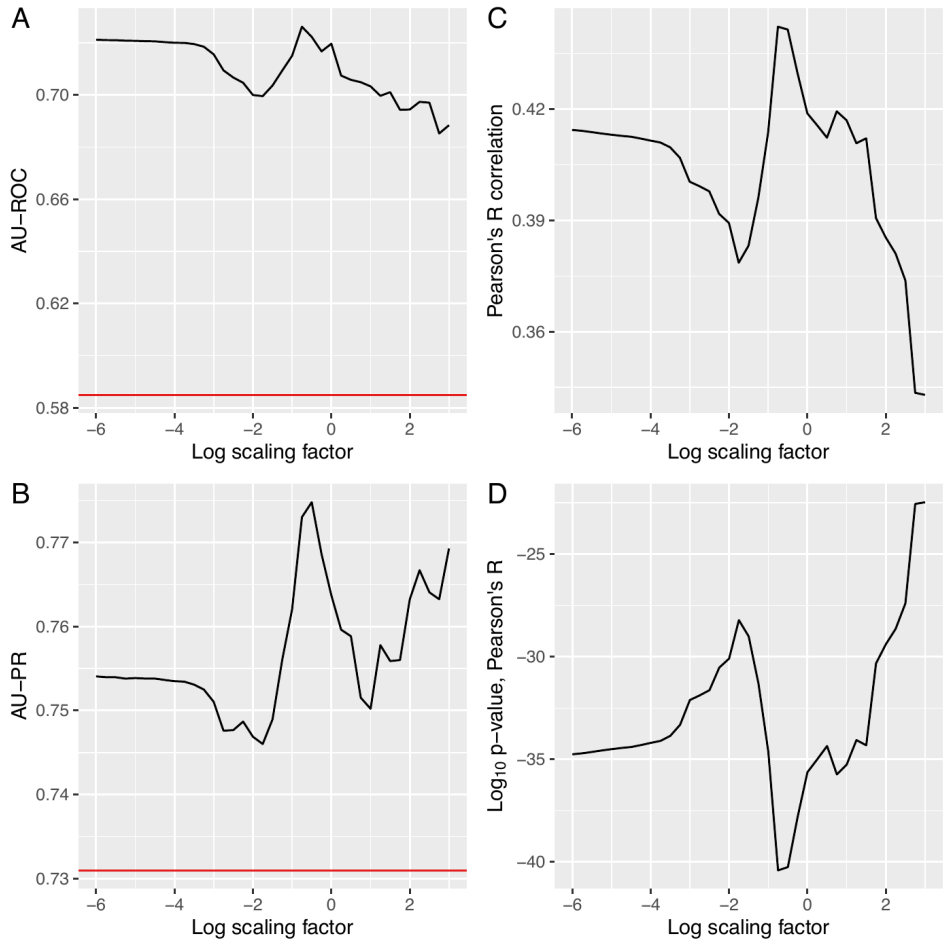


Figure 7.4: **Leave-one-out cross-validation performance metrics.** The best performance obtained across the 16 regularization strenghts tested for the LASSO regressions is shown as a function of the thermodynamic scaling factor. A) Area under the receiver operating characteristic curve (AU-ROC). The red line shows the threshold for statistically significant predictions, at the $p < 0.01$ level after correcting for the number of parameter combinations tested. B) Area under the precision-recall curve (AU-PR). C) Pearson's R linear correlation. D) Corrected p-value for the Pearson correlation observed in C), $\log_{10}$ transformed.

Figure 7.5 gives the detailed Pearson's R and AU-PR across all parameter combinations. Interestingly, it is apparent from panel B that the performance rise in AU-PR at higher scaling factors is happening at low regularization strengths. Since we are interested in relatively aggressive feature selection in order to both lead to better model generalizability and better biological interpretation of the coefficient, we confirm that a performance optimum has been reached in case of both Pearson's R and AU-PR. We did not present the detailed AU-ROC performance as it is less reliable when class imbalance exists, as already discussed in Section 4.4.2.2. The best parameters are very close for both Pearson's correlation and AU-PR, with AU-PR

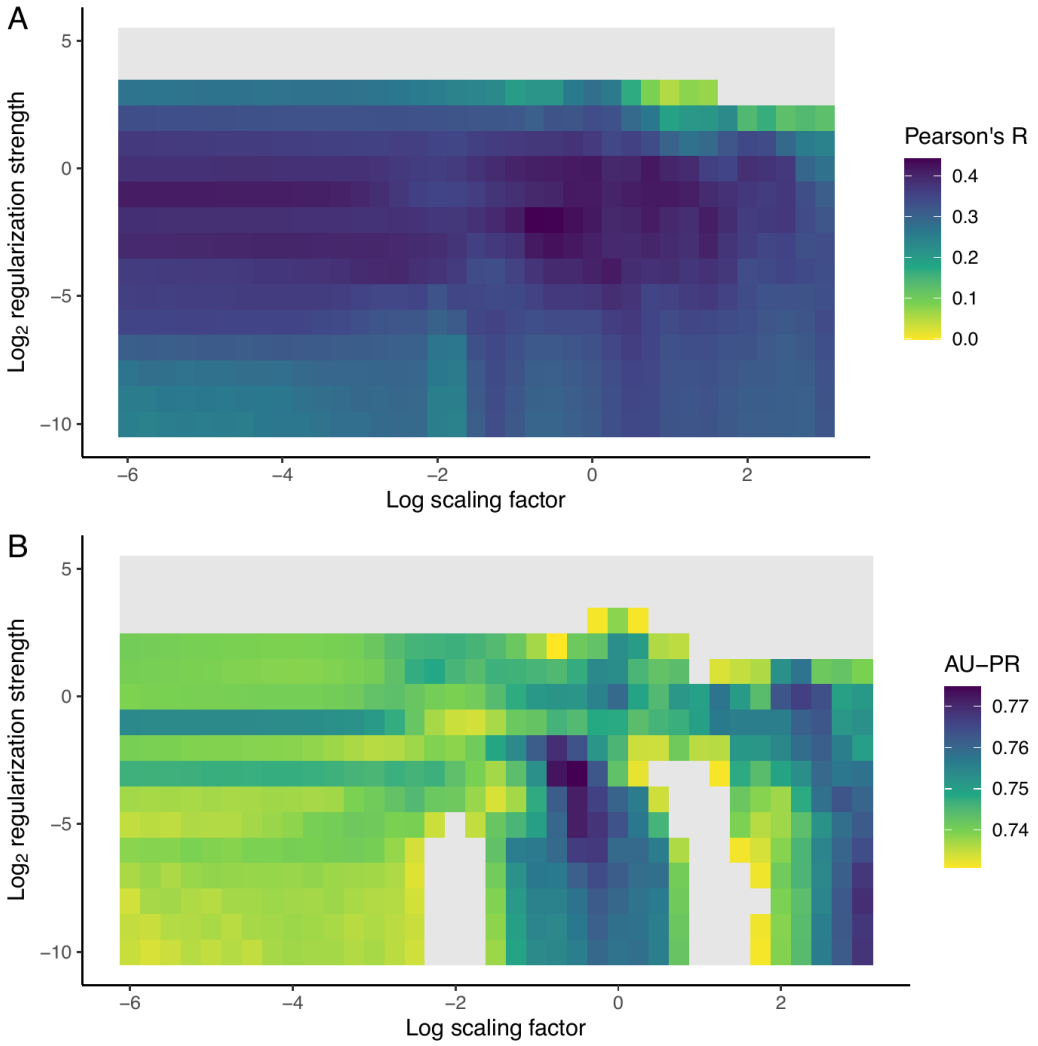


Figure 7.5: **Detailed Pearson's R and AU-PR for the LOOCV.** A) Pearson's R correlation for the pooled predictions from leave-one-out cross-validation (LOOCV) as a function of thermodynamic scaling factor for the Entropic Signatures and regularization strength for the LASSO regression. Values below zero are shown in gray. The two highest correlation coefficients of 0.442 and 0.441 are reached at $\lambda = 2^{-2}$ for both, $\beta = e^{-0.75}$ and $\beta = e^{-0.5}$ respectively. B) Same as A), for area under the precision-recall curve (AU-PR). This highest value of 0.775 is reached at $\lambda = 2^{-3}$ and $\beta = e^{-0.5}$.

benefitting from slightly lower regularization strength. After confirming that the $\lambda = 2^{-3}$ regularization strength still drives over 80% of the LASSO coefficients to zero, we selected its combination with $\beta = e^{-0.5}$ as the optimal parameters for our purposes.

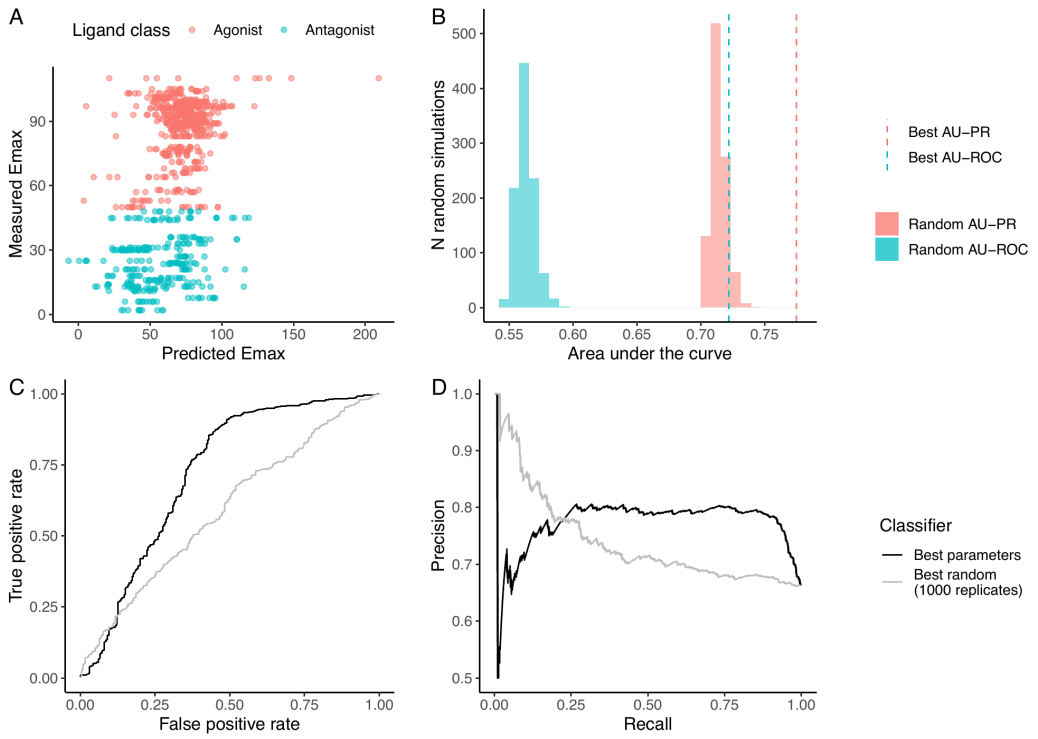


Figure 7.6: **Classification performance on the leave-one-out cross-validation.** The aggregated predictions from the 89 LASSO regression models trained on all ligands except one is shown. The parameters used are $\beta = e^{-0.5}$ and $\lambda = 2^{-3}$. A) The Emax value for the 89 MOR ligands is shown as a function of the predicted Emax. Agonists are defined as having Emax >= 50%, and antagonists as having Emax < 50%. B) Simulated distributions showing area under the receiver operating characteristic (ROC) and precision-recall (PR) curves. The distributions result from 1000 replicates of keeping the best performance from 592 samples of normally distributed noise as the predictor variable. Both AU-PR and AU-ROC are significantly better than random at $p < 0.001$. C) ROC curve, with the best random curve in gray and the curve from aggregated predictions in black. D) Same as C) for PR curve.

Figure 7.6 illustrates the classification performance from the pooled LOOCV predictions with the selected parameter combination. Both AU-ROC and AU-PR metrics are significantly better than random at $p < 0.001$, which is the lowest level our simulations can detect. The ROC curve does not show early enrichment of true positives, however the precision-recall curve shows an interesting maintenance of precision greater than 0.8 from 0.25 to 0.9 recall. This behavior is desired in the context of applying our method to enrich agonists from virtual screening campaigns, as it reflects the model's ability to maintain good precision across almost all examples in our dataset.

7.3.3 *5-fold cross-validation*

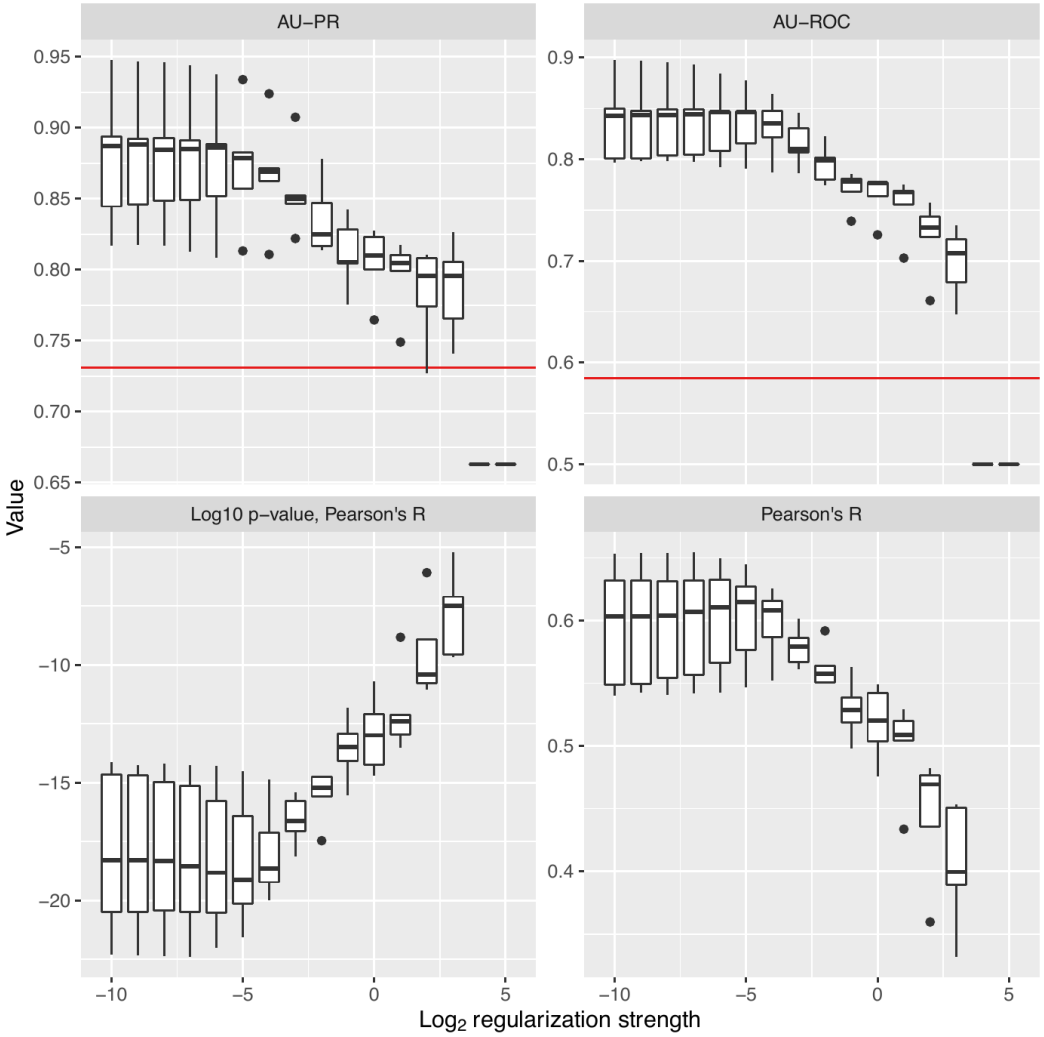


Figure 7.7: **Performance metrics for the 5-fold cross-validation.** Area under the precision-recall curev (AU-PR), area under the receiver operating characteristic curve (AU-ROC), Pearson's R correlation and its associated p-value, $\log_{10}$ transformed, are shown as boxplots across the 5 validation sets. For each metric, the performances at the LASSO regularization strength leading to the best mean performance are shown.

The 5-fold cross-validation performed represents a best-case scenario in which the training set samples a wide diversity of ligand chemotypes. Two poses from each ligand are randomly sampled to constitute every of the five test sets, without resampling. We use the thermodynamic scaling factor identified as optimal by the LOOCV, $\beta = e^{-0.5}$, without exploring other values. Figure 7.7 illustrates AU-ROC, AU-PR and Person's R across the 5 validation rounds, along with the p-value from

the Pearson correlation. A striking feature is that for this easier test, the LASSO model generally benefits from slightly lower regularization strength. However, the $\lambda = 2^{-3}$ value we identified as optimal for our purposes occupies a threshold position before significant drops in performance happen at higher values. As expected, substantial increases in performance are observed compared to the LOOCV, as the model now has examples of each ligand it can learn from. However, we remind that the docking poses selected are all unique, as we discarded duplicate poses.

Figure 7.8 shows the ranking of the 89 ligands obtained by pooling the predictions from the LOOCV (panel A) or from the 5-fold cross-validation (panel B). We observe a striking aggregation of ligands classified as antagonists (Emax < 50%) towards the bottom in both cases. The 5-fold cross-validation leads to high enrichment of agonists (Emax > 50%), with only one antagonist in the top 35 molecules.

7.3.4 *LASSO coefficients*

We trained a final LASSO model using all 890 docking poses at once and the optimal parameters identified of $\beta = e^{-0.5}$ and $\lambda = 2^{-3}$. Figure 7.9 shows the learned LASSO coefficients on the receptor's complex with morphine, as it was in our selected ligands and is arguably the most widely studied MOR ligand. In any case, the coefficient at the ligand bead is zero in the model. The sum of positive coefficients is slightly higher than the absolute sum of negative coefficients, so the model slightly favors overall gains in vibrational entropy. However, this effect is relatively small, with a softening bias (defined in Section 4.4.3) of 5.6%. Strikingly, the largest coefficients in abolute value are located at the extracellular extremity of TM6. Let us remind that TM6 is the structural element undergoing the largest conformational change between active and inactive GPCR states.

7.4 DISCUSSION

In the present chapter, we have used the ENCoM-DynaSig-ML pipeline to capture dynamical patterns which can classify μ-opioid receptor (MOR) ligands as agonists or antagonists. Different ligands occupying the receptor's binding site can effectively be seen as sequence variants of some sort, since they change the nature of atomic interactions happening inside the binding site.

The leave-one-out cross-validation we perform confirms that the LASSO models trained on ENCoM Entropic Signatures are able to learn generalizable patterns of MOR activation/antagonism. Most interestingly, Figure 7.8 shows that even in this hard test, the LASSO model with selected parameters can significantly deplete molecules classified as antagonists (Emax < 50%) from the top predictions. For instance, only 3 antagonists are present in the top 25% ligands, while we would

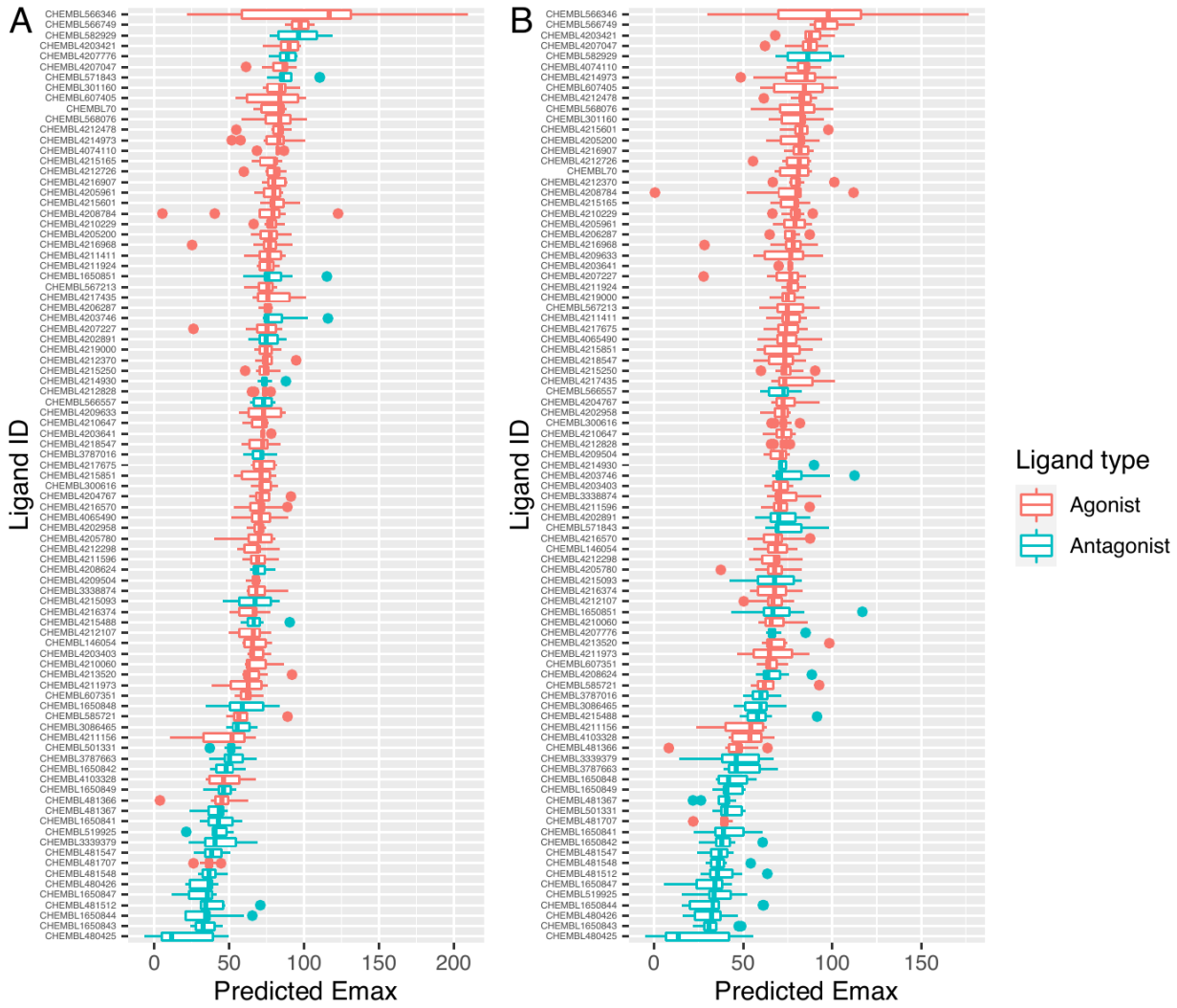


Figure 7.8: **Measured Emax as a function of predicted Emax for the LOO and 5-fold cross-validation.** The predicted Emax is shown for each of the 89 MOR ligands in the dataset for the 10 selected docking poses as a boxplot. The middle line represents the mean. The boxplots are colored blue for antagonists (Emax < 50%) and red for agonists (Emax >= 50%), and the ligands are ordered according to their mean predicted Emax. A) Aggregated prediction from the leave-on-out cross-validation. B) Aggregated predictions from the 5-fold cross-validation.

expect 7.5 of them by chance. Moreover, for the 5-fold cross-validation, which represents an ideal case where the model is presented a ligand which affects the receptor in similar ways to what was seen in the training set, only 2 antagonists are present in the top 50% ligands, while we would expect 15 by chance (half of the 30 total antagonists).

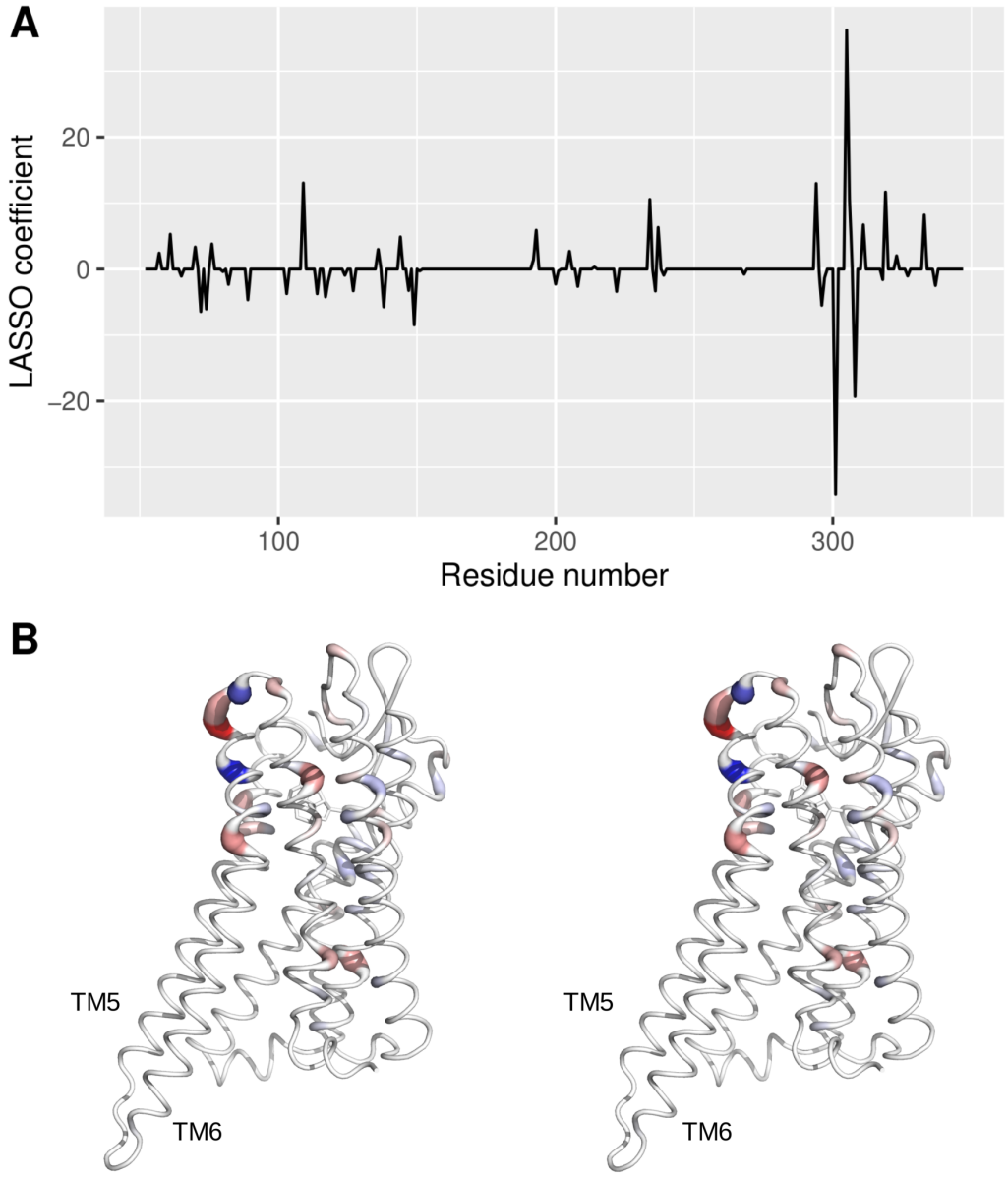


Figure 7.9: **LASSO coefficients for the selected parameters.** A) The LASSO coefficients are plotted for every MOR residue. The ligand coefficient is not shown as it is zero. The sum of coefficients is 16.1, and the absolute sum is 286.5, for a softening bias of 5.6%. B) Stereo crosseye view of the coefficients on the 3D MOR structure (PDB code 5CM1). The thickness of the backbone corresponds to the absolute value of the coefficients, and the color to the value, with a blue:white:red gradient corresponding to negative:zero:positive coefficients. Transmembrane helices 5 and 6 (TM5, TM6) are labelled.

As mentioned throughout the present thesis, one of the main advantages of ENCoM-DynaSig-ML is speed. For one ligand-MOR complex, it takes around 4 seconds CPU

time to compute the Entropic Signature and LASSO predictions. State-of-the-art virtual screening campaigns now routinely screen libraries of up to billions of make-on-demand compounds [183]. Potent ligands with affinities in the nanomolar range have been found through such ultra-large-scale library screens [184–186], however the libraries are so large that many potential hits are never tested experimentally. Thus, in a context where target dynamics are affected by ligand binding, such as MOR activation, an approach like ours would be complementary to ultra-high-throughput virtual screening in order to prioritize molecules with desired properties. Even in the absence of experimental data which can be used to train the model, the ENCoM Entropic Signatures could be used to cluster ligands according to their effects on target dynamics and experimentally tested ligands could then be selected to maximize the diversity of dynamical effects.

When looking at the LASSO coefficients mapped to the MOR 3D structure, the first apparent pattern is the clustering of high absolute value coefficients to the region at the extracellular extremity of TM6, from residue 294 to residue 319. Interestingly, this region corresponds to one of the two key regions identified as important for ligand binding and dynamical in nature from a study using accelerated molecular dynamics [187]. The authors used Gaussian accelerated molecular dynamics to study the active-inactive state transition in MOR, with the presence of either an agonist or antagonist ligand. The fact that our coarse-grained, fast approach identifies the same region as important further highlights that beyond its predictive ability, the ENCoM-DynaSig-LASSO technique is able to pinpoint regions of biological interest.

In conclusion, we presented in this chapter a study of the dynamical impacts of ligand binding to the μ-opioid receptor and how these effects allow LASSO regression models to predict ligand efficacy and classify ligands as agonists or antagonists with high enough accuracy to open the door for applications to virtual screening campaigns. In contrast with the models trained in the last chapter, the ENCoM Entropic Signatures were the only input feature fed to the LASSO models, as the docking score did not lead to significant ligand classification, as expected since affinity does not correlate with activity. The thermodynamic scaling factor leading to optimal performance was again lower than the one agreeing with the MSF Dynamical Signature, pointing to an important contribution of higher-frequency normal modes.

8

VIM-2 LACTAMASE EVOLUTIONARY FITNESS

This chapter presents the last of our three dynamics-function case studies, the evolutionary fitness provided by sequence variants of VIM-2 lactamase under antibiotic selection, which acts as a proxy for its catalytic efficiency. The biological background, as for the other two case studies, was already given in Section 1.3.3.

In contrast to the previous two chapters, there is no generally accepted role for collective, slow-timescale motions in the enzymatic activity of VIM-2 lactamase. In a recent review, Gianquinto and coworkers argue for the targeting of active site dynamics and allostery as tools against antibiotic resistance in β-lactamases [71], however the knowledge on metallo-β-lactamases dynamics as a family, and on VIM-2 lactamase dynamics specifically, is scarce. We thus view the present case study as an answer to the question of whether structural dynamics play a significant role in the function of VIM-2 lactamase.

The authors of the VIM-2 deep mutational scan dataset we study also provide a number of predictor variables that they used to train a linear model in predicting evolutionary fitness of VIM-2 sequence variants. These variables, which are sequence-, structure- and stability-based, provide us with an opportunity to test our hypothesis that ENCoM-DynaSig-ML, as the only variant effect predictor tool capturing detailed dynamics, is complementary to such approaches.

Here again, the specific methodology will be detailed first, followed by the results and discussion in the context of VIM-2 lactamase.

8.1 METHODOLOGY

8.1.1 *VIM-2 deep mutational scan dataset*

The dataset from Chen *et al.* consists of experimental measures of evolutionary fitness for a deep mutational scan of VIM-2 lactamase, at various antibiotic concentrations [70]. The authors calculate evolutionary fitness from deep sequencing, as the $\log_2$ enrichment of the variant relative to WT VIM-2. They performed two replicates for each antibiotic condition tested, and the highest agreement between the replicates is reached for the highest concentration of ampicillin tested, 128

μg/mL, with an R^2 value of 0.94. Thus, we chose the fitness score at 128 μg/mL ampicillin as the experimental outcome to predict.

We kept only missense variants affecting the 231 positions present in the VIM-2 crystal structure. 4343 out of the 4389 possible missense variants were observed in the deep mutational scan for the selected conditions, for a coverage of 99.0% of the possibilities.

These 4343 variants were modeled on the WT VIM-2 crystal structure (PDB code 4bz3) solved by Brem *el al.* [69], with MODELLER as outlined in Section 4.2.2. We used biological unit 1 since there were two units of the enzyme in the asymmetric crystal unit.

8.1.1.1 *Static predictor variables*

Chen and coworkers fitted a linear model containing sequence, structural and stability properties to predict the observed evolutionary fitness at 128 μg/mL ampicillin and report a good fit with an adjusted R^2 coefficient of 0.55. Their linear model combines the $\Delta\Delta G$ of folding computed with Rosetta [119] (see Section 2.3.1.4), the solvent acessible surface area of the mutated residue and a vector of length 40 specifying the starting amino acid and the one it is mutated to. The authors made the values of these predictors available and we will use them for establishing a baseline performance of sequence- and structure-based models.

8.1.2 *Zinc ions*

The catalytic site of VIM-2 lactamase contains two coordinated zinc ions, which play a central role both in the β-lactam hydrolyzation reaction [188] and in the interaction with some inhibitors of VIM-2 [66]. For this reason, we decided to investigate whether the inclusion of the zinc ions in the ENCoM representation of VIM-2 would lead to increased performance. To do so, we defined a new residue type, the zinc ion, with the obvious choice of a single bead representation at the position of the atom. Of the 8 atom types, we selected type 3, h-bond donor, as the closest match for the zinc cation. We added the two zinc ions after the modeling of variants with MODELLER, in the same position as the crystal structure.

We computed all ENCoM Entropic Signatures for all VIM-2 variants both with and without the zinc ions, with the same set of scaling factors and regularization strengths.

8.1.3 *EntroSigs scaling factors*

As in the previous chapters, we investigated a wide range of scaling factors for Entropic Signatures of WT VIM-2 lactamase to select the scaling factors to test.

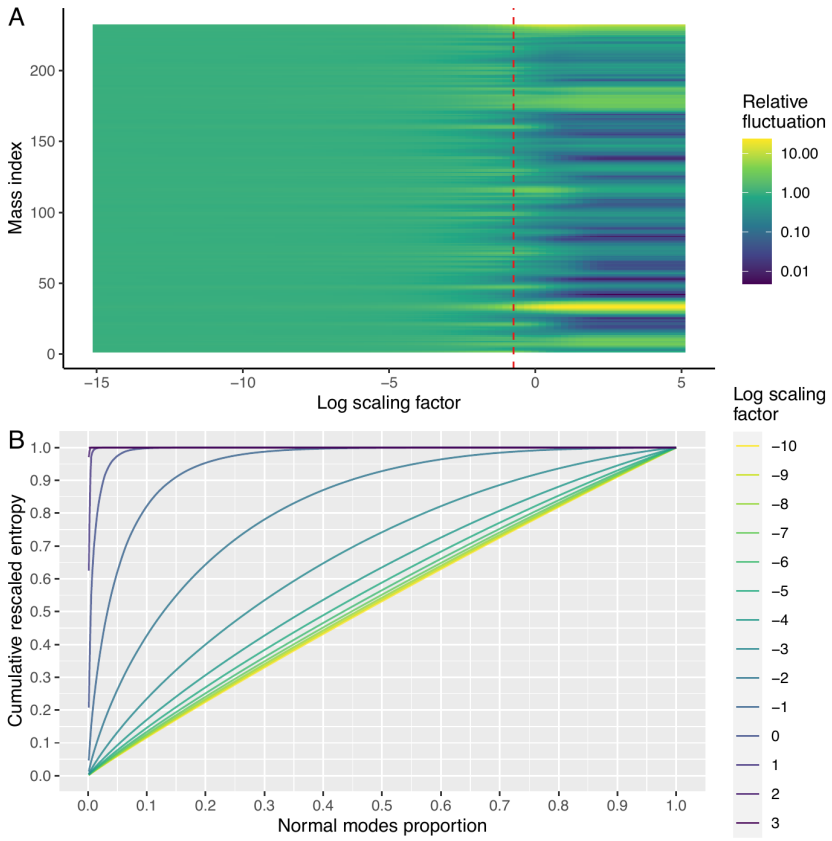


Figure 8.1: **Entropic Signatures and entropy proportions for VIM-2 lactamase across selected scaling factors.** Input structures both with and without zinc ions led to very similar patterns across the scaling factors, so only the results on the structure without zinc are shown. A) Entropic Signatures across scaling factors ranging from e^{-15} to e^{5} in log increments of 0.25. A red dashed line shows the value leading to almost perfect agreement with MSF, $e^{-0.75}$. B) The cumulative entropy across all normal modes is given for scaling factors across the selected range of e^{-10} to e^{3}, rescaled to sum to 1.

Figure 8.1 illustrates the behavior of EntroSigs and cumulative entropy. We selected scaling factors ranging from e^{-10} to e^{3}, in log increments of 0.25.

8.1.4 *Classification problem*

The evolutionary fitness provided by a VIM-2 variant under antibiotic selection can naturally be thought of as a binary classification problem: either the variant allows roughly equal or better antibiotic degradation compared to WT, or it significantly hinders it and thus survival. Moreover, Chen *et al.* specifically chose the maximal antibiotic concentrations so that they allow bacteria expressing WT VIM-2 to grow, further amplifying the binary phenomenon [70]. The authors defined three classes

of variant effects: negative, neutral and positive. They imposed a fairly stringent cutoff for the positive class of 0.7 fitness score, which corresponds to 62% better evolutionary fitness than WT VIM-2.

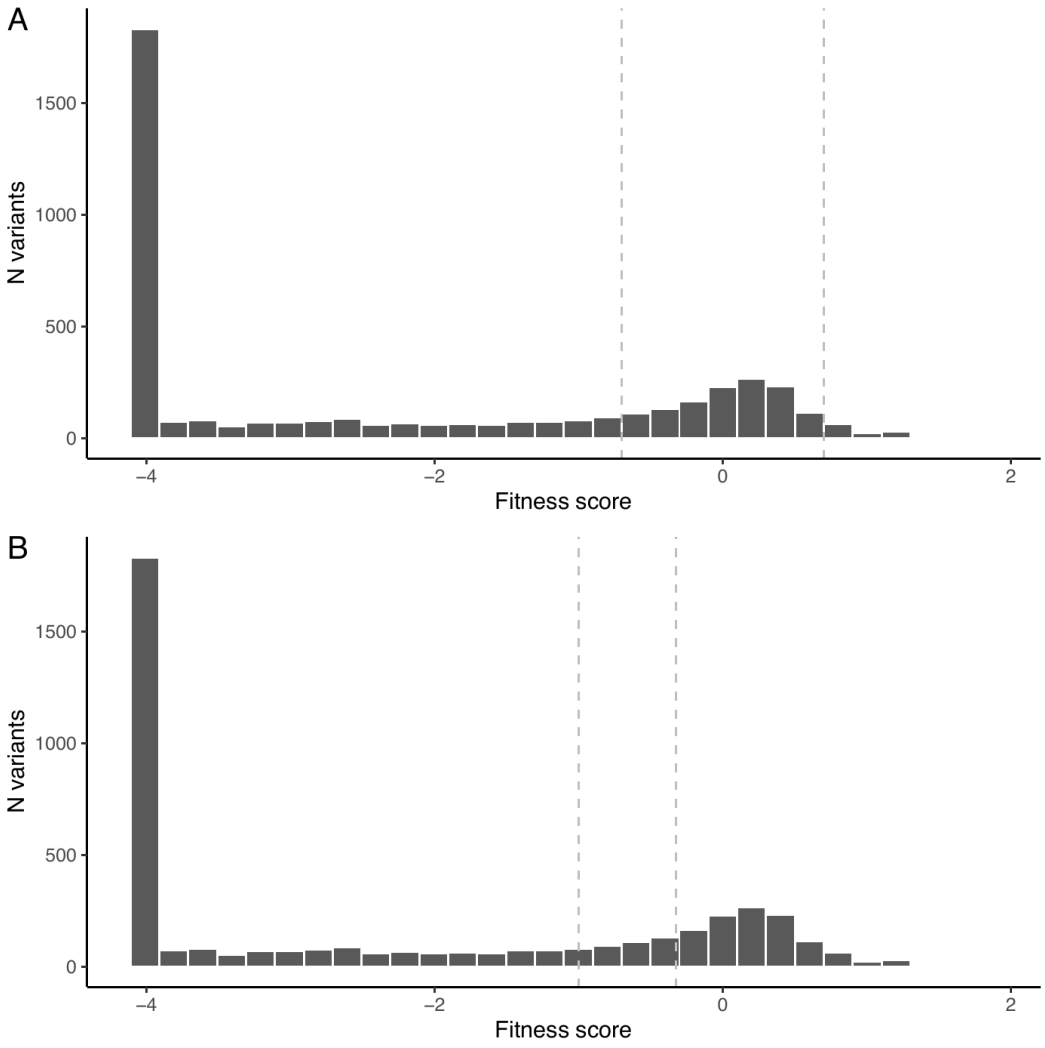


Figure 8.2: **Fitness score thresholds for positive and negative variants.** The fitness score distribution for all missense variants are shown as histograms, with the thresholds for negative and positive variants in dashed vertical lines. A) The Chen *et al.* thresholds of -0.7 for negative and 0.7 for positive variants. B) Our thresholds of -1 for negative and $\log_2 0.8$ for positive (approx. -0.322).

To partition the variants in two classes instead of three, we chose the same cutoffs as for the analysis of miR-125a maturation efficiency presented in Chapter 6, namely under 50% of the WT VIM-2 evolutionary fitness for negative variants and over 80% of the WT VIM-2 evolutionary fitness for positive variants. These values translate to fitness scores of $\log_2 0.8$ and $\log_2 0.5$ respectively, or -0.322 and -1. Our

thresholds are shown in Figure 8.2, along with the ones from Chen and coworkers. They partition the dataset in 1149 positive variants, 2837 negative variants and 358 neutral variants. The classification task thus has 28.8% positive observations.

8.1.5 *5-fold cross-validation*

Since only point mutations are part of this dataset, sequence redundancy is intrinsically absent. Thus, we split the dataset in 5 and perform 5-fold cross-validation to test the models. The maximum number of variants for a given position is 19. Apart from position 202 which has 13 sampled variants, all positions have at least 15. Thus, we sampled 3 variants from every position except position 202 for every split of the dataset, and 2 variants from position 202. The remaining 884 variants (including the WT sequence) were split randomly between the 5 sets, leaving us with four sets of 869 variants and one set of 868 variants. The splitting-by-position scheme we used ensures that both mutation-tolerant and mutation-intolerant positions are uniformly distributed between the testing sets.

8.2 RESULTS

8.2.1 *5-fold cross-validation*

We were first interested in answering two questions with the 5-fold cross-validation performance: whether the presence of the zinc ions leads to more signal in the Entropic Signatures, and whether there is complementarity between the static predictors and the EntroSigs.

Figure 8.3 shows the performance attained by the EntroSigs alone, the static predictors alone, and the combination of both. The performance metrics shown are AU-ROC, AU-PR and the predictive coefficient of determination R^2. Since we are evaluating combinations of methods, R^2 is preferable to Pearson's R for its interpretability in terms of proportion of variance explained. The EntroSigs with and without the inclusion of the active site zinc ions were tested. While the static predictors perform better than the EntroSigs by a sizeable margin across the three metrics, the two show striking complementarity. Moreover, the performance gain is the largest for predictive R^2, which is the most stringent metric of the three presented since it includes all data points (neutral variants included) and measures the goodness of fit across the whole range.

The presence of the zinc ions introduces an interesting pattern: while the performance with or without ions is closely matched for low scaling factors, the abrupt drop in performance associated with higher scaling factors happens sooner with the presence of ions. Since the performance does not seem improved by the presence of the ions, we will restrict further analyses to models without them.

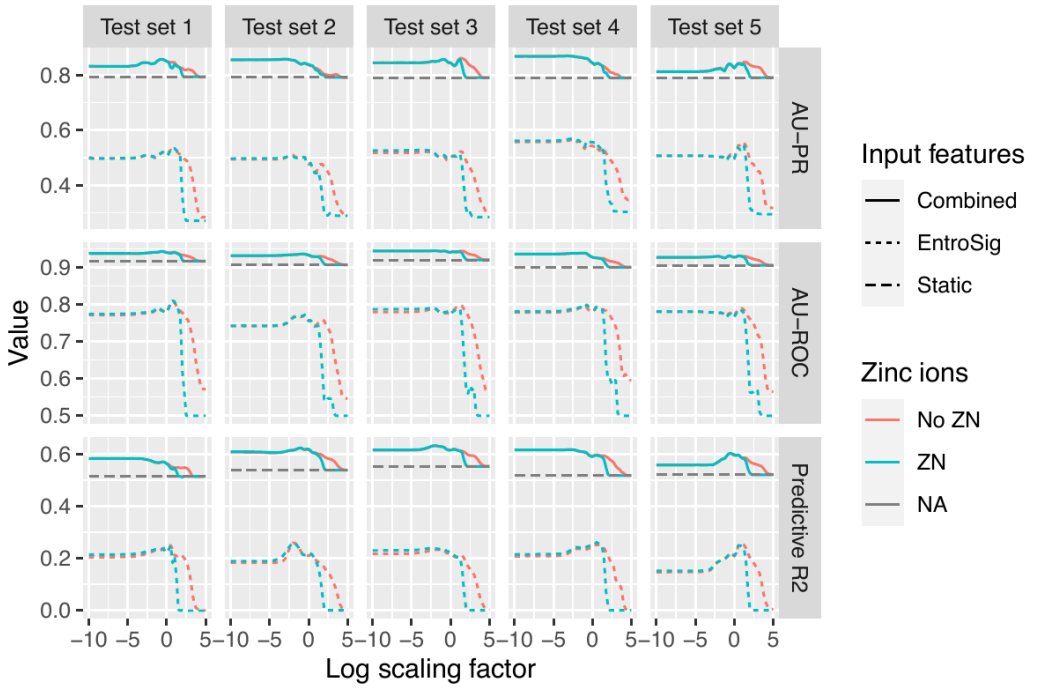


Figure 8.3: **Performance metrics for the 5-fold validation.** Area under the precision-recall curve (AU-PR), area under the receiver operating characteristic curve (AU-ROC) and predictive R^2 are given as a function of the thermodynamic scaling factor for the EntroSigs. The performance shown is the highest one across the 16 regularization strengths tested. The presence or absence of ions is denoted by the line color (NA for the static predictors).

Figure 8.4 gives the detailed R^2 and AU-PR for all tested combinations of LASSO regularization strength and thermodynamic scaling factor for the EntroSigs, averaged over the 5 test sets, for the combination of EntroSigs and static predictors. Values below the best performance from static predictors alone are shown in gray. In terms of both metrics, the best performance is attained at $\beta = e^{-0.75}$ and $\lambda = 2^{-9}$. The combination reaches predictive R^2 of 0.60, a gain of 0.07 or 7% of variance explained over the static predictors alone. It terms of AU-PR, the combination attains 0.85, with the static predictors reaching 0.79. Interestingly, the best thermodynamic scaling factor is the one equivalent to the MSF Dynamical Signature, in contrast to the two previous case studies in which it was significantly below.

8.2.2 *Generalizable static predictors*

The combined static predictors perform very well on the VIM-2 DMS dataset. The Rosetta predicted $\Delta\Delta G$ of folding is readily generalizable to variants with multiple mutations, and one could argue that the average solvent accessible area of all mutated residues could be used in replacement of the accessible solvent area

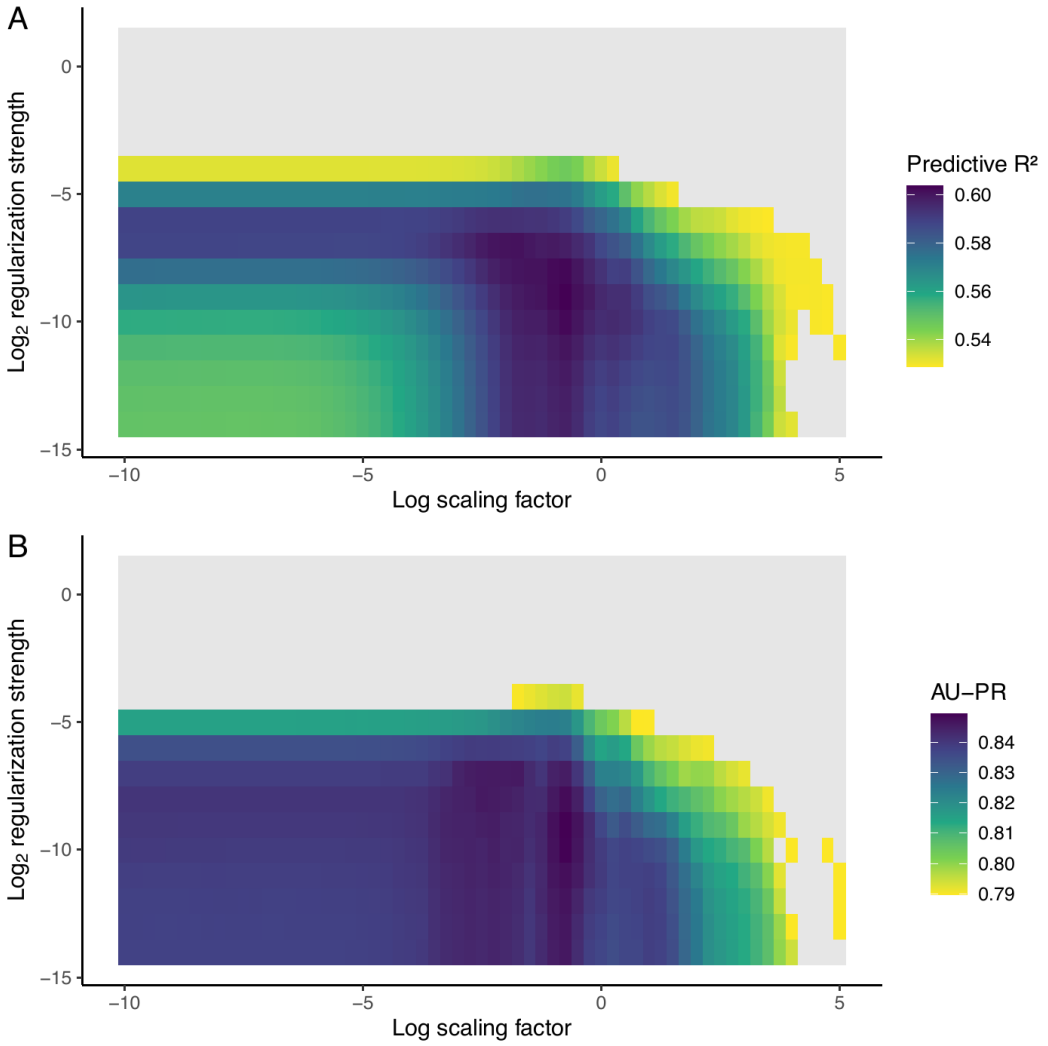


Figure 8.4: **Detailed AU-PR and R^2 for the combination of EntroSigs and static predictors.** A) Predictive R^2 as a function of both thermodynamic scaling factor and regularization strength, averaged over the 5-fold cross-validation. Values are shown for the combination of Entropic Signatures and static predictors when they exceed the performance of static predictors alone (0.53), otherwise are shown in gray. The best performance of $R^2 = 0.603$ is reached for $\beta = e^{-0.75}$ and $\lambda = 2^{-9}$. B) Same as A) for area under the precision-recall curve, with the performance of static predictors alone at 0.79. The best performance of 0.849 AU-PR is also attained at $\beta = e^{-0.75}$ and $\lambda = 2^{-9}$.

of the single mutated residue. However, it is unclear if the same can be said of the mutations identification vector. For instance, only binary values are present in this vector during training. Moreover, this vector would be unable to capture, for example, a mutation pair reversing a salt bridge interaction, which in theory should have minimal effect on fitness if it does not affect the binding site. Instead,

the vector would either sum or average the effects of both single mutations. Thus, we do not believe in the generalizability of the mutations identification vector.

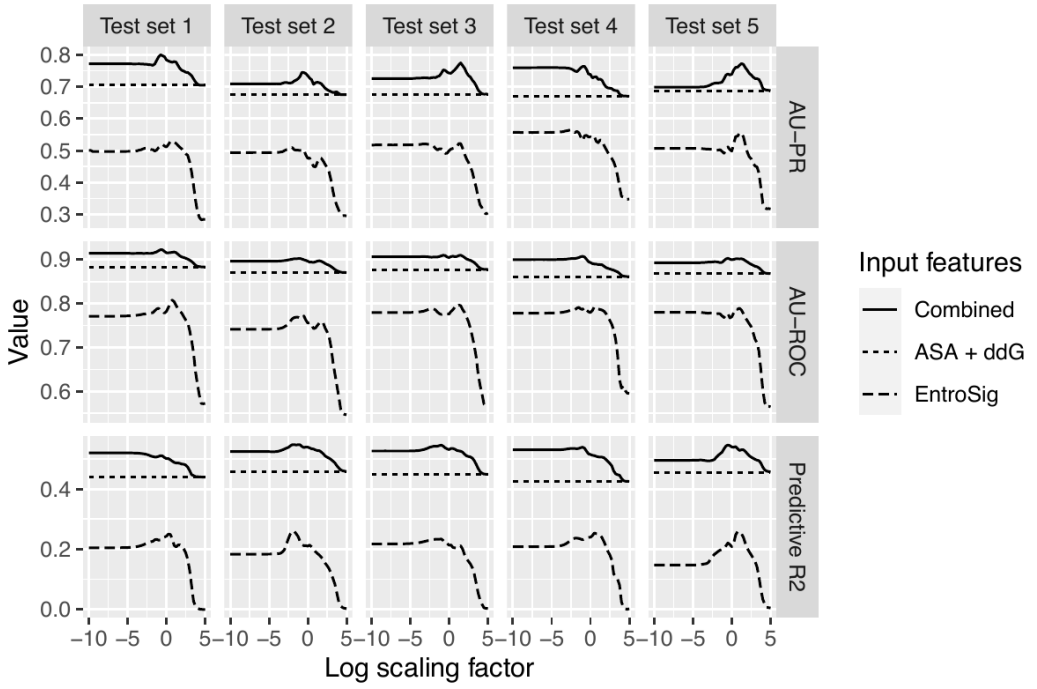


Figure 8.5: **Performance metrics for the generalizable static predictors.** Area under the precision-recall curve (AU-PR), area under the receiver operating characteristic curve (AU-ROC) and predictive R^2 are given as a function of the thermodynamic scaling factor for the EntroSigs. The performance shown is the highest one across the 16 regularization strengths tested.

Since the main application of our ENCoM-DynaSig-ML approach is the virtual screening of variants, we investigated the performance of generalizable static predictors, namely the solvent accessible area of the mutated residue and the Rosetta $\Delta\Delta G$, alone and in combination with the EntroSigs. Figure 8.5 illustrates the performance of these predictors on the 5-fold cross-validation, again showing AU-ROC, AU-PR and predictive R^2 for the EntroSigs alone, the generalizable static predictors alone and the combination of both. There is a significant drop in the performance of the static predictors from the removal of the mutations identification vector, as expected, however the combination of solvent accessible area and Rosetta $\Delta\Delta G$ still outperforms the Entropic Signatures by a sizeable margin. Nonetheless, the combination of both stays beneficial, as was expected because of the linearity of the LASSO regression.

Figure 8.6 gives the detailed performance of the EntroSigs + ASA + Rosetta $\Delta\Delta G$, averaged over the 5-fold cross-validation, for each combination of β and λ values. Here again, values of R^2 or AU-PR below the performance of the ASA

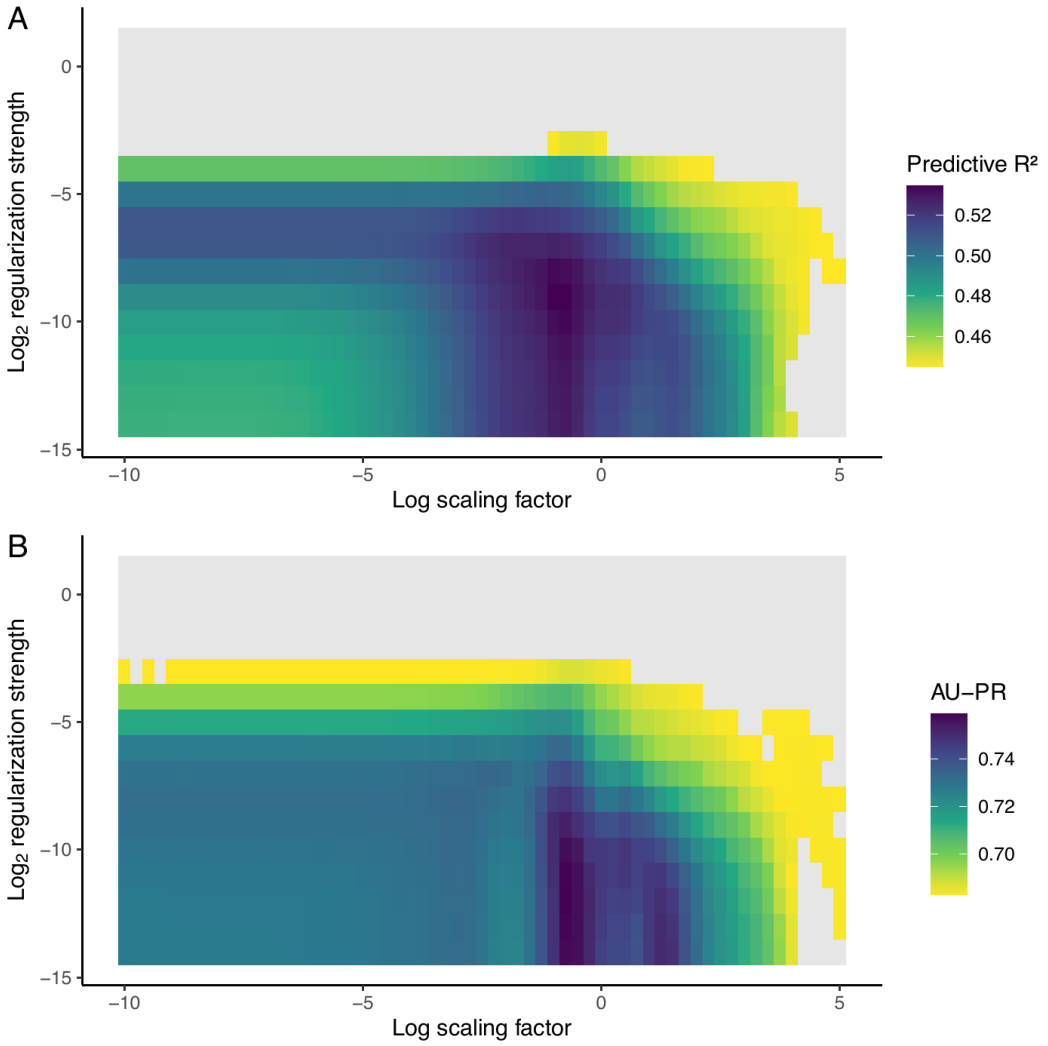


Figure 8.6: **Detailed AU-PR and R^2 for the combination of EntroSigs and generalizable static predictors.** The generalizable static predictors are the solvent accessible surface area of the mutated residue and the Rosetta ΔΔG of folding. A) Predictive R^2 as a function of both thermodynamic scaling factor and regularization strength, averaged over the 5-fold cross-validation. Values are shown for the combination of Entropic Signatures and static predictors when they exceed the performance of static predictors alone (0.44), otherwise are shown in gray. The best performance of $R^2 = 0.534$ is reached for $\beta = e^{-0.75}$ and $\lambda = 2^{-9}$. B) Same as A) for area under the precision-recall curve, with the performance of static predictors alone at 0.68. The best performance of 0.759 AU-PR is also attained at $\beta = e^{-0.75}$, but $\lambda = 2^{-12}$. It drops slightly to 0.753 at $\lambda = 2^{-9}$

+ ΔΔG predictors are shown in gray. Interestingly, the complementarity of these generalizable static predictors with the EntroSigs increases slightly, as predictive R^2 registers a gain of 0.09 compared to 0.07 with all static predictors, and AU-PR a

gain of 0.08 compared to 0.06. The best performances happen at $\beta = e^{-0.75}$ for both metrics, however AU-PR benefits from lower LASSO regularization strength.

Going from $\lambda = 2^{-9}$ to $\lambda = 2^{-7}$ leads to a drop of 0.01 in predictive R^2, and going from $\lambda = 2^{-12}$ to $\lambda = 2^{-7}$ to a drop of 0.02 in AU-PR. These drops in performance are not very large, and thus we chose the combination of $\beta = e^{-0.75}$ and $\lambda = 2^{-7}$ as the parameter combinations we are the most confident would lead to good generalizability of the model to multiple mutations. We thus train a final LASSO model with these parameters on the combination of EntroSigs + ASA + Rosetta $\Delta\Delta G$, using all sequence variants in the dataset.

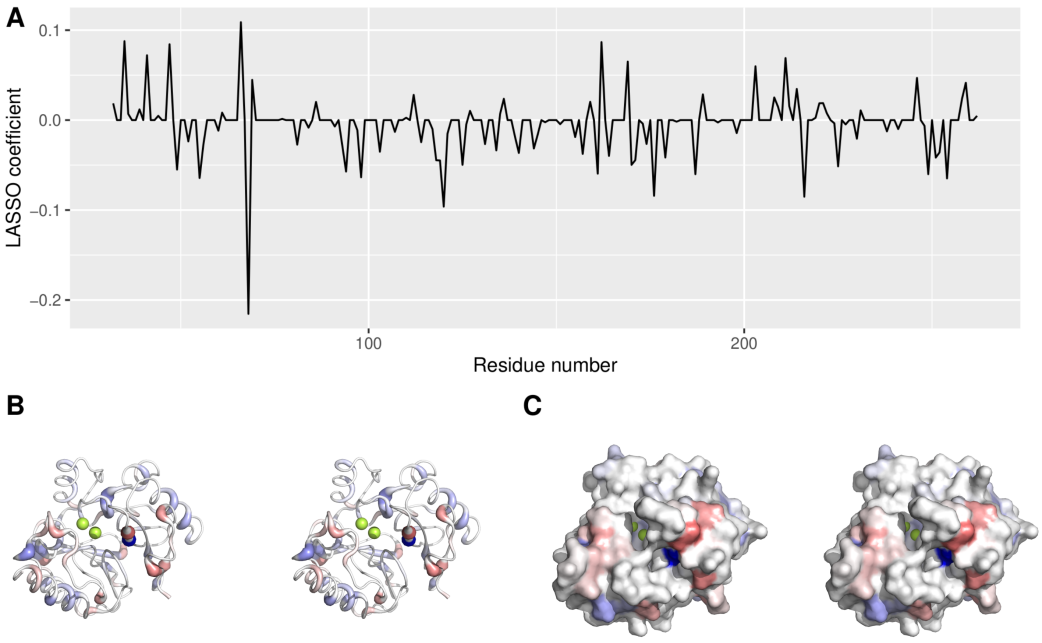


Figure 8.7: **LASSO coefficients for the selected parameters.** A) The LASSO coefficients are plotted for every VIM-2 residue. The sum of coefficients is -0.85, and the absolute sum is 3.13, for a rigidifying bias of 27%. B) Stereo crosseye view of the coefficients on the 3D structure of VIM-2 lactamase. While the active site zinc ions are not part of the model we selected, they are shown to help in visual identification of the binding site. C) Same as B), with surface view.

Figure 8.7 shows the EntroSig coefficients learned by the LASSO model, graphically for the whole VIM-2 sequence and mapped back on the VIM-2 3D structure. A striking feature is the negative sum of coefficients, leading to a rigidifying bias of 27%. Another apparent feature is the appearance of most highly positive coefficients on the protein surface.

For this last of the three dynamics-function case studies, we have investigated the evolutionary fitness provided by VIM-2 lactamase sequence variants under antibiotic selection, which acts as a proxy for both the catalytic efficiency of the enzyme and its stability [70]. We wanted to answer two key questions in this chapter: first, how complementary are the ENCoM Entropic Signatures to the static predictors used by Chen *et al.* to predict evolutionary fitness; and second, can we provide insights as to the role of structural dynamics in VIM-2 lactamase enzymatic activity.

The answer to the first question is that the Entropic Signatures do exhibit good complementarity to the sequence-, structure- and stability-based predictors tested here. In the case of generalizable static predictors, namely the solvent accessible area of the mutated residues and the Rosetta $\Delta\Delta G$, 39% of the variance explained by the Entropic Signatures is additive (0.09 out of the maximal 0.23 R^2 from the EntroSigs alone). This relatively high complementarity, combined with the very low computational cost of ENCoM-DynaSig-ML, suggests that our method could be integrated to most consensus variant effect predictors with expected gains in performance. Moreover, the DynaSig-ML Python package allows for the user-friendly integration of our tool, as outlined in Section 4.7.2.

As for the second question, we think that our results point to dynamical features playing a role in the enzymatic activity of VIM-2 lactamase, however we also think that some features learned by the LASSO model capture stabilization of the enzyme. For instance, the positive coefficients learned by the model seem to coalesce on the surface of the enzyme, in a region surrounding the binding site. To us, these patterns point to an important role for mutations affecting the opening/closing motion of the catalytic site. However, the negative coefficients appearing throughout the protein structure point towards the model capturing stabilizing effects, in a similar fashion as what has been observed for the ENCoM vibrational entropy of the whole protein [28]. Interestingly, these two seemingly opposite effects might be reconcilable: it has been suggested by Shoichet and coworkers that most enzymes make tradeoffs between activity and stability [189]. Thus, the negative coefficients, mostly positioned in alpha helices distant from the active site, favor thermal stability of the enzyme as captured by rigidification. Meanwhile, positive coefficients, which coalesce on surface residues surrounding the active site, capture the softening of the binding site, which generally leads to higher activity. In fact, Beadle and Shoichet confirmed the existence of stability-function tradeoffs in another, better characterized β-lactamase enzyme, the AmpC β-lactamase [190].

We tested the inclusion of zinc ions in the ENCoM potential and found that it led to performance on par with Entropic Signatures generated without them, and even

worsened performance at high thermodynamic scaling factors. At these high scaling factors, the Entropic Signature mostly reflects the fluctuations from the single lowest-frequency normal mode. Thus, we hypothesize that the addition of zinc ions changes this lowest-frequency mode, while the fluctuation space spanned by the few percent slowest modes remains unchanged, hence the equivalent performance at lower scaling factors. This finding illustrates one of the advantages of considering more normal modes in the computation of the Dynamical Signatures, as the relative ordering of the modes can be slightly disrupted by changes in the input structure. Thus, considering a wider ensemble of normal modes makes the predictions more robust to slight inaccuracies. However, there is a narrow range of beneficial proportion of normal modes to include. Indeed, results from the present chapter and from the other two case studies demonstrate a general drop in performance for very low scaling factors, which correspond to almost uniform contributions from all normal modes to the Entropic Signature.

In conclusion, we have investigated dynamics-function relationships in VIM-2 lactamase and have confirmed the good complementarity of ENCoM Entropic Signatures with sequence-, stability- and structure-based predictors. Here again, the mapping of the LASSO coefficients to the studied biomolecule allows interesting biological insights, and seems to confirm an important role of structural dynamics in the function of VIM-2 lactamase.

9

GENERAL DISCUSSION

In the present thesis, we have introduced the ENCoM-DynaSig-ML computational pipeline and have applied it to study dynamics-function relationships apparent in three distinct biomolecular phenomena: the first step of microRNA biogenesis (Chapter 6), the activation of the μ-opioid receptor (Chapter 7) and the catalytic efficiency and stability of the VIM-2 β-lactamase enzyme (Chapter 8). Moreover, in order to study RNA molecules with ENCoM, we have adapted it by extending its functionality to permit the streamlined inclusion of arbitrary numbers of beads per residue. This adaptation to RNA prompted us to ask whether the current parameters, optimized for proteins, are adequate for RNA molecules and whether the model is applicable as is to RNA-protein complexes.

We answered this question with a wide parameter search across seven diverse benchmarks, presented in Chapter 5. As a result of this, we confirmed that a single set of parameters can lead to good performance on protein, RNA and RNA-protein complexes. Moreover, the new set of parameters we found confirms that ENCoM's performance edge lies in its surface complementarity term. Indeed, Frappier and Najmanovich did not explore values for the σ_- and σ_+ surface complementarity parameters as part of the original ENCoM publication [27], and we saw a large increase in the σ_+ parameter as a result of our parameter search. Since σ_+ is the surface complementarity term for favorable atomic interactions in the ENCoM model, and increasing its value led to better performance, we can thus confirm the advantage provided by that term.

The aforementioned chapters each contain their own discussion, therefore, the present discussion chapter will focus on general findings and future directions, and will be relatively short.

9.1 GENERAL FINDINGS ABOUT ENCOM-DYNASIG-ML

Across the three case study chapters presented, we have found that ENCoM-DynaSig-ML led to significant gains in performance when combined to folding enthalpy, sequence-based descriptors, and structural and stability properties. To the best of our knowledge and with confirmation from a recent review of variant effect predictors discussed in Section 2.3.2 [133], the pipeline represents the first and only

computational tool in the category of dynamics-based variant effect predictors. These findings of high complementarity to other biomolecular properties reinforce the notion that as the unique member of its class, ENCoM-DynaSig-ML will lead to complementarity when integrated with other variant effect predictors. Such integration seems like a promising area for future work.

Another striking finding is that in all three case studies, the LASSO coefficients mapped on the structures of the biomolecules pointed back to known biological features. The highest coefficients in the case of pri-miR-125a happened on the mismatched GHG motif, known to favor maturation when incorporating a flexible noncanonical base pair [175]. In the case of μ-opioid receptor activation, the highest absolute value coefficients coalesce to the upper part of transmembrane helix 6 (TM6). Not only does TM6 undergo the most substantial displacement between the active and inactive conformations, the specific region with high absolute coefficients was identified as a key factor of the differential effects of agonist and antagonist binding [187]. Finally, the coefficients learned in the case of VIM-2 catalytic efficiency seem to highlight stability-function tradeoffs in the enzyme, which are thought to be a general feature of most enzymes [189].

The computational cost associated with the prediction of the effect of a sequence variant is less than 5 seconds CPU time for all three case studies, making ultra-high-throughput predictions within reach as we have shown in the case of the 30 million theoretical pri-miR-125a variants for which we have predicted maturation efficiencies in Chapter 6. Moreover, let us note that ENCoM is currently implemented in Python, an interpreted language which is not optimized for speed. While we make use of NumPy [191] for the heaviest linear algebra operations, we believe a speedup of at least 5X would be achievable by either implementing another version in a fast language, or making use of the Numba just-in-time compiler [192]. Furthermore, the lowest useful thermodynamic scaling factor selected in the case of pri-miR-125a maturation efficiency leads to around 30% of the normal modes making significant contributions to the Entropic Signature. In the other two cases, the proportion is lower than 10%. Thus, it could be possible to accelerate predictions even further by making use of the block Lanczos algorithm to extract only the relevant normal modes [193]. We thus believe that a total speedup of 10X or more would be achievable. The pursuit of such an acceleration of ENCoM-DynaSig-ML represents another interesting avenue for future work, and would allow the tackling of larger systems while maintaining ultra-high-throughput prediction ability.

9.2.1 *Untangling enthalpy-entropy compensation effects*

Previous work from Frappier and Najmanovich has found that proteins from thermophile bacteria, when compared to homologs from related mesophile organisms, exhibit reduced vibrational entropy [28]. Under the assumption that the thermophile proteins are more stable, this finding leads to the contradictory notion, from a Gibbs free energy point of view, that a loss in entropy is associated with higher stability. Indeed, lower vibrational entropy is related to higher predicted stability, whereas it should be leading to higher Gibbs free energy, thus lower predicted stability. This so-called "sign error" in the predictions from Frappier & Najmanovich's works [27, 28] has always slightly bothered us. While it does make sense from a functional perspective that thermophile proteins have evolved to be more rigid at room temperature (leading to comparable rigidity at their preferred temperature) than their mesophile homologs, it is against the laws of thermodynamics to associate lower vibrational entropy with lower Gibbs free energy. In that regard, an interesting finding from Gerasimavicius *et al.* in their study of pathogenicity predictions for missense mutations is that while all tools benefited from using the absolute value of the predicted change, ENCoM benefited the most [115]. This points to the potential existence of two distinct effects on stability captured by the ENCoM vibrational entropy:

1. The rigidification effect of mutations lowering vibrational entropy. While contradicting Gibbs free energy on an entropic level, these predictions could be partly correct due to the ENCoM potential function capturing the enthalpy-entropy tradeoff, and said tradeoff sacrificing entropy for larger gains in enthalpy.

2. "Pure" entropic effects where ENCoM predicts higher vibrational entropy, which is related to higher thermodynamic stability, as should be the case.

Or course, these two effects each have a corresponding inverse effect, namely softening mutations hurting stability through losses in enthalpy and rigidifying mutations hurting stability through loss of entropy, as should be the case according to Gibbs free energy. These reversals are probably part of what constitutes the considerable gain that Gerasimavicius *et al.* observe when taking the absolute value of the ENCoM ΔS_{vib} predictions.

In fact, an unexpected advantage of the ENCoM-DynaSig-ML pipeline is its ability to capture both types of effects at the same time. As discussed in Chapter 8, we believe that the LASSO coefficients learned for the VIM-2 lactamase reflect a tradeoff between stability and function, as was suggested as a general feature of enzymes [189, 190]. The decomposition of entropic effects by residue thus seems to

allow for the natural resolution of the "sign error" phenomenon, when experimental mutagenesis data are available.

9.2.2 *Entropic Signatures and the edge of chaos*

We have introduced Entropic Signatures as novel measures of dynamical properties at every bead in a coarse-grained normal mode analysis model, which scale the mean square fluctuations of each normal mode by its vibrational entropy. In doing so, a dependence on a thermodynamic scaling factor appears, which proves very useful in adjusting the relative contribution of normal modes to the Dynamical Signature. Indeed, we found a narrow range of scaling factors leading to improved performance in all three case studies presented. Only in the last case study did the optimal scaling factor match the widely-used, standard mean square fluctuations (MSF). In the first two, the optimal scaling factors were lower that the one equivalent to the MSF, corresponding to higher contributions from higher-frequency normal modes.

We were initially surprised to see the best performance in the miR-125 hard test emerge with $\beta = e^{-2.5}$, a very low value leading to a dominance of local effects from high-frequency normal modes. We hypothesized that the hard test benefited from the low scaling factor due to the capture of localized perturbations. Subsequent explorations of the 8 boxes benchmarks, which tested the ability of the Entropic Signatures to capture delocalized motion, confirmed that the optimal values for the scaling factor were higher for this type of motion. One value, $\beta = e^{2.25}$ was relatively close to the value leading to perfect agreement with the MSF signature ($e^{2.75}$), however the two other selected values were significantly below at $\beta = e^{1.5}$ and $\beta = e^{-1.5}$. Moreover, our ultimate test of generalizability of the LASSO models, the inverted benchmark in which the model has to predict the maturation efficiency of variants containing 3 or more mutations from variants containing only 2, has shown that the lowest of these scaling factors is the one leading to the best performance. This value of $\beta = e^{-1.5}$ corresponds to the slowest 25% of normal modes contributing 90% of the entropy, thus still exluding most high-frequency, localized motions, while allowing contributions from a lot more modes than the MSF, which have 90% of contributions coming from around 1% of the normal modes.

A similar pattern was observed in Chapter 7, where the scaling factor leading to both the best classification performance and the best Pearson correlation for the leave-one-out cross-validation was $\beta = e^{-0.5}$, lower than the MSF-agreement value of $\beta = e^{0.25}$. Interestingly, the peaking pattern of the AU-PR classification performance around this value resembles the peaking pattern observed for the selected MC-Sym 3D model of pri-miR-125a, model 61 in Figure 6.7. This optimal scaling factor for the prediction of MOR ligand activity leads to around 8% of the

normal modes contributing 90% of the entropy. Thus, it favors less higher-frequency modes than the optimal scaling factor for miR-125a, but more than the MSF.

The normal modes correspond to uncoupled classical harmonic oscillators, which do not exhibit chaotic behavior when considering the dynamics of the system. However, what the ENCoM-DynaSig-ML pipeline models is the change in Dynamical Signatures as a result of perturbations to the input structure, caused either by mutations or the binding of different ligands. These changes in the Dynamical Signatures generally fulfill the definition of chaos: slight changes in the starting conditions leading to very different outcomes [194].

Moreover, when looked at with this perspective, the most chaotic behavior is observed at very low scaling factors, for which the Entropic Signatures become almost constant. Since the signatures are standardized, the effect of a mutation on distant positions would be hard to predict based on just the observed changes in the Hessian matrix as a result of change in the surface complementarity term. At the opposite end, ordered behavior happens at very high scaling factors, which essentially result in the Entropic Signature being reduced to the square fluctuations from the single lowest-frequency normal mode. Since these low-frequency normal modes are robust to slight changes in the input configuration, it is relatively straightforward to predict the effect of a mutation on the Entropic Signature in that context: if the mutation leads to lower surface complementarity terms, the whole signature would increase in value, and *vice versa*.

Thus, we hypothesize that the performance peaks observed for miR-125a maturation and MOR activation at scaling factors lower than the value recapitulating MSF correspond to a goldilocks zone at the edge of chaos, providing a beneficial balance between contributions from slow modes and contributions from fast modes. Biological systems have been hypothesized to occupy the edge of chaos zone, as it is a zone towards which most adaptive systems converge [195, 196]. According to our hypothesis, predicting pri-miR-125a maturation efficiency benefits the most from chaos, followed by the prediction of μ-opioid receptor activation and finally VIM-2 catalytic efficiency. Maybe the biomolecules themselves occupy different zones in the chaos-order continuum. Indeed, RNAs have much more diverse conformational landscapes than proteins due to the presence of deep kinetic traps in their folding landscapes [4], and the problem of miR biogenesis is highly complex, with the Microprocessor having to recognize hundreds of substrates, all with different sequences [43]. GPCR activation also exhibits some complex behavior, with the possibility for signalling bias, receptor internalization, and the recognition of ligands with effects ranging from inverse agonism to super-agonism and allosteric modulation [197]. Finally, in many respects, VIM-2 lactamase is the studied biomolecule with the least chaotic behavior of the three. It is expressed by bacteria, and neither does it interact with partner biomolecules nor does it exhibit pronounced conformational changes as part of its function, the degradation of

β-lactam antibiotics. Therefore, it is evident that at the level of slow-timescales dynamics, VIM-2 exhibits less chaotic behavior than pri-miR-125a and MOR. This edge of chaos hypothesis for the success of our Entropic Signatures merits further investigation, which we propose as a final avenue for future research.

10

CONCLUSION

In this work, we have introduced the ENCoM-DynaSig-ML computational pipeline, the first variant effect predictor to consider biomolecular structural dynamics in a systematic way. We have also extended ENCoM to RNA molecules and have confirmed through parameter search that a single set of new optimized parameters allows for good performance on RNA, proteins and RNA-protein complexes. We have applied the ENCoM-DynaSig-ML pipeline to three diverse case studies of dynamics-function relationships, starting with datasets of experimental measures of function as a result of perturbations to the biomolecular systems (either through sequence variation or ligand binding).

The three case studies demonstrate the straightforward applicability of the ENCoM-DynaSig-ML pipeline, illustrate its ability to predict functional biomolecular properties of interest for biomolecular engineering and virtual screening, and showcase the remarkable biological insights obtainable from the mapping of learned LASSO coefficients to the structure of the studied biomolecule.

The good performance and interpretability of the pipeline is in part due to our introduction of Entropic Signatures, which scale the square fluctuations at every bead in the system by the vibrational entropy of the associated normal mode. In closing, we leave the reader to ponder our questions regarding Entropic Signatures.

It was formally shown that temperature has no effect on the standard mean square fluctuations arising from normal mode calculations, as we have discussed in Section 4.3.2. However, the Entropic Signatures depend on a thermodynamic scaling factor, which is inversely related to temperature. Moreover, specific ranges of values for the scaling factor produce Entropic Signatures decoupled from MSF, which lead to improved performance in two out of our three case studies. This raises the question, what is the effect that the Entropic Signatures capture in these cases? They could be truly capturing something akin to residue-level entropy, or they could be correcting for inaccuracies in the underlying pseudo-physical model. Maybe they are somehow correcting for the harmonic approximation of normal mode analysis. Or perhaps the higher entropy associated with lower scaling factors captures the effects of interaction with partner molecules: the Microprocessor complex in the case of microRNAs and the G protein/other intracellular interaction

partners in the case of GPCRs. Maybe none of these answers are close to the truth, which is too complex and chaotic for us to fully grasp.

Part III

APPENDIX

A

APPENDIX

Table A.1: Atom type assignations for the 4 standard ribonucleotides

Atom name	Atom type
Nucleobase atoms	
C2-C6, C8	Aromatic
N2, N4, N6	Donor
N3, N7	Acceptor
N9	Neutral
Sugar atoms	
C1′, C3′-C5′	Neutral-acceptor
C2′	Neutral
O4′	Acceptor
O2′	Hydrophilic
Phosphate atoms	
P	Neutral-acceptor
OP1-OP3, O5′, O3′	Acceptor
Specific atom types	
Adenine	
N1	Acceptor
Guanine	
N1	Donor
Cytidine	
N1	Neutral
N5	Acceptor
O2	Acceptor
Uridine	

Table A.1 (continued)

Atom name	Atom type
O2	Acceptor
O4	Acceptor

Table A.2: PDB codes for the protein B-factors benchmark

1LIT	1CNR	1GKY	2RHE
2CBA	5PTP	1AMM	1PTX
1EZM	4MT2	2END	2TGI
1ONC	1FUS	2HFT	1KNB
1PLC	1NAR	1LIS	3LZM
1MRJ	1AKY	2AYH	1BKF
4GCR	1NPK	1AHC	1RCF
1POC	1TML	5P21	1FXD
1CNV	4FGF	1PTF	1LST
1PDA	1FNC	1OSA	1NFP
1CYO	1AAC	1WHI	1HFC
2ERL	1FRD	1PPN	1JBC
1IAG	1REC	1CTJ	1RIE
1RIS	2IHL	1BPI	1CPN
1CUS	1IRO	2MHR	1POA
1GPR	1IAB	2CY3	1RRO
1AMP	1IFC	1DAD	7RSA
1ARB	1SNC	2CPL	3CHY
1RA9	2PHY	2RN2	1MJC
1UBI	3EBX	2MCM	1IGD

Table A.3: Pairs of conformations for the protein overlaps benchmark

PDB code A (chain)	PDB code B (chain)
2ZCO (A)	2ZCQ (A)
1EYM (A)	1J4R (A)
2EBF (X)	2EC5 (A)
1LYY (A)	1RE2 (A)

Table A.3 (continued)

PDB code A (chain)	PDB code B (chain)
3D97 (A)	3D97 (B)
2DDS (A)	2UYR (X)
2PLF (A)	2PMD (B)
1SBQ (B)	1U3G (A)
2QFB (H)	2QFD (C)
2OUS (B)	2OUU (A)
2DE2 (A)	2DE3 (B)
2ECK (A)	4AKE (A)
2DDB (A)	2EPF (D)
1LFH (A)	1LFI (A)
1M0Z (B)	1QYY (G)
1JYR (A)	1JYU (A)
3C6Q (B)	3C6Q (D)
1R1C (A)	1R1C (C)
1EF3 (B)	1XGD (A)
1JEJ (A)	1QKJ (A)
1T2W (A)	1T2W (C)
1Z15 (A)	1Z17 (A)
1RIF (A)	1RIF (B)
3DEO (A)	3DEP (A)
1HKA (A)	1RAO (A)
1EUT (A)	1W8N (A)
1K4K (A)	1K4M (B)
1UFP (A)	2EB8 (A)
1EX6 (B)	1GKY (A)
1GQZ (A)	2GKE (A)
1O0R (A)	1PZT (A)
2P0M (A)	2P0M (B)
1PU5 (A)	2AG4 (B)
1UHA (A)	1ULN (A)
1EIN (B)	1TIB (A)
3B8S (B)	3B9D (A)
1F1S (A)	1I8Q (A)

Table A.4: PDB codes and mutations for the protein ΔΔG benchmark

PDB code	Mutation	Experimental ΔΔG
1AJ3	D93A	-0.7
1AJ3	D93G	0.3
1AJ3	E25A	-0.1
1AJ3	H10A	-0.5
1AJ3	I23A	3.6
1AJ3	I84A	2.0
1AJ3	K26A	0.0
1AJ3	K47A	-0.4
1AJ3	K47G	0.5
1AJ3	K96A	0.4
1AJ3	K96G	1.3
1AJ3	L45A	0.2
1AJ3	L88A	2.8
1AJ3	L98A	3.8
1AJ3	N44G	0.4
1AJ3	Q86A	0.0
1AJ3	Q86G	1.3
1AJ3	T40A	-0.3
1AJ3	V30A	0.2
1AJ3	V42A	0.4
1AKY	Q48E	0.96
1APS	I75V	1.41
1BNI	I76T	2.64
1BNI	K27A	-0.44
1BNI	R59A	-0.64
1BTA	K22Q	0.79
1BTA	K60E	1.17
1BTA	R75L	-0.75
1BVC	H119F	0.68
1BVC	H24V	0.52
1BVC	H48Q	0.62
1C9O	E21A	0.29
1C9O	E50K	0.58

Table A.4 (continued)

PDB code	Mutation	Experimental ΔΔG
1C9O	F38W	-0.24
1CEY	A77G	0.31
1CEY	A80G	-0.43
1CEY	A88G	-0.04
1CEY	A99G	0.48
1CEY	D12A	-2.5
1CEY	D57A	-3.3
1CEY	F14N	-2.64
1CSE	V54A	1.58
1CUN	A126G	1.65
1CUN	A156G	1.45
1CUN	A173G	1.95
1CUN	A191G	1.35
1CUN	A212G	1.45
1CUN	F157L	1.85
1CUN	I128A	1.65
1CUN	I128V	2.85
1CUN	K152A	0.15
1CUN	K152G	1.45
1CUN	L196A	4.55
1CUN	L203A	4.05
1CUN	L214A	3.65
1CUN	M193A	2.65
1CUN	Q115G	1.15
1CUN	S201A	-0.15
1CUN	S201G	0.95
1DKT	K11A	-0.62
1DKT	M58L	0.23
1DKT	R71A	0.59
1DKT	S39A	0.6
1DKT	S9A	0.43
1DKT	V55A	0.73
1E65	A82G	3.11

Table A.4 (continued)

PDB code	Mutation	Experimental ΔΔG
1E65	H117G	2.18
1E65	I20T	2.39
1E65	I7S	3.44
1E65	L50V	0.36
1E65	V31T	1.08
1E65	V60G	3.11
1E65	V95T	-0.96
1EY0	D19F	1.28
1EY0	D21K	-1.1
1EY0	D77K	3.28
1EY0	G29V	3.11
1EY0	G55V	1.48
1EY0	G86F	1.99
1EY0	G96F	2.55
1EY0	G96V	3.74
1EY0	H124E	-0.46
1EY0	I139L	0.09
1EY0	I15M	0.15
1EY0	I72L	0.23
1EY0	I92M	1.75
1EY0	K24F	0.4
1EY0	K63Q	0.89
1EY0	K70E	0.3
1EY0	K78Q	0.15
1EY0	K9F	1.03
1EY0	L108V	3.81
1EY0	L125I	0.96
1EY0	L137V	1.42
1EY0	L14V	1.63
1EY0	L36V	3.58
1EY0	L37I	1.82
1EY0	L7V	1.15
1EY0	L89I	1.04

Table A.4 (continued)

PDB code	Mutation	Experimental ΔΔG
1EY0	M65F	1.62
1EY0	M65I	1.43
1EY0	R105C	2.55
1EY0	T13C	1.2
1EY0	T22I	0.61
1EY0	T33C	1.04
1EY0	T41I	-0.86
1EY0	T44C	0.04
1EY0	T82C	0.19
1EY0	T82I	-0.51
1EY0	V104I	-0.27
1EY0	V111I	0.74
1EY0	V111L	0.88
1EY0	V114I	0.15
1EY0	V23I	-0.03
1EY0	V23L	0.02
1EY0	V39I	-0.11
1EY0	V39L	0.9
1EY0	V51L	0.1
1EY0	V66I	0.76
1EY0	V74L	1.12
1EY0	Y27C	2.72
1EY0	Y54F	0.38
1FNA	I34V	0.11
1FNA	V50A	2.85
1FTG	A84G	1.9
1FTG	D126K	-0.81
1FTG	D65K	0.1
1FTG	D75K	-1.03
1FTG	E72K	-1.41
1FTG	I156V	3.16
1FTG	L6A	3.11
1FTG	N97A	0.58

Table A.4 (continued)

PDB code	Mutation	Experimental ΔΔG
1FTG	Q99A	-1.59
1FTG	S110A	0.73
1FTG	V160A	2.07
1G4I	F106A	1.23
1G4I	F22I	-1.43
1G4I	F22Y	-0.83
1G4I	H48A	1.93
1G4I	H48Q	0.49
1HME	G35H	0.33
1HMK	I55V	2.72
1HMK	I89V	0.86
1HMK	I95V	1.72
1HMK	L110A	0.35
1HMK	L12A	2.73
1HMK	L96A	1.75
1HMK	T29V	-2.26
1HMK	V27A	1.24
1HMK	V8A	0.83
1HMK	W60A	2.01
1HMK	Y103F	2.13
1HMS	F16S	3.98
1HMS	F4S	3.67
1HMS	F57S	2.43
1HMS	L66G	3.67
1HMS	R106T	2.84
1HMS	T40E	2.4
1IET	D60R	-0.14
1IFC	F68A	0.42
1IFC	F93A	2.37
1IFC	G65A	0.94
1IFC	L64G	2.26
1IFC	V60C	0.07
1IFC	V60N	0.83

Table A.4 (continued)

PDB code	Mutation	Experimental ΔΔG
1IFC	W6Y	0.87
1IGV	Y13F	1.08
1IHB	F37H	0.66
1IHB	F82Q	0.37
1IMQ	E31L	0.67
1IMQ	E41V	-0.89
1IMQ	V19L	1.82
1JIW	D10A	0.7
1JIW	W15F	2.3
1K9Q	L30Y	-0.27
1LNI	D79F	-2.73
1LNI	D79I	-2.85
1LNI	D79K	-2.35
1LNI	D79L	-2.65
1LNI	D79N	-1.46
1LNI	D79Y	-2.9
1LNI	H85Q	0.0
1LNI	Q94K	-0.56
1LNI	T16V	-0.3
1LNI	T56V	1.9
1LNI	V43T	0.5
1LNI	Y30F	-0.4
1LNI	Y55F	0.6
1LNI	Y80F	1.5
1LZ1	G105A	0.62
1MGR	Y54F	2.6
1MGR	Y84F	1.0
1MJC	F20L	0.31
1MJC	F20S	1.16
1MJC	F31S	1.03
1MJC	S52W	0.2
1MSI	D58N	0.2
1MSI	E25A	0.09

Table A.4 (continued)

PDB code	Mutation	Experimental ΔΔG
1MSI	R47A	0.74
1OIA	Y78F	0.0
1OIA	Y86T	2.9
1P2P	H48K	2.12
1QLP	A183I	-1.8
1QLP	A183V	-3.8
1QLP	A248F	-1.8
1QLP	A248I	-2.2
1QLP	A248L	-0.35
1QLP	A248V	-2.3
1QLP	A284I	0.0
1QLP	A284V	-0.8
1QLP	A31L	-0.9
1QLP	A70G	-1.6
1QLP	M374I	-2.3
1QLP	S330R	2.44
1QLP	S381A	-1.0
1QLP	V321I	-0.6
1QLP	V364L	0.3
1QLP	V55I	0.2
1QLP	W238F	-0.98
1QLP	Y160W	-1.18
1RG8	C16S	2.81
1RG8	L44F	-0.59
1RG8	N106G	-0.16
1RG8	V109I	0.05
1RIS	F60A	0.81
1RIS	I8A	3.56
1RIS	L21A	0.16
1RIS	L48A	0.21
1RIS	L75A	1.35
1RIS	L79A	3.91
1RIS	Y33A	-0.41

Table A.4 (continued)

PDB code	Mutation	Experimental ΔΔG
1RN1	D49A	-0.5
1RN1	G23A	1.2
1RN1	V16T	3.65
1RN1	V78S	4.73
1RN1	V78T	3.59
1RTB	A5S	0.27
1RTB	F46V	4.55
1RTB	V57A	2.85
1RTB	V57L	2.37
1RTB	V63A	2.03
1SHF	E107F	1.63
1SHF	E107H	0.99
1SHF	E107K	0.97
1SHF	E107L	3.02
1SHF	E107Y	2.4
1SHF	G128A	1.78
1SHF	I111A	2.84
1SHF	I111L	0.71
1SHF	S124D	2.02
1SHF	S124F	1.9
1SHF	S124G	1.68
1SHF	S124H	1.25
1SHF	S124L	0.37
1SHF	S124N	0.73
1SHF	V138M	0.52
1UZC	A61G	2.07
1UZC	E27A	0.62
1UZC	I44V	0.3
1UZC	Q38G	1.57
1YYJ	A20G	1.97
1YYJ	F61A	4.52
1YYJ	F65A	1.92
1YYJ	L3A	1.6

Table A.4 (continued)

PDB code	Mutation	Experimental ΔΔG
1ZNJ	F25D	-0.48
1ZNJ	H10E	-1.11
1ZNJ	H10T	-0.03
2A01	L141R	0.65
2A36	T22A	-0.4
2A36	T22F	-1.3
2A36	T22L	-0.5
2A36	T22N	-1.3
2DRI	V50E	3.5
2LZM	A146I	4.31
2LZM	F153C	3.11
2LZM	G51D	2.63
2LZM	I58Y	3.11
2NVH	C8S	3.74
2NVH	N7D	-0.09
2NVH	R4Q	0.57
2RN2	H62A	-0.44
2RN2	H83A	-0.07
3MBP	W10A	4.31
3SIL	A53L	-0.9
4LYZ	E35A	-1.24
5DFR	G121C	0.22
5DFR	G121H	0.56
5DFR	G67S	0.27
5DFR	G67T	0.62
5DFR	G95A	0.9
5DFR	I155T	2.53
5DFR	I2V	0.55
5DFR	N59T	0.05
5DFR	N59W	0.79
5DFR	V40H	2.76
5DFR	W30M	1.94
5PTI	A16T	1.7

Table A.4 (continued)

PDB code	Mutation	Experimental ΔΔG
5PTI	A16V	1.3
5PTI	Y35D	3.8

Table A.5: PDB codes for the RNA B-factors benchmark

472D	4U37	433D	405D
1CSL	1Q9A	480D	483D
5C5W	2VUQ	1I9X	1RNA
413D	280D	157D	3SZX
353D	255D	1ZEV	3GM7
1KD5	5VGW	4C40	402D
1RXB	259D	2V6W	2XSL
1DQH	7EAG	438D	5NXT
1MSY	2V7R	4U38	2OE6
1OSU	4E59		

Table A.6: Pairs of conformations for the RNA overlaps benchmark

PDB code A	PDB code B	Sequence cluster number
1J7T	1O9M	1
1J7T	2ET8	1
1J7T	3BNL	1
1J7T	4F8U	1
1LC4	3BNL	1
1LC4	4F8U	1
1MWL	1O9M	1
1MWL	2ET8	1
1MWL	3BNL	1
1MWL	4F8U	1
1O9M	1YRJ	1
1O9M	2BE0	1
1O9M	2BEE	1
1O9M	2ET3	1

Table A.6 (continued)

PDB code A	PDB code B	Sequence cluster number
1O9M	2F4S	1
1O9M	2F4T	1
1O9M	2F4U	1
1O9M	2O3X	1
1O9M	2PWT	1
1O9M	3BNL	1
1O9M	4F8U	1
1O9M	4F8V	1
1YRJ	2ET8	1
1YRJ	2F4S	1
1YRJ	2F4T	1
1YRJ	2F4U	1
1YRJ	2O3X	1
1YRJ	2PWT	1
1YRJ	3BNL	1
1YRJ	4F8U	1
1YRJ	4F8V	1
2BE0	2ET8	1
2BE0	2F4S	1
2BE0	2F4U	1
2BE0	2O3X	1
2BE0	3BNL	1
2BE0	4F8U	1
2BEE	2ET8	1
2BEE	2F4S	1
2BEE	2F4U	1
2BEE	2O3X	1
2BEE	3BNL	1
2BEE	4F8U	1
2ESJ	2F4S	1
2ESJ	3BNL	1
2ESJ	4F8U	1
2ET3	2ET8	1

Table A.6 (continued)

PDB code A	PDB code B	Sequence cluster number
2ET3	2F4S	1
2ET3	2F4T	1
2ET3	3BNL	1
2ET3	4F8U	1
2ET4	3BNL	1
2ET4	4F8U	1
2ET5	2ET8	1
2ET5	3BNL	1
2ET5	4F8U	1
2ET8	2F4S	1
2ET8	2F4T	1
2ET8	2F4U	1
2ET8	2O3X	1
2ET8	2PWT	1
2ET8	3BNL	1
2ET8	4F8U	1
2ET8	4F8V	1
2F4S	2F4T	1
2F4S	2PWT	1
2F4S	3BNL	1
2F4S	4F8U	1
2F4T	2F4U	1
2F4T	2O3X	1
2F4T	3BNL	1
2F4T	4F8U	1
2F4T	4F8V	1
2F4U	2PWT	1
2F4U	3BNL	1
2F4U	4F8U	1
2O3X	2PWT	1
2O3X	3BNL	1
2O3X	4F8U	1
2PWT	3BNL	1

Table A.6 (continued)

PDB code A	PDB code B	Sequence cluster number
2PWT	4F8U	1
3BNL	4F8U	1
3BNL	4F8V	1
3BNL	4P20	1
4F8U	4F8V	1
4F8U	4P20	1
1NLC	1XP7	2
1NLC	1XPF	2
1NLC	1Y3S	2
1NLC	1YXP	2
1NLC	1ZCI	2
1NLC	2B8S	2
1NLC	2FCX	2
1NLC	2FCY	2
1NLC	2FCZ	2
1NLC	2FD0	2
1NLC	2QEK	2
1NLC	3C44	2
1NLC	3DVV	2
1O3Z	1XP7	2
1O3Z	1XPF	2
1O3Z	1Y3S	2
1O3Z	1YXP	2
1O3Z	1ZCI	2
1O3Z	2B8S	2
1O3Z	2FCX	2
1O3Z	2FCY	2
1O3Z	2FCZ	2
1O3Z	2FD0	2
1O3Z	2QEK	2
1O3Z	3C44	2
1O3Z	3DVV	2
1XP7	2QEK	2

Table A.6 (continued)

PDB code A	PDB code B	Sequence cluster number
1XP7	3C44	2
1XP7	3DVV	2
1XP7	462D	2
1XPF	2QEK	2
1XPF	3C44	2
1XPF	3DVV	2
1XPF	462D	2
1Y3S	2QEK	2
1Y3S	3C44	2
1Y3S	3DVV	2
1Y3S	462D	2
1YXP	2QEK	2
1YXP	3C44	2
1YXP	3DVV	2
1YXP	462D	2
1ZCI	2FCX	2
1ZCI	2QEK	2
1ZCI	3C44	2
1ZCI	3DVV	2
1ZCI	462D	2
2B8S	2QEK	2
2B8S	3C44	2
2B8S	3DVV	2
2B8S	462D	2
2FCX	2QEK	2
2FCX	3C44	2
2FCX	3DVV	2
2FCX	462D	2
2FCY	2QEK	2
2FCY	3C44	2
2FCY	3DVV	2
2FCY	462D	2
2FCZ	2QEK	2

Table A.6 (continued)

PDB code A	PDB code B	Sequence cluster number
2FCZ	3C44	2
2FCZ	3DVV	2
2FCZ	462D	2
2FD0	2QEK	2
2FD0	3C44	2
2FD0	3DVV	2
2FD0	462D	2
2QEK	3C44	2
2QEK	3DVV	2
2QEK	462D	2
3C44	462D	2
3DVV	462D	2
1XPE	2OIJ	3
1XPE	2OIY	3
1XPE	2OJ0	3
1XPE	3FAR	3
2B8R	2OIJ	3
2B8R	2OIY	3
2B8R	2OJ0	3
2B8R	3FAR	3
1ZX7	1ZZ5	4
1ZX7	2A04	4
2FQN	2G5K	5
2FQN	5XZ1	5
2G5K	2O3W	5
2G5K	5XZ1	5
2O3W	5XZ1	5
2GPM	439D	6
2NOK	2PN4	7
3BNQ	3BNR	8
3BNQ	3BNS	8
3CW5	5L4O	9
3CW6	5L4O	9

Table A.6 (continued)

PDB code A	PDB code B	Sequence cluster number
3GCA	6VUH	10
3LoU	6Y3G	11
3OWI	3OWZ	12
3OWW	3OWZ	12
3OWZ	3OXo	12
3OWZ	3OXB	12
3OWZ	3OXD	12
3OWZ	3OXE	12
3OWZ	3OXJ	12
3OWZ	3OXM	12
3TDo	3TD1	13
3WRU	4PDQ	14
3WRU	6JBG	14
4GPY	4PDQ	14
4GPY	6JBG	14
4PDQ	6JBG	14
4K31	4K32	15
4L81	4OQU	16
4MSB	6Z18	17
4MSR	6WY3	17
4MSR	6Z18	17
5TDK	6Z18	17
4P3S	4P3T	18
4P3U	4P43	19
4RZD	6XKN	20
6XKN	6XKO	20
4TZX	5E54	21
4TZX	5SWD	21
4TZY	5E54	21
4TZY	5SWD	21
4XNR	5E54	21
4XNR	5SWD	21
5E54	5SWE	21

Table A.6 (continued)

PDB code A	PDB code B	Sequence cluster number
5E54	5UZA	21
5E54	6VWT	21
5E54	6VWV	21
5SWD	5SWE	21
5SWD	5UZA	21
5SWD	6VWT	21
5SWD	6VWV	21
5ZEG	5ZEI	22
5ZEG	5ZEJ	22
5ZEG	5ZEM	22
6C8D	6CAB	23
6E80	6E81	24
6E80	6E82	24
6E80	6E84	24
7EOI	7EOK	25

Table A.7: PDB codes for the RNA NCO benchmark

2LUB	1F5G	2KYD	1AFX
6U79	2M21	1QET	2N7M
5KQE	6VZC	2FDT	1JU1
6PK9	1K6G	1B36	1OW9
2JXS	6K84	2ADT	2RN1
1K4A	6XWJ	3PHP	2MI0
1TBK	6N8F	1GUC	1BN0
1LMV	1FHK	2NBZ	2GIP
2LX1	2KVN	6BG9	2Y95
2LP9	2JYH	1HLX	1E4P
2K3Z	2MXJ	1EBR	5V17
6XWW	1R2P	1N66	1ATW
2L5Z	2LPS	2JXQ	2M23
2L8F	1IDV	2LBL	2N3Q
2M24	2IXY	2N6X	2IRO

Table A.7 (continued)

1Z31	2NC1	4A4T	2N4L
2F88	2LC8	2NCI	1YNC
1TJZ	1T4X	1MFJ	2D19
2RLU	2LUN	1F7F	2LPA
6HYK	2AHT	2M22	1RNG
1ATO	2KHY	1ZIF	1WKS
1XHP	2O33	1A60	2GM0
1Q75	2LV0	1F85	5LSN
2M4W	1ZIG	2N6S	6GE1
5A17	2KY1	1YNE	1IE1
2MXL	1F7G	2FEY	1F84
1QES	2G1W	1TXS	1MFY
5VH7	6BY4	5V2R	2EUY
1AQO	1VOP	1QWA	2HNS
1A3M	2NBX	5N5C	1R7W
2NC0	6N8I	2KPC	2JXV
2KRQ	2GRW	1M82	2MFD
1P5O	5KMZ	2M18	6AAS
2L1F	2RQJ	2PCV	2LK3
1LDZ	1YSV	2MXK	1MT4
1JU7	2LJJ	1P5M	2RVO
2IXZ	2KF0	1M5L	1YMO
7DD4	2MNC	2LQZ	1N8X
2KUW	2K5Z	2LPT	2OJ8
2LDL	2MTJ	5IEM	1P5N
5VH8	1NC0	1YNG	2D17
2JSE	2KZL	1CQL	6VAR
2PCW	2KUU	1R7Z	6W3M
1OSW	1SCL	1SY4	1NBR
2IRN	4A4S	2KE6	5A18
1Z2J	2GV4	1K4B	1BVJ
2KD8	2P89	1K6H	2KRP
6VA1	2JTP	2N8V	2MQT
1F5H	1QWB	5UZT	2L3E

Table A.7 (continued)

1ATV	1IKD	1I3X	2GVO
2JR4	28SP	1FYO	1TLR
2QH4	6XXB	2JYF	2KOC
2LHP	2KEZ	1SLP	2K41
1YLG	1MFK	1KKA	1SYZ
2H49	7K4L	4A4U	2D18
2MEQ	1EBQ	2KPD	5KH8
1OQ0	1NA2	2KXZ	5WQ1
7JU1	2F87	2N6T	6BY5
2F4X	1ELH	2K96	1YN1
1JP0	2JWV	2L2J	1RRR
2K66	2M12	2N7X	1ESY
1KKS	2LU0	1A9L	2KPV
2K65	1QC8	5V16	2JYJ
2JYM	1K5I	2K95	2LAC
1S9S	6XXA	1BGZ	2XEB
2KBP	1ROQ	2N6W	2LBK
2GV3	1ZC5	1MNX	1JTJ
1JUR	1Z30	2M8K	2M57
17RA	1A51	2N2P	2GIO
1U3K	2B7G	1JOX	1HWQ
1ZIH	1UUU	1E95	2KRL
2HEM	2LDT	2QH3	2RPT
1S34	6MXQ	5UF3	2L6I
2HUA	2KUR	1D0U	2M5U
2EVY	2QH2	1JO7	2KY2
1IE2	2N2O	7LVA	1LC6
6NOA	2OJ7	1EBS	1ANR
2LBJ	1PJY	2ES5	2KY0
2NBY			

Table A.8: PDB codes for the RNA-protein NCO benchmark

7Q4L	7ZEX	6WLH	6HPJ

Table A.8 (continued)

7ACS	6SNJ	7ACT	6SDW
6SDY	6SO9	6G99	5N8M
5X3Z	5MPG	5MPL	2N8L
2N7C	2N30	2N82	2MQ0

Table A.9: Correspondence between Sybyl and Sobolev atom types

Sybyl number	Sybyl name	Condition	Sobolev type
2	Br	Any	IV (Hydrophobic)
3	C.1	Bound to at least 2 atoms of Sobolev type II or III	VI (Neutral)
		Bound to only 1 atom of Sobolev type II or III	VII (Neutral-donor)
		Any other condition	IV (Hydrophobic)
4	C.2	Bound to at least 2 atoms of Sobolev type II or III	VI (Neutral)
		Bound to only 1 atom of Sobolev type II or III	VII (Neutral-donor)
		Any other condition	IV (Hydrophobic)
5	C.3	Bound to at least 2 atom of Sobolev type II or III	VI (Neutral)
		Bound to only 1 atom of Sobolev type II or III	VII (Neutral-donor)
		Any other condition	IV (Hydrophobic)
6	C.ar	Any	V (Aromatic)
7	C.cat	Any	VI (Neutral)
9	Cl	Any	IV (Hydrophobic)
11	F	Any	VI (Neutral)
19	N.1	Any	II (Acceptor)
20	N.2	Bound to only 1 hydrogen	I (Hydrophillic)
		Any other condition	II (Acceptor)
21	N.3	Bound to only 1 hydrogen	III (Donor)
		Bound to 2 hydrogens	I (Hydrophillic)
		Any other condition	VI (Neutral)

Table A.9 (continued)

Sybyl number	Sybyl name	Condition	Sobolev type
22	N.4	Any	III (Donor)
23	N.ar	Any	V (Aromatic)
24	N.p13	Any	III (Donor)
26	O.2	Any	II (Acceptor)
27	O.3	Bound to at least 1 hydrogen	I (Hydrophillic)
		Any other condition	II (Acceptor)
28	O.co2	Any	I (Hydrophillic)
29	O.spc	Any	I (Hydrophillic)
30	O.t3p	Any	I (Hydrophillic)
32	S.2	Any	VI (Neutral)
33	S.3	Any	VI (Neutral)
34	S.o	Any	VI (Neutral)
35	S.o2	Any	VI (Neutral)

BIBLIOGRAPHY

[1] Francis Crick. "Central Dogma of Molecular Biology." In: *Nature* 227.5258 (Aug. 1970), pp. 561–563. DOI: 10.1038/227561a0. URL: https://doi.org/10.1038/227561a0.

[2] Christian B Anfinsen. "Principles that govern the folding of protein chains." In: *Science* 181.4096 (1973), pp. 223–230.

[3] Ken A Dill and Hue Sun Chan. "From Levinthal to pathways to funnels." In: *Nature structural biology* 4.1 (1997), pp. 10–19.

[4] Hashim M Al-Hashimi and Nils G Walter. "RNA dynamics: it is about time." In: *Current Opinion in Structural Biology* 18.3 (2008), pp. 321–329. DOI: 10.1016/j.sbi.2008.04.004. URL: https://doi.org/10.1016%2Fj.sbi.2008.04.004.

[5] Antonio Deiana, Sergio Forcelloni, Alessandro Porrello, and Andrea Giansanti. "Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell." In: *PloS one* 14.8 (2019), e0217889.

[6] Angelika Andrzejewska, Małgorzata Zawadzka, and Katarzyna Pachulska-Wieczorek. "On the way to understanding the interplay between the RNA structure and functions in cells: A genome-wide perspective." In: *International Journal of Molecular Sciences* 21.18 (2020), p. 6770.

[7] Naomi K. Fox, Steven E. Brenner, and John-Marc Chandonia. "SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures." In: *Nucleic Acids Research* 42.D1 (Dec. 2013), pp. D304–D309. DOI: 10.1093/nar/gkt1240. URL: https://doi.org/10.1093/nar/gkt1240.

[8] Ioanna Kalvari et al. "Rfam 14: expanded coverage of metagenomic, viral and microRNA families." In: *Nucleic Acids Research* 49.D1 (Nov. 2020), pp. D192–D200. DOI: 10.1093/nar/gkaa1047. URL: https://doi.org/10.1093/nar/gkaa1047.

[9] Gordon G. Hammes. "Multiple Conformational Changes in Enzyme Catalysis." In: *Biochemistry* 41.26 (June 2002), pp. 8221–8228. DOI: 10.1021/bi0260839. URL: https://doi.org/10.1021/bi0260839.

[10] Kausik Si and Eric R. Kandel. "The Role of Functional Prion-Like Proteins in the Persistence of Memory." In: *Cold Spring Harbor Perspectives in Biology* 8.4 (Apr. 2016), a021774. DOI: 10.1101/cshperspect.a021774. URL: https://doi.org/10.1101/cshperspect.a021774.

[11] Mackenzie W Turvey, Kristin N Gabriel, Wonbae Lee, Jeffrey J Taulbee, Joshua K Kim, Silu Chen, Calvin J Lau, Rebecca E Kattan, Jenifer T Pham, Sudipta Majumdar, et al. "Single-molecule Taq DNA polymerase dynamics." In: *Science advances* 8.10 (2022), eabl3522.

[12] Buyong Ma and Ruth Nussinov. "Enzyme dynamics point to stepwise conformational selection in catalysis." In: *Current opinion in chemical biology* 14.5 (2010), pp. 652–659.

[13] Wenwen Fang and David P. Bartel. "The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes." In: *Molecular Cell* 60.1 (2015), pp. 131–145. DOI: 10.1016/j.molcel.2015.08.015. URL: https://doi.org/10.1016%2Fj.molcel.2015.08.015.

[14] Paul Dallaire, Huiping Tan, Keith Szulwach, Christopher Ma, Peng Jin, and François Major. "Structural dynamics control the MicroRNA maturation pathway." In: *Nucleic Acids Res* (2016), gkw793. DOI: 10.1093/nar/gkw793. URL: https://doi.org/10.1093%2Fnar%2Fgkw793.

[15] Julia Winter, Stephanie Jung, Sarina Keller, Richard I. Gregory, and Sven Diederichs. "Many roads to maturity: microRNA biogenesis pathways and their regulation." In: *Nat Cell Biol* 11.3 (2009), pp. 228–234. DOI: 10.1038/ncb0309-228. URL: https://doi.org/10.1038%2Fncb0309-228.

[16] Stefan Lutz. "Beyond directed evolution—semi-rational protein engineering and design." In: *Current Opinion in Biotechnology* 21.6 (2010), pp. 734–743. DOI: 10.1016/j.copbio.2010.08.011. URL: https://doi.org/10.1016/j.copbio.2010.08.011.

[17] Stanislav Mazurenko, Zbynek Prokop, and Jiri Damborsky. "Machine Learning in Enzyme Engineering." In: *ACS Catalysis* 10.2 (Dec. 2019), pp. 1210–1223. DOI: 10.1021/acscatal.9b04321. URL: https://doi.org/10.1021/acscatal.9b04321.

[18] Scott A. Hollingsworth and Ron O. Dror. "Molecular Dynamics Simulation for All." In: *Neuron* 99.6 (Sept. 2018), pp. 1129–1143. DOI: 10.1016/j.neuron.2018.08.011. URL: https://doi.org/10.1016/j.neuron.2018.08.011.

[19] Andreas Vitalis and Rohit V Pappu. "Methods for Monte Carlo simulations of biomacromolecules." In: *Annual reports in computational chemistry* 5 (2009), pp. 49–76.

[20] David A Case. "Normal mode analysis of protein dynamics." In: *Current Opinion in Structural Biology* 4.2 (1994), pp. 285–290.

[21] Soumil Y Joshi and Sanket A Deshmukh. "A review of advancements in coarse-grained molecular dynamics simulations." In: *Molecular Simulation* 47.10-11 (2021), pp. 786–803.

[22] Zhenqin Li and Harold A Scheraga. "Monte Carlo-minimization approach to the multiple-minima problem in protein folding." In: *Proceedings of the National Academy of Sciences* 84.19 (1987), pp. 6611–6615.

[23] Osni Marques and Yves-Henri Sanejouand. "Hinge-bending motion in citrate synthase arising from normal mode calculations." In: *Proteins* 23.4 (1995), pp. 557–560. DOI: 10.1002/prot.340230410. URL: https://doi.org/10.1002%2Fprot.340230410.

[24] Monique M. Tirion. "Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis." In: *Physical Review Letters* 77.9 (1996), pp. 1905–1908. DOI: 10.1103/physrevlett.77.1905. URL: https://doi.org/10.1103/physrevlett.77.1905.

[25] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential." In: *Folding and Design* 2.3 (1997), pp. 173–181. DOI: 10.1016/s1359-0278(97)00024-2. URL: https://doi.org/10.1016%2Fs1359-0278%2897%2900024-2.

[26] A.R. Atilgan, S.R. Durell, R.L. Jernigan, M.C. Demirel, O. Keskin, and I. Bahar. "Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model." In: *Biophysical Journal* 80.1 (2001), pp. 505–515. DOI: 10.1016/s0006-3495(01)76033-x. URL: https://doi.org/10.1016%2Fs0006-3495%2801%2976033-x.

[27] Vincent Frappier and Rafael J. Najmanovich. "A Coarse-Grained Elastic Network Atom Contact Model and Its Use in the Simulation of Protein Dynamics and the Prediction of the Effect of Mutations." In: *PLoS Comput Biol* 10.4 (2014). Ed. by Alexander Donald MacKerell, e1003569. DOI: 10.1371/journal.pcbi.1003569. URL: https://doi.org/10.1371%2Fjournal.pcbi.1003569.

[28] Vincent Frappier and Rafael Najmanovich. "Vibrational entropy differences between mesophile and thermophile proteins and their use in protein engineering." In: *Protein Science* 24.4 (2014), pp. 474–483. DOI: 10.1002/pro.2592. URL: https://doi.org/10.1002%2Fpro.2592.

[29] Natália Teruel, Olivier Mailhot, and Rafael J. Najmanovich. "Modelling conformational state dynamics and its role on infection for SARS-CoV-2 Spike protein variants." In: *PLoS Comput Biol* 17.8 (2021). Ed. by Roland L. Dunbrack, e1009286. DOI: 10.1371/journal.pcbi.1009286. URL: https://doi.org/10.1371%2Fjournal.pcbi.1009286.

[30] Ivet Bahar, Timothy R. Lezon, Lee-Wei Yang, and Eran Eyal. "Global Dynamics of Proteins: Bridging Between Structure and Function." In: *Annu. Rev. Biophys.* 39.1 (2010), pp. 23–42. DOI: 10.1146/annurev.biophys.093008.131258. URL: https://doi.org/10.1146%2Fannurev.biophys.093008.131258.

[31] Olivier Mailhot and Rafael Najmanovich. "The NRGTEN Python package: an extensible toolkit for coarse-grained normal mode analysis of proteins, nucleic acids, small molecules and their complexes." In: *Bioinformatics* 37.19 (2021). Ed. by Jinbo Xu, pp. 3369–3371. DOI: 10.1093/bioinformatics/btab189. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbtab189.

[32] Olivier Mailhot, François Major, and Rafael Najmanovich. "The DynaSig-ML Python package: automated learning of biomolecular dynamics-function relationships." In: *bioRxiv* (2022). DOI: 10.1101/2022.07.06.499058. URL: https://doi.org/10.1101/2022.07.06.499058.

[33] Victor Ambros. "The functions of animal microRNAs." In: *Nature* 431.7006 (2004), pp. 350–355. DOI: 10.1038/nature02871. URL: https://doi.org/10.1038%2Fnature02871.

[34] Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. "miRBase: from microRNA sequences to function." In: *Nucleic Acids Research* 47.D1 (Nov. 2018), pp. D155–D162. DOI: 10.1093/nar/gky1141. URL: https://doi.org/10.1093/nar/gky1141.

[35] Robin C. Friedman, Kyle Kai-How Farh, Christopher B. Burge, and David P. Bartel. "Most mammalian mRNAs are conserved targets of microRNAs." In: *Genome Research* 19.1 (Oct. 2008), pp. 92–105. DOI: 10.1101/gr.082701.108. URL: https://doi.org/10.1101/gr.082701.108.

[36] Shankar Mukherji, Margaret S Ebert, Grace X Y Zheng, John S Tsang, Phillip A Sharp, and Alexander van Oudenaarden. "MicroRNAs can generate thresholds in target gene expression." In: *Nature Genetics* 43.9 (Aug. 2011), pp. 854–859. DOI: 10.1038/ng.905. URL: https://doi.org/10.1038/ng.905.

[37] Simon P R Romaine, Maciej Tomaszewski, Gianluigi Condorelli, and Nilesh J Samani. "MicroRNAs in cardiovascular disease: an introduction for clinicians." In: *Heart* 101.12 (Mar. 2015), pp. 921–928. DOI: 10.1136/heartjnl-2013-305402. URL: https://doi.org/10.1136/heartjnl-2013-305402.

[38] Camille A. Juźwik, Sienna S. Drake, Yang Zhang, Nicolas Paradis-Isler, Alexandra Sylvester, Alexandre Amar-Zifkin, Chelsea Douglas, Barbara Morquette, Craig S. Moore, and Alyson E. Fournier. "microRNA dysregulation in neurodegenerative diseases: A systematic review." In: *Progress in Neurobiology* 182 (Nov. 2019), p. 101664. DOI: 10.1016/j.pneurobio.2019.101664. URL: https://doi.org/10.1016/j.pneurobio.2019.101664.

[39] Shuibin Lin and Richard I. Gregory. "MicroRNA biogenesis pathways in cancer." In: *Nature Reviews Cancer* 15.6 (May 2015), pp. 321–333. DOI: 10.1038/nrc3932. URL: https://doi.org/10.1038/nrc3932.

[40] Minju Ha and V. Narry Kim. "Regulation of microRNA biogenesis." In: *Nature Reviews Molecular Cell Biology* 15.8 (July 2014), pp. 509–524. DOI: 10.1038/nrm3838. URL: https://doi.org/10.1038/nrm3838.

[41] David P Bartel. "MicroRNAs." In: *Cell* 116.2 (Jan. 2004), pp. 281–297. DOI: 10.1016/s0092-8674(04)00045-5. URL: https://doi.org/10.1016/s0092-8674(04)00045-5.

[42] Andrea Tanzer and Peter F Stadler. "Molecular Evolution of a MicroRNA Cluster." In: *Journal of Molecular Biology* 339.2 (May 2004), pp. 327–335. DOI: 10.1016/j.jmb.2004.03.065. URL: https://doi.org/10.1016/j.jmb.2004.03.065.

[43] Kijun Kim, S. Chan Baek, Young-Yoon Lee, Carolien Bastiaanssen, Jeesoo Kim, Haedong Kim, and V. Narry Kim. "A quantitative map of human primary microRNA processing sites." In: *Molecular Cell* 81.16 (Aug. 2021), 3422–3439.e11. DOI: 10.1016/j.molcel.2021.07.002. URL: https://doi.org/10.1016/j.molcel.2021.07.002.

[44] Tomoko Kawamata and Yukihide Tomari. "Making RISC." In: *Trends in Biochemical Sciences* 35.7 (July 2010), pp. 368–376. DOI: 10.1016/j.tibs.2010.03.009. URL: https://doi.org/10.1016/j.tibs.2010.03.009.

[45] Ranhui Duan, ChangHui Pak, and Peng Jin. "Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA." In: *Human Molecular Genetics* 16.9 (2007), pp. 1124–1131. DOI: 10.1093/hmg/ddm062. URL: https://doi.org/10.1093%2Fhmg%2Fddm062.

[46] Alexander S. Hauser, Misty M. Attwood, Mathias Rask-Andersen, Helgi B. Schiöth, and David E. Gloriam. "Trends in GPCR drug discovery: new agents, targets and indications." In: *Nature Reviews Drug Discovery* 16.12 (Oct. 2017), pp. 829–842. DOI: 10.1038/nrd.2017.178. URL: https://doi.org/10.1038/nrd.2017.178.

[47] Daniel Wacker, Raymond C. Stevens, and Bryan L. Roth. "How Ligands Illuminate GPCR Molecular Pharmacology." In: *Cell* 170.3 (July 2017), pp. 414–427. DOI: 10.1016/j.cell.2017.07.009. URL: https://doi.org/10.1016/j.cell.2017.07.009.

[48] Caroline Wilde, Jakob Mitgau, Tomáš Suchý, Torsten Schöneberg, and Ines Liebscher. "Translating the force—mechano-sensing GPCRs." In: *American Journal of Physiology-Cell Physiology* 322.6 (June 2022), pp. C1047–C1060. DOI: 10.1152/ajpcell.00465.2021. URL: https://doi.org/10.1152/ajpcell.00465.2021.

[49] Aashish Manglik, Andrew C. Kruse, Tong Sun Kobilka, Foon Sun Thian, Jesper M. Mathiesen, Roger K. Sunahara, Leonardo Pardo, William I. Weis, Brian K. Kobilka, and Sébastien Granier. "Crystal structure of the ¯-opioid receptor bound to a morphinan antagonist." In: *Nature* 485.7398 (Mar. 2012),

pp. 321–326. DOI: 10.1038/nature10954. URL: https://doi.org/10.1038/nature10954.

[50] Weijiao Huang et al. "Structural insights into ¯-opioid receptor activation." In: *Nature* 524.7565 (Aug. 2015), pp. 315–321. DOI: 10.1038/nature14886. URL: https://doi.org/10.1038/nature14886.

[51] Naomi R. Latorraca, A. J. Venkatakrishnan, and Ron O. Dror. "GPCR Dynamics: Structures in Motion." In: *Chemical Reviews* 117.1 (Sept. 2016), pp. 139–155. DOI: 10.1021/acs.chemrev.6b00177. URL: https://doi.org/10.1021/acs.chemrev.6b00177.

[52] Qingtong Zhou et al. "Common activation mechanism of class A GPCRs." In: *eLife* 8 (Dec. 2019). DOI: 10.7554/elife.50279. URL: https://doi.org/10.7554/elife.50279.

[53] Naomi R. Latorraca, Jason K. Wang, Brian Bauer, Raphael J. L. Townshend, Scott A. Hollingsworth, Julia E. Olivieri, H. Eric Xu, Martha E. Sommer, and Ron O. Dror. "Molecular mechanism of GPCR-mediated arrestin activation." In: *Nature* 557.7705 (May 2018), pp. 452–456. DOI: 10.1038/s41586-018-0077-3. URL: https://doi.org/10.1038/s41586-018-0077-3.

[54] X Edward Zhou, Karsten Melcher, and H Eric Xu. "Understanding the GPCR biased signaling through G protein and arrestin complex structures." In: *Current Opinion in Structural Biology* 45 (Aug. 2017), pp. 150–159. DOI: 10.1016/j.sbi.2017.05.004. URL: https://doi.org/10.1016/j.sbi.2017.05.004.

[55] N. Gautam, G.B. Downes, K. Yan, and O. Kisselev. "The G-Protein $\beta\gamma$ Complex." In: *Cellular Signalling* 10.7 (July 1998), pp. 447–455. DOI: 10.1016/s0898-6568(98)00006-0. URL: https://doi.org/10.1016/s0898-6568(98)00006-0.

[56] Daniel Hilger, Matthieu Masureel, and Brian K. Kobilka. "Structure and dynamics of GPCR signaling complexes." In: *Nature Structural & Molecular Biology* 25.1 (Jan. 2018), pp. 4–12. DOI: 10.1038/s41594-017-0011-7. URL: https://doi.org/10.1038/s41594-017-0011-7.

[57] A. Gaulton et al. "ChEMBL: a large-scale bioactivity database for drug discovery." In: *Nucleic Acids Research* 40.D1 (Sept. 2011), pp. D1100–D1107. DOI: 10.1093/nar/gkr777. URL: https://doi.org/10.1093/nar/gkr777.

[58] Junya Okude et al. "Identification of a Conformational Equilibrium That Determines the Efficacy and Functional Selectivity of the μ-Opioid Receptor." In: *Angewandte Chemie International Edition* 54.52 (Nov. 2015), pp. 15771–15776. DOI: 10.1002/anie.201508794. URL: https://doi.org/10.1002/anie.201508794.

[59] Rémy Sounier, Camille Mas, Jan Steyaert, Toon Laeremans, Aashish Manglik, Weijiao Huang, Brian K. Kobilka, Héléne Déméné, and Sébastien Granier. "Propagation of conformational changes during μ-opioid receptor activation." In: *Nature* 524.7565 (Aug. 2015), pp. 375–378. DOI: 10.1038/nature14680. URL: https://doi.org/10.1038/nature14680.

[60] Yun Hao and Nicholas P Tatonetti. "Predicting G protein-coupled receptor downstream signaling by tissue expression." In: *Bioinformatics* 32.22 (2016), pp. 3435–3443.

[61] C Harrison and JR Traynor. "The [35S] GTPγS binding assay: approaches and applications in pharmacology." In: *Life sciences* 74.4 (2003), pp. 489–508.

[62] Humphrey P Rang, Maureen M Dale, James M Ritter, Rod J Flower, and Graeme Henderson. *Rang & Dale's pharmacology*. Elsevier Health Sciences, 2011, pp. 9–10.

[63] Tatsuhiro Onogi, Masabumi Minami, Yoshikazu Katao, Takayuki Nakagawa, Yasuhide Aoki, Takashi Toya, Seishi Katsumata, and Masamichi Satoh. "DAMGO, a μ-opioid receptor selective agonist, distinguishes between μ- and δ-opioid receptors around their first extracellular loops." In: *FEBS letters* 357.1 (1995), pp. 93–97.

[64] Lapo Mughini-Gras, Alejandro Dorado-García, Engeline van Duijkeren, Gerrita van den Bunt, Cindy M Dierikx, Marc JM Bonten, Martin CJ Bootsma, Heike Schmitt, Tine Hald, Eric G Evers, et al. "Attributable sources of community-acquired carriage of Escherichia coli containing β-lactam antibiotic resistance genes: a population-based modelling study." In: *The Lancet Planetary Health* 3.8 (2019), e357–e369.

[65] Fahd K. Majiduddin, Isabel C. Materon, and Timothy G. Palzkill. "Molecular analysis of beta-lactamase structure and function." In: *International Journal of Medical Microbiology* 292.2 (2002), pp. 127–137. DOI: 10.1078/1438-4221-00198. URL: https://doi.org/10.1078/1438-4221-00198.

[66] Tony Christopeit, Ke-Wu Yang, Shao-Kang Yang, and Hanna-Kirsti S. Leiros. "The structure of the metallo-β-lactamase VIM-2 in complex with a triazolylthioacetamide inhibitor." In: *Acta Crystallographica Section F Structural Biology Communications* 72.11 (Oct. 2016), pp. 813–819. DOI: 10.1107/s2053230x16016113. URL: https://doi.org/10.1107/s2053230x16016113.

[67] Carola Mauri, Alberto Enrico Maraolo, Stefano Di Bella, Francesco Luzzaro, and Luigi Principe. "The Revival of Aztreonam in Combination with Avibactam against Metallo-β-Lactamase-Producing Gram-Negatives: A Systematic Review of In Vitro Studies and Clinical Cases." In: *Antibiotics* 10.8 (Aug. 2021), p. 1012. DOI: 10.3390/antibiotics10081012. URL: https://doi.org/10.3390/antibiotics10081012.

[68] Mikhail V Edelstein, Elena N Skleenova, Oksana V Shevchenko, Jimson W D'souza, Dmitry V Tapalski, Ilya S Azizov, Marina V Sukhorukova, Roman A Pavlukov, Roman S Kozlov, Mark A Toleman, et al. "Spread of extensively resistant VIM-2-positive ST235 Pseudomonas aeruginosa in Belarus, Kazakhstan, and Russia: a longitudinal epidemiological and clinical study." In: *The Lancet infectious diseases* 13.10 (2013), pp. 867–876.

[69] Jürgen Brem, Sander S van Berkel, David Zollman, Sook Y Lee, Opher Gileadi, Peter J McHugh, Timothy R Walsh, Michael A McDonough, and Christopher J Schofield. "Structural basis of metallo-β-lactamase inhibition by captopril stereoisomers." In: *Antimicrobial agents and chemotherapy* 60.1 (2016), pp. 142–150.

[70] John Z Chen, Douglas M Fowler, and Nobuhiko Tokuriki. "Comprehensive exploration of the translocation, stability and substrate recognition requirements in VIM-2 lactamase." In: *Elife* 9 (2020), e56707.

[71] Eleonora Gianquinto, Donatella Tondi, Giulia D'Arrigo, Loretta Lazzarato, and Francesca Spyrakis. "Can We Exploit β-Lactamases Intrinsic Dynamics for Designing More Effective Inhibitors?" In: *Antibiotics* 9.11 (2020), p. 833.

[72] Peter W. Rose et al. "The RCSB Protein Data Bank: views of structural biology for basic and applied research and education." In: *Nucleic Acids Research* 43.D1 (2014), pp. D345–D356. DOI: 10.1093/nar/gku1214. URL: https://doi.org/10.1093%2Fnar%2Fgku1214.

[73] Gale Thodes. *Crystallography Made Crystal Clear*. Elsevier, 1993. DOI: 10.1016/c2009-0-21368-7. URL: https://doi.org/10.1016%2Fc2009-0-21368-7.

[74] Reza Soheilifard, Dmitrii E Makarov, and Gregory J Rodin. "Critical evaluation of simple network models of protein dynamics and their comparison with crystallographic B-factors." In: *Physical Biology* 5.2 (2008), p. 026008. ISSN: 1478-3975. DOI: 10.1088/1478-3975/5/2/026008. URL: http://stacks.iop.org/1478-3975/5/i=2/a=026008?key=crossref.a294a6be7d2d86f3a99cfbc8fea36324.

[75] Takayuki Amemiya, Ryotaro Koike, Akinori Kidera, and Motonori Ota. "PSCDB: a database for protein structural change upon ligand binding." In: *Nucleic acids research* 40.D1 (2012), pp. D554–D558.

[76] W. Friedrich, P. Knipping, and M. Laue. "Interferenzerscheinungen bei Röntgenstrahlen." In: *Annalen der Physik* 346.10 (1913), pp. 971–988. DOI: 10.1002/andp.19133461004. URL: https://doi.org/10.1002/andp.19133461004.

[77] John C Kendrew, G Bodo, Howard M Dintzis, RG Parrish, Harold Wyckoff, and David C Phillips. "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis." In: *Nature* 181.4610 (1958), pp. 662–666.

[78] Veloorvalappil N Jisha, Robinson B Smitha, Sailas Benjamin, et al. "An overview on the crystal toxins from Bacillus thuringiensis." In: *Advances in Microbiology* 3.05 (2013), p. 462.

[79] Oliviero Carugo and Patrick Argos. "Protein—protein crystal-packing contacts." In: *Protein science* 6.10 (1997), pp. 2261–2263.

[80] Marat M Yusupov, Gulnara Zh Yusupova, Albion Baucom, Kate Lieberman, Thomas N Earnest, JHD Cate, and Harry F Noller. "Crystal structure of the ribosome at 5.5 Å resolution." In: *science* 292.5518 (2001), pp. 883–896.

[81] Kenneth A Jacobson and Stefano Costanzi. "New insights for drug design from the X-ray crystallographic structures of G-protein-coupled receptors." In: *Molecular pharmacology* 82.3 (2012), pp. 361–371.

[82] Andrew M Davis, Simon J Teague, and Gerard J Kleywegt. "Application and limitations of X-ray crystallographic data in structure-based ligand and drug design." In: *Angewandte Chemie International Edition* 42.24 (2003), pp. 2718–2736.

[83] Masanori Osawa, Koh Takeuchi, Takumi Ueda, Noritaka Nishida, and Ichio Shimada. "Functional dynamics of proteins revealed by solution NMR." In: *Current opinion in structural biology* 22.5 (2012), pp. 660–669.

[84] J Patrick Loria, Rebecca B Berlow, and Eric D Watt. "Characterization of enzyme motions by solution NMR relaxation dispersion." In: *Accounts of chemical research* 41.2 (2008), pp. 214–221.

[85] Algirdas Velyvis, Ying R Yang, Howard K Schachman, and Lewis E Kay. "A solution NMR study showing that active site ligands and nucleotides directly perturb the allosteric equilibrium in aspartate transcarbamoylase." In: *Proceedings of the National Academy of Sciences* 104.21 (2007), pp. 8815–8820.

[86] Zhu Liu, Zhou Gong, Xu Dong, and Chun Tang. "Transient protein–protein interactions visualized by solution NMR." In: *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1864.1 (2016), pp. 115–122.

[87] Bei Liu, Honglue Shi, and Hashim M Al-Hashimi. "Developments in solution-state NMR yield broader and deeper views of the dynamic ensembles of nucleic acids." In: *Current opinion in structural biology* 70 (2021), pp. 16–25.

[88] Vladimír Mlynárik. "Introduction to nuclear magnetic resonance." In: *Analytical Biochemistry* 529 (2017), pp. 4–9.

[89] Natalie K Goto and Lewis E Kay. "New developments in isotope labeling strategies for protein solution NMR spectroscopy." In: *Current opinion in structural biology* 10.5 (2000), pp. 585–592.

[90] Dominique Marion. "An introduction to biological NMR spectroscopy." In: *Molecular & Cellular Proteomics* 12.11 (2013), pp. 3006–3025.

[91] Lesley A Earl, Veronica Falconieri, Jacqueline LS Milne, and Sriram Subramaniam. "Cryo-EM: beyond the microscope." In: *Current opinion in structural biology* 46 (2017), pp. 71–78.

[92] Yuanchen Dong, Shuwen Zhang, Zhaolong Wu, Xuemei Li, Wei Li Wang, Yanan Zhu, Svetla Stoilova-McPhie, Ying Lu, Daniel Finley, and Youdong Mao. "Cryo-EM structures and dynamics of substrate-engaged human 26S proteasome." In: *Nature* 565.7737 (2019), pp. 49–55.

[93] Xiao-Chen Bai, Greg McMullan, and Sjors HW Scheres. "How cryo-EM is revolutionizing structural biology." In: *Trends in biochemical sciences* 40.1 (2015), pp. 49–57.

[94] Robert M Glaeser. "How good can cryo-EM become?" In: *Nature methods* 13.1 (2016), pp. 28–32.

[95] Ka Man Yip, Niels Fischer, Elham Paknia, Ashwin Chari, and Holger Stark. "Atomic-resolution protein structure determination by cryo-EM." In: *Nature* 587.7832 (2020), pp. 157–161.

[96] Takanori Nakane, Abhay Kotecha, Andrija Sente, Greg McMullan, Simonas Masiulis, Patricia MGE Brown, Ioana T Grigoras, Lina Malinauskaite, Tomas Malinauskas, Jonas Miehling, et al. "Single-particle cryo-EM at atomic resolution." In: *Nature* 587.7832 (2020), pp. 152–156.

[97] Rita P Magalhães, Henriques S Fernandes, and Sérgio F Sousa. "Modelling enzymatic mechanisms with QM/MM approaches: current status and future challenges." In: *Israel Journal of Chemistry* 60.7 (2020), pp. 655–666.

[98] Michael Levitt and Arieh Warshel. "Computer simulation of protein folding." In: *Nature* 253.5494 (1975), pp. 694–698.

[99] Wilfred F Van Gunsteren and Herman JC Berendsen. "Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry." In: *Angewandte Chemie International Edition in English* 29.9 (1990), pp. 992–1023.

[100] Martin Karplus, Michael Levitt, and Arieh Warshel. "The nobel prize in chemistry 2013." In: *Nobel Media AB 2014* (2013).

[101] MA González. "Force fields and molecular dynamics simulations." In: *École thématique de la Société Française de la Neutronique* 12 (2011), pp. 169–200.

[102] John E Lennard-Jones. "Cohesion." In: *Proceedings of the Physical Society (1926-1948)* 43.5 (1931), p. 461.

[103] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. "How fast-folding proteins fold." In: *Science* 334.6055 (2011), pp. 517–520.

[104] David E Shaw, Peter J Adams, Asaph Azaria, Joseph A Bank, Brannon Batson, Alistair Bell, Michael Bergdorf, Jhanvi Bhatt, J Adam Butts, Timothy Correia, et al. "Anton 3: twenty microseconds of molecular dynamics simulation before lunch." In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 2021, pp. 1–11.

[105] Elmar Krieger and Gert Vriend. "New ways to boost molecular dynamics simulations." In: *Journal of computational chemistry* 36.13 (2015), pp. 996–1007.

[106] Katherine Henzler-Wildman and Dorothee Kern. "Dynamic personalities of proteins." In: *Nature* 450.7172 (2007), pp. 964–972. DOI: 10.1038/nature06522. URL: https://doi.org/10.1038%2Fnature06522.

[107] Laura R Ganser, Megan L Kelly, Daniel Herschlag, and Hashim M Al-Hashimi. "The roles of structural dynamics in the cellular functions of RNAs." In: *Nature reviews Molecular cell biology* 20.8 (2019), pp. 474–489.

[108] Michael R Sawaya and Joseph Kraut. "Loop and subdomain movements in the mechanism of Escherichia coli dihydrofolate reductase: crystallographic evidence." In: *Biochemistry* 36.3 (1997), pp. 586–603.

[109] Sandro Bottaro, Kresten Lindorff-Larsen, and Robert B Best. "Variational optimization of an all-atom implicit solvent force field to match explicit solvent simulation data." In: *Journal of chemical theory and computation* 9.12 (2013), pp. 5641–5652.

[110] Shingo Okuno, Akira Hirai, and Naoto Fukumoto. "Performance Analysis of Multi-Containerized MD Simulations for Low-Level Resource Allocation." In: *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE. 2022, pp. 1014–1017.

[111] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. "Equation of state calculations by fast computing machines." In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[112] Scott H Northrup and J Andrew McCammon. "Simulation methods for protein structure fluctuations." In: *Biopolymers: Original Research on Biomolecules* 19.5 (1980), pp. 1001–1016.

[113] William L Jorgensen and Julian Tirado-Rives. "Monte Carlo vs molecular dynamics for conformational sampling." In: *The Journal of Physical Chemistry* 100.34 (1996), pp. 14508–14513.

[114] Bernard R. Brooks, Dusanka Janezic, and Martin Karplus. "Harmonic analysis of large systems. I. Methodology." In: *Journal of Computational Chemistry* 16.12 (1995), pp. 1522–1542.

[115] Lukas Gerasimavicius, Xin Liu, and Joseph A Marsh. "Identification of pathogenic missense mutations using protein stability predictors." In: *Scientific Reports* 10.1 (2020), pp. 1–10.

[116] K Abdulla Bava, M Michael Gromiha, Hatsuho Uedaira, Koji Kitajima, and Akinori Sarai. "ProTherm, version 4.0: thermodynamic database for proteins and mutants." In: *Nucleic acids research* 32.suppl_1 (2004), pp. D120–D121.

[117] Javier Delgado, Leandro G Radusky, Damiano Cianferoni, and Luis Serrano. "FoldX 5.0: working with RNA, small molecules and a new graphical interface." In: *Bioinformatics* 35.20 (2019), pp. 4168–4169.

[118] Castrense Savojardo, Piero Fariselli, Pier Luigi Martelli, and Rita Casadio. "INPS-MD: a web server to predict stability of protein variants from sequence and structure." In: *Bioinformatics* 32.16 (2016), pp. 2542–2544.

[119] Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. "The Rosetta all-atom energy function for macromolecular modeling and design." In: *Journal of chemical theory and computation* 13.6 (2017), pp. 3031–3048.

[120] Yves Dehouck, Jean Marc Kwasigroch, Dimitri Gilis, and Marianne Rooman. "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.

[121] 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.

[122] 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.

[123] 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.

[124] 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.

[125] Lars Skjærven, Xin-Qiu Yao, Guido Scarabelli, and Barry J Grant. "Integrating protein structural dynamics and evolutionary analysis with Bio3D." In: *BMC bioinformatics* 15.1 (2014), pp. 1–11.

[126] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. "The FoldX web server: an online force field." In: *Nucleic acids research* 33.suppl_2 (2005), W382–W388.

[127] Oliver Buß, Jens Rudat, and Katrin Ochsenreither. "FoldX as protein engineering tool: better than random based approaches?" In: *Computational and structural biotechnology journal* 16 (2018), pp. 25–33.

[128] Piero Fariselli, Pier Luigi Martelli, Castrense Savojardo, and Rita Casadio. "INPS: predicting the impact of non-synonymous variations on protein stability from sequence." In: *Bioinformatics* 31.17 (2015), pp. 2816–2821.

[129] 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.

[130] Ugo Bastolla, Jochen Farwer, Ernst Walter Knapp, and Michele Vendruscolo. "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.

[131] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. "Highly accurate protein structure prediction with AlphaFold." In: *Nature* 596.7873 (2021), pp. 583–589.

[132] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences." In: *Proceedings of the National Academy of Sciences* 118.15 (2021), e2016239118.

[133] Jesse Horne and Diwakar Shukla. "Recent advances in machine learning variant effect prediction tools for protein engineering." In: *Industrial & Engineering Chemistry Research* 61.19 (2022), pp. 6235–6245.

[134] She Zhang, James M Krieger, Yan Zhang, Cihan Kaya, Burak Kaynak, Karolina Mikulska-Ruminska, Pemra Doruker, Hongchun Li, and Ivet Bahar. "ProDy 2.0: increased scale and scope after 10 years of protein dynamics modelling with Python." In: *Bioinformatics* 37.20 (2021), pp. 3657–3659.

[135] Tu-Liang Lin and Guang Song. "Generalized spring tensor models for protein fluctuation dynamics and conformation changes." In: *BMC Structural Biology* 10.Suppl 1 (2010), S3. DOI: 10.1186/1472-6807-10-s1-s3. URL: https://doi.org/10.1186%2F1472-6807-10-s1-s3.

[136] Florence Tama and Y-H Sanejouand. "Conformational change of proteins arising from normal mode calculations." In: *Protein engineering* 14.1 (2001), pp. 1–6.

[137] B.J. McConkey, V. Sobolev, and M. Edelman. "Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure." In: *Bioinformatics* 18.10 (2002), pp. 1365–1373. DOI: 10.1093/bioinformatics/18.10.1365. URL: https://doi.org/10.1093%2Fbioinformatics%2F18.10.1365.

[138] Vladimir Sobolev, Rebecca C. Wade, Gert Vriend, and Marvin Edelman. "Molecular docking using surface complementarity." In: *Proteins* 25.1 (1996), pp. 120–129. DOI: 10.1002/(sici)1097-0134(199605)25:1<120::aid-prot10>3.0.co;2-m. URL: https://doi.org/10.1002%2F%28sici%291097-0134%28199605%2925%3A1%3C120%3A%3Aaid-prot10%3E3.0.co%3B2-m.

[139] Xiaogang Su, Xin Yan, and Chih-Ling Tsai. "Linear regression." In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4.3 (2012), pp. 275–294.

[140] Robert Tibshirani. "Regression Shrinkage and Selection Via the Lasso." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x. URL: https://doi.org/10.1111%2Fj.2517-6161.1996.tb02080.x.

[141] Chris M Bishop. "Neural networks and their applications." In: *Review of scientific instruments* 65.6 (1994), pp. 1803–1832.

[142] Matt W Gardner and SR Dorling. "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences." In: *Atmospheric environment* 32.14-15 (1998), pp. 2627–2636.

[143] Zuyi Wang, Yue Wang, Jianhua Xuan, Yibin Dong, Marina Bakay, Yuanjian Feng, Robert Clarke, and Eric P Hoffman. "Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data." In: *Bioinformatics* 22.6 (2006), pp. 755–761.

[144] Yuanpeng Li, Junwei Lang, Lei Ji, Jiqin Zhong, Zaiwen Wang, Yang Guo, and Sailing He. "Weather forecasting using ensemble of spatial-temporal attention network and multi-layer perceptron." In: *Asia-Pacific Journal of Atmospheric Sciences* 57.3 (2021), pp. 533–546.

[145] Lee-Wei Yang, A J Rader, Xiong Liu, Cristopher Jon Jursa, Shann Ching Chen, Hassan A Karimi, and Ivet Bahar. "oGNM: online computation of structural dynamics using the Gaussian Network Model." In: *Nucleic Acids Research* 34.suppl_2 (2006), W24–W31. DOI: 10.1093/nar/gkl084. URL: https://doi.org/10.1093%2Fnar%2Fgkl084.

[146] Giovanni Pinamonti, Sandro Bottaro, Cristian Micheletti, and Giovanni Bussi. "Elastic network models for RNA: a comparative assessment with molecular dynamics and SHAPE experiments." In: *Nucleic Acids Res* 43.15 (2015), pp. 7260–7269. DOI: 10.1093/nar/gkv708. URL: https://doi.org/10.1093%2Fnar%2Fgkv708.

[147] Edvin Fuglebakk, Nathalie Reuter, and Konrad Hinsen. "Evaluation of protein elastic network models based on an analysis of collective motions." In: *Journal of chemical theory and computation* 9.12 (2013), pp. 5618–5628.

[148] Marc Parisien and François Major. "The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data." In: *Nature* 452.7183 (2008), pp. 51–55. DOI: 10.1038/nature06684. URL: https://doi.org/10.1038%2Fnature06684.

[149] Magdalena Rother, Kristian Rother, Tomasz Puton, and Janusz M. Bujnicki. "ModeRNA: a tool for comparative modeling of RNA 3D structure." In: *Nucleic Acids Research* 39.10 (2011), pp. 4007–4022. DOI: 10.1093/nar/gkq1320. URL: https://doi.org/10.1093%2Fnar%2Fgkq1320.

[150] Narayanan Eswar, David Eramian, Ben Webb, Min-Yi Shen, and Andrej Sali. "Protein structure modeling with MODELLER." In: *Structural proteomics*. Springer, 2008, pp. 145–159.

[151] Donald A McQuarrie. *Statistical mechanics*. Sterling Publishing Company, 2000, pp. 96–97, 137.

[152] Nobuhiro Gō. "A theorem on amplitudes of thermal atomic fluctuations in large molecules assuming specific conformations calculated by normal mode analysis." In: *Biophysical Chemistry* 35.1 (1990), pp. 105–112. DOI: 10.1016/0301-4622(90)80065-f. URL: https://doi.org/10.1016/0301-4622(90)80065-f.

[153] Turkan Haliloglu, Ivet Bahar, and Burak Erman. "Gaussian Dynamics of Folded Proteins." In: *Phys. Rev. Lett.* 79.16 (1997), pp. 3090–3093. DOI: 10.1103/physrevlett.79.3090. URL: https://doi.org/10.1103%2Fphysrevlett.79.3090.

[154] Carol A Rohl, Charlie EM Strauss, Kira MS Misura, and David Baker. "Protein structure prediction using Rosetta." In: *Methods in enzymology*. Vol. 383. Elsevier, 2004, pp. 66–93.

[155] Andrew P Bradley. "The use of the area under the ROC curve in the evaluation of machine learning algorithms." In: *Pattern recognition* 30.7 (1997), pp. 1145–1159.

[156] Jesse Davis and Mark Goadrich. "The relationship between Precision-Recall and ROC curves." In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 233–240.

[157] Olivier Mailhot, Vincent Frappier, François Major, and Rafael J. Najmanovich. In: *PLOS Computational Biology* 18.12 (Dec. 2022), pp. 1–28. DOI: 10.1371/journal.pcbi.1010777. URL: https://doi.org/10.1371/journal.pcbi.1010777.

[158] Jianpeng Ma and Martin Karplus. "Ligand-induced conformational changes in ras p21: a normal mode and energy minimization analysis." In: *Journal of Molecular Biology* 274.1 (1997), pp. 114–131. DOI: 10.1006/jmbi.1997.1313. URL: https://doi.org/10.1006%2Fjmbi.1997.1313.

[159] Michael T. Zimmermann and Robert L. Jernigan. "Elastic network models capture the motions apparent within ensembles of RNA structures." In: *RNA* 20.6 (2014), pp. 792–804. DOI: 10.1261/rna.041269.113. URL: https://doi.org/10.1261%2Frna.041269.113.

[160] H. Hotelling. "Analysis of a complex of statistical variables into principal components." In: *Journal of Educational Psychology* 24.6 (1933), pp. 417–441. DOI: 10.1037/h0071325. URL: https://doi.org/10.1037%2Fh0071325.

[161] G. H. Golub and C. Reinsch. "Singular value decomposition and least squares solutions." In: *Numer. Math.* 14.5 (1970), pp. 403–420. DOI: 10.1007/bf02163027. URL: https://doi.org/10.1007%2Fbf02163027.

[162] Florian Sittel, Abhinav Jain, and Gerhard Stock. "Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates." In: *The Journal of Chemical Physics* 141.1 (2014), 07B605_1.

[163] Arfken George. *Mathematical methods for physicists*. Academic press, 1985, pp. 516–520.

[164] Andrei Kouranov, Lei Xie, Joanna de la Cruz, Li Chen, John Westbrook, Philip E Bourne, and Helen M Berman. "The RCSB PDB information portal for structural genomics." In: *Nucleic acids research* 34.suppl_1 (2006), pp. D302–D305.

[165] Sibsankar Kundu, Julia S Melton, Dan C Sorensen, and George N Phillips Jr. "Dynamics of proteins in crystals: comparison of experiment with simple models." In: *Biophysical journal* 83.2 (2002), pp. 723–732.

[166] Yves Dehouck, Aline Grosfils, Benjamin Folch, Dimitri Gilis, Philippe Bogaerts, and Marianne Rooman. "Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0." In: *Bioinformatics* 25.19 (2009), pp. 2537–2543.

[167] Magdalena A. Machnicka et al. "MODOMICS: a database of RNA modification pathways—2013 update." In: *Nucleic Acids Research* 41.D1 (2012), pp. D262–D267. DOI: 10.1093/nar/gks1007. URL: https://doi.org/10.1093%2Fnar%2Fgks1007.

[168] Saul B. Needleman and Christian D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins." In: *Journal of Molecular Biology* 48.3 (1970), pp. 443–453. DOI: 10.1016/0022-2836(70)90057-4. URL: https://doi.org/10.1016%2F0022-2836%2870%2990057-4.

[169] D. Defays. "An efficient algorithm for a complete link method." In: *The Computer Journal* 20.4 (1977), pp. 364–366. DOI: 10.1093/comjnl/20.4.364. URL: https://doi.org/10.1093%2Fcomjnl%2F20.4.364.

[170] SN Sivanandam and SN Deepa. "Genetic algorithm optimization problems." In: *Introduction to genetic algorithms*. Springer, 2008, pp. 165–209.

[171] Tito Homem-de Mello and Güzin Bayraksan. "Monte Carlo sampling-based methods for stochastic optimization." In: *Surveys in Operations Research and Management Science* 19.1 (2014), pp. 56–85.

[172] Paul Dallaire and François Major. "Exploring Alternative RNA Structure Sets Using MC-Flashfold and db2cm." In: *Methods in Molecular Biology* 1490 (2016), pp. 237–251. DOI: 10.1007/978-1-4939-6433-8.

[173] Hui Yu, David C Samuels, Ying-yong Zhao, and Yan Guo. "Architectures and accuracy of artificial neural network for disease classification from omics data." In: *BMC genomics* 20.1 (2019), pp. 1–12.

[174] S Chul Kwon, Tuan Anh Nguyen, Yeon-Gil Choi, Myung Hyun Jo, Sungchul Hohng, V Narry Kim, and Jae-Sung Woo. "Structure of human DROSHA." In: *Cell* 164.1-2 (2016), pp. 81–90.

[175] Wenxing Jin, Jia Wang, Chao-Pei Liu, Hong-Wei Wang, and Rui-Ming Xu. "Structural basis for pri-miRNA recognition by Drosha." In: *Molecular cell* 78.3 (2020), pp. 423–433.

[176] Gurman S Pall and Andrew J Hamilton. "Improved northern blot method for enhanced detection of small RNA." In: *Nature protocols* 3.6 (2008), pp. 1077–1084.

[177] Carlo Bonferroni. "Teoria statistica delle classi e calcolo delle probabilita." In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8 (1936), pp. 3–62.

[178] Francis Gaudreault and Rafael J Najmanovich. "FlexAID: revisiting docking on non-native-complex structures." In: *Journal of chemical information and modeling* 55.7 (2015), pp. 1323–1336.

[179] S Chul Kwon, S Chan Baek, Yeon-Gil Choi, Jihye Yang, Young-suk Lee, Jae-Sung Woo, and V Narry Kim. "Molecular basis for the single-nucleotide precision of primary microRNA processing." In: *Molecular cell* 73.3 (2019), pp. 505–518.

[180] Richard DS Carliss, James F Keefer, Scott Perschke, Sandra Welch, Thomas C Rich, and Arthur D Weissman. "Receptor reserve reflects differential intrinsic efficacy associated with opioid diastereomers." In: *Pharmacology Biochemistry and Behavior* 92.3 (2009), pp. 495–502.

[181] Matthew Clark, Richard D Cramer III, and Nicole Van Opdenbosch. "Validation of the general purpose tripos 5.2 force field." In: *Journal of computational chemistry* 10.8 (1989), pp. 982–1012.

[182] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. "Open Babel: An open chemical toolbox." In: *Journal of cheminformatics* 3.1 (2011), pp. 1–14.

[183] Brian J Bender, Stefan Gahbauer, Andreas Luttens, Jiankun Lyu, Chase M Webb, Reed M Stein, Elissa A Fink, Trent E Balius, Jens Carlsson, John J Irwin, et al. "A practical guide to large-scale docking." In: *Nature protocols* 16.10 (2021), pp. 4799–4832.

[184] Christoph Gorgulla, Andras Boeszoermenyi, Zi-Fu Wang, Patrick D Fischer, Paul W Coote, Krishna M Padmanabha Das, Yehor S Malets, Dmytro S Radchenko, Yurii S Moroz, David A Scott, et al. "An open-source drug discovery platform enables ultra-large virtual screens." In: *Nature* 580.7805 (2020), pp. 663–668.

[185] Reed M Stein, Hye Jin Kang, John D McCorvy, Grant C Glatfelter, Anthony J Jones, Tao Che, Samuel Slocum, Xi-Ping Huang, Olena Savych, Yurii S Moroz, et al. "Virtual discovery of melatonin receptor ligands to modulate circadian rhythms." In: *Nature* 579.7800 (2020), pp. 609–614.

[186] Jiankun Lyu, Sheng Wang, Trent E Balius, Isha Singh, Anat Levit, Yurii S Moroz, Matthew J O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, et al. "Ultra-large library docking for discovering new chemotypes." In: *Nature* 566.7743 (2019), pp. 224–229.

[187] Yeng-Tseng Wang and Yang-Hsiang Chan. "Understanding the molecular basis of agonist/antagonist mechanism of human mu opioid receptor through gaussian accelerated molecular dynamics method." In: *Scientific reports* 7.1 (2017), pp. 1–11.

[188] Andreas Ioannis Karsisiotis, CF Damblon, and Gordon CK Roberts. "A variety of roles for versatile zinc in metallo-β-lactamases." In: *Metallomics* 6.7 (2014), pp. 1181–1197.

[189] Brian K Shoichet, Walter A Baase, Ryota Kuroki, and Brian W Matthews. "A relationship between protein stability and protein function." In: *Proceedings of the National Academy of Sciences* 92.2 (1995), pp. 452–456.

[190] Beth M Beadle and Brian K Shoichet. "Structural bases of stability–function tradeoffs in enzymes." In: *Journal of molecular biology* 321.2 (2002), pp. 285–296.

[191] Travis E Oliphant. *A guide to NumPy*. Vol. 1. Trelgol Publishing USA, 2006.

[192] Siu Kwan Lam, Antoine Pitrou, and Stanley Seibert. "Numba: A llvm-based python jit compiler." In: *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*. 2015, pp. 1–6.

[193] Roger G Grimes, John G Lewis, and Horst D Simon. "A shifted block Lanczos algorithm for solving sparse symmetric generalized eigenproblems." In: *SIAM Journal on Matrix Analysis and Applications* 15.1 (1994), pp. 228–272.

[194] Edward N Lorenz and K Haman. "The essence of chaos." In: *Pure and Applied Geophysics* 147.3 (1996), pp. 598–599.

[195] Christof Teuscher. "Revisiting the edge of chaos: Again?" In: *Biosystems* (2022), p. 104693.

[196] Norman H Packard. "Adaptation toward the edge of chaos." In: *Dynamic patterns in complex systems* 212 (1988), pp. 293–301.

[197] Nicola J Smith, Kirstie A Bennett, and Graeme Milligan. "When simple agonism is not enough: emerging modalities of GPCR ligands." In: *Molecular and cellular endocrinology* 331.2 (2011), pp. 241–247.