

**UNIVERSIDADE ESTADUAL PAULISTA – UNESP**  
**Instituto de Biociências, Letras e Ciências Exatas**  
**Campus de São José do Rio Preto**

**RAFAEL GIORDANO VIEGAS**

**EXPLORANDO O RELEVO DA SUPERFÍCIE DE ENERGIA  
DE MACROMOLÉCULAS BIOLÓGICAS**

São José do Rio Preto  
2025



**RAFAEL GIORDANO VIEGAS**

**EXPLORANDO O RELEVO DA SUPERFÍCIE DE ENERGIA  
DE MACROMOLÉCULAS BIOLÓGICAS**

Tese apresentada à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas, São José do Rio Preto, para obtenção do título de Doutor em Biofísica Molecular.

Área de Concentração: Biofísica Molecular

Orientador: Prof. Dr. Vitor Barbanti Pereira Leite

São José do Rio Preto  
2025

V656e

Viegas, Rafael Giordano

Explorando o relevo da superfície de energia de macromoléculas biológicas / Rafael Giordano Viegas. -- São José do Rio Preto, 2025  
150 f. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto

Orientador: Vitor Barbanti Pereira Leite

1. Biofísica Molecular. 2. Enovelamento de biomoléculas. 3. Redução de dimensionalidade. I. Título.

**RAFAEL GIORDANO VIEGAS**

**EXPLORANDO O RELEVO DA SUPERFÍCIE DE ENERGIA  
DE MACROMOLÉCULAS BIOLÓGICAS**

Tese apresentada à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas, São José do Rio Preto, para obtenção do título de Doutor em Biofísica Molecular.

Área de Concentração: Biofísica Molecular

Data da defesa: 13/05/2025

Banca Examinadora:

---

Prof. Dr. Vitor Barbanti Pereira Leite

UNESP - Instituto de Biociências, Letras e Ciências Exatas, Campus São José do Rio Preto  
Orientador

---

Prof. Dr. Jorge Chahine

UNESP - Campus São José do Rio Preto

---

Prof. Dr. Leandro Martinez

Universidade Estadual de Campinas

---

Prof. Dr. Ronaldo Junio de Oliveira

Universidade Federal do Triângulo Mineiro

---

Prof. Dra. Thereza A. Soares

Universidade de São Paulo - Câmpus Ribeirão Preto

Aos meus filhos,  
Joaquim e Heloísa

## AGRADECIMENTOS

Ao Instituto Federal de São Paulo (IFSP), expresso meu reconhecimento e gratidão por me conceder a oportunidade de me dedicar integralmente a este projeto de doutorado.

Aos meus pais, José Milton e Mariza, e à minha irmã Taisa, que são a base sólida da minha carreira e da minha vida, expresso minha gratidão pelo apoio constante em cada etapa.

À minha esposa, Suelen, que me acompanhou desde os primeiros anos da graduação, pelo apoio, carinho e compreensão durante todo o processo que resultou na retomada da pesquisa e elaboração desta tese. Ao meu filho, Joaquim, pelo aprendizado que proporciona e pela paciência nos momentos de maior dedicação. À minha filha, Heloísa, que chegou para tornar a parte final deste projeto mais feliz.

A todos os professores e servidores do IFSP/Catanduva, pela convivência e apoio ao longo desta jornada. Em especial, agradeço àqueles que estiveram mais próximos e contribuíram direta ou indiretamente com este trabalho, seja por meio de conselhos ou discussões. Destaco o professor Osmair de Oliveira, por ter me proporcionado os conhecimentos e a estrutura necessários para retomar as pesquisas em simulação computacional. Expresso também minha gratidão ao Guilherme Prata, pela amizade e por manter vivos e em constante expansão os projetos que desenvolvemos juntos.

Aos amigos que me acompanharam durante a jornada acadêmica e não deixaram de incentivar a realização deste projeto, em especial, Rodrigo Euzébio e Marcelo Boareto.

Aos professores da Unesp e do programa de pós-graduação, pelo conhecimento compartilhado e pelas oportunidades proporcionadas desde a graduação até este doutorado. Em especial, agradeço aos professores Sidney Jurado e Jorge Chahine, que avaliaram parte deste trabalho durante o exame de qualificação e forneceram valiosos comentários.

Ao meu orientador, Prof. Dr. Vitor Leite, pela excepcional recepção em seu grupo, pelos projetos e por todo o entusiasmo científico compartilhados. Aos membros do grupo Ingrid, Murilo, Juliana, Larissa, Fernando, Antonio, Vinícius Contessoto e Vinícius de Oliveira, meu muito obrigado pela boa convivência e pelas colaborações.

Por fim, agradeço aos pesquisadores que colaboraram com nosso grupo e contribuíram para etapas importantes descritas nesta tese: Angel Garcia, Alan Chen, Rodrigo Parra, Peter Wolynes, José Onuchic, Diego Ferreira, Alexandre Suman, Sonya Hanson, Richard Lindsay, Hao Wu, Garegin Papoian, Susmita Roy, Anushree Sinha, Avijit Mainan, Jennifer Simien, Ellinor Haglund, Fernando Paulovich e Karissa Sanbonmatsu.

Meu sincero obrigado a todos!

[...]

*Também se chamava estrada  
viagem de ventania.*

L. Borges., M. Borges, M. Nascimento.

## RESUMO

A teoria da superfície de energia surgiu nas últimas décadas como um paradigma fundamental para compreender processos moleculares complexos, como o enovelamento de proteínas e RNAs. Segundo essa teoria, a dinâmica conformacional das biomoléculas é guiada por superfícies de energia que vão desde funis suaves, típicos de proteínas globulares, até superfícies rugosas com múltiplos mínimos locais, características de proteínas desordenadas. No entanto, a caracterização e visualização detalhada dessas superfícies de energia representam um desafio significativo devido à complexidade e multidimensionalidade dos sistemas biomoleculares. Nesta tese, busca-se generalizar e detalhar o *Energy Landscape Visualization Method* (ELViM), um método desenvolvido para gerar representações intuitivas e de baixa dimensão das superfícies de energia, facilitando a análise de trajetórias moleculares e *ensembles* conformacionais. Inicialmente, apresenta-se um novo código *Python* para o ELViM, avaliando seu espaço de parâmetros e sua reprodutibilidade com dados do peptídeo MP1. Em seguida, aplica-se o ELViM para analisar o enovelamento de RNAs (tetra-alça GCAA e ribocomutador SAM-II), demonstrando a eficácia do método em agrupar estruturas similares — não apenas em relação à cadeia principal de açúcar-fosfato, mas também quanto à orientação das bases. Além disso, mostra-se que a superfície efetiva gerada pelo ELViM permite identificar *ensembles* de estados de transição, caracterizados pelas interações das nucleobases. Também foram investigados *ensembles* estruturais de proteínas intrinsecamente desordenadas (histonas H1 e H4, além de dados do repositório *Protein Ensemble Database*), demonstrando como o método pode ser empregado em análises conformacionais diferenciais e na caracterização da heterogeneidade conformacional desses sistemas. Por fim, foram analisadas as dinâmicas dos peptídeos antimicrobianos MP1 e H-MP1 em solução aquosa e na presença de membranas modelo, revelando diferenças conformacionais que podem influenciar a afinidade dos peptídeos pela membrana e a cinética de adsorção. Os resultados demonstram que o ELViM gera representações robustas e intuitivas para sistemas variados, possibilitando a identificação de mecanismos moleculares não discerníveis por métodos convencionais, abrindo novas perspectivas para a caracterização do relevo da superfície de energia de biomoléculas.

**Palavras-chave:** ELViM; relevo energético; projeção multidimensional; enovelamento de biomoléculas; dinâmica conformacional.

## ABSTRACT

The energy landscape theory has emerged over recent decades as a foundational paradigm for understanding complex molecular processes, such as protein and RNA folding. According to this framework, the conformational dynamics of biomolecules are guided by energy surfaces that range from smooth funnels, typical of globular proteins, to rugged landscapes with multiple local minima, characteristic of intrinsically disordered proteins. Nonetheless, the detailed characterization and visualization of these energy surfaces remain significant challenges due to the inherent complexity and high dimensionality of biomolecular systems. This thesis aims to generalize and refine the Energy Landscape Visualization Method (ELViM), a technique developed to produce intuitive, low-dimensional representations of energy landscapes, thereby facilitating the analysis of molecular trajectories and conformational ensembles. A new Python implementation of ELViM is introduced and evaluated in terms of its parameter space and reproducibility, using data from the MP1 peptide. The method is then applied to study RNA folding (specifically, the GCAA tetra-loop and SAM-II riboswitch), demonstrating its effectiveness in clustering structurally similar conformations — considering not only the sugar-phosphate backbone but also base orientation. The effective surfaces generated by ELViM enabled the identification of transition state ensembles, characterized by nucleobase interactions. Structural ensembles of intrinsically disordered proteins (histones H1 and H4, along with data from the Protein Ensemble Database) were also analyzed, illustrating the method’s utility in differential conformational analysis and in characterizing the conformational heterogeneity inherent to these systems. Finally, the conformational dynamics of the antimicrobial peptides MP1 and H-MP1 were investigated in aqueous solution and in the presence of model membranes, revealing differences that may influence their affinity and adsorption kinetics. Overall, the results demonstrate that ELViM provides robust and insightful representations across a variety of systems, enabling the identification of molecular mechanisms that are not readily accessible through conventional approaches and offering new perspectives for exploring the biomolecular energy landscape.

**Keywords:** ELViM; energy landscape; multidimensional projection; biomolecular folding; conformational dynamics.

## LISTA DE FIGURAS

Figura 1 – Representação bidimensional do funil de enovelamento . . . . .	18
Figura 2 – Análise de componentes principais e escalonamento multidimensional . . . . .	22
Figura 3 – Esquema metodológico do ELViM . . . . .	27
Figura 4 – Esquema da métrica do ELViM . . . . .	28
Figura 5 – Iteração do algoritmo <i>Force Scheme</i> . . . . .	30
Figura 6 – Estimativa de densidade por núcleo Gaussiano . . . . .	35
Figura 7 – Evolução temporal, projeção ELViM e densidade de estados do peptídeo MP1 . . . . .	42
Figura 8 – Reprodutibilidade do ELViM e o parâmetro $\sigma_0$ . . . . .	43
Figura 9 – Reprodutibilidade do ELViM e a taxa de aprendizado. . . . .	44
Figura 10 – Estrutura secundária do RNA e pseudonó tipo H . . . . .	47
Figura 11 – Estrutura da tetra-alça GCAA e evolução temporal de $Q_w$ para as trajetórias analisadas . . . . .	50
Figura 12 – Tetra-alça GCAA: projeção ELViM, LCSs e mapas de interação média . . . . .	51
Figura 13 – Comparando ELViM com a superfície de energia livre $Q_w \times RMSD$ . . . . .	53
Figura 14 – Densidade de estado e rotas de enovelamento para a tetra-alça GCAA. . . . .	54
Figura 15 – Análise dos <i>ensembles</i> de estados de transição para a tetra-alça GCAA. . . . .	55
Figura 16 – Estruturas secundária e terciária do SAM-II, trajetória e análise via ELViM . . . . .	58
Figura 17 – Transições entre as bacias D e PE para o SAM-II . . . . .	59
Figura 18 – Densidade de estados e LCSs para o TSE (D-PE) do SAM-II . . . . .	61
Figura 19 – Evolução do número médio de íons $Mg^{2+}$ por grupo fosfato ao longo de Q . . . . .	63
Figura 20 – Nucleossomo e histonas H1 e H4. . . . .	68
Figura 21 – Projeção ELViM para as formas nativa e acetiladas da cauda H4. . . . .	71
Figura 22 – Cauda H4: fração de ocupação. . . . .	73
Figura 23 – Entropia da projeção e heterogeneidade conformacional da cauda H4. . . . .	74
Figura 24 – Análise do espaço conformacional da Histona H1 através do ELViM . . . . .	78
Figura 25 – Projeção ELViM para os <i>ensembles</i> do fragmento NUS . . . . .	82
Figura 26 – Projeção ELViM para os <i>ensembles</i> do domínio DrkN . . . . .	84
Figura 27 – Sistemas analisados, espaço de fase efetivo e LCSs para os peptídeos em solução aquosa. . . . .	88
Figura 28 – Espaço de fase efetivo e LCS para os peptídeos no sistema contendo a membrana modelo. . . . .	92
Figura 29 – Estimativa de energia livre efetiva da projeção. . . . .	94
Figura B.1 – Impacto do parâmetro $\sigma_0$ nas projeções. . . . .	122
Figura B.2 – Impacto do parâmetro $L_r$ nas projeções. . . . .	122

Figura C.1–RMSD e $R_g$ para as trajetórias da tetra-alça GCAA . . . . .	123
Figura C.2–Evolução temporal de $Q_w$ para as demais transições da tetra-alça GCAA.	124
Figura C.3–Evolução temporal das demais transições da tetra-alça GCAA através da Projeção ELViM . . . . .	125
Figura C.4–Interação A-menor do tipo I . . . . .	127
Figura C.5–RMSD para as trajetórias do ribocomutador SAM-II . . . . .	128
Figura C.6–Mapas de Interação média em função de $Q$ para o SAM-II. . . . .	129
Figura C.7–Comparação qualitativa do ELViM com PCA para o SAM-II . . . . .	130
Figura C.8–Interação RNA–Mg <sup>2+</sup> : valores médios e interação mediada . . . . .	131
Figura D.1–LCSs e mapas de contato para a forma nativa (N) da cauda H4 . . . . .	132
Figura D.2–LCSs e mapas de contato para a forma A1 <sub>A</sub> da cauda H4 . . . . .	132
Figura D.3–LCSs e mapas de contato para a forma A1 <sub>B</sub> da cauda H4. . . . .	133
Figura D.4–LCSs e mapas de contato para a forma A2 da cauda H4. . . . .	133
Figura D.5–LCSs e mapas de contato para a forma A3 da cauda H4. . . . .	133
Figura D.6–LCSs e mapas de contato para a forma A4 da cauda H4. . . . .	134
Figura D.7–Cauda H4: projeção através de PCA e t-SNE . . . . .	134
Figura D.8–Cauda H4: comparação dos subespaços gerados por diferentes métodos.	135
Figura D.9–Posição da LCS com maior fração de ocupação nos diferentes métodos.	136
Figura D.10–LCS para os diferentes setores de ligação da H1. . . . .	137
Figura D.11–Projeção ELViM para os <i>ensembles</i> do fragmento NUL. . . . .	139
Figura D.12–Projeção ELViM para os <i>ensembles</i> do amino-terminal Sic1 . . . . .	140
Figura D.13–Reprodutibilidade do ELViM para os sistemas do PED . . . . .	141
Figura E.1 –Histogramas de energia e perfil de energia livre para os peptídeos MP1 e H-MP1 . . . . .	142
Figura E.2 –Distribuição de temperaturas no espaço de fase efetivo para os peptídeos MP1 e H-MP1. . . . .	142
Figura E.3–Posição em relação à membrana para os peptídeos MP1 e H-MP1 . . . . .	143
Figura F.1 –Análise do enovelamento da proteína 4ANK . . . . .	145
Figura F.2 –Análise de rotas de entrelaçamento na proteína leptina . . . . .	148
Figura F.3 –Análise de mutação na proteína ELF3 . . . . .	150

## LISTA DE ABREVIATURAS E SIGLAS

AMP	Peptídeos Antimicrobianos (Antimicrobial Peptides)
CG	Modelo com Representação Simplificada (Coarse-Grain Model)
CTD	Domínio Carboxi-Terminal (C-Terminal Domain)
CV	Variáveis Coletivas (Collective Variables)
DR	Redução de Dimensionalidade (Dimensionality Reduction)
ELT	Teoria do Relevo Energético (Energy Landscape Theory)
ELViM	Método de Visualização da Superfície de Energia (Energy Landscape Visualization Method)
FRET	Transferência de Energia de Ressonância de Förster (Förster Resonance Energy Transfer)
GD	Domínio Globular (Globular Domain)
IDP	Proteína Intrinsecamente Desordenada (Intrinsically Disordered Protein)
KDE	Estimativa de Densidade por Núcleo (Kernel Density Estimation)
LCS	Assinatura Conformacional Local (Local Conformational Signature)
MDS	Escalonamento Multidimensional (Multidimensional Scaling)
MP	Projeção Multidimensional (Multidimensional Projection)
NTD	Domínio Amino-Terminal (N-Terminal Domain)
PCA	Análise de Componentes Principais (Principal Component Analysis)
PED	Banco de Dados de <i>ensembles</i> de Proteínas Intrinsecamente Desordenadas (Protein Ensemble Database)
PTM	Modificação Pós-Traducional (Post-Translational Modification)
REMD	Dinâmica Molecular de Troca de Réplicas (Replica-Exchange Molecular Dynamics)
RMN	Ressonância Magnética Nuclear (Nuclear Magnetic Resonance)
SBM	Modelo Baseado em Estrutura (Structure-Based Model)

SAM	S-adenosilmetionina (S-Adenosyl Methionine)
SAM-II	Ribocomutador que se Liga à S-adenosilmetionina (S-Adenosyl Methionine Riboswitch)
SAXS	Espalhamento de Raios X a Baixos Ângulos (Small-Angle X-Ray Scattering)
smFRET	Transferência de Energia de Ressonância de Förster de Molécula Única (Single-Molecule Förster Resonance Energy Transfer)
TSE	<i>Ensemble</i> do Estado de Transição (Transition State Ensemble)

## LISTA DE SÍMBOLOS

$\Delta$	erro residual por ponto projetado
$\delta_{k,l}$	dissimilaridade estrutural entre duas conformações $k$ e $l$
$d_{k,l}$	distância Euclidiana entre pontos no espaço reduzido
$dRMSD$	desvio quadrático médio das distâncias internas
$\bar{E}$	erro absoluto médio
$H$	entropia da projeção
$K_S$	função <i>stress</i> de Kruskal
$L_r$	taxa de aprendizado
$L_{r_0}$	valor inicial da taxa de aprendizado
$L_{r_{\min}}$	valor mínimo da taxa de aprendizado
$Q$	fração de contatos nativos
$Q_w$	similaridade ao estado nativo
$q_w^{k,l}$	similaridade estrutural entre duas conformações $k$ e $l$
$r_{i,j}^k$	distância Euclidiana entre os átomos $i$ e $j$ em uma conformação $k$
$R_g$	raio de giração
$RMSD$	desvio quadrático médio das posições

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO . . . . .</b>	<b>17</b>
1.1	O ENOVELAMENTO DE PROTEÍNAS E A TEORIA DA SUPERFÍCIE DE ENERGIA . . . . .	17
1.2	PROJEÇÃO MULTIDIMENSIONAL . . . . .	20
1.3	OBJETIVO E ORGANIZAÇÃO DO TRABALHO . . . . .	24
<b>2</b>	<b>METODOLOGIA . . . . .</b>	<b>26</b>
2.1	ELVIM: ENERGY LANDSCAPE VISUALIZATION METHOD . . . . .	26
2.1.1	Matriz de dissimilaridade . . . . .	26
2.1.2	O algoritmo <i>Force Scheme</i> . . . . .	29
2.2	VARIÁVEIS COLETIVAS . . . . .	31
2.3	DENSIDADE DE ESTADOS E ENERGIA LIVRE EFETIVAS . . . . .	33
2.3.1	Superfície de Energia Livre Efetiva . . . . .	33
2.3.2	Estimativa de densidade por núcleo . . . . .	35
<b>3</b>	<b>ELVIM: NOVA IMPLEMENTAÇÃO E ESTUDO DE CASO COM O PEPTÍDEO MP1 . . . . .</b>	<b>37</b>
3.1	INTRODUÇÃO . . . . .	37
3.2	NOVA IMPLEMENTAÇÃO EM PYTHON . . . . .	37
3.2.1	A taxa de aprendizado $L_r$ . . . . .	38
3.2.2	Assinatura Conformacional Local . . . . .	39
3.3	PEPTÍDEO MP1: UM ESTUDO DE CASO . . . . .	40
3.3.1	Sistema analisado . . . . .	41
3.3.2	Resultados . . . . .	41
3.3.2.1	<i>Análise qualitativa da reprodutibilidade e do impacto de parâmetros no ELViM</i> . . . . .	43
3.4	CONCLUSÃO . . . . .	45
<b>4</b>	<b>EXPLORANDO A SUPERFÍCIE DE ENERGIA DE MOLÉCULAS DE RNA: A TETRA-ALÇA GCAA E O RIBOCOMUTADOR SAM-II . . . . .</b>	<b>47</b>
4.1	INTRODUÇÃO . . . . .	47
4.2	CARACTERIZAÇÃO DOS <i>ENSEMBLES</i> DE ESTADOS DE TRANSIÇÃO DA TETRA-ALÇA GCAA . . . . .	49
4.2.1	Sistema analisado . . . . .	49
4.2.2	Resultados . . . . .	51

4.2.3	Discussão . . . . .	56
4.3	ANÁLISE DO MECANISMO DE ENOVELAMENTO DO RIBOCOMUTADOR SAM-II . . . . .	56
4.3.1	Sistema analisado . . . . .	57
4.3.2	Resultados . . . . .	58
<b>4.3.2.1</b>	<b><i>Interação com íons Mg<sup>2+</sup></i></b> . . . . .	<b>62</b>
4.3.3	Discussão . . . . .	64
4.4	CONCLUSÃO . . . . .	65
<b>5</b>	<b>PROTEÍNAS INTRINSICAMENTE DESORDENADAS . . .</b>	<b>66</b>
5.1	INTRODUÇÃO . . . . .	66
5.1.1	Nucleossomo e Histonas . . . . .	67
5.2	ANÁLISE DO RELEVO DE ENERGIA DA CAUDA H4 E SUAS FORMAS ACETILADAS . . . . .	68
5.2.1	Sistema analisado . . . . .	69
5.2.2	Resultados . . . . .	70
5.3	ANÁLISE DO RELEVO DE ENERGIA DA HISTONA H1 . . . . .	75
5.3.1	Sistema analisado . . . . .	76
5.3.2	Resultados . . . . .	77
5.3.3	Discussão . . . . .	79
5.4	ANÁLISE DE <i>ENSEMBLES</i> DE PROTEÍNAS DESORNADAS DEPOSITADOS NO <i>THE PROTEIN ENSEMBLE DATABASE (PED)</i> . . . . .	79
5.4.1	Sistema analisado . . . . .	80
5.4.2	Resultados . . . . .	81
<b>5.4.2.1</b>	<b><i>Análise dos ensembles do fragmento NUS</i></b> . . . . .	<b>81</b>
<b>5.4.2.2</b>	<b><i>Análise dos ensembles do domínio drkN</i></b> . . . . .	<b>83</b>
5.4.3	Discussão . . . . .	85
5.5	CONCLUSÃO . . . . .	85
<b>6</b>	<b>INTERAÇÃO DE PEPTÍDEOS ANTIMICROBIANOS COM MEMBRANA MODELO POR MEIO DE ANÁLISE DO RELEVO DE ENERGIA . . . . .</b>	<b>87</b>
6.1	INTRODUÇÃO . . . . .	87
6.2	SISTEMA ANALISADO . . . . .	89
6.3	RESULTADOS . . . . .	90
6.3.1	Peptídeos em solução aquosa . . . . .	90
6.3.2	Adsorção dos peptídeos em bicamada lipídica . . . . .	91
<b>6.3.2.1</b>	<b><i>Estimativa de energia livre efetiva na projeção ELViM</i></b> . . . . .	<b>93</b>
6.4	CONCLUSÃO . . . . .	95

7	CONSIDERAÇÕES FINAIS . . . . .	96
	REFERÊNCIAS . . . . .	99
	APÊNDICES . . . . .	117
	APÊNDICE A – LISTA DE PUBLICAÇÕES . . . . .	118
	APÊNDICE B – CAPÍTULO 3: RESULTADOS COMPLE- MENTARES . . . . .	120
	APÊNDICE C – CAPÍTULO 4: METODOLOGIA E RESUL- TADOS COMPLEMENTARES . . . . .	123
C.1	TETRA-ALÇA GCAA . . . . .	123
C.2	RIBOCOMUTADOR SAM-II . . . . .	126
C.2.1	Metodologia Complementar . . . . .	126
<i>C.2.1.1</i>	<i>Projeção ELViM</i> . . . . .	<i>126</i>
<i>C.2.1.2</i>	<i>Análise das interações RNA-Mg<sup>2+</sup></i> . . . . .	<i>126</i>
C.2.2	Figuras Complementares . . . . .	127
	APÊNDICE D – CAPÍTULO 5: RESULTADOS COMPLE- MENTARES . . . . .	132
D.1	CAUDA H4 . . . . .	132
D.2	HISTONA H1 . . . . .	137
D.3	OUTROS SISTEMA DO PED . . . . .	138
	APÊNDICE E – CAPÍTULO 6: FIGURAS COMPLEMEN- TARES . . . . .	142
	APÊNDICE F – APLICAÇÃO DO ELVIM A OUTROS SIS- TEMAS . . . . .	144
F.1	DISTINGUINDO ROTAS DE ENOVELAMENTO EM REPETIÇÃO DE ANQUIRINA . . . . .	144
F.2	DISTINGUINDO ROTAS DE FORMAÇÃO DE NÓ TOPOLOGICO NA PROTEÍNA LEPTINA . . . . .	146
F.3	ANÁLISE CONFORMACIONAL DOS EFEITOS DE MUTAÇÕES NA PROTEÍNA ELF3 . . . . .	149

## 1 INTRODUÇÃO

### 1.1 O ENOVELAMENTO DE PROTEÍNAS E A TEORIA DA SUPERFÍCIE DE ENERGIA

Mais de cinquenta anos se passaram desde que Anfinsen demonstrou experimentalmente que a ribonuclease pancreática bovina (RNase A), após ser desnaturada por agentes químicos, retorna ao seu estado nativo quando restabelecidas as condições fisiológicas, sem necessidade de catalisadores ou outros fatores celulares (1). Essa descoberta indicou que a sequência de aminoácidos contém todas as informações necessárias para a formação de uma proteína funcional. Com base nisso, Anfinsen formulou a hipótese termodinâmica, segundo a qual a estrutura tridimensional nativa corresponde a um mínimo da energia livre de Gibbs do sistema (2). Desde então, a elucidação do mecanismo físico-químico do enovelamento proteico se consolidou como um dos maiores desafios da ciência teórica de proteínas, sendo amplamente investigado até os dias atuais.

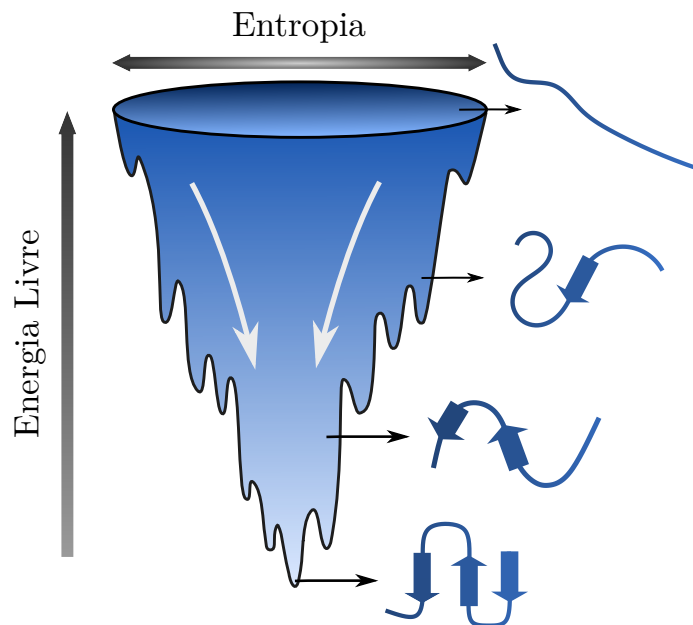
Cyrus Levinthal logo percebeu que, devido ao grande número de graus de liberdade, a cadeia polipeptídica poderia assumir um imenso número de conformações, não havendo tempo viável para a proteína realizar uma busca aleatória pela conformação mais estável (3,4). Esse fato, conhecido como paradoxo de Levinthal, o levou a propor que o enovelamento deveria seguir um caminho cinético preferencial, caracterizado por uma sequência específica de eventos que culminassem no estado nativo (4). Atraídos por esta hipótese, muitos estudos foram realizados nas décadas seguintes em busca de estados intermediários (5–7). Por outro lado, a descoberta de proteínas que enovelam de forma “tudo ou nada” sugeria um mecanismo de dois estados, caracterizado pela ausência de estado intermediário (8).

No final da década de 1980, uma nova concepção para o processo de enovelamento proteico emergiu a partir dos trabalhos realizados, principalmente, pelos grupos de J. N. Onuchic e P. G. Wolynes, culminando na formulação da Teoria da Superfície de Energia para o enovelamento de proteínas (ELT, do inglês *Energy Landscape Theory*) (5,9–11). Essa abordagem foi elaborada com base em modelos de física estatística de cadeias poliméricas, na teoria dos vidros de *spin* (*spin glasses*) e em resultados de simulações computacionais (10–12). Em suma, a ELT postula que, ao invés de uma busca aleatória, o processo de enovelamento é guiado por uma superfície de energia cujo gradiente direciona o sistema para o estado nativo.

Sob essa perspectiva, conhecida como funil de enovelamento (10), a proteína alcança o estado nativo não por uma via única, caracterizada por estados intermediários específicos, mas por um processo de difusão sobre o relevo da superfície de energia, no qual *ensembles* de estados com energias livres progressivamente menores convergem em direção ao estado

nativo (11,13). Desse modo, os estados desnaturados, que são numerosos e apresentam alta entropia configuracional, ocupam o topo do funil. À medida que a cadeia se compacta, ocorrendo a formação de contatos nativos e interações hidrofóbicas, o número de conformações possíveis diminui rapidamente até atingir um pequeno ensemble que define o estado nativo. Um esquema bidimensional representando o funil de enovelamento é apresentado na Figura 1, para o caso de uma proteína com estado nativo bem definido. Ressalta-se, no entanto, que o funil de enovelamento corresponde a uma superfície multidimensional, em que um eixo representa a energia livre do sistema e os demais eixos descrevem os diferentes graus de liberdade internos que caracterizam o espaço configuracional molecular.

Figura 1 – Representação do funil de enovelamento. Um grande número de estados desnaturados, com alta entropia configuracional ocupa a porção superior do funil. À medida que contatos presentes na estrutura nativa se formam, o número de estados conformacionais possíveis diminui, afinilando-se para um pequeno *ensemble* definindo o estado nativo. Devido a frustrações energéticas e topológicas, o funil pode apresentar topografia rugosa, marcada por diversos mínimos locais. Alguns desses mínimos podem ser profundos, formando armadilhas cinéticas que podem armadilhar a proteína em estados metaestáveis ou com enovelamento errôneo.



Fonte: Elaborada pelo autor (2025).

A ELT fornece um panorama através do qual é possível conciliar teoria e experimento. O aparente paradoxo de Levinthal tem origem na hipótese de que todos os estados, exceto o enovelado, possuem a mesma probabilidade de ocorrer, o que corresponderia a um funil de enovelamento em formato de campo de golfe (10,12,13). Contudo, as interações entre os resíduos de aminoácidos, as interações com o solvente e os efeitos topológicos, incluindo o volume excluído, conferem energias livres distintas aos estados conformacionais.

Adicionalmente, Bryngelson e Wolynes (9) demonstraram, através do estudo de *random energy models*, que heteropolímeros aleatórios exibem comportamentos similares aos dos vidros de *spin*, evidenciando a frustração energética, definida como a impossibilidade de se minimizar cada uma das interações simultaneamente.

Como consequência, a superfície de energia desses sistemas deve ser rugosa, com a existência de mínimos locais e barreiras de energia capazes de influenciar a cinética de enovelamento (9, 12, 14). Postula-se, no entanto, que as proteínas possuem sequências de aminoácidos selecionadas sob pressão evolutiva, resultando em uma superfície de energia minimamente frustrada, com um viés energético forte em sentido ao estado nativo (14, 15). Assim, a ELT é capaz de descrever desde proteínas que se enovelam rapidamente, descritas por relevos de energia afunilados e minimamente frustrados, até processos de enovelamento complexos, descritos por superfícies rugosas com maior nível de frustração.

Esse panorama fornecido pela ELT pode ainda ser estendido para outros sistemas, como o enovelamento de RNAs e a dinâmica conformacional das Proteínas Intrinsecamente Desordenadas (IDPs, do inglês Intrinsically Disordered Proteins) (16–19). As moléculas de RNA possuem um enovelamento hierárquico, que se inicia com a estabilização da estrutura secundária, seguida pela formação de interações terciárias. As interações intrínsecas entre as nucleobases conferem ao RNA um relevo energético complexo, caracterizado por múltiplos mínimos locais e diversas trajetórias de enovelamento, incluindo a possibilidade de formação de estruturas alternativas (16, 20). De modo semelhante, as IDPs, caracterizadas pela alta plasticidade estrutural e ausência de estado nativo, são descritas por uma paisagem energética mais plana, com múltiplos mínimos separados por pequenas barreiras de energia, permitindo-lhes transitar entre múltiplas conformações (17, 19). Por outro lado, o relevo energético de alguns sistemas pode apresentar mínimos locais profundos, que podem estar relacionados a armadilhas cinéticas e estados metaestáveis. Esses estados podem aprisionar as proteínas em conformações errôneas, levando à formação de agregados proteicos associados ao surgimento de patologias (21, 22).

Apesar do sucesso da ELT, o caráter multidimensional do funil de enovelamento dificulta o desenvolvimento de técnicas de visualização e caracterização dessa superfície. Uma abordagem comum consiste na adoção de coordenadas de reação definidas *a priori*, que descrevem o progresso do enovelamento sobre a superfície de energia. Dentre elas, destacam-se a similaridade ao estado nativo ( $Q_w$ ) (23) e a fração de contatos nativos ( $Q$ ) (24). Outros parâmetros de ordem — como o raio de giração ( $R_g$ ), distância ponta-à-ponta, conteúdo de hélice- $\alpha$ /folha- $\beta$  e até mesmo distâncias específicas entre átomos ou resíduos de interesse — também são comumente utilizados. Embora essa abordagem permita descrever a cinética e obter propriedades termodinâmicas a partir de um conjunto limitado de parâmetros, não existem regras definitivas para a seleção de um conjunto apropriado (23, 25). Além disso, essa escolha pode simplificar excessivamente a representação do

sistema, ocultando detalhes relevantes do relevo de energia.

Uma abordagem alternativa consiste no uso de técnicas de redução de dimensionalidade, que projetam a superfície multidimensional em coordenadas efetivas. Diferentemente das coordenadas de reação — que exigem a seleção prévia de parâmetros específicos — essas técnicas buscam identificar automaticamente coordenadas intrínsecas que melhor capturam a variação dos dados. Na próxima seção, são apresentadas algumas dessas técnicas e suas aplicações a sistemas moleculares.

## 1.2 PROJEÇÃO MULTIDIMENSIONAL

Técnicas de redução de dimensionalidade (DR, do inglês *Dimensionality Reduction*) buscam projetar um espaço original de alta dimensão em um espaço de dimensões reduzidas. Quando o objetivo é a visualização dos dados, o espaço reduzido, também denominado espaço visual, é limitado a três dimensões e o procedimento conhecido como Projeção Multidimensional (MP, do inglês *Multidimensional Projection*). Esses métodos partem da premissa de que é possível manter as informações mais relevantes do sistema no espaço de baixa dimensão, eliminando informações ruidosas ou redundantes (26,27). Especificamente, técnicas de DR podem ser desenvolvidas para preservar no espaço de baixa dimensão, por exemplo, relações de distâncias entre pares de dados ou relações de vizinhança.

Para ilustrar a natureza multidimensional do espaço de fases conformacional de uma macromolécula, considere um *ensemble* de conformações obtido por simulação computacional, no qual os graus de liberdade internos da cadeia polipeptídica são descritos pelas coordenadas dos carbonos alfa. Para uma proteína com  $N$  aminoácidos, existem  $3N$  graus de liberdade, que se reduzem a  $3N - 6$  ao eliminar os movimentos de translação e rotação, fixando o centro de massa na origem e sobrepondo as conformações a uma estrutura de referência. Desse modo, cada conformação amostrada corresponde a um ponto nesse espaço de alta dimensão, cuja posição é descrita por um vetor de dimensão  $3N - 6$ . As restrições bioquímicas, como ligações químicas e impedimentos estéricos, e movimentos correlacionados dos átomos podem ainda reduzir o número de graus de liberdade efetivos. No entanto, a topologia da superfície de energia multidimensional varia entre sistemas e não é conhecida *a priori*. A aplicação de técnicas de MP a sistemas biomoleculares procura gerar um espaço reduzido, com duas ou três dimensões efetivas, de modo que cada conformação seja representada por um ponto e que as informações relevantes sobre o sistema sejam preservadas.

Entre os métodos de redução de dimensionalidade, a Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*) é uma das técnicas mais amplamente utilizadas (28–30). Dado um conjunto de dados, a PCA determina as direções principais — conhecidas como componentes principais — ao longo das quais a variabilidade dos dados é maximizada, garantindo que sua projeção nessas direções minimize a perda

de informação. Esquemáticamente, considere uma matriz de dados  $\mathbf{A}_{nm}$ , em que cada uma das  $n$  linhas representa uma amostra (ou instância), e cada uma das  $m$  colunas corresponde a um atributo (ou coordenada) dessa amostra no espaço de alta dimensão. Inicialmente, os dados são centralizados subtraindo-se a média de cada coluna. Em seguida, calcula-se a matriz de covariância dos dados, que é diagonalizada para se obter seus autovalores e autovetores. Cada autovetor indica a direção de uma componente principal, e o autovalor associado corresponde à variância dos dados nessa direção. Um esquema desse procedimento é ilustrado, para o caso bidimensional, na Figura 2.a. Por fim, se apenas um pequeno número desses autovalores representar uma fração significativa da variância total, a dimensionalidade do sistema pode ser reduzida projetando os dados ao longo dos respectivos autovetores.

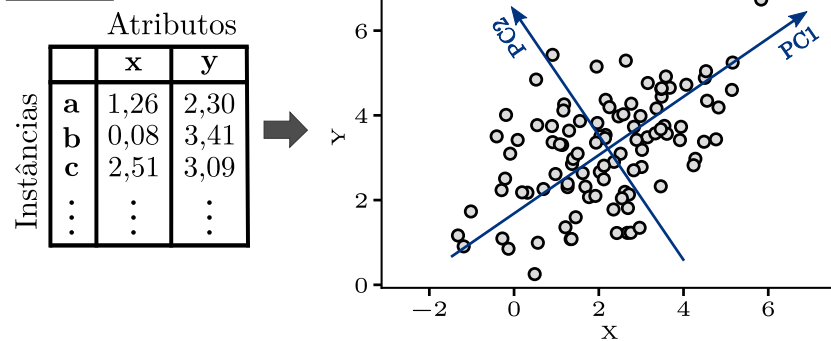
Em um trabalho seminal, García e colaboradores mostraram que essa técnica poderia ser empregada na análise de mudanças conformacionais, ou movimentos coletivos, da proteína crambina (31). A projeção dos movimentos atômicos ao longo das componentes principais revela movimentos correlacionados de baixa frequência e altas amplitudes, que ficaram conhecidos como dinâmica essencial e podem estar relacionados aos movimentos funcionais de proteínas (32). Em sistemas biomoleculares, os atributos utilizados nessa análise geralmente incluem as coordenadas cartesianas dos átomos, os ângulos diedrais ou conjuntos de distâncias internas (30). Recentemente, a técnica tem sido amplamente utilizada para caracterizar o relevo energético amostrado em simulações moleculares, analisar a dinâmica essencial, identificar estados estáveis e metaestáveis, dentre outros (30, 33–35).

Outra classe relevante de técnicas de redução de dimensionalidade é o escalonamento multidimensional (MDS, do inglês *Multidimensional Scaling*) (37, 38). O MDS visa representar relações de dissimilaridade entre objetos (ou instâncias), projetando-os em um espaço de menor dimensão. Diferentemente da PCA, que busca maximizar a variância dos dados projetando-os em componentes ortogonais, o MDS concentra-se em preservar, com a maior fidelidade possível, as dissimilaridades originais entre os dados.

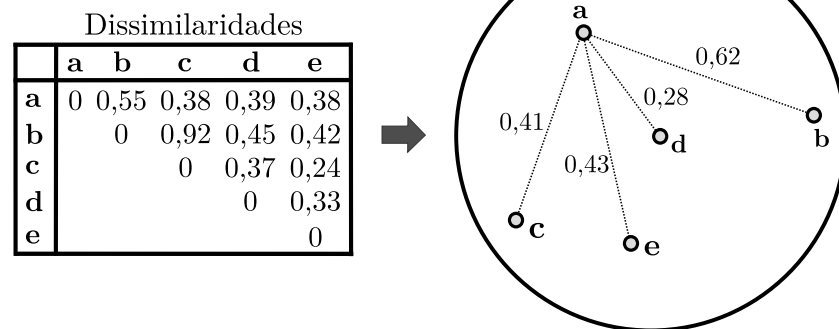
As técnicas de MDS foram inicialmente desenvolvidas nas áreas de psicofísica e análise sensorial, com o objetivo de representar visualmente estímulos em um espaço de baixa dimensionalidade, a partir das dissimilaridades percebidas entre eles (39, 40). No MDS, utiliza-se uma matriz quadrada e simétrica  $\mathbf{A}_{nn}$ , na qual  $n$  representa o número de objetos, e os elementos correspondem às medidas de dissimilaridade entre os pares. Dessa forma, o método permite obter uma representação dos objetos em um espaço de baixa dimensionalidade que preserva, o melhor possível, as relações de dissimilaridade, sem exigir, em geral, o conhecimento prévio de suas coordenadas no espaço original de alta dimensão — embora algumas medidas de dissimilaridade, como as baseadas em distâncias geométricas, possam requerer essas coordenadas para serem estimadas.

Figura 2 – Análise de Componentes Principais (PCA) e Escalonamento Multidimensional (MDS). (a) No PCA, parte-se de uma matriz de dados (ou atributos), na qual cada linha representa uma instância e cada coluna, um atributo ou coordenada. As componentes principais indicam as direções de maior variância nos dados. (b) No MDS, parte-se de uma matriz de distâncias (ou dissimilaridades) e os pontos são distribuídos no plano de forma a preservar, de maneira otimizada, as relações de distância entre eles. Nestes exemplos, as coordenadas e distâncias foram geradas aleatoriamente, e tanto o PCA quanto o MDS foram realizados utilizando a biblioteca *scikit-learn* (36).

a) PCA:



b) MDS:



Fonte: Elaborada pelo autor (2025).

O MDS clássico posiciona cada instância em um espaço de menor dimensão de modo que a distância euclidiana entre os pares de pontos nesse espaço corresponda, da maneira mais precisa possível, à dissimilaridade original entre os objetos. Para alcançar esse objetivo, o método minimiza uma função perda (ou função custo), frequentemente definida pelo erro quadrático entre as dissimilaridades originais e as distâncias no espaço de baixa dimensão (38):

$$L = \sum_{i < j} (\delta_{i,j} - d_{ij})^2. \quad (1.1)$$

Nesta equação,  $\delta_{i,j}$  representa a dissimilaridade entre os objetos  $i$  e  $j$  no espaço original de alta dimensão, enquanto  $d_{i,j}$  denota a distância euclidiana entre esses mesmos pontos no

espaço reduzido. Ao minimizar  $L$ , o MDS busca encontrar uma configuração de pontos que preserve, da melhor forma possível, a estrutura de dissimilaridades dos dados originais. Um esquema desse método é apresentado na Figura 2.b. As coordenadas absolutas dos pontos não possuem significado intrínseco nesse método, permitindo que os eixos sejam omitidos ou substituídos por circunferências de referência (26).

A aplicação do MDS em sistemas biomoleculares exige a construção de uma matriz de distâncias entre todas as conformações de uma trajetória. Frequentemente, essa matriz é baseada no RMSD entre pares de conformações. Apesar da interpretação intuitiva do RMSD, essa abordagem depende do alinhamento estrutural, o qual pode ser problemático ao analisar conformações muito distintas, como em trajetórias de enovelamento (41). Alternativamente, medidas de similaridade que dispensam alinhamento prévio, como o dRMSD (42, 43), podem ser utilizadas, embora sejam menos exploradas. Exemplos de aplicação do MDS a sistemas moleculares visam reconstruir o espaço conformacional de biomoléculas, agrupar estruturas similares e caracterizar *ensembles* conformacionais (44–47).

Um ponto a ser considerado é que tanto o PCA quanto o MDS clássico são métodos lineares, o que pode limitar sua capacidade de capturar relações não lineares entre os atributos dos dados. Diversas técnicas não lineares foram propostas para superar essas limitações, como o *Isomap* (48), *Diffusion Maps* (49), *t-SNE* (50), *UMAP* (51) e *Sketch-map* (52). Além disso, algumas abordagens têm explorado o uso de aprendizado de máquina e *deep learning* para reduzir a dimensionalidade ou gerar visualizações, como no caso do *EncoderMap* (53). Uma revisão e uma análise comparativa desses métodos podem ser encontradas, assim como suas aplicações na área de simulação molecular, nas seguintes referências (41, 52, 54).

Nos últimos anos, nosso grupo vem desenvolvendo uma abordagem alternativa denominada *Energy Landscape Visualization Method* (ELViM) (55, 56). O ELViM é um método de projeção multidimensional não linear, que foi desenvolvido especificamente para mapear o relevo da superfície de energia de proteínas. Esse método utiliza uma métrica de similaridade cuja forma funcional se baseia em um parâmetro de ordem para vidros de *spin* e foi posteriormente utilizada em potenciais de memória associativa (associative memory potentials) (57) e na coordenada de reação  $Q_w$  (23). O ELViM gera uma projeção multidimensional, semelhante ao MDS, preservando no plano, tanto quanto possível, a similaridade estrutural estimada por essa métrica. Para isso, o método *Force Scheme* (58) foi escolhido por seu bom equilíbrio entre eficácia e custo computacional.

O ELViM tem se mostrado um método promissor, capaz de mapear o espaço conformacional de proteínas de forma que parâmetros de ordem ou coordenadas de reação usuais variem, em média, suavemente ao longo do espaço efetivo projetado. Notavelmente, esse espaço se relaciona visualmente à concepção de funil de enovelamento, o que pode ser

ressaltado ao utilizar, por exemplo, a energia potencial das conformações em um terceiro eixo (56). Essa característica distingue o ELViM de outros métodos de MP, pois fornece uma interpretação intuitiva e facilita a identificação de agrupamentos estruturais (*clusters*) e relações de vizinhança entre eles (56, 59, 60).

Esta tese apresenta uma nova implementação do ELViM, com o objetivo de aumentar a eficiência computacional e integrar todas as etapas do método em um único código escrito em *Python*. São analisados os efeitos da escolha dos parâmetros sobre o resultado final, bem como sua influência na reprodutibilidade e na qualidade da projeção. O método é então aplicado a diferentes biomoléculas, com distintos níveis de complexidade, explorando aspectos metodológicos e biofísicos específicos de cada sistema.

A aplicação do ELViM a diferentes sistemas biomoleculares, como peptídeos, RNAs e proteínas intrinsecamente desordenadas, permite investigar seus espaços conformacionais, descrever preferências estruturais e identificar *ensembles* de estado de transição. O método também é empregado na construção de um espaço de fase efetivo único, capaz de englobar *ensembles* gerados por diferentes metodologias, incluindo variantes com mutações ou distintos padrões de acetilação. Essa abordagem viabiliza uma análise conformacional diferencial, apta a capturar particularidades na composição de cada *ensemble*, possivelmente relacionadas a propriedades funcionais dos sistemas biológicos. A aplicabilidade do ELViM é explorada em diferentes contextos, abrangendo simulações de dinâmica molecular e *ensembles* estruturais provenientes de repositórios públicos, o que evidencia a robustez e a versatilidade da metodologia proposta.

### 1.3 OBJETIVO E ORGANIZAÇÃO DO TRABALHO

O objetivo desta tese é reformular e generalizar o método ELViM, desenvolver novas ferramentas de análise e aplicá-lo a sistemas biomoleculares com diferentes níveis de complexidade. O foco principal está na análise de *ensembles* conformacionais obtidos a partir de simulações de dinâmica molecular ou disponibilizados por repositórios públicos. As simulações analisadas foram conduzidas por colaboradores de diferentes grupos de pesquisa, sendo a presente investigação centrada na aplicação e interpretação dos dados por meio do ELViM.

O Capítulo 2 apresenta a metodologia de análise utilizada ao longo da tese, incluindo o algoritmo ELViM e ferramentas complementares, como a fração de contatos nativos, a estimativa de densidade por núcleo e o cálculo de energia livre.

O Capítulo 3, baseado em artigo publicado no *Journal of Chemical Information and Modeling* (JCIM) (61), descreve a reformulação do ELViM, agora implementado em um único código *Python*. Essa nova versão oferece maior flexibilidade na entrada de dados e ganhos expressivos em eficiência computacional. Também é introduzida a “Assinatura

Conformacional Local” (LCS, do inglês *Local Conformational Signature*), uma ferramenta de visualização empregada nas análises subsequentes. Como exemplo de aplicação, é analisada uma trajetória de dinâmica molecular do peptídeo antimicrobiano MP1, com ênfase na exploração de parâmetros e na avaliação da reprodutibilidade do método.

A aplicabilidade do ELViM é explorada em diferentes classes de sistemas biomoleculares. O Capítulo 4 apresenta sua aplicação no estudo do enovelamento de moléculas de RNA. A Seção 4.2, baseada no artigo publicado no JCIM (62), analisa trajetórias de enovelamento da tetra-alça GCAA. A Seção 4.3 examina os mecanismos de enovelamento do ribocomutador SAM-II e explora sua interação com íons  $Mg^{2+}$ . Os resultados demonstram a capacidade do ELViM de agrupar estruturas de RNA com alta fidelidade, incluindo a orientação relativa das nucleobases, que apresenta desempenho superior a métodos convencionais. O espaço projetado é utilizado para caracterizar conformações prevalentes e identificar *ensembles* de transição.

O Capítulo 5 é dedicado ao estudo de IDPs. A Seção 5.2 emprega o ELViM em uma análise conformacional diferencial da cauda da histona H4 nativa e de cinco isoformas com diferentes padrões de acetilação. Na Seção 5.3, o método é aplicado à histona H1, investigando seu espaço conformacional e possíveis modos de ligação ao nucleossomo. Por sua vez, a Seção 5.4, baseada em artigo publicado no JCIM (63), examina *ensembles* estruturais do repositório Protein Ensemble Database (PED), gerados por abordagens híbridas que integram simulações e técnicas experimentais. Mesmo diante da elevada plasticidade das IDPs, os resultados revelam que o ELViM permite identificar padrões de amostragem e se destaca como ferramenta útil na comparação entre diferentes sistemas desordenados.

O Capítulo 6 analisa a interação dos peptídeos antimicrobianos MP1 e H-MP1 com membranas modelo, conforme descrito em artigo publicado no *Journal of Physical Chemistry B* (64). São avaliadas tanto simulações em solução aquosa quanto simulações de adsorção espontânea à membrana, com o objetivo de identificar assinaturas conformacionais relacionadas aos mecanismos de ação desses peptídeos.

Por fim, o Apêndice F apresenta brevemente outros sistemas analisados, incluindo proteínas com repetição de anquirina (publicado no *QRB Discovery* (65)), rotas de formação de nó topológico na leptina (publicado no *Journal of Structural Biology* (66)) e mutações na proteína ELF3 (publicado como *preprint* revisado no *eLife* (67)).

## 2 METODOLOGIA

### 2.1 ELViM: ENERGY LANDSCAPE VISUALIZATION METHOD

Dada uma trajetória obtida por simulação molecular (68–70) ou um *ensemble* conformacional gerado por métodos experimentais ou integrativos (71), a aplicação do ELViM envolve dois passos fundamentais:

- Extração de coordenadas atômicas e cálculo da matriz de dissimilaridade. Nesta etapa, o nível de detalhe a ser utilizado no cálculo de dissimilaridade é definido (geralmente, coordenadas dos carbonos alfa). Com base no conjunto de dados, extrai-se uma matriz de coordenadas. O ELViM aplica então sua métrica própria, descrita detalhadamente a seguir, para calcular a matriz de dissimilaridade, que estima as distâncias estruturais entre pares de conformações moleculares.
- Projeção multidimensional através do algoritmo *force scheme*. Cada conformação é representada como um ponto em um espaço bidimensional, de forma que as distâncias euclidianas entre os pares de pontos no plano cartesiano correspondam, idealmente, às dissimilaridades previamente calculadas no espaço multidimensional. Essa representação em baixa dimensão é denominada espaço de fase efetivo ou, simplesmente, projeção (ELViM).

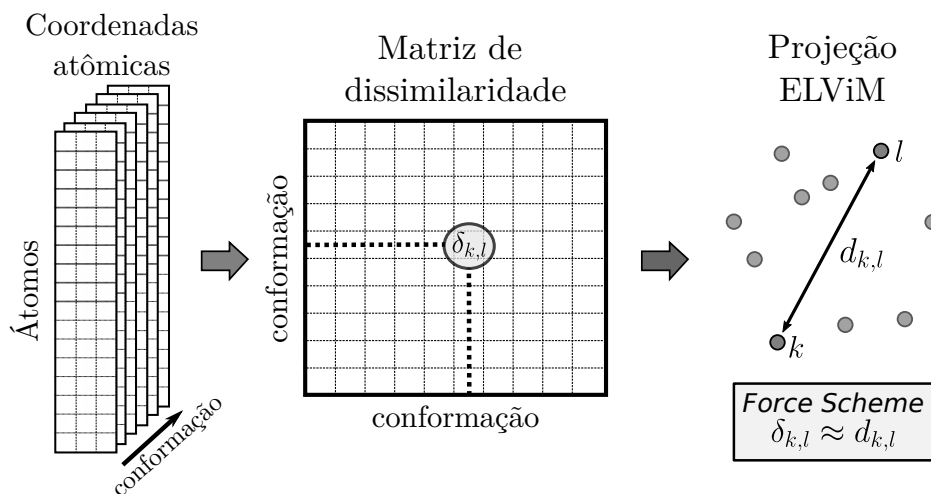
A Figura 3 ilustra a aplicação desses passos e o espaço de fase efetivo gerado. Nesse espaço, as coordenadas absolutas dos pontos não possuem significado intrínseco. Assim, pontos próximos indicam conformações estruturalmente similares, de acordo com a métrica utilizada, enquanto pontos distantes representam conformações estruturalmente distintas. Nesse contexto, é recomendado omitir os eixos coordenados para destacar exclusivamente as informações inerentes à projeção (26). Para facilitar a comparação entre diferentes projeções, em alguns casos, será adotada uma circunferência como referência, circunscrevendo a projeção, conforme sugerido na referência (26).

Nas próximas seções, serão fornecidos detalhes sobre essas duas etapas e o espaço de parâmetros do método será discutido.

#### 2.1.1 Matriz de dissimilaridade

Uma primeira versão do ELViM foi desenvolvida para o estudo de modelos de proteínas em rede (55). Nessa versão, o ELViM empregava uma métrica de dissimilaridade baseada na razão entre o índice e a distância de Jaccard calculadas sobre mapas de contato (55). Posteriormente, ao generalizar o método para sistemas biomoleculares fora de redes,

Figura 3 – Dado um *ensemble* molecular e escolhido um nível de resolução, as coordenadas atômicas do sistema são extraídas. Aplicando a métrica do ELViM, obtém-se uma matriz de dissimilaridade, que representa as distâncias estruturais entre os pares. O algoritmo *Force scheme* é então utilizado para realizar a projeção multidimensional, resultando em um espaço efetivo bidimensional, no qual cada ponto representa uma conformação e a distância euclidiana entre os pontos aproxima-se otimamente das dissimilaridades estimadas. Usualmente, os eixos coordenados são omitidos, uma vez que as coordenadas absolutas dos pontos não são significativas.



Fonte: Elaborada pelo autor (2025).

Oliveira *et al.* (56) propuseram uma nova métrica fundamentada em um parâmetro de ordem derivado da física de vidros de *spin*, amplamente utilizado no estudo do envelhecimento de biomoléculas (57, 72, 73).

Para duas conformações quaisquer,  $k$  e  $l$ , a similaridade estrutural entre elas é obtida comparando-se todas as distâncias internas, conforme esquematizado na Figura 4. Matematicamente, a similaridade é definida como:

$$q^{k,l} = \frac{1}{N_p} \sum_{i,j \in S} \exp \left[ \frac{-(r_{i,j}^k - r_{i,j}^l)^2}{2\sigma_{i,j}^2} \right], \quad (2.1)$$

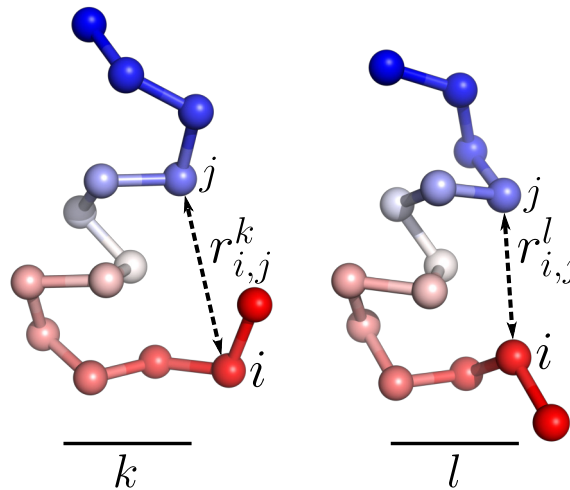
sendo  $r_{i,j}^k$  e  $r_{i,j}^l$  as distâncias entre os átomos  $i$  e  $j$  nas conformações  $k$  e  $l$ , respectivamente. A soma é realizada sobre o conjunto  $S$  dos  $N_p$  pares de átomos. O parâmetro  $\sigma_{i,j}$  é definido como:

$$\sigma_{i,j} = \sigma_0 |n_i - n_j|^\epsilon, \quad (2.2)$$

sendo  $n_i$  e  $n_j$  os números dos resíduos aos quais os átomos  $i$  e  $j$  pertencem. O parâmetro

$\sigma_{i,j}$  define a resolução da medida de similaridade, sendo que  $\sigma_0$  indica a largura de banda (*bandwidth*) inicial, enquanto  $\epsilon$  regula a dependência de  $\sigma_{i,j}$  com a distância ao longo da cadeia. Usualmente,  $\epsilon$  é mantido fixo em 0,15 (23), e considera-se o intervalo  $0,5 \leq \sigma_0 \leq 3$ .

Figura 4 – Dadas duas conformações,  $k$  e  $l$ , a similaridade estrutural é obtida ao comparar todas as distâncias internas das duas conformações. Este cálculo não exige alinhamento estrutural prévio e é realizado por meio da Equação 2.1.



Fonte: Elaborada pelo autor (2025).

Embora as projeções do ELViM sejam geralmente robustas ao uso desses valores típicos, valores atípicos podem distorcer os resultados da projeção. Valores muito pequenos de  $\sigma_0$  levam a menores similaridades, dificultando a representação de todos os pontos no plano. Em contraste, valores elevados de  $\sigma_0$  resultam em maior similaridade, mesmo entre estruturas bastante distintas, fazendo com que essas estruturas sejam projetadas em regiões vizinhas na projeção. Um exemplo do impacto de diferentes valores de  $\sigma_0$  no aspecto final da projeção será apresentado na Seção 3.3.2.1.

Por fim, a dissimilaridade entre o par de conformações  $k$  e  $l$  é definida por:

$$\delta_{k,l} = 1 - q^{k,l}. \quad (2.3)$$

Essa medida de dissimilaridade é adimensional e varia no intervalo  $0 \leq \delta_{k,l} < 1$ . Um valor de dissimilaridade igual a zero é alcançado apenas quando as duas conformações são idênticas. É importante ressaltar que essa medida de dissimilaridade depende exclusivamente das distâncias internas, não exigindo qualquer alinhamento estrutural.

Após o cálculo dos valores de dissimilaridade entre cada par de conformações, os resultados são armazenados em uma matriz de dissimilaridade, que será utilizada pela técnica de projeção multidimensional.

### 2.1.2 O algoritmo *Force Scheme*

A matriz de dissimilaridade fornece informações sobre as distâncias entre cada par de conformações, conforme estimado no espaço conformacional multidimensional do sistema biomolecular. O ELViM utiliza essa matriz para realizar um processo de projeção multidimensional, gerando um espaço de fase efetivo bidimensional, no qual cada conformação é representada por um ponto no plano. Para isso, um processo de otimização é empregado, garantindo que as distâncias euclidianas entre os pares de pontos no plano correspondam, tanto quanto possível, aos valores de dissimilaridade originais. Esse procedimento é realizado pelo algoritmo *Force Scheme* (FS), introduzido por Tejada *et al.* (58), que foi escolhido devido ao seu equilíbrio entre acurácia e eficiência computacional (56).

O FS é um método de projeção multidimensional orientado por forças. Nesse método, os pontos no plano podem ser concebidos como massas conectadas por molas, sendo o comprimento natural de cada mola igual à dissimilaridade entre os pares conectados (27,74). Assim, pares de pontos cujas distâncias euclidianas sejam maiores do que as dissimilaridades estão sob ação de forças atrativas, enquanto pares de pontos cujas distâncias sejam menores do que as dissimilaridades estão sob ação de forças repulsivas.

Embora alguns métodos orientados por forças utilizem equações diferenciais para alcançar o estado de equilíbrio do sistema, o FS emprega um método iterativo capaz de convergir com um número menor de iterações quando comparado a outros métodos baseados em forças (74). Nesse algoritmo, cada conformação  $k$ , com coordenadas  $\vec{x}_k$  no espaço de fase multidimensional, é projetada para um ponto  $x'_k$ , representado no plano cartesiano pelas coordenadas  $\vec{x}'_k$ . Inicialmente, um número de pontos igual ao número de conformações é disposto aleatoriamente sobre o plano, formando a projeção inicial.

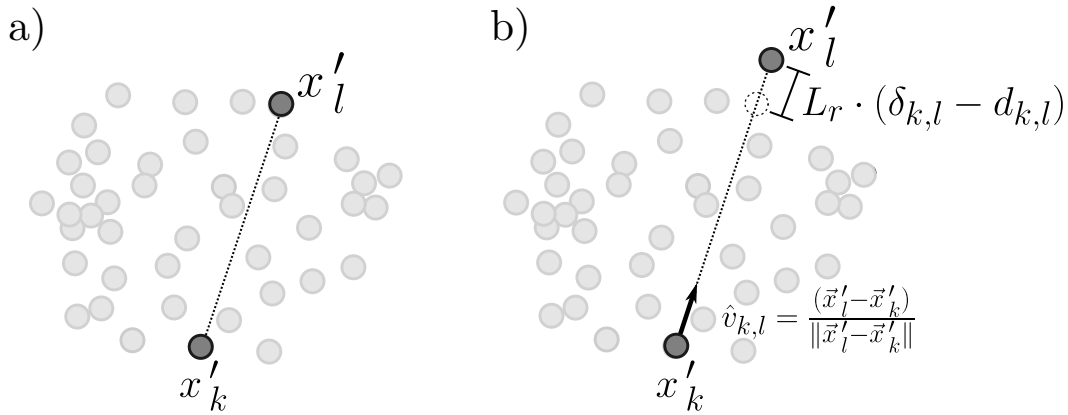
Em cada passo do algoritmo, uma conformação  $k$  é escolhida como referência, e as posições de todas as outras conformações  $l$  ( $l \neq k$ ) são ajustadas de modo que a distância euclidiana entre os pares de pontos  $d_{k,l}(\vec{x}'_k, \vec{x}'_l)$  tenda aos valores de dissimilaridade estimados no espaço multidimensional  $\delta_{k,l}(\vec{x}_k, \vec{x}_l)$ . Esse ajuste ocorre na direção do vetor  $\vec{v}_{k,l} = (\vec{x}'_l - \vec{x}'_k)$ , como ilustrado na Figura 5, e é dado pelo produto entre uma taxa de aprendizado  $L_r$  e o erro residual  $\Delta_{k,l}$  definido como:

$$\Delta_{k,l} = \delta_{k,l} - d_{k,l}. \quad (2.4)$$

Dessa forma, o método busca minimizar a função custo (ou função perda) definida como:

$$E = \sum_{(k,l)} |\delta_{k,l} - d_{k,l}|. \quad (2.5)$$

Figura 5 – Iteração do algoritmo *Force Scheme*. Em cada passo, uma conformação  $k$  é escolhida aleatoriamente como referência, e as posições de todas as outras conformações são ligeiramente ajustadas em relação a ela. a) O algoritmo calcula a distância euclidiana entre um par selecionado,  $x'_k$  e  $x'_l$ , e a compara com a dissimilaridade entre as conformações  $k$  e  $l$ . b) O ponto  $x'_l$  é aproximado ou afastado de  $x'_k$  na direção definida por  $\hat{v}_{k,l}$ , de modo que  $dk,l$  tenda a  $\delta k,l$ .



Fonte: Elaborada pelo autor (2025).

O algoritmo FS empregado no ELViM é descrito da seguinte forma:

---

**Algoritmo 1:** Algoritmo *Force Scheme*

---

**Entrada:** Matriz de dissimilaridades

**Saída:** Coordenadas da projeção bidimensional

```

1 Inicializa projeção inicial  $X_0$  aleatoriamente
2 enquanto  $i \leq i_{max}$  faça
3     para cada  $x'_k$  (ordem aleatória) faça
4         para cada  $x'_{l \neq k}$  (ordem aleatória) faça
5              $\delta_{k,l} \leftarrow \delta(\vec{x}_k, \vec{x}_l)$  ; // Dissimilaridade
6              $d_{k,l} \leftarrow \|\vec{x}'_l - \vec{x}'_k\|$  ; // Distância euclidiana
7              $\vec{x}'_l \leftarrow \vec{x}'_l + L_r \cdot (\delta_{k,l} - d_{k,l}) \cdot (\vec{x}'_l - \vec{x}'_k) / \|\vec{x}'_l - \vec{x}'_k\|$ 
8         fim
9     fim
10 fim
11 retorna Projeção Final X

```

---

Note que, para cada ponto  $x'_k$  escolhido como referência, a posição de todos os outros pontos é ajustada, movendo-os por um fator  $L_r$  do erro residual  $\Delta_{k,l}$ , na direção que une os dois pontos, dada pelo versor  $\hat{v}_{k,l} = (\vec{x}'_l - \vec{x}'_k) / \|\vec{x}'_l - \vec{x}'_k\|$ . Uma iteração é

considerada completa quando todos os pontos foram escolhidos como referência. Assim, cada ponto tem sua posição ajustada  $(N - 1)$  vezes por iteração, sendo  $N$  o número de conformações utilizadas. Usualmente, um número de iterações igual à raiz quadrada do número de pontos ( $i_{max} = \sqrt{N}$ ) é suficiente para alcançar a convergência. Nas primeiras versões do ELViM, uma taxa de aprendizado constante,  $L_r = 1/8$ , era utilizada. Mais detalhes sobre esse parâmetro serão apresentados na Seção 3.2.1.

## 2.2 VARIÁVEIS COLETIVAS

Devido à alta dimensionalidade dos dados gerados em simulações moleculares, é comum descrever a dinâmica conformacional do sistema ou acompanhar o processo de enovelamento por meio de um conjunto de grandezas conhecido como variáveis coletivas (CVs, do inglês *collective variables*). As CVs são geralmente definidas como funções das coordenadas atômicas e têm como objetivo capturar os aspectos macroscópicos mais relevantes do comportamento do sistema em estudo. Um parâmetro de ordem corresponde a uma CV, ou a uma combinação de CVs, capaz de caracterizar o grau de organização de um sistema físico. Esses parâmetros evidenciam propriedades estruturais ou geométricas importantes de sistemas biomoleculares, permitindo distinguir, por exemplo, entre diferentes estados metaestáveis (75).

As coordenadas de reação, por sua vez, constituem uma classe especial de CVs, empregadas para descrever o progresso de uma reação química ou do enovelamento de uma proteína (75,76). Deve-se notar que esses conceitos são inter-relacionados, e a terminologia associada nem sempre é utilizada de forma consistente na literatura. Exemplos de CVs incluem a quantidade de contatos nativos, o raio de giração, a distância ponta-a-ponta, a quantidade de estrutura secundária, entre outras propriedades relevantes para a análise conformacional.

O ELViM é um método que não depende de CVs, nem de conformações de referência, para projetar o espaço de fase efetivo. No entanto, é possível obter informações adicionais a partir dessas projeções utilizando mapas de cores para representar, sobre os pontos projetados, os valores de CVs estimados para cada conformação. Essa abordagem não apenas facilita a interpretação das regiões do espaço de fase efetivo, como também permite identificar padrões, tendências e correlações entre diferentes regiões da projeção e os comportamentos estruturais descritos pelas CVs. É importante destacar que o uso do mapa de cores não influencia as posições dos pontos na projeção, as quais são determinadas exclusivamente pela métrica de dissimilaridade adotada. A seguir, serão listadas as CVs utilizados com esse propósito ao longo desta tese.

O parâmetro  $Q_w$  pode ser entendido como uma medida da similaridade de uma conformação ao estado nativo (ou a um estado de referência arbitrário). Esta coordenada é usualmente utilizada para descrever o processo de enovelamento de proteínas (23).

Matematicamente, considerando uma conformação qualquer e um estado de referência, denotados respectivamente por  $k$  e  $N$ ,  $Q_w$  é definido da seguinte forma:

$$Q_w(k) = \frac{1}{N_p} \sum_{i,j \in S} \exp \left[ \frac{-(r_{i,j}^k - r_{i,j}^N)^2}{2\sigma_{i,j}^2} \right]. \quad (2.6)$$

Note que  $Q_w(k) = q^{k,N}$  (Equação 2.1), sendo utilizado para estimar a similaridade de todas as conformações ao estado nativo.  $Q_w$  é uma medida adimensional, sendo igual a 1 se a conformação for idêntica ao estado nativo e tendendo a zero para conformações estruturalmente muito distintas do estado nativo.

A fração de contatos nativos,  $Q$  (24), é uma variável coletiva amplamente utilizada como coordenada de reação em estudos de enovelamento de proteínas. Essa grandeza é definida como:

$$Q(k) = \frac{1}{N_S} \sum_{i,j \in S} \frac{1}{1 + \exp [\beta^0 (r_{ij}^k - \lambda r_{ij}^N)]}, \quad (2.7)$$

sendo, novamente,  $r_{ij}^k$  a distância entre os átomos  $i$  e  $j$  na conformação  $k$ , e  $r_{ij}^N$  a distância correspondente no estado nativo.  $S$  denota o conjunto dos  $N_S$  pares de átomos pesados que estão separados no mínimo por três resíduos e estão em contato na estrutura nativa. O parâmetro  $\beta^0$  controla a suavização da função sigmoide, que define se o contato está ou não formado, enquanto  $\lambda$  representa um fator de tolerância para que o contato seja considerado formado. Esses parâmetros devem ser ajustados de acordo com o modelo utilizado, sendo usualmente adotado para modelos CG  $\beta^0 = 50 \text{ nm}^{-1}$  e  $\lambda = 1,2$  (77).

O raio de giração ( $R_g$ ) quantifica o grau de compactação da cadeia polipeptídica em relação ao seu centro de massa. Conformações mais compactas apresentam valores menores de  $R_g$ , enquanto conformações mais abertas apresentam valores maiores. Dada uma conformação, o raio de giração é definido por:

$$R_g = \sqrt{\frac{\sum_{i=1}^N m_i \|\vec{r}_i - \vec{r}_{CM}\|^2}{\sum_{i=1}^N m_i}}, \quad (2.8)$$

em que  $N$  é o número de átomos,  $m_i$  e  $\vec{r}_i$  correspondem, respectivamente, à massa e à posição do átomo  $i$ , e  $\vec{r}_{CM}$  é a posição do centro de massa da proteína, calculada como:

$$\vec{r}_{CM} = \frac{\sum_{i=1}^N m_i \vec{r}_i}{\sum_{i=1}^N m_i}. \quad (2.9)$$

O RMSD, por sua vez, é uma medida de dissimilaridade estrutural que pode ser calculada entre pares de conformações ou em relação a um estado de referência. Como o cálculo do RMSD considera as coordenadas atômicas das estruturas, é necessário

realizar um alinhamento estrutural prévio para eliminar efeitos de rotação e translação. Considerando uma conformação  $k$  e o estado nativo  $N$ , com alinhamento estrutural prévio, o RMSD é definido por:

$$RMSD(k) = \sqrt{\frac{1}{N} \sum_{i=1}^N \|\vec{r}_i^k - \vec{r}_i^N\|^2}, \quad (2.10)$$

sendo que  $\vec{r}_i^k$  e  $\vec{r}_i^N$  correspondem às posições do átomo  $i$  nas conformações  $X$  e  $N$ , respectivamente. Valores menores de RMSD indicam maior proximidade da conformação ao estado nativo.

Essas métricas complementam  $Q_w$  e  $Q$ , fornecendo informações adicionais sobre a compactação estrutural e o desvio em relação ao estado nativo. No restante deste trabalho, essas coordenadas serão utilizadas para explorar a dinâmica e o envelhecimento de biomoléculas através do espaço de fase gerado pelo ELViM.

### 2.3 DENSIDADE DE ESTADOS E ENERGIA LIVRE EFETIVAS

Em gráficos de dispersão contendo muitos pontos, a sobreposição de dados pode dificultar a percepção e comparação de regiões com diferentes densidades. No entanto, a distribuição de densidade na projeção ELViM contém informação relevante, pois esse método busca preservar as relações de distância estimada no espaço multidimensional. Assim, espera-se que a projeção ELViM mantenha, tanto quanto possível, a estrutura da superfície de energia do sistema no espaço reduzido.

Sob essa hipótese e assumindo que a amostragem siga uma distribuição de Boltzmann, a densidade de pontos no espaço efetivo gerado pelo ELViM reflete a distribuição de estados do sistema. Em uma simulação de dinâmica molecular clássica, conduzida sob temperatura absoluta constante, sem viés e com amostragem suficiente, regiões de alta densidade na projeção correspondem idealmente a mínimos da superfície de energia livre.

A densidade do espaço reduzido pode ser estimada por histogramas ou pela técnica de Estimativa de Densidade por Núcleo (KDE, do inglês *Kernel Density Estimation*). Enquanto os histogramas particionam o espaço em células de tamanho fixo e contam o número de pontos em cada uma, o KDE fornece uma estimativa contínua da densidade ao aplicar funções núcleo (como a Gaussiana) ao redor de cada ponto. Neste trabalho, as densidades serão estimadas por KDE com núcleo Gaussiano, como definido a seguir.

#### 2.3.1 Superfície de Energia Livre Efetiva

Como previamente discutido, é comum projetar superfícies de energia livre de sistemas biomoleculares em coordenadas de reação ou coordenadas efetivas. Para um

sistema com amostragem suficiente, a probabilidade de se observar um microestado com coordenadas  $\vec{r}$  e energia  $E(\vec{r})$  é dada por

$$P(\vec{r}) = \frac{\Omega(E(\vec{r}))e^{-\beta E(\vec{r})}}{Z}, \quad (2.11)$$

onde  $\beta = 1/(k_B T)$ ,  $k_B$  é a constante de Boltzmann,  $T$  é a temperatura absoluta,  $\Omega(E)$  representa o número de microestados com energia  $E$  e  $Z$  é a função de partição do sistema, dada por

$$Z = \sum_E \Omega(E)e^{-\beta E}. \quad (2.12)$$

Utilizando as definições de entropia,  $S(E) = k_B \ln \Omega(E)$ , e de energia livre,  $F = E - TS$ , a Equação 2.11 pode ser reescrita como

$$P(\vec{r}) = \frac{e^{-\beta F(\vec{r})}}{Z}, \quad (2.13)$$

de onde se obtém

$$F(\vec{r}) = -k_B T \ln P(\vec{r}) - k_B T \ln Z. \quad (2.14)$$

Note que o termo  $-k_B T \ln Z$  é uma constante, utilizada para definir a escala e que não influencia no cálculo das diferenças de energia livre entre macroestados. Considerando agora uma coordenada de reação  $\xi = f(\vec{r})$ , a distribuição de probabilidade associada é dada por

$$P(\xi) = \int P(\vec{r}) \delta(\xi - \xi(\vec{r})) d\vec{r} = \frac{Z(\xi)}{Z}, \quad (2.15)$$

sendo  $\delta(x)$  a função Delta de Dirac. Aplicando-se o logaritmo e multiplicando todos os termos por  $k_B T$ , obtém-se

$$F(\xi) = -k_B T \ln Z(\xi) = -k_B T \ln P(\xi) - k_B T \ln Z. \quad (2.16)$$

Assim, a energia livre em função da coordenada de reação pode ser estimada a partir de histogramas ou KDE da distribuição  $P(\xi)$ . De forma análoga, uma superfície bidimensional pode ser obtida considerando a distribuição conjunta  $P(\xi, \zeta)$ , onde  $\zeta$  é uma segunda coordenada de reação, ou por coordenadas intrínsecas obtidas via redução de dimensionalidade.

### 2.3.2 Estimativa de densidade por núcleo

O KDE é uma técnica não paramétrica para estimar a função densidade de probabilidade de uma variável aleatória a partir de um conjunto de amostras. Para  $n$  observações  $\{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n\}$  no espaço  $d$ -dimensional, a estimativa KDE é dada por

$$\rho(\vec{x}) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{\|\vec{x} - \vec{x}_i\|}{h}\right), \quad (2.17)$$

sendo  $\vec{x}$  a posição na qual a densidade é estimada,  $h > 0$  o parâmetro de largura de banda (*bandwidth*) e  $K(u)$  a função núcleo, que é simétrica e satisfaz

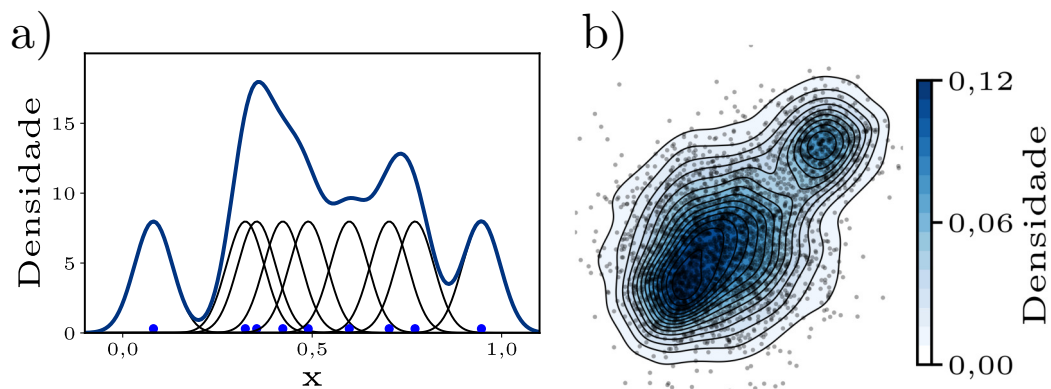
$$\int_{\mathbb{R}^d} K(u) du = 1, \quad (2.18)$$

Para o núcleo Gaussiano, a função  $K(u)$  é definida como

$$K(u) = \frac{1}{(2\pi)^{d/2}} e^{-u^2/2}. \quad (2.19)$$

Em resumo, essa técnica aplica um núcleo Gaussiano centrado em cada ponto de dado para obter uma estimativa suavizada da densidade de probabilidade. A largura de banda controla o grau de suavização e deve ser ajustada em cada caso para evitar suavização excessiva ou insuficiente. A densidade de probabilidade estimada é calculada como a soma normalizada de todos os núcleos.

Figura 6 – Estimativa de densidade por núcleo Gaussiano. (a) Um núcleo Gaussiano centrado em cada ponto de dado. A densidade estimada resulta da soma normalizada de todas as contribuições. A curva em azul escuro indica, para facilitar a visualização, a soma não normalizada dos núcleos. (b) Aplicação da técnica em uma projeção bidimensional.



Fonte: Elaborada pelo autor (2025).

A Figura 6.a ilustra a aplicação da técnica para um conjunto de pontos gerados aleatoriamente em uma dimensão. Os núcleos Gaussianos individuais são representados em

preto, enquanto a densidade estimada é mostrada em azul. Para facilitar a visualização, a densidade estimada não foi normalizada. Já a Figura 6.b ilustra a aplicação da técnica em uma dispersão bidimensional, como no caso da projeção ELViM, com a densidade de probabilidade representada por um mapa de curvas de nível.

### 3 ELViM: NOVA IMPLEMENTAÇÃO E ESTUDO DE CASO COM O PEPTÍDEO MP1

Este capítulo é baseado no artigo: “*ELViM: Exploring Biomolecular Energy Landscapes through Multidimensional Visualization* (61).”

#### 3.1 INTRODUÇÃO

Em sua primeira implementação (55,56), o ELViM utilizava um conjunto de códigos e rotinas para extração das coordenadas atômicas, cálculo da matriz de dissimilaridade e automatização das etapas. Além disso, um código para o FS em linguagem R era utilizado (uma versão similar encontra-se disponível em <https://rdrr.io/cran/mp/man/forceScheme.html>). No entanto, tal implementação tinha um elevado custo computacional, considerando os sistemas biomoleculares de interesse.

Durante este projeto de doutorado, uma nova versão do código ELViM foi desenvolvida, visando o aumento da eficácia computacional e a integração de todas as etapas em um único programa. Além disso, alguns parâmetros, como a taxa de aprendizado e a métrica de similaridade foram generalizados para permitir uma maior adaptação a diversos sistemas. Este trabalho resultou na criação de um repositório público e de artigo científico (61), com intuito de detalhar a metodologia e explorar o espaço de parâmetros do ELViM. Este capítulo detalha tais mudanças e apresenta um estudo de caso, indicando a reprodutibilidade e o impacto de diferentes parâmetros no aspecto final da projeção.

#### 3.2 NOVA IMPLEMENTAÇÃO EM PYTHON

O aumento de eficiência computacional partiu de uma versão do algoritmo FS elaborada pelo grupo do Dr. Fernando V. Paulovich (Eindhoven University of Technology, Holanda), disponível em <https://github.com/fpaulovich/dimensionality-reduction>. Esta nova versão, escrita em *Python*, se beneficia do processamento paralelo ao incorporar funções do módulo Numba (78), que otimiza o código *Python* e o executa de forma eficiente em *hardware* CPU. De modo a incorporar todas as funções necessárias em um único código, foram desenvolvidas neste trabalho as seguintes funcionalidades:

- Implementação de rotinas que utilizam o MDtraj (79), uma biblioteca *Python* para análise de dinâmica molecular, para manipular as trajetórias (ou *ensembles*) moleculares e extrair a matriz de coordenadas atômicas.

- Implementação de rotina para cálculo da matriz de dissimilaridade, utilizando as mesmas bibliotecas do módulo Numba.
- Generalização do parâmetro de aprendizado  $L_r$ , permitindo o uso de regimes variável e constante.
- Implementação de código executável em linha de comando. O programa possui comando *help* que lista as possíveis opções de entrada e saída e indica os comandos para personalização dos parâmetros. Nesta versão, que está disponível no repositório (<https://github.com/VLeiteGroup/ELViM>), todos os parâmetros de similaridade e da rotina do FS podem ser controlados pelo usuário.
- Implementação de nova ferramenta de seleção (*lasso*), que permite obter os índices das conformações dentro uma região de formato arbitrário, definida pelo usuário.
- Implementação de algoritmo para obtenção de estruturas prevalentes em uma dada região da projeção.

O ELViM aceita como entrada arquivos de trajetória molecular em diversos formatos compatíveis com o MDTraj. A entrada pode ser fornecida como um único arquivo contendo a trajetória completa em formato PDB ou como uma trajetória binária (por exemplo, XTC ou DCD) acompanhada de um arquivo de topologia de referência, como um PDB. Além disso, o ELViM permite salvar a matriz de dissimilaridade calculada em formato binário *Python*, possibilitando seu uso posterior com diferentes parâmetros de projeção. Também é possível utilizar matrizes de dissimilaridade pré-calculadas, incluindo aquelas obtidas com outras métricas. A saída do programa consiste em um arquivo de texto contendo as coordenadas cartesianas que representam as posições das conformações biomoleculares no espaço de fase efetivo.

### 3.2.1 A taxa de aprendizado $L_r$

A taxa de aprendizado  $L_r$  é o parâmetro do FS responsável por controlar a fração do erro residual (Equação 2.4) utilizada para ajustar as posições dos pontos a cada iteração. Dessa forma,  $L_r$  exerce influência tanto na convergência quanto no aspecto final da projeção. De maneira geral, valores muito baixos de  $L_r$  restringem o movimento dos pontos, o que pode confiná-los em vizinhanças inadequadas e exigir um número elevado de iterações até a convergência da projeção. Por outro lado, valores muito altos de  $L_r$  podem causar instabilidades e distorções na projeção.

A forma funcional de  $L_r$  pode ser escolhida de diferentes maneiras, incluindo valores constantes ou, mais comumente, decrescentes. Neste trabalho, propomos uma abordagem que permite combinar dois regimes: variável e constante. No regime variável, ou de *annealing*, a taxa de aprendizado inicia com o valor  $L_{r_0}$  e decresce ao longo das

iterações até atingir um valor mínimo  $L_{r_{min}}$ . Esse regime permite iniciar com valores mais elevados de  $L_r$ , em comparação aos regimes constantes, com o objetivo de evitar que os pontos fiquem presos em regiões incorretas. Em seguida, um regime constante, no qual  $L_r = L_{r_{min}}$ , pode ser utilizado para realizar um ajuste fino. Matematicamente, a taxa de aprendizado na  $i$ -ésima iteração é definida como:

$$L_r[i] = \max \left\{ L_{r_0} \left( 1 - \frac{i}{i_{max}} \right)^D, L_{r_{min}} \right\}, \quad (3.1)$$

em que  $D$  é o expoente de decaimento e  $i_{max}$  é o número máximo de iterações. Observa-se que é possível optar por uma taxa de aprendizado constante, definindo  $D = 0$  e  $L_{r_0} = L_{r_{min}}$ , ou por um regime de decaimento puro, fazendo  $L_{r_{min}} = 0$ . As faixas típicas de valores utilizadas para tais parâmetros são:  $1/8 < L_{r_0} < 0,5$ ,  $0,95 < D < 3$ , e  $0 < L_{r_{min}} < 1/8$ . Essas combinações fornecem maior flexibilidade para diferentes sistemas.

Um aspecto adicional importante é a escolha de  $L_{r_{min}}$  no regime variável, ou de  $L_r$  no caso de um regime constante. Sistemas complexos, que apresentam grandes mudanças conformacionais, podem gerar conformações que se comportam como *outliers* em relação à métrica do ELViM. Tais conformações não possuem vizinhos próximos na matriz de dissimilaridade ( $\delta_{i,j} < 0,2$ ). Representar essas relações de vizinhança no plano é inerentemente difícil, o que leva a pares de pontos com maiores valores de  $\Delta_{k,l}$ . Nesse contexto, se a taxa  $L_r$  mantiver valores significativos nas últimas iterações, pode ocorrer a formação de regiões semelhantes a ilhas, com os *outliers* isolados no centro. Esse processo pode distorcer ligeiramente a vizinhança local, sem comprometer os aspectos globais da projeção. Contudo, recomenda-se uma busca criteriosa de parâmetros para minimizar esses efeitos. Um exemplo ilustrativo do impacto da escolha desses parâmetros no aspecto final da projeção será apresentado na Seção 3.3.2.1.

### 3.2.2 Assinatura Conformacional Local

Para caracterizar regiões arbitrárias da projeção ELViM, foi desenvolvido um algoritmo para obter estruturas representativas de uma dada região da projeção, denominadas assinatura conformacional local (LCS, do inglês Local Conformational Signature). Este método visa identificar as conformações predominantes em uma região específica da projeção, que pode ter formato arbitrário. A LCS leva em consideração tanto a conformação centroide quanto um número  $n$  arbitrário de vizinhos próximos, com base na similaridade estrutural, aferida pelo desvio médio quadrático das distâncias interatômicas (dRMSD).

O dRMSD entre duas conformações  $k$  e  $l$  é dado por:

$$\text{dRMSD}(k, l) = \sqrt{\frac{1}{N_p} \sum_{i < j} (r_{i,j}^k - r_{i,j}^l)^2}, \quad (3.2)$$

sendo que  $r_{i,j}^k$  e  $r_{i,j}^l$  são as distâncias interatômicas entre os átomos  $i$  e  $j$  das conformações  $k$  e  $l$ , respectivamente, e  $N_p$  é o número total de pares de átomos considerados. O dRMSD foi escolhido por sua capacidade de capturar diferenças estruturais sem exigir alinhamento entre as conformações, além de apresentar boa correlação com termos energéticos (42).

O processo para calcular a LCS envolve os seguintes passos:

- Geração do gráfico da projeção: Utiliza-se a biblioteca *Matplotlib* (80) para gerar o gráfico da projeção.
- Seleção da região de interesse: O usuário seleciona, com a ajuda da ferramenta *lasso*, uma região de contorno arbitrário, identificando os índices das conformações contidas nessa área.
- Cálculo da matriz de dRMSD: Com as conformações selecionadas, calcula-se a matriz de  $dRMSD$ , e a conformação centroide é determinada. A centroide é a conformação que minimiza o valor médio do  $dRMSD$  em relação às demais conformações da região selecionada.
- Formação da LCS: A LCS é formada pela conformação centroide sobreposta aos seus  $n$  primeiros vizinhos, selecionados com base nos valores de dRMSD, ou seja, aqueles com as menores diferenças de distância em relação à centroide.

É importante destacar que, ao contrário de uma análise de agrupamento (*clustering*), a determinação da LCS visa representar as conformações predominantes em uma região da projeção, sem a necessidade de separações explícitas entre diferentes grupos. Uma das principais vantagens dessa implementação é a capacidade de selecionar regiões com formato arbitrário, permitindo seguir as distribuições naturais dos pontos na projeção. Além disso, o número de conformações  $n$  incluídas na LCS pode ser ajustado, oferecendo um reflexo da heterogeneidade estrutural em torno da conformação centroide, o que serve como um indicativo qualitativo do nível de agrupamento estrutural local.

### 3.3 PEPTÍDEO MP1: UM ESTUDO DE CASO

Para ilustrar a metodologia e explorar o espaço de parâmetros do método, será analisada, como estudo de caso, uma trajetória de dinâmica molecular clássica do peptídeo Polybia-MP1, ou simplesmente MP1 (81). O MP1 é um peptídeo de ação antimicrobiana pertencente à família dos mastoparanos, extraído do veneno da vespa brasileira *Polybia paulista*. Detalhes adicionais sobre a ação desse peptídeo serão abordados no Capítulo 6, no qual também será discutida a sua interação com uma membrana modelo. Neste capítulo, uma trajetória do MP1 isolado em água será analisada para exemplificar a metodologia, ilustrar o impacto de diferentes parâmetros no aspecto final e na reprodutibilidade da projeção ELViM.

### 3.3.1 Sistema analisado

As trajetórias que serão analisadas nesta seção foram obtidas pela Dra. Ingrid B. S. Martins. O MP1 foi construído na configuração de hélice alfa utilizando o módulo *Molefactory* do VMD (82) e foi simulado com todos os átomos, em caixa de água com adição de NaCl à concentração de 150 mM. O sistema foi mantido à temperatura de 330 K e pressão de 1 atm, utilizando respectivamente o termostato (83) e pistão de Langevin (84). O campo de força CHARMM 36 (85) foi adotado e o sistema foi simulado no NAMD (86), sendo integrado com um passo de 2 fs e condições de contorno periódicas. O algoritmo SHAKE (87) foi utilizado para restringir os comprimentos das ligações covalentes, e a geometria das moléculas de água foi preservada utilizando o algoritmo SETTLE (88). As interações de van der Waals foram calculadas com um corte de 12 Å, com uma distância de transição de 10 Å, e a interação de longo alcance foi tratada utilizando o método *Particle Mesh Ewald* (PME) (89). Mais detalhes metodológicos são fornecidos na referência (61). Ao longo de 600 ns de simulação, algumas transições entre os estados desnaturado e parcialmente enovelado foram observadas. Da trajetória obtida, 30000 conformações foram selecionadas, de forma igualmente espaçadas ao longo da trajetória, para gerar a projeção ELViM. A execução do código ELViM, neste caso, levou cerca de 27 minutos em um computador equipado com um processador Intel Core i7 de 6 núcleos e 12 *threads*, operando a 2,6 GHz, e 32 GB de RAM.

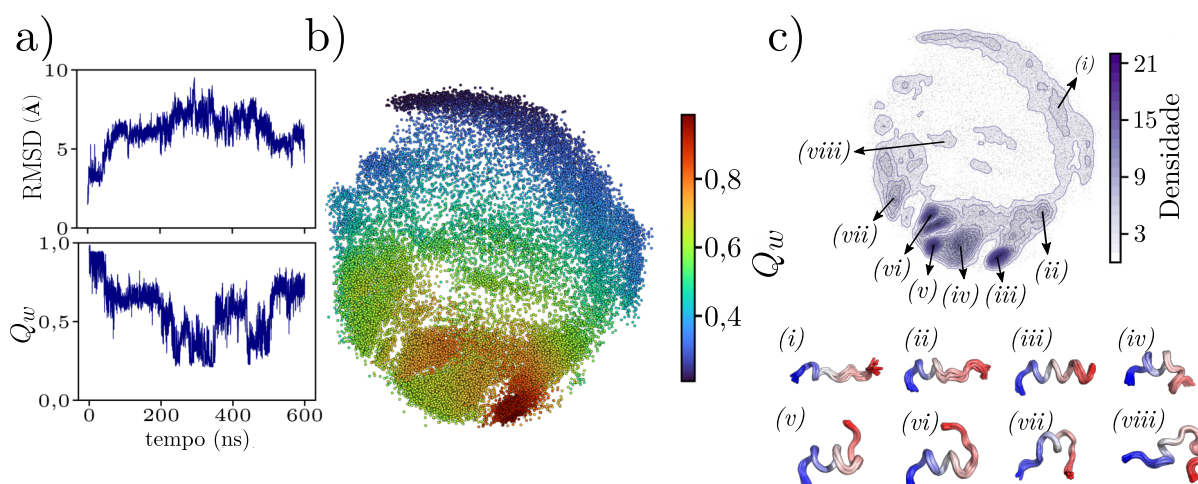
### 3.3.2 Resultados

Considerando os 30000 quadros selecionados para a projeção ELViM, a Figura 7.a ilustra a evolução temporal de  $Q_w$ , calculado em relação à conformação inicial. Nesse contexto,  $Q_w$  assume o valor 1 para uma hélice alfa ideal e atinge 0,2 para conformações completamente desenoveladas. Além disso, a evolução temporal dos valores de RMSD, indicando o desvio estrutural em relação à conformação inicial, também é apresentada.

A projeção ELViM gerada é apresentada na Figura 7.b. A dissimilaridade foi calculada utilizando os seguintes parâmetros:  $\sigma_0 = 1$  Å,  $L_{r_0} = 0,3$ ,  $D = 0,95$  e  $L_{r_{\min}} = 0$ . Conforme descrito no Capítulo 2, cada ponto da projeção representa uma conformação específica. Um mapa de cores foi aplicado, atribuindo a cada ponto uma cor correspondente ao valor de  $Q_w$  calculado para a respectiva conformação. Os eixos coordenados foram removidos, uma vez que, na projeção ELViM, as coordenadas absolutas não possuem significado intrínseco; apenas as distâncias relativas entre os pontos refletem a dissimilaridade estrutural.

Ao analisar a projeção ELViM (Figura 7.b), observa-se que as conformações mais estendidas (representadas em azul escuro) concentram-se na porção superior da projeção, enquanto os estados mais colapsados (em vermelho escuro) situam-se na região inferior. Um aspecto notável da projeção ELViM é que, ao ser colorida com base em coordenadas

Figura 7 – Análise da trajetória de dinâmica molecular (MD) para o peptídeo MP1 em solução aquosa. (a) Evolução temporal da coordenada de reação  $Q_w$  e do RMSD. (b) Projeção ELViM baseada em 30000 conformações selecionadas de forma igualmente espaçada ao longo da trajetória. Os pontos estão coloridos de acordo com os valores de  $Q_w$ . Os eixos foram omitidos, pois apenas as distâncias entre os pontos refletem a similaridade estrutural. (c) Estimativa da densidade de estados por meio de KDE com núcleo Gaussiano. Algumas regiões de alta densidade foram selecionadas, e estruturas representativas, obtidas a partir da análise da assinatura conformacional local (LCS), são exibidas com o amino-terminal representado em azul.



Fonte: Adaptada de (61).

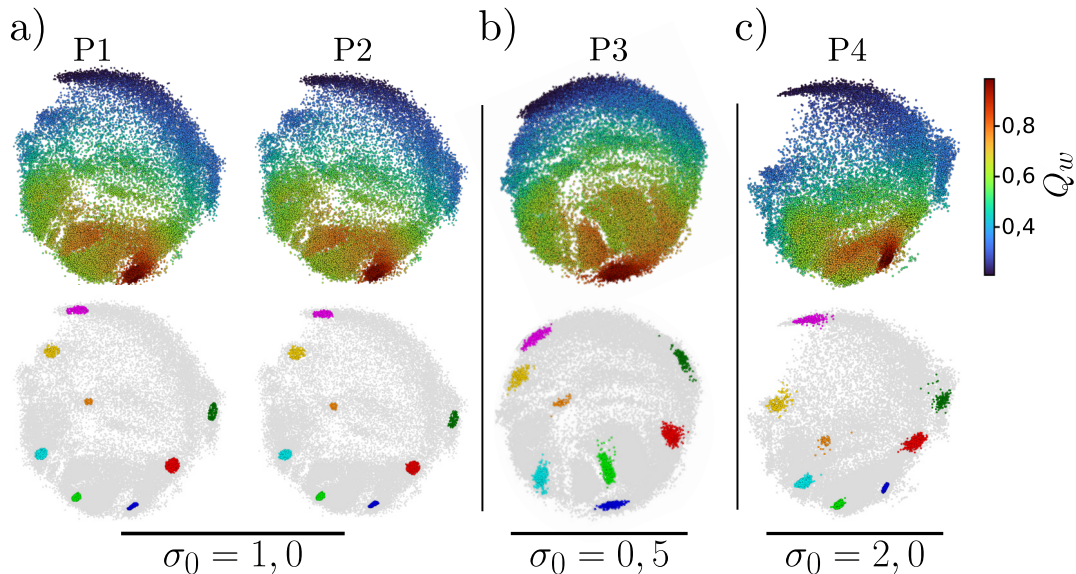
de reação convencionais, as cores variam de maneira aproximadamente contínua ao longo do espaço projetado. Contudo, verifica-se que a densidade de pontos não é homogênea, revelando uma estrutura inerente, com a formação de agrupamentos mais densos. Cabe destacar também que a região associada aos maiores valores de  $Q_w$ , correspondendo à porção inicial e não completamente equilibrada da trajetória, encontra-se ligeiramente deslocada da bacia principal.

Para caracterizar as conformações predominantes no espaço efetivo gerado, a densidade de estados efetiva foi estimada por meio de KDE com núcleo Gaussiano, conforme introduzido na Seção 2.3. O resultado é apresentado por meio de um mapa de curvas de nível na Figura 7.c. Além disso, oito regiões de alta densidade foram selecionadas, e conformações representativas, obtidas com base na análise LCS (Seção 3.2.2), estão indicadas. Cada LCS é composta por 10 conformações sobrepostas, com a extremidade amino-terminal representada em azul.

### 3.3.2.1 Análise qualitativa da reprodutibilidade e do impacto de parâmetros no ELViM

O ELViM é uma técnica essencialmente estocástica, dado que os pontos são inicialmente distribuídos aleatoriamente sobre o plano, e as distâncias são ajustadas em ordem também aleatória. Esse processo resulta em projeções ligeiramente diferentes a cada execução do algoritmo. Entretanto, espera-se que o método convirja para sistemas que não sejam excessivamente complexos, de modo que as variações entre projeções independentes sejam pequenas e não comprometam, em média, a vizinhança local ou a estrutura global da projeção. É importante destacar que diferentes execuções do algoritmo podem produzir rotações e inversões da projeção em torno de eixos arbitrários, o que não afeta a interpretação final dos resultados. Para comparar diferentes réplicas, é possível realizar um alinhamento sem perda de generalidade.

Figura 8 – Reprodutibilidade do ELViM em relação a diferentes execuções e à escolha dos valores de  $\sigma_0$ . (a) Duas réplicas com  $\sigma_0 = 1$ , (b)  $\sigma_0 = 0,5$  e (c)  $\sigma_0 = 2$ . Todas as projeções foram geradas com a mesma taxa de aprendizado ( $L_{r_0} = 0,3$ ,  $D = 0,95$ ,  $L_{r_{\min}} = 0$ ) e um número de iterações  $i_{\max} = \sqrt{N} = 173$ . A projeção P1 é uma reprodução da Figura 7.b.



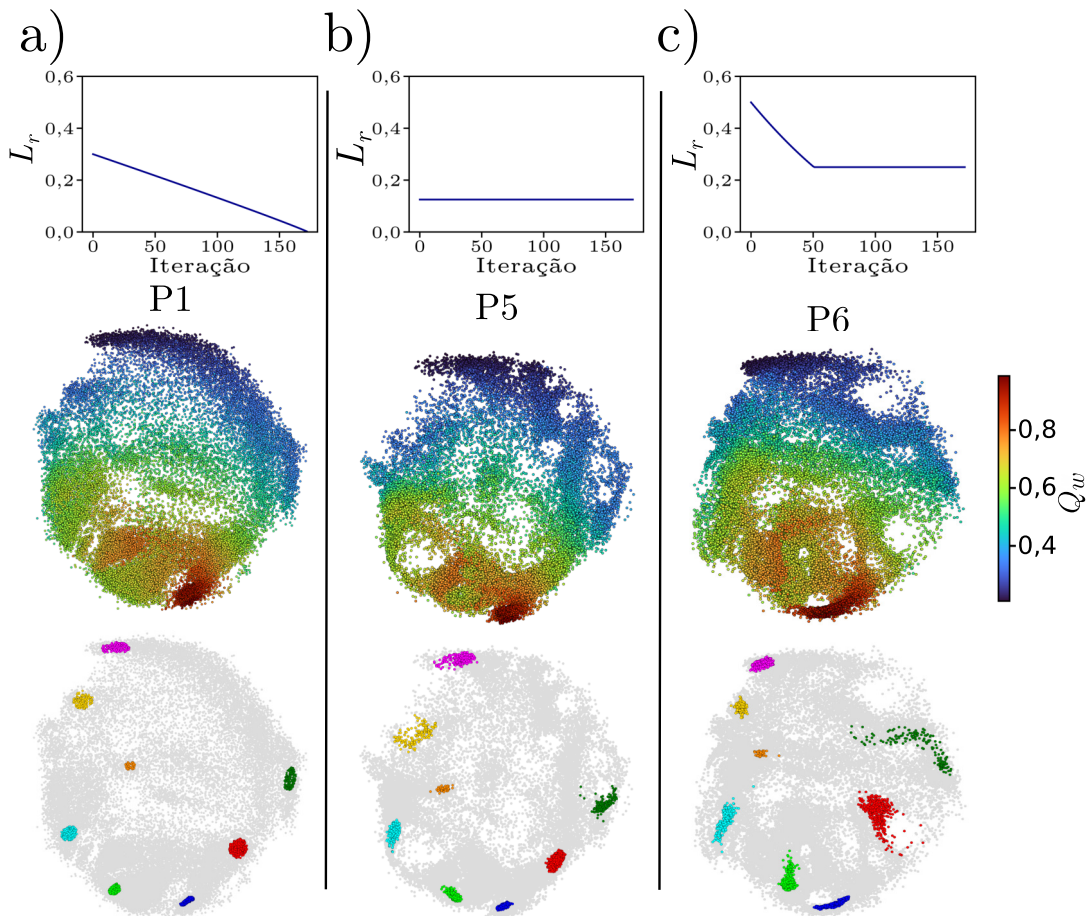
Fonte: Adaptada de (61).

Para ilustrar a reprodutibilidade do ELViM, a Figura 8.a, compara a projeção apresentada anteriormente (Figura 7.b), denominada P1, com uma nova execução independente do programa (projeção P2). O resultado destaca o alto nível de similaridade entre ambas as projeções. Para facilitar uma comparação visual, oito grupos de pontos — que foram selecionados na projeção original e marcados por diferentes cores — são reproduzidos nas demais projeções analisadas. A coesão desses grupos reflete a manutenção da estrutura local, enquanto o ordenamento relativo entre eles indica a preservação da estrutura global

da projeção.

Adicionalmente, as projeções P3 e P4, mostradas nas Figuras 8.b e 8.c, ilustram o impacto de diferentes valores do parâmetro  $\sigma_0$  no aspecto final e na reprodutibilidade do ELViM. Como discutido na Seção 2.1.1,  $\sigma_0$  modifica a matriz de dissimilaridade inicial, de modo que, para um mesmo par de estruturas, a dissimilaridade seria maior com  $\sigma_0 = 0,5$  e menor com  $\sigma_0 = 2$ , em relação ao valor de referência  $\sigma_0 = 1$ . Na projeção P3 ( $\sigma_0 = 0,5$ ), observa-se maior espaçamento entre os pontos da bacia de estados enveloados (em vermelho). Por outro lado, na projeção P4 ( $\sigma_0 = 2$ ), há um maior colapso dessa mesma bacia. Apesar dessas mudanças, a análise das distribuições dos grupos selecionados revela que tanto a vizinhança local de cada grupo quanto o ordenamento relativo entre os grupos são pouco impactados pelas variações em  $\sigma_0$ .

Figura 9 – Reprodutibilidade do ELViM em relação a diferentes escolhas da taxa de aprendizado. Os seguintes parâmetros foram utilizados: (a)  $L_{r_0} = 0,3$ ,  $D = 0,95$ ,  $L_{r_{\min}} = 0$ ; (b)  $L_{r_0} = 0,125$ ,  $D = 0$ ,  $L_{r_{\min}} = 0,125$ ; e (c)  $L_{r_0} = 0,5$ ,  $D = 2,0$ ,  $L_{r_{\min}} = 0,25$ . Todas as projeções utilizaram  $\sigma_0 = 1,0$  e um número de iterações  $i_{\max} = \sqrt{N} = 173$ .



Fonte: Adaptada de (61).

Por fim, para ilustrar os impactos de diferentes escolhas para o parâmetro de

aprendizado na reprodutibilidade e no aspecto global da projeção ELViM, a projeção P1 foi novamente utilizada como referência e duas outras projeções, P5 e P6, foram geradas utilizando-se a mesma matriz de dissimilaridade. Três regimes de taxa de aprendizado foram comparados: um regime de *annealing*, com  $L_r$  tendendo a zero (P1, Figura 9.a); um regime de taxa de aprendizado constante e igual a  $1/8$  (P5, Figura 9.b); e um regime de *annealing* com taxa de aprendizado mínima igual a  $0,25$  (P6, Figura 9.c). Este último caso possui valor não usual de  $L_{r_{\min}}$ , que foi escolhido para enfatizar a formação de ilhas previamente discutida. Note que a formação de ilhas perturba ligeiramente a vizinhança local, mas tem pouco impacto na distribuição das distâncias relativas entre os grupos destacados.

Nesta seção, o impacto dos parâmetros e a reprodutibilidade do método foram explorados por meio de uma análise qualitativa e visual. Algumas métricas de qualidade têm sido propostas para quantificar o grau em que distâncias ou relações de vizinhança são preservadas em projeções multidimensionais (26, 90, 91). No Apêndice B, duas dessas métricas foram aplicadas às projeções apresentadas neste capítulo, considerando abordagens tanto local quanto global. Contudo, vale destacar que tais métricas avaliam predominantemente a eficácia do procedimento de otimização, sem oferecer uma medida direta da capacidade do método em capturar a topologia, muitas vezes desconhecida, do espaço de fase multidimensional (92). Essa limitação torna a comparação entre diferentes métodos de projeção uma tarefa intrinsecamente desafiadora. Neste contexto, apresentaremos nos capítulos 4 e 5 comparações qualitativas da projeção ELViM com superfícies de energia reconstruídas através do uso de coordenadas de reação e de outras técnicas de redução dimensional.

### 3.4 CONCLUSÃO

Neste capítulo, foi apresentada uma reformulação do código do ELViM, agora disponibilizado em um repositório público de acesso livre. Utilizando o peptídeo MP1 como estudo de caso, foram demonstradas as funcionalidades do método e explorado o impacto de diferentes parâmetros sobre o aspecto final e a reprodutibilidade das projeções. As regiões com alta densidade de pontos foram caracterizadas por LCSs, nas quais as conformações apresentaram alto grau de alinhamento estrutural, sugerindo uma vizinhança local bem preservada.

Erros de projeção ou distorções representam uma limitação comum em técnicas de redução de dimensionalidade e visualização multidimensional (26, 27). Essas distorções são causadas por pontos posicionados incorretamente na projeção e podem ser decorrentes de limitações matemáticas inerentes à projeção de dados de alta dimensionalidade em um espaço bidimensional (26). Ao analisar a projeção ELViM, pontos posicionados incorretamente podem se comportar como ruído estrutural na vizinhança local. Nesse sentido,

foi sugerido que a interpretação da projeção seja baseada nas assinaturas estruturais predominantes dentro de uma vizinhança local.

A ferramenta LCS, desenvolvida para este fim, permite identificar conformações prevalentes em regiões arbitrariamente selecionadas da projeção. A análise dessas conformações ao longo das regiões de alta densidade proporciona uma visão intuitiva da distribuição de estados conformacionais representativos. Embora essa abordagem seja qualitativa, ela auxilia na interpretação da projeção, podendo também ilustrar a heterogeneidade estrutural em torno de pontos específicos.

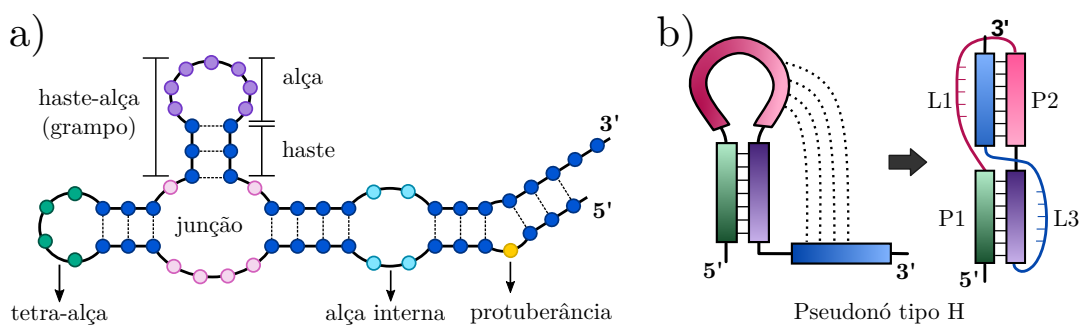
Adicionalmente, o impacto dos parâmetros  $\sigma_0$  e  $L_r$  na reprodutibilidade e na estrutura global da projeção ELViM foi avaliado qualitativamente. Demonstrou-se que o posicionamento relativo de grupos selecionados manteve-se consistente entre diferentes projeções, apesar de pequenas perturbações locais, como a formação de ilhas. O conjunto de características aqui discutido reforçou o potencial do ELViM como uma ferramenta promissora para produzir visualizações intuitivas do relevo energético de biomoléculas, podendo auxiliar na investigação de detalhes e mecanismos moleculares.

## 4 EXPLORANDO A SUPERFÍCIE DE ENERGIA DE MOLÉCULAS DE RNA: A TETRA-ALÇA GCAA E O RIBOCOMUTADOR SAM-II

### 4.1 INTRODUÇÃO

No início do desenvolvimento da Biologia Molecular, o ácido ribonucleico (RNA) era visto apenas como um intermediário passivo na transferência de informação genética, enquanto o ácido desoxirribonucleico (DNA) e as proteínas eram considerados os agentes principais. No entanto, diversas descobertas realizadas nas últimas três décadas demonstraram que o RNA desempenha funções essenciais em múltiplos processos celulares. Além de atuar como mensageiro na síntese proteica (mRNA), o RNA desempenha papéis estruturais — como no RNA transportador (tRNA) e no RNA ribossômico (rRNA) — e apresenta atividade catalítica, como evidenciado pelas ribozimas. Adicionalmente, o RNA regula a expressão gênica por meio de mecanismos que envolvem pequenos RNAs nucleares (snRNAs, do inglês small nuclear RNAs), micro-RNAs (miRNAs), pequenos RNAs interferentes (siRNAs, do inglês small interfering RNAs) e outros RNAs não codificantes (ncRNAs, do inglês non-coding RNAs) (93).

Figura 10 – Estrutura secundária do RNA. (a) Esquema ilustrando a formação de motivos estruturais em fita de RNA. Regiões de hélice são representadas em azul escuro e as demais cores destacam os motivos indicados; (b) Pseudonó do tipo H. Bases não pareadas presentes na alça de um grampo (haste-alça) interagem com outras bases fora do grampo original. A estrutura final é composta por duas regiões de hélice (P1 e P2) e duas alças (L1 e L2).



Fonte: Elaborado pelo autor (2025).

Comparado às proteínas, o RNA possui estrutura química mais simples, sendo composto por um grupo fosfato, uma ribose e quatro bases nitrogenadas: adenina (A), guanina (G), citosina (C) e uracila (U). No entanto, as fitas simples de RNA podem dobrar-se sobre si mesmas, formando estruturas secundárias e terciárias complexas. A estrutura secundária do RNA resulta de interações de pareamento (*base pairing*) entre as bases, originando os motivos estruturais como os ilustrados na Figura 10.a. Os pareamentos entre as bases

C-G e A-U são ditos canônicos (ou de Watson-Crick). No entanto, outros pareamentos não-canônicos são frequentes, como o pareamento G-U. Na estrutura secundária dos RNAs, as regiões formadas pelo pareamento de bases complementares correspondem a hélices e hastes, enquanto as regiões com bases não pareadas formam protuberâncias e alças. A região composta pelo encontro de duas ou mais hélices é denominada junção. Estruturas secundárias mais complexas emergem do pareamento canônico entre bases distantes ao longo da fita, como nos pseudonós. A Figura 10.b ilustra um pseudonó do tipo H, no qual as bases da alça interagem com as bases localizadas fora da haste-alça original.

A estrutura terciária do RNA refere-se à conformação tridimensional que a molécula adquire após a formação dos elementos secundários. Essa organização é estabilizada por interações de longo alcance, incluindo o empilhamento de bases (*base stacking*), pareamentos não canônicos e a associação com íons metálicos, os quais auxiliam na neutralização da carga negativa da cadeia principal de açúcar-fosfato (94). Em geral, a estrutura secundária apresenta alta estabilidade e sofre poucas alterações após a formação da estrutura terciária (95). Dessa forma, os RNAs são considerados moléculas que se enovelam de maneira hierárquica e sequencial, com a estrutura secundária geralmente se formando primeiro e, em grande parte, de maneira independente da estrutura terciária (20,96).

Uma das interações terciárias mais comuns é denominada interação A-menor (do inglês *A-minor*) (97,98). Nessa interação, um resíduo de adenina não pareado interage com o sulco menor de uma hélice canônica distante, estabelecendo diversas ligações de hidrogênio, que contribuem com a estabilização da estrutura terciária. Essas interações ocorrem com frequência em várias moléculas de RNA não codificantes e constituem componentes essenciais de motivos estruturais maiores, como os pseudonós (98). Existem quatro variações dessa interação, sendo a mais comum denominada A-menor tipo I, na qual os átomos N1, N3 e O2' da adenina se posicionam entre os dois átomos O2' do sulco menor do par canônico. Um exemplo dessa interação é apresentado na Figura C.4 do Apêndice C. Usualmente, regiões contendo adeninas consecutivas não pareadas tendem a realizar interações A-menor empilhadas, aumentando a estabilidade dessa interação terciária. Esse caso é usualmente referido com motivo A-menor (98).

Considerando a complexidade da estrutura terciária do RNA, compreender o processo de enovelamento, incluindo *ensembles* de transição e estados armadilhados, é essencial para entender aspectos funcionais relevantes dessas moléculas. Nos últimos anos, diversos métodos teóricos foram desenvolvidos para simular tais sistemas, incluindo novos ou adaptados campos de força, métodos de *enhanced sampling* e modelos *coarse-grain* (96).

Neste capítulo, o ELViM será empregado para estudar o relevo da superfície de energia de duas moléculas de RNA: a tetra-alça GCAA e o ribocomutador SAM-II. As tetra-alças são motivos estruturais conservados e altamente estáveis, que estão presentes

em diversas moléculas de RNA e desempenham papéis cruciais na estabilidade e função dessas moléculas (99). Os ribocomutadores, como o SAM-II, alteram sua estrutura tridimensional ao se ligarem a um metabólito ou íon específico, modulando a expressão gênica ao influenciar processos como a iniciação ou interrupção da tradução (100). O ELViM será utilizado para analisar trajetórias de enovelamento desses sistemas e investigar a dinâmica conformacional de cada uma delas. Os resultados indicam que o ELViM é capaz de construir um espaço efetivo, com bom agrupamento estrutural, permitindo caracterizar os ensembles de estados de transição e as rotas de enovelamento dessas moléculas.

## 4.2 CARACTERIZAÇÃO DOS *ENSEMBLES* DE ESTADOS DE TRANSIÇÃO DA TETRA-ALÇA GCAA

Esta seção é baseada no artigo “*Characterizing the Folding Transition-State Ensembles in the Energy Landscape of an RNA Tetraloop (62)*.”

As tetra-alças são hiperestáveis, devido a extensivas interações de empilhamento e pareamento de bases presentes na estrutura, e apresentam enovelamento rápido em comparação com moléculas maiores de RNA. Sugere-se que as tetra-alças atuem como componentes evolutivamente especializados, auxiliando o enovelamento de moléculas maiores de RNA, pré-organizando estruturas locais e mediando interações terciárias de longo alcance (96, 101, 102).

Nesta seção, o ELViM será empregado para analisar trajetórias de enovelamento da tetra-alça GCAA, com o objetivo de caracterizar a superfície de energia, descrever as rotas de enovelamento e identificar os *ensembles* de estados de transição.

### 4.2.1 Sistema analisado

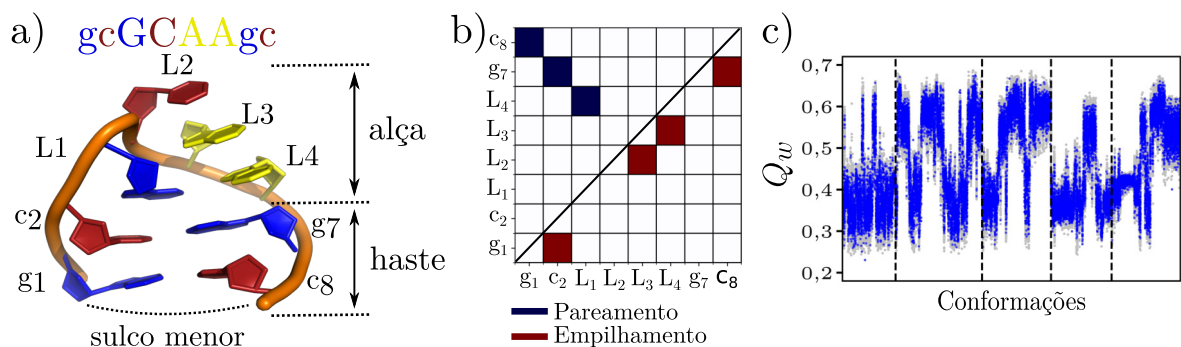
A tetra-alça GCAA (gcGCAA<sub>gc</sub>) apresenta uma estrutura secundária do tipo haste-alça, na qual a haste é formada pelo pareamento de bases complementares e a alça consiste em quatro nucleotídeos. Sua estrutura terciária é ilustrada na Figura 11.a, enquanto a Figura 11.b exibe um mapa de contatos com as interações de pareamento e empilhamento características da estrutura nativa.

Nesta seção, serão analisadas trajetórias de dinâmica molecular clássica, simuladas no supercomputador PSC ANTON-1 (103), pelo grupo do Dr. Angel Garcia (*Los Alamos National Laboratory, Los Alamos, EUA*). Serão consideradas cinco trajetórias independentes, totalizando cerca de 120  $\mu$ s, obtidas com um modelo contendo todos os átomos pesados e solvente explícito. Todas as simulações partiram de estados desnaturados, arbitrariamente selecionados a partir de simulações anteriores de REMD (104). A tetra-alça foi posicionada em uma caixa cúbica contendo 6.428 moléculas de água TIP3P, 133 íons  $K^+$  e 125 íons  $Cl^-$

(com concentração aproximada de 1 M de KCl em excesso), resultando em um sistema com 19.741 átomos. O sistema foi simulado em condições NVT a uma temperatura de 350 K, com o campo de força AMBER-99 para ácidos nucleicos (105), acrescido de modificações propostas por Chen e Garcia (104). O controle de temperatura foi realizado utilizando o termostato de Nosé-Hoover (106, 107), com um passo de integração de 2 fs. Mais detalhes computacionais sobre as simulações podem ser encontrados na referência (62).

Para a projeção ELViM, foram selecionados um a cada quinze quadros da simulação. Além disso, as transições entre os estados desnaturado e enovelado foram amostradas sequencialmente, resultando em 28.560 pontos na projeção. Na figura 11.c, as conformações selecionadas para a projeção ELViM são indicadas por pontos azuis e as linhas verticais tracejadas separam as diferentes simulações independentes. A trajetória oscila entre estados desnaturados ( $Q_w \sim 0,35$ ) e enovelados ( $Q_w \sim 0,6$ ). Foram amostradas 32 transições entre os estados desnaturado e enovelado. Os valores mais baixos de  $Q_w$  observados para o estado enovelado são devido, em parte, à escolha do parâmetro  $\sigma_0$ , sendo adotado o mesmo formalismo aplicado na métrica de similaridade do ELViM. A evolução temporal das coordenadas de reação usuais, RMSD e  $R_g$ , são mostradas na Figura C.1 (Apêndice C).

Figura 11 – (a) Estrutura cristalográfica (PDB:1ZIH) da tetra-alça GCAA. O esquema de cores indica a base nitrogenada de cada nucleotídeo, sendo amarelo para adenina (A), vermelho para citosina (C) e azul para guanina (G). (b) Mapa de interações da estrutura nativa. Os pareamentos são indicados em azul, acima da diagonal principal, e os empilhamentos são mostrados em vermelho, abaixo da diagonal. (c) Evolução da similaridade à estrutura cristalográfica,  $Q_w$ , para as cinco trajetórias de MD. que estão separadas por linhas pontilhadas na vertical. As conformações selecionadas para a projeção ELViM são mostradas como pontos azuis.



Fonte: Adaptada de (62).

Para o cálculo da dissimilaridade (Eq. 2.1), foram consideradas as posições de todos os átomos exceto hidrogênios. O parâmetro  $\sigma_{i,j}$  (Equação 2.2) foi ajustado para:

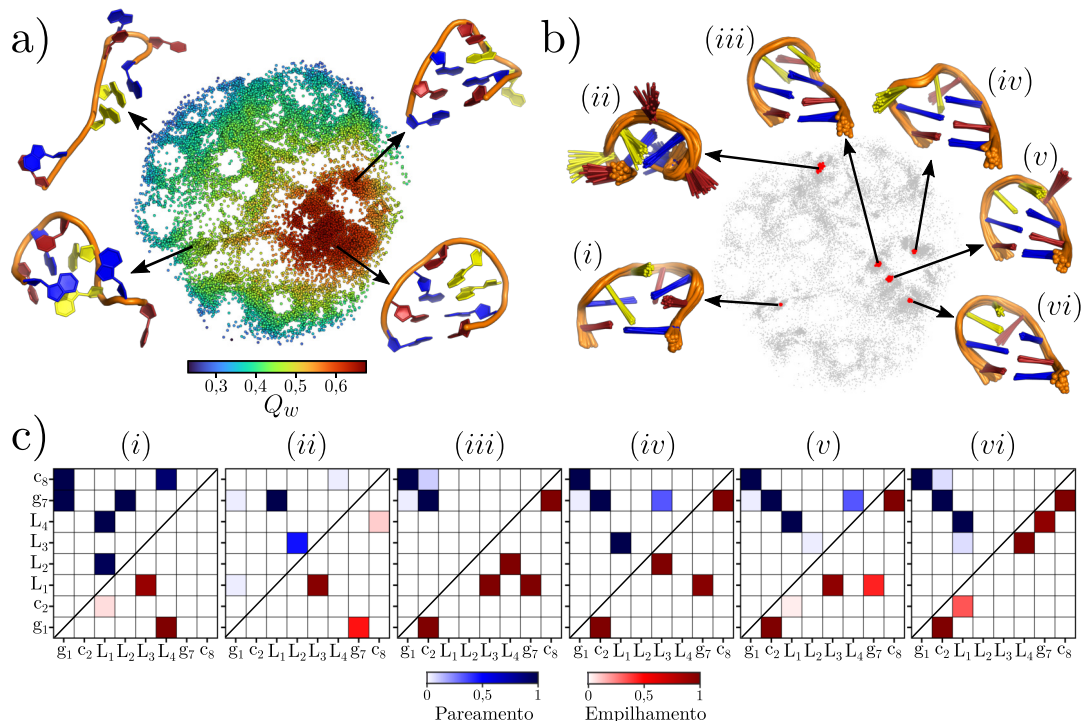
$$\sigma_{i,j} = \sigma_0 \max(|n_i - n_j|^\epsilon, 1), \quad (4.1)$$

sendo que  $n_i$  e  $n_j$  representam os índices do nucleotídeo  $i$  e  $j$ , respectivamente,  $\sigma_0 = 1 \text{ \AA}$  e  $\epsilon = 0,15$ . Nesse modelo,  $\sigma_{i,j}$  assume o valor de  $1 \text{ \AA}$  para átomos pertencentes ao mesmo nucleotídeo e aumenta gradativamente com a distância ao longo da cadeia.

#### 4.2.2 Resultados

Aplicando o ELViM às conformações selecionadas, obtém-se o espaço de fase efetivo representado na Figura 12.a, na qual os pontos estão coloridos de acordo com os respectivos valores de  $Q_w$ . A Figura C.1 (Apêndice C) reproduz esse espaço de fase efetivo com os pontos coloridos de acordo com os valores de RMSD e  $R_g$ . Para ilustrar a organização desse espaço, quatro conformações, arbitrariamente selecionadas, são indicadas por setas.

Figura 12 – Espaço de fase efetivo projetado pelo ELViM. (a) Os pontos foram coloridos de acordo valores de  $Q_w$ , calculada em relação à estrutura cristalográfica (PDB: 1ZIH). Quatro conformações são mostrada para ilustrar diferentes níveis de enovelamento ao longo do espaço de fase efetivo. (b) Análise LCS para as regiões de alta densidade de pontos. Cada LCS indicado é formado por 30 conformações, mostradas como pontos vermelhos na projeção. As bases são mostradas em *cartoon*, para facilitar a visualização das orientações relativas, seguindo o mesmo esquema de cor da Figura 11. (c) Mapa de interação média para cada LCS. Os mapas indicam a frequência das interações de pareamento e empilhamento calculadas para cada LCS.



Fonte: Adaptada de (62).

Seis regiões de alta densidade de pontos foram escolhidas para análise das LCSs (Seção 3.2.2). Na Figura 12.b, cada LCS é composta por trinta conformações, que estão

destacadas em vermelho sobre a projeção em cinza. Nessa figura, as bases estão representadas em *cartoon* para facilitar a visualização das orientações relativas. As LCSs (i) e (ii) pertencem à bacia de estados desnaturados, enquanto (iii) a (vi) estão associadas à bacia de estados enovelados. Esse resultado evidencia a eficiência do ELViM na identificação e agrupamento de conformações similares de RNA. Neste caso, uma maior variabilidade estrutural é observada apenas para a LCS (ii).

Para o RNA, as LCSs serão caracterizadas com base no pareamento e empilhamento das bases. Para isso, o pacote computacional Barnaba (108) foi utilizado para classificar as interações de pareamento e empilhamento de cada conformação. Mapas de interação média foram obtidos realizando-se a média das interações presentes em todas as conformações de uma LCS. A Figura 12.c mostra o resultado para as seis LCSs selecionadas (Figura 12.b).

Nestes mapas, o valor 1 indica que a interação ocorre em todas as conformações do grupo, enquanto valores próximos a zero indicam que apenas algumas conformações no grupo realizam a interação. Esse caso pode ocorrer devido a erros e distorções intrínsecos aos métodos de projeção multidimensional (pontos mal projetados) (26), à alta heterogeneidade conformacional ou, ainda, como consequência da adoção de valores de corte (*cut-off*) nos critérios de classificação das interações.

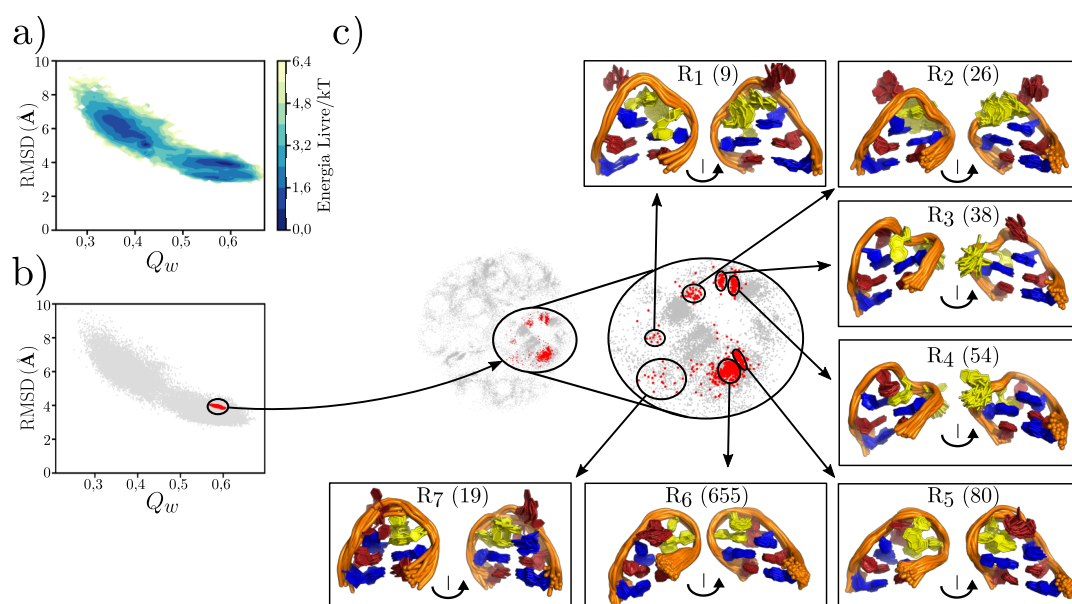
Os mapas de interação média (Figura 12.c) indicam que as conformações na LCS (i) apresentam a haste parcialmente formada e alça colapsada, com  $L_4$  empilhado sobre  $g_1$ . A LCS (ii) também é composta por conformações colapsadas, com pareamento entre  $g_7$  e  $L_1$ , e  $L_3$  empilhado sobre  $L_1$ . As LCSs (iii) a (vi) abrangem a bacia nativa e consistem em diferentes conformações parcialmente enoveladas, com haste bem definida e variações nas interações da alça, incluindo empilhamentos nativos e não-nativos.

Para comparar a superfície de energia efetiva gerada pelo ELViM com métodos usuais, a Figura 13 apresenta a projeção da energia livre nas coordenadas de reação  $Q_w$  e RMSD (109). A superfície de energia livre obtida (Figura 13.a) apresenta dois mínimos principais, sendo um para o estado desnaturado e outro para o estado nativo. Os pontos pertencentes ao mínimo do estado enovelado foram selecionados e destacados em vermelho, sobre os outros pontos em cinza, na Figura 13.b.

Esses pontos também são indicados na projeção ELViM (Figura 13.c), onde se dispersam em diferentes grupos dentro da bacia do estado enovelado. Cada um desses grupos foi selecionado e as conformações pertencentes a cada um foram sobrepostas e indicadas no grupos  $R_1$  a  $R_7$ . Nesses grupos, a figura à esquerda mostra o sulco menor da haste, enquanto a figura à direita mostra o sulco maior. Os números entre parenteses indicam o número de conformações sobrepostas e circuladas na projeção.

A análise desses grupos, mostra que o ELViM possui uma resolução maior ao distinguir conformações tanto pelo alinhamento do esqueleto de açúcar-fosfato, quanto

Figura 13 – Comparação do ELViM com superfície de energia livre efetiva projetada em coordenadas de reação usuais. (a) Superfície de energia baseada nas coordenadas de reação  $Q_w \times RMSD$ . (b) Os pontos do mínimo contendo o estado nativo foram selecionados arbitrariamente e destacados na projeção ELViM. (c) As conformações pertencentes a diferentes grupos na projeção são indicadas superpostas com o sulco menor mostrado à esquerda e o maior à direita. O número de conformações em cada grupo está indicado entre parênteses.



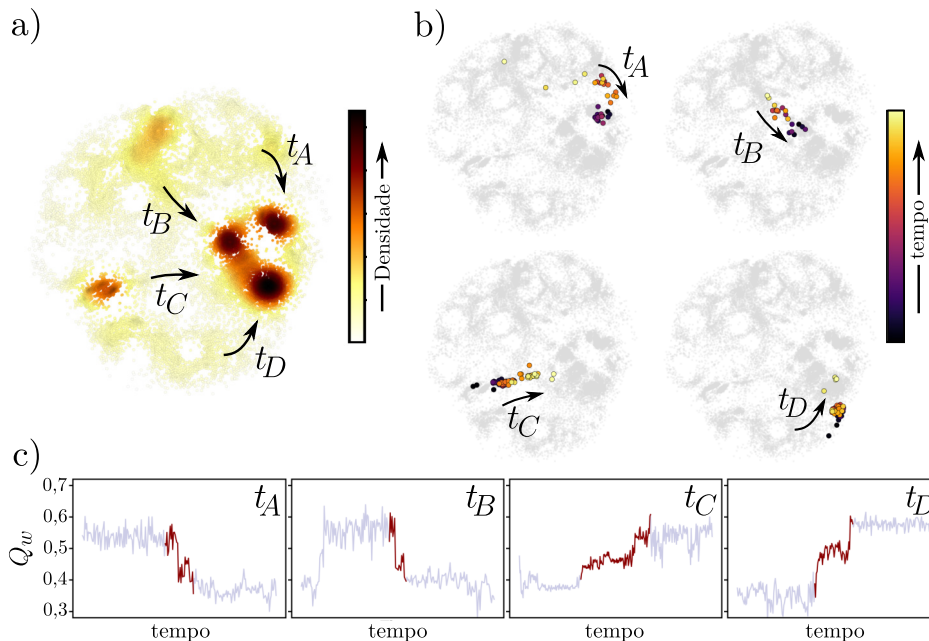
Fonte: Adaptada de (62).

pela orientação relativa das bases na alça. Por exemplo, a base  $L_2$  está orientada em direção ao sulco maior nos grupos  $R_1$ ,  $R_2$ ,  $R_3$ ,  $R_5$  e  $R_7$ , enquanto nos demais grupos está voltada para o sulco menor. O grupo  $R_6$  apresenta um alinhamento estrutural marcante, com 655 conformações sobrepostas, formando uma região coesa e densamente populada na projeção ELViM. Em contrapartida, alguns pontos se dispersam, não constituindo grupos bem definidos na projeção.

Para caracterizar as regiões de maior densidade e possíveis rotas conectando tais regiões, o KDE com núcleo gaussiano (Seção 2.3) foi novamente utilizado para estimar a densidade de estados efetiva da projeção. A Figura 14.a apresenta o resultado, com os pontos coloridos de acordo com a densidade local. Neste caso, quatro rotas distintas – indicadas pelas setas curvas e nomeadas  $t_A$ ,  $t_B$ ,  $t_C$  e  $t_D$  – conectam as bacias dos estados desnaturado e nativo.

Para verificar a consistência destas rotas, as 32 transições amostradas foram analisadas em detalhes. Para isso, as conformações pertencentes às transições foram coloridas de acordo com a evolução temporal sobre o espaço de fase efetivo em cinza, de modo a permitir uma visualização da rota seguida. Como exemplo, uma transição foi

Figura 14 – Densidade de estados e rotas de envelhecimento. (a) Densidade de estados calculada através de KDE gaussiano. Cada ponto foi colorido com o valor local da densidade estimada. As setas indicam quatro rotas que ligam as bacias dos estados desnovelado e envelado. (b) Uma transição seguindo cada rota é mostrada, com os pontos selecionados coloridos de acordo com a evolução temporal. (c) Detalhe da evolução temporal da coordenada  $Q_w$  durante as transições indicadas.



Fonte: Adaptada de (62).

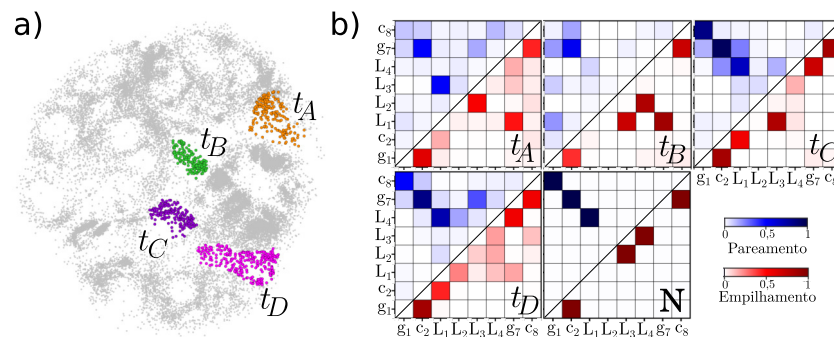
escolhida para cada rota e as conformações envolvidas foram destacadas sobre a projeção ELViM, sendo coloridas de acordo com a sequência temporal de amostragem (Figura 14b). Além disso, a Figura 14c, mostra em detalhe, a evolução de  $Q_w$  durante a transição indicada.

Desta análise, observa-se que transições partindo de qualquer parte da bacia de estados desnaturados podem seguir por uma das quatro rotas destacadas e atingir diferentes pontos da bacia do estado nativo. No entanto, não são observados saltos ou mistura de rotas na região de transição. Deste modo, todas as 32 transições podem ser classificadas como pertencentes a uma das quatro rotas indicadas. Especificamente, o número de transições que seguiram as rotas  $t_A$ ,  $t_B$ ,  $t_C$  e  $t_D$  foi igual a 10, 5, 8 e 9, respectivamente. A mesma análise para as demais 28 transições é apresentada nas Figuras C.2 e C.3 do Apêndice C.

Selecionando os pontos na região que conecta as bacias do estado nativo e desnaturado, obtém-se os *ensembles* de estados de transição (TSE, do inglês *Transition-State Ensemble*). Note que cada TSE é composto por conformações que foram amostradas em diversas transições e classificados em uma mesma rota de acordo com a métrica do ELViM. As conformações presentes em cada TSE foram manualmente selecionadas e caracterizadas

através das interações presentes em cada grupo, como apresentado na Figura 15. É importante observar que a definição de cada região é arbitrária, devido à ausência de limites na projeção que definam tais regiões. Dessa forma, ao se mudar os limites aqui estabelecidos, pequenas alterações poderiam ocorrer no mapa de frequência mostrado. No entanto, as assinaturas de cada região, definidas pelas interações de maior frequência, são robustas a pequenas variações nas fronteiras dos *ensembles*.

Figura 15 – Análise dos *ensembles* de estados de transição. (a) TSE para cada rota de enovelamento mostrada na figura 14 foi selecionado observando a densidade de estado e as transições amostradas. (b) Mapa de interação média para cada TSE. O mapa de interações do estado nativo (N) é reproduzido novamente para facilitar a comparação.



Fonte: Adaptada de (62).

A análise da Figura 15 mostra que o TSE  $t_A$  é o único que apresenta pareamento entre as bases  $L_1-L_3$ . Neste TSE, embora o par de bases  $c_2-g_7$  esteja bem definido, o primeiro par da haste ainda não está bem formado. As interações de empilhamento nativas, com exceção de  $L_4-L_3$ , possuem frequência significativa. Além disso, o empilhamento de bases não-nativo  $L_1-g_7$  está presente. Para o TSE  $t_B$ , apenas o pareamento  $c_2-g_7$  está formado na maioria das conformações, que apresentam três empilhamentos não-nativos ( $L_1-g_7, L_1-L_3$  e  $L_2-L_4$ ). Os TSEs  $t_C$  e  $t_D$  têm em comum o pareamento das bases  $L_1-L_4$  e o empilhamento não-nativo  $L_4-g_7$ , mas falta o empilhamento nativo  $L_3-L_2$ . A haste está melhor formada em  $t_C$  do que em  $t_D$ . O empilhamento não-nativo  $L_1-L_3$  está presente nos TSEs  $t_B$  e  $t_C$ , mas não em  $t_D$ .

Chen e Garcia (104), a partir de análise de trajetórias obtidas por REMD, identificaram duas rotas distintas de enovelamento para a tetra-alça GCAA. Na primeira rota, ocorre o pareamento  $L_1-L_4$  após o colapso da alça, e o enovelamento prossegue rapidamente. Na segunda rota, um pareamento não-nativo é formado primeiro entre  $L_1$  e  $L_3$ . Tal pareamento deve ser quebrado e  $L_3$  se afastar para que o pareamento nativo  $L_1-L_4$  possa se formar. A análise realizada revela que o TSE ( $t_A$ ) é caracterizado pelo pareamento  $L_1-L_3$ , enquanto que ( $t_C$ ) e ( $t_D$ ) são caracterizados pelo pareamento  $L_1-L_4$ . Em contraste, a maioria das conformações no TSE ( $t_B$ ) não possui nenhum desses pareamentos.

### 4.2.3 Discussão

Neste estudo, o ELViM foi utilizado, pela primeira vez, para investigar o relevo de energia de uma molécula de RNA, a tetra-alça GCAA. Os resultados obtidos demonstram que o método possui boa capacidade de agrupamento de estruturas semelhantes e manutenção de aspectos globais, como a variação suave de coordenadas de reação. Notavelmente, a comparação com a superfície de energia livre ( $Q_{WxRMSD}$ ) demonstra que o ELViM proporciona maior resolução estrutural, agrupando conformações com melhor alinhamento tanto do esqueleto de açúcar-fosfato quanto da orientação das bases nitrogenadas. Por fim, a análise das rotas de enovelamento na projeção permitiu a identificação de quatro *ensembles* de estados de transição que conectam as bacias desnaturada e nativa. Esses *ensembles* acrescentam detalhes estruturais às rotas de enovelamento previamente descritas por Chen e Garcia (104), reforçando a utilidade do ELViM na investigação da dinâmica conformacional e enovelamento de RNA.

### 4.3 ANÁLISE DO MECANISMO DE ENOVELAMENTO DO RIBOCOMUTADOR SAM-II

O ribocomutador SAM-II, localizado na região 5' não codificadora do mRNA de genes da biossíntese da metionina em proteobactérias alfa (110), regula a expressão gênica em resposta à ligação da S-adenosilmetionina (SAM) (110). Sua estrutura, um pseudonó do tipo H (Figura 10.b)(111, 112), compreende um aptâmero de alta especificidade para SAM e uma plataforma de expressão na extremidade 3' (A46-G51), contendo a sequência Shine-Dalgarno de início da tradução.

Ao se ligar à SAM, o ribocomutador SAM-II sofre uma mudança conformacional que compacta a estrutura de pseudonó, sequestrando a sequência Shine-Dalgarno. Consequentemente, a ligação do ribossomo a expressão dos genes a jusante é interrompida (111, 113). Na ausência da SAM, por outro lado, o ribocomutador assume uma conformação parcialmente enovelada (ou parcialmente aberta) (113–115), permitindo a expressão dos genes a jusante ao expor a sequência de início da tradução.

No SAM-II, o pseudonó é formado por duas hélices do tipo Watson-Crick, P1 e P2a, pela hélice P2b, e pelas alças L1 e L3 que interagem, respectivamente, com o sulco maior de P2b e com o sulco menor de P1 (Figura 16.a-b). Adicionalmente, a alça L3 possui quatro adeninas não pareadas, que se empilham entre si na estrutura terciária e formam múltiplas interações de hidrogênio com o sulco menor da hélice P1. No entanto, conforme descrito por Gilbert e colaboradores (111), embora a interação entre P1 e L3 se assemelhe a um motivo A-menor, as adeninas em L3 não formam interações triplas planares, mas apresentam uma torção de aproximadamente 70° em relação a P1, permitindo a interação com pares de bases sucessivos da hélice. Este motivo estrutural é denominado *A-minor twist*. Interações de hidrogênio não usuais resultantes deste motivo são indicadas por linhas

pontilhadas na Figura 16.a.

Evidências experimentais e estudos computacionais sugerem que a estrutura terciária do SAM-II depende da presença do ligante e da estabilização por íons  $\text{Mg}^{2+}$  (114–117). Além disso, esses íons desempenham um papel crucial na pré-organização do estado parcialmente aberto, favorecendo a ligação da SAM (114). No entanto, o mecanismo exato pelo qual esses íons influenciam o enovelamento e a pré-organização do estado parcialmente aberto não está totalmente esclarecido. Nesta seção, o ELViM será empregado na análise de trajetórias de enovelamento do SAM-II, com o intuito de investigar em detalhes atomísticos o enovelamento do ribocomutador e sua interação com íons  $\text{Mg}^{2+}$ .

#### 4.3.1 Sistema analisado

As trajetórias que serão analisadas foram obtidas pelo grupo da Dra. Susmita Roy (IISER Kolkata, Índia) utilizando-se um modelo baseado em estrutura (SBM, do inglês *Structure Based Model*) (118,119). Tal modelo, proposto por Hayes e colaboradores (120), considera todos os átomos pesados do RNA e inclui íons de  $\text{Mg}^{2+}$  de forma explícita, enquanto trata os íons monovalentes,  $\text{K}^+$  e  $\text{Cl}^-$ , implicitamente. O sistema foi simulado em uma caixa cúbica de comprimento 75 nm, com condições de contorno periódicas aplicadas. Cada simulação foi executada por 200 milhões de passos de integração, utilizando dinâmica de *Langevin* em regime subamortecido, com passo de integração de  $0,001 \tau_R$ . Cargas efetivas de  $-1$  são utilizadas para cada grupo fosfato e de  $+2$  para cada íon  $\text{Mg}^{2+}$ . Para partículas explícitas, foi utilizada uma massa reduzida de  $1 \mu_R$  e um coeficiente de arrasto de  $1 \tau_R^{-1}$ . Os parâmetros reduzidos são:  $\tau_R = 2$  ps,  $\mu_R = 15$  uma,  $T_R = 3,75$  K,  $\epsilon_R = 3,75$  kJ/mol. Mais detalhes sobre o modelo e conjunto de parâmetros utilizados são descritos nas referências. (121,122).

Para explorar completamente o espaço de fases do SAM-II, foram realizadas 28 trajetórias, cada uma contendo 200 milhões de passos de integração, com temperaturas efetivas entre  $96,5T_R$  e  $107,0T_R$  e concentração de  $\text{Mg}^{2+}$  igual a 2 mM. Todas as trajetórias partiram do estado enovelado na ausência da SAM, e cinco transições completas (enovelado até completamente estendida) foram amostradas com equilíbrio suficiente. Devido à ausência do ligante, a hélice P2a atinge apenas um estado parcialmente estruturado (123).

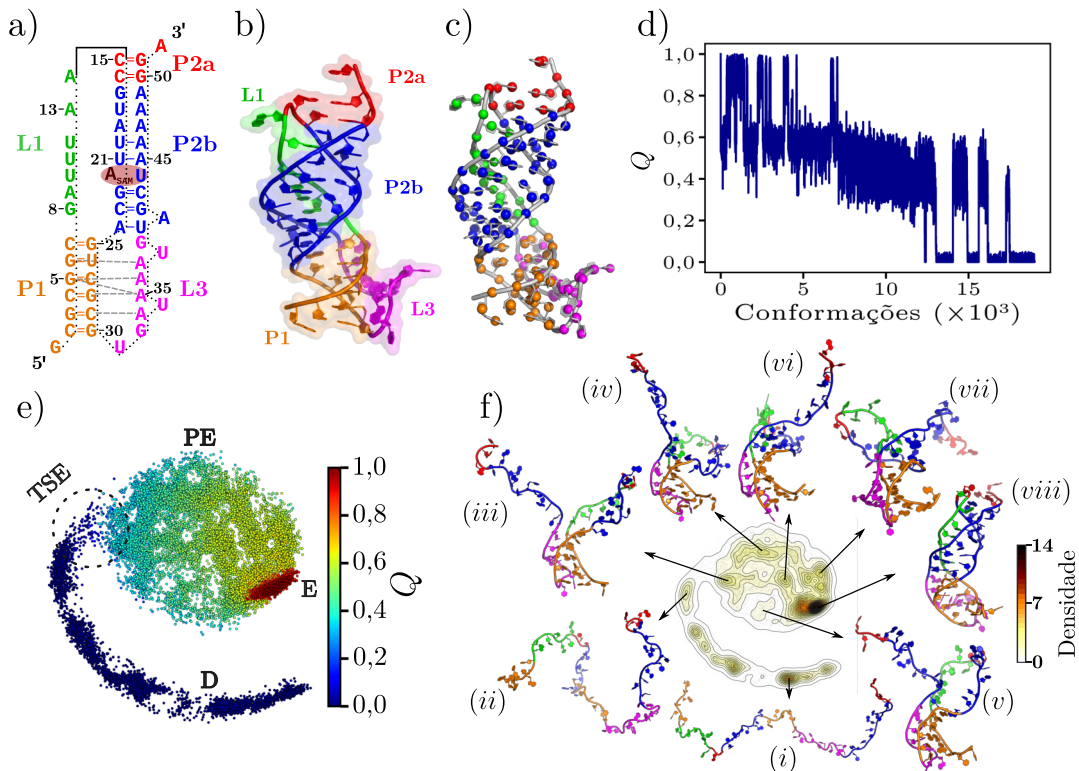
Dessas simulações, 19.005 conformações foram selecionadas de 16 trajetórias ( $96,5T_R \leq T \leq 104,6T_R$ ) para compor a projeção ELViM. Mais detalhes sobre a metodologia são dados no Apêndice C.

Para o cálculo da matriz de dissimilaridade, algumas adaptações também foram necessárias. Devido à importância da orientação das bases na estrutura dos RNAs, a análise da similaridade não pode se limitar ao esqueleto principal da molécula. Além disso, devido ao grande número de átomos presentes no ribocomutador, foi proposto calcular a dissimilaridade utilizando um modelo do tipo *coarse-grain* (CG). Para isso, foi utilizada

a ferramenta *Coarse Grain Builder* (124), implementada no VMD (82), para criar, para cada conformação amostrada, uma representação contendo três átomos virtuais localizados no centro de massa dos grupos fosfato, ribose e base nitrogenada. A Figura 16.c ilustra o modelo obtido sobreposto à estrutura tridimensional. Por fim, a dissimilaridade foi calculada utilizando o parâmetro  $\sigma_{i,j}$  conforme definido pela equação 4.1.

#### 4.3.2 Resultados

Figura 16 – (a) Estrutura secundária do SAM-II. O SAM-II possui estrutura de pseudonó do tipo H. Adeninas da alça L3 interagem com o sulco menor de P1, formando o motivo *A-minor twist*. Interações de hidrogênio não usuais nos pareamentos são indicadas por linhas pontilhadas (111). (b) Estrutura cristalográfica (PDB: 2QWY, cadeia A) (111). (c) Modelo “*coarse-grain*” utilizado para calcular a dissimilaridade: pseudoátomos correspondem aos centros de massa do fosfato, ribose e base nitrogenada. (d) Evolução temporal de  $Q$  para conformações selecionadas para a projeção ELViM. As trajetórias foram concatenadas em ordem crescente de temperatura efetiva. (e) Projeção ELViM para as trajetórias de enovelamento do SAM-II. São indicadas as bacias correspondentes aos estados desenovelado (D), parcialmente enovelado (PE) e enovelado (E). O círculo pontilhado destaca a região da transição D-PE, que contém o *ensemble* de estados de transição. (f) A densidade de estados, estimada por KDE com núcleo gaussiano, é representada por um mapa de curvas de nível. Regiões de alta densidade foram selecionadas para análise LCS. Para facilitar a visualização, apenas a estrutura centroide é indicada.

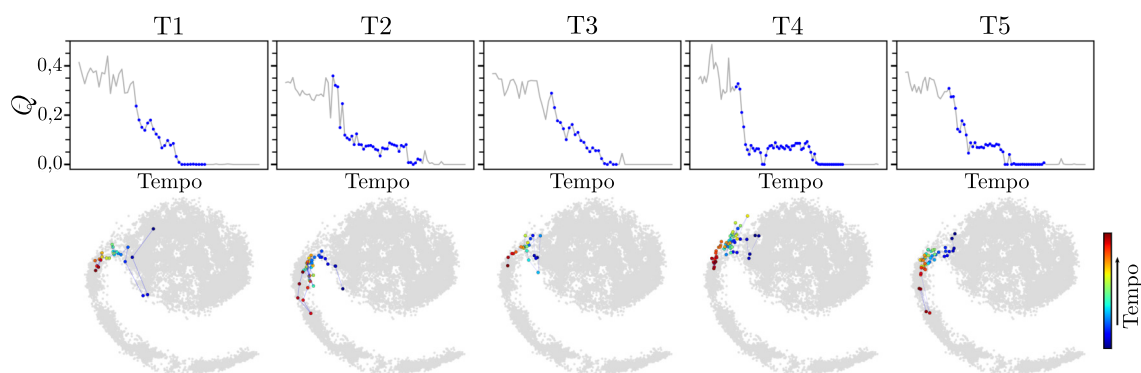


Fonte: Elaborada pelo autor (2025).

Nesta seção, será analisado o espaço conformacional efetivo gerado pelo ELViM a partir das trajetórias de desenovelamento da forma não ligada do SAM-II. Os valores de  $Q$ , calculados para cada conformação selecionada em relação à estrutura cristalográfica, são indicados na Figura 16.d, com as trajetórias concatenadas em ordem crescente de temperatura efetiva.

A Figura 16.e ilustra o espaço de fases efetivo gerado pelo ELViM, com os pontos coloridos de acordo com os respectivos valores de  $Q$ . Valores de RMSD para as conformações selecionadas, assim como a projeção ELViM colorida com base em tais valores, são apresentados na Figura C.5 (Apêndice C). Nesta projeção, são identificadas três regiões distintas: a bacia do estado enovelado (E:  $Q > 0,8$ , em vermelho), uma região central correspondente ao estado parcialmente enovelado (PE:  $0,2 < Q < 0,8$ , com coloração variando do azul claro ao laranja) e a bacia do estado desnaturado (D:  $Q < 0,2$ , em azul escuro).

Figura 17 – Transições entre D e PE para o SAM-II. Cinco transições foram amostradas sob regime de equilíbrio. Na linha superior, é indicada a evolução temporal de  $Q$  ao longo das transições. O Eixo horizontal corresponde à evolução temporal das simulações. Na linha inferior, conformações sequencialmente amostradas nas transições são coloridas na projeção de acordo com a evolução temporal.



Fonte: Elaborada pelo autor (2025).

Notavelmente, nesta projeção, a bacia D concentra-se em uma região que se curva em direção ao corpo principal da projeção e está conectada a ele por meio de uma região estreita (indicada por uma circunferência pontilhada). Esta região contém o TSE para a transição entre as bacias D e PE. Por outro lado, a bacia E está conectada à região PE através de uma área ampla, sugerindo a existência de grande heterogeneidade conformacional do estado parcialmente enovelado e de diversas rotas para a transição PE-E. A seguir, o TSE da transição D-PE, denominado simplesmente TSE, será analisado em detalhes.

A Figura 16.f apresenta a densidade de pontos na projeção, estimada por KDE de núcleo gaussiano, e a análise das assinaturas LCSs para algumas regiões de alta densidade.

Nessa análise, todos os pontos contidos na curva de nível com maior densidade local foram selecionados e apenas a estrutura centroeide foi mostrada, para facilitar a visualização da estrutura do RNA.

Partindo da bacia D, a LCS (*i*) representa uma região ocupada por conformações totalmente estendidas, enquanto a LCS (*ii*) indica estados mais colapsados, nos quais o segmento central de P1 se aproxima de L3. Considerando a bacia PE, as LCSs (*iii*) a (*vii*) ilustram a heterogeneidade estrutural do estado parcialmente enovelado, no qual grande parte das interações de pareamento/empilhamento presentes no motivo *A-minor twist* estão estabelecidas (hélice P1 e as interações terciárias entre P1-L3). Adicionalmente, nota-se que ao se aproximar da bacia E (LCSs (*iii*) a (*vii*)), aumenta-se o número de pareamentos na hélice P2b e de interações P2b-L1, com a extremidade 3' assumindo diferentes orientações em relação à hélice P1. Finalmente, a LCS (*viii*) representa o estado enovelado, no qual a hélice P2a está parcialmente formada. Esta região, que é a última a se formar, contém o sítio de ligação do ribossomo, sendo portanto a conformação responsável por interromper o início da tradução dos genes a jusante.

Para as trajetórias analisadas, cinco eventos de desenovelamento foram identificados em partes das trajetórias suficientemente equilibradas. A Figura 17 ilustra estas transições, indicando a ordem sequencial das conformações envolvidas. Para cada transição, o gráfico superior apresenta a evolução sequencial de  $Q$ , enquanto o painel inferior exibe essa evolução na projeção ELViM. Todas as transições passam pelo TSE sem saltos, confirmando a hipótese de um TSE bem definido. No entanto, um número maior de eventos seria necessário para excluir outros caminhos de menor probabilidade.

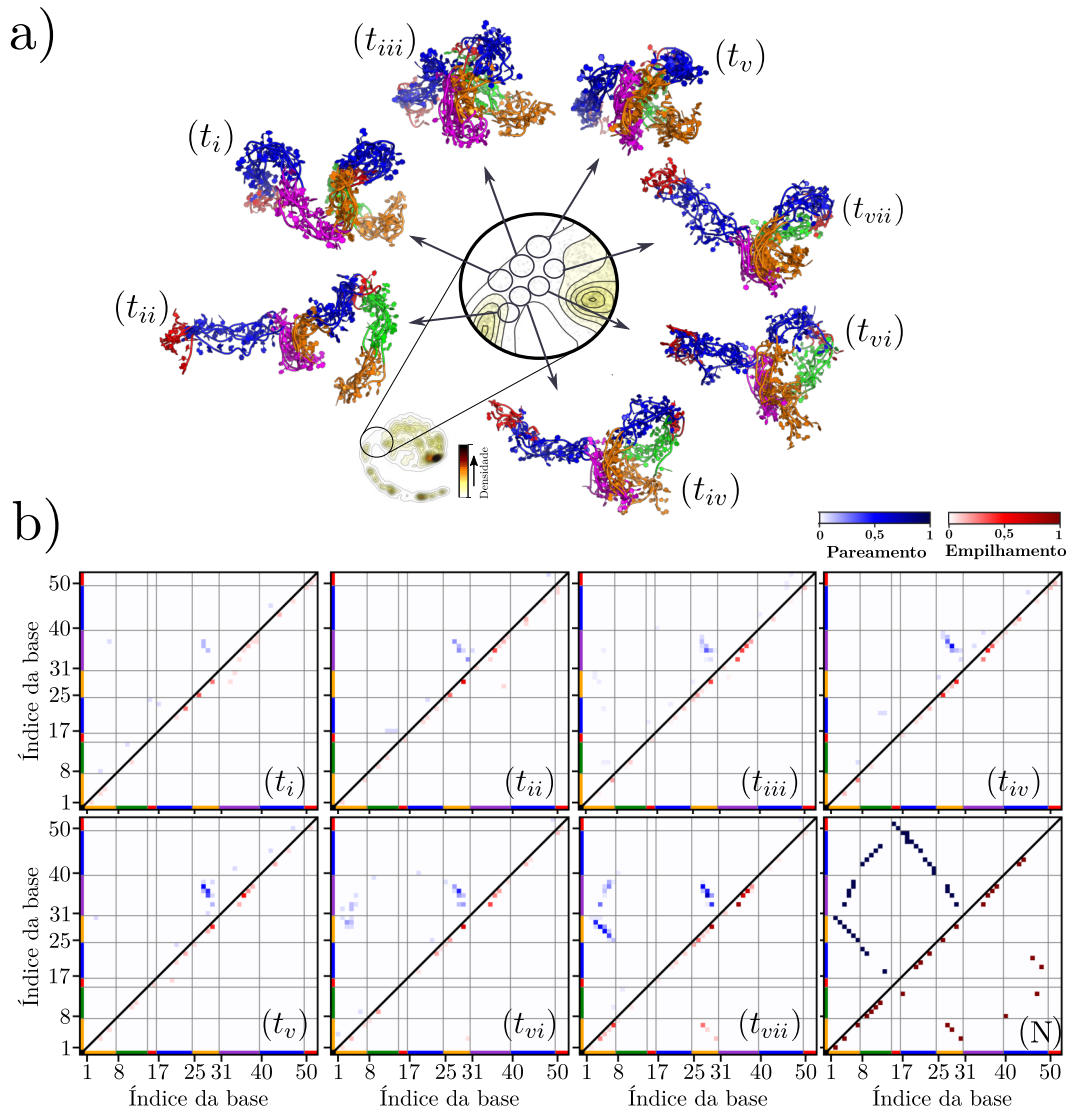
A Figura 18.a mostra uma aproximação da região TSE, na qual seis sub-regiões foram arbitrariamente selecionadas e LCSs foram geradas, cada uma com cinco conformações. De modo geral, as LCSs indicam que o enovelamento se inicia por interações entre o segmento interno de P1 e a alça L3, seguidas pela formação da hélice P1 e demais interações P1-L3.

A LCS ( $t_i$ ) (Figura 18.a) representa uma região ocupada por conformações que, apesar de sugerir um colapso inicial, possuem poucas interações significativas. Por sua vez, a LCS ( $t_{ii}$ ) indica uma aproximação entre a alça L3 e o segmento interno de P1, permitindo a formação residual de interações de pareamento de bases entre P1 e L3 (Figura 18.b). Essas interações tornam-se ligeiramente mais pronunciadas nas LCSs ( $t_{iii}$ ) e ( $t_v$ ), que apresentam as extremidades 3' e 5' assumindo diferentes orientações sem, no entanto, formar interações significativas. Nessas regiões, as adeninas desemparelhadas na alça L3 começam a apresentar algumas das interações de empilhamento nativas.

Nas regiões do TSE representadas pelas conformações ( $t_{iv}$ ), ( $t_{vi}$ ), os pareamentos mais frequentes incluem G28–A35/A36 e C27–A36/A37. Essas interações, juntamente com o empilhamento das adeninas em L3, contribuem para estabilizar a região central.

Em seguida, a extremidade 5' dobra-se sobre este núcleo, permitindo a formação dos pareamentos de bases da hélice P1 (G3–C29, C4–G28, G5–C27) e das interações terciárias do motivo *A-minor twist*. Por fim, a LCS ( $t_{vii}$ ), localizada no início da bacia PE, mostra esse motivo parcialmente formado, estando presentes os pareamentos nativos da hélice P1, parte das interações P1-L3, além dos empilhamentos nativos característicos do *A-minor twist*.

Figura 18 – Densidade de estados e LCSs para a transição D-PE. (a) A região do estado de transição (TSE) foi dividida em seis sub-regiões, representadas pelas LCSs. Cada LCS é composta por cinco conformações sobrepostas. (b) Mapa de interação média para cada sub-região. As interações de pareamento e empilhamento de cada conformação foram identificadas pelo programa Barnaba (108) e a média foi calculada considerando todas as conformações de cada sub-região. O mapa de interações da estrutura nativa (N) é apresentado para comparação.



Fonte: Elaborada pelo autor (2025).

De forma complementar, foram analisados mapas de interação média para intervalos definidos da coordenada  $Q$ , considerando todas as 28 simulações. As conformações foram agrupadas em intervalos iguais de  $Q = 0,1$ , e a interação média foi calculada para cada intervalo. O resultado dessa análise está apresentado na Figura C.6 do Apêndice C. Observa-se que as interações relativas ao motivo *A-minor twist* se completam até  $Q = 0,4$ . Para  $Q > 0,4$  iniciam-se os pareamentos na hélice P2b, seguidos pelas interações do pseudonó P2b-L1. Quando  $Q > 0,9$ , o pareamento C16-G50 na hélice P2a torna-se significativo, levando à formação parcial dessa hélice.

Para realizar uma comparação qualitativa, também foi projetado o espaço conformacional efetivo utilizando o método PCA. O PCA foi aplicado às coordenadas cartesianas ponderadas pela massa de todos os átomos pesados. Efeitos de translação e rotação foram removidos por meio da centralização das coordenadas na estrutura média, refinando esse processo até atingir convergência. A superfície PCA resultante é apresentada na Figura C.7 do Apêndice C. As posições das conformações representativas indicadas na Figura 16.f são também mostradas nas duas projeções (Figura C.7.b). Nota-se que, em ambas as projeções, as conformações representativas exibem uma distribuição relativa semelhante. No entanto, as bacias PE e E são mais compactas na projeção PCA, enquanto a bacia D e o TSE apresentam uma distribuição mais difusa.

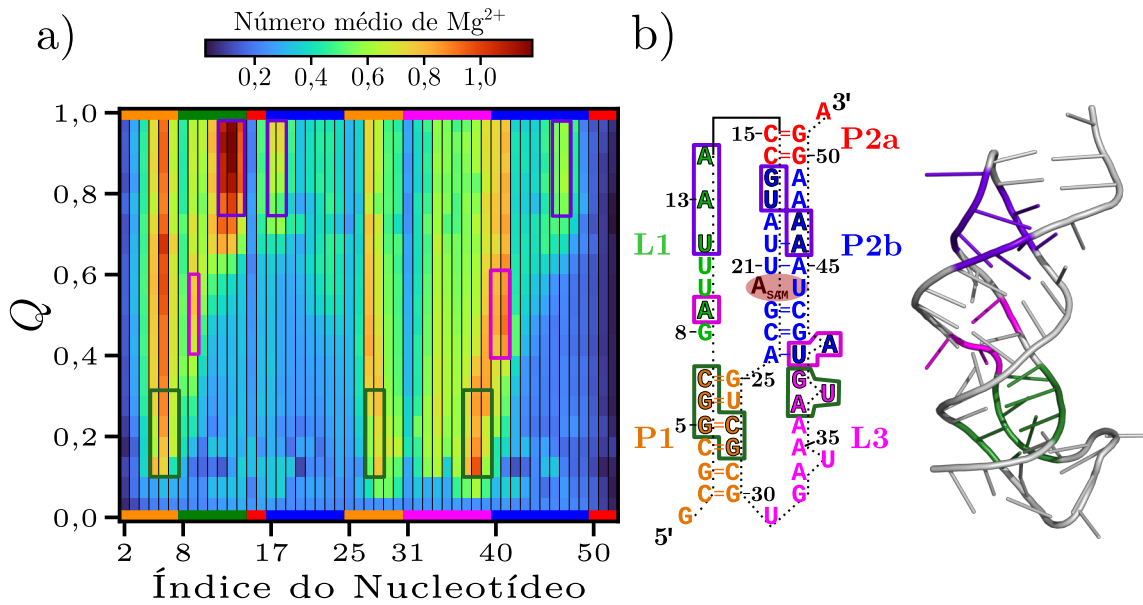
A mesma análise foi realizada para as conformações representativas do TSE (Figura 18.a), conforme apresentado na Figura C.7.c. Nesse caso, a projeção PCA mantém parcialmente as orientações relativas das LCSs, como visto na projeção ELViM, ocorrendo sobreposição entre alguns grupos. O caso mais notável refere-se a LCS ( $t_{viii}$ ), cuja estrutura, embora próxima ao estado parcialmente enovelado, aparece no PCA mais distante da bacia enovelada e parcialmente misturada com outras LCSs mais desestruturadas. Esse contraste destaca as vantagens do ELViM, que, em suma, facilita a identificação da região do estado de transição e fornece uma caracterização mais detalhada da bacia PE, incluindo os múltiplos mínimos locais ocupados por estruturas parcialmente enoveladas.

#### 4.3.2.1 Interação com íons $Mg^{2+}$

Para investigar o papel dos íons  $Mg^{2+}$  na estabilização da estrutura do SAM-II, a população média de  $Mg^{2+}$  associada a cada grupo fosfato será analisada. Para isso, foi quantificado, para cada quadro da simulação, o número de íons  $Mg^{2+}$  localizados a uma distância inferior a 8 Å do centro de cada grupo fosfato. A população de  $Mg^{2+}$  média foi obtida utilizando-se todas 28 simulações, conforme detalhado no Apêndice C.

Para analisar como a população desses íons varia ao longo da coordenada de reação, os quadros simulados foram agrupados em um histograma contendo 20 intervalos idênticos em  $Q$ . Para cada intervalo, o número médio de íons  $Mg^{2+}$  dentro da distância limite de cada grupo fosfato foi calculado. O resultado é mostrado na Figura 19.a, sendo que o eixo

Figura 19 – Evolução do número médio de íons  $Mg^{2+}$  por grupo fosfato ao longo de Q. (a) O número médio de íons  $Mg^{2+}$  em torno de cada grupo fosfato foi determinado em 20 intervalos da coordenada Q. O eixo x corresponde ao índice de cada grupo fosfato, enquanto o eixo y representa os intervalos na coordenada Q, e o número médio de íons é indicado pelo código de cores. (b) Estruturas secundária e terciária, destacando os resíduos com maior visitação de íons  $Mg^{2+}$ .



Fonte: Elaborada pelo autor (2025).

horizontal representa o índice do grupo fosfato, o eixo vertical mostra os intervalos em  $Q$  e o mapa de cores indica a população média de íons  $Mg^{2+}$  em cada intervalo. Dessa forma, é possível analisar como a população média em torno de cada resíduo evolui ao longo da coordenada de reação.

A primeira linha, na parte inferior do histograma (Figura 19.a), corresponde ao intervalo  $0 < Q < 0,05$  (cadeia totalmente estendida), no qual o número médio de íons  $Mg^{2+}$  é insignificante em torno de todos os grupos fosfato. Conforme o enovelamento progride, por volta de  $Q \approx 0,2$ , o número médio de íons  $Mg^{2+}$  aumenta significativamente em três grupos formados pelos resíduos G5-C7 e C27-G28 da hélice P1, e A37-G39 da alça L3. Para facilitar a visualização, esses mesmos grupos foram destacados, em verde, no histograma e nas estruturas secundária e terciária (Figura 19.b). Note que esses grupos correspondem aos primeiros contatos formados na região de nucleação. Para  $Q \approx 0,5$ , o número médio de íons  $Mg^{2+}$  aumenta para os resíduos G8 da alça L1 e U40-A41 da hélice P2b (destacados em magenta), próximo ao sítio de ligação da SAM. Finalmente, durante a transição PE-E,  $Q \approx 0,8$ , o número médio de íons  $Mg^{2+}$  aumenta em três regiões adicionais: resíduos U12-A14 da alça L1, G17-U18 e A46-A48 da hélice P2b.

É importante notar que apenas na região  $Q > 0,9$  as conformações são representa-

tivas do estado enovelado, sendo a estrutura terciária estabilizada pelos íons de  $Mg^{2+}$ . A análise global da distribuição de íons revela um efeito cumulativo ao longo do processo de enovelamento, sendo os principais sítios de ligação correspondentes às três regiões destacadas. A população média máxima é observada em torno do resíduo A13, sugerindo um papel crucial dos íons  $Mg^{2+}$  na estabilização da conformação parcialmente enovelada, pelo mesmo na ausência do ligante SAM.

A Figura C.8 do Apêndice C, também apresenta esses valores para cada resíduo, com a média calculada sobre todos os valores de  $Q$ . Além disso, essa Figura apresenta uma análise de contatos entre grupos fosfatos que são mediados por íons  $Mg^{2+}$ , conforme definido no Apêndice C. Essa análise, que foi realizada por Avijit Mainan, mostra que, no estado enovelado, o contato G28-U38 é mediado por um íon  $Mg^{2+}$ , com maior frequência relativa, em todas as simulações que amostram a transição D-PE. Esse contato parece sustentar a aproximação da alça L3 à fita interna de P1, favorecendo sua formação e, posteriormente, contribuindo para a estabilidade do motivo *A-minor twist*.

### 4.3.3 Discussão

Neste estudo, o ELViM foi utilizado para investigar o mecanismo de enovelamento do ribocomutador SAM-II. Os resultados demonstram que a projeção ELViM distingue as bacias D, PE e E, identificando regiões de transição e evidenciando a heterogeneidade conformacional do estado PE. Essa descrição está em concordância com resultados prévios obtidos por técnicas experimentais (113, 117, 125) e computacionais (114). Ao examinar as conformações representativas das bacias mais densas, foram descritas as interações de pareamento e empilhamento presentes em diferentes etapas do enovelamento. Em especial, foram fornecidos detalhes para a transição D-PE, para a qual as LCSs demonstraram um aumento progressivo dessas interações, que culmina na formação do motivo *A-minor twist* e na estabilização do estado PE.

Além disso, foram analisadas cinco transições entre os estados D-PE, que sugerem uma rota de enovelamento específica durante o colapso inicial do ribocutador. Embora eventos adicionais de enovelamento sejam necessários para descartar outras rotas menos prováveis, o caminho proposto é compatível com evidências descritas na literatura. Por exemplo, Haller e colaboradores (113) sugeriram, com base em dados de RMN e smFRET, que os íons  $Mg^{2+}$  estabilizam o segmento P1-L3, favorecendo a formação da hélice P1, bem como a estabilização da alça L1 e do sítio de ligação. A estabilização completa da estrutura ocorre, após a ligação da SAM, com a formação da hélice P2a. Experimentos de sondagem química (*chemical probing*) (111) demonstraram que a ligação da SAM contribui para a estabilização de L1 e P2a/b, enquanto a interação P1-L3 permanece praticamente inalterada. Esse resultado sugere que essa região já se encontra pré-organizada na ausência do ligante. Em contrapartida, Xue e colaboradores (126), através de simulações

de REMD sem íons de  $Mg^{2+}$ , relataram uma rota de enovelamento alternativa, na qual a formação de P2a/b precede a da hélice P1; contudo, essa rota ainda não foi confirmada experimentalmente.

Por fim, foi apresentada uma análise detalhada das interações RNA- $Mg^{2+}$ . Diversos estudos têm demonstrado que a presença de  $Mg^{2+}$  é crucial para que o SAM-II atinja sua conformação enovelada (115, 116, 125, 127, 128). Nesse contexto, os sítios preferenciais de ligação desses íons foram analisados, avaliando a média do número de interações entre  $Mg^{2+}$  e cada grupo fosfato durante o enovelamento. Os resultados evidenciaram um efeito cumulativo dos íons, que acompanha as mudanças conformacionais, destacando os sítios preferenciais em cada etapa. Notavelmente, a maior média de  $Mg^{2+}$ , para valores de  $Q$  superiores a 0,8, foi observada nas proximidades dos nucleotídeos U12 e A13 (Figura 19). Chen e colaboradores (115) demonstraram, por meio de titulação por RMN, que a ligação de  $Mg^{2+}$  induz perturbações significativas em U12 na forma ligada do SAM-II, facilitando a formação de novos pareamentos de bases e promovendo interações terciárias entre alças e hastes, o que impulsiona a transição do estado PE para uma conformação mais compacta. Por fim, foi sugerido que um contato mediado por  $Mg^{2+}$ , entre os resíduos G28 (P1) e U38 (L3), pode contribuir para a estabilização adicional do motivo *A-minor twist*.

#### 4.4 CONCLUSÃO

Neste estudo, o ELViM foi aplicado para mapear a superfície de energia e investigar as rotas de enovelamento de duas moléculas de RNA com tamanhos e complexidades distintas: a tetra-alça GCAA e o ribocomutador SAM-II. Nesta análise da tetra-alça, a avaliação das LCSs evidenciou a alta capacidade do método de agrupar conformações com base tanto na estrutura da cadeia principal de açúcar-fosfato quanto na orientação das bases, permitindo a identificação de múltiplos *ensembles* de estados de transição que conectam as bacias desnaturada e nativa.

No caso do SAM-II, a projeção ELViM distinguiu as bacias conformacionais (D, PE e E), capturando detalhes do estado de transição D-PE, incluindo a formação progressiva do motivo *A-minor twist*. Adicionalmente, foi realizada uma análise detalhada dos sítios específicos de interação RNA- $Mg^{2+}$ , indicando o efeito cumulativo dessas interações ao longo do processo de enovelamento e ressaltando seu papel na estabilização da estrutura terciária, essencial para a função do ribocomutador.

Em síntese, os resultados apresentados reforçam o potencial do ELViM como uma ferramenta promissora para a análise de simulações de RNA, capaz de fornecer detalhes sobre os mecanismos de enovelamento e dinâmica conformacional de tais moléculas. Embora a métrica empregada possa precisar de ajustes para aplicações em sistemas mais complexos, nossa investigação contribui significativamente para a compreensão dos processos de enovelamento e abre caminho para novos estudos nessa área.

## 5 PROTEÍNAS INTRINSICAMENTE DESORDENADAS

### 5.1 INTRODUÇÃO

Nas últimas décadas, a hipótese de que uma proteína funcional requer uma estrutura tridimensional bem definida foi desafiada pela descoberta de partes de proteínas, ou mesmo sua totalidade, que apresentam alta flexibilidade, permanecendo desordenadas em condições fisiológicas, pelo menos *in vitro* (129–131). Essas proteínas são denominadas Proteínas Intrinsecamente Desordenadas (IDPs, do inglês *Intrinsically Disordered Proteins*) e as regiões não estruturadas presentes em proteínas com domínios enovelados são denominadas Regiões Intrinsecamente Desordenadas (IDRs, do inglês *Intrinsically Disordered Regions*). Apesar de sua plasticidade estrutural, essas proteínas desempenham papéis fundamentais em processos como a transcrição, as vias de sinalização celular, a regulação do ciclo celular e diversas outras funções (132–135). Por outro lado, a sobreexpressão e agregação de tais proteínas estão ligados ao surgimento de condições patológicas, tais como câncer e doenças neurodegenerativas (132, 136–139).

Diferente das proteínas globulares, as IDPs não possuem um estado nativo bem definido. Em vez disso, são geralmente descritas por *ensembles* de estruturas interconvertíveis (132), ou seja, essas IDPs existem como uma coleção dinâmica de conformações distintas que podem transitar continuamente entre si devido à sua alta flexibilidade. Embora algumas IDPs se comportem como polímeros aleatórios, outras podem exibir algum tipo de ordem estrutural, representada por *ensembles* preferenciais, caracterizados por regiões do espaço de fase separadas por pequenas barreiras energéticas, o que resulta em um contínuo heterogêneo de estruturas (19, 140). A capacidade de conversão entre esses *ensembles* está relacionada à multifuncionalidade das IDPs, já que diferentes *ensembles* preferenciais podem estar associados a funções biológicas distintas. Essas funções podem ser ativadas ou inibidas por modificações pós-traducionais (PTMs), como a acetilação, que induzem alterações em propriedades físico-químicas, como carga e flexibilidade, remodelando o relevo energético das IDPs e regulando, assim, sua atividade (19). Além disso, diferentes *ensembles* podem ser favorecidos em condições fisiológicas, dependendo de interações com o meio celular, como solvente, íons ou outras moléculas.

Do ponto de vista da ELT, espera-se que o relevo da superfície de energia das IDPs seja complexo e altamente rugoso (17), situando-se mais próximo do perfil típico de um polímero aleatório do que de uma proteína globular com enovelamento “tudo-ou-nada”. Assim, as IDPs tendem a ser caracterizadas por relevos energéticos com múltiplos mínimos locais separados por barreiras energéticas possivelmente baixas (17, 19, 141). Essa característica torna a exploração de seu espaço de fase particularmente desafiadora,

---

exigindo o desenvolvimento de novos campos de força, técnicas de amostragem aprimorada (*enhanced sampling*) e abordagens integrativas que incorporem dados experimentais para orientar as simulações computacionais (71, 142, 143).

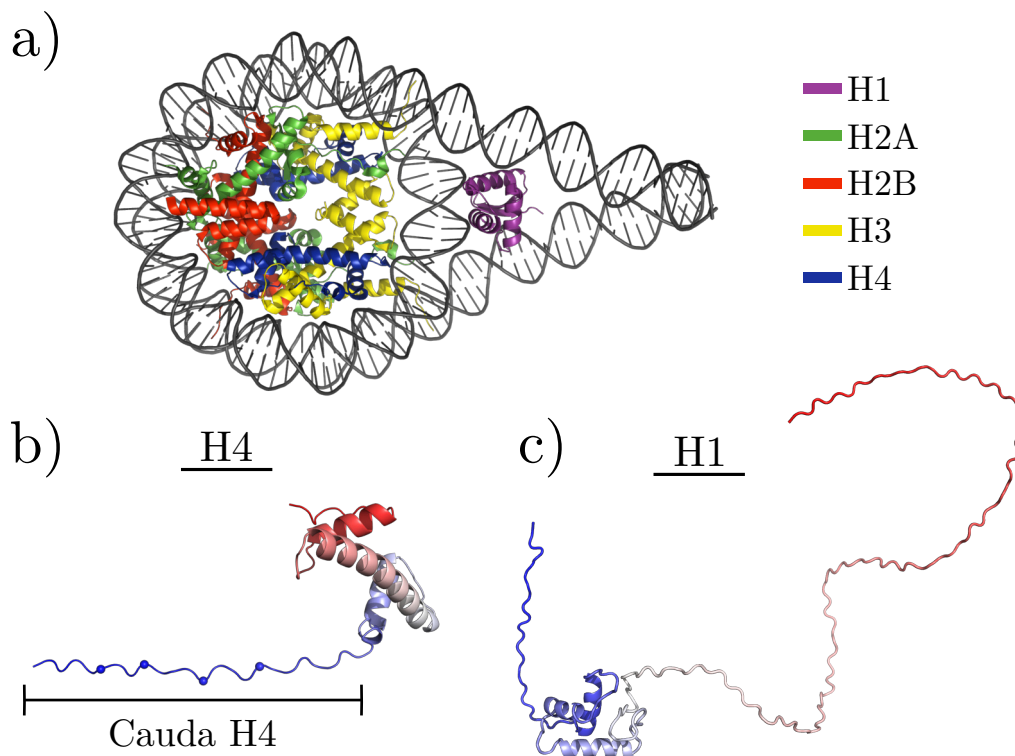
Nesse capítulo, o ELViM será utilizado para analisar o relevo de energia de IDPs, explorando dados de simulações com todos átomos em solvente explícito, modelos *coarse-grain* e *ensembles* obtidos por métodos integrativos. Como a métrica do ELViM pode ser aplicada considerando apenas os carbonos alfa, é possível construir um espaço efetivo único, abrangendo diversos *ensembles* de IDPs. Nesse espaço, os *ensembles* são organizados de acordo com a dissimilaridade estrutural estimada pelo ELViM, permitindo uma análise diferencial detalhada.

Primeiramente, essa abordagem será empregada para investigar o impacto de diferentes níveis de acetilação na cauda desordenada da histona H4. Para isso, serão introduzidas novas técnicas de análise da projeção ELViM, como o cálculo de entropia da projeção, para se estimar a heterogeneidade estrutural de cada variante da cauda H4. Em seguida, o espaço de fase efetivo da histona de ligação H1 será projetado e utilizado para investigar como o modo de ligação ao nucleossomo influencia as regiões acessíveis do espaço efetivo. Por fim, o ELViM será utilizado para analisar *ensembles* estruturais obtidos por métodos integrativos e depositados no banco de dados denominado *Protein Ensemble Database (PED, [proteinensemble.org](http://proteinensemble.org))* (144).

### 5.1.1 Nucleossomo e Histonas

No núcleo das células eucarióticas, o DNA associa-se a proteínas chamadas histonas, formando um complexo conhecido como cromatina. Além de sua função estrutural, as histonas desempenham papéis essenciais na compactação da cromatina, na replicação do DNA e na regulação gênica (145–148). A cromatina organiza-se em unidades fundamentais chamadas nucleossomos, conectadas por segmentos de DNA denominados DNA de ligação. Cada nucleossomo é composto por um núcleo proteico constituído por um octâmero de histonas, formado por duas cópias de cada uma das histonas H2A, H2B, H3 e H4. Ao redor desse octâmero, cerca de 146 pares de bases de DNA se enrolam (148), formando em torno de 1,65 voltas, como ilustrado na Figura 20.a. As histonas do núcleo possuem uma estrutura composta por um domínio globular altamente conservado e caudas intrinsecamente desordenadas localizadas na extremidade amino-terminal, que se estendem para fora do nucleossomo. Essas caudas permitem interações com regiões ácidas do DNA e com outros nucleossomos, desempenhando um papel crucial na dinâmica da cromatina.

Figura 20 – Estrutura do nucleossomo e das histonas H1 e H4. (a) Estrutura do nucleossomo em complexo com a histona de ligação H1 (PDB: 5NL0). O nucleossomo é composto por um octânomo central contendo duas cópias de cada uma das histonas H2A, H2B, H3 e H4. A fita de DNA se enrola sobre o octânomo central completando aproximadamente 1,65 voltas. A histona de ligação H1 se liga à parte externa da partícula nucleossomal, contribuindo com a estabilização do nucleossomo e influenciando na compactação final da cromatina. Apenas os domínios globares são mostrados para as proteínas. (b) Estrutura prevista pelo *alphafold* para histona H4 da espécie homo sapiens (AF-P62805-F1). A figura evidencia a cauda desordenada presente no domínio amino-terminal (azul) e o domínio globular presente no carboxi-terminal (vermelho). As principais lisinas sujeitas a acetilação (K5, K8, K12 e K16) são mostradas como esferas. (c) Histona H1 (PDB:5NL0) com caudas desordenadas modeladas através do MODELLER (149).



Fonte: Elaborada pelo autor (2025).

## 5.2 ANÁLISE DO RELEVO DE ENERGIA DA CAUDA H4 E SUAS FORMAS ACETILADAS

Apesar da alta flexibilidade das caudas H4, estudos teóricos sugerem que seu relevo energético não seja completamente desordenado, mas revele certa organização estrutural. Potoyan e colaboradores (150), utilizando simulações REMD com todos os átomos explícitos, demonstraram que a cauda H4 se desvia do comportamento de cadeia aleatória, adotando um relevo energético multi-bacias, enriquecido por conformações em grampos  $\beta$ . Röder e colaboradores (151), utilizando a técnica conhecida como *Path*

*Sampling* (152), também caracterizaram o relevo de energia da cauda H4 como um sistema multi-funil, composto por cinco mínimos distintos, separados por altas barreiras de energia.

As modificações pós-traducionais (PTMs) nas caudas das histonas do núcleo desempenham papéis fundamentais na regulação da estrutura da cromatina e da expressão gênica, impactando processos essenciais como replicação, transcrição e reparação do DNA (153–155). Essas caudas sofrem diversas PTMs, como acetilação, metilação e fosforilação, que modulam a dinâmica da cromatina ao controlar o acesso ao DNA. Em particular, a acetilação na cauda H4 tem sido amplamente investigada devido ao seu impacto significativo na função e na estrutura geral da cromatina (18, 73, 151, 156–158). Os efeitos de acetilações específicas também são bem descritos na literatura. Por exemplo, a acetilação da lisina K16 da cauda H4 impacta fortemente sua heterogeneidade conformacional, reduzindo sua flexibilidade e promovendo conformações exclusivas, que não são observadas na forma nativa (156–158).

Nesta seção, o ELViM será aplicado para realizar uma análise diferencial dos *ensembles* conformacionais da cauda H4 nativa e suas formas acetiladas. Para isso, um espaço de fase conformacional efetivo será projetado, permitindo analisar como os *ensembles* individuais de cada variante se distribuem por esse espaço efetivo global. Através desta análise, estados estruturais serão caracterizados e a heterogeneidade estrutural de cada variante será estimada.

### 5.2.1 Sistema analisado

Para avaliar o impacto de diferentes níveis de acetilação no relevo conformacional da cauda H4, serão analisadas simulações atomísticas de dinâmica molecular da cauda H4 em sua forma nativa (N) e de cinco variantes que apresentam diferentes níveis e locais de acetilação em lisinas. As variantes consideradas são H4-K16<sub>ac</sub>, H4-K5<sub>ac</sub>, H4-K8<sub>ac</sub>K16<sub>ac</sub>, H4-K8<sub>ac</sub>K12<sub>ac</sub>K16<sub>ac</sub> e H4-K5<sub>ac</sub>K8<sub>ac</sub>K12<sub>ac</sub>K16<sub>ac</sub>. Para simplificar a notação, essas variantes serão denominadas, respectivamente, como A1<sub>A</sub>, A1<sub>B</sub>, A2, A3 e A4.

As dinâmicas a serem analisadas foram realizadas pelo grupo do Prof. Dr. Garegin Papoian (University of Maryland) e publicadas em um estudo prévio (157). As simulações foram realizadas no pacote AMBER12 e utilizaram o campo de força amber99SB\* (159) para proteínas, ions94 (160) para íons, e TIP3P (161) para água, com condições NVT, passo de integração de 2 fs e termostato de Langevin. A conformação inicial da forma nativa da cauda H4 foi obtida de um trabalho anterior (150), e os modelos acetilados foram preparados a partir dela utilizando a ferramenta *xleap* do AmberTool12. O sistema foi simulado através de protocolo de REMD (162), com 60 réplicas entre as temperaturas de 300 K a 450 K. Cada réplica foi simulada por 100 ns, totalizando 6  $\mu$ s de trajetórias para cada sistema. Os últimos 90 ns de cada trajetória à temperatura de 300 K foi utilizado para análise.

Para obter um espaço de fase efetivo único para todas as variantes, foram consideradas apenas as coordenadas dos carbonos alfa no cálculo da dissimilaridade. De cada trajetória da cauda H4, selecionou-se uma conformação a cada 14 quadros, após o descarte dos primeiros 4.000 quadros, considerados uma fase adicional de equilíbrio. Dessa forma, foram selecionadas 6.000 conformações de cada modelo, totalizando uma projeção ELViM com 36.000 conformações.

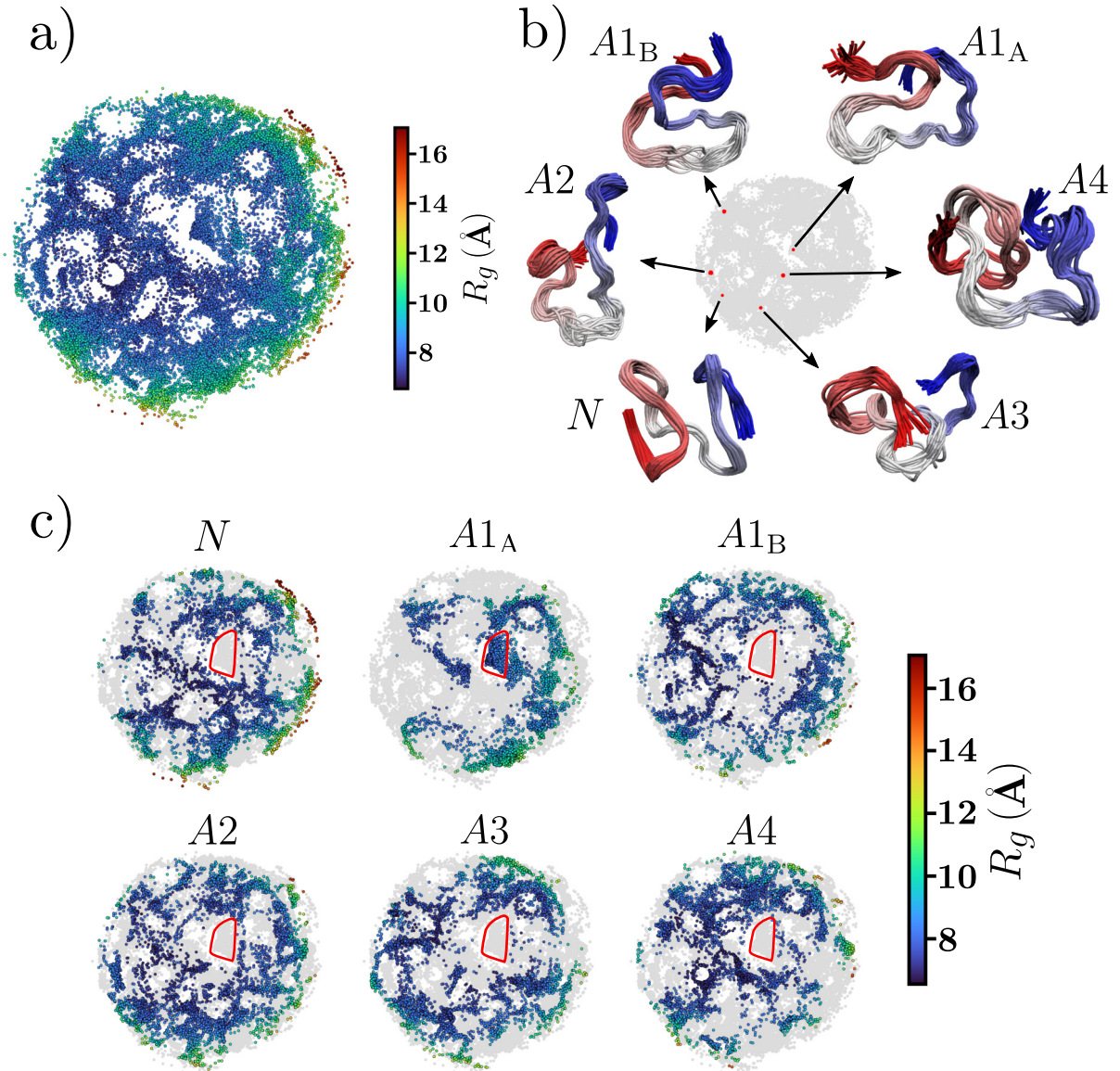
### 5.2.2 Resultados

Considerando-se apenas as posições do carbonos alfa da cauda H4 em suas formas nativa (N) e acetiladas (A1<sub>A</sub>, A1<sub>B</sub>, A2, A3 e A4), o ELViM foi utilizado para gerar o espaço de fase efetivo mostrado na figura 21.a, na qual os pontos foram coloridos de acordo com o raio de giração ( $R_g$ ) de cada conformação. Os valores de  $R_g$  variam de forma suave através da projeção ELViM, com valores mais baixos próximo à região centro-esquerda e valores mais altos na borda superior direita da projeção. A figura 21.b mostra, como exemplo, uma LCS de cada variante. Cada LCS é composta por 30 conformações sobrepostas, e as conformações que formam a LCS são destacadas como pontos vermelhos na projeção. A análise de tais LCSs ressalta a capacidade do ELViM de agrupar estruturas similares, mesmo no caso de proteínas altamente flexíveis.

Após a projeção do espaço de fase efetivo, é possível analisar como os pontos correspondentes à cada modelo se distribuem através do espaço global. Esse resultado é mostrado na figura 21.c, na qual os pontos de cada modelo rotulado são coloridos de acordo com seus valores de  $R_g$ , enquanto as conformações dos outros modelos são mostradas em cinza. Nota-se que os *ensembles* de cada variante não estão distribuídos de forma uniforme, com alguns modelos acessando apenas determinadas regiões do espaço efetivo. Os modelos N e A2 são os que se encontram mais distribuídos através da projeção, refletindo uma maior heterogeneidade estrutural. Por outro lado, observa-se que a forma monoacetilada A1<sub>A</sub> encontra-se consideravelmente menos distribuída pelo espaço, refletindo uma redução na heterogeneidade estrutural do modelo. Os modelos A3 e A4 também tendem a concentrar conformações em regiões específicas do espaço de fase.

A Figura 21.c ainda revela que algumas regiões do espaço efetivo são exclusivamente ocupadas por um único modelo. O caso mais importante está destacado por um contorno vermelho e indica região exclusivamente ocupada pelo modelo A1<sub>A</sub>. Como discutido anteriormente, tem-se seguido que a acetilação não apenas afeta a heterogeneidade conformacional da H4, mas também remodela seu relevo energético, dando origem a novos mínimos locais. A existência de tais regiões na projeção ELViM mostra que o método possui sensibilidade de identificar tais conformações exclusivas, separando-as em um agrupamento específico, sem a necessidade de informação adicional, além de sua métrica usual. Do ponto de vista biológico, a reorganização de mínimos locais e o surgimento de estruturas

Figura 21 – Projeção ELViM para cauda H4 nativa e suas formas acetiladas. (a) Espaço de fase efetivo gerado para o ELViM contendo a forma nativa da cauda H4 e as formas acetiladas  $A1_A$ ,  $A1_B$ ,  $A2$ ,  $A3$  e  $A4$ . (b) Um LCS de cada variante é mostrado para ilustrar o espaço gerado. (c) O espaço de fase efetivo de cada variante é mostrado, colorido com base no raio de giração, sobre todo o espaço efetivo em cinza.



Fonte: Elaborada pelo autor (2025).

exclusivas podem se relacionar a diferentes funções biológicas acessadas por acetilações específicas.

Devido à importâncias dessas conformações exclusivas e considerando que pequenas regiões da projeção ELViM possuem uma assinatura conformacional específica, foi proposta uma métrica, denominada fração de ocupação, para identificar tais grupos ou regiões no espaço de fase efetivo. Para isso, subdividimos a projeção ELViM em uma grade de  $27 \times 27$

compartimentos e definimos a fração de ocupação,  $F_{i,j}^w$ , do modelo  $w$  no compartimento  $(i, j)$  como:

$$F_{i,j}^w = \frac{n_{i,j}^w}{N_{i,j}}, \quad (5.1)$$

sendo  $n_{i,j}^w$  o número de conformações do modelo  $w$  no compartimento  $(i, j)$  e  $N_{i,j}$  o número total de conformações presentes nesse compartimento. O número de compartimentos foi escolhido levando-se em conta o tamanho dos agrupamentos observados no espaço gerado e a quantidade de pontos contidos na projeção. Dessa forma,  $F_{i,j}^w$  é uma medida adimensional que representa a proporção de ocupação de cada modelo em cada compartimento. Para evitar vieses associados às bordas da projeção, compartimentos periféricos com estatística insuficiente foram desconsiderados.

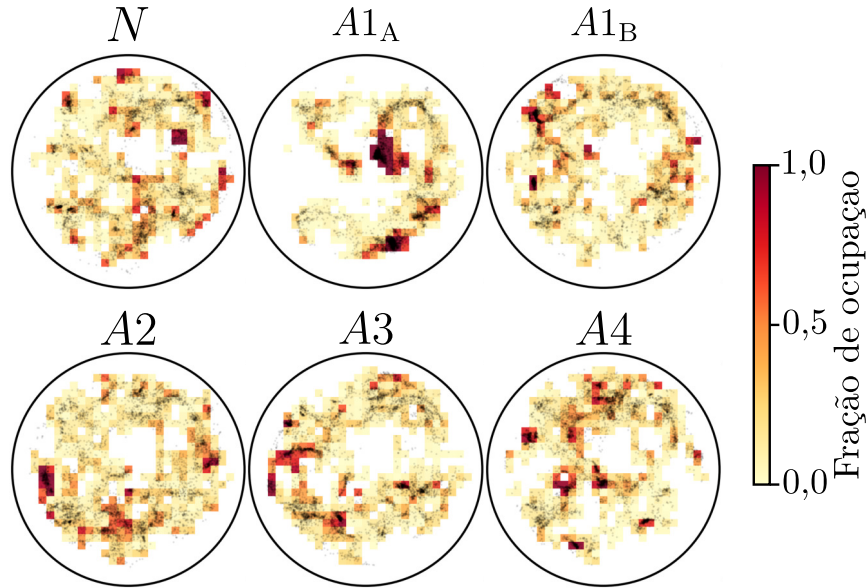
A fração de ocupação de cada modelo analisado é mostrada na Figura 22. Observa-se que a maioria das regiões do espaço de fase é compartilhada entre todos os modelos, com cada um apresentando uma fração de ocupação em torno de 0,2. No entanto, cada modelo também ocupa diferentes regiões do espaço de fase com alta exclusividade. Para o modelo A1<sub>A</sub>, identificam-se duas grandes regiões, que abrigam cerca de 5% de suas conformações em compartimentos com fração de ocupação igual a 1. Para os demais modelos, pequenos agrupamentos de pontos também apresentam fração de ocupação próxima a 1. Essas regiões evidenciam conformações que ganham estabilidade devido às PTMs presentes em cada modelo.

Uma caracterização mais detalhada de cada *ensemble* é fornecida no Apêndice D.1. Para tanto, a densidade de pontos em cada espaço efetivo individual foi calculada, identificando as regiões com alta densidade. Para cada variante, seis LCSs foram extraídos de regiões de alta densidade ou fração de ocupação elevada. Cada LCS é representado com 20 estruturas alinhadas, e o respectivo mapa de frequência de contato para cada região selecionada também é apresentado.

A análise dessas regiões revela um padrão interessante: a maioria das áreas com alta fração de ocupação também coincide com regiões de elevada densidade no relevo projetado para cada modelo. Vale ressaltar que a projeção ELViM organiza as conformações de forma contínua, conforme sua similaridade estrutural. Dessa forma, muitas dessas regiões não são isoladas, sugerindo que conformações semelhantes foram amostradas em outros modelos. No entanto, o fato dessas regiões exclusivamente ocupadas apresentarem alta densidade indica que correspondem a mínimos locais específicos de cada variante, reforçando a hipótese de que a acetilação remodela o relevo energético de cada uma delas.

Por fim, como uma forma de estimar a heterogeneidade conformacional de cada *ensemble*, foi calculada a entropia de ocupação de cada modelo sobre o espaço efetivo gerado pelo ELViM. Para isso, leva-se em consideração a distribuição das conformações

Figura 22 – Fração de ocupação. O espaço efetivo global foi dividido em uma grade com 27x27 e a fração de ocupação de cada variante foi calculada para cada compartimento da grade.



Fonte: Elaborada pelo autor (2025).

nos compartimentos da grade. Especificamente, para cada modelo  $w$ , foi estimada a distribuição de probabilidade de encontrar  $n_{i,j}^w$  conformações no compartimento  $(i, j)$  como  $p_{i,j}^w = n_{i,j}^w / N^w$ , sendo  $N^w$  o número total de conformações do modelo  $w$ . A entropia da projeção é então calculada como:

$$H^w = - \sum_{i,j} p_{i,j}^w \ln p_{i,j}^w. \quad (5.2)$$

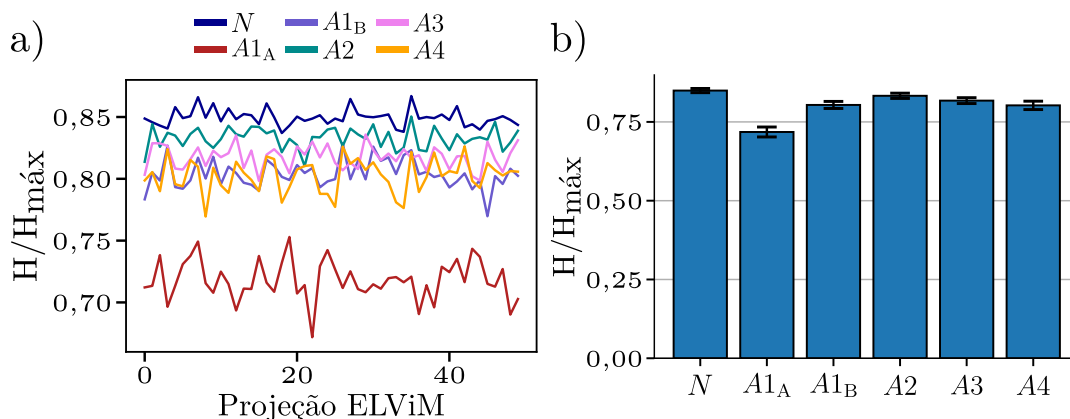
Essa grandeza é adimensional e, para o tamanho de grade adotado, atinge um valor máximo de  $H_{\text{máx}} = 6,59$ , correspondente à distribuição uniforme das conformações entre os compartimentos. Para facilitar a comparação entre os modelos, os valores de entropia foram normalizados pela divisão por  $H_{\text{máx}}$ .

Como discutido no Capítulo 2, o ELViM possui caráter estocástico, estando sujeito a pequenas variações que, entretanto, não impactam as propriedades globais da projeção (61, 63). Para estimar a robustez da medida de entropia proposta, foram realizadas 50 projeções ELViM independentes (não mostradas) e a entropia da projeção foi calculada para cada uma delas, conforme ilustrado na Figura 23.a. A entropia média normalizada e o desvio padrão de cada projeção são apresentados na Figura 23.b.

Analisando os resultados, observa-se que a forma nativa apresenta a maior entropia média, enquanto a forma monoacetilada  $A1_A$  possui a menor, refletindo a maior redução em seu espaço conformacional acessível. A variante  $A1_B$  também exibe entropia inferior à da

forma nativa. Entre os modelos A2, A3 e A4, identifica-se uma leve tendência decrescente. Contudo, as diferenças são estatisticamente pequenas, com sobreposição das barras de erro, especialmente entre as variantes A3 e A4.

Figura 23 – Entropia da projeção e heterogeneidade conformacional da cauda H4. (a) Entropia da projeção calculada para cada variante da cauda H4 em 50 projeções ELViM independentes. (b) Média e desvio padrão calculados através das 50 projeções.



Fonte: Elaborada pelo autor (2025).

Em um trabalho anterior, Winogradoff e colaboradores (157) exploraram o espaço conformacional dessas trajetórias utilizando um método distinto de redução dimensional. Inicialmente, os autores conduziram uma análise de agrupamento, utilizando como métrica os valores de RMSD entre pares, seguida por uma projeção PCA para os centros dos grupos previamente identificados. A heterogeneidade conformacional foi quantificada por meio de duas abordagens: (i) o cálculo do número de grupos, ordenados por tamanho, necessários para conter uma determinada fração acumulada das conformações amostradas; e (ii) a distância média entre o centro de cada grupo e o centro de massa da projeção PCA. Essas análises indicaram, também, que a variante nativa e A1<sub>A</sub> apresentam, respectivamente, a maior e a menor heterogeneidade conformacional. Os autores sugeriram ainda que as acetilações progressivas promovem uma tendência de redução na heterogeneidade. Embora não seja possível realizar uma comparação quantitativa direta, os resultados apresentados aqui estão em concordância qualitativa com os obtidos por Winogradoff e colaboradores.

Como vantagem do método aqui proposto, o espaço efetivo único oferece uma visualização intuitiva que permite analisar, de forma integrada, a heterogeneidade estrutural e a presença de conformações exclusivas. Além disso, a estimativa de heterogeneidade não depende de técnicas de agrupamento, sendo o número de compartimentos da grade o único parâmetro ajustável, que pode influenciar os resultados.

Conforme discutido anteriormente, a comparação de métodos de projeção multi-dimensional em sistemas biomoleculares apresenta desafios significativos. A ausência de

conhecimento prévio da topografia da superfície de energia e a inexistência de uma métrica absoluta de similaridade estrutural complicam essa tarefa. Contudo, para evidenciar as diferenças intrínsecas entre os métodos, a Figura D.7 do Apêndice D.1 apresenta uma comparação dos espaços efetivos gerados por PCA, dPCA (163) e t-SNE (50).

O dPCA consiste na aplicação da PCA aos ângulos diedros,  $\Phi$  e  $\Psi$ , que são mapeados no círculo unitário por meio das funções seno e cosseno devido à sua periodicidade. O t-SNE, por sua vez, é uma técnica de redução de dimensionalidade não-linear que preserva a estrutura local dos dados, otimizando distribuições de probabilidade para refletir as similaridades entre os pontos. Embora estudos demonstrem a eficácia do t-SNE na identificação de agrupamentos em *ensembles* heterogêneos de IDPs, o método prioriza a preservação da estrutura local, podendo não capturar a organização global da paisagem conformacional (164).

Ao colorir os espaços efetivos (Figura D.7) pelos valores de  $R_g$ , observa-se maior sobreposição das bacias desnaturada e nativa nas técnicas PCA e dPCA. Em contraste, o t-SNE segrega o espaço em múltiplos agrupamentos locais, sem organização global em relação ao raio de giração. Os subespaços ocupados por cada modelo são comparados na Figura D.8 do Apêndice D.1, sendo também possível observar regiões exclusivamente povoadas, principalmente para a variante A1<sub>A</sub>, nos espaços projetados por dPCA e t-SNE.

Por fim, a região de maior ocupação relativa de cada modelo foi selecionada para obter uma LCS, mostrada em todas as projeções na Figura D.9. Apesar de estruturalmente distintas, as LCSs (A2) e (A1<sub>B</sub>) são projetadas como sobrepostas pela PCA, enquanto as LCSs (N) e (A3) se sobrepõem no espaço dPCA. O t-SNE identifica todas as LCSs selecionadas como agrupamentos distintos, mas não preserva a estrutura global da paisagem conformacional. O ELViM, por outro lado, não visa a identificação de agrupamentos, mas parece oferecer maior equilíbrio entre aspectos locais e globais, garantindo a variação suave de coordenadas de reação ao longo da projeção, mantendo uma organização contínua das conformações. Isso facilita a comparação dos *ensembles* e a identificação de regiões exclusivamente povoadas.

### 5.3 ANÁLISE DO RELEVO DE ENERGIA DA HISTONA H1

A histona de ligação H1 associa-se ao DNA de ligação e à região de transição entre o DNA enrolado no nucleossomo e o DNA de ligação (Figura 20). Rica em lisinas e altamente carregada positivamente, a H1 desempenha um papel essencial na compactação da cromatina (165, 166). Em eucariotos, a H1 é composta por um domínio amino-terminal curto (NTD), um domínio globular central (GD) e um domínio carboxi-terminal longo e altamente desordenado (CTD) (167).

O CTD desordenado é crucial para a afinidade da associação da H1 ao nucleossomo,

além de influenciar diretamente a estruturação e compactação da cromatina (166,168–170). Para se analisar de forma detalhada os modos de ligação da histona H1 sobre o nucleossomo, deve-se levar em conta a geometria do nucleossomo. Considera-se que o eixo de simetria do nucleossomo passa sobre a região central, onde as duas metades do octâmero de histonas se encontram, e próximo aos pontos de entrada e saída do DNA de ligação. A H1 pode se associar ao nucleossomo em dois modos distintos: (i) axialmente, ou sobre o eixo de simetria (*on-dyad*); e (ii) lateralmente, ou fora do eixo de simetria (*off-dyad*) (171,172). Esses diferentes modos de ligação podem variar durante o enovelamento da cromatina, resultando em distintos níveis de condensação (166,173).

Nesta seção, o ELViM será aplicado para analisar os dados provenientes de uma simulação computacional de um nucleossomo completo, incluindo a forma completa da histona H1, previamente descrita por Wu e colaboradores (170). Neste trabalho, os autores demonstraram que a presença do longo domínio desordenado CTD é fundamental para afinidade da histona ao nucleossomo. Além disso, mostraram através de um histograma de coordenadas esféricas centradas no nucleossomo, que a histona H1 possui modos preferenciais de ligação. Nesta seção, uma trajetória obtida para a forma completa da H1 será revisitada com o objetivo de mapear o espaço conformacional, buscando estados conformacionais preferenciais. Além disso, o espaço conformacional gerado pelo ELViM será utilizado para investigar se os modos preferenciais de ligação ao nucleossomos se relacionam a assinaturas conformacionais específicas.

### 5.3.1 Sistema analisado

As trajetórias de simulação *coarse-grain* da histona H1 foram realizadas por Hao Wu, no grupo do Prof. Dr. Garegin Papoian, e previamente analisadas por outros métodos (170). Para realizar a simulação, Wu e colaboradores desenvolveram um modelo AWSEM (174) para simular a H1, incluindo suas extremidades desordenadas, ligada ao nucleossomo completo. Este campo de força híbrido, denominado AWSEM-DNA, integra o AWSEM (174), sua ramificação para proteínas intrinsecamente desordenadas AWSEM-IDP (175), e o campo de força CG de DNA 3SPN.2 (176). A conformação inicial do sistema H1-nucleossomo foi obtida a partir de uma estrutura experimental (PDB: 5NL0) (177), com os terminais modelados através do MODELLER (149). Um total de 50 réplicas foi simulado, utilizando um passo de integração de 5 fs, ensemble NVT e um termostato do tipo Langevin, totalizando 3  $\mu$ s de trajetórias de produção – correspondendo a, aproximadamente milissegundos de simulação em escalas atômicas.

As coordenadas dos carbonos alfa da H1 foram extraídas para as análises neste estudo. Para a projeção ELViM, foi selecionada uma conformação a cada 13 quadros, após descartar os 4.000 quadros iniciais, totalizando 30.005 conformações.

### 5.3.2 Resultados

O espaço de fase conformacional efetivo gerado pelo ELViM para a histona H1 está ilustrado na Figura 24.a. Embora a alta flexibilidade das caudas da H1 seja refletida em uma grande variação nos valores de  $R_g$ , o ELViM projeta um espaço conformacional no qual os valores de  $R_g$  variam, novamente, de forma aproximadamente suave por todo o espaço.

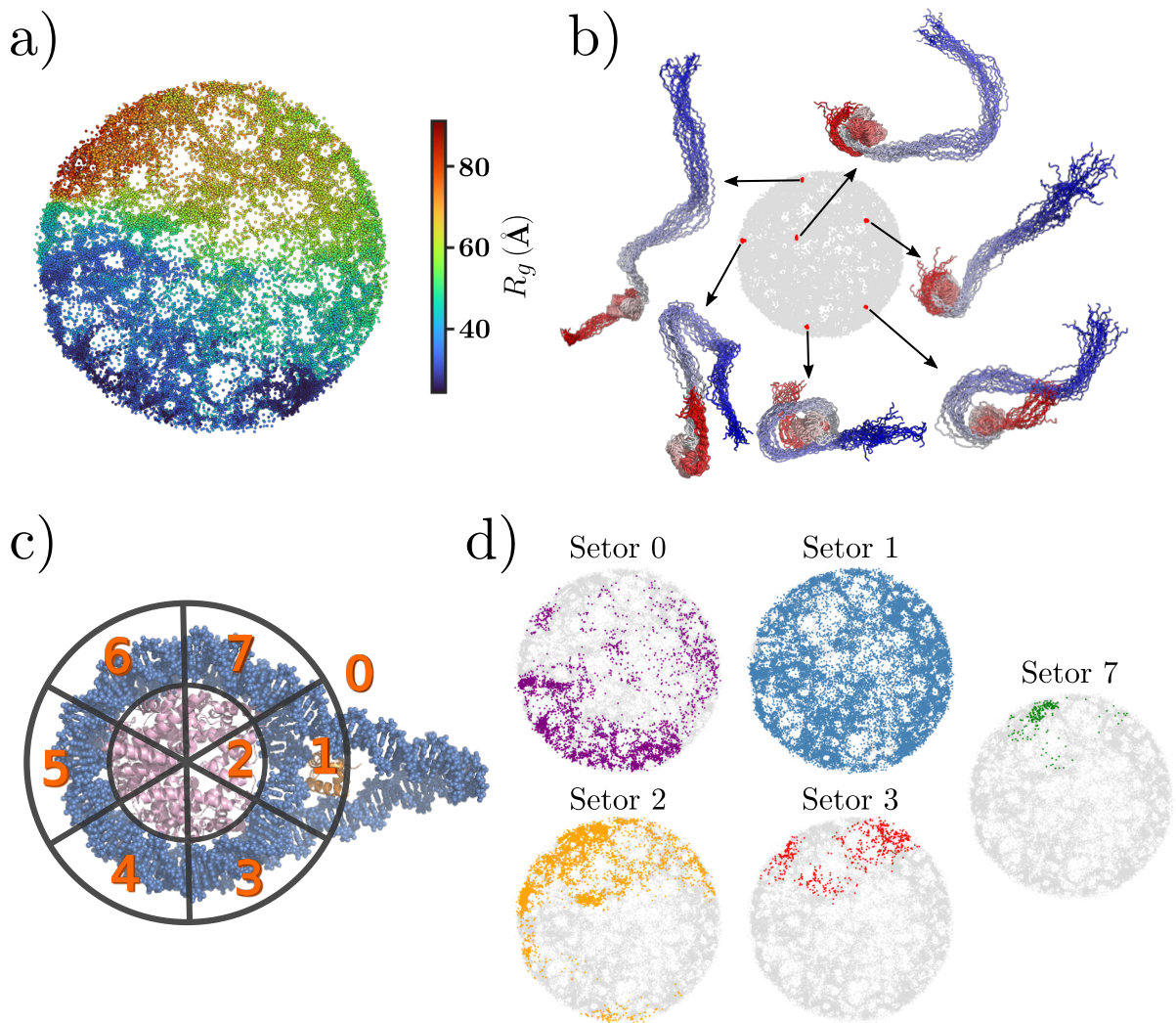
Para ilustrar a variabilidade conformacional no espaço de fase efetivo, a Figura 24.b apresenta cinco LCSs selecionadas de regiões com alta densidade de pontos. Cada LCS é composta por 20 conformações, que estão destacadas em vermelho na projeção, e representadas com a extremidade amino-terminal em vermelho. O alto alinhamento estrutural observado nas vizinhanças de cada LCS sugere a existência de conformações preferenciais, diferindo do comportamento de um polímero aleatório.

Adicionalmente, o espaço conformacional gerado pelo ELViM foi utilizado para investigar se o modo de ligação da histona ao nucleossomo se relaciona a assinaturas conformacionais específicas. Para isso, foi estabelecido um sistema de coordenadas cilíndricas centradas no nucleossomo e o espaço em seu entorno foi dividido em seis setores cilíndricos iguais (setores 1, 3-7), como mostrado na figura 24.c. O setor 0 abrange regiões distantes do nucleossomo, enquanto o setor 2, adjacente ao núcleo de histonas, compreende o modo de ligação lateral. Cada quadro da simulação foi atribuído a um setor, de acordo com código desenvolvido e aplicado por Hao Wu.

Na Figura 24.d, a projeção ELViM é reproduzida e cores distintas são utilizadas para destacar o setor do qual cada conformação foi amostrada. Considerando todos os quadros simulados, a frequência de ocupação dos setores foi: setor 0: 10,1%, setor 1: 78,3%, setor 2: 9,8%, setor 3: 1,3%, setor 7: 0,5%. Os setores 4, 5 e 6 não foram visitados durante a simulação. Dado que a maioria das conformações foi amostrada do setor 1, era esperado que o espaço de fase efetivo fosse amplamente ocupado por este setor. Entretanto, é interessante notar que as conformações estão distribuídas por todo o espaço, sendo ausentes as regiões exclusivamente ocupadas, como observado para os modelos da cauda H4 discutidos anteriormente. Este resultado sugere que a histona H1 mantém grande heterogeneidade conformacional, mesmo quando ligada próximo ao eixo de simetria do nucleossomo. Apesar disso, a densidade de pontos não é homogênea através do espaço efetivo, sendo possível a identificação de mínimos locais, que refletem a existência de conformações preferenciais, como mostrado no Apêndice D.2.

Outro resultado interessante é que conformações amostradas nos demais setores estão restritas a pequenas vizinhanças do espaço de fase. No entanto, não é possível distinguir se tal restrição na heterogeneidade conformacional resulta, de fato, de preferências conformacionais induzidas pela ligação ou efeitos secundários tais como efeitos estéricos ou

Figura 24 – Espaço de fase conformacional efetivo da histona H1. (a) Projeção ELViM colorida de acordo com valores de  $R_g$ . (b) Assinaturas Conformacionais Locais, formadas por 20 conformações alinhadas. Conformações selecionadas foram destacadas em vermelho na projeção. Resíduos próximos ao amino-terminal foram desenhados em vermelho. (c) O espaço ao redor do nucleossomo foi dividido em oito setores cilíndricos. (d) Conformações amostradas de cada setor são exibidas em cores distintas sobre o espaço de fase completo, mostrado em cinza. Os setores não ilustrados nesta figura indicam aqueles que não foram visitados durante a simulação. Frequências de ocupação dos setores, considerando todos os quadros simulados, são: setor 0: 10,1%, setor 1: 78,3%, setor 2: 9,8%, setor 3: 1,3%, setor 7: 0,5%.



Fonte: Elaborada pelo autor (2025).

amostragem limitada dentro desses setores. Vale ressaltar, que as simulações analisaram modos de ligação espontâneos ao nucleossomo, não tendo por objetivo amostrar amplamente outros modos de ligação. Comparando-se os setores 0, 1 e 2, que possuem maior amostragem, pode-se notar que as regiões de maior densidade estão distribuídas em diferentes partes do espaço de fase. Tal resultado corrobora a ideia de que o setor de ligação esteja associado a

diferentes preferências conformacionais, pelo menos para a amostragem realizada. Detalhes sobre as preferências conformacionais observadas em cada setor são apresentados na Figura D.10 do Apêndice D.2.

### 5.3.3 Discussão

Neste estudo, o ELViM foi utilizado para analisar a dinâmica conformacional das caudas intrinsecamente desordenadas das histonas H4 e H1. Os resultados obtidos destacam os efeitos da acetilação e da ligação ao nucleossomo sobre o relevo conformacional das caudas H4 e da histona H1, respectivamente.

A projeção do ELViM para a cauda da H4 revelou variações significativas no espaço conformacional acessível entre a forma nativa e as diversas formas acetiladas. Notavelmente, a acetilação tende a afetar a heterogeneidade conformacional, provocando alterações nas regiões acessível do espaço global e na distribuição dos mínimos locais. Todas as variantes analisadas apresentaram grupos locais com alta fração de ocupação no espaço, sugerindo que padrões específicos de acetilação podem induzir preferências conformacionais distintas e, potencialmente, funções biológicas únicas. A projeção ELViM se mostra eficaz em identificar tais grupos, fornecendo um panorama único para realização de uma análise conformacional diferencial entre os *ensembles*. Por fim, foi sugerido um novo método para estimativa de heterogeneidade estrutural, calculando a entropia da projeção ELViM de cada variante. O resultado obtido concorda qualitativamente com análises anteriores, apresentando sensibilidade estatística semelhante à previamente relatada. (157).

A análise da projeção ELViM para a histona de ligação H1 sugere que o modo de ligação pode induzir mudanças nas preferências conformacionais, restringindo a dinâmica conformacional a regiões específicas do espaço de fase. Nesse contexto, a ligação a diferentes setores ao redor do nucleossomo limita a dinâmica conformacional da H1, destacando o papel da interação na definição dos estados funcionais das IDPs.

## 5.4 ANÁLISE DE *ENSEMBLES* DE PROTEÍNAS DESORNADAS DEPOSITADOS NO *THE PROTEIN ENSEMBLE DATABASE (PED)*

Esta seção é baseada no artigo: *Understanding the Energy Landscape of Intrinsically Disordered Protein Ensembles*, JCIM, 2024. (63).

Do ponto de vista teórico, o principal desafio em estudos de IDPs é obter *ensembles* conformacionais confiáveis, dada a plasticidade estrutural dessas proteínas. Para isso, nos últimos anos, diversos estudos focaram no desenvolvimento ou aprimoramento de campos de força, modelos de solvente, modelos CG, além de técnicas de amostragem capazes de explorar eficientemente um vasto espaço de fase (178–184).

Por outro lado, diversos métodos biofísicos podem ser utilizados para caracterizar estados desordenados de proteínas, incluindo ressonância magnética nuclear (NMR), dicroísmo circular (CD), espalhamento de raios X a baixo ângulo (SAXS) e transferência de energia de ressonância Förster (FRET) (185). Esses métodos oferecem valiosas informações sobre a organização estrutural das biomoléculas. No entanto, tais técnicas geralmente refletem propriedades médias que, isoladamente, são insuficientes para definir de maneira única o *ensemble* estrutural subjacente (186).

Neste cenário, uma alternativa para se obter conjuntos conformacionais mais confiáveis é integrar dados experimentais com técnicas computacionais (71, 142, 143, 187). Recentemente, um banco de dados denominado *Protein Ensemble Database (PED, protei-nensemble.org)* foi criado para o depósito de *ensembles* estruturais de IDPs, gerados por tais métodos integrativos (144). Posteriormente, *ensembles* conformacionais gerados por métodos baseados em inteligência artificial generativa foram também incluídos (188).

Como uma forma de verificar a capacidade do ELViM de realizar a análise comparativa de *ensembles* conformacionais gerados em diversas condições, quatro *ensembles* diferentes, depositados no PED, foram selecionados para análise (63). Nesta seção, serão detalhados os resultados obtidos para dois desses sistemas: um fragmento da proteína do complexo do poro nuclear Nup153, referido por NUS, e o domínio SH3 amino-terminal da proteína Drk (*Downstream of Receptor Kinase*). Os outros dois sistemas serão brevemente apresentados no Apêndice D.3. Por fim, vale destacar que este estudo não tem por objetivo validar ou avaliar criticamente os *ensembles* conformacionais, mas sim ilustrar e avaliar o emprego do ELViM como ferramenta para facilitar a análise comparativa dos mesmos.

#### 5.4.1 Sistema analisado

Todos os *ensembles* analisadas neste estudo foram gerados através de abordagens experimentais e computacionais integradas e estão disponíveis no PED. Detalhes adicionais sobre como os *ensembles* foram obtidos estão disponíveis nas respectivas entradas PED e nos correspondentes artigos de referência. Os *ensembles* analisados são:

- NUS: fragmento (1313-1390) da proteína do complexo do poro nuclear Nup153 (UniProtKB P49790), disponível sob os códigos de acesso PED00149 e PED00150 e descritos na referência (189). Os *ensembles* são: NUS1 (PED00149: e001, 9482 conformações), NUS2 (PED00149: e002, 9405 conformações), NUS3 (PED00149: e003, 9473 conformações), NUS4 (PED00150: e001, 9255 conformações), NUS5 (PED00150: e002, 9248 conformações) e NUS6 (PED00150: e003, 9277 conformações). Estes *ensembles* foram baseados em simulações atomísticas de *Replica Exchange* com solvatação implícita utilizando ABSINTH e CAMPARI (190, 191). Em seguida, os *ensembles* foram ponderados através do programa COPER (192) para corresponder aos valores experimentais de  $Rg^2$  e diferentes eficiências médias de FRET. Os

valores experimentais foram obtidos em condições desnaturantes (solução salina tamponada com fosfato (*buffer PBS*) com 6M de ureia) e nativas (*buffer PBS*) por meio de smFRET e SAXS. Para a projeção ELViM, foi selecionada uma a cada duas conformações, resultando em uma projeção com 28.078 conformações.

- drkN SH3: Domínio SH3 amino-terminal da proteína Drk, descritos na referência (193). Neste caso, *ensembles* consistentes com dados de NMR, SAXS e smFRET foram geradas pelos autores utilizando uma nova abordagem denominada *Extended Experimental Inferential Structure Determination (X-EISD)*. Os *ensembles* conformacionais se referem ao estado desnaturado e são descritos no PED como: SH3-1 (PED00156, 100 conformações), um *ensemble* otimizado a partir de um conjunto aleatório gerado pelo TraDES (194,195); SH3-2 (PED00157, 100 conformações), um *ensemble* conformacional inicialmente ponderado pelo ENSEMBLE (196); e SH3-3 (PED00158, com 88 conformações), um *ensemble* misto contendo conformações obtidas dos dois *ensembles* anteriores. Todas as conformações foram utilizadas na projeção ELViM, totalizando 288 pontos.

#### 5.4.2 Resultados

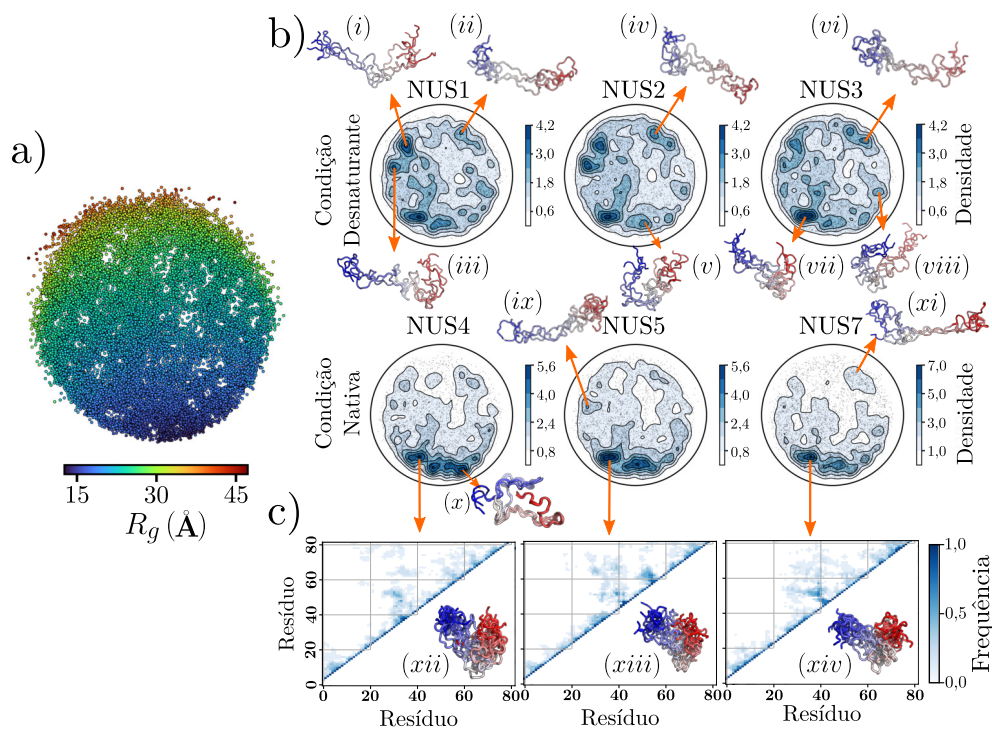
Nesta seção, o ELViM será adotado para analisar *ensembles* conformacionais gerados por métodos integrativos, que utilizam pesos estatísticos ou extraem subconjuntos de conformações com o intuito de reproduzir valores experimentais médios. Para este estudo, foram escolhidos sistemas que possuíam *ensembles* gerados para condições biofísicas distintas, tais como nativa e desnaturante, ou que foram gerados por diferentes métodos para corresponder aos mesmos dados experimentais. Neste caso, a análise diferencial previamente discutida será utilizada para avaliar o relevo conformacional de cada *ensemble* analisado.

##### 5.4.2.1 Análise dos ensembles do fragmento NUS

Seis *ensembles* do fragmento NUS foram analisados, incluindo três gerados sob condição desnaturante (NUS1, NUS2 e NUS3) e três em condição nativa (NUS4, NUS5 e NUS6), como descrito na seção 5.4.1. Todos os *ensembles* da NUS apresentam, aproximadamente, o mesmo raio de giração médio: (23,38, 23,78, 23,51, 21,45, 21,72, 21,69 Å para NUS1 a NUS6, respectivamente).

A Figura 25.a ilustra o espaço de fase efetivo projetado pelo ELViM, contendo as conformações dos seis *ensembles*. Os pontos foram coloridos de acordo com os valores de  $R_g$ . A Figura 25.b indica como cada *ensemble* se distribui através do espaço global gerado pelo ELViM. Desta vez, a densidade de pontos foi estimada, através de um KDE com núcleo gaussiano, e indicada separadamente para cada caso. Em cada figura, apenas

Figura 25 – Projeção ELViM para os *ensembles* do fragmento NUS. (a) Espaço de fase efetivo ELViM gerado para os *ensembles* do fragmento NUS. Os *ensembles* foram reamostradas utilizando uma em cada duas conformações, resultando em 28.078 conformações. (b) Densidade de estados estimada para os *ensembles* individuais. As estruturas representativas das conformações de regiões arbitrariamente selecionadas são mostradas com o amino-terminal em azul. (c) Frequência de contatos para região de alta densidade similar nos três *ensembles* gerados sob condição nativa.



Fonte: Adaptada de (63).

a densidade do espaço efetivo individual é mostrada e uma circunferência, cujo centro coincide com o centro da projeção total, foi incluída como referencial.

Como discutido anteriormente, em simulações não-enviesadas e bem amostradas, regiões de maior densidade correspondem a mínimos na superfície de energia livre. No entanto, isso não ocorre quando são utilizados vieses ou técnicas de amostragem que favorecem a amostragem de regiões específicas do espaço de fase. Do mesmo modo, técnicas de reamostragem integrativa podem resultar em *ensembles* que não seguem uma distribuição canônica, e os picos de densidade deixam de representar mínimos de energia livre do sistema. Ainda assim, essas regiões destacam as partes do espaço efetivo preferencialmente amostradas, revelando as conformações predominantes em cada *ensemble*.

Ao se comparar os mapas de densidades mostrados na figura 25.b, é possível observar que cada *ensemble* exibe uma distribuição única de picos de densidade ao longo

de toda a projeção ELViM. Uma mudança drástica no perfil de densidade pode ser observada ao comparar os conjuntos correspondentes às condições desnaturante e nativa. Considerando apenas uma condição (desnaturante ou nativa), é possível observar como os picos de densidade variam entre os diferentes ensembles, revelando regiões do espaço de fase privilegiadas na construção de cada um. A alta semelhança no raio de giro para os *ensembles* analisados destaca a utilidade desta abordagem, que permite identificar diferenças nas conformações prevalentes, mesmo entre conjuntos que se ajustam ao mesmo observável experimental médio.

Na Figura 25.b, são apresentadas LCSs para algumas regiões de alta densidade. Note que as LCSs, cada uma composta por três estruturas, apresentam uma variação estrutural maior em comparação aos casos anteriores, indicando que o espaço gerado é mais heterogêneo estruturalmente e com poucos mínimos energéticos bem definidos. Sob condições desnaturantes, as LCSs (i)–(viii) exemplificam regiões que são consistentemente mais densamente povoadas em todos os três *ensembles*, mas com valores de densidade variados. Outras pequenas bacias também ganham ou perdem densidade em diferentes *ensembles*, destacando diferenças de composição entre eles. Os mapas de densidade gerados sob condição nativa diferem significativamente dos gerados para condição desnaturante, ocupando majoritariamente a porção inferior da projeção ELViM. Curiosamente, a LCS (x) representa uma região com muitas conformações bem alinhadas estruturalmente, mas que perde densidade nos outros dois *ensembles* nativos, os quais apresentam maior heterogeneidade conformacional mesmo nas regiões de alta densidade.

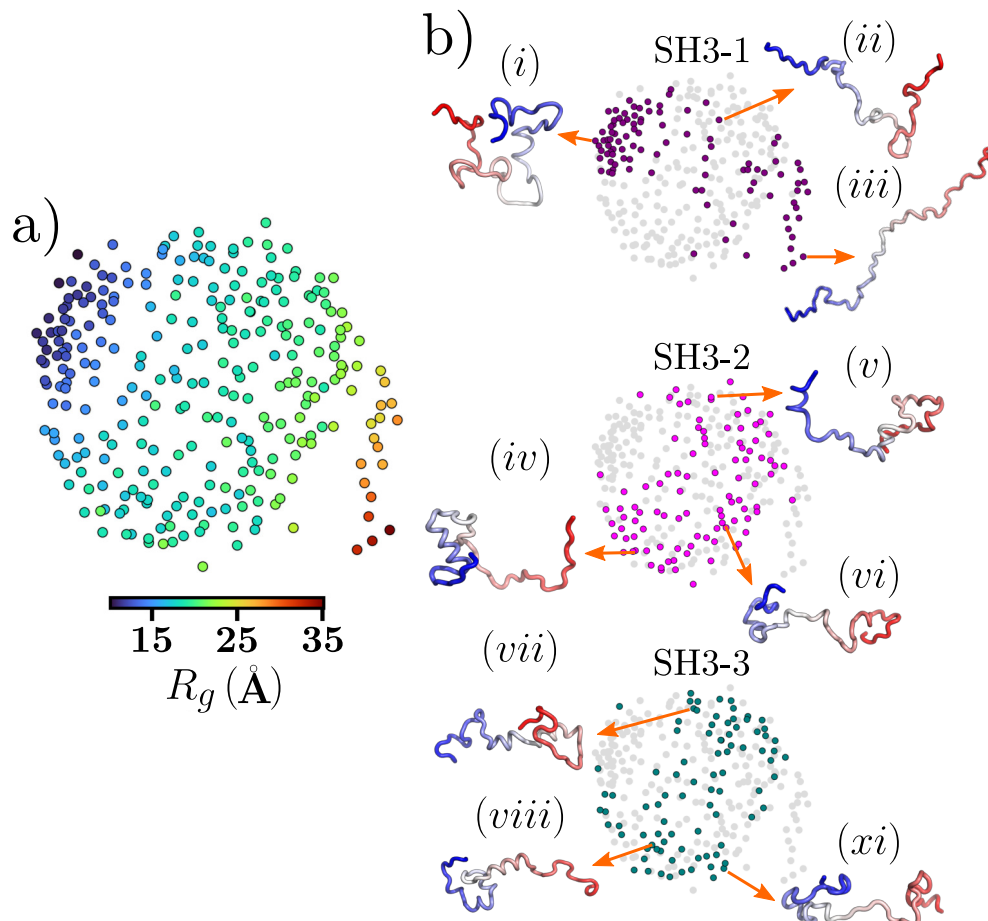
Algumas bacias de alta densidade também aparecem consistentemente em todos os *ensembles* analisados, como as regiões indicadas na figura 25.c. A análise das conformações pertencentes à tais bacias permite identificar características estruturais que são mantidas ou alteradas em cada *ensemble*. Como um exemplo, foram selecionadas todas as conformações dentro da curva de nível com maior densidade nos três *ensembles* gerados sob a condição nativa e a frequência de contatos observadas para tais conformações foi calculada (Figura 25.c). As estruturas contidas nessas regiões são mostradas na parte inferior do mapa de contatos e ilustram a heterogeneidade estrutural característica de cada *ensemble*.

#### 5.4.2.2 *Análise dos ensembles do domínio drkN*

Por fim, foram analisados três *ensembles* do domínio drkN. A projeção ELViM, abrangendo todas as 288 conformações, é apresentada na Figura 26. Os pontos estão novamente coloridos com base nos valores de  $R_g$ . Observa-se que as conformações mais estendidas (pontos vermelhos) formam novamente uma estrutura semelhante a uma cauda, em oposição às conformações com menor raio de giração (pontos azul-escuros).

Para analisar o modo como cada *ensemble* se distribui sobre o espaço efetivo, os espaços individuais são mostrados na Figura 26.b. Como ilustração, algumas conformações

Figura 26 – Análise ELViM do domínio SH3 DrkN. (a) Projeção ELViM gerada para os três *ensembles* do domínio SH3 DrkN, totalizando 288 conformações, com pontos codificados por cor de acordo com os valores de  $R_g$ . (b) As conformações de cada conjunto são destacadas em roxo (SH3-1), fúcsia (SH3-2) e ciano escuro (SH3-3). Em cada caso, algumas conformações selecionadas arbitrariamente são apresentadas, com o amino-terminal em azul.



Fonte: Adaptada de (63).

foram arbitrariamente selecionadas e mostradas com a extremidade amino-terminal em azul. Ao contrário do caso anterior, estimar de forma confiável a densidade de estados torna-se desafiador com poucos ou esparsos pontos de dados. No entanto, uma análise cuidadosa da Figura 26.b revela alguns grupos locais com pontos (quase) sobrepostos.

Comparando-se os três *ensembles*, nota-se que o SH3-1 apresenta um padrão de amostragem distinto dos demais, sendo formado principalmente por conformações mais estendidas e mais compactas, com apenas algumas conformações na porção central da projeção. Por outro lado, os *ensembles* SH3-2 e SH3-3 ocupam a mesma região central da projeção ELViM, sugerindo uma amostragem semelhante. No entanto, é possível identificar diferenças na densidade de pontos de dados e na ocorrência de pequenos grupos locais. Tal

resultado revela um comportamento interessante, uma vez que segundo a documentação disponível, o ensemble SH3-3 foi gerado a partir de um *ensemble* maior composto por uma mistura dos outros dois *ensembles* originais dos quais SH3-1 e SH3-2 foram obtidos.

### 5.4.3 Discussão

No presente estudo, foram abordados *ensembles* de IDPs obtidos por diferentes métodos integrativos. A análise diferencial da ocupação do espaço de fase efetivo gerado pelo ELViM revela diferentes relevos conformacionais, destacando diferenças nos padrões de amostragem e identificando as conformações mais prevalentes dentro de cada *ensemble*.

Deve-se notar que este estudo não tem por objetivo discutir as metodologias empregadas para obtenção dos *ensembles* ou validar os mesmos. Por outro lado, esses dados foram usados para testar a capacidade do ELViM de discernir entre os *ensembles* com o mínimo de informação fornecida. Como resultado interessante, o método releva diferentes preferências conformacionais e uma grande variabilidade estrutural, mesmo para *ensembles* que levam ao mesmo observável médio.

Os sistemas aqui analisados possuem uma heterogeneidade estrutural maior, como evidenciado pelo alinhamento das LCSs indicadas, e, em alguns casos, um número limitado de estruturas. Apesar disso, o ELViM é capaz de convergir para uma representação robusta e reprodutível do espaço de fase, como indicado na Figura D.13 do Apêndice D.3. Essa figura indica como a posição de alguns grupos locais varia em três execuções independentes do algoritmo. Para facilitar a comparação, rotações e inversões das projeções foram retiradas através de alinhamento dos pontos. O resultado indica a coesão de tais grupos locais, além da manutenção do ordenamento relativo entre eles.

## 5.5 CONCLUSÃO

Neste capítulo, o ELViM foi utilizado para investigar o espaço conformacional de sistemas desordenados, incluindo *ensembles* gerados por simulações moleculares tradicionais e métodos integrativos. Especificamente, foi realizada uma análise diferencial avaliando como cada *ensemble* ocupa o espaço efetivo global, indicando preferências conformacionais e vieses de amostragem. Além disso, foi discutido como o ELViM pode ser utilizado para a obtenção de detalhes estruturais adicionais, como mapas de frequência de contatos e propriedades médias para regiões específicas do espaço de fase efetivo.

Recentemente, outras abordagens para comparar *ensembles* conformacionais foram sugeridas na literatura (197–201). No entanto, uma abordagem comparativa continua complexa, devido à diversidade de parâmetros e abordagens, à ausência de dados experimentais específicos e à natureza multidimensional dos sistemas. Esse desafio é particularmente evidente na avaliação de representações multidimensionais (91,92).

Neste contexto, uma comparação qualitativa entre ELViM e técnicas usuais, tais como PCA e t-SNE, foi apresentada para o espaço efetivo gerado para as variantes da cauda H4. Os resultados indicam que o ELViM apresenta um bom equilíbrio entre a preservação da vizinhança local, evidenciada pela análise de LCSs, e a estrutura global do espaço, indicada pela variação suave das coordenadas de reação através da projeção. Por outro lado, a tentativa de preservar ambos os aspectos pode fazer com que algumas bacias locais apareçam mais dispersas no ELViM, por exemplo, quando comparado ao resultado obtido pelo t-SNE.

Em suma, este estudo reforça a utilidade do ELViM na dissecação de paisagens conformacionais complexas de IDPs. A flexibilidade do método, permitindo o ajuste de parâmetros como a dissimilaridade e a taxa de aprendizado, amplia seu potencial para análise de sistemas ainda mais complexos. O ordenamento global do espaço e o alinhamento local, evidenciado pelas LCSs, facilitam a análise diferencial, auxiliando na distinção entre *ensembles* e na elucidação de mecanismos atomísticos subjacentes à função das IDPs.

## 6 INTERAÇÃO DE PEPTÍDEOS ANTIMICROBIANOS COM MEMBRANA MODELO POR MEIO DE ANÁLISE DO RELEVO DE ENERGIA

Este capítulo é baseado no artigo “*Probing Mastoparan-like Antimicrobial Peptides Interaction with Model Membrane Through Energy Landscape Analysis (JPCB)*” (64).

### 6.1 INTRODUÇÃO

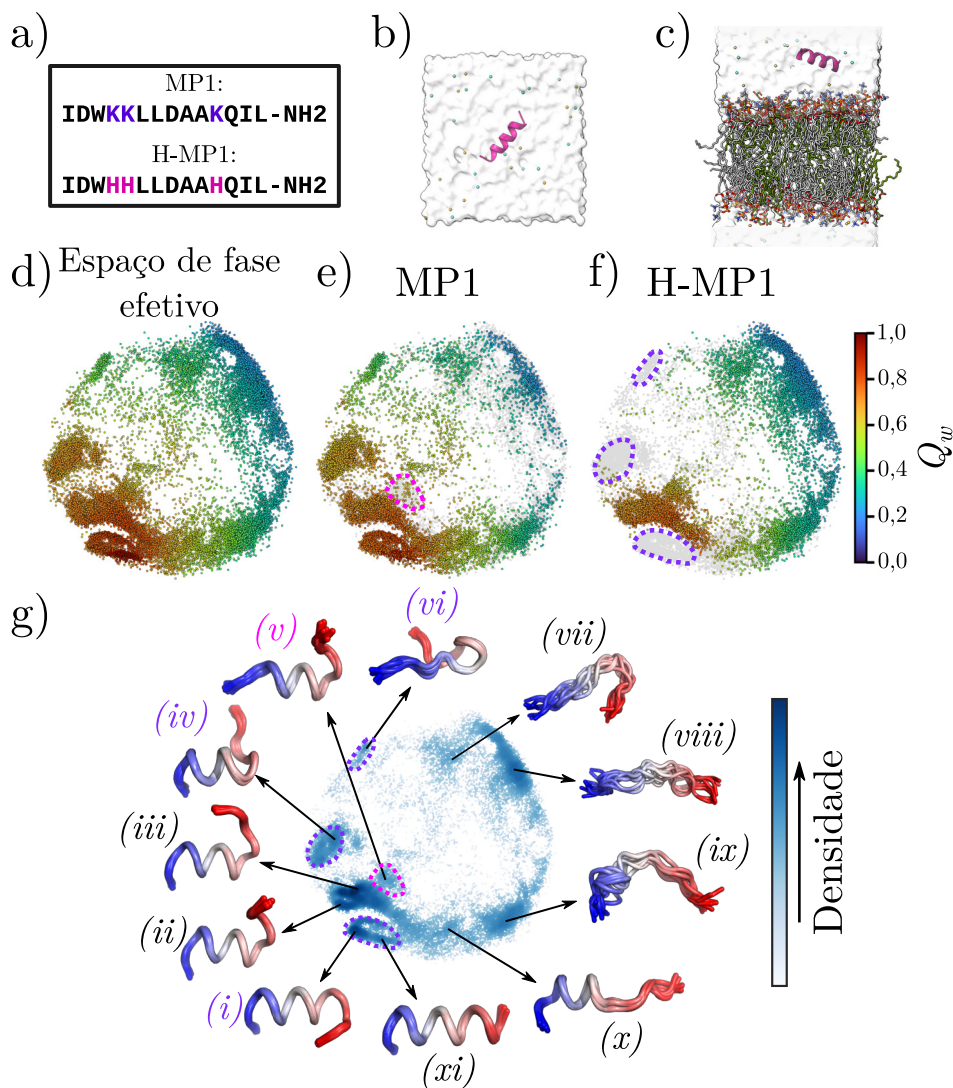
Os Peptídeos Antimicrobianos (AMPs, do inglês, *Antimicrobial Peptides*) emergem como uma alternativa promissora frente ao aumento da resistência aos antibióticos vivenciada nas últimas décadas. Tais peptídeos atuam através da adsorção na membrana plasmática bacteriana (202, 203), que perturba o empacotamento lipídico e leva à morte celular devido ao vazamento citoplasmático.

O Polybia-MP1, ou simplesmente MP1, é um AMP pertencente à família dos mastoparanos, que é extraído do veneno da vespa brasileira *Polybia paulista*. O MP1 possui ação bactericida de amplo espectro e, diferente de outros matoparanos, não apresenta atividade hemolítica ou citotóxica (81). Além disso, tem sido demonstrado que o MP1 possui atividade antitumoral (203, 204), apresentando seletividade entre linfócitos leucêmicos e saudáveis (205).

Como uma estratégia de aumentar a seletividade do peptídeo para ambientes ácidos (206–208), um análogo do MP1, denominado H-MP1, foi sintetizado substituindo-se as lisinas por histidinas (209). Tal modificação estrutural conferiu ao H-MP1 uma maior sensibilidade a variações de pH próximas ao pH fisiológico, principalmente devido à dinâmica de protonação das cadeias laterais de histidina. Embora ambos os peptídeos, MP1 e H-MP1, possuam a mesma carga positiva (+2) em pH ligeiramente ácido, o MP1 é mais eficiente em perturbar o empacotamento lipídico de bicamadas modelo (210). Resultados experimentais mostraram que o MP1 apresenta uma afinidade com a membrana 1,5 vezes maior do que o H-MP1 (210). Além disso, o H-MP1 possui uma inserção mais rasa e cerca de três vezes menos atividade lítica do que o MP1 (210).

Neste estudo, o ELViM será utilizado para analisar simulações moleculares dos peptídeos MP1 e H-MP1 (figura 27.a) realizadas em solução aquosa (Figura 27.b) e na presença de uma bicamada modelo (Figura 27.c). A metodologia de análise conformacional diferencial discutida no capítulo anterior será empregada para investigar os impactos da

Figura 27 – Sistemas analisados, espaço de fase efetivo e LCSs para os peptídeos em solução aquosa. (a) Estrutura primária dos peptídeos MP1 e H-MP1 destacando as substituições de aminoácidos. (b) Representação do sistema simulado contendo o peptídeo em solução aquosa. (c) Representação do sistema simulado com peptídeo em presença de membrana modelo. (d) Projeção ELViM contendo os *ensembles* estruturais amostrados para ambos MP1 e H-MP1 em solução aquosa. A projeção contém conformações amostradas em todas as temperaturas. Os pontos são coloridos de acordo com a similaridade ( $Q_w$ ) à estrutura de referência. Os *ensembles* estruturais obtidos para (e) MP1 e (f) H-MP1 são mostrados com pontos coloridos de acordo com  $Q_w$  sobreposto à projeção completa, mostrada em cinza. (g) A densidade de pontos na projeção foi estimada através de KDE gaussiano e os pontos foram coloridos de acordo com o logaritmo natural da densidade. As setas indicam, para algumas regiões de alta densidade, a LCS composta por 10 estruturas sobrepostas, representadas com amino-terminal em azul.



Fonte: Adaptada de (64).

substituição de aminoácidos e da adsorção na membrana na heterogeneidade conformacional amostrada para cada peptídeo.

## 6.2 SISTEMA ANALISADO

Nesse estudo, serão analisadas simulações realizadas pela Dra. Ingrid B. S. Martins durante seu doutorado e pós-doutorado no IBILCE/UNESP. Inicialmente, ambos peptídeos foram simulados, isoladamente, em solução aquosa ( $[\text{NaCl}] = 150 \text{ mM}$ ) à pressão de 1 atm em diversas temperaturas. Para o MP1, foram realizadas simulações a 300, 315, 330, 345 e 360 K, enquanto para o H-MP1, foram utilizadas temperaturas de 280, 290, 300, 315 e 330 K. Cada simulação foi conduzida por 600 ns. Como os peptídeos possuem estabilidade diferentes, foram necessários distintos intervalos de temperatura para explorar adequadamente os estados enovelado e desnaturado. A temperatura para comparação dos dois sistemas e cálculo da energia livre foi definida como 300 K para corresponder aos dados experimentais.

Para estudar o comportamento dos peptídeos na presença de uma bicamada lipídica, as simulações utilizadas na referência (210) também foram incluídas nesta análise. Essas simulações foram realizadas utilizando uma bicamada lipídica equilibrada composta por POPC:POPG (razão de 70:30) com 60 moléculas por camada, à temperatura de 300 K e uma pressão de 1 atm. A bicamada lipídica foi gerada utilizando o *Membrane Builder plugin* (211), disponível através do CHARMM-GUI (212).

Para obter a bicamada lipídica equilibrada, ela foi inicialmente solvatada em uma caixa de água TIP3P (213) com dimensões de aproximadamente  $65 \times 65 \times 120 \text{ \AA}$ . A energia foi minimizada com 10.000 passos de gradiente conjugado e 100 ns de MD de equilíbrio até que o valor da área por lipídio parasse de variar. Os peptídeos foram adicionados na fase aquosa da bicamada previamente equilibrada, a cerca de  $10 \text{ \AA}$  de distância da superfície da membrana (Figura 27.c). Íons suficientes foram adicionados para manter nula a carga líquida da caixa de simulação. O sistema peptídeo–bicamada foi novamente equilibrado, realizando 10.000 passos de minimização de energia, seguidos por 10 ns de dinâmica molecular, com restrições de posição aplicadas ao esqueleto do peptídeo para relaxamento do solvente. As simulações de produção, sem restrições, foram realizadas por 1.200 ns para cada peptídeo.

Todas as simulações foram realizadas através do NAMD (86), utilizando o campo de força CHARMM 36 (85), condições NPT, passo de integração de 2 fs e condições de contorno periódicas. Em todas as simulações, o estado de ionização de cada resíduo foi escolhido com base em um pH ligeiramente ácido, com as lisinas protonadas no MP1, histidinas protonadas no H-MP1 e o ácido aspártico deprotonado para ambos os peptídeos. Esses estados foram escolhidos para comparar com resultados experimentais já publicados (208, 210).

O ELViM foi usado para projetar um espaço de fase conformacional efetivo que abrangesse os conjuntos conformacionais de ambos os peptídeos MP1 e H-MP1 amostrados em solução aquosa. As conformações foram selecionadas de cada trajetória de forma igualmente espaçada após descartar os 100 ns iniciais, considerados como equilíbrio adicional. Para projeção, o número de conformações selecionadas para cada temperatura foi: 5000 (300 K), 3125 (315 K), 3125 (330 K), 1250 (345 K) e 1250 (360 K) para MP1; e 3125 (280 K), 3125 (290 K), 5000 (300 K), 1250 (315 K), 1250 (330 K) para H-MP1.

Uma segunda projeção independente foi gerada incluindo os conjuntos conformacionais amostrados em simulações de adsorção espontânea. Nesse caso, foram adicionadas às conformações anteriores 3.335 conformações para o MP1 e 3.221 para o H-MP1, simuladas à temperatura de 300 K.

## 6.3 RESULTADOS

### 6.3.1 Peptídeos em solução aquosa

Inicialmente, o ELViM foi utilizado para analisar o espaço de fase efetivo amostrado por ambos os peptídeos em solução aquosa sob diferentes temperaturas. O intervalo de temperaturas foi selecionado de modo a garantir uma amostragem adequada para cada peptídeo. A Figura E.1 do Anexo E apresenta histogramas de energia amostrados em cada temperatura e o perfil de energia livre obtido para cada peptídeo. Considerando, portanto, as conformações amostradas para ambos os peptídeos em todas as temperaturas, obtém-se o espaço de fase efetivo ilustrado na Figura 27.d, com os pontos coloridos de acordo com os valores de  $Q_w$ , calculados em relação a uma conformação gerada em hélice alfa. A Figura E.2 do Anexo E reproduz a projeção ELViM, indicando, por meio de um mapa de cores, as temperaturas nas quais cada conformação foi amostrada.

As Figuras 27.e e 27.f exibem o espaço de fase individual amostrado para cada peptídeo, colorido de acordo com os valores de  $Q_w$  e sobreposto ao espaço de fase completo, representado em cinza. É possível observar que os peptídeos ocupam regiões distintas do espaço de fase: enquanto o MP1 visita todos os grupos da bacia do estado enovelado, o H-MP1 ocupa apenas a região central. Por outro lado, o H-MP1 ocupa a região do estado desenovelado com maior densidade do que o MP1. Adicionalmente, foi destacado em rosa, na Figura 27.e, uma região densamente populada pelo H-MP1, mas esparsamente ocupada pelo MP1. Do mesmo modo, na Figura 27.f, foram destacadas em roxo três regiões exclusivamente ocupadas pelo MP1. Como discutido anteriormente, a restrição no espaço de fase individual revela impactos estruturais decorrentes da substituição de aminoácidos, indicando regiões do espaço de fase global que ganham ou perdem estabilidade.

Para melhor descrever o espaço de fase efetivo obtido, a densidade de pontos na projeção foi estimada através de KDE gaussiano. Na Figura 27.g, os pontos foram coloridos

de acordo com o logaritmo natural da densidade local, utilizando uma escala que varia de branco, para densidade zero, a azul escuro, para a região mais densa. Em seguida, algumas regiões de alta densidade foram selecionadas para a análise das LCSs, nas quais dez estruturas representativas de cada região selecionada são exibidas em sobreposição. A comparação dessas assinaturas revela que a bacia nativa é composta principalmente por conformações que mantêm a porção central da hélice alfa, enquanto os terminais se desenovelam e adotam diferentes orientações. Alguns exemplos também são fornecidos para a bacia do estado desnaturado, que exhibe considerável heterogeneidade conformacional, contendo conformações totalmente estendidas. Finalmente, deve-se observar que o MP1 popula mais densamente, em relação ao H-MP1, as regiões com maior quantidade de estrutura helicoidal (maiores valores de  $Q_w$ ). Esse resultado está em concordância qualitativa com observações experimentais obtidas por calorimetria diferencial (210), as quais indicam que o MP1 e o H-MP1 apresentam, respectivamente, cerca de 46% e 7% de conformação helicoidal em solução aquosa ácida.

### 6.3.2 Adsorção dos peptídeos em bicamada lipídica

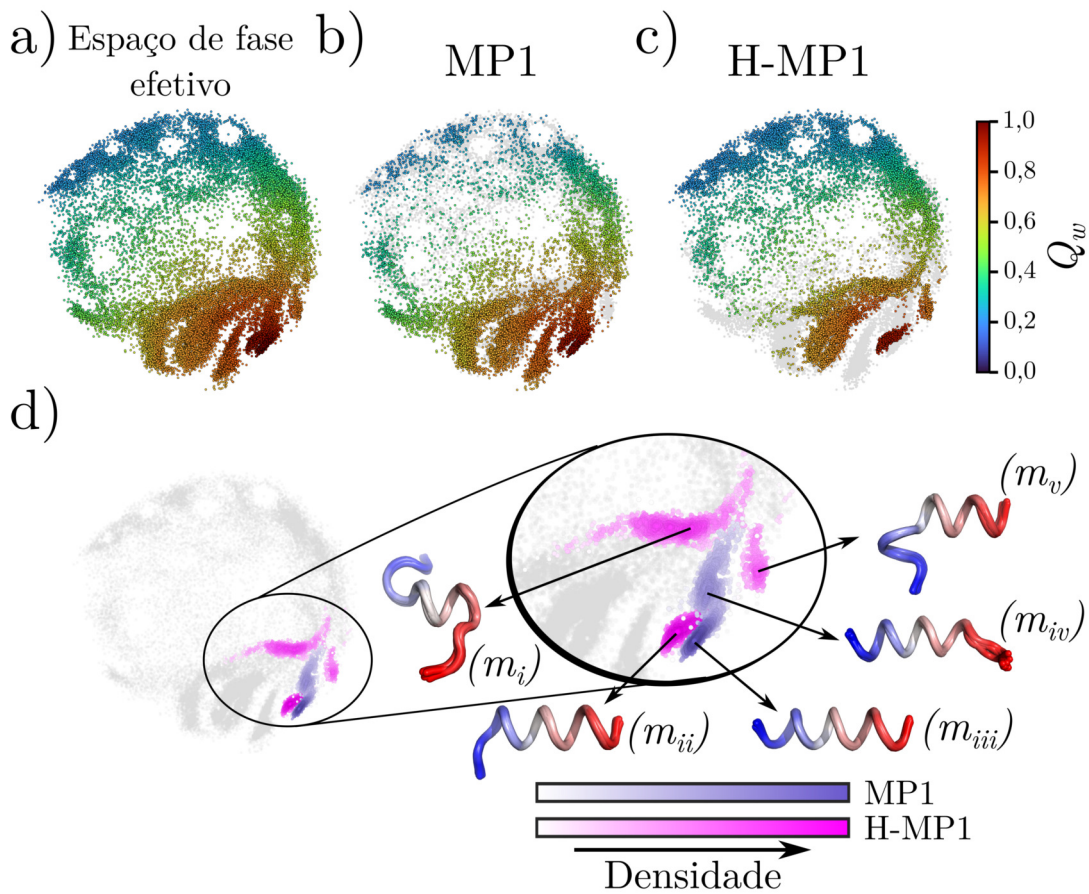
Com o objetivo de realizar uma análise comparativa dos estados estruturais amostrados pelos peptídeos na ausência e presença de membrana modelo, foram incorporadas ao sistema analisado na seção anterior conformações de ambos os peptídeos obtidas na presença da membrana modelo, conforme descrito na Seção 6.2. Essas trajetórias foram previamente descritas por Martins e colaboradores (210). O espaço de fase efetivo resultante é apresentado na Figura 28.

A Figura 28.a exhibe o espaço de fase efetivo composto por conformações de ambos os sistemas (membrana modelo ausente e presente). Os *ensembles* individuais do MP1 e H-MP1 são representados nas Figuras 28.b e 28.c, respectivamente. A inclusão de conformações do sistema com membrana insere novos agrupamentos locais na bacia do estado enovelado. Esse resultado indica que, após a adsorção, os peptídeos adquirem conformações distintas das amostradas apenas em solução aquosa, como esperado. Similarmente ao resultado da Figura 27, o espaço efetivo ocupado pelo H-MP1 continua restrito a apenas uma porção da bacia do estado enovelado.

A Figura E.3 (Apêndice E) reproduz os espaços individuais (Figura 28), separando as conformações amostradas na ausência e na presença da membrana modelo. Para o sistema com membrana, as conformações foram ainda classificadas com base na posição do peptídeo em relação à membrana, distinguindo três estados: peptídeo em solução aquosa próximo à membrana, peptídeo interagindo com a membrana ou parcialmente adsorvido e peptídeo completamente adsorvido. A Figura 28.d apresenta a densidade estimada considerando-se apenas as conformações amostradas no sistema com membrana.

De modo geral, os resultados obtidos evidenciam o impacto significativo do campo

Figura 28 – Espaço de fase efetivo e LCSs para o sistema peptídeo–membrana (a) Espaço de fase efetivo contendo conformações amostradas em várias temperaturas em solução aquosa e em temperatura ambiente no sistema com membrana. Os pontos são coloridos de acordo com os valores de  $Q_w$ . Os *ensembles* individuais do (b) MP1 e (c) H-MP1 foram projetados sobre toda a projeção, colorida em cinza. (d) Conformações amostradas no sistema de membrana, com conformações de MP1 em roxo e de H-MP1 em fúcsia. A densidade de pontos, representando os *ensembles* de cada peptídeo, foi estimada utilizando-se KDE gaussiano. LCSs são mostradas para ilustrar conformações representativas.



Fonte: Adaptada de (64).

elétrico da membrana sobre a heterogeneidade conformacional dos peptídeos, restringindo a visitação a uma pequena região do espaço de fase conformacional efetivo. Além disso, observa-se pouca sobreposição entre as regiões ocupadas nos sistemas com e sem membrana, reforçando as alterações conformacionais induzidas pelo processo de adsorção.

A análise da Figura 28 revela ainda que o MP1 ocupa uma região contínua do espaço de fase efetivo, com dois picos de densidade caracterizados pelas LCSs  $(m_{iii})$  e  $(m_{iv})$ . A LCS  $(m_{iv})$ , associada a conformações nas quais o MP1 interage com a membrana ou está parcialmente adsorvido, indica que o carboxi-terminal permanece desenovelado. Por sua vez, a LCS  $(m_{iii})$  representa estruturas nas quais o peptídeo encontra-se completamente

adsorvido, adotando uma hélice alfa bem definida.

Por outro lado, o H-MP1 apresenta maior heterogeneidade estrutural, ocupando três regiões distintas do espaço de fase efetivo. A LCS ( $m_i$ ), que exhibe apenas a porção central da hélice alfa formada, representa estados nos quais o H-MP1 se encontra em solução aquosa próximo à membrana. Estruturas parcialmente adsorvidas distribuem-se principalmente nas regiões correspondentes às LCSs ( $m_{ii}$ ) e ( $m_v$ ), enquanto as conformações totalmente adsorvidas estão restritas à região da LCS ( $m_{ii}$ ). Dessa forma, a projeção ELViM sugere que, à medida que o H-MP1 se aproxima da membrana, a formação da hélice alfa progride do carboxi- para o amino-terminal. Este último, no entanto, mantém certa flexibilidade mesmo após a adsorção. Notavelmente, as LCSs associadas aos estados adsorvidos do MP1 e do H-MP1 são distintas, sem sobreposição significativa entre suas respectivas densidades.

### 6.3.2.1 Estimativa de energia livre efetiva na projeção ELViM

O espaço de fases efetivo mostrado na Figura 27 contém conformações amostradas em diferentes temperaturas, selecionadas para garantir uma melhor cobertura do espaço conformacional. Para estimar a energia livre efetiva da projeção, é necessário ponderar essas conformações levando em conta a probabilidade de se observar cada estado em temperatura ambiente. Para isso, foi empregado o método dos histogramas ponderados (WHAM, do inglês *Weighted Histogram Analysis Method*) (119,214), que permite estimar a densidade de estados  $\Omega(E)$  a partir das trajetórias realizadas em diferentes temperaturas. Cada trajetória analisada é composta por 30.000 conformações.

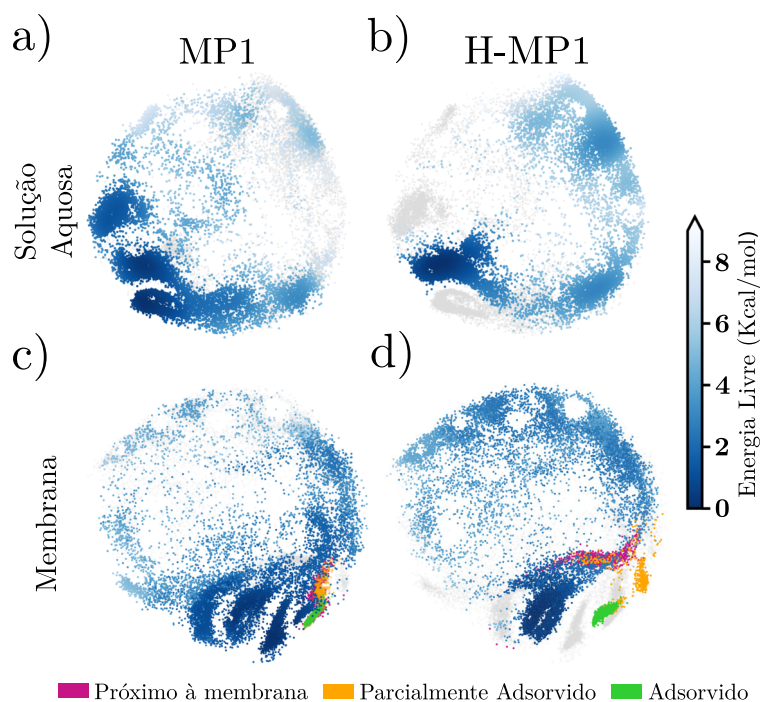
A probabilidade de se amostrar uma conformação com energia  $E_i$  a uma temperatura inversa  $\beta = (k_B T)^{-1}$  é dada por

$$P(E_i, \beta) = A \Omega(E_i) e^{-\beta E_i}, \quad (6.1)$$

na qual  $A$  é uma constante de normalização. Posteriormente, foi empregado um estimador de densidade por núcleos (KDE) gaussiano ponderado, atribuindo a cada ponto um peso igual a  $P(E_i, \beta)$ . Esse procedimento fornece uma estimativa não enviesada da densidade de probabilidade  $\rho$  e da energia livre, definida como  $F = -\frac{1}{\beta} \ln(\rho)$  (Seção 2.3). Um procedimento similar foi previamente empregado na referência (60).

As Figuras 29.a–b apresentam a energia livre estimada para a primeira projeção do ELViM (Figura 27.d), correspondente à simulação dos peptídeos apenas em solução aquosa. O mínimo de energia livre do MP1 e do H-MP1 está localizado, respectivamente, nas proximidades das LCSs ( $i$ ) e ( $iii$ ) (Figura 27.g). A bacia contendo o mínimo de energia livre para o MP1 inclui estruturas com um maior número de resíduos em hélice, enquanto o mínimo de energia livre para o H-MP1 compreende conformações em que apenas a região central da hélice está formada.

Figura 29 – Estimativa da energia livre efetiva. A densidade de pontos foi estimada por KDE gaussiano ponderado, com os pesos definidos pela Equação 6.1. (a) O mapa de cores representa a energia livre efetiva, calculada conforme descrito no texto, para o sistema em solução aquosa. (b) Os valores de energia livre obtidos para as conformações em solução aquosa são atribuídos aos pontos correspondentes na projeção do sistema com membrana, enquanto as conformações do sistema com membrana são classificadas de acordo com o nível de interação com a bicamada.



Fonte: Adaptada de (64).

Nas Figuras 29.c–d, os valores de energia livre calculados em solução aquosa foram utilizados para colorir os respectivos pontos na segunda projeção do ELViM (sistema com membrana). As conformações do sistema com membrana foram coloridas de acordo com o nível de interação do peptídeo com a bicamada. A análise dessas figuras revela que, no caso do MP1, as regiões de menor energia livre em solução aquosa estão conectadas de forma contínua às regiões populadas por conformações interagindo ou adsorvidas na membrana. Em contraste, para o H-MP1, as conformações parcialmente adsorvidas ou totalmente adsorvidas encontram-se distantes de seus mínimos de energia livre em solução aquosa.

Além disso, no caso do H-MP1, as regiões parcialmente adsorvidas e adsorvidas não estão diretamente conectadas às regiões populadas pelo peptídeo em água. Isso indica que o H-MP1, sendo menos estável que o MP1, atinge sua conformação enovelada apenas durante o processo de adsorção. Por outro lado, o MP1 naturalmente amostra conformações

que favorecem a adsorção, como evidenciado pela sobreposição das regiões ocupadas nas simulações com e sem membrana. Esses resultados sugerem que o H-MP1 precisa superar uma barreira de energia mais alta durante sua adsorção na bicamada lipídica, o que pode influenciar a dinâmica e a cinética desse processo.

## 6.4 CONCLUSÃO

Nesse estudo, o ELViM foi utilizado para analisar os *ensembles* conformacionais do MP1 e H-MP1, comparando-os em solução aquosa e na presença de uma membrana modelo. No primeiro caso, a análise diferencial permitiu identificar regiões específicas do espaço efetivo individual de cada peptídeo que ganharam ou perderam estabilidade devido à substituição de aminoácidos. Na presença da membrana modelo, as regiões acessíveis na projeção ELViM foram significativamente reduzidas para ambos os peptídeos, os quais se tornaram mais estruturados do que no sistema aquoso. A análise de estruturas representativas através das LCSs mostrou que o amino-terminal do MP1 se enovela antes da adsorção, enquanto o H-MP1 exibe o carboxi-terminal enovelado durante a adsorção parcial e total, mantendo o amino-terminal parcialmente desenovelado mesmo após a adsorção. A comparação dos mínimos de energia observados no sistema aquoso com as regiões ocupadas durante a adsorção espontânea fornece um indicativo das mudanças conformacionais envolvidas no processo e do papel da membrana no mecanismo de enovelamento final.

Detalhes sobre como a distribuição de carga, o tamanho das cadeias laterais carregadas e a estrutura secundária dos AMPs afetam a ligação à membrana ainda não são completamente esclarecidos (215–217). Outros aspectos que podem influenciar o processo de adsorção, como a agregação de peptídeos e a regulação de carga próxima à membrana, estão além do escopo deste estudo e não foram abordados. Estabelecer um panorama da relação estrutura-função que explique os mecanismos de ação dos AMPs ainda é um desafio. No entanto, o presente estudo mostra que o ELViM constitui uma metodologia intuitiva e eficaz para distinguir cenários possíveis.

## 7 CONSIDERAÇÕES FINAIS

Esta tese investigou o uso do ELViM para explorar o relevo da superfície de energia e a dinâmica conformacional de biomoléculas de interesse biológico. Ao longo do trabalho, uma nova implementação do método foi desenvolvida, permitindo a integração de todas as etapas em um único código e proporcionando maior eficiência computacional. A exploração inicial do espaço de parâmetros do método foi realizada de forma qualitativa e quantitativa, e generalizações do método, especialmente para o cálculo de dissimilaridade e da taxa de aprendizado, permitiram sua aplicação em sistemas mais complexos, como RNAs e IDPs.

Ao analisar moléculas de RNA, os resultados indicaram que o ELViM, assim como para proteínas, é capaz de gerar um espaço de fase efetivo com bom equilíbrio entre aspectos locais e globais. A análise das assinaturas LCSs para esses sistemas indica que o método possui bom alinhamento estrutural em vizinhanças locais, distinguindo corretamente não apenas o alinhamento da cadeia principal de açúcar fosfato, mas também das bases nitrogenadas. Os resultados obtidos, comparados com técnicas tradicionais como PCA e a projeção em coordenadas de reação ( $Q$ , RMSD), mostraram-se qualitativamente superiores. Os espaços de fase gerados foram utilizados para caracterizar rotas de enovelamento, mínimos locais na superfície de energia e *ensembles* do estado de transição.

Para o estudo de IDPs, o ELViM foi empregado para construir um espaço efetivo único, incluindo *ensembles* de diversas variantes ou gerados sob diferentes condições físico-químicas. Uma abordagem semelhante foi utilizada para realizar uma análise diferencial da dinâmica conformacional dos peptídeos análogos MP1 e H-MP1, quando simulados em solução aquosa ou em trajetórias de adsorção espontânea à membrana modelo. Essa metodologia permite visualizar como cada *ensemble* ocupa o espaço global, evidenciando estados conformacionais que ganham ou perdem estabilidade e indicando qualitativamente a heterogeneidade estrutural de cada um. Novamente, é interessante notar que, mesmo para esses casos, o ELViM produz um espaço efetivo no qual parâmetros de ordem variam de forma aproximadamente suave, mantendo bom alinhamento da vizinha local, como evidenciado pelas assinaturas LCSs.

Uma análise comparativa entre os resultados obtidos para a cauda H4 através do ELViM, de PCA de coordenadas cartesianas e diedrais e de um protocolo recentemente desenvolvido para o uso de t-SNE foi apresentado no Apêndice D.1. Nesse caso foi demonstrado que LCSs, caracterizadas por alto alinhamento estrutural e obtidas de regiões com alta fração de ocupação, aparecem sobrepostas e dispersas nos espaços gerados por PCA. Deve-se notar, entretanto, que as técnicas possuem objetivos distintos. Enquanto o ELViM busca manter as dissimilaridades, o PCA busca a direção efetiva que maximiza a variância das coordenadas. Já a aplicação de t-SNE sobre uma matriz de RMSD levou a

um espaço efetivo caracterizado por grupos segregados, com ótimo alinhamento estrutural interno. Esse resultado decorre do fato conhecido de que o t-SNE possui melhor desempenho na manutenção da estrutura da vizinhança local, podendo não preservar a orientação relativa global dos grupos. Para ambos os casos, a falta de estrutura global dificulta a interpretação do espaço gerado como um espaço de fase efetivo. Por outro lado, alguns grupos com alinhamento estrutural bem definido identificados pelo t-SNE aparecem mais dispersos no ELViM, o que poderia dificultar a identificação de mínimos locais menos populados.

Contudo, é importante ressaltar que essas dificuldades estão relacionadas às limitações da projeção de espaços de alta dimensão em planos bidimensionais, o que pode resultar em distorções, como aglomeração excessiva de pontos (*crowding*) no espaço projetado, ou até mesmo na criação de “buracos”. Essas distorções são características comuns dos métodos de redução dimensional (26,27). Estabelecer critérios abrangentes para quantificar e interpretar seu impacto na representação do espaço projetado permanece uma tarefa essencial e desafiadora.

No campo biofísico, avaliar a qualidade das projeções multidimensionais é uma tarefa complexa, que envolve duas etapas principais: verificar o quão bem as similaridades no espaço multidimensional são capturadas na matriz de dissimilaridade de entrada e quão eficientemente essa matriz é reproduzida no espaço efetivo. Embora várias métricas de qualidade tenham sido propostas para avaliar a qualidade das projeções (26,91), comparar os resultados obtidos por diferentes métodos e métricas continua sendo desafiador, especialmente devido à topografia desconhecida da superfície de energia (92). Uma abordagem comum na literatura consiste em analisar como os grupos identificados por diferentes técnicas de agrupamento se distribuem em distintas projeções (92,218). No entanto, avaliar a qualidade das métricas e técnicas de agrupamento recaí em problema semelhante (41). Dessa forma, estabelecer critérios mais precisos para se acessar a qualidade das representações em espaço reduzido e avaliar o impacto de distorções em grandezas de interesse biofísico continua sendo campo para pesquisas adicionais.

Neste contexto, o ELViM se consolida como uma ferramenta promissora para gerar visualizações intuitivas do relevo energético de sistemas biomoleculares a partir de *ensembles* estruturais obtidos por diferentes técnicas. Entre suas características encontra-se o bom equilíbrio dos aspectos locais e globais e dependência de poucos parâmetros, que podem ser utilizados para ajustar a resolução da métrica de dissimilaridade e a taxa de aprendizado do método. A métrica utilizada pode ser ainda ajustada à resolução de representação do sistema, podendo ser utilizada com todos os átomos ou em modelos mais simplificados onde uma partícula virtual pode representar diversos átomos ou resíduos. Entre as limitações mais fundamentais está o custo computacional para o cálculo da matriz de dissimilaridade, que limita a quantidade de conformações utilizadas e o tamanho

do sistema. Além disso, em sistemas excessivamente complexos, as projeções podem ser afetadas por distâncias degeneradas na métrica, resultando em matrizes cuja distribuição das dissimilaridades não convirja no plano ou que apresentem distorções excessivas. Neste contexto, esta tese contribui ao explorar sistemas de complexidade variada e comparar o desempenho do ELViM com outros métodos usuais, destacando suas potencialidades e ressaltando os desafios envolvidos.

## REFERÊNCIAS

- 1 ANFINSEN, C. B.; HABER, E.; SELA, M.; WHITE, F. H. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. **Proceedings of the National Academy of Sciences of the United States of America**, v. 47, n. 9, p. 1309–1314, 1961.
- 2 ANFINSEN, C. B. Principles that govern the folding of protein chains. **Science**, v. 181, n. 4096, p. 223–230, 1973.
- 3 LEVINTHAL, C. Are there pathways for protein folding? **Journal de Chimie Physique**, v. 65, p. 44–45, 1968.
- 4 LEVINTHAL, C. How to fold gracefully. *In*: DEBRUNNER, P.; TSIBRIS, J.; MUNCK, E. (ed.). **Mossbauer spectroscopy in biological systems**. Urbana, IL: University of Illinois Press, 1969.
- 5 CONTESSOTO, V. d. G.; JUNIOR, A. B. d. O.; CHAHINE, J.; OLIVEIRA, R. J. d.; LEITE, V. B. P. Introdução ao problema de enovelamento de proteínas: uma abordagem utilizando modelos computacionais simplificados. **Revista Brasileira de Ensino de Física**, SciELO Brasil, v. 40, p. e4307, 2018.
- 6 KIM, P. S.; BALDWIN, R. L. Intermediates in the folding reactions of small proteins. **Annual Review of Biochemistry**, v. 59, n. 1, p. 631–660, 1990.
- 7 MATTHEWS, C. R. Pathways of protein folding. **Annual Review of Biochemistry**, v. 62, n. 1, p. 653–683, 1993.
- 8 JACKSON, S. E.; FERSHT, A. R. Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. **Biochemistry**, v. 30, n. 43, p. 10428–10435, 1991.
- 9 BRYNGELSON, J. D.; WOLYNES, P. G. Spin glasses and the statistical mechanics of protein folding. **Proceedings of the National Academy of Sciences of the United States of America**, v. 84, n. 21, p. 7524–7528, 1987.
- 10 LEOPOLD, P. E.; MONTAL, M.; ONUCHIC, J. N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. **Proceedings of the National Academy of Sciences of the United States of America**, v. 89, n. 18, p. 8721–8725, 1992.
- 11 ONUCHIC, J. N.; LUTHEY-SCHULTEN, Z.; WOLYNES, P. G. Theory of protein folding: the energy landscape perspective. **Annual Review of Physical Chemistry**, v. 48, n. 1, p. 545–600, 1997.
- 12 BRYNGELSON, J. D.; ONUCHIC, J. N.; SOCCI, N. D.; WOLYNES, P. G. Funnels, pathways, and the energy landscape of protein folding: a synthesis. **Proteins: Structure, Function, and Bioinformatics**, v. 21, n. 3, p. 167–195, 1995.
- 13 DILL, K. A.; CHAN, H. S. From levinthal to pathways to funnels. **Nature Structural Biology**, v. 4, n. 1, p. 10–19, 1997.

- 
- 14 FERREIRO, D. U.; KOMIVES, E. A.; WOLYNES, P. G. Frustration in biomolecules. **Quarterly Reviews of Biophysics**, v. 47, n. 4, p. 285–363, 2014.
- 15 GIANNI, S.; FREIBERGER, M. I.; JEMTH, P.; FERREIRO, D. U.; WOLYNES, P. G.; FUXREITER, M. Fuzziness and frustration in the energy landscape of protein folding, function, and assembly. **Accounts of Chemical Research**, v. 54, n. 5, p. 1251–1259, 2021.
- 16 MINER, J. C.; CHEN, A. A.; GARCÍA, A. E. Free-energy landscape of a hyperstable RNA tetraloop. **Proceedings of the National Academy of Sciences of the United States of America**, v. 113, p. 6665–6670, 2016.
- 17 CHEBARO, Y.; BALLARD, A. J.; CHAKRABORTY, D.; WALES, D. J. Intrinsically disordered energy landscapes. **Scientific Reports**, v. 5, n. 1, p. 10386, 2015.
- 18 MOLLER, J.; LEQUIEU, J.; DE PABLO, J. J. The Free Energy Landscape of Internucleosome Interactions and Its Relation to Chromatin Fiber Structure. **ACS Central Science**, v. 5, n. 2, p. 341–348, 2019.
- 19 CHOI, U.; SANABRIA, H.; SMIRNOVA, T.; BOWEN, M.; WENINGER, K. Spontaneous Switching among Conformational Ensembles in Intrinsically Disordered Proteins. **Biomolecules**, v. 9, n. 3, p. 114, 2019.
- 20 TINOCO, I.; BUSTAMANTE, C. How RNA folds. **Journal of Molecular Biology**, v. 293, n. 2, p. 271–281, 1999.
- 21 DOBSON, C. M. Protein folding and misfolding. **Nature**, v. 426, n. 6968, p. 884–890, 2003.
- 22 HARTL, F. U. Protein misfolding diseases. **Annual Review of Biochemistry**, v. 86, p. 21–26, 2017.
- 23 CHO, S. S.; LEVY, Y.; WOLYNES, P. G. P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. **Proceedings of the National Academy of Sciences of the United States of America**, v. 103, n. 3, p. 586–591, 2006.
- 24 BEST, R. B.; HUMMER, G.; EATON, W. A. Native contacts determine protein folding mechanisms in atomistic simulations. **Proceedings of the National Academy of Sciences of the United States of America**, v. 110, n. 44, p. 17874–17879, 2013.
- 25 DILL, K. A.; BROMBERG, S.; YUE, K.; CHAN, H. S.; FTEBIG, K. M.; YEE, D. P.; THOMAS, P. D. Principles of protein folding — a perspective from simple exact models. **Protein Science**, v. 4, n. 4, p. 561–602, 1995.
- 26 NONATO, L. G.; AUPETIT, M. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. **IEEE Transactions on Visualization and Computer Graphics**, v. 25, n. 8, p. 2650–2673, 2019.
- 27 ORTIGOSSA, E. S.; DIAS, F. F.; NASCIMENTO, D. C. do. Getting over high-dimensionality: How multidimensional projection methods can assist data science. **Applied Sciences**, v. 12, n. 13, p. 6799, 2022.
- 28 JOLLIFFE, I. T. **Principal Component Analysis**. 2. ed. New York: Springer, 2002. (Springer Series in Statistics).

- 
- 29 GREENACRE, M.; GROENEN, P. J. F.; HASTIE, T.; D'ENZA, A. I.; MARKOS, A.; TUZHILINA, E. Principal component analysis. **Nature Reviews Methods Primers**, v. 2, n. 1, p. 100, 2022.
- 30 PALMA, J.; PIERDOMINICI-SOTTILE, G. On the Uses of PCA to Characterise Molecular Dynamics Simulations of Biological Macromolecules: Basics and Tips for an Effective Use. **ChemPhysChem**, v. 24, n. 2, 2023.
- 31 GARCÍA, A. E. Large-amplitude nonlinear motions in proteins. **Physical Review Letters**, v. 68, p. 2696–2699, 1992.
- 32 AMADEI, A.; LINSSEN, A. B. M.; BERENDSEN, H. J. C. Essential dynamics of proteins. **Proteins: Structure, Function, and Bioinformatics**, v. 17, n. 4, p. 412–425, 1993.
- 33 LEVY, Y.; JORTNER, J.; BECKER, O. M. Solvent effects on the energy landscapes and folding kinetics of polyalanine. **Proceedings of the National Academy of Sciences**, v. 98, n. 5, p. 2188–2193, 2001.
- 34 ALTIS, A.; NGUYEN, P. H.; HEGGER, R.; STOCK, G. Dihedral angle principal component analysis of molecular dynamics simulations. **The Journal of Chemical Physics**, v. 126, n. 24, p. 244111, 2007.
- 35 JAIN, A.; STOCK, G. Identifying metastable states of folding proteins. **Journal of Chemical Theory and Computation**, v. 8, n. 10, p. 3810–3819, 2012.
- 36 PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.
- 37 TORGERSON, W. S. Multidimensional scaling: I. Theory and method. **Psychometrika**, v. 17, n. 4, p. 401–419, 1952.
- 38 BORG, I.; GROENEN, P. J. F. **Modern Multidimensional Scaling: Theory and Applications**. 2. ed. New York: Springer, 2005.
- 39 MEAD, A. Review of the development of multidimensional scaling methods. **The Statistician**, v. 41, n. 1, p. 27, 1992.
- 40 FRANCE, S. L.; CARROLL, J. D. Two-Way Multidimensional Scaling: A Review. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, v. 41, n. 5, p. 644–661, 2011.
- 41 GLIELMO, A.; HUSIC, B. E.; RODRIGUEZ, A.; CLEMENTI, C.; NOÉ, F.; LAIO, A. Unsupervised learning methods for molecular simulation data. **Chemical Reviews**, v. 121, n. 16, p. 9722–9758, 2021.
- 42 WALLIN, S.; FARWER, J.; BASTOLLA, U. Testing similarity measures with continuous and discrete protein models. **Proteins**, v. 50, n. 1, p. 144–157, 2003.
- 43 COSSIO, P.; LAIO, A.; PIETRUCCHI, F. Which similarity measure is better for analyzing protein structures in a molecular dynamics trajectory? **Physical Chemistry Chemical Physics**, v. 13, n. 22, p. 10421, 2011.

- 
- 44 FEHER, M.; SCHMIDT, J. M. Metric and multidimensional scaling: Efficient tools for clustering molecular conformations. **Journal of Chemical Information and Computer Sciences**, v. 41, n. 2, p. 346–353, 2001.
- 45 ANDRECUT, M. Molecular dynamics multidimensional scaling. **Physics Letters A**, v. 373, n. 23, p. 2001–2006, 2009.
- 46 PISANI, P.; CAPORUSCIO, F.; CARLINO, L.; RASTELLI, G. Molecular dynamics simulations and classical multidimensional scaling unveil new metastable states in the conformational landscape of CDK2. **PLOS ONE**, v. 11, n. 4, p. e0154066, 2016.
- 47 SRIVASTAVA, A.; BALA, S.; MOTOMURA, H.; KOHDA, D.; TAMA, F.; MIYASHITA, O. Conformational ensemble of an intrinsically flexible loop in mitochondrial import protein Tim21 studied by modeling and molecular dynamics simulations. **Biochimica et Biophysica Acta (BBA)**, v. 1864, n. 2, p. 129417, 2020.
- 48 TENENBAUM, J. B.; SILVA, V. de; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. **Science**, v. 290, n. 5500, p. 2319–2323, 2000.
- 49 COIFMAN, R. R.; LAFON, S.; LEE, A. B.; MAGGIONI, M.; NADLER, B.; WARNER, F.; ZUCKER, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 21, p. 7426–7431, 2005.
- 50 MAATEN, L. Van der; HINTON, G. Visualizing data using t-SNE. **Journal of machine learning research**, v. 9, n. 11, 2008.
- 51 MCINNES, L.; HEALY, J.; SAUL, N.; GROßBERGER, L. UMAP: Uniform manifold approximation and projection. **Journal of Open Source Software**, v. 3, n. 29, p. 861, 2018.
- 52 CERIOTTI, M.; TRIBELLO, G. A.; PARRINELLO, M. Simplifying the representation of complex free-energy landscapes using sketch-map. **Proceedings of the National Academy of Sciences of the United States of America**, v. 108, n. 32, p. 13023–13028, 2011.
- 53 LEMKE, T.; PETER, C. EncoderMap: Dimensionality reduction and generation of molecule conformations. **Journal of Chemical Theory and Computation**, v. 15, n. 2, p. 1209–1215, 2019.
- 54 TROZZI, F.; WANG, X.; TAO, P. UMAP as a dimensionality reduction tool for molecular dynamics simulations of biomacromolecules: A comparison study. **The Journal of Physical Chemistry B**, v. 125, n. 19, p. 5022–5034, 2021.
- 55 OLIVEIRA, A. B.; FATORE, F. M.; PAULOVICH, F. V.; OLIVEIRA, O. N.; LEITE, V. B. P. Visualization of protein folding funnels in lattice models. **PLoS ONE**, v. 9, n. 7, p. e100861, 2014.
- 56 OLIVEIRA, A. B.; YANG, H.; WHITFORD, P. C.; LEITE, V. B. P. Distinguishing biomolecular pathways and metastable states. **Journal of Chemical Theory and Computation**, v. 15, n. 11, p. 6482–6490, 2019.

- 57 HARDIN, C.; EASTWOOD, M. P.; PRENTISS, M. C.; LUTHEY-SCHULTEN, Z.; WOLYNES, P. G. Associative memory Hamiltonians for structure prediction without homology: alpha/beta proteins. **Proceedings of the National Academy of Sciences of the United States of America**, v. 100, n. 4, p. 1679–1684, 2003.
- 58 TEJADA, E.; MINGHIM, R.; NONATO, L. G. On improved projection techniques to support visual exploration of multi-dimensional data sets. **Information Visualization**, v. 2, n. 4, p. 218–231, 2003.
- 59 JUNIOR, A. B. O.; LIN, X.; KULKARNI, P.; ONUCHIC, J. N.; ROY, S.; LEITE, V. B. Exploring energy landscapes of intrinsically disordered proteins: Insights into functional mechanisms. **Journal of Chemical Theory and Computation**, v. 17, n. 5, p. 3178–3187, 2021.
- 60 SANCHES, M. N.; KNAPP, K.; OLIVEIRA, A. B.; WOLYNES, P. G.; ONUCHIC, J. N.; LEITE, V. B. P. Examining the ensembles of amyloid- $\beta$  monomer variants and their propensities to form fibers using an energy landscape visualization method. **The Journal of Physical Chemistry B**, v. 126, n. 1, p. 93–99, 2022.
- 61 VIEGAS, R. G.; MARTINS, I. B. S.; SANCHES, M. N.; JUNIOR, A. B. O.; CAMARGO, J. B. d.; PAULOVICH, F. V.; LEITE, V. B. P. ELViM: Exploring Biomolecular Energy Landscapes through Multidimensional Visualization. **Journal of Chemical Information and Modeling**, v. 64, n. 8, p. 3443–3450, 2024a.
- 62 VIEGAS, R. G.; SANCHES, M. N.; CHEN, A. A.; PAULOVICH, F. V.; GARCIA, A. E.; LEITE, V. B. P. Characterizing the folding transition-state ensembles in the energy landscape of an RNA tetraloop. **Journal of Chemical Information and Modeling**, v. 63, n. 17, p. 5641–5649, 2023.
- 63 VIEGAS, R. G.; MARTINS, I. B. S.; LEITE, V. B. P. Understanding the energy landscape of intrinsically disordered protein ensembles. **Journal of Chemical Information and Modeling**, v. 64, n. 10, p. 4149–4157, 2024b.
- 64 MARTINS, I. B. S.; VIEGAS, R. G.; SANCHES, M. N.; ARAUJO, A. S. de; LEITE, V. B. P. Probing mastoparan-like antimicrobial peptides interaction with model membrane through energy landscape analysis. **The Journal of Physical Chemistry B**, v. 128, n. 1, p. 163–171, 2024.
- 65 SANCHES, M. N.; PARRA, R. G.; VIEGAS, R. G.; OLIVEIRA, A. B.; WOLYNES, P. G.; FERREIRO, D. U.; LEITE, V. B. Resolving the fine structure in the energy landscapes of repeat proteins. **QRB Discovery**, v. 3, p. e7, 2022.
- 66 DA SILVA, F. B.; SIMIEN, J. M.; VIEGAS, R. G.; HAGLUND, E.; LEITE, V. B. P. Exploring the folding landscape of leptin: Insights into threading pathways. **Journal of Structural Biology**, v. 216, n. 1, p. 108054, 2024.
- 67 LINDSAY, R. J.; VIEGAS, R. G.; LEITE, V. B.; WIGGE, P. A.; HANSON, S. M. Molecular dynamics simulations illuminate the role of sequence context in the ELF3-PrD-based temperature sensing mechanism in plants. eLife Sciences Publications, Ltd, 2024.

- 
- 68 DROR, R. O.; DIRKS, R. M.; GROSSMAN, J.; XU, H.; SHAW, D. E. Biomolecular simulation: A computational microscope for molecular biology. **Annual Review of Biophysics**, v. 41, n. 1, p. 429–452, 2012.
- 69 LAZIM, R.; SUH, D.; CHOI, S. Advances in molecular dynamics simulations and enhanced sampling methods for the study of protein systems. **International Journal of Molecular Sciences**, v. 21, n. 17, p. 6339, 2020.
- 70 SINHA, S.; TAM, B.; WANG, S. M. Applications of molecular dynamics simulation in protein study. **Membranes**, v. 12, n. 9, p. 844, 2022.
- 71 WARD, A. B.; SALI, A.; WILSON, I. A. Integrative structural biology. **Science**, v. 339, n. 6122, p. 913–915, 2013.
- 72 CHO, S. S.; LEVY, Y.; WOLYNES, P. G. P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. **Proceedings of the National Academy of Sciences of the United States of America**, v. 103, n. 3, p. 586–591, 2006.
- 73 POTOYAN, D. A.; PAPOIAN, G. A. Regulation of the H4 tail binding and folding landscapes via Lys-16 acetylation. **Proceedings of the National Academy of Sciences of the United States of America**, v. 109, n. 44, p. 17857–17862, 2012.
- 74 PAULOVICH, F. V. **Mapeamento de dados multi-dimensionais - integrando mineração e visualização**. 2008. Tese (Doutorado) — Universidade de São Paulo, 2008.
- 75 ROGAL, J. Reaction coordinates in complex systems—a perspective. **The European Physical Journal B**, v. 94, n. 11, p. 223, 2021.
- 76 PETERS, B. Reaction coordinates and mechanistic hypothesis tests. **Annual Review of Physical Chemistry**, v. 67, p. 669–690, 2016.
- 77 HABIBI, M.; ROTTLE, J.; PLOTKIN, S. S. As simple as possible, but not simpler: Exploring the fidelity of coarse-grained protein models for simulated force spectroscopy. **PLOS Computational Biology**, Public Library of Science, v. 12, n. 11, p. 1–31, 11 2016.
- 78 LAM, S. K.; PITROU, A.; SEIBERT, S. Numba: a LLVM-based Python JIT compiler. *In: Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*. New York, NY, USA: Association for Computing Machinery, 2015.
- 79 MCGIBBON, R. T. *et al.* Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. **Biophysical Journal**, v. 109, n. 8, p. 1528–1532, 2015.
- 80 HUNTER, J. D. Matplotlib: A 2D graphics environment. **Computing in Science & Engineering**, v. 9, n. 3, p. 90–95, 2007.
- 81 SOUZA, B. M.; MENDES, M. A.; SANTOS, L. D.; MARQUES, M. R.; CESAR, L. M. M.; ALMEIDA, R. N. A.; PAGNOCCA, F. C.; KONNO, K.; PALMA, M. S. U. Structural and functional characterization of two novel peptide toxins isolated from the venom of the social wasp *Polybia paulista*. **Peptides**, p. 2157–2164, 2005.
- 82 HUMPHREY, W.; DALKE, A.; SCHULTEN, K. VMD: Visual molecular dynamics. **Journal of Molecular Graphics**, v. 14, n. 1, p. 33–38, 1996.

- 
- 83 DAVIDCHACK, R. L.; HANDEL, R.; TRETYAKOV, M. V. Langevin thermostat for rigid body dynamics. **The Journal of chemical physics**, v. 130, n. 23, p. 234101, 2009.
- 84 FELLER, S. E.; ZHANG, Y.; PASTOR, R. W.; BROOKS, B. R. Constant pressure molecular dynamics simulation: The Langevin piston method. **The Journal of chemical physics.**, v. 103, n. 11, p. 4613–4621, 1995.
- 85 HUANG, J.; MACKERELL, A. D. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. **Journal of Computational Chemistry**, v. 34, n. 25, p. 2135–2145, 2013.
- 86 PHILLIPS, J. C. *et al.* Scalable molecular dynamics with NAMD. **Journal of Computational Chemistry**, v. 26, n. 16, p. 1781–1802, 2005.
- 87 ELBER, R.; RUYMGAART, A. P.; HESS, B. SHAKE parallelization. **The European Physical Journal. Special topics.**, v. 200, n. 1, p. 211–223, 2011.
- 88 MIYAMOTO, S.; KOLLMAN, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. **Journal of computational chemistry**, v. 13, n. 8, p. 952–962, 1992.
- 89 DARDEN, T.; YORK, D.; PEDERSEN, L. Particle mesh Ewald: An N.log(N) method for Ewald sums in large systems. **The Journal of Chemical Physics**, v. 98, n. 12, p. 10089–10092, 1993.
- 90 PAGLIOSA, P.; PAULOVICH, F. V.; MINGHIM, R.; LEVKOWITZ, H.; NONATO, L. G. Projection inspector: Assessment and synthesis of multidimensional projections. **Neurocomputing.**, v. 150, p. 599–610, 2015.
- 91 ESPADOTO, M.; MARTINS, R. M.; KERREN, A.; HIRATA, N. S. T.; TELEA, A. C. Toward a quantitative survey of dimension reduction techniques. **IEEE Transactions on Visualization and Computer Graphics**, v. 27, n. 3, p. 2153–2173, 2021.
- 92 TRIBELLO, G. A.; GASPAROTTO, P. Using Dimensionality Reduction to Analyze Protein Trajectories. **Frontiers in Molecular Biosciences**, v. 6, p. 1–11, 2019.
- 93 QU, Z.; ADELSON, D. L. Evolutionary conservation and functional roles of ncRNA. **Frontiers in Genetics**, v. 3, 2012.
- 94 WESTHOF, E.; FRITSCH, V. Rna folding: beyond watson–crick pairs. **Structure**, v. 8, n. 3, p. R55–R65, 2000.
- 95 WYATT, J. R.; PUGLISI, J. D.; TINOCO, I. RNA pseudoknots: Stability and loop size requirements. **Journal of Molecular Biology**, v. 214, n. 2, p. 455–470, 1990.
- 96 ŠPONER, J. *et al.* RNA structural dynamics as captured by molecular simulations: A comprehensive overview. **Chemical Reviews**, v. 118, p. 4177–4338, 2018.
- 97 NISSEN, P.; IPPOLITO, J. A.; BAN, N.; MOORE, P. B.; STEITZ, T. A. RNA tertiary interactions in the large ribosomal subunit: The a-minor motif. **Proceedings of the National Academy of Sciences**, v. 98, n. 9, p. 4899–4903, 2001.
- 98 BAULIN, E. F. Features and Functions of the A-Minor Motif, the Most Common Motif in RNA Structure. **Biochemistry (Moscow)**, v. 86, n. 8, p. 952–961, 2021.

- 
- 99 WOESE, C. R.; WINKER, S.; GUTTEL, R. R. Architecture of ribosomal RNA: Constraints on the sequence of tetra-loops. **Proceedings of the National Academy of Sciences of the United States of America**, v. 87, n. 21, p. 8467–8471, 1990.
- 100 PESELIS, A.; SERGANOV, A. Themes and variations in riboswitch structure and function. **Biochimica et Biophysica Acta - Gene Regulatory Mechanisms**, v. 1839, n. 10, p. 908–918, 2014.
- 101 UHLENBECK, O. C. Tetraloops and RNA folding. **Nature**, v. 346, n. 6285, p. 613–614, 1990.
- 102 FIORE, J. L.; NESBITT, D. J. An RNA folding motif: GNRA tetraloop–receptor interactions. **Quarterly Reviews of Biophysics**, v. 46, n. 3, p. 223–264, 2013.
- 103 SHAW, D. E. *et al.* Millisecond-scale molecular dynamics simulations on Anton. *In: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*. New York, NY, USA: Association for Computing Machinery, 2009.
- 104 CHEN, A. A.; GARCÍA, A. E. High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. **Proceedings of the National Academy of Sciences of the United States of America**, v. 110, n. 42, p. 16820–16825, 2013.
- 105 CORNELL, W. D. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. **Journal of the American Chemical Society**, v. 117, n. 19, p. 5179–5197, 1995.
- 106 NOSE, S. A molecular dynamics method for simulations in the canonical ensemble. **Molecular Physics**, v. 52, n. 2, p. 255–268, 1984.
- 107 HOOVER, W. G. Canonical dynamics: Equilibrium phase-space distributions. **Physical review A**, v. 31, p. 1695–1697, 1985.
- 108 BOTTARO, S.; BUSSI, G.; PINAMONTI, G.; REIßER, S.; BOOMSMA, W.; LINDORFF-LARSEN, K. Barnaba: software for analysis of nucleic acid structures and trajectories. **RNA**, v. 25, n. 2, p. 219–231, 2019.
- 109 ZERZE, G.; PIAGGI, P.; DEBENEDETTI, P. A computational study of RNA tetraloop thermodynamics, including misfolded states. **The Journal of Physical Chemistry B**, v. 125, p. 13685–13695, 2021.
- 110 CORBINO, K. A.; BARRICK, J. E.; LIM, J.; WELZ, R.; TUCKER, B. J.; PUSKARZ, I.; MANDAL, M.; RUDNICK, N. D.; BREAKER, R. R. Evidence for a second class of s-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. **Genome biology**, v. 6, p. 1–10, 2005.
- 111 GILBERT, S. D.; RAMBO, R. P.; TYNE, D. V.; BATEY, R. T. Structure of the SAM-II riboswitch bound to S-adenosylmethionine. **Nature Structure & Molecular Biology**, v. 15, n. 2, p. 177–182, 2008.
- 112 PRICE, I. R.; GRIGG, J. C.; KE, A. Common themes and differences in SAM recognition among SAM riboswitches. **Biochimica et Biophysica Acta - Gene Regulatory Mechanisms**, v. 1839, n. 10, p. 931–938, 2014.

- 
- 113 HALLER, A.; RIEDER, U.; AIGNER, M.; BLANCHARD, S. C.; MICURA, R. Conformational capture of the SAM-II riboswitch. **Nature Chemical Biology**, v. 7, n. 6, p. 393–400, 2011.
- 114 ROY, S.; LAMMERT, H.; HAYES, R. L.; CHEN, B.; LEBLANC, R.; DAYIE, T. K.; ONUCHIC, J. N.; SANBONMATSU, K. Y. A magnesium-induced triplex pre-organizes the SAM-II riboswitch. **PLOS Computational Biology**, v. 13, n. 3, p. 1–19, 2017.
- 115 CHEN, B.; ZUO, X.; WANG, Y.-X.; DAYIE, T. K. Multiple conformations of SAM-II riboswitch detected with SAXS and NMR spectroscopy. **Nucleic Acids Research**, v. 40, n. 7, p. 3117–3130, 2011.
- 116 HAMMOND, M. C. A tale of two riboswitches. **Nature Chemical Biology**, v. 7, n. 6, p. 342–343, 2011.
- 117 CHEN, B.; ZUO, X.; WANG, Y.-X.; DAYIE, T. K. Multiple conformations of SAM-II riboswitch detected with SAXS and NMR spectroscopy. **Nucleic Acids Research**, v. 40, n. 7, p. 3117–3130, 2012.
- 118 WHITFORD, P. C.; NOEL, J. K.; GOSAVI, S.; SCHUG, A.; SANBONMATSU, K. Y.; ONUCHIC, J. N. An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. **Proteins: Structure, Function, and Bioinformatics**, v. 75, n. 2, p. 430–441, 2009.
- 119 NOEL, J. K.; LEVI, M.; RAGHUNATHAN, M.; LAMMERT, H.; HAYES, R. L.; ONUCHIC, J. N.; WHITFORD, P. C. Smog 2: A versatile software package for generating structure-based models. **PLoS Computational Biology**, v. 12, n. 3, p. 1–14, 2016.
- 120 HAYES, R. L.; NOEL, J. K.; MANDIC, A.; WHITFORD, P. C.; SANBONMATSU, K. Y.; MOHANTY, U.; ONUCHIC, J. N. Generalized manning condensation model captures the RNA ion atmosphere. **Physical Review Letters**, v. 114, n. 25, p. 258105, 2015.
- 121 RAMACHANDRAN, V.; MAINAN, A.; ROY, S. Dynamic effects of the spine of hydrated magnesium on viral RNA pseudoknot structure. **Physical Chemistry Chemical Physics**, v. 24, n. 39, p. 24570–24581, 2022.
- 122 MAINAN, A.; ROY, S. Dynamic counterion condensation model decodes functional dynamics of RNA pseudoknot in SARS-CoV-2: Control of ion-mediated pierced lasso topology. **The Journal of Physical Chemistry Letters**, v. 14, p. 10402–10411, 2023.
- 123 ROY, S.; HENNELLY, S. P.; LAMMERT, H.; ONUCHIC, J. N.; SANBONMATSU, K. Y. Magnesium controls aptamer-expression platform switching in the SAM-I riboswitch. **Nucleic Acids Research**, v. 47, n. 6, p. 3158–3170, 2019.
- 124 SHIH, A. Y.; ARKHIPOV, A.; FREDDOLINO, P. L.; SCHULTEN, K. Coarse grained protein-lipid model with application to lipoprotein particles. **The Journal of Physical Chemistry B**, v. 110, n. 8, p. 3674–3684, 2006.
- 125 CHEN, B.; LEBLANC, R.; DAYIE, T. K. SAM-II Riboswitch Samples at least Two Conformations in Solution in the Absence of Ligand: Implications for Recognition. **Angewandte Chemie**, v. 128, n. 8, p. 2774–2777, 2016.

- 
- 126 XUE, X.; YONGJUN, W.; ZHIHONG, L. Folding of SAM-II riboswitch explored by replica-exchange molecular dynamics simulation. **Journal of Theoretical Biology**, v. 365, p. 265–269, 2015.
- 127 SARKAR, R.; JAISWAR, A.; HENNELLY, S. P.; ONUCHIC, J. N.; SANBONMATSU, K. Y.; ROY, S. Chelated magnesium logic gate regulates riboswitch pseudoknot formation. **The Journal of Physical Chemistry B**, v. 125, n. 24, p. 6479–6490, 2021.
- 128 SARKAR, R.; MAINAN, A.; ROY, S. Influence of ion and hydration atmospheres on RNA structure and dynamics: insights from advanced theoretical and computational methods. **Chemical Communications**, v. 60, p. 3624–3644, 2024.
- 129 WRIGHT, P. E.; DYSON, H. J. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. **Journal of Molecular Biology**, v. 293, n. 2, p. 321–331, 1999.
- 130 DUNKER, A. K. *et al.* Intrinsically disordered protein. **Journal of Molecular Graphics and Modelling**, v. 19, n. 1, p. 26–59, 2001.
- 131 UVERSKY, V. N.; KULKARNI, P. Intrinsically disordered proteins: Chronology of a discovery. **Biophysical Chemistry**, v. 279, p. 106694, 2021.
- 132 KULKARNI, P. *et al.* Intrinsically disordered proteins: Critical components of the wetware. **Chemical Reviews**, v. 122, n. 6, p. 6614–6633, 2022.
- 133 TANTOS, A.; HAN, K.-H.; TOMPA, P. Intrinsic disorder in cell signaling and gene transcription. **Molecular and Cellular Endocrinology**, v. 348, n. 2, p. 457–465, 2012.
- 134 WRIGHT, P. E.; DYSON, H. J. Intrinsically disordered proteins in cellular signalling and regulation. **Nature Reviews Molecular Cell Biology**, v. 16, n. 1, p. 18–29, 2015.
- 135 BONDOS, S. E.; DUNKER, A. K.; UVERSKY, V. N. Intrinsically disordered proteins play diverse roles in cell signaling. **Cell Communication and Signaling**, v. 20, n. 1, p. 20, 2022.
- 136 SANTOFIMIA-CASTAÑO, P.; RIZZUTI, B.; XIA, Y.; ABIAN, O.; PENG, L.; VELÁZQUEZ-CAMPOY, A.; NEIRA, J. L.; IOVANNA, J. Targeting intrinsically disordered proteins involved in cancer. **Cellular and Molecular Life Sciences**, v. 77, n. 9, p. 1695–1707, 2020.
- 137 COSKUNER-WEBER, O.; MIRZANLI, O.; UVERSKY, V. N. Intrinsically disordered proteins and proteins with intrinsically disordered regions in neurodegenerative diseases. **Biophysical Reviews**, v. 14, n. 3, p. 679–707, 2022.
- 138 IAKOUCHEVA, L. M.; BROWN, C. J.; LAWSON, J.; OBRADOVIĆ, Z.; DUNKER, A. Intrinsic disorder in cell-signaling and cancer-associated proteins. **Journal of Molecular Biology**, v. 323, n. 3, p. 573–584, 2002.
- 139 KULKARNI, P.; UVERSKY, V. N. Intrinsically disordered proteins in chronic diseases. **Biomolecules**, v. 9, n. 4, p. 147, 2019.
- 140 KULKARNI, P. *et al.* Intrinsically disordered proteins: Ensembles at the limits of Anfinsen's dogma. **Biophysics Reviews**, v. 3, n. 1, p. 011306, 2022.

- 
- 141 JUNIOR, A. B. O.; LIN, X.; KULKARNI, P.; ONUCHIC, J. N.; ROY, S.; LEITE, V. B. P. Exploring energy landscapes of intrinsically disordered proteins: Insights into functional mechanisms. **Journal of Chemical Theory and Computation**, v. 17, n. 5, p. 3178–3187, 2021.
- 142 EVANS, R.; RAMISETTY, S.; KULKARNI, P.; WENINGER, K. Illuminating intrinsically disordered proteins with integrative structural biology. **Biomolecules**, v. 13, n. 1, p. 124, 2023.
- 143 BONOMI, M.; HELLER, G. T.; CAMILLONI, C.; VENDRUSCOLO, M. Principles of protein structural ensemble determination. **Current Opinion in Structural Biology**, v. 42, p. 106–116, 2017.
- 144 LAZAR, T. *et al.* PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. **Nucleic Acids Research**, v. 49, n. D1, p. D404–D411, 2020.
- 145 PEPENELLA, S.; MURPHY, K. J.; HAYES, J. J. Intra- and inter-nucleosome interactions of the core histone tail domains in higher-order chromatin structure. **Chromosoma**, v. 123, p. 3 – 13, 2013.
- 146 CUTTER, A. R.; HAYES, J. J. A brief review of nucleosome structure. **FEBS Letters**, v. 589, n. 20, Part A, p. 2914–2922, 2015.
- 147 SHAYTAN, A. K.; ARMEEV, G. A.; GONCEARENCO, A.; ZHURKIN, V. B.; LANDSMAN, D.; PANCHENKO, A. R. Coupling between Histone Conformations and DNA Geometry in Nucleosomes on a Microsecond Timescale: Atomistic Insights into Nucleosome Functions. **Journal of Molecular Biology**, v. 428, n. 1, p. 221–237, 2016.
- 148 ZHOU, K.; GAULLIER, G.; LUGER, K. Nucleosome structure and dynamics are coming of age. **Nature Structural & Molecular Biology**, v. 26, n. 1, p. 3–13, 2019.
- 149 WEBB, B.; SALI, A. Comparative protein structure modeling using MODELLER. **Current Protocols in Bioinformatics**, v. 54, n. 1, p. 5–6, 2016.
- 150 POTOYAN, D. A.; PAPOIAN, G. A. Energy landscape analyses of disordered histone tails reveal special organization of their conformational dynamics. **Journal of the American Chemical Society**, v. 133, n. 19, p. 7405–7415, 2011.
- 151 RÖDER, K. Is the H4 histone tail intrinsically disordered or intrinsically multifunctional? **Physical Chemistry Chemical Physics**, v. 23, n. 9, p. 5134–5142, 2021.
- 152 WALES, D. J. Discrete path sampling. **Molecular Physics**, v. 100, n. 20, p. 3285–3305, 2002.
- 153 NG, M. K.; CHEUNG, P. A brief histone in time: understanding the combinatorial functions of histone PTMs in the nucleosome context. **Biochemistry and Cell Biology**, v. 94, n. 1, p. 33–42, 2016.
- 154 ANDRÉS, M.; GARCÍA-GOMIS, D.; PONTE, I.; SUAÚ, P.; ROQUE, A. Histone H1 Post-Translational Modifications: Update and Future Perspectives. **International Journal of Molecular Sciences**, v. 21, n. 16, 2020.

- 155 LIU, R. *et al.* Post-translational modifications of histones: Mechanisms, biological functions, and therapeutic targets. **MedComm**, v. 4, n. 3, p. e292, 2023.
- 156 COLLEPARDO-GUEVARA, R.; PORTELLA, G.; VENDRUSCOLO, M.; FRENKEL, D.; SCHLICK, T.; OROZCO, M. Chromatin unfolding by epigenetic modifications explained by dramatic impairment of internucleosome interactions: A multiscale computational study. **Journal of the American Chemical Society**, v. 137, n. 32, p. 10205–10215, 2015.
- 157 WINOGRADOFF, D.; ECHEVERRIA, I.; POTOYAN, D. A.; PAPOIAN, G. A. The Acetylation Landscape of the H4 Histone Tail: Disentangling the Interplay between the Specific and Cumulative Effects. **Journal of the American Chemical Society**, v. 137, n. 19, p. 6245–6253, 2015.
- 158 SHABANE, P. S.; ONUFRIEV, A. V. Significant compaction of H4 histone tail upon charge neutralization by acetylation and its mimics, possible effects on chromatin structure. **Journal of Molecular Biology**, v. 433, n. 6, p. 166683, 2021.
- 159 BEST, R. B.; HUMMER, G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. **Journal of Physical Chemistry B**, v. 113, n. 26, p. 9004–9015, 2009.
- 160 CORNELL, W. D. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. **Journal of the American Chemical Society**, v. 117, n. 19, p. 5179–5197, 1995.
- 161 JORGENSEN, W. L.; CHANDRASEKHAR, J.; MADURA, J. D.; IMPEY, R. W.; KLEIN, M. L. Comparison of simple potential functions for simulating liquid water. **Journal of Chemical Physics**, v. 79, n. 2, p. 926–935, 1983.
- 162 SUGITA, Y.; OKAMOTO, Y. Replica-exchange molecular dynamics method for protein folding. **Chemical Physics Letters**, v. 314, n. 1-2, p. 141–151, 1999.
- 163 MU, Y.; NGUYEN, P. H.; STOCK, G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. **Proteins: Structure, Function, and Bioinformatics**, v. 58, n. 1, p. 45–52, 2005.
- 164 APPADURAI, R.; KONERU, J. K.; BONOMI, M.; ROBUSTELLI, P.; SRIVASTAVA, A. Clustering Heterogeneous Conformational Ensembles of Intrinsically Disordered Proteins with t-Distributed Stochastic Neighbor Embedding. **Journal of Chemical Theory and Computation**, v. 19, n. 14, p. 4711–4727, 2023.
- 165 SAHA, A.; DALAL, Y. A glitch in the snitch: the role of linker histone H1 in shaping the epigenome in normal and diseased cells. **Open Biology**, v. 11, n. 8, p. 210124, 2021.
- 166 SHUKLA, S.; AGARWAL, P.; KUMAR, A. Disordered regions tune order in chromatin organization and function. **Biophysical Chemistry**, v. 281, p. 106716, 2022.
- 167 HARTMAN, P. G.; CHAPMAN, G. E.; MOSS, T.; BRADBURY, E. M. Studies on the role and mode of operation of the very-lysine-rich histone h1 in eukaryote chromatin. **European Journal of Biochemistry**, v. 77, n. 1, p. 45–51, 1977.

- 168 WHITE, A. E.; HIEB, A. R.; LUGER, K. A quantitative investigation of linker histone interactions with nucleosomes and chromatin. **Scientific Reports**, v. 6, n. 1, p. 19122, 2016.
- 169 HENDZEL, M. J.; LEVER, M. A.; CRAWFORD, E.; TH'NG, J. P. The C-terminal Domain Is the Primary Determinant of Histone H1 Binding to Chromatin in Vivo. **Journal of Biological Chemistry**, v. 279, n. 19, p. 20028–20034, 2004.
- 170 WU, H.; DALAL, Y.; PAPOIAN, G. A. Binding Dynamics of Disordered Linker Histone H1 with a Nucleosomal Particle. **Journal of Molecular Biology**, v. 433, n. 6, p. 166881, 2021.
- 171 ÖZTÜRK, M. A.; COJOCARU, V.; WADE, R. C. Toward an ensemble view of chromatosome structure: a paradigm shift from one to many. **Structure**, v. 26, n. 8, p. 1050–1057, 2018.
- 172 HAO, F.; KALE, S.; DIMITROV, S.; HAYES, J. J. Unraveling linker histone interactions in nucleosomes. **Current Opinion in Structural Biology**, v. 71, p. 87–93, 2021.
- 173 ZHOU, B.-R.; JIANG, J.; FENG, H.; GHIRLANDO, R.; XIAO, T. S.; BAI, Y. Structural mechanisms of nucleosome recognition by linker histones. **Molecular Cell**, v. 59, n. 4, p. 628–638, 2015.
- 174 DAVTYAN, A.; SCHAFER, N. P.; ZHENG, W.; CLEMENTI, C.; WOLYNES, P. G.; PAPOIAN, G. A. AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. **The Journal of Physical Chemistry B**, v. 116, n. 29, p. 8494–8503, 2012.
- 175 WU, H.; WOLYNES, P. G.; PAPOIAN, G. A. AWSEM-IDP: a coarse-grained force field for intrinsically disordered proteins. **The Journal of Physical Chemistry B**, v. 122, n. 49, p. 11115–11125, 2018.
- 176 HINCKLEY, D. M.; FREEMAN, G. S.; WHITMER, J. K.; PABLO, J. J. D. An experimentally-informed coarse-grained 3-site-per-nucleotide model of DNA: Structure, thermodynamics, and dynamics of hybridization. **The Journal of Chemical Physics**, v. 139, n. 14, p. 144903, 2013.
- 177 BEDNAR, J. *et al.* Structure and Dynamics of a 197 bp Nucleosome in Complex with Linker Histone H1. **Molecular Cell**, v. 66, n. 3, p. 384–397, 2017.
- 178 CHONG, S.-H.; CHATTERJEE, P.; HAM, S. Computer simulations of intrinsically disordered proteins. **Annual Review of Physical Chemistry**, v. 68, n. 1, p. 117–134, 2017.
- 179 MU, J.; LIU, H.; ZHANG, J.; LUO, R.; CHEN, H.-F. Recent force field strategies for intrinsically disordered proteins. **Journal of Chemical Information and Modeling**, v. 61, n. 3, p. 1037–1047, 2021.
- 180 FAGERBERG, E.; SKEPÖ, M. Comparative performance of computer simulation models of intrinsically disordered proteins at different levels of coarse-graining. **Journal of Chemical Information and Modeling**, v. 63, n. 13, p. 4079–4087, 2023.

- 181 KASAHARA, K.; TERAZAWA, H.; TAKAHASHI, T.; HIGO, J. Studies on molecular dynamics of intrinsically disordered proteins and their fuzzy complexes: A mini-review. **Computational and Structural Biotechnology Journal**, v. 17, p. 712–720, 2019.
- 182 CHEN, J.; LIU, X.; CHEN, J. Targeting intrinsically disordered proteins through dynamic interactions. **Biomolecules**, v. 10, n. 5, p. 743, 2020.
- 183 WANG, W. Recent advances in atomic molecular dynamics simulation of intrinsically disordered proteins. **Physical Chemistry Chemical Physics**, v. 23, n. 2, p. 777–784, 2021.
- 184 GONG, X.; ZHANG, Y.; CHEN, J. Advanced sampling methods for multiscale simulation of disordered proteins and dynamic interactions. **Biomolecules**, v. 11, n. 10, p. 1416, 2021.
- 185 NAG, N.; CHETRI, P. B.; UVERSKY, V. N.; GIRI, R.; TRIPATHI, T. Experimental methods to study intrinsically disordered proteins. *In*: TRIPATHI, T.; DUBEY, V. K. (ed.). **Advances in Protein Molecular and Structural Biology Methods**. [*S.l.: s.n.*]: Academic Press, 2022. p. 505–533.
- 186 RAVERA, E.; SGHERI, L.; PARIGI, G.; LUCHINAT, C. A critical assessment of methods to recover information from averaged data. **Physical Chemistry Chemical Physics**, v. 18, p. 5686–5701, 2016.
- 187 PIETREK, L. M.; STELZL, L. S.; HUMMER, G. Structural ensembles of disordered proteins from hierarchical chain growth and simulation. **Current Opinion in Structural Biology**, v. 78, p. 102501, 2023.
- 188 GHAFOURI, H.; LAZAR, T.; DEL CONTE, A.; TENORIO KU, L. G.; CONSORTIUM, P.; TOMPA, P.; TOSATTO, S. C. E.; MONZON, A. M. PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins. **Nucleic Acids Research**, v. 52, n. D1, p. D536–D544, 2023.
- 189 FUERTES, G. *et al.* Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in saxs vs. fret measurements. **Proceedings of the National Academy of Sciences of the United States of America**, v. 114, n. 31, p. E6342–E6351, 2017.
- 190 VITALIS, A.; PAPPU, R. V. ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions. **Journal of Computational Chemistry**, v. 30, n. 5, p. 673–699, 2009.
- 191 VITALIS, A.; PAPPU, R. V. Chapter 3 methods for monte carlo simulations of biomacromolecules. *In*: WHEELER, R. A. (ed.). [*S.l.: s.n.*]: Elsevier, 2009, (Annual Reports in Computational Chemistry, v. 5). p. 49–76.
- 192 LEUNG, H. T. A.; BIGNUCOLO, O.; AREGGER, R.; DAMES, S. A.; MAZUR, A.; BERNÈCHE, S.; GRZESIEK, S. A rigorous and efficient method to reweight very large conformational ensembles using average experimental data and to determine their relative information content. **Journal of Chemical Theory and Computation**, v. 12, n. 1, p. 383–394, 2016.

- 193 LINCOFF, J.; HAGHIGHATLARI, M.; KRZEMINSKI, M.; TEIXEIRA, J. M. C.; GOMES, G.-N. W.; GRADINARU, C. C.; FORMAN-KAY, J. D.; HEAD-GORDON, T. Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states. **Communications chemistry**, v. 3, p. 74, 2020.
- 194 FELDMAN, H. J.; HOGUE, C. W. A fast method to sample real protein conformational space. **PROTEINS: Structure, Function, and Bioinformatics**, v. 39, n. 2, p. 112–131, 2000.
- 195 FELDMAN, H. J.; HOGUE, C. W. Probabilistic sampling of protein conformations: New hope for brute force? **PROTEINS: Structure, Function, and Bioinformatics**, v. 46, n. 1, p. 8–23, 2002.
- 196 KRZEMINSKI, M.; MARSH, J. A.; NEALE, C.; CHOY, W.-Y.; FORMAN-KAY, J. D. Characterization of disordered proteins with ENSEMBLE. **Bioinformatics**, v. 29, n. 3, p. 398–399, 2012.
- 197 RÖDER, K.; WALES, D. J. The energy landscape perspective: Encoding structure and function for biomolecules. **Frontiers in Molecular Biosciences**, v. 9, 2022.
- 198 GONZÁLEZ-DELGADO, J.; SAGAR, A.; ZANON, C.; LINDORFF-LARSEN, K.; BERNADÓ, P.; NEUVIAL, P.; CORTÉS, J. WASCO: a wasserstein-based statistical tool to compare conformational ensembles of intrinsically disordered proteins. **Journal of Molecular Biology**, v. 435, n. 14, p. 168053, 2023.
- 199 LINDORFF-LARSEN, K.; FERKINGHOFF-BORG, J. Similarity measures for protein ensembles. **PLOS ONE**, v. 4, n. 1, p. 1–13, 2009.
- 200 LAZAR, T.; GUHARROY, M.; VRANKEN, W.; RAUSCHER, S.; WODAK, S. J.; TOMPA, P. Distance-based metrics for comparing conformational ensembles of intrinsically disordered proteins. **Biophysical Journal**, v. 118, n. 12, p. 2952–2965, 2020.
- 201 ARDEVOL, A.; PALAZZESI, F.; TRIBELLO, G. A.; PARRINELLO, M. General protein data bank-based collective variables for protein folding. **Journal of Chemical Theory and Computation**, v. 12, n. 1, p. 29–35, 2016.
- 202 CHEN, Y.; MANT, C. T.; FARMER, S. W.; HANCOCK, R. E.; VASIL, M. L.; HODGES, R. S. Rational design of  $\alpha$ -helical antimicrobial peptides with enhanced activities and specificity/therapeutic index. **Journal of Biological Chemistry**, v. 280, n. 13, p. 12316–12329, 2005.
- 203 WANG, K. rong; ZHANG, B. zhi; ZHANG, W.; YAN, J. xi; LI, J.; WANG, R. Antitumor effects, cell selectivity and structure–activity relationship of a novel antimicrobial peptide polybia-mpi. **Peptides**, v. 29, n. 6, p. 963–968, 2008.
- 204 WANG, K.; YAN, J.; ZHANG, B.; SONG, J.; JIA, p.; WANG, R. Novel mode of action of polybia-MPI, a novel antimicrobial peptide, in multi-drug resistant leukemic cells. **Cancer Letters**, v. 278, n. 1, p. 65–72, 2009.
- 205 CABRERA, M. P. dos S.; ARCISIO-MIRANDA, M.; GORJÃO, R.; LEITE, N. B.; SOUZA, B. M. de; CURI, R.; PROCOPIO, J.; NETO, J. R.; PALMA, M. S. Influence of the bilayer composition on the binding and membrane disrupting effect of Polybia-MP1,

- an antimicrobial mastoparan peptide with leukemic T-lymphocyte cell selectivity. **Biochemistry**, v. 51, n. 24, p. 4898–4908, 2012.
- 206 VOGT, T. C. B.; BECHINGER, B. The interactions of histidine-containing amphipathic helical peptide antibiotics with lipid bilayers. The effects of charges and pH. **Journal of Biological Chemistry**, v. 274, n. 41, p. 29115–29121, 1999.
- 207 MAKOVITZKI, A.; SHAI, Y. pH-dependent antifungal lipopeptides and their plausible mode of action. **Biochemistry**, v. 44, n. 28, p. 9775–9784, 2005.
- 208 CABRERA, M. P. dos S.; COSTA, S. T. B.; SOUZA, B. M. de; PALMA, M. S.; RUGGIERO, J. R.; NETO, J. R. Selectivity in the mechanism of action of antimicrobial mastoparan peptide polybia-mp1. **European Biophysics Journal**, v. 37, p. 879–891, 2008.
- 209 ALVARES, D. d. S.; MARTINS, I. B. S.; VIEGAS, T. G.; PALMA, M. S.; ARAUJO, A. S. de; CARVALHO, S. J. de; NETO, J. R. Modulatory effects of acidic pH and membrane potential on the adsorption of pH-sensitive peptides to anionic lipid membrane. **Membranes**, v. 11, n. 5, p. 307, 2021.
- 210 MARTINS, I. B. S.; VIEGAS, T. G.; dos SANTOS ALVARES, D.; SOUZA, B. M. de; PALMA, M. S.; Ruggiero Neto, J.; ARAUJO, A. S. de. The effect of acidic pH on the adsorption and lytic activity of the peptides Polybia-MP1 and its histidine-containing analog in anionic lipid membrane: a biophysical study by molecular dynamics and spectroscopy. **Amino Acids**, v. 53, n. 5, p. 753–767, 2021.
- 211 JO, S.; LIM, J. B.; KLAUDA, J. B.; IM, W. CHARMM-GUI Membrane Builder for Mixed Bilayers and Its Application to Yeast Membranes. **Biophysical Journal**, v. 97, n. 1, p. 50–58, 2009.
- 212 JO, S.; KIM, T.; IYER, V. G.; IM, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. **Journal of Computational Chemistry**, v. 29, n. 11, p. 1859–1865, ago. 2008.
- 213 MARK, P.; NILSSON, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. **Journal of Physical Chemistry A**, v. 105, n. 43, p. 9954–9960, nov. 2001.
- 214 KUMAR, S.; ROSENBERG, J. M.; BOUZIDA, D.; SWENDSEN, R. H.; KOLLMAN, P. A. The weighted histogram analysis method for free energy calculations on biomolecules. I. The method. **Journal of Computational Chemistry**, v. 13, n. 8, p. 1011–1021, 1992.
- 215 HUANG, Y.; HUANG, J.; CHEN, Y. Alpha-helical cationic antimicrobial peptides: relationships of structure and function. **Protein Cell**, v. 1, p. 143–152, 2010.
- 216 AHMED, T. A.; HAMMAMI, R. Recent insights into structure–function relationships of antimicrobial peptides. **Journal of Food Biochemistry**, v. 43, n. 1, p. e12546, 2019.
- 217 XUAN, J.; FENG, W.; WANG, J.; WANG, R.; ZHANG, B.; BO, L.; CHEN, Z.-S.; YANG, H.; SUN, L. Antimicrobial peptides for combating drug-resistant bacterial infections. **Drug Resistance Updates**, v. 68, p. 100954, 2023.

- 218 BHATTACHARYA, S.; CHAKRABARTY, S. Mapping conformational landscape in protein folding: Benchmarking dimensionality reduction and clustering techniques on the trp-cage mini-protein. **Biophysical Chemistry**, v. 319, p. 107389, 2025.
- 219 KRUSKAL, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. **Psychometrika**, v. 29, n. 1, p. 1–27, 1964.
- 220 NIX, J.; SUSSMAN, D.; WILSON, C. The 1.3 Å crystal structure of a biotin-binding pseudoknot and the basis for RNA molecular recognition. **Journal of Molecular Biology**, v. 296, n. 5, p. 1235–1244, 2000.
- 221 GOMES, G.-N. W.; KRZEMINSKI, M.; NAMINI, A.; MARTIN, E. W.; MITTAG, T.; HEAD-GORDON, T.; FORMAN-KAY, J. D.; GRADINARU, C. C. Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. **Journal of the American Chemical Society**, v. 142, n. 37, p. 15697–15710, 2020.
- 222 ARRÍAS, P. N.; OSMANLI, Z.; PERALTA, E.; CHINESTRAD, P. M.; MONZON, A. M.; TOSATTO, S. C. E. Diversity and structural-functional insights of alpha-solenoid proteins. **Protein Science**, v. 33, n. 11, p. e5189, 2024.
- 223 BARRICK, D.; FERREIRO, D. U.; KOMIVES, E. A. Folding landscapes of ankyrin repeat proteins: experiments meet theory. **Current Opinion in Structural Biology**, v. 18, n. 1, p. 27–34, 2008.
- 224 MOSAVI, L. K.; CAMMETT, T. J.; DESROSIERS, D. C.; PENG, Z.-Y. The ankyrin repeat as molecular architecture for protein recognition. **Protein Science**, v. 13, n. 6, p. 1435–1448, 2004.
- 225 WERBECK, N. D.; ITZHAKI, L. S. Probing a moving target with a plastic unfolding intermediate of an ankyrin-repeat protein. **Proceedings of the National Academy of Sciences of the United States of America**, v. 104, n. 19, p. 7863–7868, 2007.
- 226 AKSEL, T.; BARRICK, D. Direct observation of parallel folding pathways revealed using a symmetric repeat protein system. **Biophysical Journal**, v. 107, n. 1, p. 220–232, 2014.
- 227 CLEMENTI, C. Coarse-grained models of protein folding: toy models or predictive tools? **Current Opinion in Structural Biology**, v. 18, n. 1, p. 10–15, 2008.
- 228 DAVTYAN, A.; SCHAFER, N. P.; ZHENG, W.; CLEMENTI, C.; WOLYNES, P. G.; PAPOIAN, G. A. AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. **The Journal of Physical Chemistry B**, v. 116, n. 29, p. 8494–8503, 2012.
- 229 YEATES, T. O.; NORCROSS, T. S.; KING, N. P. Knotted and topologically complex proteins as models for studying folding and stability. **Current Opinion in Chemical Biology**, v. 11, n. 6, p. 595–603, 2007.
- 230 FAÍSCA, P. F. Knotted proteins: A tangled tale of structural biology. **Computational and Structural Biotechnology Journal**, v. 13, p. 459–468, 2015.

- 
- 231 HAGLUND, E.; SULKOWSKA, J. I.; HE, Z.; FENG, G.-S.; JENNINGS, P. A.; ONUCHIC, J. N. The unique cysteine knot regulates the pleiotropic hormone leptin. **PLOS ONE**, v. 7, n. 9, p. 1–13, 09 2012.
- 232 HAGLUND, E.; SULKOWSKA, J. I.; NOEL, J. K.; LAMMERT, H.; ONUCHIC, J. N.; JENNINGS, P. A. Pierced lasso bundles are a new class of knot-like motifs. **PLoS Computational Biology**, v. 10, n. 6, p. 1–11, 06 2014.
- 233 SIMIEN, J. M.; HAGLUND, E. Topological twists in nature. **Trends in Biochemical Sciences**, v. 46, n. 6, p. 461–471, 2021.
- 234 SULKOWSKA, J. I.; SULKOWSKI, P.; ONUCHIC, J. Dodging the crisis of folding proteins with knots. **Proceedings of the National Academy of Sciences**, v. 106, n. 9, p. 3119–3124, 2009.
- 235 NOEL, J. K.; WHITFORD, P. C.; SANBONMATSU, K. Y.; ONUCHIC, J. N. SMOG@ctbp: simplified deployment of structure-based models in gromacs. **Nucleic Acids Research**, v. 38, n. suppl\_2, p. W657–W661, 2010.
- 236 NOEL, J. K.; WHITFORD, P. C.; ONUCHIC, J. N. The shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function. **The Journal of Physical Chemistry B**, v. 116, n. 29, p. 8692–8702, 2012.
- 237 NIEMYSKA, W.; DABROWSKI-TUMANSKI, P.; KADLOF, M.; HAGLUND, E.; SULKOWSKI, P.; SULKOWSKA, J. I. Complex lasso: new entangled motifs in proteins. **Scientific Reports**, v. 6, n. 1, p. 36895, 2016.
- 238 ZHAO, H.; XU, D.; TIAN, T.; KONG, F.; LIN, K.; GAN, S.; ZHANG, H.; LI, G. Molecular and functional dissection of EARLY-FLOWERING 3 (ELF3) and ELF4 in Arabidopsis. **Plant Science**, v. 303, p. 110786, 2021.
- 239 JUNG, J.-H. *et al.* A prion-like domain in elf3 functions as a thermosensor in arabidopsis. **Nature**, v. 585, n. 7824, p. 256–260, 2020.

## APÊNDICES

## APÊNDICE A – LISTA DE PUBLICAÇÕES

## ARTIGOS PUBLICADOS

- SANCHES, M. N., PARRA, R. G., VIEGAS, R. G., de OLIVEIRA JR, A. B., WOLYNES, P. G., FERREIRO, D. U., LEITE, V. B. P. Resolving the fine structure in the energy landscapes of repeat proteins. **QRB Discovery**, v. 3, p. e7, 2022.
- VIEGAS, R. G., SANCHES, M. N., CHEN, A. A., PAULOVICH, F. V., GARCIA, A. E., LEITE, V. B. P. Characterizing the Folding Transition-State Ensembles in the Energy Landscape of an RNA Tetraloop. **Journal of Chemical Information and Modeling**, v. 63, n. 17, p. 5641-5649, 2023.
- MARTINS, I. B. S., VIEGAS, R. G., SANCHES, M. N., de ARAUJO, A. S., LEITE, V. B. P. Probing Mastoparan-like Antimicrobial Peptides Interaction with Model Membrane Through Energy Landscape Analysis. **The Journal of Physical Chemistry B**, v. 128, n. 1, p. 163-171, 2023.
- DA SILVA, F. B., SIMIEN, J. M., VIEGAS, R. G., HAGLUND, E., LEITE, V. B. P. Exploring the folding landscape of leptin: Insights into threading pathways. **Journal of Structural Biology**, v. 216, n. 1, p. 108054, 2024.
- VIEGAS, R. G., MARTINS, I. B. S., SANCHES, M. N., DE OLIVEIRA JUNIOR, A. B., DE CAMARGO, J. B., PAULOVICH, F. V., LEITE, V. B. P. ELViM: Exploring Biomolecular Energy Landscapes through Multidimensional Visualization. **Journal of Chemical Information and Modeling**, v. 64, n. 8, p. 3443-3450, 2024.
- VIEGAS, R. G., MARTINS, I. B. S., LEITE, V. B. P. Understanding the Energy Landscape of Intrinsically Disordered Protein Ensembles. **Journal of Chemical Information and Modeling**, v. 64, n. 10, p. 4149-4157, 2024.

*PREPRINT* REVISADO

- LINDSAY, R. J., VIEGAS, R. G., LEITE, V. B. P., WIGGE, P. A.; HANSON, S. M. Molecular dynamics simulations illuminate the role of sequence context in the ELF3-PrD-based temperature sensing mechanism in plants. **eLife**, v. 13, e102410.1, 2024.

## CAPÍTULO DE LIVRO

- VIEGAS, R. G., SANCHES, M. N., LEITE, V. B. P. Cancer-related Intrinsic Disordered Proteins: Functional Insights from Energy Landscape Analysis. *In: SALGIA, R. et al. Cancer Systems Biology and Translational Mathematical Oncology.* Oxford: Oxford University Press. *In press.*

## ARTIGOS SUBMETIDOS PARA ANÁLISE

- VIEGAS, R. G., WU, H., SANCHES, M. N., PAPOIAN, G. A., LEITE, V. B. P. Energy Landscapes and Structural Plasticity of Intrinsically Disordered Histones. 2025. Submetido.
- VIEGAS, R. G., SINHA, A., MAINAN A., SANBONMATSU, K. Y., ONUCHIC, J. N., ROY, S., LEITE, V. B. P. Critical role of  $Mg^{2+}$  ions in folding transitions: anchoring the A-minor twist in the SAM-II riboswitch. 2025. Submetido.
- ROS, J., ARLEO, A., VIEGAS, R. G., LEITE, V. B. P., PAULOVICH, F. V. Challenges and opportunities for the visualization of protein energy landscapes. 2025. Submetido.

## APÊNDICE B – CAPÍTULO 3: RESULTADOS COMPLEMENTARES

Como forma de quantificar a qualidade das projeções geradas com diferentes parâmetros, serão utilizadas algumas métricas de qualidade, que comparam as distâncias originais,  $\delta_{ij}$ , com as distâncias projetadas,  $d_{ij}$ .

O *Kruskal stress* ( $K_S$ ) (219) é amplamente utilizado em técnicas de MDS para avaliar a preservação global das distâncias e é definido como:

$$K_S = \sqrt{\frac{\sum_{i<j} (d_{ij} - \delta_{ij})^2}{\sum_{i<j} \delta_{ij}^2}}, \quad (\text{B.1})$$

sendo que valores menores de  $K_S$  indicam maior preservação das distâncias. Como uma forma de estimar a preservação das distâncias localmente, será também considerado o valor do *Kruskal stress* calculado para cada ponto, como:

$$K_S(i) = \sqrt{\frac{\sum_j (d_{ij} - \delta_{ij})^2}{\sum_j \delta_{ij}^2}}. \quad (\text{B.2})$$

Como este valor é calculado para cada ponto na projeção, é possível indicar tais valores através de um mapa de cores. Dessa forma, a variação de  $K_S$  ao longo da projeção pode ser analisada, indicando regiões nas quais a estrutura do espaço original é melhor preservada.

O Erro Absoluto Médio ( $\bar{E}$ ), que avalia a discrepância absoluta média entre  $d_{ij}$  e  $\delta_{ij}$ , é definido como:

$$\bar{E} = \frac{1}{N_p} \sum_{i<j} |d_{ij} - \delta_{ij}|, \quad (\text{B.3})$$

sendo que  $N_p$  é o número total de pares únicos  $(i, j)$ . Valores menores de  $\bar{E}$  indicam uma maior correspondência entre  $d_{ij}$  e  $\delta_{ij}$ . De forma análoga, o Erro Absoluto Médio local é definido por:

$$\bar{E}(i) = \frac{1}{N-1} \sum_j |d_{ij} - \delta_{ij}|. \quad (\text{B.4})$$

Os valores globais de  $K_S$  e  $\bar{E}$ , para todas as projeções analisadas no Capítulo 3, são sumarizados na Tabela 1, que também lista os parâmetros utilizados em cada uma delas. As projeções P1 e P2 são réplicas independentes com todos os parâmetros iguais e possuem apenas uma pequena diferença em  $\bar{E}$ .

A Figura 8 analisa o impacto do parâmetro  $\sigma_0$  nas projeções P1 a P4. Comparando-se as métricas globais para tais projeções, nota-se que a projeção P3 possui os maiores valores de  $K_S$  e  $\bar{E}$ , enquanto P4 possui os menores. Esse resultado deve-se ao fato que o parâmetro  $\sigma_0$  utilizado para a projeção P4 leva a dissimilaridades menores, sendo mais fácil de acomodar a distribuição de distâncias no plano. No entanto, a mudança no  $\sigma_0$  impacta a resolução da similaridade. Assim, embora P4 possua menores valores de  $K_S$ , diferenças estruturais próximas à bacia de estado envelado são perdidas.

A Figura 9 analisa o impacto do parâmetro  $L_r$  nas projeções P1, P5 e P6. De acordo com os dados da Tabela 1, a projeção P5, gerada em regime constante, possui valores ligeiramente maiores de  $K_S$  e  $\bar{E}$ . Esses valores são ainda maiores para a projeção P6, que combina os dois regimes da taxa de aprendizado, mas possui um valor elevado de  $L_{r_{\min}}$ . Apesar disso, a variação nos indicadores de qualidade não é excessiva e, juntamente com a análise qualitativa realizada no texto principal, mostra que a estrutura global da projeção é mantida com algumas distorções.

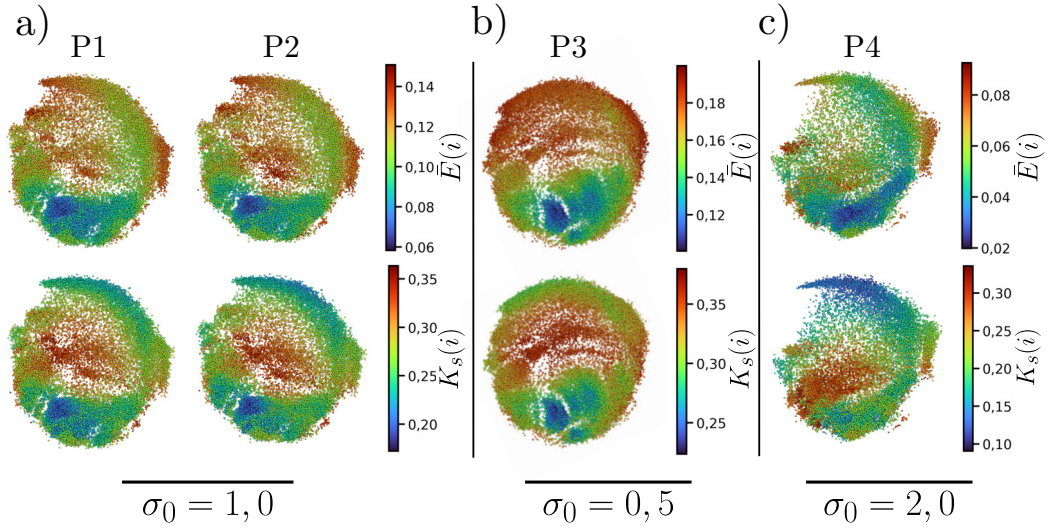
Tabela 1 – Parâmetros e Métricas de Qualidade global para as projeções apresentadas no Capítulo 3.  $K_S$  é *Kruskal stress* e  $\bar{E}$  é o erro absoluto médio.

	$\sigma_0$	$L_{r_0}$	$L_{r_{\min}}$	D	$K_S$	$\bar{E}$
P1	1,0	0,3	0	0,95	0,258	0,101
P2	1,0	0,3	0	0,95	0,258	0,102
P3	0,5	0,3	0	0,95	0,310	0,154
P4	2,0	0,3	0	0,95	0,181	0,047
P5	1,0	0,125	0,125	0	0,271	0,105
P6	1,0	0,5	0,25	2,00	0,289	0,111

Fonte: Elaborada pelo autor (2025).

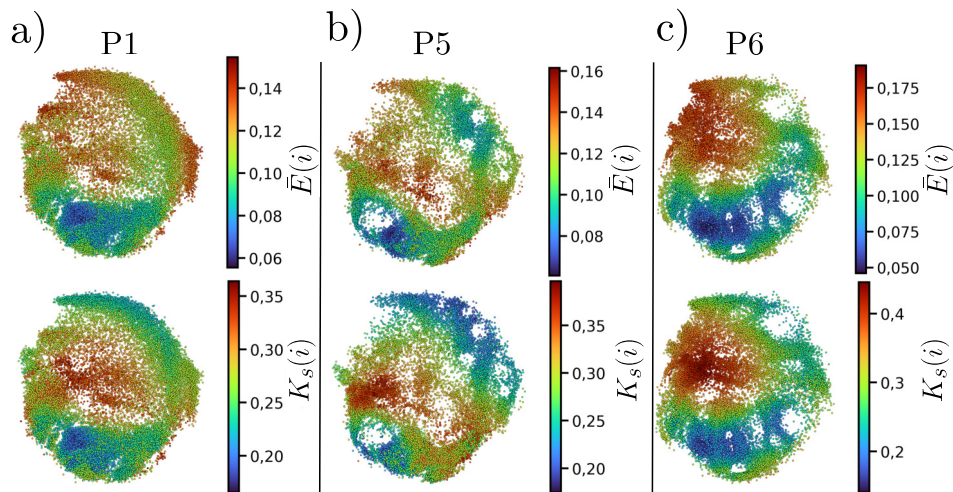
Por fim, são analisados como as métricas de qualidade variam localmente, utilizando os valores calculados para colorir as respectivas projeções. O resultado desta análise está apresentado nas Figuras B.1 e B.2.

Figura B.1 – Impacto do parâmetro  $\sigma_0$  nas projeções P1 a P4. A análise detalhada mostra como diferentes valores de  $\sigma_0$  afetam a preservação local e global das distâncias.



Fonte: Elaborada pelo autor (2025).

Figura B.2 – Impacto do parâmetro  $L_r$  nas projeções P1, P5 e P6. A análise detalhada mostra como diferentes valores e regimes de  $L_r$  afetam a preservação local e global das distâncias.

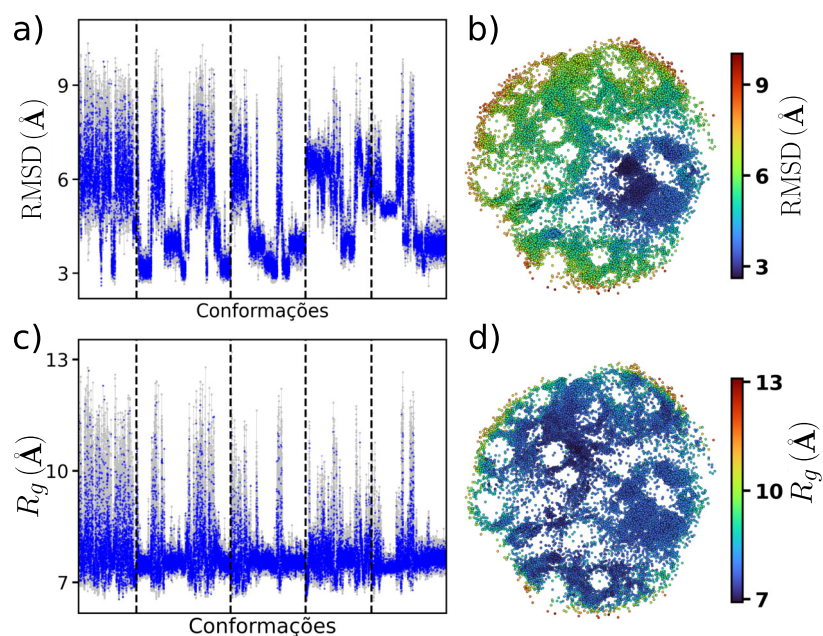


Fonte: Elaborada pelo autor (2025).

## APÊNDICE C – CAPÍTULO 4: METODOLOGIA E RESULTADOS COMPLEMENTARES

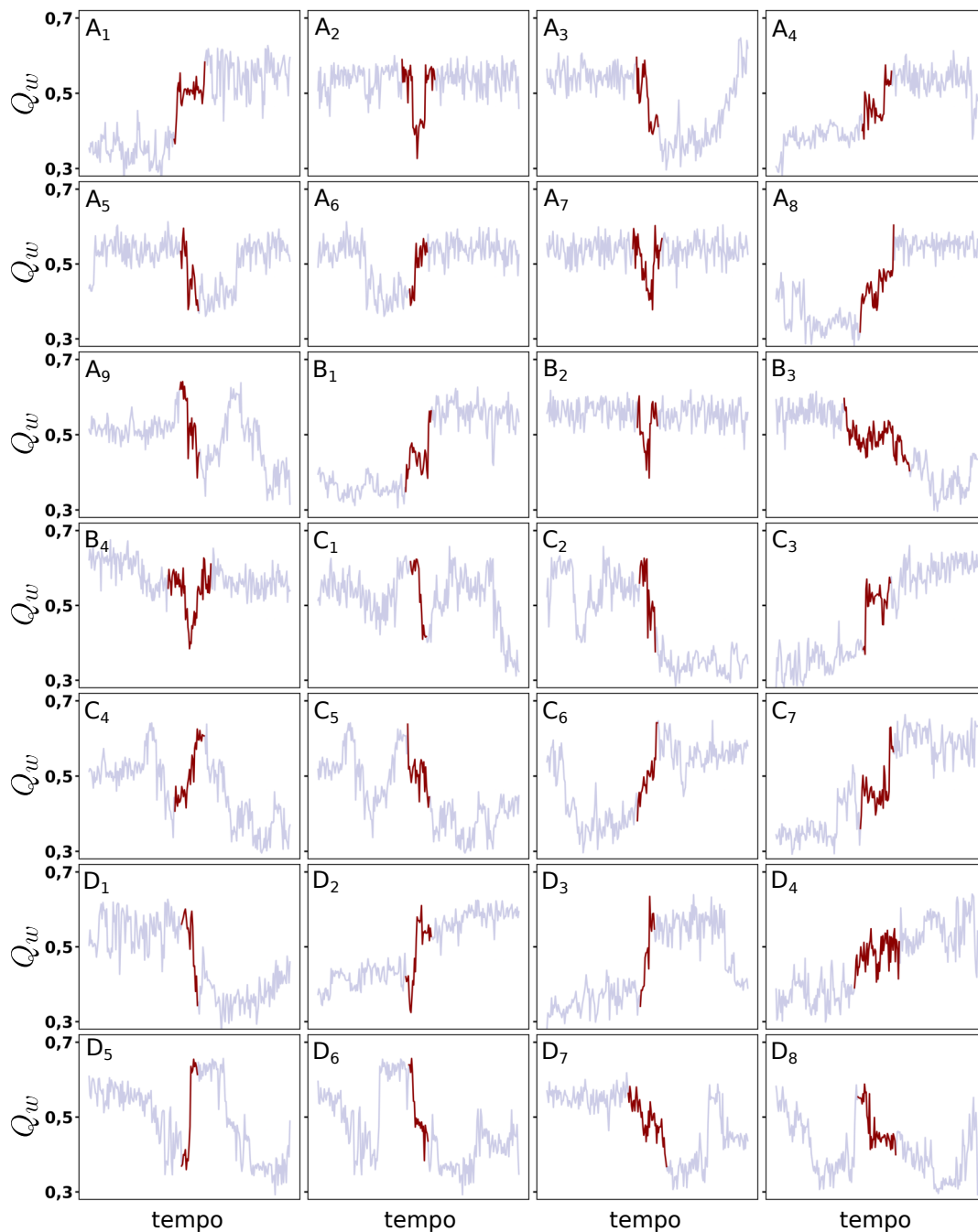
### C.1 TETRA-ALÇA GCAA

Figura C.1 – Evolução temporal do RMSD, calculado em relação à estrutura cristalográfica (PDB: 1ZIH), e do raio de giração ( $R_g$ ) são mostradas nos gráficos à esquerda. Os valores correspondentes aos pontos selecionados para a projeção ELViM são destacados em azul, enquanto os dados de todos os pontos simulados são representados em cinza. À direita, reproduzimos a projeção ELViM (Figura 12.a), com os pontos coloridos de acordo com seus respectivos valores de RMSD e  $R_g$ .



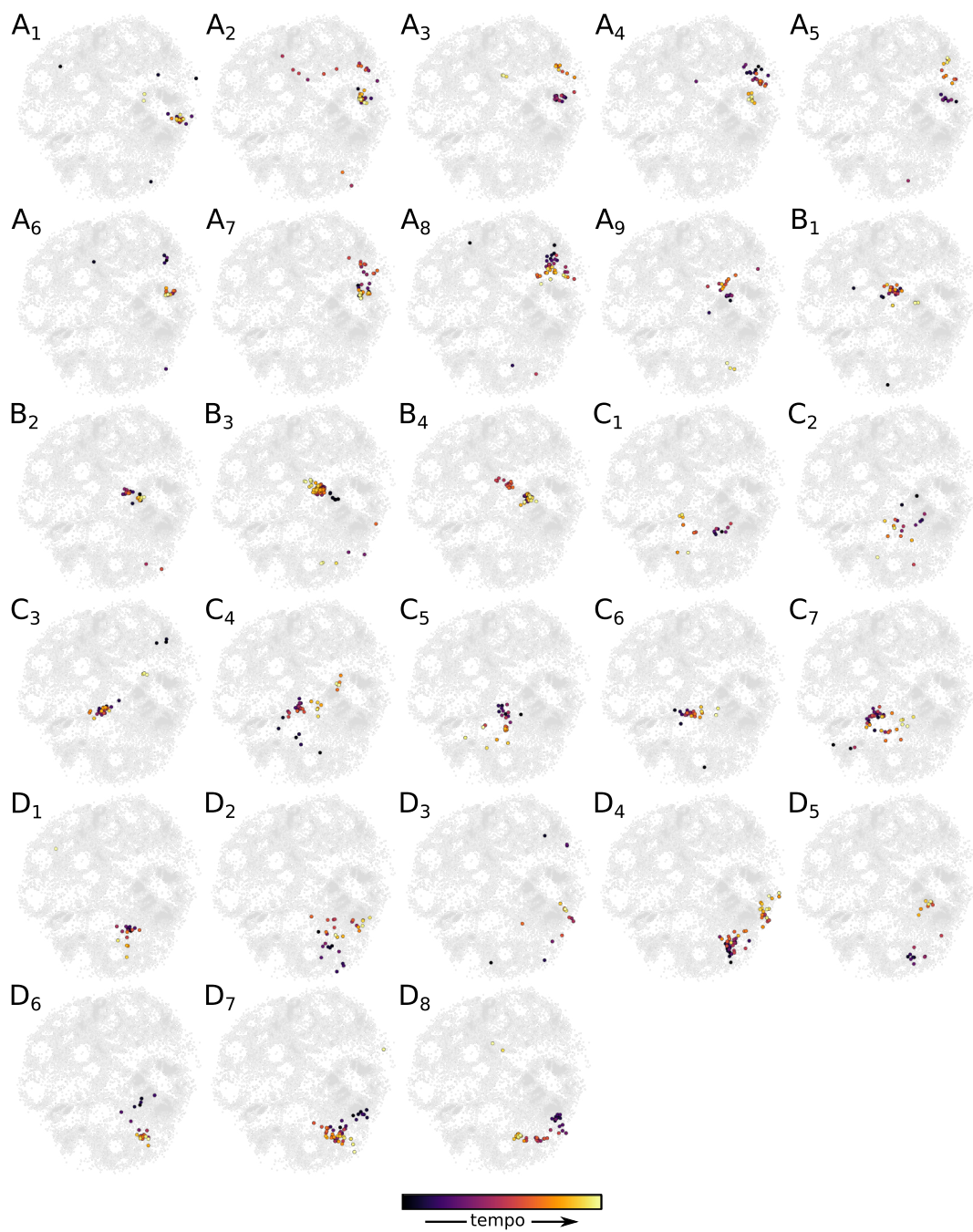
Fonte: Adaptada de (62).

Figura C.2 – Evolução temporal de  $Q_w$  durante as 28 transições desnaturado/enovelado restantes (Figura 14).



Fonte: Adaptada de (62).

Figura C.3 – Evolução temporal das demais transições da tetra-alça GCAA (Figura 14) através do espaço de fase efetivo gerado pelo ELViM. Os pontos da transição são destacados e coloridos conforme o tempo, permitindo observar a dinâmica das transições em cada trajetória.



Fonte: Adaptada de (62).

## C.2 RIBOCOMUTADOR SAM-II

### C.2.1 Metodologia Complementar

#### C.2.1.1 *Projeção ELViM*

Para gerar a projeção ELViM analisada, 17 trajetórias, compostas por 5.000 quadros cada uma foram utilizadas. As simulações amostrando a transição entre os estados parcialmente enovelado–enovelado (PE-E) foram realizadas às temperaturas  $96,5T_R$ ,  $97,0T_R$ ,  $97,5T_R$ ,  $98,0T_R$ ,  $98,5T_R$ . Simulações explorando a bacia do estado parcialmente enovelado (PE) foram realizadas às temperaturas  $100,0T_R$ ,  $100,5T_R$ ,  $101,0T_R$ ,  $101,5T_R$ ,  $102,0T_R$ ,  $102,5T_R$ . A transição entre os estados desnaturado e parcialmente enovelado (D-PE) foram amostradas às temperaturas de  $103T_R$ ,  $103,2T_R$ ,  $104T_R$ ,  $104,4T_R$ ,  $106,6T_R$ . As conformações foram selecionadas para a projeção descartando-se os 700 quadros iniciais e mantendo-se um a cada três quadros para os casos PE-E e D-PE e um a cada sete quadros para o caso PE. Além disso, os quadros foram amostrados sequencialmente durante cada uma das cinco transições D-PE utilizadas, resultando em 954 quadros adicionais (o corresponde a 5% dos pontos na projeção). Os pontos selecionados totalizam 19.005 conformações.

#### C.2.1.2 *Análise das interações RNA-Mg<sup>2+</sup>*

A interação RNA-Mg<sup>2+</sup> foi investigada a partir da distribuição de íons nas proximidades de cada grupo fosfato ao longo da simulação. A função de distribuição radial (RDF, do inglês Radial Distribution Function) das distâncias Mg<sup>2+</sup>-P foi calculada, estabelecendo-se um raio de corte de 8 Å como limite da camada de solvatação, como mostrado em um trabalho anterior (122). Em cada quadro da simulação, registrou-se o número de íons dentro dessa região esférica em torno de cada grupo fosfato  $P_i$ .

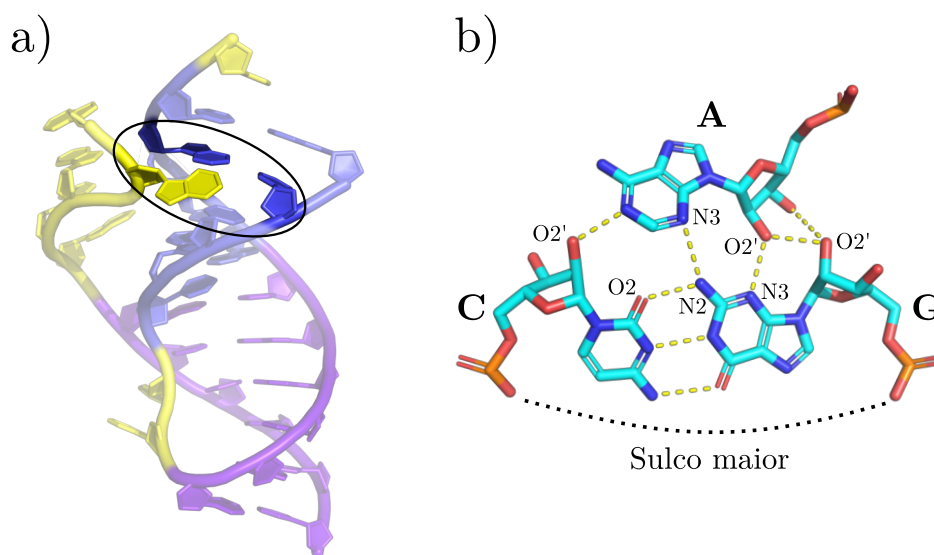
A frequência de visitação ao longo da trajetória de enovelamento foi analisada classificando os quadros conforme seus valores de  $Q$  e organizando-os em compartimentos de largura  $Q = 0,05$ . A média do número de íons para cada  $P_i$  foi então calculada dentro de cada compartimento. Para melhorar a estatística, essa análise incluiu mais 12 simulações realizadas em temperaturas efetivas entre  $104,8T_R$  e  $107,0T_R$ , totalizando 140.000 quadros de simulação.

Além disso, contatos entre dois grupos fosfato não locais mediados por íons Mg<sup>2+</sup> foram analisados em uma trajetória representativa do estado enovelado, amostrada a uma temperatura de  $96,5T_R$ . Para essa análise, foram examinadas as distâncias entre um íon Mg<sup>2+</sup> e os grupos fosfato de dois resíduos, denotadas como Mg<sup>2+</sup>- $P_i$  e Mg<sup>2+</sup>- $P_j$ . A restrição de que  $j$  seja maior ou igual a  $i + 2$  é imposta para garantir o comportamento não-local da interação. Além disso, um segundo limite foi aplicado à distância entre os grupos fosfato, denominada  $P_i$ - $P_j$ , assegurando proximidade espacial desses grupos. Os

valores dos limites de distância foram determinados com base na função RDF  $Mg^{2+}$ -P e P-P, sendo respectivamente iguais a 8 Å e 7,8 Å (122). Para cada contato mediado encontrado, foi acumulado o número de quadros nos quais o contato estava estabelecido e a frequência relativa entre os contatos mediados foi calculada.

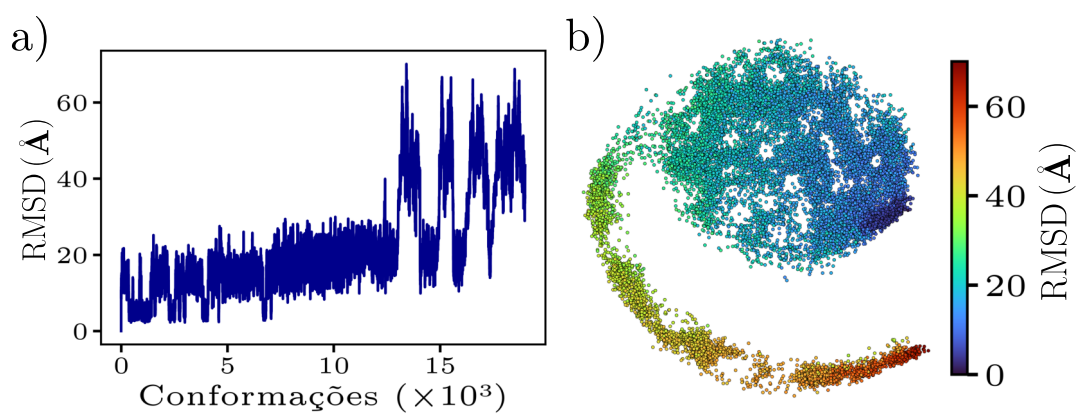
### C.2.2 Figuras Complementares

Figura C.4 – Exemplo de interação A-menor do tipo I presente em um pseudonó do tipo H. (a) Estrutura cristalográfica de um RNA, com estrutura de pseudonó, que se liga à biotina (PDB: 1F27) (220). As duas hélices são representadas em azul e roxo e as alças em amarelo. A interação A-menor encontra-se destacada, com a adenina (em amarelo) interagindo com o sulco menor do par canônico C-G. (b) Detalhes da interação A-menor evidenciando os átomos envolvidos.



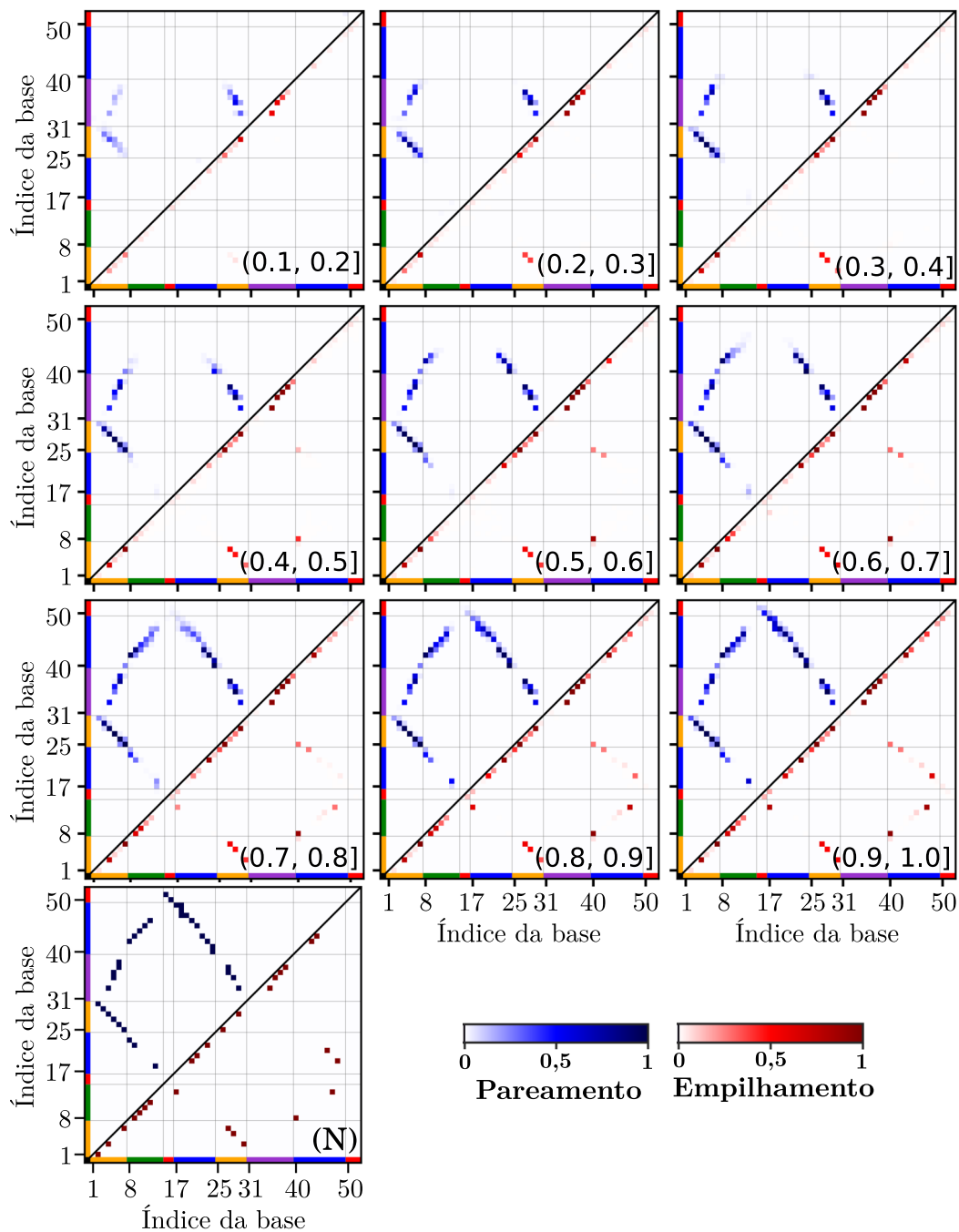
Fonte: Elaborada pelo autor (2025).

Figura C.5 – (a) Evolução temporal do RMSD calculado em relação a estrutura cristalográfica (PDB: 2QWY, cadeia A) para conformações selecionadas para a projeção ELViM. As trajetórias foram concatenadas em ordem crescente de temperatura efetiva. (b) Espaço conformacional efetivo colorido no base nos valores de RMSD.



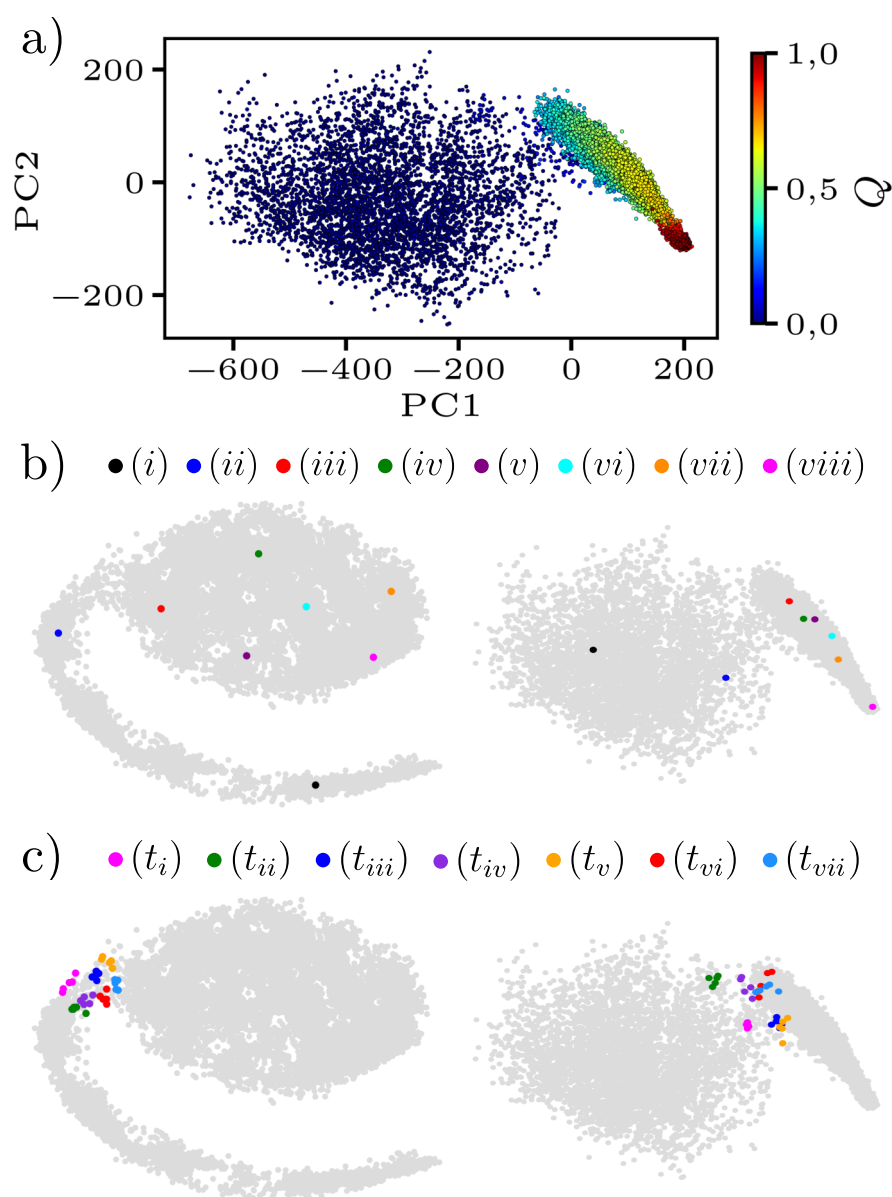
Fonte: Elaborada pelo autor (2025).

Figura C.6 – Nesta análise todas as 28 simulações (140.000 quadros) foram utilizados. As conformações foram categorizados em intervalos de  $Q = 0,1$ , e a frequência das interações foi calculada dentro de cada intervalo. A matriz triangular superior indica interações de pareamento, enquanto a inferior indica empilhamento. O mapa de cores é utilizado para indicar a frequência. As interações foram classificadas através do programa Barnaba (108). O mapa de interações da estrutura cristalográfica (N) é fornecido para comparação.



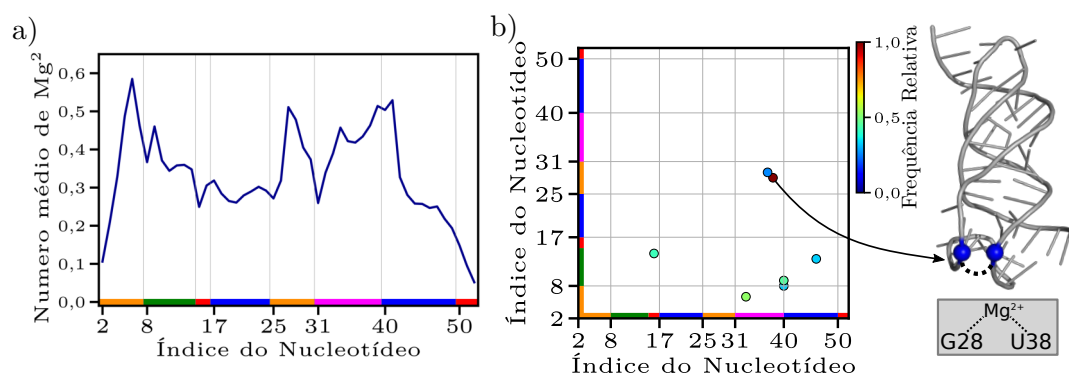
Fonte: Elaborada pelo autor (2025).

Figura C.7 – Comparação qualitativa do ELViM com PCA para o SAM-II. (a) A análise de componentes principais (PCA) foi realizada sobre a matriz de coordenadas cartesianas de todos os átomos pesados. Para remover translação e rotação, as estruturas foram centradas na estrutura média, refinando esse processo 10 vezes até a convergência. A análise foi conduzida com a biblioteca *Scikit-learn* (36). Para facilitar a comparação com a projeção ELViM, a superfície PCA foi refletida e cada ponto foi colorido de acordo com  $Q$ . Os dois primeiros componentes acumulam 68% da variância total. (b) A localização das conformações representativas mostradas na Figura 16.f está indicada nas projeções ELViM (à esquerda) e PCA (à direita). (c) As estruturas que estão presentes em cada LCS indicada na Figura 17 são mostradas nas projeções ELViM (à esquerda) e PCA (à direita).



Fonte: Elaborada pelo autor (2025).

Figura C.8 – Interação RNA-Mg<sup>2+</sup> (a) Número médio de átomos de Mg<sup>2+</sup> em torno de cada fosfato, considerando uma distância limite de 8 Å. As 28 simulações foram incluídas nesta análise. (b) Mapa de contatos mediados por Mg<sup>2+</sup> para trajetória representativa do estado enovelado (96,5T<sub>R</sub>). Resultado obtido conforme descrito na Seção C.2.1.2, segundo análise realizada por Avijit Mainan.



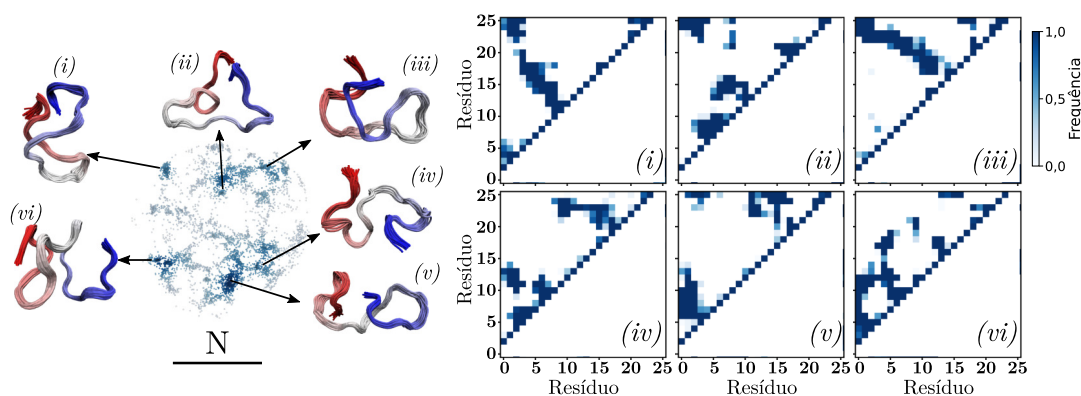
Fonte: Elaborada pelo autor (2025).

## APÊNDICE D – CAPÍTULO 5: RESULTADOS COMPLEMENTARES

### D.1 CAUDA H4

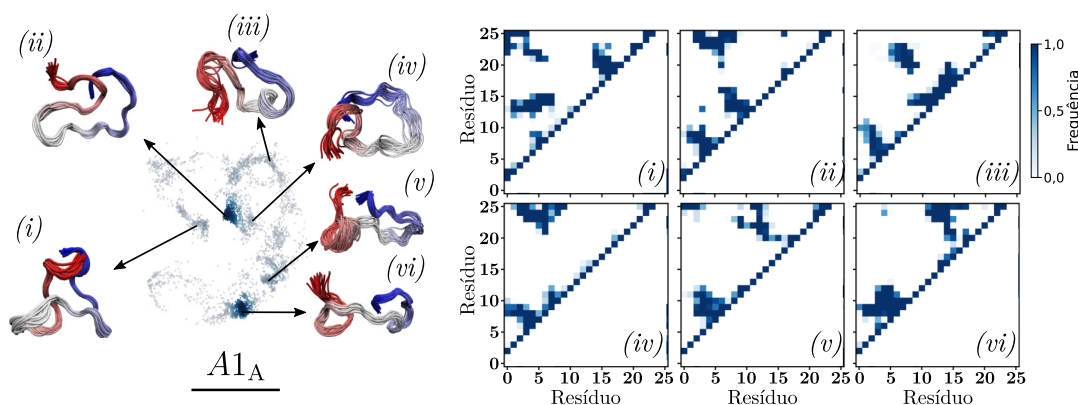
Nesta seção, serão apresentadas LCSs e mapas de contato para cada modelo da cauda H4. Cada figura apresenta novamente o espaço de fase efetivo de cada modelo, com os pontos coloridos de acordo com a densidade local, calculada através de KDE com núcleo gaussiano. Algumas regiões, incluindo as de maiores densidades, foram arbitrariamente selecionadas para análise. As LCSs indicadas são compostas por 20 conformações sobrepostas com o amino-terminal em vermelho. Além disso, fornecemos mapas de contato médios, ou frequências de contato, para cada LCS. Os contatos foram definidos entre pares de carbonos alfa, usando um critério de distância de 8 Å.

Figura D.1 – LCSs e mapas de contato para a forma nativa (N) da cauda H4.



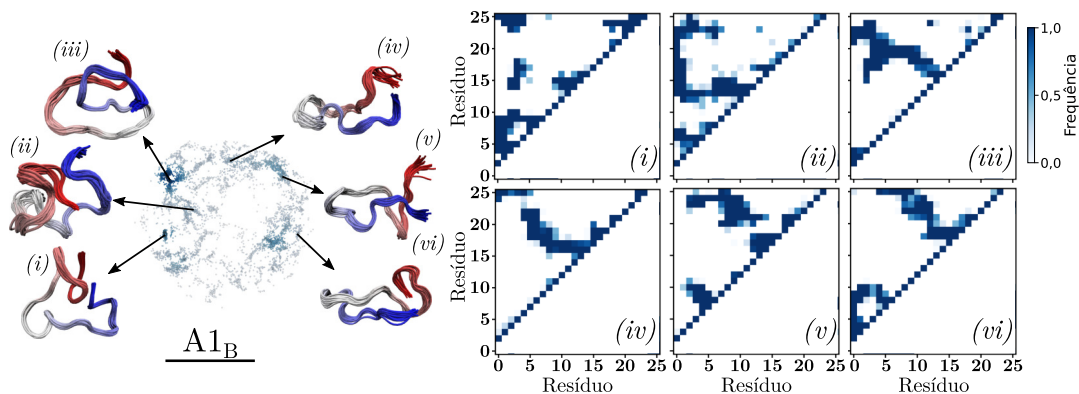
Fonte: Elaborada pelo autor (2025).

Figura D.2 – LCSs e mapas de contato para a forma A1<sub>A</sub> da cauda H4.



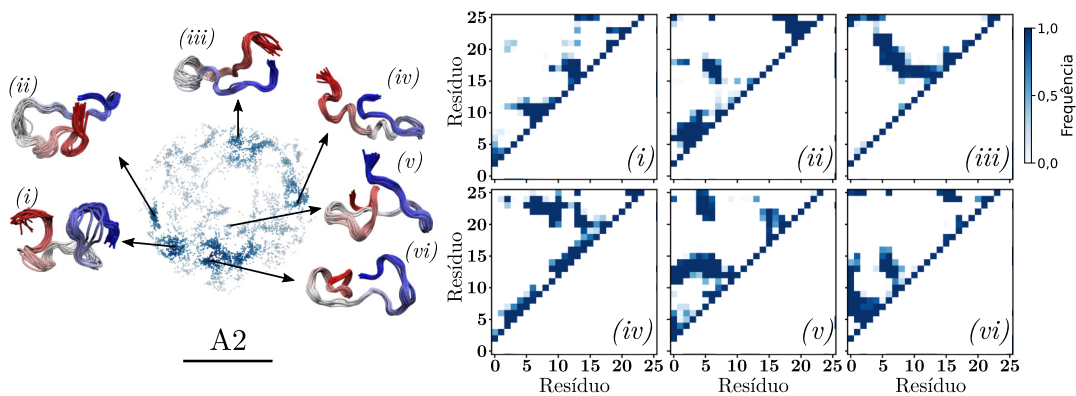
Fonte: Elaborada pelo autor (2025).

Figura D.3 – LCSs e mapas de contato para a forma A1<sub>B</sub> da cauda H4.



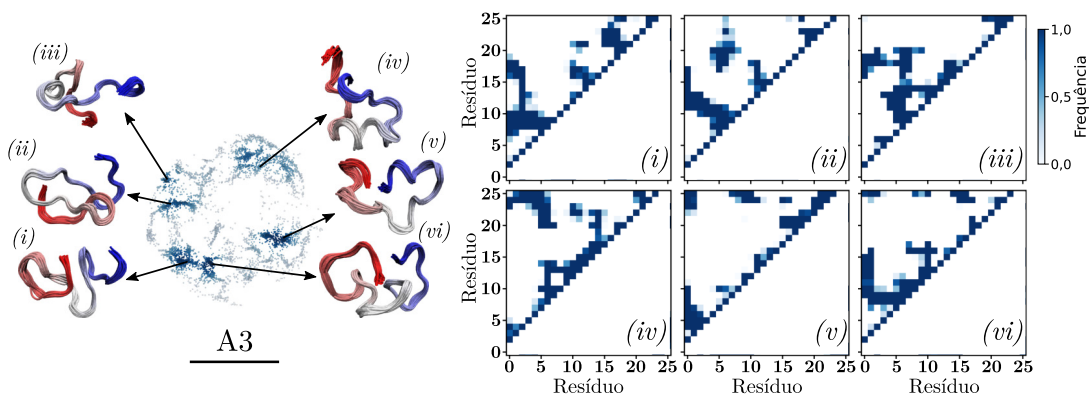
Fonte: Elaborada pelo autor (2025).

Figura D.4 – LCSs e mapas de contato para a forma A2 da cauda H4.



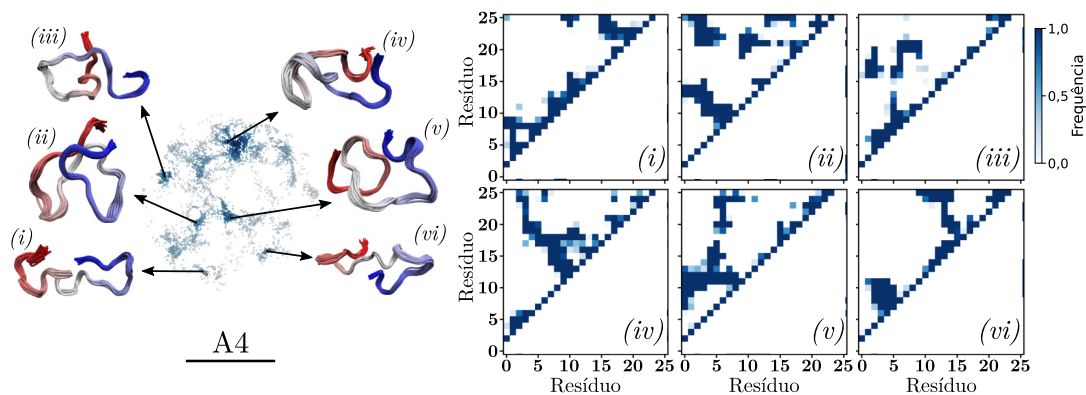
Fonte: Elaborada pelo autor (2025).

Figura D.5 – LCSs e mapas de contato para a forma A3 da cauda H4.



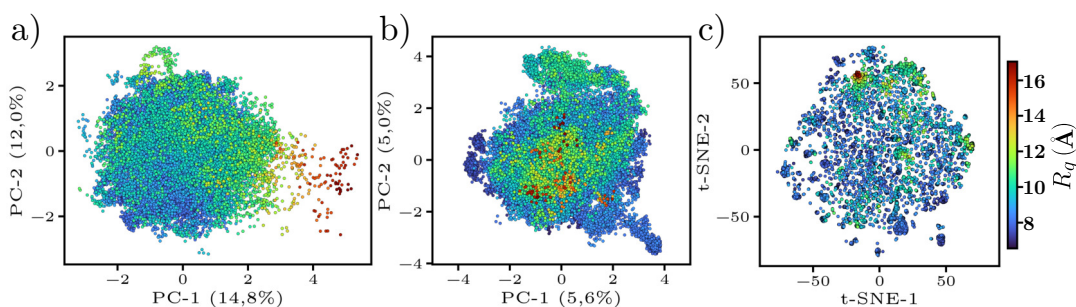
Fonte: Elaborada pelo autor (2025).

Figura D.6 – LCSs e mapas de contato para a forma A4 da cauda H4.



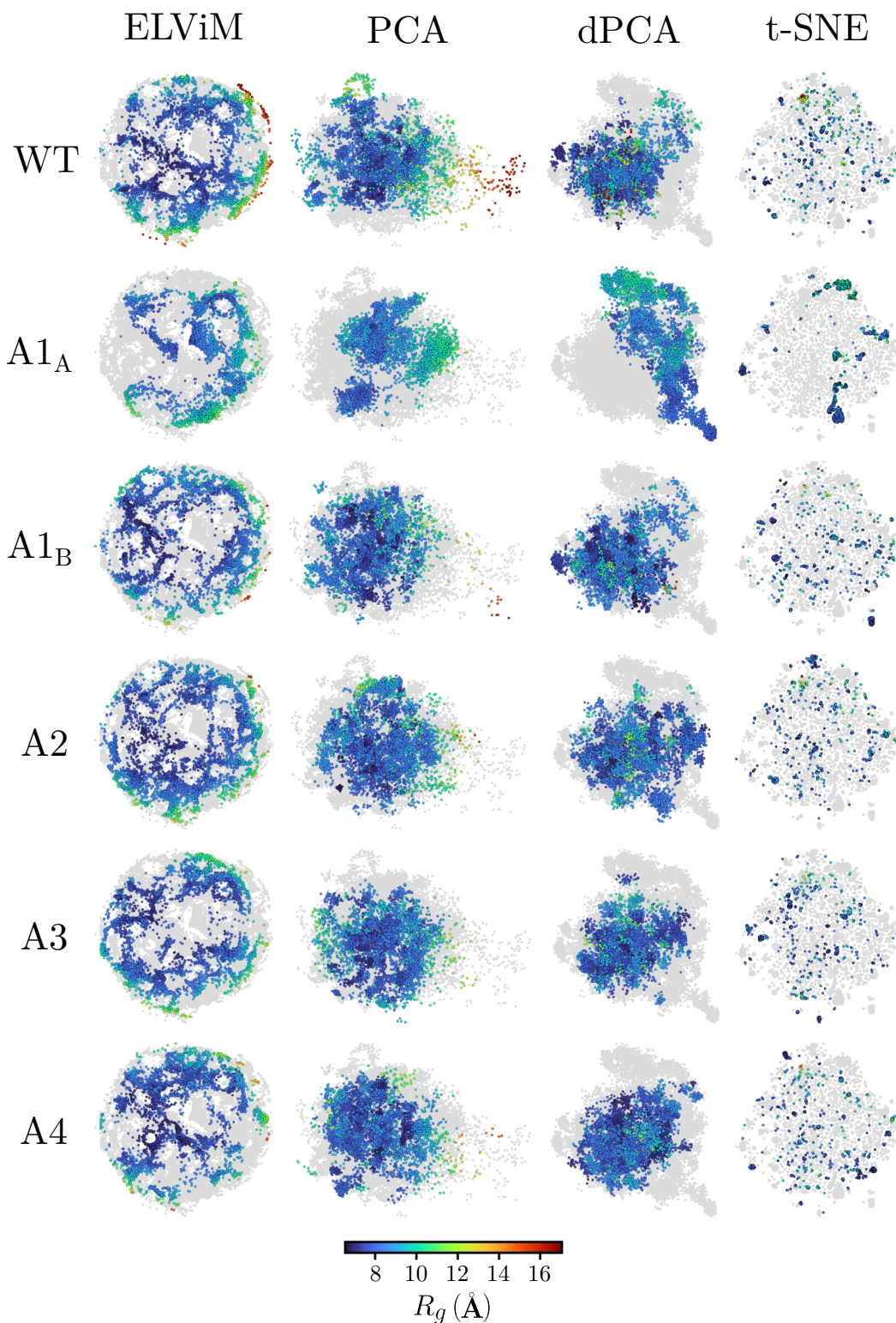
Fonte: Elaborada pelo autor (2025).

Figura D.7 – Cauda H4: projeção através de PCA e t-SNE. (a) PCA das coordenadas cartesianas dos carbonos alfa. Para remover translação e rotação, as conformações foram centralizadas na estrutura média da trajetória de MD, e esse processo foi refinado até atingir a convergência. (b) PCA baseado nos ângulos de torção  $\Phi$  e  $\Psi$ . Devido à periodicidade, os ângulos diedrais foram mapeados no círculo unitário, calculando-se o seno e o cosseno de cada ângulo. A variância explicada por cada componente principal é indicada nos rótulos dos eixos. Ambas as análises PCA foram realizadas utilizando a biblioteca *Scikit-learn* (36). (c) t-SNE foi aplicado seguindo o protocolo descrito na ref. (164). A matriz de distâncias de entrada foi calculada como o RMSD dos carbonos alfa. A análise do coeficiente de silhueta (ref. (164)) resultou em uma perplexidade de 400 para o t-SNE e  $K=100$  para o algoritmo K-means. Os demais parâmetros foram mantidos conforme descrito na ref. (164). Para facilitar a comparação, os pontos foram coloridos com base nos valores de  $R_g$ , como na projeção ELViM.



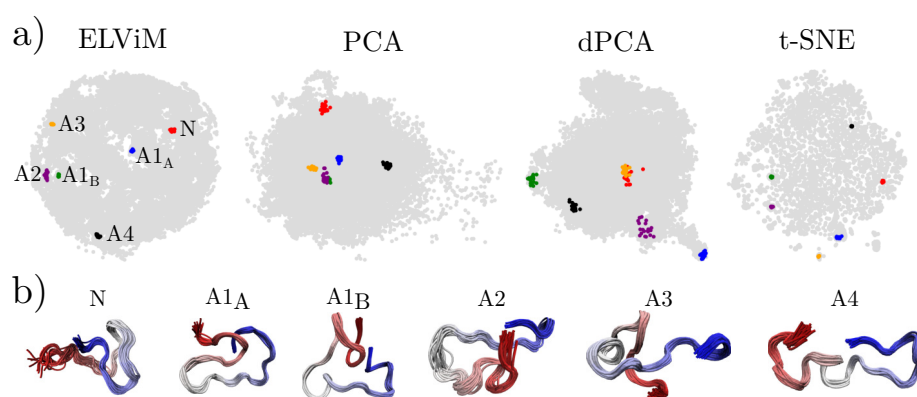
Fonte: Elaborada pelo autor (2025).

Figura D.8 – Espaço de fase efetivo de cada variante da cauda H4 gerado por diferentes métodos de redução de dimensionalidade. Para facilitar a comparação visual, todos os eixos foram omitidos. Os pontos que representam cada modelo estão coloridos de acordo com os valores de  $R_g$ , enquanto o espaço global é mostrado em cinza.



Fonte: Elaborada pelo autor (2025).

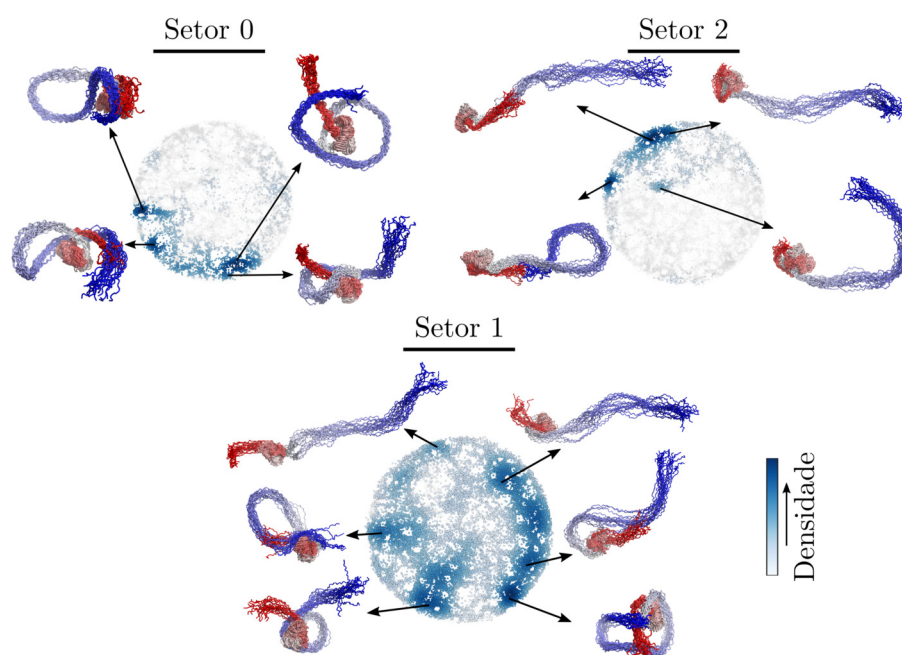
Figura D.9 – Posição relativa das conformações que compõe a assinatura LCS com maior fração de ocupação em cada modelo. (a) Os pontos que representam as conformações de cada LCS são indicados com as mesmas cores em todas as projeções. (b) As LCSs são indicadas para facilitar a comparação.



Fonte: Elaborada pelo autor (2025).

## D.2 HISTONA H1

Figura D.10 – LCS para os diferentes setores de ligação da H1. O espaço efetivo de cada setor indicado é representado com os pontos coloridos de acordo com a densidade local, calculada através de KDE com núcleo gaussiano. As LCSs indicadas são compostas por 20 conformações sobrepostas com o amino-terminal em vermelho.



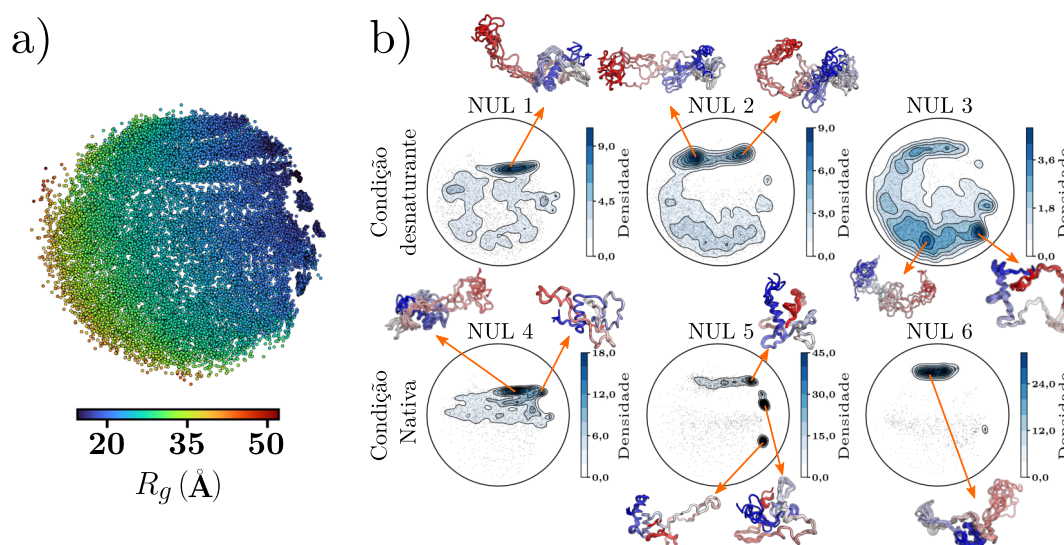
Fonte: Elaborada pelo autor (2025).

### D.3 OUTROS SISTEMA DO PED

Os sistemas adicionais analisados neste estudo foram:

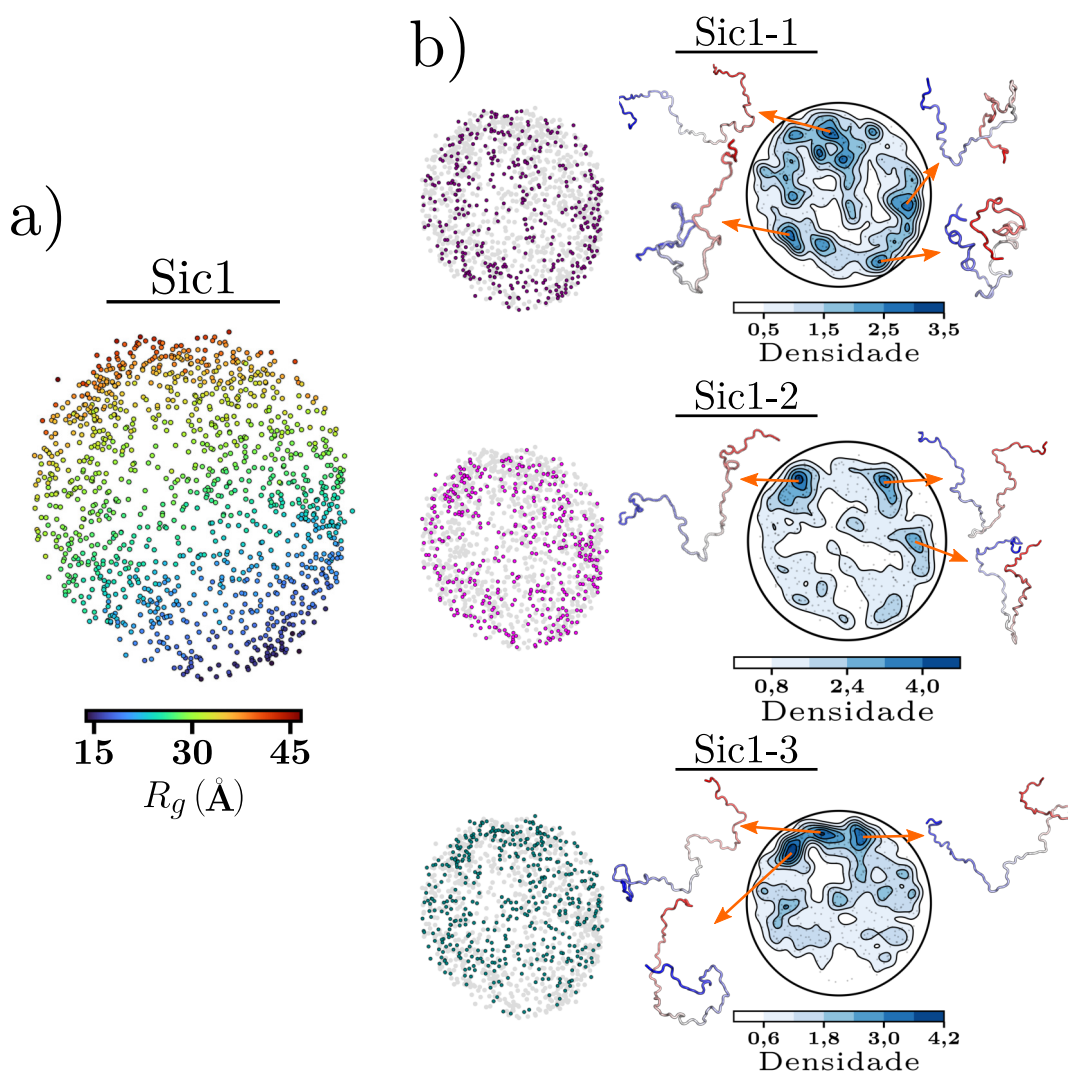
- **NUL (884-993)** (189): um fragmento da proteína do complexo do poro nuclear Nup153 (UniProtKB P49790), disponível em PED00144 e PED00145. Os *ensembles* analisados são: NUL1 (PED00144: e001, 8003 conformações), NUL2 (PED00144: e002, 8930 conformações), NUL3 (PED00144: e003, 7293 conformações), NUL4 (PED00145: e001, 5938 conformações), NUL5 (PED00145: e002, 6620 conformações) e NUL6 (PED00145: e003, 5528 conformações). Estes ensembles foram baseados em simulações atomísticas de *Replica Exchange* com solvatação implícita utilizando ABSINTH e CAMPARI (190, 191). Em seguida, os ensembles foram ponderados através do programa COPER (192) para corresponder aos valores experimentais de  $Rg^2$  e diferentes eficiências médias de FRET. Os valores experimentais foram obtidos em condições desnaturantes (solução salina tamponada com fosfato (*buffer PBS*) com 6M de ureia) e nativas (*buffer PBS*) por meio de smFRET e SAXS. Para a projeção ELViM, foi selecionada uma a cada duas conformações, resultando em 21.157 pontos.
- **Sic1** (221): região amino-terminal Sic1 (1-90). Os conjuntos foram gerados através de abordagens integrativas, na qual *ensembles* de conformações aleatórias geradas pelo TraDES (195) foram refinados através do ENSEMBLE (196), para satisfazer vínculos experimentais. Os *ensembles* analisados são: Sic1-1 (PED00159, 500 conformações) composto pela forma não fosforilada de Sic1, gerado usando dados de SAXS e relaxação paramagnética (PRE); Sic1-2 (PED00160, 500 conformações) utilizando SAXS, PRE e deslocamentos químicos (CS); e Sic1-3 (PED00161, 500 conformações), um conjunto para a forma fosforilada da região amino-terminal da Sic1 (com fosforilação nos resíduos Thr2, Thr5, Thr33, Thr45, Ser69, Ser76 e Ser80), gerado integrando dados de SAXS e PRE. Todos os dados foram validados por smFRET para os três conjuntos. Cada conjunto é composto por 500 conformações.

Figura D.11 – Espaço de fase efetivo ELViM gerado para os *ensembles* do fragmento NUL. Os *ensembles* foram reamostradas utilizando uma a cada duas conformações. (a) O espaço efetivo global, contendo 21.157 conformações, é mostrado com os pontos coloridos com base nos valores de  $R_g$ . (b) A figura mostra a densidade de estados estimada para os *ensembles* individuais NUL1 a NUL6. Conformações representativas, obtidas através da análise LCS, são indicadas para regiões de alta densidade. Cada LCS é composta por cinco conformações, representadas com amino-terminal em azul.



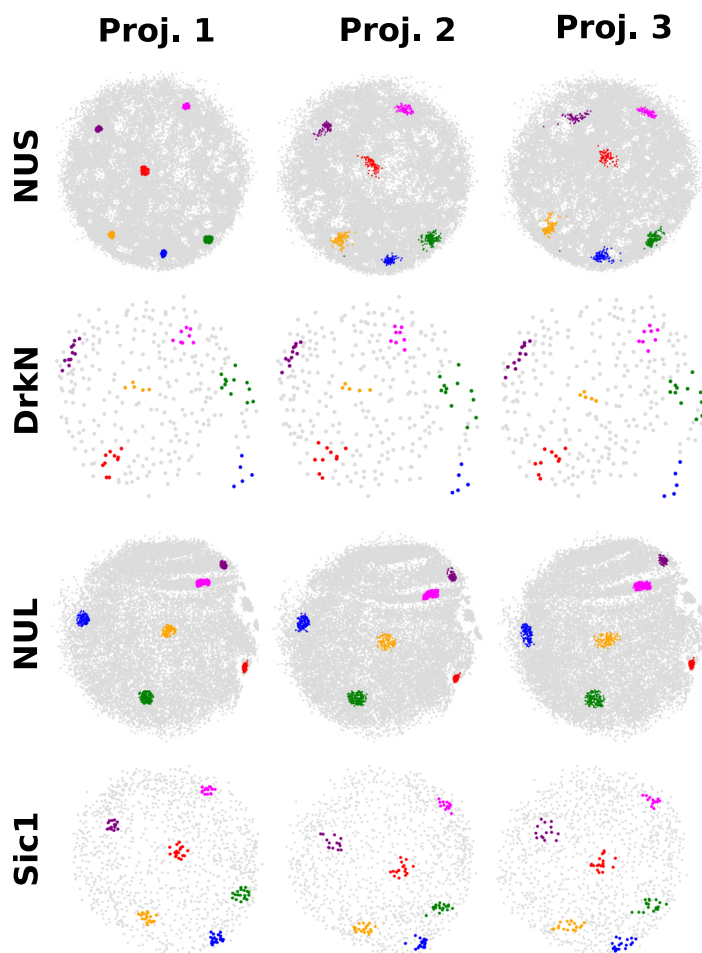
Fonte: Adaptada (63).

Figura D.12 – Análise ELViM para a região amino-terminal desordenada de Sic1 (1-90). (a) A figura mostra a projeção ELViM gerada para os três conjuntos de Sic1, totalizando 1500 conformações, com os pontos coloridos com base nos valores de  $R_g$ . (b) As conformações de cada conjunto são exibidas na coluna à esquerda em roxo (Sic1-1), fúcsia (Sic1-2), e ciano escuro (Sic1-3). Para cada caso, também são fornecidas, na coluna à direita, a densidade de pontos e LCSs (centróide) de algumas regiões de alta densidade.



Fonte: Adaptada de (63).

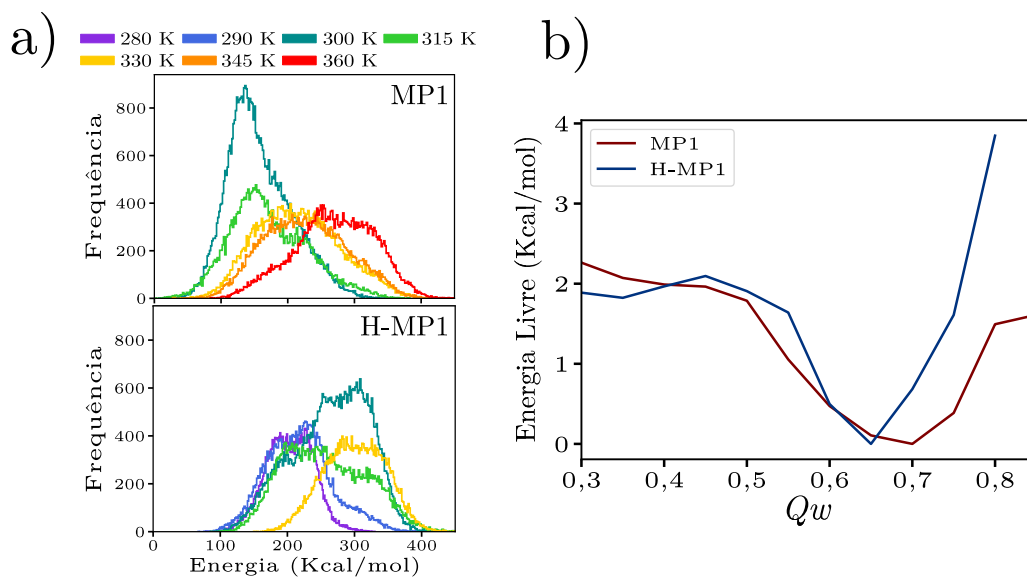
Figura D.13 – Análise qualitativa da reprodutibilidade do ELViM para os *ensembles* analisados do PED. Para cada sistema, três execuções independentes do ELViM foram realizadas. Seis pequenas regiões foram arbitrariamente selecionadas e coloridas com as mesmas cores em todas execuções. Para facilitar a comparação, as projeções foram alinhadas à projeção 1. Os resultados indicam manutenção da vizinha local e da estrutura global da projeção.



Fonte: Adaptada de (63).

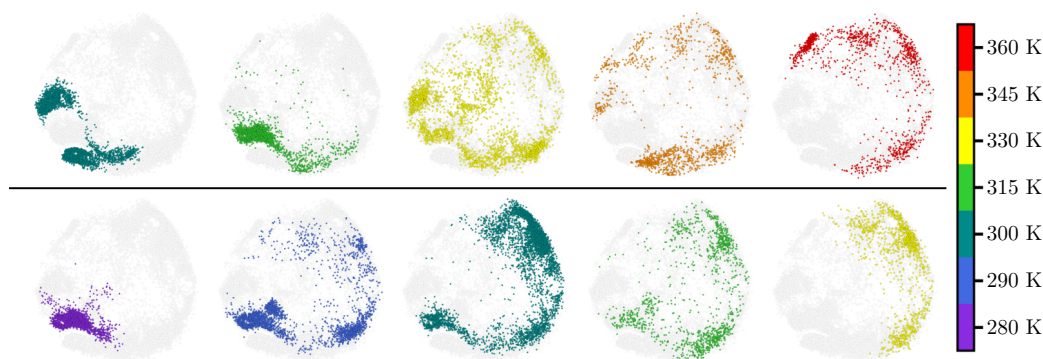
## APÊNDICE E – CAPÍTULO 6: FIGURAS COMPLEMENTARES

Figura E.1 – Histogramas de energia e perfil de energia livre para os peptídeos MP1 e H-MP1. (a) Histogramas de energia amostradas em cada simulação para os peptídeos MP1 e H-MP1. (b) Perfil de energia livre obtido através do WHAM (214), em função de  $Q_w$  para a temperatura de 300 K.



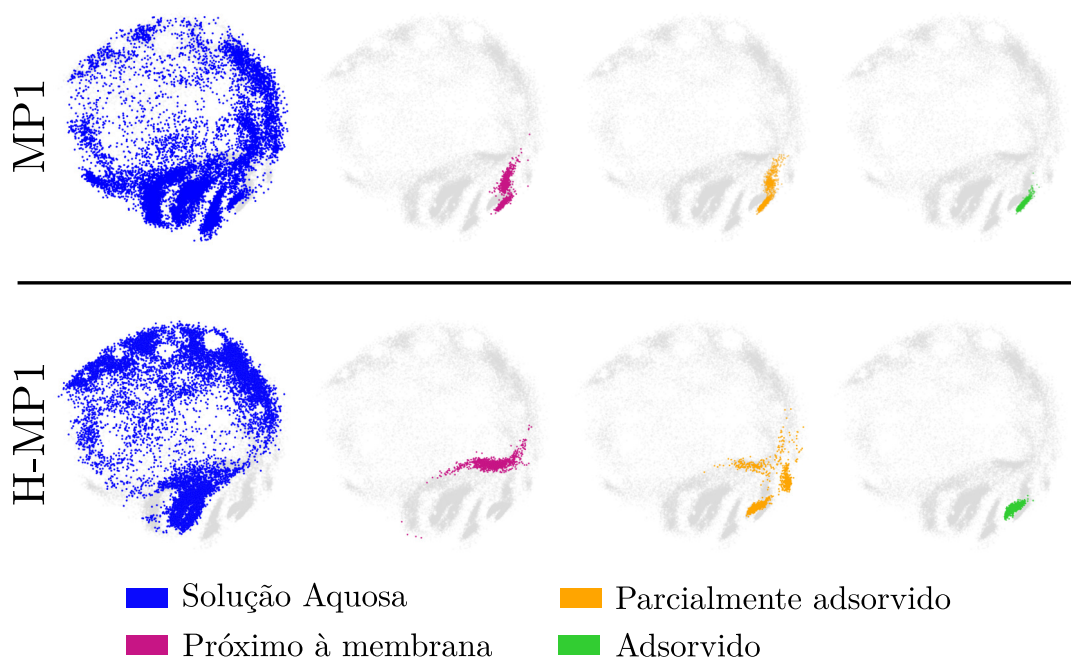
Fonte: Adaptada de (64).

Figura E.2 – Distribuição de temperaturas no espaço de fase efetivo para os peptídeos MP1 e H-MP1. O espaço de fase efetivo para os peptídeos MP1 e H-MP1 em solução aquosa (Figura 27) é reproduzido, destacando-se as conformações amostradas em uma temperatura específica (pontos coloridos), sobrepostas ao espaço de fase global (pontos em cinza).



Fonte: Adaptada de (64).

Figura E.3 – Posição em relação à membrana para os peptídeos MP1 e H-MP1. O espaço de fase efetivo para o sistema peptídeos–membrana (Figura 28) é reproduzido, destacando-se a classificação dos pontos em relação à membrana modelo. A classificação foi realizada analisando as distâncias entre os terminais e centro de massa do peptídeo e a componente z do centro de massa da membrana. Três estados são considerados: peptídeo em solução aquosa próximo à membrana (magenta), peptídeo interagindo com a membrana ou parcialmente adsorvido (laranja) e peptídeo completamente adsorvido (verde).



Fonte: Adaptada de (64).

## APÊNDICE F – APLICAÇÃO DO ELViM A OUTROS SISTEMAS

Neste Apêndice, será brevemente discutida a aplicação do ELViM a três outros sistemas. Os correspondentes trabalhos podem ser encontrados nas referências (65–67). Nas seções a seguir, será apresentado um exemplo de cada trabalho, focando nos resultados obtidos a partir da aplicação do ELViM a tais sistemas.

### F.1 DISTINGUINDO ROTAS DE ENOVELAMENTO EM REPETIÇÃO DE ANQUIRINA

Proteínas de repetição em tandem (TRP, do inglês *Tandem Repeat Protein*) são caracterizadas pela presença de motivos estruturais que se repetem ao longo da cadeia polipeptídica e desempenham papéis cruciais em diversas funções biológicas. Dentre os diversos tipos de TRPs, destacam-se os solenoides alfa, uma classe especial na qual as repetições se empilham lateralmente, formando estruturas alongadas e flexíveis, essenciais para interações proteína-proteína e reconhecimento molecular (222). Um exemplo proeminente desses motivos é a repetição de anquirina (ANK), composta por cerca de 33 aminoácidos que se repetem em tandem, formando domínios estruturais altamente semelhantes (223,224). O enovelamento dessas proteínas é um processo complexo e frequentemente cooperativo que geralmente inicia-se com uma nucleação localizada, seguida pela propagação estrutural aos elementos adjacentes. Se os vínculos energéticos locais e entre as repetições forem semelhantes, podem emergir vias paralelas de enovelamento (225,226). Por outro lado, a redução dos vínculos energéticos entre repetições pode comprometer a cooperatividade, originando subdomínios estruturais (227).

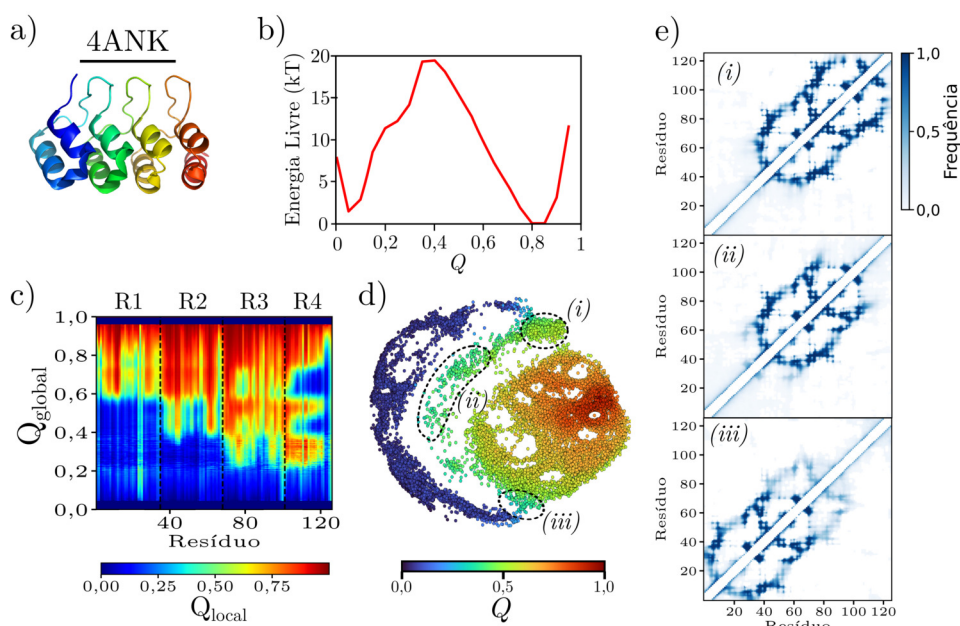
No estudo relatado na referência (65), o ELViM foi empregado para explorar o relevo de energia de proteínas da família ANK constituídas por 3, 4 e 6 repetições. Os dados analisados foram obtidos a partir de simulações realizadas com modelo SBM, utilizando o campo de força AWSEM (228). Para cada proteína, o espaço conformacional foi amostrado por meio de uma simulação clássica (sem viés) e 40 janelas com potencial de *umbrella sampling* centradas em valores igualmente espaçados de  $Q$ . Na Figura F.1 exemplifica-se o resultado obtido para a proteína 4ANK, com quatro repetições.

A Figura F.1.a mostra a estrutura tridimensional da proteína 4ANK (PDB: 1N0R), uma proteína projetada com quatro repetições consensuais idênticas. O perfil de energia livre, obtido ao longo da coordenada  $Q$  através do WHAM (214), é apresentado na Figura F.1.b. Inicialmente, o processo de enovelamento foi mapeado analisando a fração de contatos nativos formada por cada resíduo ( $Q_{local}$ ) em função da fração total de contatos nativos ( $Q_{global}$ ), como ilustrado na Figura F.1.c. Essa análise indicava que o enovelamento

iniciava-se na repetição R4, embora essa formação parecesse se desfazer e ser retomada somente na etapa final, um fenômeno conhecido como *backtracking*.

A projeção ELViM para o sistema é apresentada na Figura F.1.d, com os pontos coloridos com base nos valores de  $Q$ . Três regiões específicas, que ligam as bacias dos estados desnaturado e enovelado, são indicadas delimitadas por linhas pontilhadas. Um mapa de frequências de contatos para cada região é apresentado na Figura F.1.e. De forma interessante, a região (i) apresenta conformações nas quais somente a região carboxi-terminal está formada, enquanto a região (iii) é composta por conformações em que apenas o amino-terminal se encontra formado. Esse resultado sugere a existência de duas rotas distintas de enovelamento, conforme observado por outros métodos de análise e simulação (23). Análises de simulações não enviesadas demonstraram que o enovelamento procede, com maior probabilidade, pelo caminho (iii). Por outro lado, na região (ii), apenas a região central da proteína está formada, um fenômeno descrito por métodos teóricos e experimentais como *fraying*.

Figura F.1 – Análise do enovelamento da proteína 4ANK. (a) Estrutura tridimensional da proteína 4ANK (PDB: 1N0R), uma proteína projetada com quatro repetições consensuais idênticas, representada em *cartoon* com amino-terminal em azul. Cada repetição é formada por um motivo hélice-alça-hélice. (b) Perfil de energia livre obtido através do WHAM (214). (c) Histograma indicando a relação entre os valores globais e locais de  $Q$ . Para um resíduo  $i$ , o valor local  $Q^i$  corresponde à fração dos contatos nativos formados com o resíduo  $i$ . (d) Projeção ELViM obtida para o sistema a partir de simulações de *umbrella sampling*. (e) Mapa de frequência de contatos para regiões indicadas.



Fonte: Adaptada de (65).

## F.2 DISTINGUINDO ROTAS DE FORMAÇÃO DE NÓ TOPOLÓGICO NA PROTEÍNA LEPTINA

Proteínas com nós topológicos exibem um entrelaçamento complexo da cadeia polipeptídica, conferindo-lhes estabilidade e funcionalidades únicas (229,230). A leptina, um hormônio crucial na regulação do gasto energético e da saciedade, exemplifica essa complexidade estrutural (231,232). Sua arquitetura consiste em um feixe de quatro hélices alfa, complementado por uma hélice curta inserida em um loop quase perpendicular ao feixe principal, conforme ilustrado na Figura F.2.a. Estudos experimentais revelaram que, em solução, a leptina adota uma topologia singular, na qual uma ponte dissulfeto, entre a cisteína carboxi-terminal C146 e a C59, forma um laço covalentemente ligado, através do qual as hélices se entrelaçam (231). Essa descoberta levou à identificação de um novo tipo de proteína com nó, denominado *Pierced Lasso Topology* (PLT) (232,233). Após a formação desse laço covalente no estado oxidado, o entrelaçamento pode ocorrer por dois mecanismos distintos, esquematizados na Figura F.2.b: *Plugging* (P), no qual uma extremidade livre atravessa o laço, e *Slipknotting* (S), no qual um resíduo interno cruza o laço (234). Estudos anteriores demonstraram que a leptina se entrelaça através do mecanismo (S) em 99% dos casos (232).

No estudo descrito na referência (66), o ELViM foi utilizado para investigar simulações de enovelamento da leptina, previamente analisadas na referência (232). As simulações foram realizadas com um modelo SBM de resolução  $C_\alpha$ , preparado no SMOG (235) e utilizando o mapa de contatos *Shadow map* (236).

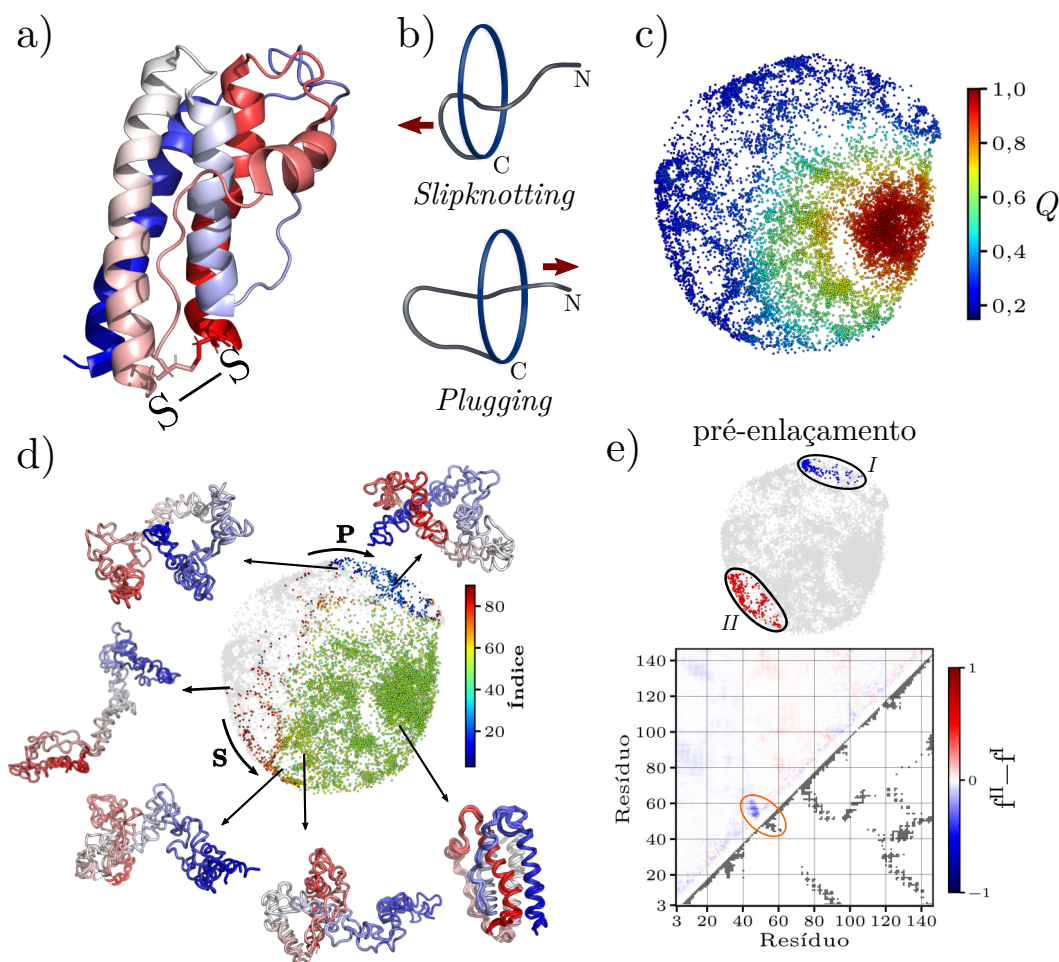
A Figura F.2.c apresenta o espaço de fase efetivo projetado pelo ELViM, colorido de acordo com a fração de contatos nativos. Apesar da complexidade introduzida pelo nó topológico, o ELViM projeta um espaço no qual  $Q$  varia de forma aproximadamente contínua, sem distorções evidentes.

Para investigar as rotas de enovelamento, a Figura F.2.d exhibe novamente a projeção, agora colorida pelo índice do resíduo que cruza a superfície de área mínima definida pelo laço (237). Conformações sem entrelaçamento são mostradas em cinza. Notavelmente, é possível observar uma separação entre os caminhos, que agrupam conformações seguindo os mecanismos (P) e (S). Alguns pontos ruidosos ou mal posicionados, segundo o critério do índice de cruzamento, também estão presentes, mas não comprometem a análise do comportamento efetivo do sistema.

Duas regiões, *I* e *II*, abrangendo as vizinhanças imediatas dos pontos onde o entrelaçamento se inicia foram selecionadas de acordo com o índice de cruzamento. As conformações nessas regiões foram separadas em dois grupos: pré e pós-entrelaçamento. Mapas de frequência de contato foram gerados para cada um desses casos, sendo as conformações pré-entrelaçadas mostradas na Figura F.2.e. Como resultado interessante, ao

comparar as regiões associadas aos mecanismos (P) e (S), verifica-se que um conjunto de contatos nativos — envolvendo os resíduos L45-L49 e K53-L59, destacados em laranja na Figura F.2.e — apresenta frequência significativa no caminho (P), mas não no dominante (S). Essa observação sugere detalhe atômico que pode ser testado em estudos futuros, modificando a intensidade dessas interações para investigar se, de fato, desempenham um papel relevante na probabilidade de ocorrência de cada mecanismo.

Figura F.2 – Análise de rotas de entrelaçamento na proteína leptina. (a) Estrutura tridimensional da leptina obtida por RMN (PDB: 8K6Z, modelo 1), representada em *cartoon* com a extremidade amino-terminal em azul. No estado oxidado, uma ponte dissulfeto entre as cisteínas C59 e C146 (carboxi-terminal) forma um laço covalente através do qual o feixe de hélices se entrelaça. (b) Esquema de dois possíveis caminhos de entrelaçamento. (c) Projeção ELViM para trajetórias de enovelamento obtidas por modelos simplificados (SBM), com pontos coloridos de acordo com os valores de  $Q$ . (d) Projeção ELViM colorida segundo o índice do resíduo que atravessa a superfície de área mínima definida pelo laço covalente. Conformações sem entrelaçamento são mostradas em cinza. Algumas estruturas representativas são fornecidas para exemplificar o processo de entrelaçamento. (e) As regiões *I* e *II* abrangem a vizinhança do início do entrelaçamento, conforme indicado pelo índice de cruzamento. A matriz triangular superior apresenta um mapa da diferença nas frequências de contato entre as duas regiões, enquanto a matriz inferior exibe o mapa dos contatos nativos. Em laranja, destacam-se os contatos nativos que podem favorecer o mecanismo (S) em detrimento da rota dominante (P).



Fonte: Adaptada de (66).

### F.3 ANÁLISE CONFORMACIONAL DOS EFEITOS DE MUTAÇÕES NA PROTEÍNA ELF3

A planta *Arabidopsis thaliana*, amplamente utilizada em estudos de genética e bioquímica, é um organismo chave na investigação dos mecanismos do ciclo circadiano em plantas (238). Esse relógio biológico interno permite que a planta sincronize seus processos fisiológicos com as mudanças ambientais diárias, como os períodos de luz e escuridão. Um dos componentes desse relógio é o Complexo da Noite (EC, *Evening Complex*), composto pelas proteínas ELF3, ELF4 e pelo fator de transcrição LUX ARRYPATH, que regula a expressão de genes essenciais para o ritmo circadiano, florescimento e resposta a sinais ambientais como luz e temperatura (238).

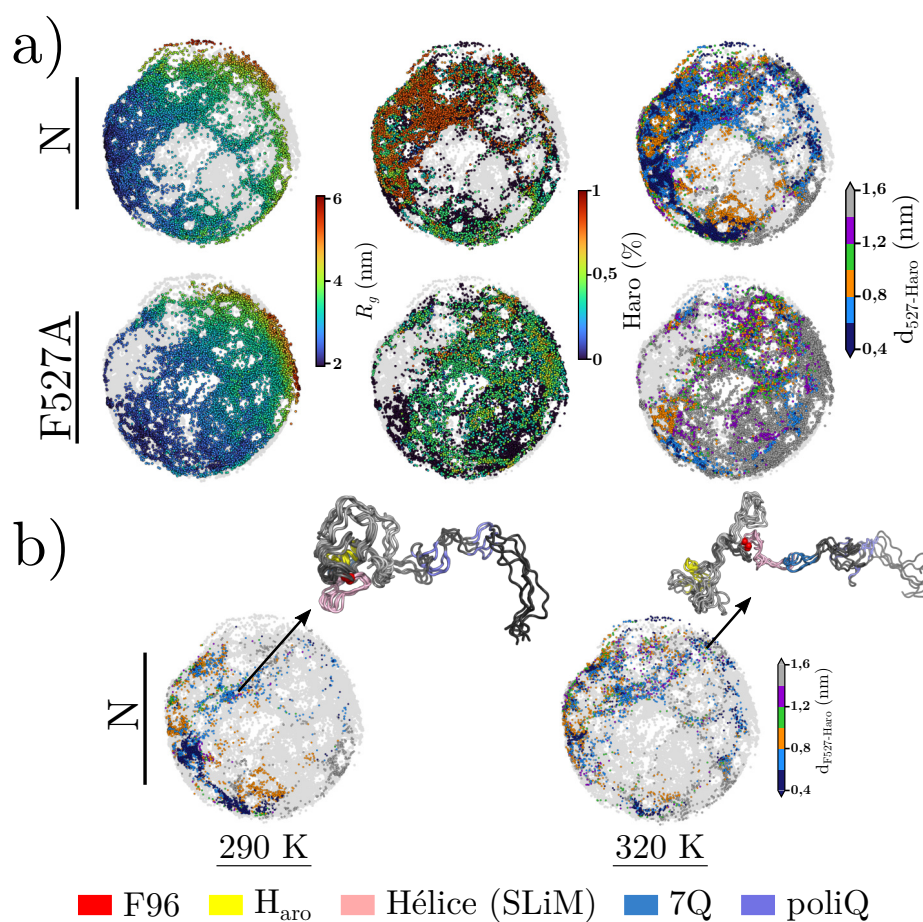
A ELF3 é uma IDP composta por 695 resíduos e apresenta um domínio carboxi-terminal semelhante a príon (PrD), inicialmente identificado por Jung e colaboradores (239). O PrD corresponde a um quarto da sequência total da ELF3 (resíduos 492 a 664) e está associado à formação de condensados moleculares (239). Estudos realizados com a gramínea *Brachypodium distachyon*, na qual a ELF3 não apresenta o domínio PrD, demonstraram que a taxa de crescimento da planta é insensível a elevações de temperatura, sugerindo que o PrD desempenha um papel crucial na sensibilidade à temperatura do EC. A estrutura da ELF3 apresenta três regiões contendo tratos poli-Q, compostos por uma sequência de glutaminas, sendo que uma delas possui comprimento variável em diferentes populações de *Arabidopsis thaliana*. Estudos preliminares de dinâmica conformacional destacaram a importância de uma hélice ( $H_{ARO}$ , resíduos 494 a 504), composta por resíduos aromáticos que estabelecem uma ligação de longo alcance com a fenilalanina F527 (67).

No estudo descrito na referência (67), diversos métodos computacionais foram empregados para investigar os possíveis mecanismos responsáveis pela resposta à temperatura. A Figura F.3 ilustra a aplicação do ELViM para analisar trajetórias do ELF3-PrD, obtidas por REST2 (*Replica Exchange with Solute Tempering*) para as temperaturas de 290, 300, 320 e 405 K. A Figura F.3.a ilustra o espaço de fase efetivo gerado pelo ELViM, que abrange a forma nativa (N) e a mutação F527A, em todas as temperaturas simuladas. As conformações de cada variante são representadas no mapa de cores sobre o espaço global em cinza. Como discutido em outros capítulos desta tese, a análise permite visualizar regiões exclusivamente ocupadas por cada modelo, destacando-se o impacto conformacional da mutação.

Além disso, está indicada, sobre o mesmo espaço, a fração da hélice  $H_{ARO}$  formada, bem como a distância mínima entre o resíduo F527 (ou A527) e a hélice aromática. Ao colorir a projeção ELViM com diferentes CVs, torna-se possível relacionar tais parâmetros com as assinaturas conformacionais que caracterizam cada região da projeção. Duas dessas assinaturas LCSs são indicadas na Figura F.3.b, mostrando como a distância de ligação entre o resíduo F527 e a região aromática é perdida com o aumento da temperatura.

Por fim, destaca-se que outros métodos de simulação e análise também foram explorados ao longo deste trabalho, cujos detalhes podem ser consultados na referência (67).

Figura F.3 – Análise de mutação na proteína ELF3: (a) Espaço conformacional efetivo gerado pelo ELViM, abrangendo a forma nativa (N) e a mutação F527A em diferentes temperaturas. Os pontos foram coloridos com base no raio de giração ( $R_g$ ) (coluna à esquerda), na porcentagem de formação da hélice aromática  $H_{ARO}$  (coluna central) e na distância mínima entre o resíduo F527 (ou A527) e  $H_{ARO}$ . (b) A figura mostra o subespaço da forma nativa amostrado às temperaturas de 290 e 320 K. Duas LCSs são indicadas, ilustrando a perda da interação F527- $H_{ARO}$ .



Fonte: Adaptada de (67).