

Title	Molecules to medicine: advancing integrated computational frameworks for biomedical applications
Authors	Mi, Yanlin
Publication date	2025
Original Citation	Mi, Y. 2025. Molecules to medicine: advancing integrated computational frameworks for biomedical applications. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2025, Yanlin Mi. - https://creativecommons.org/licenses/by/4.0/
Download date	2026-09-25 02:44:11
Item downloaded from	https://hdl.handle.net/10468/18409

Ollscoil na hÉireann, Corcaigh
National University of Ireland, Cork



Molecules to Medicine: Advancing Integrated Computational Frameworks for Biomedical Applications

Thesis presented by

Yanlin Mi

for the degree of

Doctor of Philosophy

University College Cork

School of Computer Science and Information Technology

Head of School: Prof. Utz Roedig

Under supervision of:

Dr. Sabin Tabirca and Prof. Barry A O'Reilly

2025

Contents

Abstract	viii
Acknowledgements	ix
List of Tables	ix
List of Figures	xi
List of Acronyms	xvi
1 Introduction	1
1.1 Translation as Infrastructure	1
1.2 Key Challenges in Biomedical Translation	4
1.2.1 Patient System Translation	6
1.2.2 Data Diagnosis Translation	7
1.2.3 Molecular Clinical Translation	8
1.2.4 Protein Infrastructure Translation	10
1.3 Research Methodology and Computational Framework	12
1.3.1 Design Principles of Translation-as-Infrastructure	12
1.3.1.1 Semantic Continuity	13
1.3.1.2 Structural Modularity	13
1.3.1.3 Interoperability	14
1.3.1.4 Interpretability and Actionability	14
1.3.2 Layered Computational Architecture	15
1.3.3 System Implementation Across Layers	16
1.3.3.1 Patient System Translation Layer	16
1.3.3.2 Data Diagnosis Translation Layer	17
1.3.3.3 Molecular Clinical Translation Layer	17
1.3.3.4 Protein Infrastructure Translation Layer	18
1.3.4 Scope and Limitations	19
1.4 Contributions and Thesis Structure	21
1.4.1 Theory and Framework	22
1.4.2 System Implementation	22
1.4.3 Discussion and Conclusion	24
2 Bridging Patient Narratives to Structured Healthcare Services: The MSaaVS Framework	26
2.1 Abstract	26
2.2 Introduction	27
2.3 Methods	29
2.3.1 Questionnaire System Module	30
2.3.2 Virtual Agent Module	30
2.3.3 Privacy and Data Security	32
2.3.4 Technical Implementation	32
2.3.5 Current Development Status of MSaaVS	33
2.4 Results	34

2.5	Discussion	39
2.5.1	MSaaVS as a Telemedicine SaaS Platform	39
2.5.2	Challenges	40
2.5.3	Outlook	41
2.5.4	Limitations and Considerations in AI-Driven Telemedicine	41
2.6	Conclusion	43
3	Enhancing Patient Comprehension and Data Structuring for Pelvic Health: The PHPlace Platform	44
3.1	Abstract	44
3.2	Introduction	45
3.3	Methods	47
3.3.1	General Architecture	48
3.3.2	Animation of Symptoms and Treatment	49
3.3.3	Questionnaire Design and Automated Scoring System	50
3.3.4	PDF Generation and Patient Management	54
3.3.5	Application Localisation	56
3.4	Results	57
3.5	Discussion and Outlook	59
3.6	Conclusion	60
4	Strengthening Diagnostic Translation: Data Augmentation Strategies for POP Prediction	61
4.1	Abstract	61
4.2	Introduction	62
4.3	Methodology	64
4.3.1	Data Preprocessing	65
4.3.2	Data Augmentation: The Innovations	66
4.3.3	Model Selection and Training	66
4.4	Results	67
4.4.1	Data Exploration and Preprocessing	67
4.4.2	Data Augmentation: The Innovations	67
4.5	Discussion and Conclusion	71
5	Translating Molecular Insights to Clinical Relevance: The Extracellular Matrix Protein Predictor for Pelvic Organ Prolapse System	73
5.1	Abstract	73
5.2	Introduction	74
5.3	Materials and Methods	77
5.3.1	Data Collection and Preprocessing	77
5.3.1.1	ECM and Non-ECM Protein Datasets	77
5.3.1.2	POP-Related Proteomic Data	80
5.3.2	Model Development and Training	80
5.3.2.1	EPOP Model: Fine-Tuning ESM-2	80
5.3.2.2	Baseline Models	82
5.3.3	Prediction and Functional Analysis of POP-Associated ECM Proteins	83
5.3.4	Model Interpretability Analysis	85
5.4	Results	86
5.4.1	Performance Evaluation of ECM Function Prediction Models	86

5.4.2	Prediction Results of POP-Associated ECM Proteins	89
5.4.3	Model Interpretability Analysis	96
5.4.3.1	Single Amino Acid Importance Analysis	96
5.4.3.2	Amino Acid Mutation Effect Analysis	96
5.4.3.3	Sliding Window Analysis	98
5.4.3.4	Aggregated Interpretability Insights Across Multiple ECM Proteins	99
5.5	Discussion	100
5.5.1	Implications for Understanding POP Pathogenesis and ECM Roles . . .	100
5.5.2	Advantages and Challenges of AI-Based Disease Protein Discovery . .	101
5.5.3	Future Directions	103
5.6	Conclusion	105
5.7	Data and Material Availability	106
6	Protein Design Enters the AI Era	108
6.1	Abstract	108
6.2	Landscape of Proteins-AI	109
6.3	Key AI Technologies in Protein Design	114
6.3.1	Deep Learning Architectures	114
6.3.2	Diffusion-Based Generative Models	116
6.3.3	Protein Language Models	118
6.3.4	Knowledge Graph Technologies	120
6.3.5	Integration of Molecular Dynamics and AI	121
6.4	AI-Driven Protein Design Tools	123
6.4.1	AlphaFold 2 Architecture and Implementation	123
6.4.2	RoseTTAFold Technical Implementation	124
6.4.3	ESM-2 System Architecture	125
6.4.4	Comprehensive Ecosystem of AI Tools	126
6.5	Applications and Impact	126
6.5.1	Therapeutic Protein Development	126
6.5.2	Food Protein Engineering	128
6.5.3	Industrial Applications	129
6.5.4	Future Market Trajectory and Implementation Challenges	130
6.6	Challenges and Future Directions	131
6.6.1	Technical Challenges	131
6.6.2	Future Directions	134
6.7	Conclusion	135
7	Protein Fragment and Structure Analysis Workstation	137
7.1	Abstract	137
7.2	Introduction	138
7.2.1	Myriad of Computational Tools for Tools	138
7.2.2	One tool to rule them all	140
7.3	Materials and Methods	141
7.3.1	Architecture design	141
7.3.2	Front end design	143
7.3.3	Back end design	147
7.4	Result	149

7.5	Discussion and Outlook	154
7.5.1	As a single UI protein workstation	154
7.5.2	Bottlenecks and challenges for PROFASA	157
7.5.3	Ambitions and future outlook	157
7.6	Data and Software Availability	158
8	PS-GO Parametric Protein Search Engine	160
8.1	Abstract	160
8.2	Introduction	161
8.2.1	Topology of protein landscape	161
8.2.2	Parametric protein design	163
8.2.3	Challenges and Bottlenecks in Searching the Protein Space	166
8.2.4	PS-GO - Transforming Protein Exploration with a Breakthrough Search Method	168
8.2.5	Parametric Search Strategy	170
8.3	Methods	171
8.3.1	Parametric Search Algorithm	171
8.3.2	Protein Database Construction	175
8.3.3	Silver Surfer	177
8.4	Results	179
8.4.1	Key Functionalities and User Interface of PS-GO	179
8.4.2	Advantages and Potential Impact of PS-GO	182
8.4.3	Performance Evaluation	184
8.5	Discussion and Outlook	185
9	Generative Protein Sequence Design via Parametric Optimization: The Silver Surfer Approach	190
9.1	Abstract	190
9.2	Introduction	191
9.2.1	Protein Engineering	191
9.2.2	AI Predictors	192
9.2.3	Our Contribution	192
9.3	Materials and Methods	193
9.3.1	Silver Surfer	193
9.3.2	Implementaiton	194
9.3.3	Protein Search	195
9.3.4	PS-GO	198
9.4	Results	198
9.4.1	Validating the Correctness of the Protein Properties	199
9.4.2	Parametric Search	199
9.5	Conclusions and Further Research	201
10	Building Foundational Protein Infrastructure: The ProteinFlow Framework	203
10.1	Abstract	203
10.2	Introduction	204
10.3	Methods	206
10.3.1	Data Collection	206
10.3.1.1	Multi-Source Data Integration	206
10.3.1.2	Data Retrieval and Validation Strategies	207

10.3.2	Data Preprocessing	208
10.3.2.1	Data Cleaning	208
10.3.2.2	Data Conversion	210
10.3.2.3	Feature Extraction	210
10.3.3	Database Construction	214
10.3.4	Flow Optimization	216
10.4	Results	216
10.4.1	Efficiency in Data Collection and Preprocessing	216
10.4.2	Advanced Feature Extraction and Data Integration	217
10.4.3	Scalability and Performance	217
10.4.4	Application of ProteinFlow	219
10.5	Discussion and Conclusion	219
11	Discussion	221
11.1	Limitations	225
11.1.1	Semantic alignment	225
11.1.2	Causal inference capabilities	226
11.1.3	Automated linkage between systems	226
11.1.4	Domain generalization and adaptability	226
11.1.5	Non-computational factors	227
11.2	Future Work	227
12	Conclusion	231
	Appendices	233
A	Clinical Research Ethics Committee of the Cork Teaching Hospitals	234
B	PROFASA User Guide	236
C	PS-GO User Manual	239
D	Other Published Works during PhD	243
	References	245

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism and intellectual property.

Abstract

With the increasing complexity of biomedical knowledge and its interdisciplinary nature, effective translation of knowledge faces significant challenges. From molecular discovery to therapeutic application, it is challenging to ensure conceptual continuity and operational consistency due to the fragmented and domain-specific nature of current technology. In this thesis, we propose a unified computational framework, Translation-as-Infrastructure, to systematically support the stable migration of biomedical knowledge across cognitive levels, data forms, and operating systems.

Guided by the principles of semantic fidelity, structural unity, and output interpretability, the framework addresses four major translational bottlenecks: translating patient narratives into structured system logic, translating heterogeneous clinical data into trustworthy diagnostic reasoning, translating molecular insights into clinically actionable knowledge, and translating protein properties into scalable, computable representations. To operationalize this framework, some specific systems and computational tools have been developed and validated, including patient service platforms (MSaaVS, PHPlace), data modeling systems (Strengthening Diagnostic Translation), cross-level inference models (EPOP, AI-driven protein function analysis), and a modular protein computing infrastructure (PROFASA, PS-GO, Silver Surfer, ProteinFlow).

By integrating these systems, the thesis constructs a continuous, multi-layered translational pipeline from molecules to medicine. While this work represents an initial foundational effort, primarily focusing on structured knowledge migration rather than fully addressing semantic transparency, causal inference, and dynamic system self-adaptation, it provides a scalable architecture for future development of adaptive biomedical knowledge systems. By rethinking translation as a computational and infrastructural mechanism, this study develops a new paradigm for multidisciplinary knowledge engineering, providing methodological improvements and practical tools to assist the next generation of biomedical informatics and translational research.

Acknowledgements

First and foremost, I would like to express my most profound gratitude to my supervisor, Dr. Sabin Tabirca. His unwavering support, timely guidance, and encouragement have been the cornerstone of my academic journey. Throughout these years, he has consistently believed in my work, helped me build confidence during moments of self-doubt, and offered generous, insightful feedback whenever I needed it. His dedication to his students, compassion, and integrity have profoundly motivated me. They have shaped not only my academic pursuits but also my personal aspirations. I hope to become a mentor who genuinely cares for and empowers their students, just as he has done for me.

I would also like to sincerely thank Prof. Barry A. O'Reilly, my second supervisor, whose clinical expertise and generous support greatly enriched my research. As a practicing clinician and academic, he provided not only crucial domain knowledge but also access to clinical data, without which this work would not have been possible. His practical insights bridged the gap between theoretical modeling and real-world application, helping me maintain a strong clinical relevance throughout my study.

I am also grateful to Dr. Venkata Vamsi Bharadwaj Yallapragada, my collaborator and friend. From the very beginning of my studies in biological sciences, he has been a source of patience, encouragement, and unwavering support. He actively shared his expertise with me, taught me how to be a researcher and believed in my abilities, which provided a solid foundation for my multidisciplinary development.

I would like to extend my appreciation to Ms. Janet Choi, the manager of CRT-AI, whose kindness, wisdom, and dedication made even the most challenging periods feel lighter and more hopeful. More than just a manager, she is the "mother" of the CRT-AI community, tirelessly creating a nurturing environment where we could learn, connect, and grow beyond academic boundaries. I am also grateful for the financial support provided by Taighde Éireann–Research Ireland under Grant No. 18/CRT/6223, without which this research would not have been possible.

Finally, my parents deserve my sincere gratitude. My development and fortitude have been built on their unwavering love, unwavering support, and silent strength. Their steadfast faith in me has allowed me the flexibility to follow my goals, make my own decisions, and work toward being the person I want to be. Their presence continues to be my biggest source of inspiration and strength, no matter where life leads me.

This four-year PhD journey has not only been an academic experience, but it has also had a significant impact on who I am now. I have become more confident, more independent, and more resilient. I have learned to think critically, to accept uncertainty, and to persevere in the face of difficulties. Every moment of staying up late and struggling, every tiny victory out of doubt, and every person I met along the way has left a deep mark on my heart. I am sincerely grateful for all of this, and for those who have always accompanied and supported me.

List of Tables

1.1	Mapping of Translational Layers, Computational Modules, and Thesis Chapters	21
2.1	MSaaVS Usability Metrics	38
3.1	Questionnaires for POP	52
4.1	Optimized Model Performance After Hyperparameter Tuning	68
5.1	Performance comparison of Extracellular Matrix (ECM) function prediction models on the test set.	87
5.3	Proteins not explicitly associated with Pelvic Organ Prolapse (POP) in the provided literature. Functional annotations were adapted from UniProtKB (“UniProt: the universal protein knowledgebase in 2025”).	91
5.2	Cross-validation of Literature Can Support Proteins.	107
6.1	Overview of Recent Major AI-Driven Drug Discovery Partnerships . . .	111
6.2	Comparative Overview of Key Deep Learning Architectures in Protein Design	117
6.3	Commercial Artificial Intelligence (AI) Platforms for Protein Design . .	123
6.4	Overview of Additional AI-Driven Protein Design Tools	126
7.1	Model generation timetable	141
7.2	F2F plot parameters	148
7.3	MaxASA of amino acids	149
8.1	Computable parameters and description	165
8.2	The genetic algorithm employed by Silver Surfer	180
8.3	Comparison of features and characteristics between PS-GO and other protein search methods	183
8.4	The average performance of PS-GO in handling different tasks	184
9.1	Evaluation of the Instability Index, GRAVY Score, Monoisotopic mass and Isoelectric Point properties for different proteins	200

List of Tables

10.1 Comparison of data retrieval time and data quality metrics.	217
10.2 Comparative Analysis of Feature Extraction Performance	218

List of Figures

1.1	Structural Faultlines in Biomedical Translation	5
2.1	Overview of the MSaaVS platform	29
2.2	Virtual agent interaction in Medical Software as a Virtual Service (MSaaVS)	31
2.3	Question set building interface and logic editor interface in MSaaVS . .	36
2.4	Virtual doctor interface in MSaaVS	36
2.5	Virtual Agent Trial Results: Patient Interaction Metrics	37
2.6	Key features and benefits of the MSaaVS platform	39
3.1	PHPlace features	47
3.2	The first and last frames of the animation	50
3.3	Multiple questionnaire score visualisation chart	53
3.4	Example of the generated pdf showing the visualisation of the animation	55
3.5	Patient management UI	56
4.1	Enhancing the performance of ML model for POP prediction via data augmentation	63
4.2	A boxplot for each variable in the dataframe visualizes the distribution of the data, showing the minimum, first quartile, median, third quartile, and maximum values, along with potential outliers.	68
4.3	Determine the optimal number of clusters by identifying the "elbow" point where adding more clusters does not significantly improve the model.	69
4.4	Average cross-validation scores of each model under different data en- hancement techniques.	70
5.1	AI-driven ECM protein prediction in POP. The figure illustrates the key components of our analytical framework. Left panel lists six key features of the approach: Extracellular Matrix, AI-Driven, Systematic Prediction, Mechanism Elucidation, Experimental Validation, and Novel Perspective. Center shows a protein structure visualization with multiple domains in different colors. Right panel displays five key analytical outputs: Evolutionary Scale Modeling 2 (ESM-2) Model Prediction, Differential Expression, Key Sites & Mutations, POP Proteomics Data, and Diagnostic & Therapeutic Insights, each represented by a distinctive icon and box style.	76

5.2	Workflow overview showing the four main steps of the Extracellular Matrix Protein Predictor (EPOP) methodology. Step 1 (Data Collection and Preprocessing) shows the retrieval of data from Gene Ontology (GO) :0031012 and UniProt. Initially, this resulted in raw datasets of 40,155 ECM-related proteins and 46,427 non-ECM proteins. These were subsequently cleaned, balanced to 40,000 ECM and 40,000 non-ECM sequences, and then split into a training set (48,000 sequences), a validation set (16,000 sequences), and a test set (16,000 sequences) for model training. Additionally, proteomic data from PRIDE project ID PXD011467, comprising 22,341 Peptide-Spectrum Matches (PSMs) from patient samples, was utilized for external validation and application. Step 2 (Model Development and Training) illustrates three parallel models: EPOP (fine-tuning of pre-trained ESM-2), Transformer Architecture Model (self-attention mechanism), and Long Short-Term Memory (LSTM) Model (basic neural network). Step 3 (Prediction and Analysis) shows EPOP's input-output process. Specifically, the trained EPOP model was applied to the PRIDE proteomic data, identifying 40 unique UniProt IDs of proteins predicted as ECM based on the optimal classification threshold, which will be further analyzed and validated against existing literature. Step 4 (Model Interpretability Analysis) details three analysis approaches: Single Amino Acid Importance Analysis, Amino Acid Substitution Impact Assessment, and Sliding Window Analysis, which were applied to the high-scoring peptides to identify key functional regions and residues.	78
5.3	Quality control metrics for POP proteomic data. A: Peptide length distribution histogram showing the frequency distribution of peptide lengths, with peak frequency around 10-15 amino acids. B: Scatter plot showing the relationship between calculated m/z values (x-axis, 400-1800) and -log10 Percolator Posterior Error Probability (PEP) scores (y-axis, 0-14), displaying a broad distribution of peptide identifications. C: Scatter plot demonstrating the correlation between Mascot scores (x-axis, log scale) and Percolator PEP values (y-axis, log scale), showing a strong negative correlation.	81

List of Figures

5.4	This figure presents a detailed evaluation of EPOP’s performance on the test set, combining core classification metrics with a thorough analysis of threshold effects. A: Precision-Recall (PR) curve (AUPRC = 0.999) demonstrating near-perfect performance with high precision maintained across most recall values. B: Receiver Operating Characteristic (ROC) curve (AUROC = 0.999) illustrating excellent classification performance, substantially outperforming random classification. C: Score distribution by class, showing distinct prediction score distributions for ECM and non-ECM proteins. D: Confusion matrix at the optimal classification threshold (0.168), providing a detailed breakdown of classification outcomes. E: Performance metrics (Accuracy, F1-score, Precision, Recall) across varying classification thresholds, highlighting the optimal threshold (0.168) that maximizes F1-score. F: Cumulative score distributions for ECM and non-ECM proteins, visually demonstrating the model’s high confidence and clear class separation.	88
5.5	Key aspects of EPOP’s protein identification and validation. A: Top functional enrichment terms identified among the 40 predicted ECM proteins. B: Distribution of EPOP prediction confidence for the identified 40 proteins, highlighting the confidence levels of both literature-supported and novel candidate proteins.	89
5.6	Interpretability Analysis of Collagen Alpha-1(VI) Chain (P12109) and Aggregated Insights Across Multiple ECM Proteins. A: Amino acid importance score plot showing varying contributions of different residues, with peaks indicating highly important positions. B: Mutation effects heatmap displaying the impact of amino acid substitutions at each position, with color intensity representing effect magnitude (blue to red scale). C: Functional region score plot showing the prediction scores across different sequence windows, revealing regions of particular functional importance. D: Overall Amino Acid Importance. Average importance score for each amino acid type, aggregated across all positions in all 16 proteins. E: Overall Amino Acid Mutation Sensitivity. Average mutation effect for each amino acid type, aggregated across all positions in all 16 proteins.	97
6.1	A comprehensive overview of AI-driven protein design	113
6.2	Three deep learning architectures used in protein design	115
6.3	The conceptual parallel between natural language models and protein language models	119
7.1	PROFASA, combines all the major computational tools and provides a single platform interface for computational protein works.	140

List of Figures

7.2	The number of parameters per tool	142
7.3	PROFASA key features	142
7.4	PROFASA architectur	143
7.5	Ramachandran Plot and F2F Plot	146
7.6	Database design	150
7.7	PROFASA flowchart	150
7.8	Analysis user interfac	151
7.9	RMSD and GDT-TS calculations are used to compare the differences between two superimposed 3D structures	152
7.10	Structure blast	153
7.11	PROFASA functional classification	156
7.12	Ambitions and future outlook	158
8.1	The steps in parametric protein design.	162
8.2	Issues and Challenges of Existing Protein Search Methods.	166
8.3	How PS-GO search works and its key features.	169
8.4	PS-GO protein search engine	181
8.5	Structural visualization of protein models and graphical representation of parameters using PS-GO and PROFASA.	182
8.6	Future development outlook of PS-GO	187
9.1	Silver Surfer stages for parametric protein search	196
9.2	Mutation search process across 4-length protein space	202
10.1	ProteinFlow: A Novel Method for Restructuring Protein Databases . . .	209
10.2	ProteinFlow Data Statistics (PDB and UniProt Data)	215
A.1	Ethical approval letter from the Clinical Research Ethics Committee (CREC) of the Cork Teaching Hospitals for the PHPlace platform. . . .	235
C.1	PS-GO User Manual Step 1	239
C.2	PS-GO User Manual Step 2	240
C.3	PS-GO User Manual Step 3	241
C.4	PS-GO User Manual Step 5	241
C.5	PS-GO User Manual Step 6	242

List of Acronyms

2D	Two Dimensional
3D	Three Dimensional
AES	Advanced Encryption Standard
AI	Artificial Intelligence
API	Application Programming Interface
AR	Augmented Reality
ASA	Accessible Surface Area
CAGR	Compound Annual Growth Rate
CDD	Conserved Domain Database
CREC	Clinical Research Ethics Committee
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DSSP	Dictionary of Secondary Structure of Proteins
DL	Deep Learning
DTR	Decision Tree Regression
ECM	Extracellular Matrix
EHR	Electronic Health Records
EPOP	Extracellular Matrix Protein Predictor
ESM-2	Evolutionary Scale Modeling 2
F2F	Function2Form
FDR	False Discovery Rate
GDT_TS	Global Distance Test - Total Score
GDT_PN	Global Distance Test - Percent Nativeness

List of Acronyms

GNN	Graph Neural Network
GO	Gene Ontology
GPU	Graphics Processing Unit
GRAVY	Grand Average of Hydropathicity
HCI	Human-Computer Interaction
HIPAA	Health Insurance Portability and Accountability Act
ICIQ-UI	International Consultation on Incontinence Questionnaire-Urinary Incontinence Short Form
IoT	Internet of Things
KHQ	King's Health Questionnaire
kNN	k-Nearest Neighbors
LFQ	Label-Free Quantification
LSTM	Long Short-Term Memory
ML	Machine Learning
MSaaS	Medical Software as a Virtual Service
MSA	Multiple Sequence Alignments
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
NLP	Natural Language Processing
ORM	Object Relational Mapping
PDB	Protein Data Bank
PEP	Posterior Error Probability
PFDI-20	Pelvic Floor Distress Inventory-20
PFIQ-7	Pelvic Floor Impact Questionnaire-7
PGI	Patient Global Impression of Improvement

List of Figures

PGI-C	Patient Global Impression of Change
PGI-I	Patient Global Impression of Improvement
PGI-S	Patient Global Impression of Severity
PHPlace	Pelvic Health Place
PR	Precision-Recall
PROFASA	Protein Fragment and Structure Analysis
POP	Pelvic Organ Prolapse
POP-Q	Pelvic Organ Prolapse Quantification
PISQ-12	Pelvic Organ Prolapse/Urinary Incontinence Sexual Function Questionnaire
PS-GO	Parametric Search, GO!
PSMs	Peptide-Spectrum Matches
Q&A	Questions and Answers
RAM	Random Access Memory
RFR	Random Forest Regression
RMSD	Root-mean-square Deviation
ROC	Receiver Operating Characteristic
SaaS	Software-as-a-Service
SCM	Structural Causal Models
SSL	Secure Sockets Layer
SSD	Solid State Drive
SVR	Support Vector Regression
TLS	Transport Layer Security
UI	User Interface
US	United States of America

List of Acronyms

UX	User Experience
XAI	Explainable Artificial Intelligence

Contents

1 Introduction

1.1 Translation as Infrastructure

From the broadest perspective, translation is more than just a conversion between languages. It is increasingly understood as a fundamental cognitive and operational mechanism (K. Bennett 2024; Gutt 2014; Forsman and Solitander 2003). As Paul Ricoeur said, "Understanding is translation" reveals the complex entanglement and negotiation between meaning, medium and context (K. Bennett 2024; Ricoeur 2007). From this perspective, translation is no longer just a conversion of words, but a generative process. In other words, meaning is negotiated and reconstructed between different systems, actors and symbolic frameworks. Every translation involves the faithful transmission of information and the structural reshaping in the context, constituting the dual task of semantic fidelity and functional integration (Gutt 2014; Forsman and Solitander 2003). As a result, translation becomes the mechanism for the flow of knowledge across conceptual, disciplinary and institutional boundaries, supporting mutual understanding and knowledge coordination between human and artificial systems (Drolet and Lorenzi 2011; L. Xue et al. 2020).

Although translation aims to achieve semantic fidelity, the process is highly vulnerable to disruption. In different symbolic systems and cultural contexts, translation behaviour is often accompanied by semantic distortion, ambiguity or loss. These translation failures can be roughly divided into three categories: **misrepresentation**, that is, the original meaning is changed due to the lack of context; **overtranslation**, that is, the addition of an interpretation layer beyond the original intention; and **untranslatability**, that is, the concept corresponding to the source domain does not exist in the target system (K. Bennett 2024; Gutt 2014). The common point of these forms of failure is that they undermine the cognitive consistency and structural docking pursued by translation. Moreover, these deviations are not simply technical issues. They will

1 Introduction

profoundly affect the epistemological structure of knowledge and interfere with inter-subjective understanding and application in practice (Plebani 2016; L. Xue et al. 2020). Therefore, maintaining semantic integrity in the translation process is not a marginal technical issue, but a basic premise for the circulation and operation of knowledge between systems.

In the biomedical field, the translation challenge is even more apparent. The translation must cross multiple heterogeneous levels such as molecular biology, clinical practice, regulatory systems and patient experience (Drolet and Lorenzi 2011; Rajesh Krishna, Schaefer, and Bjerrum 2007). Each level has unique representation logic, terminology system and verification standards. Therefore, translation between levels is not just a transformation of information, but also requires the negotiation of meaning between different knowledge cultures. For example, molecular discoveries must be translated into clinically actionable solutions; the outputs of statistical models must be understood and trusted by physicians; and patients' subjective narratives must be encoded into standardized electronic health records. In these contexts, translation means connecting worlds with very different symbolic systems and operational logics. (Forsman and Solitander 2003; L. Xue et al. 2020) Unlike isolated language translation, translation in biomedicine is structurally embedded and must continuously bridge the differences in goals, timescales, and worldviews between different contexts.

It is precisely because of the above structural differences that biomedical translation often has specific discontinuities in reality, which damages the integrity of science and weakens clinical effectiveness. For example, when complex research results are translated into clinical guidelines, key contextual details may be ignored, resulting in oversimplified intervention plans (K. Bennett 2024). Health policies are often misinterpreted due to different interpretive frameworks during cross-institutional or cross-regional implementation, weakening their due influence (Solano-Flores, Backhoff, and Contreras-Niño 2009). When patients' subjective narratives are included in the standardized scope of electronic health records, they often lose their semantic richness and emotional tension (Lord and Gale 2014; M. Quinn et al. 2019). These failures are not minor deviations, but profound epistemological distortions that hinder communication and mislead practice, and sometimes even directly harm patients themselves. Facing these translation bottlenecks means we must view "translation" as a structural mechanism embedded in medical practice rather than a post-process technical fix (K. Bennett 2024; Plebani 2016).

1.1 Translation as Infrastructure

In recent years, the rapid development of AI has expanded the possible boundaries of "translation" as a systemic mechanism. Natural Language Processing (NLP) systems, represented by Google Translate, have demonstrated exceptional capabilities in handling large-scale semantic ambiguity and syntactic complexity (James 2023). More groundbreaking is that AI systems in scientific computing, such as AlphaFold 2, have achieved a "structural translation". In other words, it can convert linear amino acid sequences into reasonable Three Dimensional (3D) protein conformations (James 2023). These results show that translation across representation systems is not only possible, but also computationally feasible. However, these systems also reveal a new translation bottleneck. Many AI models operate as "black boxes" that can output statistically reliable results but lack semantic transparency (H. Xu and Shuttleworth 2024). Clinicians may be reserved about their predictions because they cannot understand their internal logic. At the same time, these models rely on highly structured data inputs screened by domain knowledge to be effective (Abgrall et al. 2024). Therefore, although AI expands the ability to translate across symbolic systems, it also exposes the structural deficiencies of existing computing infrastructure in terms of interpretability, bidirectional integration, and semantic traceability (P. Liu et al. 2021; Abgrall et al. 2024).

In the face of the above challenges, we must re-understand the concept of translation. Translation should not be seen as a marginal language-mediated activity, but should be recognized as a core infrastructure mechanism that can achieve the crossing, alignment and operationalization of meaning between heterogeneous fields. This study regards **computational translation** as an inevitable proposition at the computational level, rather than a simple metaphorical expression. Its fundamental task is to explore how to achieve cognitive consistency and semantic docking of representational structures, knowledge organization and practical decisions through systematic design. In other words, translation must be computable. The translation is embedded in a formalized architecture that supports semantic integrity, two-way understandability and knowledge traceability (K. Bennett 2024; Abgrall et al. 2024). The following chapter will analyze the most prominent structural translation rupture in the biomedical system, laying the foundation for the hierarchical and interoperable computational framework proposed in this study.

1.2 Key Challenges in Biomedical Translation

In the contemporary biomedical knowledge ecology, translation is no longer a simple act of information transmission, but a multi-level cognitive and structural process that determines how meaning is transferred and reconstructed between different fields, modalities and institutional systems (Liverman et al. 2013; Nagarajan et al. 2015). The present translation failure in biomedicine involves more than just a loss of data or a semantic misunderstanding, as was indicated in Chapter 1.1. The underlying cause is the absence of efficient structural support systems across diverse knowledge systems, which makes it challenging to guarantee semantic accuracy and operational logic continuity (Vodovotz and G. An 2014). Although ethical, institutional and policy barriers are equally important, this chapter will focus on translation bottlenecks that can be structured and analyzed and have potential computational modeling value (Flier and Loscalzo 2017).

In order to establish a systematic observation framework, this thesis identifies four structural breakpoints throughout the biomedical knowledge migration path, from molecular discovery to clinical implementation, from patient experience to system response. They are:

- **Patient System Translation**

Translating patient needs into service logic

- **Data Diagnosis Translation**

Translating clinical data into diagnostic reasoning

- **Molecular Clinical Translation**

Translating molecular insights into clinical meaning

- **Protein Infrastructure Translation**

Translating protein properties into computable representations

These four layers are summarized in Figure 1.1. Although analytically independent, these four types of challenges are highly coupled in practice and together form a continuous chain of knowledge transformation, whose breaks seriously weaken the coherence, accessibility, and responsiveness of biomedical systems (Cannataro et al. 2022; Q. Chen 2024). The following sections will analyze the structural failures of these four levels one

1.2 Key Challenges in Biomedical Translation

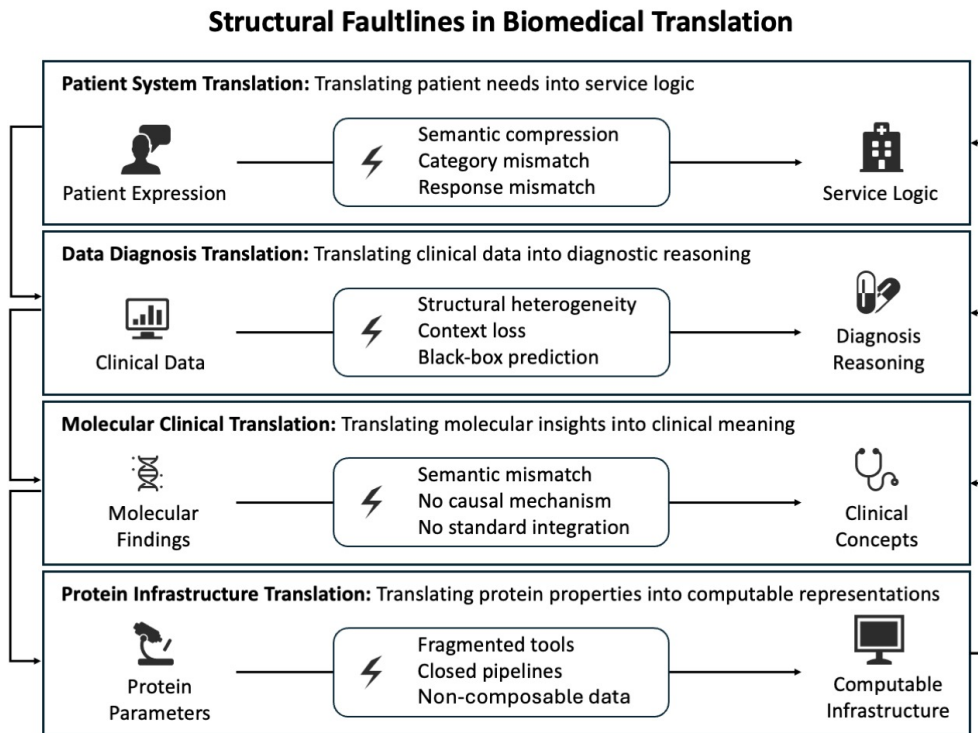


Figure 1.1: It illustrates four major layers of translational breakdown across the biomedical knowledge flow: (1) Patient System Translation, where subjective needs are compressed into rigid system logic; (2) Data Diagnosis Translation, where heterogeneous clinical data fails to integrate into coherent reasoning due to semantic opacity; (3) Molecular Clinical Translation, where statistically derived molecular findings remain disconnected from clinical concepts; and (4) Protein Infrastructure Translation, where protein-level information cannot be reused or composed computationally due to fragmented tool ecosystems. The arrows between layers indicate cross-layer dependencies, highlighting how upstream semantic misalignments cascade into downstream interpretability and implementation challenges.

1 Introduction

by one, laying the theoretical and practical foundation for the integrated computational framework proposed in Chapter 1.3.

1.2.1 Patient System Translation

The first critical challenge in biomedical translation arises at the interface between patient expression and healthcare system logic. Patients often convey their needs through multifaceted narratives, emotions, physical reactions, and sociocultural contexts (Kerosuo 2010; Heath and Porter 2019). In stark contrast, healthcare systems operate on standardized terminology, predefined diagnostic categories, rigid operating procedures, and structured management processes. This fundamental ontological divergence leads to a structural translation breakdown: the rich, subjective, and context-dependent patient experience is forcibly compressed into a simplified, often reductive, format for system input (Côté 2024; Jordan et al. 2009; Heath and Porter 2019).

This semantic disparity fundamentally limits the system’s responsiveness to patient narratives, often reducing rich, contextual information into sterile system inputs. As highlighted by Kerosuo, patient expressions are frequently recorded from a professional and organizational perspective, making it challenging to restore their original semantic integrity (Kerosuo 2010). Similarly, Pedersen emphasizes that patient narratives are not merely information carriers but symbolic devices capable of revealing systemic blind spots and inspiring local structural innovations (Pedersen 2016). Yet, current medical systems often prioritize formalism, overlooking the profound cognitive value embedded within patient experiences.

Translational failures in this domain primarily manifest in three forms: semantic compression, where complex patient narratives are oversimplified into limited structured fields within Electronic Health Records (EHR), leading to significant loss of contextual information (S. B. Johnson et al. 2008); category mismatch, where standardized classification systems struggle to accurately capture the true intent or nuance of patients’ subjective symptoms (Stausberg et al. 2008); and response mismatch, where despite formal process completion, the system’s intervention logic fails to genuinely address the patient’s actual needs (Cahan et al. 2019).

These issues are particularly acute for vulnerable populations, including the elderly, non-native speakers, chronic disease patients, and individuals with cognitive impair-

1.2 Key Challenges in Biomedical Translation

ments. Their expressions tend to be more ambiguous, variable, and resistant to standard classification, frequently resulting in their marginalization or systemic neglect (Terui 2017; Memon et al. 2016). Consequently, patients often bear the burden of repeated, cross-level translation of their needs across multiple visits to maintain care continuity (Kerosuo 2010).

Despite localized efforts to enhance patient expression, such as symptom self-assessment questionnaires, shared decision-making tools, and AI-based conversational systems, these approaches remain fragmented. They are often difficult to systematically integrate into the core logic of healthcare operations (Desomer et al. 2018; Consolo et al. 2023). Without an organized and semantically sensitive translation mechanism, the persistent patient-system translation gap remains a major barrier to the widespread adoption of experience-oriented services, equitable healthcare, and precision medicine. This foundational mismatch hinders not only individualized care delivery but also the development of responsive, human-centered medical systems. Without resolving this layer, upstream information flow remains fundamentally compromised.

1.2.2 Data Diagnosis Translation

The second fundamental challenge in biomedical translation lies in effectively transforming raw clinical data into explanatory, credible, and actionable diagnostic reasoning. Ideally, multi-dimensional information—ranging from physical signs, laboratory tests, and imaging to lifestyle data—should integrate seamlessly to support continuous medical judgment. However, in practice, data structure heterogeneity, semantic opacity, and a cognitive mismatch between algorithmic models and human clinicians create multi-level obstacles in this translational process (Q. Cai et al. 2019; X. Xu et al. 2024; Moya 2024; Reyero Lobo et al. 2023; Lo 2024).

Structural heterogeneity represents a primary impediment. Clinical data spans diverse formats (A. Ahmed et al. 2023; Sarwar et al. 2022; I. Li et al. 2022): structured indicators (e.g., vital signs), semi-structured content (e.g., medical record summaries), unstructured text (e.g., physician’s notes), and multimodal sources (e.g., images, genomic sequences). This fragmented information typically lacks a unified framework, making comprehensive cross-modal integration within diagnostic systems exceptionally difficult.

1 Introduction

Equally critical is the lack of semantic context. Current health information systems predominantly focus on numerical data collection, often neglecting the underlying generation background, clinical context, and practical significance of these values (Gubert, Costa, and Righi 2020; A. Y. Lin et al. 2023). For instance, the diagnostic significance of mild fever differs profoundly between postoperative monitoring and chronic infection screening, a semantic distinction most models fail to discern (Pile 2006).

Furthermore, many algorithmic models exhibit considerable semantic opacity. Despite their statistical validity, the absence of traceable logical paths and transparent causal chains undermines clinicians' ability to establish trust in their predictions (T. P. Quinn et al. 2022; Durán and Jongsma 2021). This "diagnostic black box" phenomenon not only weakens the foundation of medical reasoning but also complicates ethical review and safety verification in real-world clinical environments.

While advancements in interpretable modeling methods—such as multimodal fusion, knowledge graphs, and attention mechanisms—have been noted (Holzinger, Malle, et al. 2021; Soenksen et al. 2022; Zongbao Yang et al. 2023), these approaches frequently remain at the level of technical optimization. They often fail to fully align with the intricate clinical semantic structures and reasoning processes, making it challenging to completely restore the reasoning chain from data acquisition to medical decision-making (Kline et al. 2022; G. Huang et al. 2024; Hossain et al. 2025).

In essence, this translational failure reflects a profound discontinuity between the data system's symbolic structure and the logical structure of clinical cognition. Achieving accurate diagnostic translation necessitates not only robust data processing capabilities but also the construction of context-sensitive systems that possess semantic preservation, causal traceability, and alignment with human reasoning paradigms. Without bridging this gap, data-driven models will continue to suffer from a lack of clinical trust and utility, thereby limiting their role in real-world diagnostic workflows.

1.2.3 Molecular Clinical Translation

The third translational challenge resides in bridging molecular research findings with clinical semantics and actionable decisions. Despite rapid advancements in genomics, proteomics, metabolomics, and related fields, a significant portion of molecular-level information remains confined to the research domain, proving difficult to systematically

1.2 Key Challenges in Biomedical Translation

transform into actionable bases for clinical judgment (Chongyang Chen et al. 2023; Tong et al. 2023; Correa-Aguila, Alonso-Pupo, and Hernández-Rodríguez 2022).

A fundamental obstacle is the mismatch between semantic structure and representation. Molecular data are typically expressed in abstract forms, such as vectors, matrices, or sequences, whereas clinical language centers on symptoms, observable indicators, and established diagnostic pathways (Vitorino 2024; Zhiqiang Zhang et al. 2019). Without an intermediary system for semantic alignment, this inherent "language distance" renders effective translation unattainable.

More critically, most molecular studies primarily focus on correlation identification rather than mechanistic elucidation. For example, a protein might be detected as up-regulated in a specific diseased tissue and subsequently labeled a potential biomarker. However, this statistical correlation often lacks explanatory power, failing to directly guide clinicians on precise interventional strategies (Rosati et al. 2024). This gap between statistical association and causal understanding severely limits clinical utility.

Furthermore, molecular research outcomes often lack a structured embedding mechanism. Current research frequently employs inconsistent analytical processes, model formats, and evaluation criteria. This variability leads to a severe lack of interoperability between findings, hindering their ability to form reusable system modules (Misra et al. 2019). Without unified interface mechanisms, these discoveries struggle to transcend academic publications and integrate effectively into clinical practice.

While tools like disease databases, medical knowledge graphs, and multi-omics platforms strive to bridge the research-to-clinic divide (Paananen and Fortino 2020; Y. Pan, Lei, and Yuchen Zhang 2022), most still rely on static name-matching mechanisms—directly linking gene or protein names to disease labels. Such simplistic approaches overlook the critical need for semantic variation, contextual dependence, and the modeling of underlying causal mechanisms, thereby challenging their ability to genuinely support complex clinical reasoning (Tanabe and Wilbur 2002).

In essence, this translational barrier is not merely a problem of data scale or algorithmic performance. It fundamentally represents a misalignment between epistemological frameworks and operational systems. Without a robust mechanism capable of achieving semantic transfer and causal interpretation from molecular mechanistic discoveries to clinical language, the full promise of precision medicine will remain constrained by structural untranslatability. Addressing this barrier is crucial for unlocking the

1 Introduction

translational potential of molecular research and enabling its integration into clinically actionable knowledge pipelines.

1.2.4 Protein Infrastructure Translation

The fourth structural translation challenge occurs at the level of protein information expression, processing, and manipulation. Proteins are increasingly pivotal in analyzing disease mechanisms, developing targeted therapies, and designing novel biomaterials (Dai and Y. Chen 2023; Zhidong Chen et al. 2023; Duarte et al. 2019). However, the current protein computing ecosystem remains characterized by isolated tools, incompatible formats, fragmented analysis processes, and a pervasive lack of semantic interoperability and common frameworks (Whisstock and Lesk 2003).

The core of this problem is the absence of a unified computational expression language for protein properties. Most existing tools are designed to perform localized structure or function predictions for single sequences, lacking a structured semantic organization mechanism for their outputs. This limitation severely impedes the reusability of computational results and their seamless integration into subsequent diagnostic modeling, sequence generation, or clinical workflows.

Furthermore, a majority of current protein analysis tools operate as closed-loop processes. This means their inputs are typically fixed, outputs are static, and they lack crucial features such as flexible parameter control, interactive querying capabilities, and reverse generation mechanisms (Allouche et al. 2014; Marcos and D.-A. Silva 2018). This inherent rigidity diminishes their practical applicability and severs the potential transformation path of protein data towards upstream sequence design and downstream therapeutic development.

Therefore, there is an urgent need to construct a composable and semantic protein computational representation system. Such a system must be capable of encoding functional properties, supporting goal-oriented intelligent queries, and serving as both input and output within larger biomedical systems. This indispensable infrastructure should feature open interfaces, customizability, and cross-system adaptability to enable multi-layer docking from initial computational analysis to eventual clinical deployment.

In essence, the protein translation problem transcends purely biochemical or algorithmic dimensions; it is fundamentally a deficiency in systemic infrastructure. Without

1.2 Key Challenges in Biomedical Translation

a comprehensive protein semantic system that ensures computability, transferability, and generability, biological knowledge will remain confined to local interpretations, unable to truly enter the clinical or engineering practice design loop. Translation, in this specific context, necessitates building a robust, cross-layer information system that possesses both linguistic and logical coherence, and is capable of robust inter-system docking. As protein-based therapeutics and diagnostics become increasingly central to modern medicine, building a computationally expressive and interoperable infrastructure is imperative for future innovation and scalability.

Overall, this chapter identifies and systematically analyzes four structural breakpoints in translating biomedical knowledge: from patients to systems, from data to diagnosis, from molecular discovery to clinical applications, and from protein properties to computational infrastructure. Although each dimension exhibits different focuses on translation failures, covering issues such as semantic incoherence, operational logic breaks, and non-interoperability between systems, they are not isolated from each other, but together constitute a continuous chain of translation, from human experience to digital records, from molecular encoding to therapeutic design. The fragility of this chain seriously hinders the coherence and usability of biomedical knowledge at different levels (Cannataro et al. 2022; Q. Chen 2024).

More importantly, these translation challenges are not isolated incidents but structural, reproducible, and systematic obstacles. They repeatedly appear at key nodes of representational domain transformation, ontological structural conflicts, and interdisciplinary semantic migration. Although local tool optimization and single-point modeling methods can alleviate individual problems, without systematic alignment between structural levels, the closed loop from data to action and from scientific research to medical care cannot be truly established. Therefore, the next chapter will propose an integrated computational framework that aims to support biomedical translation pathways that are cross-environmental, cross-system, semantically maintainable, and structurally interoperable.

1.3 Research Methodology and Computational Framework

This thesis proposes a unified computational translation framework for the four structural biomedical translation gaps identified in Chapter 1.2. The core principle of the framework is to view translation as infrastructure. In other words, rather than viewing translation as a local, one-off data mapping task or a simple system interface, it conceptualizes it as a hierarchical, system-level computational mechanism that aims to systematically bridge these critical gaps by achieving semantic continuity, structural modularity, and operational interoperability of knowledge in the biomedical context through intelligent modeling and data engineering. The overall architectural overview of the framework is shown in Table 1.1.

The proposed framework aims to address four significant challenges in biomedical knowledge flow, which are instantiated into four logical layers: patient system translation, data diagnostic translation, molecular clinical translation, and protein infrastructure translation. Each layer focuses on solving a unique translation breakpoint and supports information flow and integration between layers through standardized interfaces and data protocols.

To elaborate on the working mechanism and implementation details of this framework, the rest of this chapter is divided into four sections: Chapter 1.3.1 outlines the core design principles of the framework, which guide the construction of the entire framework. Chapter 1.3.2 defines its layered computational architecture and the logic of knowledge flow between layers. Chapter 1.3.3 details the implementation of the core computational systems of the framework, which are distributed in various layers and discussed in depth in subsequent chapters (Chapters 2 to 10). Chapter 1.3.4 discusses the current scope and limitations of the framework, clarifying the core functions that have been implemented and the structural foundation laid for future expansion.

1.3.1 Design Principles of Translation-as-Infrastructure

To ensure that the proposed framework is effective in addressing both isolated translation barriers and systemic, multi-level knowledge misalignments, its design is guided by four core principles. These principles not only guide the framework but also underlie

1.3 Research Methodology and Computational Framework

the actual design and implementation of each computational system under it, directly responding to the challenges outlined in Chapter 1.2 and running through the entire work in subsequent chapters (Chapters 2 to 10).

1.3.1.1 Semantic Continuity

The loss or distortion of meaning when knowledge is transferred across levels is a fundamental cause of failure in biomedical translation. Whether it is the oversimplification of rich patient narratives into mechanistic system labels or the separation of molecular discoveries from their clinical context, semantic discontinuity can severely impair the interpretability and integration of information.

To address this challenge, each layer of this framework aims to preserve the semantic fidelity between input and output by designing explicit semantic mappings and intermediate representation layers. MSaaVS (Chapter 2) establishes an innovative virtualized service structure that aims to efficiently capture patients' subjective expression information through multi-channel data access. Its core function is to combine a dynamic questionnaire system and an interaction module to convert patients' natural language narratives into standardized intent labels and try to retain contextual information in multiple rounds of interaction. This ensures the computability and executableness of patients' expressions and provides structured patient intent data for downstream systems. In addition, EPOP (Chapter 5) realizes the mapping of protein sequences to disease-related, biologically meaningful functional categories, effectively bridging the semantic gap between molecular-level abstractions and clinical concepts. These designs ensure the human readability and medical validity of all computational translation results within the framework, so that the core meaning of knowledge is not diluted or distorted during the translation process.

1.3.1.2 Structural Modularity

To adapt to the inherent complexity of biomedical tasks, promote the scalability of the framework, and support collaborative development, this framework adopts a strict modular architecture. It avoids the traditional rigid pipeline processing flow and implements each translation layer as a highly cohesive and clearly bounded, independent subsystem, which can be developed, deployed, and iterated independently.

1 Introduction

For example, MSaaVS (Chapter 2), PHPlace (Chapter 3), EPOP (Chapter 5), PRO-FASA (Chapter 7), PS-GO (Chapter 8), and Silver Surfer (Chapter 9) are all designed as well-encapsulated computing modules that can be independently run and integrated. In addition, ProteinFlow (Chapter 10), as a general feature engineering backend, can independently and efficiently support a variety of upper-level computing tasks such as diagnostic modeling and generative design. This structural modular design not only greatly supports the independent evolution and reuse of each subsystem but also allows for flexible integration and functional expansion without requiring a redesign of the entire framework's core architecture.

1.3.1.3 Interoperability

In biomedical information systems, semantic incompatibility and data fragmentation are persistent and frequent issues that lead to ineffective data and knowledge transformation. This framework defines open Application Programming Interface (API), common semantic representation protocol, and single data format to essentially address these issues by enforcing interoperability between systems.

MSaaVS (Chapter 2) and PHPlace (Chapter 3) follow a unified intent classification system at the design level when processing patient intent, ensuring semantic consistency and understanding of patient expressions across different application scenarios. In addition, ProteinFlow (Chapter 10) provides standardized and structured feature outputs, serving as the core data bus for upstream model training and downstream design tasks. These design choices transform previously independent data and systems into a coordinated computing ecosystem, laying a solid foundation for the local interoperability currently achieved and for deeper system integration in the future.

1.3.1.4 Interpretability and Actionability

In critical biomedical applications, relying solely on statistical performance or model accuracy is not enough. Model outputs must be understandable and trustworthy to clinical experts, and be directly applicable and guide decision-making in actual diagnostic and treatment workflows.

This framework incorporates interpretability and actionability as core design constraints throughout. Systems at the protein infrastructure layer, such as PS-GO (Chap-

1.3 Research Methodology and Computational Framework

ter 8), allow users to perform parameterized searches based on biophysical properties, and PROFASA (Chapter 7) provides intuitive visual analysis. Protein generators such as Silver Surfer (Chapter 9) allow users to perform biophysical property-based queries and multi-constraint tracking, which together ensure the biological plausibility and engineering controllability of the results. In addition, PHPlace (Chapter 3) uses visual interaction diagrams and Two Dimensional (2D) animations to support joint decision-making and understanding of complex surgical plans between doctors and patients, significantly improving its operability in clinical communication. These mechanisms deeply embed interpretability into the system interface and core semantic structure, thereby maximizing the clinical practical value of computational tools.

1.3.2 Layered Computational Architecture

To enable the “translation as infrastructure” paradigm to operate effectively, this thesis implements a four-layer computational architecture that precisely maps and is designed to address the structural discontinuities identified in Chapter 1.2. These layers (Patient System Translation, Data Diagnosis Translation, Molecular Clinical Translation, and Protein Infrastructure Translation) correspond to key breakpoints in the transfer of biomedical knowledge, spanning the entire process from human experience expression to molecular design.

Crucially, these layers are not entirely isolated in the architecture. The architecture encodes the partial semantic and structural links that have been implemented to support unidirectional or conditional data and knowledge transfer. Structured intent data generated from patient narratives (via MSaaVS and PHPlace) can be used as input to guide diagnostic analysis; diagnostic output can be logically used to guide molecular target analysis (EPOP); and molecular predictions can further drive protein-level retrieval, analysis, or generation (PROFASA, PS-GO, Silver Surfer, and ProteinFlow). Although all theoretically fully automated bidirectional or closed-loop flows envisioned in the framework have not yet been fully implemented, the implemented interfaces between layers and standardized data schemas ensure smooth critical knowledge flows and lay a solid foundation for more comprehensive system integration and automated processes in the future.

1 Introduction

This architecture allows for scalable integration and a high degree of modular evolution. The framework's different computing subsystems can be separately developed and implemented while yet being structurally compatible. As a result, the entire system may be easily modified to accommodate novel use cases, different kinds of data, or sophisticated modeling techniques. The following will provide a detailed description of how this layered design is specifically instantiated. Each system is designed to address a specific breakpoint in the biomedical knowledge flow, and its core principles, implementation methods, and experimental verification will be explained in depth in Chapters 2 to 10.

1.3.3 System Implementation Across Layers

Based on the above layered architecture logic, this study has implemented eight computational systems and a review article, which are carefully distributed in four translation levels. Each system focuses on solving a specific semantic break and contributes to the overall continuity of knowledge flow together. Although these systems are developed independently, their design strictly follows shared data structures, interface standards, and semantic patterns to support the local interoperability achieved and lay a solid foundation for deeper integration in the future.

1.3.3.1 Patient System Translation Layer

At the first level of the framework, two core systems, MSaaVS (Chapter 2) and PHPlace (Chapter 3), jointly construct the computational interface between the patient's subjective experience expression and the standardized medical system logic. They aim to address the problem of missing patient narrative information and insufficient system understanding.

MSaaVS (Chapter 2) establishes an innovative virtualized service structure that aims to efficiently and completely capture patients' subjective expressions through multi-channel data access. Its core function is to combine a dynamic questionnaire system and an interaction module to intelligently convert complex natural language narratives into standardized intent labels while intelligently retaining context in multiple rounds of interaction. This ensures the computability and executable nature of patient expressions, providing semantically rich, structured patient intent data to downstream systems.

1.3 Research Methodology and Computational Framework

PHPlace (Chapter 3) builds further on the foundation of MSaaVS (Chapter 2), focusing on the specific clinical scenario of female pelvic floor health management. It enhances the semantic structuring capability by introducing a rule-based dynamic question-answering engine to capture more detailed symptom descriptions, lifestyle factors, and related psychosocial information. In addition, it innovatively integrates the Pelvic Organ Prolapse Quantification (POP-Q) measurement module and surgical 2D animation visualization capabilities to promote effective communication between doctors and patients about complex conditions and assist patients in understanding surgical plans. The system significantly improves the operability of patient system understanding, transforming patient information from vague narratives into clear, actionable system inputs.

1.3.3.2 Data Diagnosis Translation Layer

The second layer consists of a core computational modeling system that is designed to address the challenge of transforming heterogeneous clinical data into interpretable diagnostic reasoning. This system aims to enhance the transparency and actionability of diagnostic predictions while ensuring model robustness in a data-constrained clinical setting.

Strengthening Diagnostic Translation (Chapter 4) introduces a semantic consistency-constrained data augmentation mechanism that is particularly suitable for small sample and imbalanced clinical environments. This method effectively enhances the generalization and robustness of the model by strictly preserving the rationality of clinical semantics when expanding the feature space, significantly reducing overfitting without introducing any data distortion. It provides key support for building reliable diagnostic models in data-scarce scenarios.

This approach builds a semantic bridge from raw clinical data to robust diagnostic models, aiming to enhance physicians' trust in data-driven diagnoses and provide understandable and trustworthy risk and prediction information for downstream analysis.

1.3.3.3 Molecular Clinical Translation Layer

The third layer aims to connect protein-level data with specific disease-related functional insights, thereby enhancing the clinical interpretation of molecular discoveries and

1 Introduction

providing information for potential target identification. It addresses the challenge that molecular research results are difficult to apply directly to clinical practice.

The EPOP (Chapter 5) is a core implementation of this layer, which is based on the ESM-2 protein language model and fine-tuned on a domain-specific proteomics dataset related to POP. The primary contribution of EPOP (Chapter 5) is to provide an interpretable mapping of protein sequences to disease-related functional labels, thereby transforming abstract molecular data into clinically meaningful information by revealing clear associations between sequence features and clinical relevance. This dramatically improves the interpretability and practical value of molecular data in a clinical context.

Review of Protein Design in the AI Era (Chapter 6) serves as the theoretical foundation for this layer and critically reviews contemporary AI-driven protein design methods. This review highlights the common problem of excessive structural optimization and insufficient functional or application-driven goals in current models. It highlights the need to deeply integrate biological semantics and practical use-case relevance in future protein modeling strategies. This lays an important theoretical foundation for subsequent protein engineering and computational biology work, and guides how to transform molecular-level insights into clinically applicable knowledge more effectively.

Through the combination of the above-mentioned empirical system and theoretical reflection, this layer aims to systematically enhance the ability to translate knowledge from molecular mechanisms to clinical interpretations, providing a key perspective for understanding the deep molecular basis of diseases and developing novel therapies.

1.3.3.4 Protein Infrastructure Translation Layer

The last layer focuses on supporting the transition from static protein sequence and structure representations to reusable, generative, and task-oriented computational assets, providing a solid infrastructure for advanced protein design, analysis, and functional understanding. It aims to address the fragmentation of traditional protein data, the isolation of analysis tools, and the lack of a unified computational paradigm.

PROFASA (Chapter 7) provides an intuitive and interactive visualization platform for protein fragment analysis, property evaluation, and structure visualization. It enables researchers to visually explore the local properties of proteins, thereby accelerating the understanding and reasoning process of complex molecular mechanisms.

1.3 Research Methodology and Computational Framework

PS-GO (Chapter 8) implements an efficient parameterized protein search engine that can accurately retrieve proteins or functional fragments based on a user-defined range of biophysical properties, rather than just sequence or structural similarity. This greatly improves the efficiency and flexibility of function-driven protein structure discovery.

Silver Surfer (Chapter 9) further extends the capabilities of PS-GO (Chapter 8), enabling the automated generation of novel protein sequences through an innovative genetic algorithm. The system is capable of targeted optimization based on multiple user-specified biophysical properties, thus supporting exploratory protein design and engineering and accelerating the transformation process from concept to actual molecule.

ProteinFlow (Chapter 10) builds a comprehensive, large-scale protein feature engineering platform that aims to systematically integrate multi-source, heterogeneous data, including protein sequences, physicochemical properties, and structure. Through parallel data collection, strict quality control, and a multidimensional feature extraction pipeline, it ensures standardized representation and high scalability of all protein data, serving as the core data and feature bus for upper-level protein computing, design, and diagnostic applications.

Together, these systems build a highly modular, composable and scalable computing backend. It not only supports data-driven protein retrieval and novel molecule generation tasks, but also maintains structural compatibility with upstream molecular biology reasoning and downstream potential diagnostic analysis through standardized data interfaces and representation logic. This layer of work has extensively promoted the transformation of protein knowledge from static description to dynamic, engineerable computational assets, providing unprecedented computing power for biomedical research and drug discovery.

1.3.4 Scope and Limitations

Although the proposed computational translation framework offers a unified architectural response to the major translation breakpoints in biomedical knowledge systems, its specific implementation remains partial and is still evolving. The development and validation of all computational systems presented in this thesis (Chapters 2 to 10) are strictly based on the implemented core functions and experimental results.

1 Introduction

Currently, some theoretically envisioned, fully automated cross-layer integration paths in the framework have not yet been fully operational. For example, the seamless feeding of patient intent structures directly to diagnostic models to drive personalized diagnostic logic, or the automatic guidance of molecular target analysis by diagnostic outputs, and the automatic initiation of protein design processes using molecular predictions, these fully automated, closed-loop complex integration processes are currently still in the conceptual stage or structural planning. However, multiple core modules under this framework have achieved the sharing of local semantic patterns and interface logic. For example, MSaaVS (Chapter 2) and PHPlace (Chapter 3) rely on a common intent label dictionary, while PROFASA (Chapter 7), PS-GO (Chapter 8), Silver Surfer (Chapter 9), and ProteinFlow (Chapter 10) share a unified protein property parameter space. These achieved interoperabilities are structurally encoded through carefully designed data formats, semantic schemas, and modular interfaces, which not only ensure the smooth flow of critical knowledge at present, but also lay a solid foundation for more comprehensive and deeper system integration and automated processes in the future.

This framework focuses on translation challenges that can be solved through computational modeling and structured methods. Specifically, it focuses on the level where data representation, knowledge logic, and interface semantics can be formally modeled and enforced. Therefore, the contribution of this paper does not involve translation barriers rooted in non-computational levels such as institutions, policies, ethics, or socio-cultural factors. However, these factors are crucial in actual clinical applications and the adoption of systems.

In addition, although the layered architecture design has general applicability, the existing implementation of this paper primarily focuses on two specific biomedical fields: patient system translation and data-diagnosis translation applications, as represented by pelvic floor health management, and molecular clinical translation and protein infrastructure translation tasks centered on proteins. Adaptation to other biomedical fields will require customization and domain-adaptive development of corresponding data types, representation structures, and task interfaces.

These limitations should not obscure the core contributions of this thesis. This chapter has demonstrated that the paradigm of "translation as infrastructure" can be concretely instantiated as a modular, layered, and semantically aligned computational system by describing the design principles, layered architecture, and core computational system

implementation of the computational translation framework. The core contribution of this paper focuses on addressing the key structural challenges in patient-system translation and protein infrastructure translation identified in Chapter 1.2. It provides important computational methodologies and conceptual proofs for data diagnostic translation and molecular clinical translation.

1.4 Contributions and Thesis Structure

Table 1.1: Mapping of Translational Layers, Computational Modules, and Thesis Chapters

Translation Layer	Computational Modules	Key Functions	Chapter(s)
Patient System Translation	MSaaVS PHPlace	Capture patient narratives through dynamic questionnaires; structure intent data for downstream computation; visualize pelvic health states to support clinical discussion.	Ch. 2, 3
Data Diagnosis Translation	Strengthening Diagnostic Translation	Perform semantic-preserving data augmentation for clinical data; improve model generalizability under data imbalance; support interpretable diagnostic modeling.	Ch. 4
Molecular Clinical Translation	EPOP AI-Driven Protein Design Review	Translate protein sequence data into disease-relevant features; enable interpretable molecular-to-clinic linkages; theorize semantic alignment in protein-based prediction.	Ch. 5, 6
Protein Infrastructure Translation	PROFASA PS-GO Silver Surfer ProteinFlow	Support fragment-level property analysis and visualization; enable parameterized protein structure search; generate optimized protein sequences for design; integrate structured features for downstream modeling.	Ch. 7-10

This thesis aims to systematically address the structural translational fractures that are prevalent in biomedical knowledge flows by introducing the core concept of "translation as infrastructure" and building a hierarchical, modular, and interoperable computational framework. The core contribution of this study is to provide a unified theoretical and computational response that structurally integrates the various independent systems from patient interaction to protein design into a coherent architecture, thereby achieving computational and semantically aligned knowledge translation from molecules to clinics.

The core contributions of this research include:

1 Introduction

- **Conceptual Innovation:** Establishing Translation-as-Infrastructure as a conceptual foundation for enabling stable, interpretable, and computable knowledge translation across biomedical contexts.
- **Framework Construction:** Proposing and operationalizing a four-layer computational framework corresponding to distinct translational faultlines: Patient System, Data Diagnosis, Molecular Clinical, and Protein Infrastructure Translation.
- **System Realization:** Eight functional computational systems and a review article were implemented, distributed across different levels including clinical interface design, diagnostic modeling, molecular reasoning, and protein characterization, with each system instantiated as a level-specific translation module.
- **Interoperability Advancement:** Semantic continuity and interface compatibility are encoded between layers, laying a solid foundation for building a progressively integrable, semi-closed-loop biomedical translation process.

The overall structure of this thesis is divided into three main parts, which gradually and in-depth explain this study.

1.4.1 Theory and Framework

- **Chapter 1:** This chapter lays the conceptual and architectural foundation for the entire study. It first identifies and outlines four major structural challenges in biomedical knowledge translation, proposes the core concept of "translation as infrastructure", and elaborates a hierarchical computational translation framework. In addition, this chapter defines four core principles that guide the design and implementation of the entire framework and outlines the overall structure of this paper.

1.4.2 System Implementation

This part is the core of the thesis, consisting of eight core computational systems and a review article. They specifically address the translation gap at various levels through empirical research and innovative methods. The research results in these chapters, seven

of which have been published in peer-reviewed journals and another two are currently under review, together form a solid empirical foundation for this thesis.

- **Chapter 2:** This chapter delves into the design and implementation of MSaaVS, a generic virtualized service structure that captures and structures patients' subjective experiences at the patient system translation layer by translating their natural language narratives into standardized intent labels through dynamic questionnaires and multiple rounds of interactions, and attempts to preserve contextual information.
- **Chapter 3:** This chapter is based on MSaaVS and focuses on the field of female pelvic floor health management. PHPlace further enhances the patient information structuring capability and the operability of doctor-patient communication through a rule-driven dynamic question-answering engine, POP-Q measurement module, and 2D animation visualization function. It is a domain-specific implementation of the patient-system translation layer.
- **Chapter 4:** This chapter introduces a semantic consistency constrained data augmentation mechanism for small sample clinical data. This method effectively improves the generalization ability and robustness of the diagnostic model by retaining clinical semantic rationality while expanding the feature space, providing key support for reliable modeling of the data diagnosis translation layer.
- **Chapter 5:** This chapter describes the EPOP system, which implements an interpretable mapping of protein sequences to disease-associated functional labels based on the ESM-2 protein language model. The system transforms abstract molecular data into clinically meaningful information at the molecular clinical translation level, enhancing the clinical interpretability of molecular discoveries.
- **Chapter 6:** As a theoretical supplement to the molecular clinical translation layer, this chapter critically reviews contemporary AI-driven protein design methods. The review emphasizes the importance of deep integration of biological semantics and practical use case relevance for future protein modeling strategies, laying a theoretical foundation for subsequent protein engineering research.

1 Introduction

- **Chapter 7:** This chapter introduces PROFASA, an intuitive and interactive visualization platform for protein fragment analysis, property evaluation, and structure visualization. It accelerates researchers' understanding and reasoning about complex molecular mechanisms at the protein infrastructure translation layer.
- **Chapter 8:** This chapter describes PS-GO, an efficient parameterized protein search engine. The system can accurately retrieve proteins or functional fragments according to user-defined biophysical property ranges at the protein infrastructure translation layer, greatly improving the efficiency and flexibility of function-driven protein structure discovery.
- **Chapter 9:** This chapter describes in detail Silver Surfer, which automates the generation of novel protein sequences through an innovative genetic algorithm. The system supports targeted optimization based on multiple biophysical properties at the protein infrastructure translation layer, accelerating exploratory protein design and engineering.
- **Chapter 10:** This chapter introduces ProteinFlow, a comprehensive large-scale protein feature engineering platform. It provides a standardized and highly scalable core data and feature bus for upper-level protein computing, design, and diagnostic applications through parallel data collection, quality control, and multi-dimensional feature extraction, and is the cornerstone of the protein infrastructure translation layer.

1.4.3 Discussion and Conclusion

This part critically reflects on the overall work of the paper, summarizes the main contributions, and looks forward to future research directions.

- **Chapter 11:** This chapter comprehensively discusses the computational translation framework proposed in the paper, and deeply analyzes its boundaries and potential in achieving structural closure, practical deployment potential, and theoretical construction. On this basis, this chapter will also look forward to and propose future research directions, including deeper system integration, application in new fields, and more advanced automation capabilities, aiming to further improve the efficiency and effectiveness of biomedical knowledge translation.

1.4 Contributions and Thesis Structure

- **Chapter 12:** This chapter provides a final summary of the main research contributions of this thesis, reaffirming the value of the computational translation framework and its significance in addressing biomedical translation challenges, and emphasizing the long-term impact of this research on the development of the field.

The subsequent chapters will be developed in the above order, elaborating on the specific design, implementation details, and experimental verification of each subsystem and its specific role in the overall framework.

2 Bridging Patient Narratives to Structured Healthcare Services: The MSaaVS Framework

This chapter was published as: **Yanlin Mi***, Barry A O'Reilly and Sabin Tabirca. (2024). *Medical software as a virtual service: A multifaceted approach to telemedicine through software-as-a-service and user-centric features*. Digital Health, 10.

DOI:10.1177/20552076241295441

Publisher: Sage Publishing

JCR Quartile: Q1

Impact Factor: 3.3

I made the primary contribution to this work through conceptualization, data curation, formal analysis, investigation, methodology development, resource provision, software creation, validation, visualization, and writing the original and final draft.

2.1 Abstract

This paper presents the MSaaVS platform, a comprehensive telemedicine solution integrating Software-as-a-Service (SaaS) principles with user-centric features to enhance healthcare service efficiency, accessibility, and patient-provider interaction. MSaaVS employs a multifaceted approach, incorporating a robust questionnaire system for data collection and an empathetic virtual agent module to facilitate nuanced patient-provider interactions. The platform prioritizes privacy and data security through Advanced Encryption Standard (AES) encryption and data anonymization, aligning with Health Insurance Portability and Accountability Act (HIPAA) standards. The initial deployment of MSaaVS demonstrates significant improvements in data collection efficiency, patient

engagement, and healthcare service quality. The platform's adaptability is evidenced by its successful application in specialized fields such as radiology and prolapse. User feedback underscores the system's ease of use and potential for reducing healthcare providers' workload. Despite its promising outcomes, MSaaVS faces challenges, including the need for in-person treatment and limited virtual agent customization. Short-term improvements aim to enhance appointment scheduling, speech recognition, and agent personalization. Long-term goals include integrating AI for diagnostic assistance, Internet of Things (IoT) for comprehensive remote care, and advanced virtual agents for improved patient interaction. MSaaVS emerges as a transformative telemedicine solution, effectively addressing contemporary healthcare challenges. Its continuous evolution, marked by the integration of advanced technologies and user-centered design, holds the potential to reshape the landscape of remote healthcare delivery. Future research will focus on refining platform features and examining the broader impact on healthcare systems and patient outcomes.

2.2 Introduction

In the global healthcare domain, the advent of telemedicine marks a significant shift in the paradigm of medical service delivery. Particularly during the COVID-19 pandemic, telemedicine not only ensured the continuity of medical services but also substantially reduced direct contact between patients and healthcare workers, mitigating the risk of virus transmission (Haleem et al. 2021; Omboni et al. 2022). However, while telemedicine has introduced convenience in the provision and reception of medical services, it has also posed a series of challenges, including ensuring data accuracy, protecting patient privacy, optimizing the interaction between patients and healthcare providers, and enhancing the overall quality of medical services (Nittari et al. 2020).

Data plays a crucial role in modern healthcare, supporting daily medical decisions and forming the basis for personalized treatment and precision medicine. In the telemedicine environment, the effective collection, processing, and analysis of vast amounts of medical data to ensure quality and security become key factors driving the development of remote medical services (Zeadally, Siddiqui, et al. 2020). Moreover, as patients' expectations for the quality of medical services continue to rise, providing an equivalent or superior user experience through telemedicine platforms compared to traditional face-to-face services

2 Bridging Patient Narratives to Structured Healthcare Services

has become a focal point for healthcare providers (Hao Jin et al. 2019; Price and Cohen 2019).

To address these challenges, this study introduces the innovative concept of MSaaVS. By integrating advanced SaaS models, MSaaVS offers a powerful, flexible, and easy-to-manage platform for healthcare providers and significantly enhances the efficiency and accuracy of data collection while optimizing patient interaction through its carefully designed questionnaire system and virtual agent module. The questionnaire system allows healthcare providers to design and customize questionnaires according to specific needs, and its efficient data processing and analysis capabilities ensure the accuracy and integrity of patient information. Meanwhile, the virtual agent module simulates human interaction, providing a more humanized and empathetic interaction environment that enhances patient engagement and satisfaction. These innovations enable MSaaVS to improve not only the convenience and satisfaction of patients but also the efficiency and quality of the entire healthcare service system.

Although integrating AI technology into MSaaVS is seen as an important pathway to enhance system performance and service quality, the current phase of the study focuses more on enhancing the overall performance of the system through the optimization of core functions such as the questionnaire system and virtual agent module. In the future, with the advancement of technology and the accumulation of practical experience, AI is expected to play a more significant role in MSaaVS, bringing deeper data insights and more intelligent interaction experiences to healthcare providers and patients (Schwalbe and Wahl 2020; Rajkomar, Dean, and Kohane 2019; Obermeyer and Emanuel 2016).

As shown in Figure 2.1, MSaaVS assists healthcare providers in the development of their intelligent telemedicine system. It collects patients' data and assists in diagnosis. The SaaS platform for healthcare providers can lower development, maintenance, and learning costs, enabling more healthcare providers to deliver services to their patients (Cusumano 2010). Moreover, MSaaVS does not rely on direct communication between healthcare providers and patients but instead provides a virtual agent that can communicate with patients and collect patient data. The agent allows a healthcare provider to help multiple patients at once and enables patients to request consultations anytime, anywhere. It increases the capacity of healthcare providers to handle a broader range of patients and expands the scope of telemedicine. As a result, healthcare providers from various fields can use MSaaVS, allowing patients to access consultations for most diseases in the

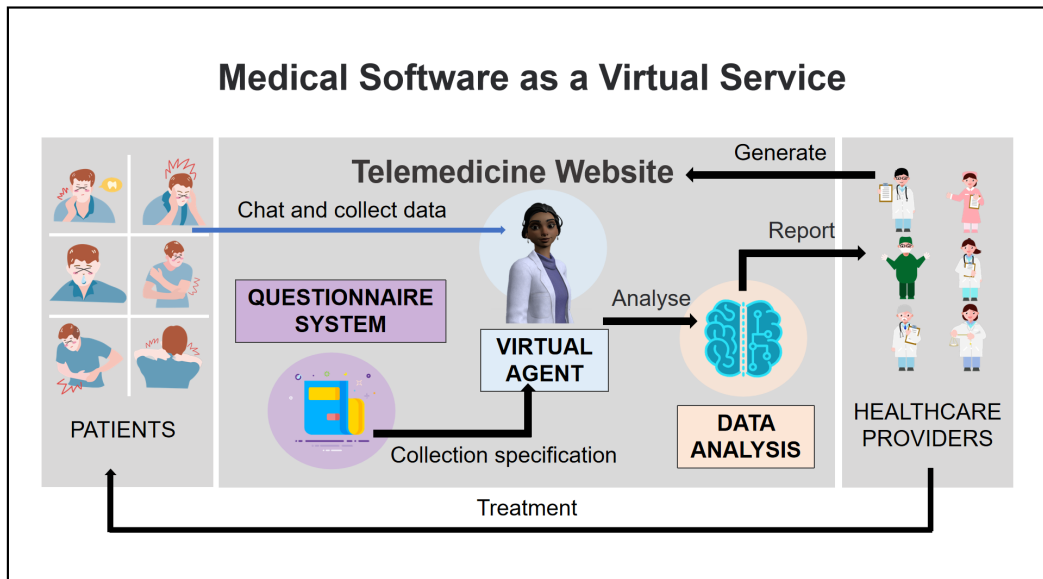


Figure 2.1: Overview of the MSaaVS platform

future, thus making telemedicine more widespread. MSaaVS is specifically designed to be user-friendly and compatible with web browsers, ensuring that patients and healthcare providers can access the platform and its services regardless of their preferred device.

2.3 Methods

MSaaVS is a SaaS platform designed to offer healthcare providers an innovative telemedicine solution (Ali et al. 2019). Healthcare providers can configure and manage their telemedicine websites through the platform. Compared to traditional healthcare approaches, MSaaVS offers several advantages. Patients can directly access the healthcare provider's telemedicine website or install the application for essential consultations without waiting to visit a hospital or clinic, resulting in time savings and increased efficiency. Furthermore, healthcare providers can easily create a telemedicine website using MSaaVS without investing significant time and material resources, making it easier for patients to access the right healthcare provider, thus creating a more transparent and fair healthcare market.

The core features of MSaaVS include a questionnaire system and virtual agents, as shown in Figure 2.1. These features can greatly improve the efficiency and accuracy of telemedicine consultations. The questionnaire system helps healthcare providers

create questionnaires, while virtual agents can use pre-created questionnaires to replace healthcare providers for initial consultations and data collection, thus reducing their workload.

2.3.1 Questionnaire System Module

The most effective method for collecting data is through questionnaires, ensuring patients provide consistent data suitable for training machine learning models. The questionnaire system module is one of the essential modules for defining the patient-entered data format (Tsang, Royse, and Terkawi 2017). The module includes the development of question sets and a logic editor, allowing healthcare providers to add and modify the questionnaire's content using MSaaVS.

Question sets involve adding or removing questions, managing options, and sorting them. Four question types are currently supported: single choice, multiple choice, text input, and color selection. The logic editor provides more flexibility in the questionnaire system, relying on visualization and drag-and-drop operations. Along with typical screen zooming and location dragging, healthcare providers can connect an answer from one question to another to reduce cognitive load. If a specific option is chosen, the patient is directed to the related question, ensuring the questionnaire system's ease of use and integrity. However, it is essential to consider that patients may find impersonal questionnaires unappealing or unpleasant, particularly if there are many questions. To address this, MSaaVS uses the virtual agent module to handle the questionnaire output and gather responses.

2.3.2 Virtual Agent Module

Since the virtual agent always represents the patient's initial impression of the MSaaVS platform, it can be considered the face of MSaaVS. Healthcare providers can make the most basic customizations to the virtual agent through three configuration items: gender selection, 3D avatar selection, and voice selection.

As a direct contact between the healthcare provider and the patient, the virtual agent collects and communicates information, as shown in Figure 2.2. The physical characteristics of the virtual agent become the patient's first impression. The appearance of the virtual agent in MSaaVS is defined by its appearance, speech, facial expressions,

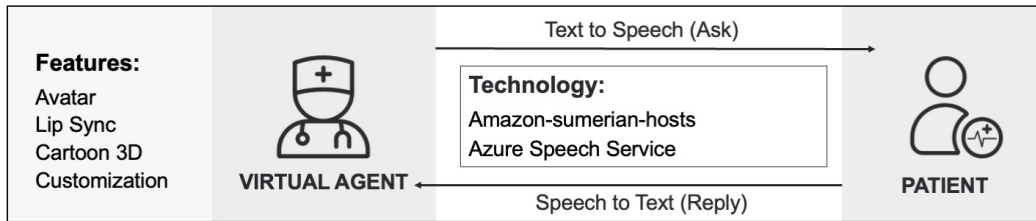


Figure 2.2: Virtual agent interaction in MSaaVS

and body movements. Due to the performance limitations of browser-based 3D model rendering, the virtual agent's appearance cannot be entirely realistic like those seen in next-generation games (S. E. Kirkley and J. R. Kirkley 2005). However, using a realistic appearance can lead to the "uncanny valley" effect (Tinwell 2014), which refers to the discomfort and eeriness humans can feel when robots or other human-like objects look almost, but not exactly, like real humans. Therefore, the virtual agent of MSaaVS uses a 3D cartoon model with slightly exaggerated facial features and a head-to-body ratio to reduce this effect while adjusting the patient's mentality during the consultation through formal dressing and expressions.

Furthermore, the virtual agent must engage in meaningful patient interactions, which require verbal and nonverbal communication. To achieve effective communication, the virtual agent must convert the text from the questionnaire into speech, which current technology can easily accomplish. However, the real challenge lies in ensuring that the virtual agent's facial expressions and body language are synchronized with the spoken words to create an authentic and distraction-free experience for the patient during the consultation. Moreover, the virtual agent must be able to receive input from the user through various means, such as text, clickable options, or voice commands, to ensure accessibility for all patients. Achieving this requires leveraging advanced NLP techniques, such as machine learning algorithms and neural networks, which enable the system to comprehend the meaning conveyed by spoken words. For this purpose, MSaaVS relies on Microsoft's speech-to-text technology, which employs cutting-edge deep learning models to enhance transcription accuracy and ensure optimal performance for the virtual agent.

2.3.3 Privacy and Data Security

The implementation of robust security measures and privacy protocols is recognized as a cornerstone of MSaaS. Due to the sensitive nature of healthcare data, a comprehensive, multi-tiered approach to data security and privacy has been devised, giving priority to both the safeguarding of user information and adherence to regulatory compliance.

One of the critical steps in this procedure is data encryption, both during transit and while stored. The AES with a key size of 256 bits is employed, which is currently seen as the industry standard and is even used by the United States of America (US) federal government for encrypting classified information (Kaur, Lamba, and P. Saini 2021). This powerful encryption method ensures that user data is made unreadable to unauthorized individuals, protecting it from potential breaches.

In addition to encryption, data anonymization techniques are utilized during data processing. Data anonymization is the process of either encrypting or removing personally identifiable information from data sets, so the individuals whom the data describe remain anonymous. This technique not only strengthens privacy but also ensures compliance with the HIPAA (Cohen and Mello 2018), a crucial legal requirement in the healthcare sector.

However, privacy measures don't stop at data protection. Stringent access controls are implemented to restrict data access to only authorized personnel. This means that even within the system, user data is kept protected and inaccessible to those without the necessary permissions. This layer of security adds another barrier to potential internal breaches.

2.3.4 Technical Implementation

MSaaS is primarily a software-based solution, leveraging various modern software tools and technologies to deliver its functionality. The core components of the system include:

- **Backend Services:** Utilizing GO and the GIN framework to ensure high performance and scalability
- **Frontend Interface:** Employing React.js to provide a smooth, responsive user experience

- Database: Using MySQL and MongoDB to support flexible data models and efficient queries
- Virtual Agent: Integrating TensorFlow.js for NLP, enabling intelligent interactions
- Security Layer: Implementing AES-256 encryption standard and OAuth 2.0 authentication protocol to protect data transmission and storage
- API Integration: Designing RESTful APIs to facilitate integration with other medical systems
- Data Analysis: Incorporating Apache Spark for large-scale data processing and analysis

While MSaaVS itself does not directly include specific hardware components, it is designed to be compatible with various standard computing devices, including computers, tablets, and smartphones commonly used by healthcare providers and patients. The system's web-based nature allows for broad accessibility without the need for specialized hardware, aligning with our goal of making telemedicine services widely available and easily adoptable.

2.3.5 Current Development Status of MSaaVS

The MSaaVS platform has progressed to an advanced prototype stage, surpassing initial conceptualization and basic functionality. Core functional modules, including the questionnaire system and virtual agent module, have been fully implemented and undergone preliminary deployment and testing in actual healthcare environments. The platform's development follows an iterative approach, continuously optimizing and refining based on user feedback and performance metrics.

Currently, MSaaVS enables healthcare providers to create and manage customized telemedicine interfaces, facilitating efficient patient data collection and processing. The system architecture has been designed with future scalability in mind, laying the groundwork for integrating more advanced features such as AI-assisted diagnostics and IoT device connectivity.

While MSaaVS has already demonstrated significant practical value, we recognize the ongoing challenges posed by the rapidly evolving field of telemedicine. As such,

our development team is actively exploring the application of new technologies, such as the use of Augmented Reality (AR) in remote diagnostics and the optimization of deep learning-based NLP techniques, to further enhance the system's intelligence and user experience.

2.4 Results

The MSaaVS platform has demonstrated significant potential and tangible benefits in the field of remote healthcare during its preliminary deployment and testing phases. Its core advantage is primarily manifested in the Questionnaire System Module.

This module has successfully implemented the collection of patient data in a standardized format, which not only lays the foundation for training future machine learning models but also provides reliable decision support for healthcare providers. Currently, the platform has generated 23 distinct questionnaires, each meticulously designed to meet the diagnostic requirements of specific conditions. These questionnaires include, but are not limited to:

- Bladder Function Assessment
- Bowel Function Assessment
- Sexual Function Assessment
- King's Health Questionnaire (KHQ)
- Pelvic Organ Prolapse/Urinary Incontinence Sexual Function Questionnaire (PISQ-12)
- International Consultation on Incontinence Questionnaire-Urinary Incontinence Short Form (ICIQ-UI)
- Patient Global Impression of Improvement (PGI-I)
- Patient Global Impression of Change (PGI-C)
- Patient Global Impression of Severity (PGI-S)

The level of understanding MSaaVS has for each question depends on the logic and rules input by healthcare providers during questionnaire creation. The system's intelligent design allows for complex logical branching, ensuring that the virtual agent can dynamically adjust the order and content of questions based on patient responses. This flexibility enables MSaaVS to simulate the understanding level of trained medical professionals, thereby providing a personalized consultation experience.

For instance, when assessing bladder function, the system can automatically adjust subsequent questions based on the frequency and severity of symptoms reported by the patient, delving deeper into possible causes or complications. This intelligent consultation process not only improves the efficiency of data collection but also provides doctors with more comprehensive and targeted patient information to aid clinical decision-making.

The sufficiency and effectiveness of these questionnaires depend on the specific needs of each healthcare provider and the characteristics of their patient population. The flexibility of MSaaVS allows medical professionals to continuously add and optimize questionnaires as needed. For example, some specialists may regularly update their questionnaire content based on the latest clinical guidelines or research findings to ensure the collection of the most relevant and up-to-date patient information.

It is worth noting that MSaaVS itself does not perform diagnoses, so disease prediction accuracy is not directly applicable to platform evaluation. However, our partner healthcare institutions report that by using the systematized patient data collected through MSaaVS, they have been able to conduct preliminary symptom assessments and risk screenings more efficiently.

As illustrated in Figure 2.3, the interface of the Questionnaire System Module prioritizes intuitive user interaction, empowering healthcare providers of varied technical backgrounds to effortlessly create and manage questionnaires. This aspect has been widely acknowledged in user feedback, with numerous medical professionals recognizing MSaaVS for significantly reducing the workload involved in data collection and processing, allowing them to focus more on patient diagnosis and treatment.

Furthermore, the Virtual Agent Module, as depicted in Figure 2.4, has successfully collected responses and offered personalized experiences to patients. The high level of interaction and personalization provided by the virtual agent not only enhances patient satisfaction but also offers healthcare providers a powerful tool to serve patients more efficiently and in a human-centric manner.

2 Bridging Patient Narratives to Structured Healthcare Services

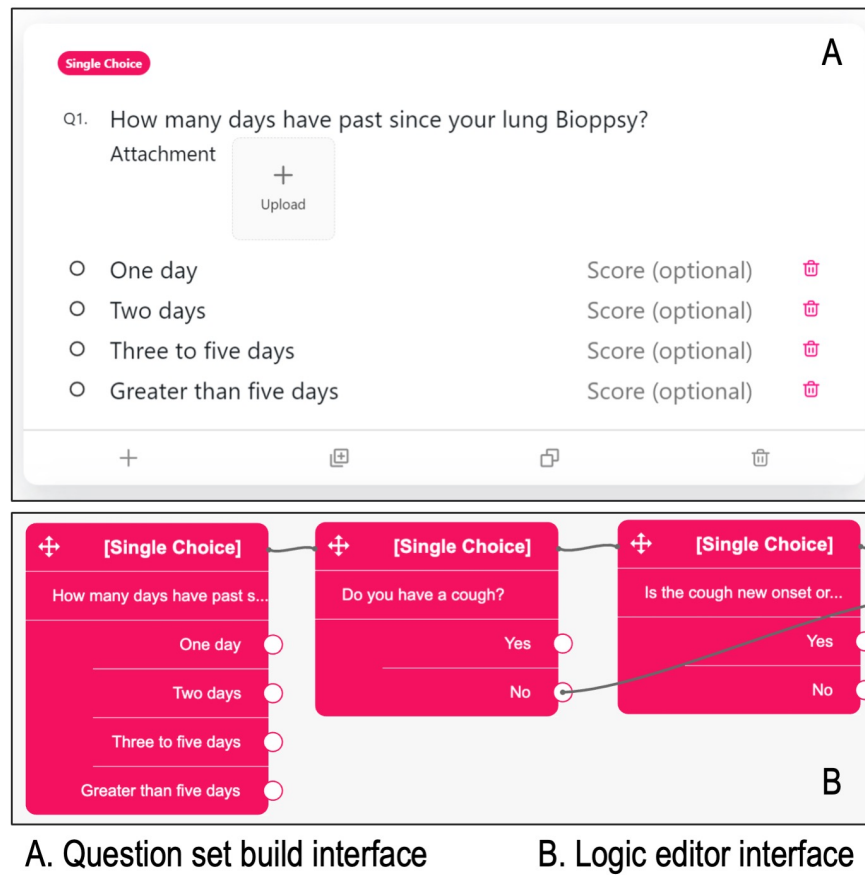


Figure 2.3: Question set building interface and logic editor interface in MSaaVS

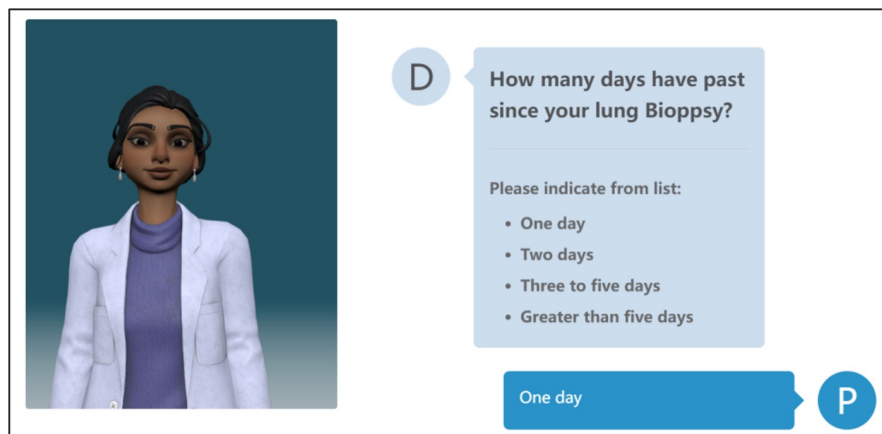


Figure 2.4: Virtual doctor interface in MSaaVS

During the initial trial phase of MSaaVS, our virtual agent module interacted with a total of 50 patients from various medical specialties. Considering the importance of patient privacy protection, we cannot provide detailed information of individual patients. However, we can present aggregated data on patient interactions and satisfaction rates to demonstrate the effectiveness of the virtual agent.

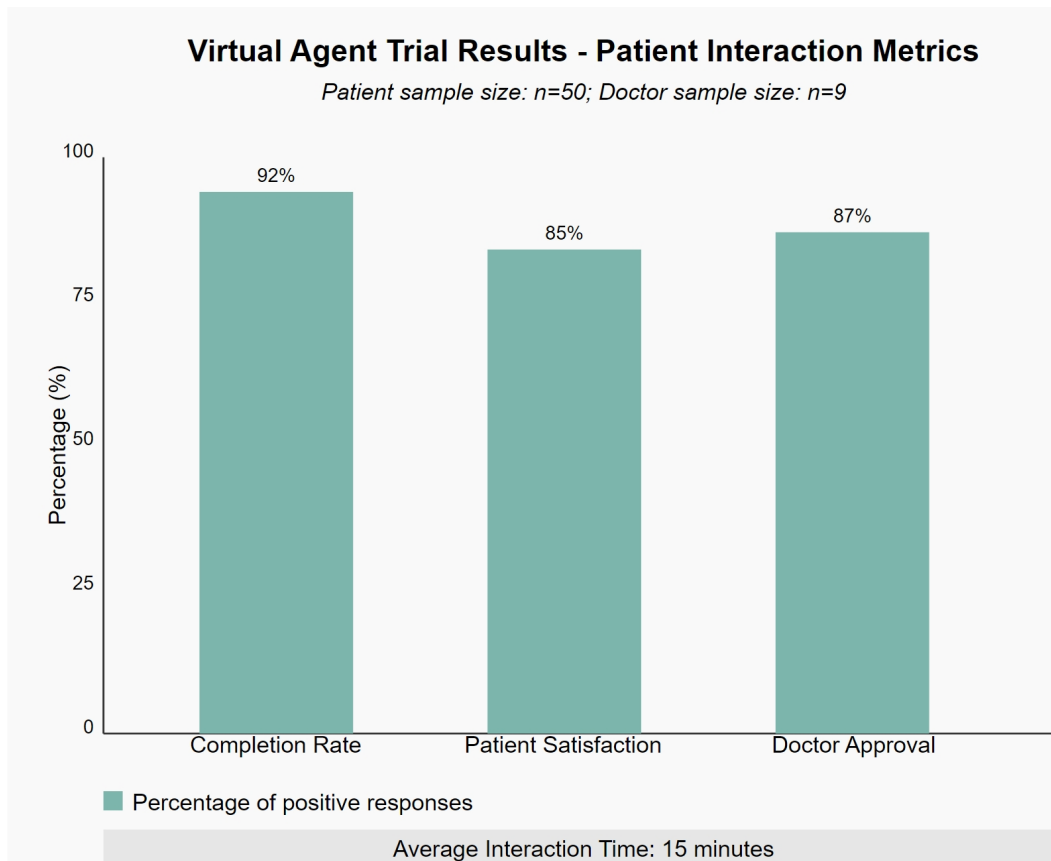


Figure 2.5: Virtual Agent Trial Results: Patient Interaction Metrics

Figure 2.5 illustrates key metrics from the virtual agent trials, including average interaction time, patient satisfaction scores, and successful completion rates of questionnaires. The data shows that among the 50 patients who participated in the trial, 92% successfully completed the entire consultation process, with an average interaction time of 15 minutes. Patient satisfaction survey results indicate that 85% of patients were satisfied or very satisfied with the virtual agent's performance, particularly in terms of accuracy in understanding and responding to questions.

Table 2.1: MSaaVS Usability Metrics

Usability Metric	Satisfaction Rate (%)
Navigation ease	87
Questionnaire clarity	83
Virtual agent interaction	86
Information input convenience	84
Response speed	88
Overall user experience	85

Furthermore, we collected feedback from healthcare providers. Of the 9 doctors who participated in the trial, 87% reported that the quality of information collected by the virtual agent was equal to or better than traditional face-to-face initial consultations. This result highlights the potential of MSaaVS in improving healthcare efficiency and quality.

The 85% usability rate of MSaaVS was derived from a comprehensive user satisfaction survey conducted after the trial period. The survey covered multiple aspects of the system, including ease of use, interface intuitiveness, questionnaire clarity, virtual agent interaction quality, and overall user experience.

Table 2.1 provides a detailed breakdown of the usability metrics that contribute to the overall 85% satisfaction rate. These data are based on survey results from 50 patients.

In specialized medical fields such as radiology and prolapse, MSaaVS has demonstrated its efficacy and applicability. Its flexible design and robust functionality enable it to seamlessly adapt to various medical scenarios and needs, indicating its potential applicability across a broader spectrum of medical fields. It is important to note that the results presented here are based on data collected during the initial testing phase. As the system continues to be optimized and more data is accumulated, the performance and scope of application of MSaaVS are expected to improve further.

Beyond the significant achievements already made, the design of the MSaaVS platform has been envisioned with future development and expansion in mind. With technological advancements and an increase in medical data volume, MSaaVS is expected to integrate more advanced features, such as more intelligent data analytics, more accurate diagnostic support, and more efficient resource management. These potential enhancements and expansions will further solidify the leading position of MSaaVS in the remote healthcare

service domain, bringing additional value and convenience to healthcare providers and patients.

2.5 Discussion

2.5.1 MSaaVS as a Telemedicine SaaS Platform

Telemedicine is a rapidly evolving field transforming healthcare by providing patients with increased access to medical care and enabling healthcare providers to deliver care more efficiently and effectively. Despite the increasing interest among healthcare providers to offer remote care services, the cost of developing telemedicine platforms remains a major barrier, limiting the widespread adoption of telemedicine. Additionally, with limited healthcare providers available to serve many patients, long wait times are common and prevent patients from receiving timely medical attention. MSaaVS, a telemedicine SaaS platform, was developed to address these challenges. As shown in Figure 2.6, MSaaVS can revolutionize healthcare delivery by offering numerous benefits to healthcare providers and patients, enhancing the accessibility and quality of healthcare provision.

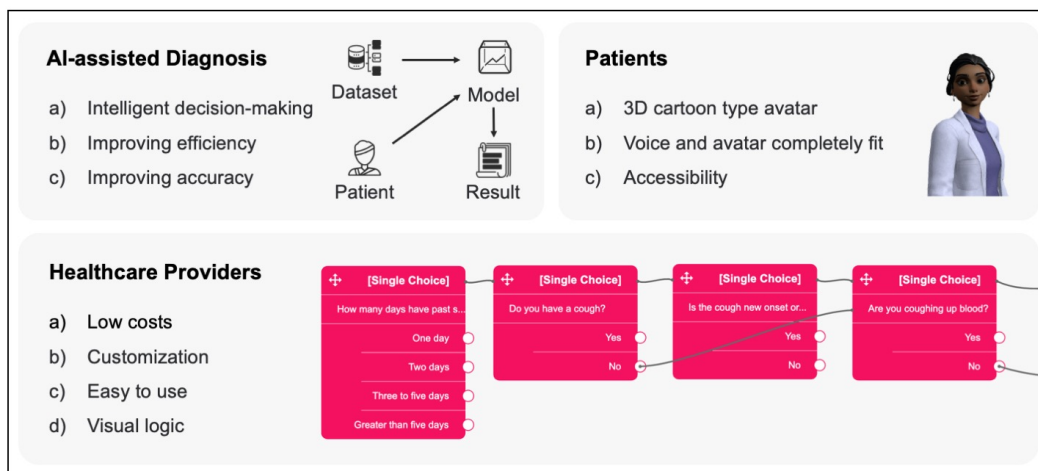


Figure 2.6: Key features and benefits of the MSaaVS platform

For healthcare providers, MSaaVS adheres to the SaaS concept and design principles, helping to lower the costs associated with setting up and operating a telemedicine system.

2 Bridging Patient Narratives to Structured Healthcare Services

This makes the platform more accessible to healthcare providers, promoting wider adoption of telemedicine. Moreover, the user-friendly interface and intuitive interaction logic of MSaaVS significantly reduce the learning curve for healthcare providers. This level of convenience frees healthcare providers from the burden of technology, allowing them to focus on providing high-quality patient care.

For patients, MSaaVS offers a way to access medical care without the limitations of time and place. Patients can use the platform to consult with healthcare providers from the comfort of their homes without worrying about transportation or wait times. Additionally, the virtual agent can question patients about their symptoms 24 hours a day, making it more convenient for patients to provide information about their health status. With MSaaVS, patients can receive medical care that is both convenient and effective, improving their overall health outcomes.

Overall, MSaaVS provides several advantages to the field of telemedicine, making it an important tool for both healthcare providers and patients. By reducing costs, providing intelligent diagnostic advice, and offering a convenient and user-friendly platform, MSaaVS is helping to revolutionize how medical care is delivered.

2.5.2 Challenges

Despite the significant contributions of MSaaVS to the healthcare industry, it faces various challenges that require attention. One of the most notable challenges is that consultation is not the only aspect of healthcare. Even if healthcare providers can accurately diagnose patients through teleconsultation, patients may still require in-person treatment at a hospital or medical facility. Additionally, the current approach to remote consultation in MSaaVS is relatively simplistic, limited to surveys that may not meet the needs of patients who prefer personalized and intelligent conversations with their healthcare providers. Another significant challenge faced by MSaaVS is the limited level of customization for its virtual agents. Currently, healthcare providers cannot obtain virtual agents that closely resemble themselves or meet their specific expectations, which can be challenging for those relying heavily on virtual agents to assist with their workload.

2.5.3 Outlook

Although MSaaVS is already a full-fledged telemedicine platform, our goals go beyond its current state. In the short term, MSaaVS has identified several improvements to enhance its telehealth services. To address the first challenge, MSaaVS plans to add the ability for patients to make appointments online, allowing them to check the business hours of their healthcare provider and schedule an appointment at a specific time that works for them. To tackle the second challenge, MSaaVS intends to enhance its speech-to-text recognition capabilities by incorporating a dictionary of medically relevant terms, helping virtual agents better understand patients' voices. Finally, to address the third challenge, MSaaVS plans to offer more virtual agent avatars and customizable clothing options to provide a more personalized experience for users.

In the long term, MSaaVS aims to leverage cutting-edge technologies to offer more comprehensive telehealth services. One of the most important goals is to collect enough data from the same questionnaires to train AI models using deep learning techniques, realizing the vision of AI-assisted diagnosis. Additionally, MSaaVS plans to explore the use of IoT, remote care, and remote surgery technologies to eliminate the time and space limitations of visiting a healthcare provider (Zeadally and Bello 2021). These technologies will allow patients to receive full telemedicine services from the comfort of their homes. Furthermore, MSaaVS plans to use NLP techniques and chatbots trained with vast medical datasets to enable virtual agents to gather patient data through more natural communication with patients (Thirunavukarasu et al. 2023). This will provide a more personalized and intuitive experience for patients. Moreover, MSaaVS plans to deploy facial scanning and emotion capture technologies to enable completely customizable anthropomorphic virtual agent avatars, giving healthcare providers more control over their virtual agents' appearance and improving the overall user experience.

2.5.4 Limitations and Considerations in AI-Driven Telemedicine

While MSaaVS offers significant advantages in telemedicine, it is crucial to acknowledge certain limitations inherent to AI-driven healthcare systems. One key consideration is the potential for somatization of health problems, where patients may present physical symptoms that are manifestations of psychological distress.

2 Bridging Patient Narratives to Structured Healthcare Services

Moreover, certain specific health issues may be overlooked in an AI-driven interaction, particularly those that require subtle interpretation or physical examination. These limitations underscore the importance of positioning MSaaVS as a complementary tool to traditional healthcare rather than a complete replacement for in-person medical evaluation.

To mitigate these risks and enhance the system's capabilities, we propose the following strategies:

- Regular review and updating of questionnaires to include prompts that may help identify somatization and other complex presentations. This process should involve collaboration with mental health professionals to ensure comprehensive coverage of psychosomatic symptoms.
- Incorporation of open-ended questions that allow patients to express concerns that may not fit into standardized responses. This feature aims to capture nuanced information that might be missed by structured questionnaires.
- Implementation of advanced NLP algorithms to analyze free-text responses, potentially identifying patterns indicative of somatization or other complex health issues.
- Clear communication to patients that MSaaVS is a support tool and does not replace comprehensive in-person medical evaluation when necessary. This message should be integrated into the user interface and reiterated at key points during the virtual consultation.
- Ongoing training for healthcare providers on how to effectively use MSaaVS in conjunction with their clinical expertise. This training should emphasize the system's strengths and limitations, ensuring that providers can critically evaluate the AI-generated insights.
- Development of a hybrid model that combines AI-driven initial assessments with human clinician review, especially for cases flagged as complex or potentially involving somatization.

2.6 Conclusion

In conclusion, this study has demonstrated the potential of the MSaaVS platform as an innovative telemedicine SaaS solution that addresses several critical challenges in the remote healthcare sector. By providing personalized recommendations and a user-friendly interface, MSaaVS aims to enhance accessibility, reduce barriers to entry, and improve overall patient care for healthcare providers and patients alike.

As the demand for remote healthcare services continues to grow, platforms like MSaaVS will play an increasingly important role in ensuring that patients receive timely and accurate diagnoses and treatment. Future research should focus on refining the algorithms and features of the MSaaVS platform, as well as investigating the long-term impacts of such platforms on patient outcomes and healthcare systems. Additionally, exploring the integration of MSaaVS with EHR systems and fostering interdisciplinary communication could contribute to better-informed decision-making and improved patient outcomes.

3 Enhancing Patient Comprehension and Data Structuring for Pelvic Health: The PHPlace Platform

This chapter was published as: **Yanlin Mi***, Reut Rotem, Yair Daykan, Barry A O'Reilly and Sabin Tabirca. (2023). *PHPlace: A New Perspective on Managing Pelvic Organ Prolapse Through Mobile Applications*. In International Conference on Wireless Mobile Communication and Healthcare (pp. 386-399).

DOI:10.1007/978-3-031-60665-6_29

Publisher: Springer Nature Switzerland

I made the primary contribution to this work through conceptualization, data curation, formal analysis, investigation, methodology development, resource provision, software creation, validation, visualization, and writing the original and final draft.

3.1 Abstract

In the rapidly evolving field of digital health, the use of mobile health apps is increasing, which not only allows patients to be more actively involved in their health management and treatment but also significantly improves the efficiency of healthcare professionals. Yet despite the success of mHealth apps in a large portion of the field, there is room for improvement thrown in the field of POP. Pelvic Health Place (PHPlace) is an example of a new mHealth app designed specifically for POP. It aims to improve patient comprehension and healthcare provider efficiency. These features include engaging animated presentations, a groundbreaking algorithm-based scoring system to measure the severity of a condition and versatile medical information management tools. In addition,

by effectively localising the application, PHPlace transcends geographic constraints and extends healthcare services globally. Initial user feedback shows an impressive 90% improvement in users' understanding of their POP condition and its associated treatments, signaling the success of the design implementation. This paper's in-depth exploration of PHPlace provides a strategic blueprint for the design and development of future mHealth applications, setting a new standard for digital healthcare platforms.

3.2 Introduction

As technology continues to advance, mobile health applications have become a major force for change in healthcare. These apps are profoundly changing the way we understand, diagnose, and treat disease, as well as reshaping the way patients interact with healthcare professionals (Rowland et al. 2020). The unique innovation of mHealth applications lies in their ability to cleverly blend knowledge and technology from computer science, data science, medicine, and health care. With real-time health monitoring and personalised disease prevention information, patients are able to more actively engage and manage their health status, potentially improving treatment outcomes and quality of life. At the same time, these apps provide an integrated platform for healthcare professionals to more efficiently manage patient information, monitor patient health status and provide timely feedback and interventions, which not only enhances the quality of healthcare delivery but also helps to reduce the workload of healthcare professionals (Kay, J. Santos, Takane, et al. 2011). However, while mHealth applications have made significant progress in most areas, there are still many challenges in the area of POP.

POP is a common gynecological condition that refers to the prolapse of pelvic organs from their original position in the pelvis and includes Bladder prolapse, Rectal prolapse, Small bowel prolapse, Urethral prolapse, Uterine prolapse, and Vault prolapse (Jelovsek, Maher, and Barber 2007). It is a widespread public health problem, a phenomenon that has a major impact on women's physical health worldwide, as well as a profound impact on the quality of life of those affected. Numerous scientific studies have reported that approximately half of all women worldwide will suffer from POP during their lifetime (Wilkins and J. M. Wu 2016). POP not only causes physical problems, but can also lead to a range of complications, including sexual life disorders, mental health

3 Enhancing Patient Comprehension and Data Structuring for Pelvic Health

problems, and dysfunction related to activities of daily living (Iglesia and Smithling 2017). The effects extend far beyond the realm of physical health and reach into the mental life and social activities of the patient, thus causing a serious impact on their quality of life.

In the medical field, despite the considerable efforts that have been invested in the research and treatment of POP (Jelovsek, Maher, and Barber 2007), many problems remain. One of the most prominent of these problems is the complexity and subjectivity of disease assessment. As current disease assessment relies mainly on the patient's self-report and the doctor's clinical judgment, this approach is susceptible to a number of factors, such as the patient's memory and ability to express themselves, and the doctor's professional experience and judgment, which may lead to biased assessment results (R. Raju and Linder 2021). Another issue is the diversity of treatment options. treatment options for POP often involve multiple modalities, such as surgery, medication, and physiotherapy, which makes it difficult for patients to understand and remember information about the condition and treatment to ensure compliance (Weintraub, Gliner, and Marcus-Braun 2019). In addition, workflow optimisation for healthcare professionals is an important challenge. Current working practices often require physicians to perform extensive manual data recording, analysis, and sharing, which is not only inefficient but can also lead to missed or misunderstood information (Grimes et al. 2020). In addition, it is a major challenge for healthcare to better serve and manage patients from different language and cultural backgrounds on a global scale, and to provide the same quality of service to patients worldwide.

In response to these issues, we have designed and implemented PHPlace, a mobile health application for POP (see Fig. 3.1). PHPlace is an innovative medical software that presents symptoms and treatment processes in a visual and animated format, allowing patients to better understand their condition and treatment plan, thus improving adherence to treatment. In addition, it automates the management of patient records and the scoring of diseases, which is extremely helpful for doctors in terms of improving efficiency, optimising workflow, and improving the quality and efficiency of care. Moreover, the software is also localised to accommodate different languages and cultures, thus increasing the accessibility of healthcare services worldwide. We expect PHPlace to drive digital transformation in healthcare and have a real impact on improving the quality of life for women around the world.

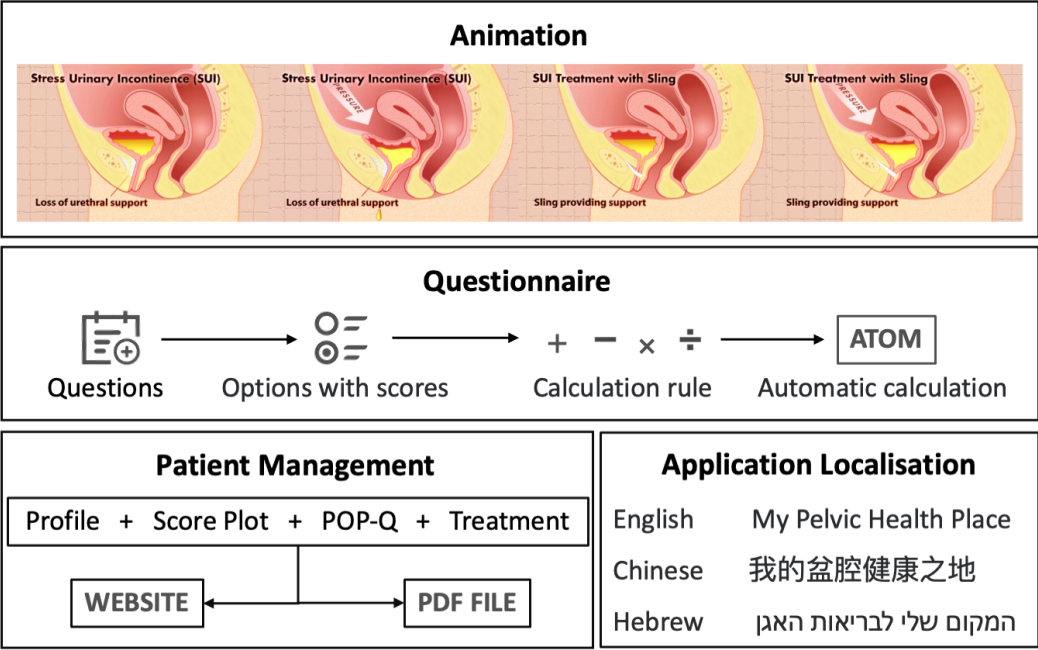


Figure 3.1: PHPlace features

3.3 Methods

We initiated the development of the PHPlace app with a qualitative study, involving 20 in-depth interviews with patients who have experienced POP and 15 consultations with healthcare professionals specializing in urogynecology. The thematic analysis of these interviews highlighted a significant communication gap and identified specific areas where technology could intervene. The patients expressed a need for visual and interactive content to better understand their condition, while professionals emphasized the importance of personalized and dynamic information sharing to enhance patient engagement and compliance.

To convert these insights into app features, we employed a co-design approach. Five workshops were conducted with the participation of both patients and professionals to brainstorm, prioritize, and validate the initial feature set of PHPlace. PHPlace incorporates integrated mobile technology and automation tools to greatly enhance the efficiency of POP disease management. To foster a better understanding of POP disease in patients, animated presentations detailing disease symptoms and treatments are designed and

3 Enhancing Patient Comprehension and Data Structuring for Pelvic Health

implemented. Disease assessment is further automated through the creation of questionnaires and automated scoring systems. Optimisation of the medical information management process also takes place for both doctors and patients, encompassing the generation of PDF files and patient management functions. Localised design of the application is implemented to heighten the accessibility of healthcare services. Further discussion will delve into the process of designing, implementing and optimising these key features.

3.3.1 General Architecture

Designing an application that can effectively manage the POP disease treatment process requires consideration of various usage environments and user roles. Therefore, a decentralised, multi-sided architecture was adopted, comprising a server side, a doctor side, and a patient-side, each working independently and in collaboration with the other.

The server side is responsible for processing and managing the data and is developed in Go, a language whose strength lies in its excellent concurrency and performance, enabling it to process large volumes of requests quickly and ensure real-time data accuracy. Data security and privacy protection are taken into account when handling sensitive medical information of patients. The server side is responsible for processing and managing the data, and advanced encryption algorithms and Secure Sockets Layer (SSL)/Transport Layer Security (TLS) technology are introduced for encrypted data transmission to ensure data security during transmission and storage (Naylor et al. 2014). At the same time, the stored data is hashed and salted, making it difficult to crack even if the data is compromised. In addition, the system ensures that only authorised users and systems can access sensitive information, thus protecting patient privacy.

The doctor and patient sides are designed and developed with the needs of doctors and patients in mind, with the User Interface (UI) and User Experience (UX) in mind. The doctor side uses the React framework, based on its rich component library and flexible state management (Gackenhimer 2015). Its interface is designed to provide intuitive and easy-to-use operations based on the doctors' daily workflow and habits and offers rich interactive elements to optimise the doctors' working experience. The patient side uses the Flutter framework and its ability to develop cross-platform applications ensures a consistent and high-quality UX for both iOS and Android users (Tashildar et al. 2020).

The interface was designed with a particular focus on making it simple and easy to understand and suitable for patients of all ages. The use of large fonts, clear instructions, and a simplified flow of operations make it easier for patients to understand and use.

3.3.2 Animation of Symptoms and Treatment

There is a clear correlation between the effectiveness of treatment for POP diseases and the level of understanding of the disease and treatment modalities by the patient. With this in mind, a set of symptom and treatment animations was developed specifically to help patients gain a deeper understanding of the disease and its treatment.

The 2D animation capabilities of Unity3D were used in depth in the construction of this animation system, resulting in a practical application of computer graphics and programming to the visualisation of diseases and the presentation of treatment options (Thorn 2015). The animation system first generates a series of accurate 2D vector graphics that accurately depict the morphology of biological structures associated with POP, such as the urethra, bladder, rectum and other organs. These graphics are shown in detail in a healthy state as well as in the morphology of POP in its different disease states. In the Unity3D environment, these graphics are further transformed into manipulable objects or sprites. Skeletal and skinning techniques are then applied to the dynamic manipulation of these 2D objects.

In order to give the patient a better visualisation of the condition, the animation system shows the changes in the position of the pelvic organs in different states of POP-Q. For example, the degree of prolapse or prolapse of the organ position is shown in detail in the animation, thus enabling the patient to better understand the specific impact of POP on quality of life, such as difficulty in urination, and pelvic pressure.

In the treatment animation section, the animation system shows changes in the position of the pelvic organs before and after treatment, and how treatment can improve the symptoms of POP. For example, in the animation on surgical treatment, it is shown in detail how the surgery repairs damaged ligaments and muscles to restore the pelvic organs to their normal position.

Finally, through parametric design, the animation system can generate corresponding animation effects based on different POP-Q parameter values. The POP-Q system contains several parameters, such as Aa, Ba, Ap, Bp, C, D, TVL, GH, PB, which

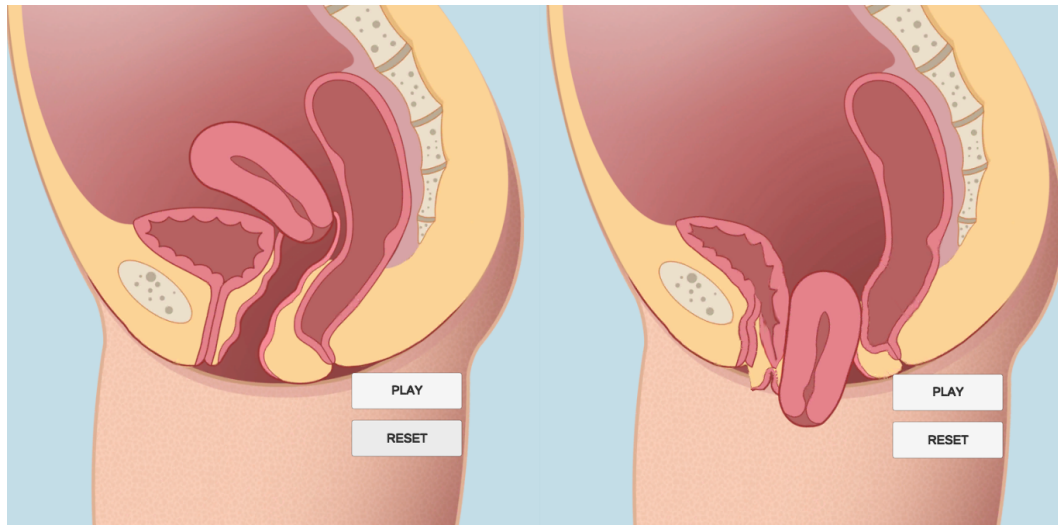


Figure 3.2: The first and last frames of the animation

represent the position of organs in different parts of the pelvis. When the value of each parameter changes, the animation system is designed to present the corresponding animation effect. For example, when the value of C (the maximum distance from the cervix or vaginal tip to the external anal opening) changes, the animation system shows the corresponding change in the position of the cervix or vaginal tip (see Fig. 3.2). In this way, both for the patient and the doctor, the pathological changes of the disease and the treatment process can be understood more visually.

3.3.3 Questionnaire Design and Automated Scoring System

Questionnaire assessment is an important part of POP disease management. Commonly used assessment questionnaires include the Patient Global Impression of Improvement (PGI), the Pelvic Floor Distress Inventory-20 (PFDI-20) and the Pelvic Floor Impact Questionnaire-7 (PFIQ-7) (see Table 3.1 for more information). We collaborated with ten senior urogynecologists to ensure the clinical relevance and accuracy of these tools within the app. Each questionnaire was reviewed, and its scoring algorithm was adapted to the app's interactive and dynamic interface. Traditional assessment methods are time-consuming for both physicians and patients, and the scoring can be subjective due

to the human factors involved. To improve the efficiency and accuracy of assessments, PHPlace has developed an automated questionnaire scoring system.

The system first digitises all possible questionnaire responses and each response is coded into a specific number which corresponds to a specific score or range of scores. For example, in the PFDI-20 questionnaire, possible responses to a question about incontinence include "none", "occasionally", "often" and "always", which are coded as 0,1,2,3, corresponding to different scores. After this step of processing, each questionnaire response could be translated into a clear score or range of scores.

An automated scoring algorithm was then developed using the Go language. The algorithm stores the scoring rules in the form of a dictionary, where the key is the identifier of the question and the value is a mapping function that maps the responses to the question (which have been digitised) to the corresponding score value. When scoring, the algorithm traverses each questionnaire response, finds the corresponding mapping function and calculates the score based on the key value of the response, and finally adds up the scores for all questions or a more complex combination to obtain the total score (see Fig. 3.3).

In addition, the automated scoring system takes into account all possible exceptions. Missing questionnaire responses are treated as invalid input and excluded from the final scoring. For specific borderline cases, such as questions on the impact of sexuality in the PFIQ-7 questionnaire, the system sets up specific conditional judgment logic so that responses to these questions are only included in the final score if the patient indicates that he or she has a sexual life.

Finally, to verify the accuracy and stability of this automated scoring system, the consistency of the system's scoring rules and manual scoring was verified by comparing the system's automated scoring results with historical manual scoring data.

This automated questionnaire scoring system has greatly improved the efficiency of the POP disease assessment, while ensuring accuracy through accurate calculations and comprehensive exception handling.

3 Enhancing Patient Comprehension and Data Structuring for Pelvic Health

Table 3.1: Questionnaires for POP

Questionnaires	Cite
Pelvic Floor Distress Inventory-20 (PFDI-20)	(Teleman et al. 2011)
Pelvic Floor Impact Questionnaire-7 (PFIQ-7)	(Teleman et al. 2011)
Pelvic Organ Prolapse/Urinary Incontinence Sexual Function Questionnaire (PISQ-12)	(Rogers et al. 2003)
The Questionnaire for female Urinary Incontinence Diagnosis (QUID)	(C. S. Bradley et al. 2010)
Female sexual function index scoring (FSFI)	(Sen et al. 2006)
Patient Global Impression of Improvement (PGI)	(Srikrishna, Robinson, and Cardozo 2010)
King's Health Questionnaire(KHQ)	(Hebbbar, Pandey, and Chawla 2015)
International Consultation on Incontinence Questionnaire-Urinary Incontinence Short Form(ICIQ-UI)	(Klovning et al. 2009)

3.3 Methods

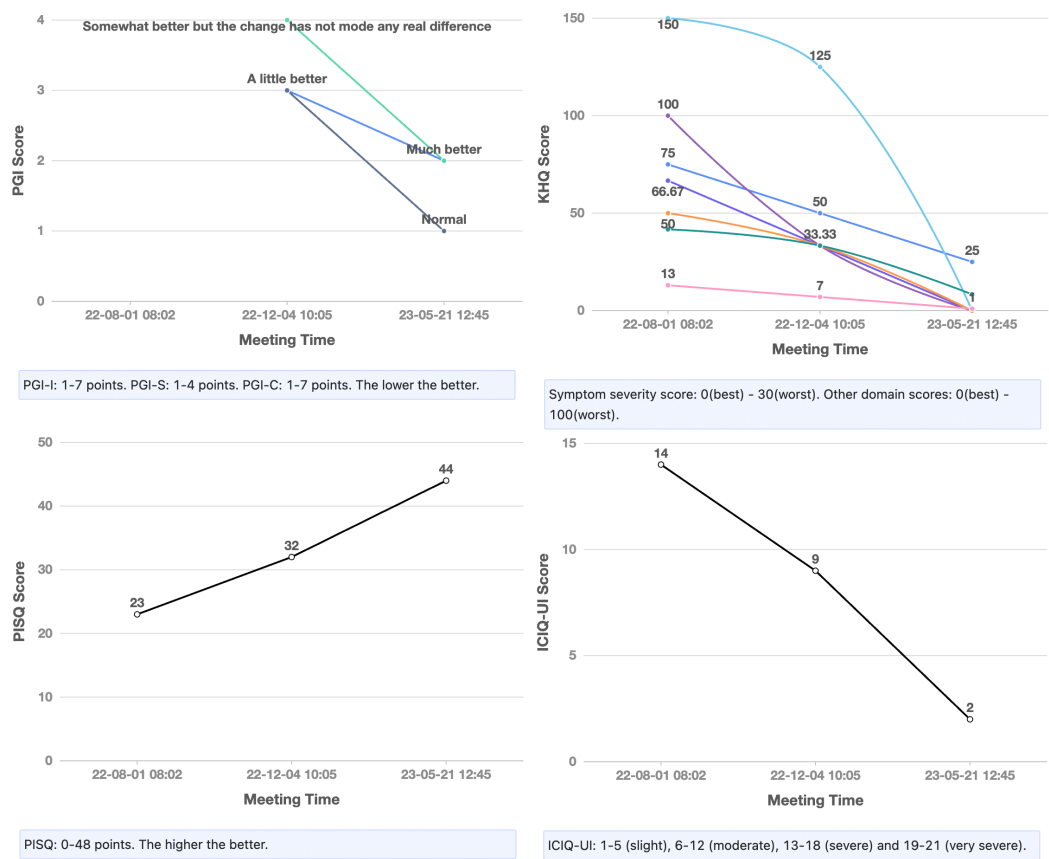


Figure 3.3: Multiple questionnaire score visualisation chart

3.3.4 PDF Generation and Patient Management

Effective medical information processing, access and management is essential in POP disease management. To address this, PHPlace has developed two key features: a file PDF generation function and a patient management function.

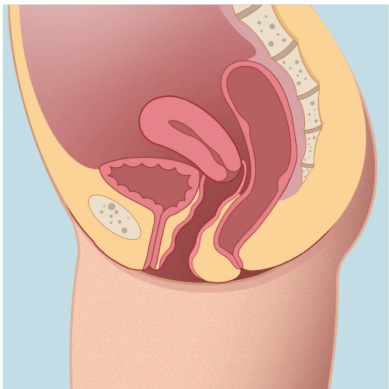
The file PDF generation function uses react-pdf, an open source PDF library based on React.js, to enable the centralised visualisation of patient medical information. Specific PDF templates were developed which followed the design standards of the health-care industry and were designed with HTML/CSS to ensure their readability and user-friendliness. The template covers various sections such as basic patient information, disease assessment questionnaire, questionnaire scores, POP-Q scores, surgical recommendations and medication information. Each section is designed with a separate visual presentation module to improve the clarity and ease of printing the information (see Fig. 3.4). The system is able to automatically update PDF files when new medical information is generated or when existing information changes, and in this way provides up-to-date and accurate medical information.

For the patient management function, a doctor-side web application has been developed to provide a unified and convenient interface for doctors to manage their patients' medical information. The application accepts the POP-Q scores entered by the doctor and calls the back-end API to translate the scores into parametric animations. The animations are rendered directly in Canvas via WebGL technology, ensuring smoothness and adaptability to different devices and resolutions. After the animation has been generated, the system inserts specific frames of the animation into the patient's profile PDF, enabling doctors and patients to understand the disease status more visually. In addition to this, the application also allows doctors to view and manage basic patient information, questionnaire details, and questionnaire scores (see Fig. 3.5). The app allows doctors to access real-time, accurate medical information, which in turn improves the efficiency of diagnosis and treatment.

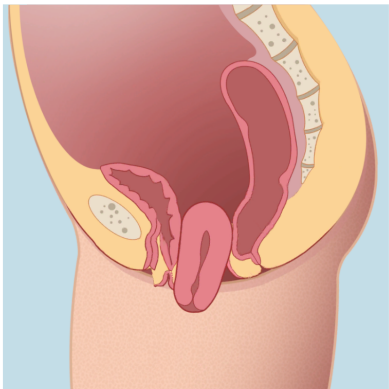
These two features have been designed and implemented to not only improve the efficiency of medical information management, but also to improve the understanding of disease status for both doctors and patients through the visual presentation of information.

Clinical Examination

Meeting	Follow Up Meeting I 2023-03-28 22:37:17
Uterus	With Uterus
Comments	None
POP-Q	233 / 200



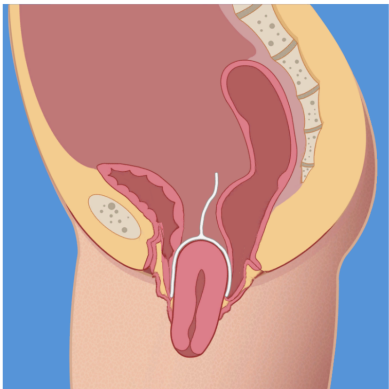
Normal



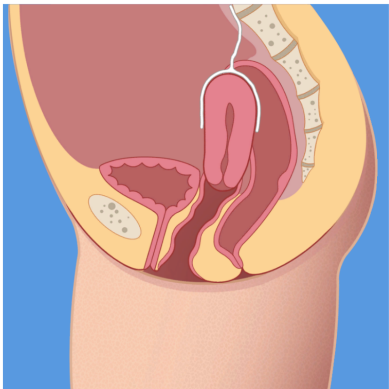
Current

Treatment Plan

Meeting	Follow Up Meeting I 2023-03-28 22:37:17
Surgical Option	Hysteropexy
Comments	None



Before



After

Figure 3.4: Example of the generated pdf showing the visualisation of the animation

3 Enhancing Patient Comprehension and Data Structuring for Pelvic Health

The screenshot displays two sections of a patient management interface. The first section, titled 'Clinical Examination', contains a 'Meeting' dropdown menu with the value 'Follow Up Meeting : 2023-05-21 12:45:03', a 'Uterus' dropdown menu with the value 'With Uterus', and a 'POP-Q' section with six dropdown menus containing the values 2, 3, 3, 2, 0, and 0, followed by a blue circular icon with a white plus sign. Below these is a 'Comments' text area and a blue 'Submit' button. The second section, titled 'Treatment Plan', contains a 'Meeting' dropdown menu with the same value, a 'Surgical Options' dropdown menu with the value 'Hysteropexy', and a 'Comments' text area.

Figure 3.5: Patient management UI

3.3.5 Application Localisation

In PHPlace, particular attention was paid to the localised design of the application in order to improve the experience of patients and doctors in different parts of the world, involving two key components: internationalisation (i18n) and localisation (l10n) (Couto and Miranda 2022).

In terms of internationalisation, the application leverages the internationalisation support of the React and Flutter frameworks. Specifically, by collecting all the string information that needs to be displayed and storing them in a separate language file. This string information is then translated into different language versions depending on the language needs of the target user group. When patients or doctors use the application, depending on their language preference, the system automatically loads the file in the corresponding language, thus providing an interface adapted to the user's language environment. As for dynamic content that may be updated frequently, such as questions and answers of the questionnaire, when this information is returned by the server, the API of Google translate is called for machine translation according to the corresponding language tokens, returning the questionnaire in the corresponding language version.

In terms of localisation, emphasis is placed on the specific details and habits of different cultures and regions. This includes date and time presentation formats, currency units, and units of measure. Localisation tools for React and Flutter make this process much easier. These tools automatically recognise the user's language and locale settings and then provide the corresponding date, time, number and currency formats based on these settings. This way, users get a localised presentation that matches their habits, no matter what region they are from.

This careful design and implementation enables the application to provide a more user-friendly and personalised service to users around the world, effectively improving the accessibility and UX of healthcare services. This localised design not only enhances the global reach of POP Disease Management, but also allows more patients and doctors to access and understand medical information more accurately and easily in a language and format they are familiar with.

3.4 Results

The design and development of PHPlace is complete and is expected to bring significant benefits to POP patients and healthcare professionals. Some positive results can already be anticipated through a series of tests conducted in a simulated environment. Initially, a heuristic evaluation by five usability experts identified potential usability issues, which were addressed before the user testing phase. In the user testing phase, 30 patients and 15 healthcare professionals were involved. They were tasked with specific scenarios to explore the app's features, especially focusing on the animated presentations and questionnaire functionalities. User interactions were logged, and heat maps of interaction hotspots provided insights into the app's usability strengths and areas for improvement. Post-testing interviews captured qualitative feedback, indicating that users found the animations to be particularly enlightening, but desired more interactive and personalized content.

The animated presentation feature showed clear advantages in the simulation tests. These tests covered 45 simulation users worldwide, including patients with POP disease and medical staff. Test results showed that 93.3% of users reported that the animated presentation enhanced their understanding of POP disease and treatment options. This clear visual presentation is expected to improve patient adherence to treatment. The

3 Enhancing Patient Comprehension and Data Structuring for Pelvic Health

superiority of the animated presentations was evident not only in the users' subjective feedback but also in objective measures. We introduced a comprehension test post-interaction, consisting of 20 questions designed to assess the users' understanding of POP and its treatments. The group exposed to animated presentations scored, on average, 75% higher than the control group, who were provided with traditional text-based information.

In a simulated environment, PHPlace's automated scoring system showcased significant efficiency and accuracy. A comparative analysis, involving 100 questionnaires, revealed a drastic reduction in scoring time from an average of 5 minutes per questionnaire in traditional manual methods to 1.2 seconds with PHPlace, marking a 99.6% improvement. However, the system's merit isn't confined to its speed. A concordance correlation coefficient of 0.98, obtained from comparing the automated scores with expert manual assessments, attests to the system's accuracy and consistency, ensuring that the rapidity of assessments doesn't undermine their quality. Furthermore, PHPlace is equipped with a feature to automatically generate PDF reports. This addition enhances the ease and precision in recording and sharing scores, promising to further elevate the efficiency and reliability of clinical assessments in real-world applications.

Also in the simulation test, the questionnaire scoring function demonstrated good ease of use. 84.4% of the simulated users reported that they could easily complete the questionnaire scoring and quickly understand and record the results.

PHPlace has also shown significant potential for healthcare information management. Clinicians in our team reported that they were 60% more efficient at processing patient information using PHPlace than without it. This is backed by a detailed time-motion study. The study observed clinicians during their consultation sessions, measuring the time taken to explain the POP conditions and treatments to the patients. With PHPlace, the explanation time reduced from an average of 15 minutes to 6 minutes per session without compromising the quality of information shared, as affirmed by the patients' feedback.

The localization features were further evaluated in terms of the accuracy of translation and cultural appropriateness of the content. A separate evaluation involving 30 non-English speaking users from five different countries was conducted. Feedback indicated not just ease of use, but also a high level of satisfaction with the accuracy and relevance of the translated content.

Overall, PHPlace demonstrates significant potential in terms of educational effectiveness, scoring efficiency, ease of use, and information management efficiency, and is expected to provide significant benefits to POP patients and healthcare professionals. All of these tests and studies were conducted by a team of experienced healthcare professionals and data scientists to ensure their impartiality and accuracy. In the future, expect further real-world application testing to validate these expected results and enable PHPlace to better serve POP patients and physicians through ongoing optimisation.

3.5 Discussion and Outlook

Today's healthcare environment is being rapidly transformed by digital tools and apps, particularly those focused on disease management and patient education. For example, apps such as MyChart provide a convenient way for patients to view their medical records, communicate with their doctors and even perform some basic health education in one centralised place (Milani et al. 2017). However, these apps tend to take a more traditional and text-heavy approach to education, especially for those complex medical concepts such as POP, which can lead to patients struggling to understand their condition and treatment plan.

PHPlace introduces an animated presentation approach to present these complex medical concepts in a more visual and vivid manner. This approach is expected to improve patients' understanding of their condition and treatment plan, thereby increasing their adherence to treatment. Such anticipated effects may play an important role in reducing relapse and deterioration of the condition. However, whether these effects can be validated in practice will need to be determined through additional clinical studies and patient feedback. Long-term follow-up and research may reveal the specific impact of this educational approach on improving patient adherence to treatment and quality of life.

On the flip side, PHPlace also attempts to address some common problems that physicians face in disease management. Specifically, it integrates a range of automated tools, such as automated questionnaire scoring and information management systems, to improve the efficiency of doctors in POP management. This innovation may help doctors save a lot of time relative to existing applications, allowing them to focus more on caring for their patients' conditions and improving the quality of care. However, the actual

3 Enhancing Patient Comprehension and Data Structuring for Pelvic Health

effectiveness of PHPlace in improving doctors' efficiency needs to be further tested and validated in a real-world environment.

While the innovation and functionality of the app are encouraging, there are limitations that need to be recognised. The app is somewhat dependent on the digital skills of the user, which may pose a challenge for some patients, especially older patients or those with weaker digital skills. In addition, privacy and security issues in handling personal health data must be adequately addressed. These issues will need to be addressed by future research and improvements.

Overall, the development of PHPlace demonstrates the potential for positive impact through innovative approaches to integrating mobile technology and automated tools into disease management. The design and expected results of PHPlace contribute to the digital transformation of the healthcare sector and provide new ideas for POP management. However, in order to achieve these expected results, more practical research and clinical testing are needed to address the issues and challenges and to continuously optimise the functionality and performance of the application.

3.6 Conclusion

The study highlights the potential of a new mobile app to optimise POP disease management. PHPlace demonstrates the key role of digital technology in enhancing the efficiency and accessibility of healthcare delivery through a blend of intuitive animations, automated disease assessment, efficient healthcare information management, and a localised design adapted to multiple cultures.

The innovation of PHPlace is reflected in its new approach, which aims to improve patients' understanding of diseases and treatments, save doctors' assessment time, standardise the assessment process and provide easy management of medical information. Particularly noteworthy is the localised design that significantly improves the accessibility of healthcare services to meet diverse global needs.

While work in this area still needs to be deepened, PHPlace has provided a viable pathway towards digitising healthcare services and offers valuable lessons and insights for future research. With more such research and further refinement of the application, it is expected that digitisation and automation of healthcare services will be achieved in a wider range of areas to provide more efficient and accurate healthcare services.

4 Strengthening Diagnostic Translation: Data Augmentation Strategies for POP Prediction

This chapter was published as: **Yanlin Mi***, Sabin Tabirca and Barry A O'Reilly. (2023, August). *Enhancing the Performance of Multi-Objective Regression for Pelvic Organ Prolapse Prediction via Data Augmentation*. In 2023 4th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE) (pp. 49-54).

DOI:10.1109/ICBASE59196.2023.10303157

Publisher: IEEE

I made the primary contribution to this work through conceptualization, data curation, formal analysis, investigation, methodology development, resource provision, software creation, validation, visualization, and writing the original and final draft.

4.1 Abstract

In healthcare, Machine Learning (ML) has been increasingly applied to predictive models, but the efficacy of these models is often compromised due to limitations in data quality, diversity, and metrics. In other domains, such as image recognition and natural language processing, data augmentation techniques have been successfully applied to mitigate these challenges, but in healthcare such strategies have not been widely applied. Therefore, our research actively explores how these data augmentation techniques can be applied to ML models for predicting the outcome of pelvic organ prolapse surgery. We first performed in-depth data preprocessing and then tried innovative data enhancement techniques such as noise injection and self-sampling. The results show that the application of data enhancement techniques significantly improves the performance of predictive models and

effectively addresses data scarcity and quality issues, which opens up new possibilities for wider application of data enhancement techniques in the medical field in the future.

4.2 Introduction

AI and its subfield ML have become omnipresent forces driving transformation across numerous sectors. One sector that has prominently reaped the benefits of this technological revolution is healthcare (Bini 2018). With the power to process vast amounts of data, identify patterns, and make accurate predictions, these innovative technologies have given rise to new approaches in healthcare management. Among these novel applications, using ML models for predicting surgical outcomes has emerged as a particularly exciting development (Elfanagely et al. 2021). This burgeoning field harnesses the analytical prowess of ML to deliver predictions that are increasingly overcoming the limitations of traditional methods. As such, it has the potential to profoundly transform disease diagnosis, treatment planning, and prognosis prediction, paving the way for improved patient outcomes and more efficient healthcare systems.

In the complex sphere of gynecological health, POP exemplifies the need and potential for ML applications. POP is a common condition affecting millions of women worldwide (Iglesia and Smithling 2017). This condition, often debilitating and significantly impairing women's quality of life, occurs when the pelvic floor muscles and connective tissues weaken, causing one or more of the pelvic organs to descend and potentially protrude from the vagina. POP is characterized by a wide range of variables including severity, organs involved, and related symptoms, all of which contribute to the challenge of surgical management and prognosis prediction (Iglesia and Smithling 2017). The precise prediction of surgical outcomes is, therefore, of paramount importance. Accurate predictions set patient expectations, thereby enhancing their satisfaction with treatment. Healthcare providers, provide valuable insights, enabling the optimization of resource allocation and the development of more effective and personalized treatment strategies.

Traditionally, as shown in Fig. 4.1, the prediction of POP surgical outcomes has relied heavily on conventional statistical methods. These methods have played a crucial role in informing clinical decisions over the years. However, they grapple with limitations, particularly when dealing with complex, high-dimensional, and non-linear data typical of medical datasets. These characteristics of healthcare data demand methods capable

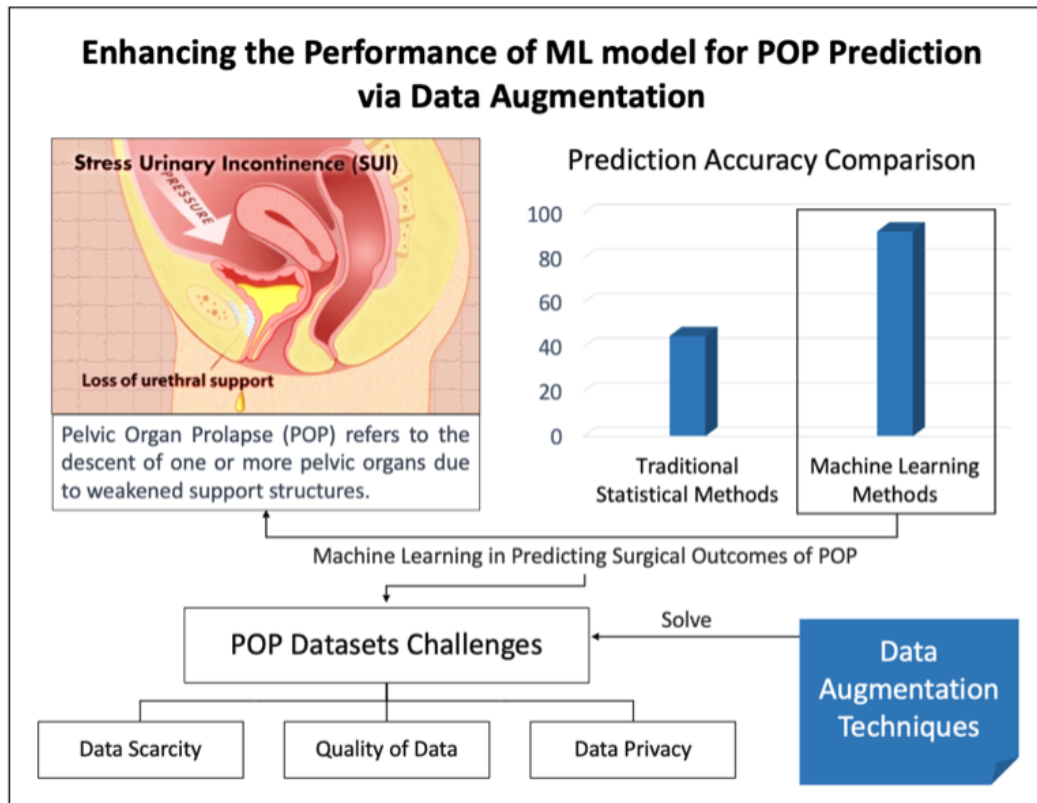


Figure 4.1: Enhancing the performance of ML model for POP prediction via data augmentation

of navigating their intricacies and capturing the underlying patterns with precision. Over recent years, a promising shift toward the use of ML models such as Decision Trees, Support Vector Machines, and Random Forests has been observed. These models have shown potential in handling high-dimensional and non-linear data, with studies demonstrating their superior predictive accuracy over traditional statistical methods (Jelovsek, Chagin, et al. 2018).

Despite the immense promise and growing popularity of ML models, their effective deployment in the context of POP outcome prediction is not without challenges. One of the most significant challenges is the lack of comprehensive, high-quality, and diverse datasets for training robust ML models. The unique characteristics of each POP case, the wide variety of clinical presentations, and the often-subjective nature of medical observations make the collection of representative and extensive datasets a daunting task

4 Strengthening Diagnostic Translation

(Abdi Reyhan et al. 2021). This scarcity of data, coupled with issues of privacy and interoperability of health data systems, severely impedes the development of accurate, reliable, and generalizable ML-based prediction models (R. J. Chen et al. 2021).

In the quest to overcome these challenges, as shown in Fig. 4.1, data augmentation techniques have emerged as a promising solution. Initially utilized in the field of image analysis, where they have significantly boosted the performance of deep learning models (Shorten and Khoshgoftaar 2019), these techniques have begun to find their application in other domains. At their core, data augmentation techniques create synthetic data based on existing samples, enriching the dataset without the need for additional data collection efforts. Although still in the early stages, data augmentation in healthcare has shown promise as an effective strategy to overcome data scarcity, with studies indicating improved model performance in tasks like disease detection and prognosis prediction (R. J. Chen et al. 2021).

However, despite these encouraging advancements, the specific application of data augmentation techniques to improve the performance of multi-target regression models in predicting POP surgical outcomes remains largely unexplored. This uncharted territory warrants exploration, given the potential of data augmentation to bridge the gaps imposed by limited and low-quality data. By shedding light on this under-explored aspect, we can unlock new possibilities for healthcare ML, extending beyond POP to encompass a wide range of other applications grappling with data limitations. In conclusion, the exploration of data augmentation techniques in healthcare ML promises to open new avenues for research and application, bringing us closer to a future where ML is an integral part of healthcare decision-making.

4.3 Methodology

The methodology advanced in this research showcases a novel integration of data augmentation techniques into multi-target regression models. Focusing on postoperative outcomes of POP treatment, this methodology not only contributes to enhancing the prediction accuracy but also mitigates the challenges posed by limited or poor-quality data in the medical domain.

4.3.1 Data Preprocessing

Data preprocessing acts as the foundation for any robust machine-learning model. An array of strategies ensures dataset quality, including handling missing data, managing outliers, performing feature selection, and data scaling.

Missing data is a common issue, and the mean imputation method replaces missing values with the mean value of the respective feature (W.-C. Lin and Tsai 2020). More formally, if $X = \{x_1, x_2, \dots, x_n\}$ represents a feature with m missing values, mean imputation is performed as (4.1).

$$x_i = \frac{1}{n - m} \sum_{j=1, x_j \neq null}^n x_j \quad (4.1)$$

where x_i are the missing values in feature X .

To manage outliers, the Isolation Forest method is adopted, a well-regarded algorithm for outlier detection (Z. Cheng, Zou, and Dong 2019). Given a dataset D of d dimensions, the Isolation Forest isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. The number of such random partitions is equivalent to the path length from the root node to the terminating node. This path length averaged over a forest of such random trees, is a measure of normality and is used for anomaly detection.

Feature selection using Random Forest Regression (RFR) allows the models to focus on the most relevant predictors, reducing overfitting and improving performance (Reme-seiro and Bolon-Canedo 2019). In a dataset with p features, the RFR assigns importance scores $I = \{i_1, i_2, \dots, i_p\}$ to each feature, based on which the most predictive features are selected.

Data scaling is performed using the MinMaxScaler, which transforms features by scaling them to a range between zero and one, ensuring the equal contribution of all features to the prediction (V. G. Raju et al. 2020). The transformation is given by (4.2):

$$x'_i = \frac{x_i - \min(X)}{\max(X) - \min(X)} \quad (4.2)$$

where $X = \{x_1, x_2, \dots, x_n\}$ is a feature, x'_i is the scaled value and x_i is the original value.

4.3.2 Data Augmentation: The Innovations

The data augmentation techniques constitute the main innovation in this research, directly tackling the issue of limited and low-quality medical datasets by generating realistic, synthetic data.

The first strategy involves adding Gaussian noise to the data. A small random variation following a Gaussian distribution is added to each data point, thus creating new, subtly varied instances (Qizhe Xie et al. 2020). Given a feature $X = \{x_1, x_2, \dots, x_n\}$, Gaussian noise of mean 0 and standard deviation σ is added as (4.3):

$$x'_i = x_i + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2) \quad (4.3)$$

and x'_i represents the augmented feature.

The second strategy employs a bootstrapping process. This statistical resampling technique generates a new dataset by randomly sampling from the existing data with replacement, effectively increasing the size of the training set and helping reduce overfitting (Bandara et al. 2021). The bootstrap sample, from a dataset $D = \{d_1, d_2, \dots, d_n\}$, is constructed as (4.4):

$$D' = \{d'_1, d'_2, \dots, d'_n\} \quad (4.4)$$

where d'_i are the data points sampled with replacement from D .

4.3.3 Model Selection and Training

A spectrum of regression models is trained on the augmented dataset. These models include Linear Regression, Decision Tree Regression (DTR), RFR, Support Vector Regression (SVR), and k-Nearest Neighbors (kNN) Regression. Each of these models is trained using the training set obtained post data-augmentation. The models are then evaluated on the validation set, and the performance of each model is recorded. This allows for model selection based on performance, ensuring that the best model is selected for predicting the postoperative outcomes.

The RFR and SVR models have hyperparameters that can greatly affect their performance. These hyperparameters are optimized using a grid search strategy (Liao et al. 2022). The grid search strategy is an exhaustive search over a specified parameter space.

For the RFR, the parameter space might include the number of trees in the forest, the maximum depth of the trees, etc. For the SVR, parameters such as the type of kernel used, the penalty parameter of the error term, etc. are tuned.

After the best hyperparameters are found using grid search, these hyperparameters are used to train the models on the complete training dataset. The models are then evaluated on the test dataset to obtain an unbiased estimate of the generalization performance of the models.

This novel methodology provides a significant contribution to the application of ML in healthcare predictions. By integrating a series of robust preprocessing steps with innovative data augmentation techniques and meticulous model training, this methodology offers an enhanced strategy for using data to improve healthcare prediction outcomes.

4.4 Results

This chapter encapsulates the findings and discoveries derived from various methodical experiments, emphasizing the remarkable role of data augmentation techniques in enhancing the performance of multi-target regression models for predicting POP postoperative outcomes.

4.4.1 Data Exploration and Preprocessing

The journey commenced with an exploration of the dataset's structure, attributes, and distribution, as illustrated by box plots shown in Fig. 4.2. The analysis underlined the need for careful preprocessing, an integral step that involved outlier identification and treatment, as well as standardization of the dataset, to eliminate potential biases.

4.4.2 Data Augmentation: The Innovations

Upon refining the dataset, feature selection was the subsequent significant step. Here, the pivotal attributes contributing to the model's predictive power were identified with the help of the RFR algorithm, and the top 40 features were further validated by an elbow plot shown in Fig. 4.3.

The comparative performance analysis of various regression models before and after data augmentation revealed a significant boost in their efficiency, underscoring the

4 Strengthening Diagnostic Translation

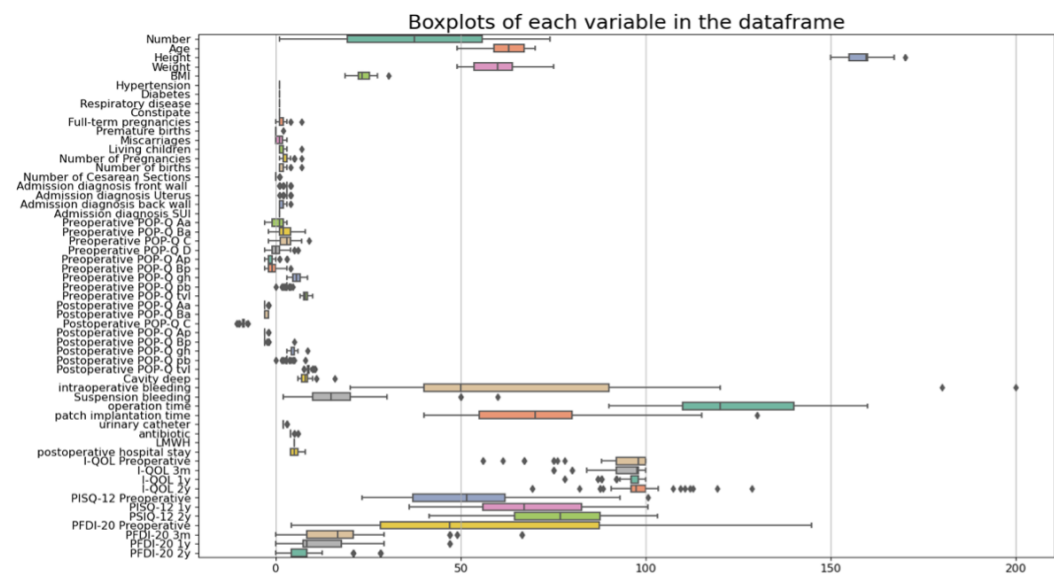


Figure 4.2: A boxplot for each variable in the dataframe visualizes the distribution of the data, showing the minimum, first quartile, median, third quartile, and maximum values, along with potential outliers.

effectiveness of the implemented augmentation techniques. Graphical representations of the improvements can be observed in Fig. 4.4.

Hyperparameter tuning using GridSearchCV led to optimal parameters for the best-performing models, resulting in further performance enhancement. The final performance scores are presented in Table 4.1.

Table 4.1: Optimized Model Performance After Hyperparameter Tuning

Model	Score
Support Vector Regression	0.942
Random Forest Regression	0.728

The results affirm the proposed hypothesis, emphasizing the role of data augmentation techniques in improving the prediction of POP surgical outcomes. The next section further elucidates the implications of these findings within the broader healthcare context.

4 Strengthening Diagnostic Translation

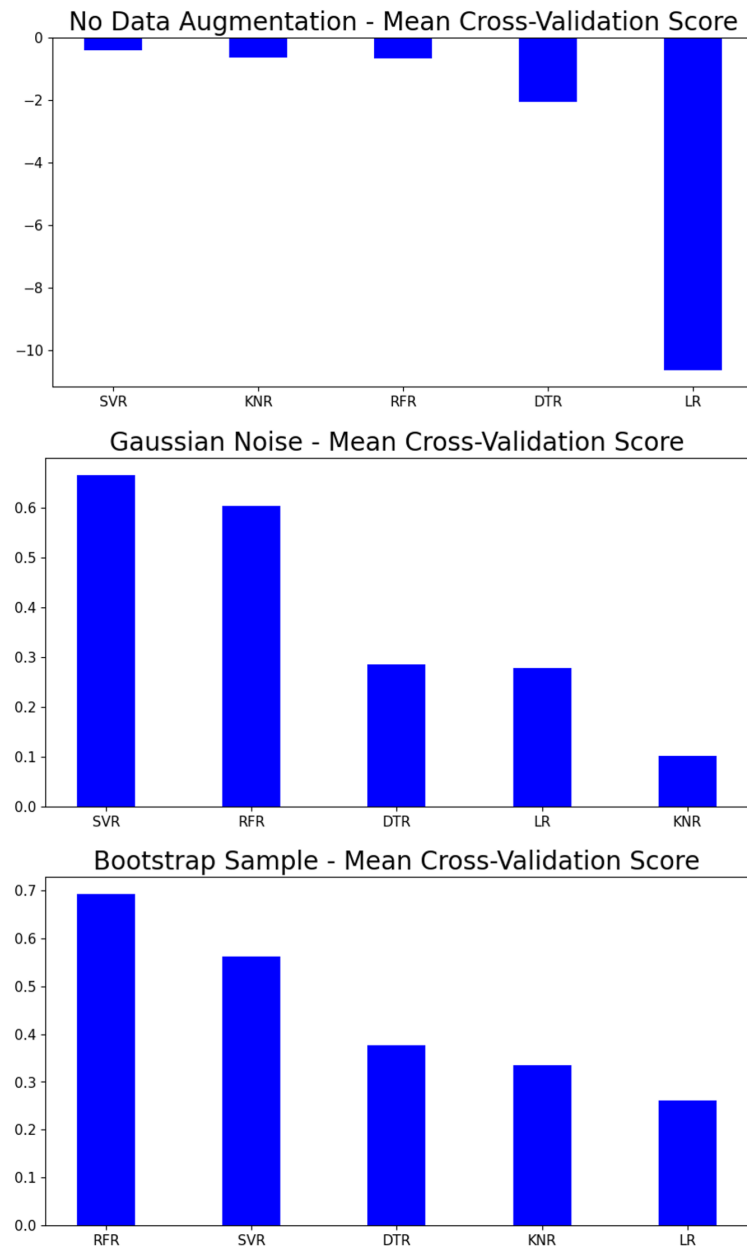


Figure 4.4: Average cross-validation scores of each model under different data enhancement techniques.

4.5 Discussion and Conclusion

Grounded in the quest for better predictive performance in modeling POP surgical outcomes, this research project unfolds a new path by utilizing data augmentation techniques. By addressing the core challenge of data scarcity, the work presented here offers a novel solution that elevates the capabilities of multi-target regression models.

Various existing research has previously proposed different methodologies to mitigate data limitations (Caro et al. 2022; Maharana, Mondal, and Nemade 2022), yet the issue of data scarcity remained. The innovation of this study lies in the direct confrontation of this problem. The data augmentation techniques, namely Gaussian noise addition and bootstrap sampling, not only addressed the limitation of data scarcity but also mitigated the risk of overfitting in ML models, making it superior to previous solutions.

However, it is worth acknowledging the boundaries of these techniques. The Gaussian noise method's efficacy might dwindle when applied to datasets with discrete variables or nonlinear relationships. Similarly, the bootstrap sampling method, while beneficial for this study, may not be universally applicable due to its inherent assumptions. Consequently, future research should explore alternative data augmentation techniques or refine existing ones to cater to different types of datasets.

The study's robustness is further underpinned by the application of an array of multi-output regression models, including Linear Regression, DTR, RFR, SVR, and kNN Regressor. This comprehensive approach allowed for the best-performing models to be identified. Still, the potential for even superior performance from other unexplored models or techniques can't be dismissed, inviting future explorations in this direction.

While these techniques demonstrated significant improvements in model performance, the study also identified potential areas for further research. Exploring other data augmentation techniques such as SMOTE (L. Camacho, Douzas, and Bacao 2022), ADASYN (Zhewei Chen, L. Zhou, and W. Yu 2021), or GAN-based techniques (Karras et al. 2020) could be one such avenue. Each of these techniques presents unique methods for generating synthetic data and could further enhance model performance.

Moreover, the examination of other ML algorithms, including neural networks or ensemble methods, could yield valuable insights for this prediction task. Especially, Deep learning models, given their ability to model complex relationships, might be more

4 Strengthening Diagnostic Translation

adept at providing accurate predictions. However, the data-demanding nature of these models emphasizes the importance of data augmentation techniques.

Finally, expanding the research to other areas of healthcare, where predicting surgical outcomes could be valuable, holds immense potential. It could aid in the development of advanced decision-making tools, leading to improved patient outcomes and more efficient healthcare systems.

In conclusion, this research hopes to inspire further exploration into these promising areas and continues to drive advancements in the application of ML to healthcare. By presenting a viable solution to the challenge of data scarcity, this research project paves the way for improvements in patient care outcomes and efficiencies in healthcare delivery.

5 Translating Molecular Insights to Clinical Relevance: The Extracellular Matrix Protein Predictor for Pelvic Organ Prolapse System

This chapter is based on the manuscript: **Yanlin Mi***, Ben Cahill, Venkata VB Yallapragada, Reut Rotem, Barry A O'Reilly and Sabin Tabirca. *AI-Driven Discovery of Novel Extracellular Matrix Biomarkers in Pelvic Organ Prolapse*.

The manuscript has been accepted for publication in *PLOS Computational Biology*.

Publisher: PLOS

JCR Quartile: Q1

Impact Factor: 3.6

I made the primary contribution to this work through conceptualization, data curation, formal analysis, investigation, methodology development, resource provision, software creation, validation, visualization, and writing of both the original and revised drafts.

5.1 Abstract

Deep Learning (DL) for protein function prediction faces significant challenges in identifying disease-specific proteins. We present EPOP, an advanced transfer learning framework leveraging protein language models to decode disease mechanisms. Focusing on POP, which affects up to 50% of women worldwide, EPOP demonstrates AI's power to reveal novel therapeutic targets. We developed a sophisticated fine-tuning protocol for the ESM-2 model, optimized for ECM protein prediction. Our architecture

5 Translating Molecular Insights to Clinical Relevance

integrates specialized attention mechanisms with interpretability modules, trained on expertly curated and balanced datasets totaling 80,000 proteins (40,000 ECM and 40,000 non-ECM). The framework employs a novel validation strategy using a 16,000-sample independent test set and clinical proteomics data. EPOP achieved unprecedented performance (99.40% accuracy) in ECM protein classification, significantly surpassing traditional DL architectures (10.81% improvement over Transformer models, 21.71% over Long Short-Term Memory). Applied to clinical samples, our model revealed a previously unknown pattern of ECM remodeling, identifying 24 novel disease-associated proteins. Model interpretability analysis uncovered specific sequence motifs and structural features critical for ECM protein function, providing mechanistic insights into disease progression. EPOP demonstrates how advanced AI bridges molecular analysis and clinical applications, uncovering novel therapeutic targets. Its success suggests broader applications across ECM-related disorders, potentially transforming approaches to diseases affecting connective tissue architecture.

5.2 Introduction

Recent advancements in AI have transformed biomarker discovery across medical fields (Bonio et al. 2021). ML, particularly DL, now identifies disease-specific biomarkers inaccessible to conventional methods (Marcos-Zambrano et al. 2021; Ogunjobi et al. 2024). For example, breast cancer studies leveraging recursive feature elimination revealed nine diagnostic biomarkers (e.g., COL10A1 and S100P), validated by qRT-PCR with an AUC of 0.989 (Mirza et al. 2023). And integrating multi-omics data through XGBoost classifiers uncovered cardiovascular biomarkers (RPL36AP37, HBA1) with 100% predictive accuracy (DeGroat et al. 2024), while metabolomic analysis via logistic regression outperformed traditional amyloid-beta detection in Alzheimer's diagnosis using choline and lactate levels (G. (S. Raikar, A. S. Raikar, and Somnache 2023)). Protein language models further exemplify AI's impact, ESM-2 (Z. Lin et al. 2023) and transformer architectures (Rives et al. 2021) enable high-throughput structure-function analysis, culminating in tools like ESM Metagenomic Atlas (Z. Lin et al. 2023) and AlphaFold 2 (Jumper, R. Evans, Pritzel, T. Green, Figurnov, Ronneberger, et al. 2021), that achieve near-experimental prediction accuracy. These AI-driven approaches are

particularly valuable for complex disorders, offering screening of thousands of potential biomarkers simultaneously.

POP represents a significant health challenge, affecting millions of women worldwide with a lifetime risk of surgery estimated at 12.6% to 19% (Fatton et al. 2020). This disorder, characterized by the descent of pelvic organs into the vaginal canal, not only impairs quality of life but also poses substantial economic burdens on healthcare systems (Collins and Lewicky-Gaupp 2022). Despite its prevalence and impact, the molecular underpinnings of POP remain elusive, hampering the development of effective diagnostic and therapeutic strategies (Kluiwers et al. 2023).

At the heart of POP pathogenesis lies the ECM, a complex network of proteins, glycoproteins, and proteoglycans that forms the structural scaffold of pelvic floor supportive tissues (Deng et al. 2021; Guler and Roovers 2022). The ECM's role extends beyond mere structural support; it orchestrates cellular behavior, tissue homeostasis, and biomechanical properties crucial for pelvic organ support (Karamanos et al. 2021). Mounting evidence suggests that disruptions in ECM metabolism and organization are pivotal in POP development, manifesting as alterations in collagen composition, elastin degradation, and dysregulation of matrix metalloproteinases (Tian et al. 2024; Marcu et al. 2020).

However, systematically unraveling the complex web of ECM proteins involved in POP has proven to be challenging. Traditional experimental approaches, while valuable, are often limited by their throughput, bias toward known proteins, and inability to capture the full spectrum of ECM alterations (Yu Li et al. 2021). The complex nature of the ECM, with its diverse protein components and dynamic remodeling processes, further complicates comprehensive analysis using conventional methods (Naba 2023). Additionally, the heterogeneity of pelvic support tissues presents analytical challenges that can obscure subtle but significant alterations in ECM composition and organization (Söderberg 2008; Meijerink, Van Rijssel, and Van Der Linden 2013; R. Gong and Z. Xia 2019; Alperin et al. 2016). This knowledge gap calls for innovative strategies to comprehensively map the ECM landscape in POP and identify novel therapeutic targets with potential clinical applications (Z. Lin et al. 2023).

The advent of artificial intelligence, particularly protein language models like the ESM-2, offers a transformative approach to these challenges (Z. Lin et al. 2023). Based on Transformer architecture, ESM-2 is trained on vast, evolutionarily diverse protein

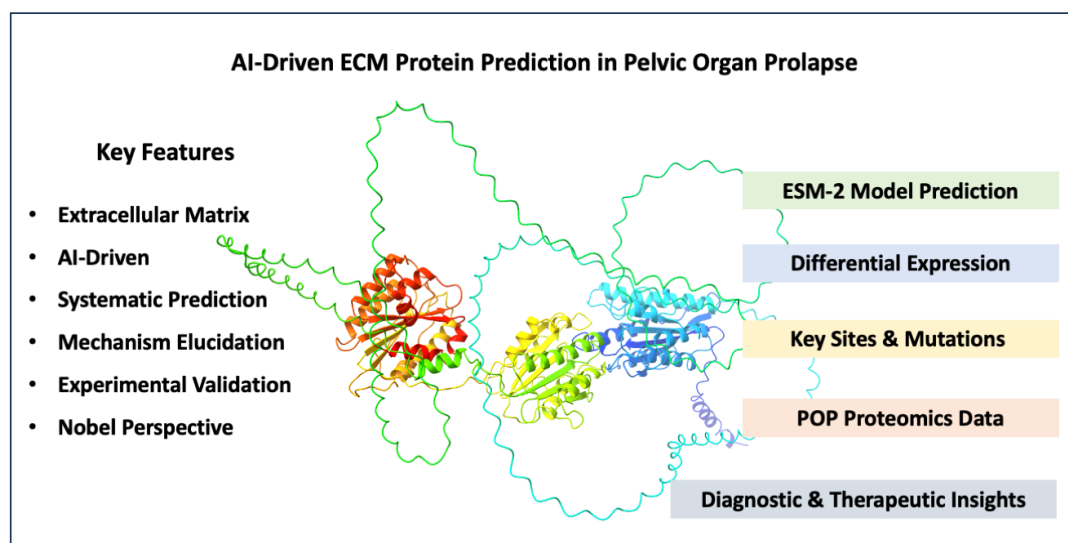


Figure 5.1: AI-driven ECM protein prediction in POP. The figure illustrates the key components of our analytical framework. Left panel lists six key features of the approach: Extracellular Matrix, AI-Driven, Systematic Prediction, Mechanism Elucidation, Experimental Validation, and Novel Perspective. Center shows a protein structure visualization with multiple domains in different colors. Right panel displays five key analytical outputs: ESM-2 Model Prediction, Differential Expression, Key Sites & Mutations, POP Proteomics Data, and Diagnostic & Therapeutic Insights, each represented by a distinctive icon and box style.

sequence databases through self-supervised learning to capture complex amino acid contexts and intricate patterns within protein sequences (Kouba et al. 2023; Cioletti et al. 2025). The generated representations integrate both sequence conservation patterns and implicit structural features, enabling high-throughput unbiased prediction of protein functions with unprecedented precision, approaching experimental precision in certain contexts (Rives 2025; C. Zhang et al. 2024). With its innovative architecture that integrates structural priors, ESM-2 represents a significant leap forward in our ability to decode protein sequences and infer functional implications (Schmirler, Heinzinger, and Rost 2023; Boadu and J. Cheng 2024).

In this study, we used the power of ESM-2 to develop a new computational framework, the EPOP, which aims to systematically identify and characterize POP-associated ECM proteins (Fig 5.1). By fine-tuning ESM-2 on expertly curated datasets of ECM and non-

ECM proteins, we create a highly accurate tool capable of predicting ECM functionality from amino acid sequences alone. We then apply EPOP to analyze proteomics data from POP patient samples, uncovering potential key ECM proteins and evaluating their known associations with POP pathogenesis through literature cross-validation and functional enrichment.

Our approach not only deepens our understanding of the ECM-mediated molecular pathogenesis of POP but also demonstrates the potential of AI-driven strategies in uncovering novel disease-associated proteins. By bridging the gap between large-scale protein sequence data and disease-specific molecular mechanisms, we pave the way for more targeted experimental investigations and potential therapeutic interventions. Moreover, the EPOP framework holds promise for extension to other ECM-related disorders, offering a versatile tool for the systematic discovery of pathogenic mechanisms and therapeutic targets across a spectrum of diseases.

This study represents a significant step forward in leveraging AI to decode the complexities of the ECM in disease contexts, potentially revolutionizing our approach to understanding and treating POP and related disorders. As we unravel the intricate relationships between ECM composition, function, and POP pathogenesis, we open new avenues for personalized medicine and targeted therapies in women's health.

5.3 Materials and Methods

Our study employed a comprehensive workflow (Fig 5.2) integrating data curation, advanced ML techniques, and biological validation to predict and characterize POP-associated ECM proteins.

5.3.1 Data Collection and Preprocessing

5.3.1.1 ECM and Non-ECM Protein Datasets

We curated two high-confidence datasets from UniProt ("UniProt: the universal protein knowledgebase in 2025" 2025) for model training and evaluation, resulting in a balanced dataset of 80,000 proteins (40,000 ECM and 40,000 non-ECM) to ensure class balance and computational efficiency.

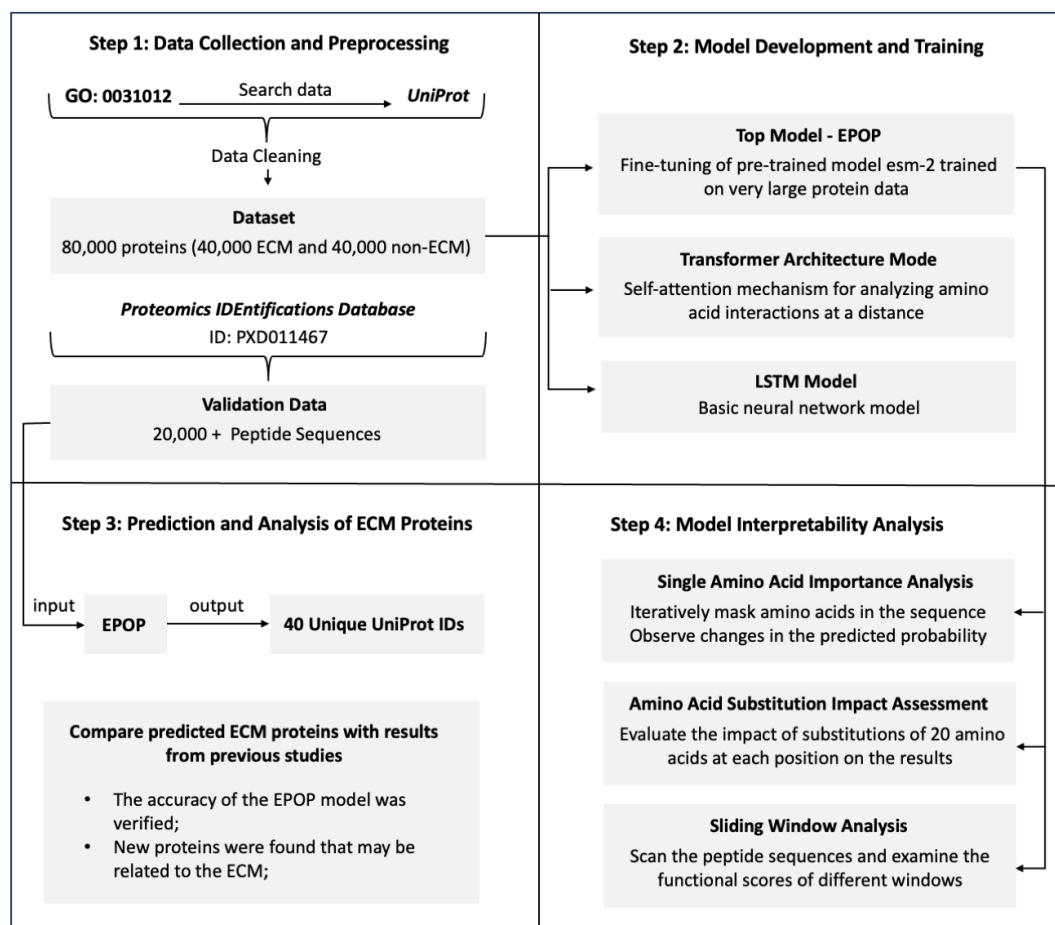


Figure 5.2: Workflow overview showing the four main steps of the EPOP methodology. Step 1 (Data Collection and Preprocessing) shows the retrieval of data from GO :0031012 and UniProt. Initially, this resulted in raw datasets of 40,155 ECM-related proteins and 46,427 non-ECM proteins. These were subsequently cleaned, balanced to 40,000 ECM and 40,000 non-ECM sequences, and then split into a training set (48,000 sequences), a validation set (16,000 sequences), and a test set (16,000 sequences) for model training. Additionally, proteomic data from PRIDE project ID PXD011467, comprising 22,341 PSMs from patient samples, was utilized for external validation and application. Step 2 (Model Development and Training) illustrates three parallel models: EPOP (fine-tuning of pre-trained ESM-2), Transformer Architecture Model (self-attention mechanism), and LSTM Model (basic neural network). Step 3 (Prediction and Analysis) shows EPOP's input-output process. Specifically, the trained EPOP model was applied to the PRIDE proteomic data, identifying 40 unique UniProt IDs of proteins predicted as ECM based on the optimal classification threshold, which will be further analyzed and validated against existing literature. Step 4 (Model Interpretability Analysis) details three analysis approaches: Single Amino Acid Importance Analysis, Amino Acid Substitution Impact Assessment, and Sliding Window Analysis, which were applied to the high-scoring peptides to identify key functional regions and residues.

Positive dataset (ECM proteins): We initially retrieved 40,155 ECM-related proteins annotated with the GO (Consortium 2004) term “extracellular matrix organization” (GO:0031012) and its descendant terms (Consortium 2004). While this GO term broadly captures biological processes involved in ECM dynamics, not all annotated entries represent true ECM components. To refine the dataset, we applied a rigorous two-step curation strategy. Computational biologists filtered proteins based on reviewed annotation status and high-confidence experimental evidence codes. Then, an experienced cell biologist manually evaluated each candidate using UniProt functional descriptions and supporting literature. Proteins were retained only if they demonstrated clear extracellular localization and experimentally supported roles in ECM structure or regulation. Proteins with exclusively intracellular localization, membrane-bound proteins without ECM function, or non-secreted factors unrelated to matrix organization were excluded. From this curated set, we randomly selected 40,000 proteins to constitute the positive class for downstream modeling.

Negative dataset (Non-ECM proteins): We initially collected 46,427 proteins from UniProt that lacked any ECM-related GO annotations. To further minimize the inclusion of overlooked ECM components, we applied an additional two-step filtering procedure. SignalP (Teufel et al. 2022) was used to exclude proteins predicted to contain signal peptides, which often indicate secretion into the extracellular space. Then, TMHMM (Hallgren et al. 2022) was used to remove proteins with predicted transmembrane domains, thereby eliminating integral membrane proteins that may interact with the ECM but are not themselves its structural components. From this filtered pool, we randomly selected 40,000 proteins to constitute the negative class for downstream modeling.

To reduce sequence redundancy and enhance dataset diversity, we applied CD-HIT (L. Fu et al. 2012) with a 30% sequence identity threshold to both datasets. This threshold is a widely accepted heuristic in protein bioinformatics for balancing redundancy reduction with the preservation of distinct biological information, allowing us to group likely homologous sequences while retaining a diverse representation of protein families. By removing highly similar sequences, this step mitigates the risk of model overfitting to specific sequence variations and ensures that the model learns more generalizable features indicative of ECM or non-ECM characteristics. This approach results in a dataset that facilitates robust model training and provides a more reliable evaluation of classification

5 Translating Molecular Insights to Clinical Relevance

performance by preventing artificial inflation of metrics due to highly similar sequences in training and test sets.

In the end, the comprehensive balanced dataset of 80,000 proteins (40,000 ECM and 40,000 non-ECM) was randomly partitioned into three mutually exclusive sets: a training set of 48,000 sequences, a validation set of 16,000 sequences, and an independent test set of 16,000 sequences. This random partitioning ensured that each subset was representative of the overall protein diversity and that the test set remained completely unseen during model training and validation phases, thereby allowing for an unbiased assessment of generalization performance.

5.3.1.2 POP-Related Proteomic Data

We obtained proteomic data comparing vaginal wall tissues of POP patients and healthy controls from Li et al. (X.-J. Li et al. 2019) (PRIDE (Perez-Riverol et al. 2022): PXD011467). Raw mass spectrometry data were processed using MaxQuant (Pranichnikov et al. 2020) with the following parameters:

1. False Discovery Rate (FDR) threshold: 1% at both peptide and protein levels
2. Label-Free Quantification (LFQ) for protein abundance estimation

To ensure data reliability, we implemented comprehensive quality control measures (Fig 5.3):

1. Peptide length distribution analysis
2. Percolator PEP evaluation
3. Correlation analysis between Mascot scores and PEP values

5.3.2 Model Development and Training

5.3.2.1 EPOP Model: Fine-Tuning ESM-2

We leveraged the state-of-the-art ESM-2 protein language model (Z. Lin et al. 2023) as the foundation for our ECM function prediction task. Specifically, we used the pre-trained ESM-2 (650M parameters, 33 layers, UR50D) checkpoint from Meta ECM, which

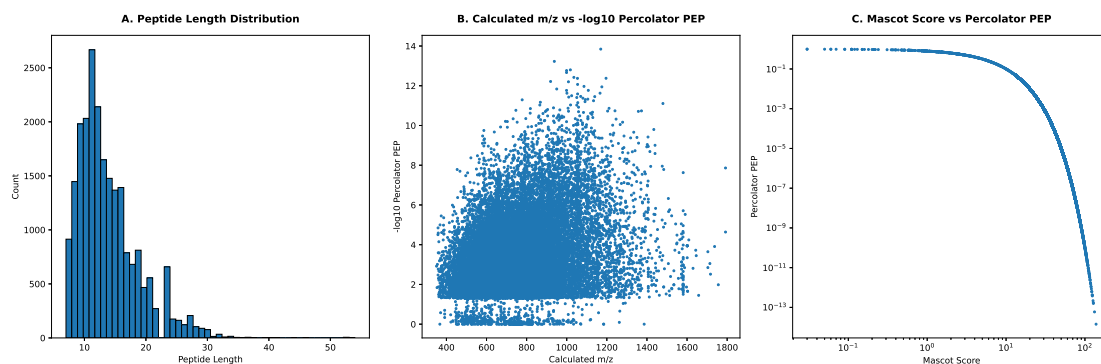


Figure 5.3: Quality control metrics for POP proteomic data. **A:** Peptide length distribution histogram showing the frequency distribution of peptide lengths, with peak frequency around 10-15 amino acids. **B:** Scatter plot showing the relationship between calculated m/z values (x-axis, 400-1800) and $-\log_{10}$ Percolator PEP scores (y-axis, 0-14), displaying a broad distribution of peptide identifications. **C:** Scatter plot demonstrating the correlation between Mascot scores (x-axis, log scale) and Percolator PEP values (y-axis, log scale), showing a strong negative correlation.

provides a powerful foundation with a large parameter count for complex biological sequence modeling.

Our fine-tuning process employed a two-stage transfer learning approach on the prepared training dataset (48,000 sequences):

1. Initial Adaptation

- **Dataset:** The entire training set (48,000 sequences) was used for this stage.
- **Training parameters:** The classification head was trained for 4 epochs.
- **Batch size:** 32.
- **Optimizer:** AdamW (Loshchilov and Hutter 2017) (learning rate: $2e-5$, weight decay: 0.01).
- **Objective:** This stage aimed to quickly adapt the newly added classification layer to the ECM prediction task while keeping the pre-trained ESM-2 layers frozen.

2. Full Fine-Tuning

5 Translating Molecular Insights to Clinical Relevance

- Dataset: The entire training set (48,000 sequences) was again used.
- Training parameters: The entire ESM-2 model (including pre-trained layers and the classification head) was fine-tuned for 2 epochs.
- Batch size: 8 (with a gradient accumulation of 2, resulting in an effective batch size of 16).
- Optimizer: AdamW (learning rate: 1e-5, weight decay: 0.02).
- Objective: This stage allowed for more subtle adjustments across all layers of the pre-trained model, enabling it to capture highly specific features relevant to ECM protein classification.

The selection of these training parameters (epochs, batch size, AdamW optimizer with its learning rate and weight decay) was guided by empirical exploration. This approach aimed to mitigate catastrophic forgetting and prevent overfitting, common challenges when adapting large pre-trained models like ESM-2 to a specific classification task. While a formal hyperparameter optimization sweep was not conducted, these settings were carefully determined through preliminary experimentation to achieve satisfactory convergence and ensure the reproducibility of our fine-tuning process.

Throughout both stages, model performance was continuously monitored on a held-out validation set of 16,000 samples to assess convergence and generalization ability. The model checkpoint achieving the best performance on the validation set was saved for final evaluation.

5.3.2.2 Baseline Models

To benchmark EPOP’s performance, we developed two baseline deep learning models, both trained on the same balanced dataset (training set: 48,000 sequences, validation set: 16,000 sequences).

1. Transformer model

- Architecture: Our custom Transformer model consisted of a single transformer encoder layer with 8 attention heads and a hidden dimension (embedding size) of 128. Sequence tokenization used a vocabulary size of 27 (20 standard amino acids plus special tokens).

- Input: Tokenized amino acid sequences.
- Output: Binary classification (ECM-related or not) via a final dense layer.

2. LSTM model

- Architecture: This model featured a Bidirectional LSTM layer with 32 hidden units, followed by a dense classification layer.
- Input: Same tokenized amino acid sequences as the Transformer model.
- Output: Binary classification (ECM-related or not).

The Transformer model used the Adam optimizer with a learning rate of $5e-5$ and a batch size of 16. The LSTM model used the Adam optimizer with a learning rate of $1e-4$ and a batch size of 128. Both baseline models were trained on the full training dataset for 15 epochs.

The training parameters for the baseline models (epochs, batch size, and Adam optimizer with its learning rate) were empirically determined, also taking into account the practical limitations of the Graphics Processing Unit (GPU) environment, especially given the comparatively slower training speed of the LSTM model. These settings were found to facilitate effective convergence and robust performance for these deep learning architectures on the protein sequence classification task. Similar to EPOP, while a formal hyperparameter optimization sweep was not performed, these choices were informed by preliminary experimentation to ensure satisfactory results and reproducibility.

5.3.3 Prediction and Functional Analysis of POP-Associated ECM Proteins

We applied the EPOP model to predict the ECM-relatedness of peptides identified in the POP proteomic dataset. Our analysis pipeline systematically processed this data to yield a high-confidence list of POP-associated ECM proteins, which then underwent further functional analysis.

Our analysis pipeline consisted of the following steps:

1. Peptide-level Prediction: The trained EPOP model was applied to each peptide sequence identified in the PRIDE dataset (PXD011467), yielding a predicted probability of ECM-relatedness.

5 Translating Molecular Insights to Clinical Relevance

2. **Threshold Application:** Peptides with predicted probabilities greater than 0.168 were selected as high-confidence candidates. This threshold was previously optimized on our validation dataset to maximize the F1-score, ensuring a balanced consideration of both precision and recall for ECM classification.
3. **Protein Identification:** A multi-step process was employed to map these high-confidence peptides to their corresponding proteins and consolidate the results.
 - **Peptide-to-Protein Mapping:** Initial associations between PSMs and protein accessions (e.g., Ensembl IDs) were extracted directly from the raw MaxQuant processed XML data.
 - **Peptide Deduplication and Consolidation:** Redundant peptide entries were deduplicated, representing each unique peptide by its best EPOP score. All associated protein accessions for multi-mapping peptides were retained and consolidated.
 - **Ensembl to UniProt Mapping:** Consolidated protein accessions (primarily Ensembl IDs) were mapped to UniProtKB entries using the UniProt ID mapping service, facilitating retrieval of standardized annotations.
 - **UniProt Annotation Retrieval:** Detailed functional annotations for all mapped UniProt proteins were retrieved from the UniProt REST API.
 - **Protein-level Aggregation:** Peptide-level predictions and UniProt annotations were aggregated. For each unique protein (UniProt ID), the maximum EPOP score among its corresponding peptides was assigned as its representative ECM probability, yielding a final protein-centric table.
4. **Functional Enrichment Analysis:** To gain deeper biological insights into the identified POP-associated ECM proteins, functional enrichment analysis was performed using Enrichr (Zhuorui Xie et al. 2021). Genes corresponding to the predicted proteins were analyzed for overrepresentation in established biological pathways and processes, leveraging various public databases including GO, KEGG (Kanehisa et al. 2025), and Reactome (Milacic et al. 2024). This analysis was performed on the full set of identified proteins, with subsequent interpretation focusing on terms and pathways explicitly related to ECM.

5. Literature-based Validation: To assess the biological relevance and potential novelty of our predicted ECM proteins, we conducted a comprehensive literature review. This process involved manual searches using Google Scholar and PubMed, combining general terms such as “pelvic organ prolapse”, “extracellular matrix”, “collagen”, “elastin”, “fibronectin”, and “tenascin” with specific protein names or UniProt identifiers derived from our proteomic predictions. For each protein, we carefully examined the retrieved publications to determine whether there was clear evidence of a functional role or direct association with POP. This systematic review allowed us to classify our predicted ECM proteins into those previously reported in POP literature and novel candidates for the condition.

This multi-faceted comparison allowed us to assess both the reproducibility of our findings and the potential of our approach to uncover novel POP-associated ECM proteins.

5.3.4 Model Interpretability Analysis

To gain insights into the sequence features driving EPOP’s predictions, we conducted three complementary analyses on high-confidence ECM-related peptides:

1. Single amino acid importance:
 - Method: Iteratively masked each position in the peptide sequence.
 - Metric: Change in predicted probability compared to original sequence.
 - Visualization: Line plots of importance scores.
2. Amino acid substitution impact:
 - Method: For each position, we systematically tested all possible single amino acid substitutions.
 - Metric: Change in predicted probability for each substitution.
 - Visualization: Heatmaps of substitution effects.
3. Sliding window analysis:
 - Method: Extracted fixed-length subsequences and obtained predictions for each window.

5 Translating Molecular Insights to Clinical Relevance

- Metric: Predicted ECM-relatedness probability for each window.
- Visualization: Score curves along peptide sequences.

These analyses provide a comprehensive view of the sequence determinants of ECM function, offering valuable insights for future experimental studies and potential therapeutic interventions.

By integrating cutting-edge AI techniques with rigorous biological validation and interpretability analyses, our methodology offers a powerful framework for uncovering novel insights into the molecular basis of POP and other ECM-related disorders.

5.4 Results

5.4.1 Performance Evaluation of ECM Function Prediction Models

The performance of the EPOP model on the ECM function prediction task was comprehensively evaluated using an independent test set of 16,000 samples (8,000 positive and 8,000 negative) that did not overlap with the training data.

EPOP achieved outstanding performance on the test set, with an accuracy of 99.40%, precision of 99.50%, recall of 99.30%, and an F1-score of 99.40%. These results demonstrate the model's robust ability to distinguish between ECM-related and non-ECM proteins with extremely high accuracy and sensitivity, showcasing its strong generalization ability.

To further assess the superiority of EPOP, we compared its performance with two baseline models: a Transformer model and an LSTM model. As shown in Table 5.1, EPOP significantly outperformed both baseline models across all evaluation metrics. The EPOP model improves accuracy by 10.81% compared to the Transformer model and by 21.71% compared to the LSTM model, highlighting the advantages of leveraging pretrained protein language models in capturing evolutionary information and structural patterns from sequences.

A more detailed evaluation, particularly focusing on the balance between precision and recall, is provided by the F1-score. As seen in Table 5.1, EPOP (ESM-2) achieved an F1-score of 99.40%, indicating an excellent trade-off between precision (99.50%) and recall (99.30%). In contrast, the Transformer's F1-score was 87.74%, reflecting a

Table 5.1: Performance comparison of ECM function prediction models on the test set.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
EPOP (ESM-2)	99.40	99.50	99.30	99.40
Transformer	88.59	94.78	81.67	87.74
LSTM	77.69	74.52	84.15	79.04

disparity between its high precision (94.78%) and moderate recall (81.67%), while the LSTM, despite achieving a recall of 84.15%, showed lower precision (74.52%) leading to an F1-score of only 79.04%, suggesting a higher false positive rate. This comprehensive analysis demonstrates EPOP’s superior ability to maintain high performance across multiple critical metrics compared to baseline methods.

To provide a comprehensive visualization of EPOP’s performance characteristics, we present a figure encompassing key evaluation aspects. This includes the PR curve (Fig 5.4A) and the ROC curve (Fig 5.4B) at different probability thresholds. EPOP achieved an area under the PR curve (AUPRC) of 0.999 and an area under the ROC curve (AUROC) of 0.999, demonstrating near-perfect performance. These metrics confirm EPOP’s outstanding ability to identify ECM-related proteins, maintaining high accuracy and sensitivity even in the presence of imbalanced positive and negative sample distributions.

Furthermore, Fig 5.4C illustrates the distinct distribution of EPOP’s prediction scores for both ECM and non-ECM proteins on the test set. The model exhibited a clear separation of scores, with positive (ECM) class scores averaging 0.992 (median: 1.000) and negative (non-ECM) class scores averaging 0.005 (median: 0.000). This pronounced distinction, quantified by a score separation of 0.987 (positive-negative), underscores the model’s high predictive confidence. Fig 5.4D presents the confusion matrix at the optimal classification threshold of 0.168. At this threshold, the model achieved high True Positives and True Negatives, alongside very low False Positives and False Negatives, confirming its exceptional accuracy in classifying both positive and negative instances.

To complement the primary performance metrics and guide practical application, we further analyzed model performance across varying classification thresholds and examined the cumulative score distributions. Fig 5.4E illustrates how accuracy, precision, recall, and F1-score vary across different probability thresholds, highlighting the inherent trade-offs. An optimal threshold of 0.168 was identified, which maximizes the F1-score.

5 Translating Molecular Insights to Clinical Relevance

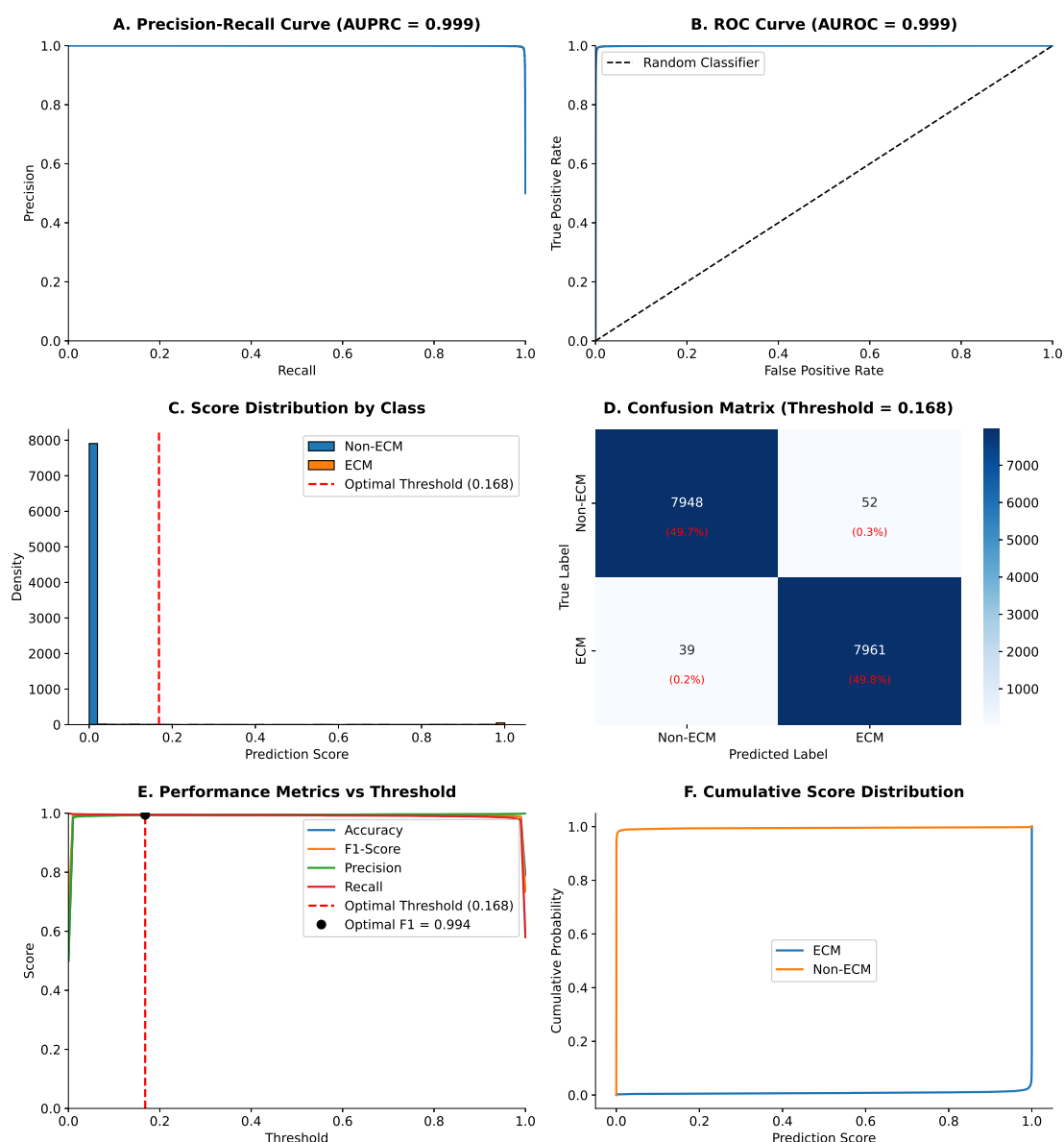


Figure 5.4: This figure presents a detailed evaluation of EPOP’s performance on the test set, combining core classification metrics with a thorough analysis of threshold effects. A: PR curve (AUPRC = 0.999) demonstrating near-perfect performance with high precision maintained across most recall values. **B:** ROC curve (AUROC = 0.999) illustrating excellent classification performance, substantially outperforming random classification. **C:** Score distribution by class, showing distinct prediction score distributions for ECM and non-ECM proteins. **D:** Confusion matrix at the optimal classification threshold (0.168), providing a detailed breakdown of classification outcomes. **E:** Performance metrics (Accuracy, F1-score, Precision, Recall) across varying classification thresholds, highlighting the optimal threshold (0.168) that maximizes F1-score. **F:** Cumulative score distributions for ECM and non-ECM proteins, visually demonstrating the model’s high confidence and clear class separation.

Correspondingly, Fig 5.4F displays the cumulative distribution functions for both ECM and non-ECM scores, further emphasizing the clear separation in prediction confidence and reinforcing the choice of optimal threshold.

5.4.2 Prediction Results of POP-Associated ECM Proteins

The EPOP model was applied to mass spectrometry-based proteomic data from POP patients and healthy controls (PRIDE: PXD011467) to predict the ECM-relatedness of identified peptides. High-confidence peptides with predicted probabilities above 0.168 were selected, a threshold optimized for maximizing F1-score. These peptides were subsequently mapped to proteins through a multi-step pipeline, consolidating unique UniProt entries.

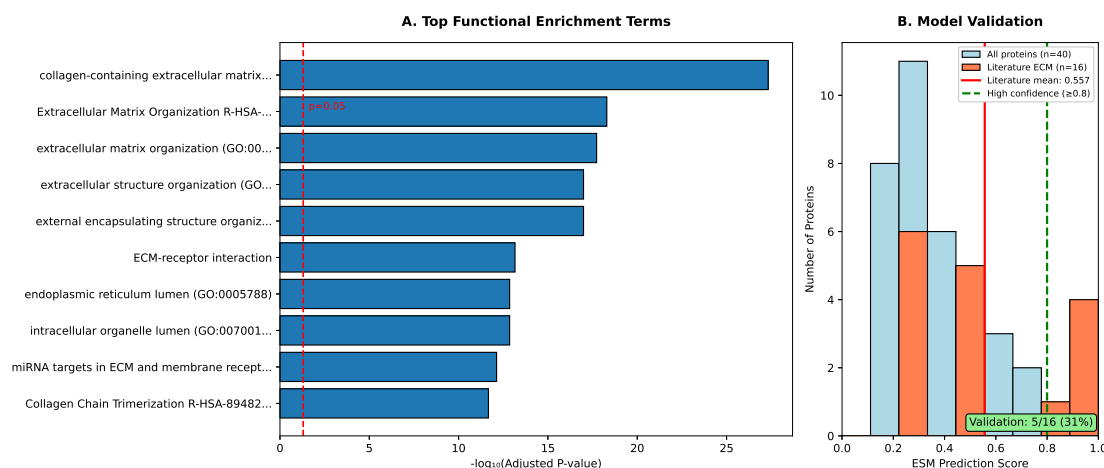


Figure 5.5: Key aspects of EPOP's protein identification and validation. A: Top functional enrichment terms identified among the 40 predicted ECM proteins. **B:** Distribution of EPOP prediction confidence for the identified 40 proteins, highlighting the confidence levels of both literature-supported and novel candidate proteins.

This pipeline identified 63 unique high-confidence peptides. When aggregated to the protein level, these peptides yielded a final set of 40 unique, reviewed UniProt proteins confidently predicted by EPOP as ECM proteins. The distribution of EPOP prediction confidence for these 40 identified proteins is visually presented in Figure 5.5B, showcasing a strong bias towards high confidence scores. The mean EPOP score for these proteins was 0.451, with 6 proteins achieving a score of 0.8 or higher. These

5 Translating Molecular Insights to Clinical Relevance

identified ECM proteins offer crucial insights into the molecular landscape of POP within the context of patient samples.

To biologically validate EPOP's predictions, we conducted a comprehensive functional enrichment analysis on the 40 identified proteins using Enrichr. As anticipated, these proteins were overwhelmingly enriched in GO terms and biological pathways directly related to the extracellular matrix. The most compelling evidence emerged from the GO Biological Process 2021 database, where "extracellular matrix organization" (GO:0031012) ranked as the top enriched term with an exceptionally low Adjusted P-value of $1.88\text{e-}18$. This profound enrichment strongly confirms the core identity of our predicted proteins as central to ECM dynamics.

Furthermore, other highly significant and biologically pertinent enrichments were consistently observed across various databases. For instance, GO Cellular Component 2021 highlighted "collagen-containing extracellular matrix" (GO:0062023, Adjusted P-value: $4.72\text{e-}28$) and "endoplasmic reticulum lumen" (GO:0005788), reflecting the synthesis and localization of ECM components. Pathway analyses further solidified these findings: KEGG 2021 Human showed strong enrichment in "ECM-receptor interaction" and "Focal adhesion", underscoring the critical interplay between cells and their surrounding matrix. Similarly, Reactome 2022 terms like "Extracellular Matrix Organization", "Collagen Chain Trimerization", and "Collagen Biosynthesis And Modifying Enzymes" were highly significant, detailing the intricate processes of collagen formation and assembly. WikiPathways 2019 Human also pointed to "miRNA targets in ECM and membrane receptors" and "Focal Adhesion" pathways. The comprehensive results of these enrichment analyses are summarized in Figure 5.5A, visually demonstrating the extensive network of ECM-related functions and pathways represented by our identified proteins. In total, 148 significant terms were found across all queried databases, with 15 explicitly recognized as ECM-related, providing robust biological validation for EPOP's predictions.

We further cross-validated our identified ECM proteins against existing literature concerning POP pathogenesis. Our comprehensive set of 40 predicted proteins was categorized based on their prior association with POP in published studies (Table 5.2 and Table 5.3).

Table 5.2 summarizes 16 proteins previously investigated in the context of POP, revealing diverse associations including various expression patterns, structural alterations,

or observations of no significant change in affected pelvic tissues. This set largely comprises fundamental ECM components and key regulatory molecules. Our model successfully identified several collagen chains, which are critical for tissue mechanical integrity (X. Wang et al. 2002; M. Page, Hogg, and Ashhurst 1986; Drishtant Singh, Rai, and Agrawal 2023). Literature reports varied expression profiles for these collagens, ranging from downregulation to no significant change. Such findings underscore the complex and heterogeneous nature of collagen remodeling in POP, where specific collagen subtypes exhibit distinct responses to disease progression (X. Wang et al. 2002; M. Page, Hogg, and Ashhurst 1986; Drishtant Singh, Rai, and Agrawal 2023). Furthermore, our predictions include Fibronectin (P02751) and Tenascin (P24821), consistently reported as upregulated, and Vitronectin (P04004) and Fibromodulin (Q06828), which showed downregulation. These proteins are known to be involved in tissue repair, cell adhesion, and matrix organization (Drishtant Singh, Rai, and Agrawal 2023; Shao et al. 2020). The varied expression patterns of these non-collagenous ECM proteins suggest a dynamic and intricate process of tissue repair and compensatory responses to altered biomechanics in POP (Drishtant Singh, Rai, and Agrawal 2023; Shao et al. 2020; Hynes and Naba 2012). The accurate identification of these established ECM players, encompassing a spectrum of reported involvements, affirms our model’s robust capacity to recognize proteins central to POP pathogenesis and validates its biological relevance (X. Wang et al. 2002; M. Page, Hogg, and Ashhurst 1986; Drishtant Singh, Rai, and Agrawal 2023).

Table 5.3: Proteins not explicitly associated with POP in the provided literature. Functional annotations were adapted from UniProtKB (“UniProt: the universal protein knowledgebase in 2025” 2025).

UniProt ID	Gene	Function
Q9HDC9	APMAP	Involved in adipocyte differentiation and related to fat metabolism, which may indirectly influence tissue composition.
Q15063	POSTN	Matrisome-associated protein promoting cell adhesion and ECM restructuring, enhancing collagen fibril formation and wound repair.

5 Translating Molecular Insights to Clinical Relevance

UniProt ID	Gene	Function
Q14112	NID2	Cell adhesion glycoprotein, widely distributed in basement membranes, binds to collagens I and IV, perlecan, and laminin 1, likely playing a role in cell-extracellular matrix interactions.
P62820	RAB1A	Key regulator of intracellular membrane trafficking to the cell surface, playing a role in cell adhesion and migration through protein transport.
P55290	CDH13	Calcium-dependent cell adhesion protein connecting cells, potentially contributing to cell type sorting.
P46939	UTRN	Primarily involved in anchoring the cytoskeleton to the plasma membrane.
P43121	MCAM	Cell adhesion glycoprotein mediating endothelial monolayer cohesion in vascular tissue, relevant for cell-ECM interactions in vascular remodeling.
P26038	MSN	Connects actin cytoskeleton to plasma membrane, regulating cell shape and membrane transport, impacting cellular structural integrity.
P25092	GUCY2C	Guanylyl cyclase involved in cGMP synthesis, a signaling molecule.
P23946	CMA1	Secreted protease of mast cells, implicated in extracellular matrix degradation and inflammation.
P06681	C2	Precursor in the complement pathway, mediating pathogen breakdown and adaptive immune system strengthening, relevant for inflammatory responses.
P02768	ALB	Main function is blood colloidal osmotic pressure regulation and transport, indirectly influencing tissue fluid balance.

UniProt ID	Gene	Function
P02749	APOH	Binds negatively charged substances like phospholipids, potentially influencing blood coagulation.
Q9UBG3	CRNN	Promotes cell proliferation and cell cycle progression, regulating proliferation via inflammatory cytokine responses.
Q9H0U4	RAB1B	Key regulator of intracellular membrane trafficking for vesicle formation and fusion.
P02679	FGG	Forms insoluble fibrin matrix, crucial for hemostasis and early wound repair, stabilizing lesions and guiding cell migration.
P00751	CFB	A precursor of the catalytic component of the alternative complement pathway's C3 and C5 convertase complexes, CFB leads to pathogen phagocytosis and breakdown, and strengthens the adaptive immune system upon cleavage by CFD.
P00746	CFD	A serine protease that initiates the alternative pathway of the complement system by cleaving CFB, leading to pathogen phagocytosis and breakdown, and strengthening the adaptive immune system.
P00505	GOT2	Catalyzes L-kynurenine transamination, important for NAD(H) redox balance, amino acid metabolism, and fatty acid uptake.
O00159	MYO1C	Actin-based motor molecule for intracellular movements, linking actin cytoskeleton to membranes and regulating vesicular transport.
Q9UGT4	SUSD2	Putative cytokine receptor with tumor suppressor activity, may regulate cell proliferation.

5 Translating Molecular Insights to Clinical Relevance

UniProt ID	Gene	Function
P01024	C3	Complement C3 is a central component of the complement system, contributing to pathogen breakdown and immune response, processes that are indirectly linked to inflammation and tissue remodeling within the extracellular matrix.
Q9Y6Q9	NCOA3	A nuclear receptor coactivator that stimulates hormone-dependent transcriptional activities and participates in the NF-kappa-B pathway.
P08294	SOD3	Protect the extracellular space from toxic effect of reactive oxygen intermediates by converting superoxide radicals into hydrogen peroxide and oxygen.

Table 5.3 presents 24 proteins that, to the best of our knowledge, have not been explicitly associated with POP in the current literature. Nevertheless, their annotated biological functions, retrieved from UniProt, suggest plausible involvement in ECM integrity, cell adhesion, tissue remodeling, immune modulation, or inflammation, positioning them as promising novel candidates for further POP-related investigation (“UniProt: the universal protein knowledgebase in 2025” 2025).

Notably, several proteins are directly implicated in ECM architecture or cell–matrix interactions (Shao et al. 2020). For example, POSTN (Q15063) is a well-characterized matrisome-associated protein. Research indicates that periostin is associated with extracellular matrix remodeling during wound repair, where it is linked to enhanced cell adhesion and promotes fibrogenic differentiation (Jackson-Boeters, W. Wen, and Hamilton 2009; Shao et al. 2020; Hynes and Naba 2012). These functions are highly pertinent to pelvic floor tissue remodeling. Similarly, NID2 (Q14112), a basement membrane glycoprotein, profoundly influences matrix cohesion through its interactions with multiple ECM components like collagens I, IV, and laminin, suggesting a central scaffolding role in matrix cohesion (Shao et al. 2020; Hynes and Naba 2012). MCAM (P43121) and CDH13 (P55290) further contribute to crucial cell–cell and cell–ECM adhesion

processes, which are fundamental for maintaining tissue structural integrity (Kozlova and Sytnyk 2024).

In addition, several proteins participate in inflammation and immune-related matrix remodeling. For instance, CMA1 (P23946), a mast cell-derived protease, is directly implicated in ECM degradation and the regulation of inflammatory responses, with studies showing its extensive effects on ECM remodeling, including the suppression of ECM-related genes and degradation of fibronectin (X. O. Zhao et al. 2022). Furthermore, core components of the complement cascade, including C3 (P01024), CFB (P00751), and CFD (P00746), modulate immune responses and are increasingly recognized for their involvement in tissue remodeling processes (Pejler et al. 2010).

Proteins involved in oxidative stress response or vesicle transport also emerge as functionally relevant. SOD3 (P08294), for instance, provides critical defense against oxidative matrix damage in inflamed tissues by neutralizing extracellular reactive oxygen species (Fujiwara et al. 2016). Meanwhile, RAB1A (P62820) and RAB1B (Q9H0U4), though classically recognized as intracellular trafficking regulators, are involved in protein transport pathways essential for secretion and membrane dynamics, thereby indirectly influencing cell adhesion and migration (Hongyuan Jin et al. 2021; Guadagno and Progida 2019).

Other notable identified proteins encompass diverse functions that may contribute to tissue homeostasis. UTRN (P46939) and MSN (P26038) are actin-binding proteins crucial for anchoring the cytoskeleton to the plasma membrane, impacting cellular structural integrity (Dowling et al. 2021; Ruo Wu et al. 2022). GUCY2C (P25092), a guanylyl cyclase, plays signaling roles in intestinal epithelial cells by orchestrating intestinal homeostatic mechanisms (Pattison et al. 2016). Even metabolic enzymes such as GOT2 (P00505) and ALB (P02768), while not canonical ECM components, may exert systemic effects on matrix composition through redox or fluid balance regulation (Tatsukawa and Hitomi 2021; Gonzales et al. 2022).

Collectively, the identification of these 24 proteins, many with functions related to the extracellular matrix, immune modulation, or structural support that have not yet been directly studied in the context of POP, demonstrates the unique potential of the EPOP model to reveal biologically relevant but previously unrecognized molecular candidates. These results provide a valuable foundation for future mechanistic investigations and the discovery of novel biomarkers in POP research beyond currently known associations.

5.4.3 Model Interpretability Analysis

To gain deeper insights into the sequence features driving EPOP's predictions, we conducted a series of complementary interpretability analyses on 16 literature-validated ECM proteins identified in our study (listed in Table 5.2). These analyses aimed to pinpoint critical amino acids, mutation-sensitive positions, and functionally important regions within their peptide sequences. We present the detailed results for the Collagen Alpha-1(VI) chain (P12109) as a representative example, given its high confidence score (0.9974) and established relevance in ECM (Fig 5.6).

5.4.3.1 Single Amino Acid Importance Analysis

We systematically evaluated the importance of each amino acid at every position for predicting protein function. By iteratively masking individual amino acids in the sequence and observing the resulting change in predicted ECM probability, we quantified the contribution of each residue to the model's discrimination.

For the representative Collagen Alpha-1(VI) Chain (P12109), the results (Fig 5.6A) showed varying importance across its sequence. Notably, residues such as Isoleucine (I) at position 26 (importance: 0.0090), Methionine (M) at position 27 (importance: 0.0078), and Glutamic acid (E) at position 21 (importance: 0.0056) exhibited consistently higher importance scores. Other impactful residues included Proline (P) at position 9 (importance: 0.0053) and Proline (P) at position 2 (importance: 0.0051). This suggests these residues, even those outside the classical Gly-X-Y motif, contribute significantly to the overall ECM functional prediction.

While previous literature highlights Gly-X-Y repeats as critical for collagen's triple-helical structure where Gly residues (e.g., Glycine at position 10) are essential (Bella et al. 1994; Shoulders and Raines 2009), our analysis also reveals the importance of other residues and regions (e.g., positions 26-27), potentially reflecting nuances in ECM protein folding or interactions learned by the model.

5.4.3.2 Amino Acid Mutation Effect Analysis

To explore the sensitivity of critical positions to sequence alterations and their role in maintaining protein function, we systematically evaluated the impact of substituting each

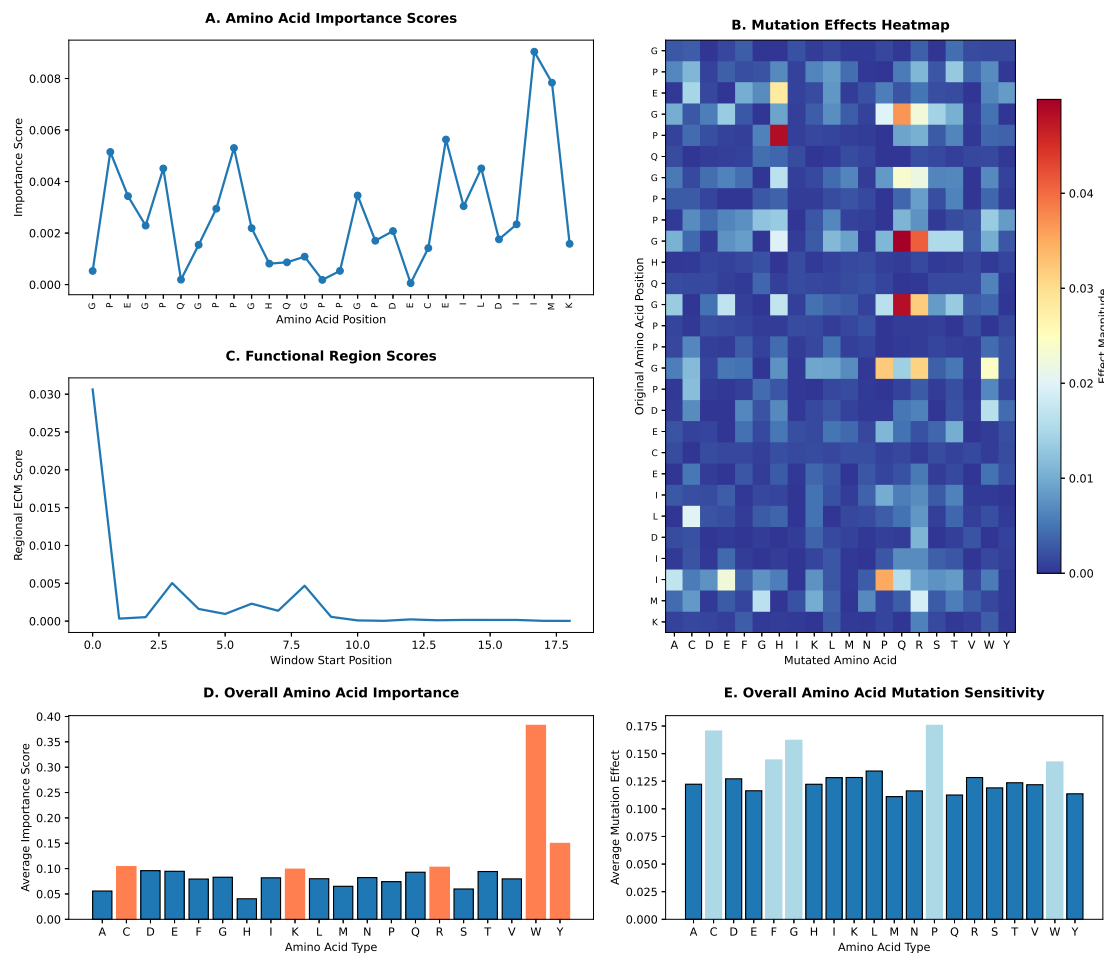


Figure 5.6: Interpretability Analysis of Collagen Alpha-1(VI) Chain (P12109) and Aggregated Insights Across Multiple ECM Proteins. **A:** Amino acid importance score plot showing varying contributions of different residues, with peaks indicating highly important positions. **B:** Mutation effects heatmap displaying the impact of amino acid substitutions at each position, with color intensity representing effect magnitude (blue to red scale). **C:** Functional region score plot showing the prediction scores across different sequence windows, revealing regions of particular functional importance. **D:** Overall Amino Acid Importance. Average importance score for each amino acid type, aggregated across all positions in all 16 proteins. **E:** Overall Amino Acid Mutation Sensitivity. Average mutation effect for each amino acid type, aggregated across all positions in all 16 proteins.

5 Translating Molecular Insights to Clinical Relevance

amino acid at every position with all possible alternative amino acids on the prediction results. The mutation effect heatmap for P12109 is presented in Fig 5.6B.

The analysis revealed that at key positions, even single amino acid substitutions could significantly alter the predicted ECM probability. For instance, at position 10, a Glycine (G) to Glutamine (Q) mutation (G→Q) caused the largest decrease in ECM prediction probability (effect: 0.0499), followed closely by Glycine to Arginine (G→R) (effect: 0.0411). Similarly, a Proline (P) at position 5 mutating to Histidine (H) (P→H) showed a strong effect (0.0485).

These findings underscore the critical conservation of residues like Glycine and Proline, which are essential for the structural integrity and stability of collagen's triple helix (Ramshaw, Shah, and Brodsky 1998; Brodsky et al. 2008). This alignment with previous site-directed mutagenesis and functional studies (Kuivaniemi, Tromp, and Prockop 1991) demonstrates that the model accurately captures biologically relevant mutational sensitivities that impact collagen folding, assembly, and ECM function.

5.4.3.3 Sliding Window Analysis

To investigate the contribution of local regions to overall protein function, we employed a sliding window approach to scan the peptide sequences and examine the functional prediction scores of different windows. The sliding window score curve for sequence P12109 is shown in Fig 5.6C.

The results revealed specific regions with consistently high prediction scores, indicating areas of particular functional importance. For P12109, the region from positions 1-10 (GPEGPQGPPG) exhibited the highest regional ECM score (0.0306), suggesting it carries key sequence features. Further comparison of these high-scoring regions with resolved collagen protein structures in the Protein Data Bank (PDB) database indicates that they frequently correspond to repeating motifs involved in forming the collagen triple-helix structure, as well as regions containing critical functional sites such as cleavage sites and ligand-binding sites (Orgel et al. 2006).

5.4.3.4 Aggregated Interpretability Insights Across Multiple ECM Proteins

Beyond the detailed single-protein analysis, we performed an aggregated interpretability analysis across all 16 literature-validated ECM proteins to identify common sequence determinants influencing EPOP's predictions (Fig 5.6).

As shown in Fig 5.6D), the aggregated analysis of single amino acid importance revealed that Tryptophan (W) consistently held the highest average importance score (0.3825) across the 16 proteins, significantly higher than other amino acids. This highlights Tryptophan's critical role in ECM protein function or recognition. Other highly important amino acids included Tyrosine (Y) (0.1497), Cysteine (C) (0.1041), and Arginine (R) (0.1027). Interestingly, while Glycine (G) is fundamental to collagen structure, its average importance (0.0830) was lower than these specific residues, suggesting that while indispensable for core structure, the model might assign higher discriminative power to rarer or more functionally specialized amino acids.

Also, as shown in Fig 5.6E), the analysis of mutation effects revealed that Proline (P) exhibited the highest overall average mutation sensitivity (0.1757), meaning that substitutions involving Proline (either as the original or mutated amino acid) generally caused the largest changes in prediction probability. Other highly sensitive amino acids included Cysteine (C) (0.1705), Glycine (G) (0.1621), and Phenylalanine (F) (0.1443). The high sensitivity of Proline and Glycine aligns with their crucial structural roles in collagen. The sensitivity of Cysteine suggests the importance of disulfide bonds in many ECM proteins, while Phenylalanine's impact may relate to hydrophobic interactions critical for protein folding and stability. These aggregated findings provide a broader understanding of sequence determinants universally recognized by EPOP across various ECM proteins.

These results not only demonstrate the potential of machine learning methods in revealing protein sequence-function relationships but also provide valuable directions and clues for experimental research and drug development in the context of POP and other ECM-related disorders.

5.5 Discussion

5.5.1 Implications for Understanding POP Pathogenesis and ECM Roles

Our study represents a significant advancement in unraveling the molecular underpinnings of POP, a condition that affects millions of women worldwide (R. Gong and Z. Xia 2019). Using the power of advanced AI techniques, specifically the ESM-2 protein language model, we have developed EPOP, a highly accurate tool for predicting and characterizing POP-associated ECM proteins.

The ECM plays a crucial role in maintaining the structural integrity and function of pelvic floor tissues (Alarab et al. 2014). Our findings provide a more nuanced understanding of ECM alterations in POP, stemming from the identification of 40 unique ECM proteins within patient proteomic samples. This intricate pattern suggests a fundamental disruption of the ECM structural framework affecting the biomechanical properties of pelvic floor tissue. For instance, the literature-reported downregulation of specific collagen chains, such as Collagen alpha-1(VI) chain (P12109) in POP patients, suggests a direct contribution to the comprehensive weakening of the ECM structural framework. Each collagen subtype serves distinct mechanical functions – type I provides tensile strength, type III offers elasticity, and types V and VI regulate fibril organization – indicating that POP involves the multifaceted deterioration of tissue integrity rather than isolated protein defects (Nimni and Harkness 2018; Sorushanova et al. 2021; B. Sun 2021; Rossert and Crombrughe 2002; Kuivaniemi and Tromp 2019; Darvish 2022). This aligns with previous studies reporting decreased collagen content in POP tissues (R. Gong and Z. Xia 2019), and our results further refine this understanding by identifying specific collagen subtypes affected.

Conversely, the reported upregulation of tenascin (P24821) and fibronectin (P02751) in POP patients indicates an active, albeit potentially maladaptive, remodeling process. These matricellular proteins are known to play pivotal roles in tissue repair and cell-ECM interactions (Guler and Roovers 2022), suggesting their increased expression may represent a compensatory attempt to stabilize the weakened pelvic floor tissues amidst altered biomechanics. Similarly, the observed downregulation of Vitronectin (P04004) and Fibromodulin (Q06828) in POP tissues highlights further complex dysregulations impacting tissue integrity and repair.

In addition to validating known proteins, our study identified 24 new ECM proteins potentially associated with POP. These proteins have not been widely reported in POP-related studies but are known to play important roles in ECM structure and function. For example, POSTN (Q15063) helps organize the extracellular matrix and supports collagen fiber formation and wound healing. NID2 (Q14112) is a key glycoprotein in basement membranes, contributing to matrix stability and cohesion. SOD3 (P08294) protects tissues from oxidative damage, especially in inflamed environments. Also, Some of the newly predicted proteins are involved in immune or inflammatory pathways that affect ECM remodeling. These include CMA1 (P23946), C3 (P01024), CFB (P00751), and CFD (P00746). Such proteins may contribute to matrix degradation through immune-mediated mechanisms. These findings expand our understanding of the molecular changes in POP and provide new targets for future research and biomarker development.

5.5.2 Advantages and Challenges of AI-Based Disease Protein Discovery

Our study showcases the immense potential of AI-based methods in advancing our understanding of complex diseases. The EPOP model's ability to predict ECM-related proteins with high accuracy (99.40%) and its superior performance compared to traditional machine learning approaches highlight the power of transfer learning and protein language models in capturing the intricate patterns within protein sequences.

The approach offers numerous advantages for protein discovery in disease contexts. The unbiased nature of our method enables the identification of novel POP-associated proteins beyond those targeted in hypothesis-driven approaches. EPOP's computational efficiency allows rapid screening of entire proteomes, making comprehensive analyses feasible at scales impractical with conventional techniques. The interpretability analyses we implemented provide valuable insights into sequence features driving predictions, bridging the gap between computational outputs and biological understanding. Additionally, by leveraging ESM-2's pre-training on millions of evolutionarily diverse protein sequences, EPOP captures subtle sequence-function relationships that would remain invisible to conventional analysis techniques (Z. Lin et al. 2023; Meier et al. 2021).

Despite these strengths, several limitations remain. A major limitation is the absence of direct wet-lab or histological validation for the newly identified ECM protein candidates.

5 Translating Molecular Insights to Clinical Relevance

Although our computational framework offers strong predictions of disease-associated proteins, their specific biological roles and clinical relevance in POP still require rigorous experimental confirmation.

Another important challenge lies in the construction of the negative training set. Our approach defined non-ECM proteins based on the absence of ECM-related GO annotations, further filtered by SignalP and TMHMM to exclude proteins with signal peptides or transmembrane domains. While this strategy enhances label clarity, it introduces the risk of annotation bias: proteins that are truly ECM-related but lack comprehensive or up-to-date annotations (e.g., newly discovered ECM regulators or atypical secreted factors) may have been mistakenly labeled as negative. Consequently, the model might be penalized during training for correctly identifying such proteins, potentially limiting its ability to generalize to unannotated or unconventional ECM components. This highlights a fundamental challenge in supervised learning based on imperfect biological annotations.

Furthermore, the interpretability analyses, while insightful, are also subject to potential biases originating from both the model and the datasets. The ESM-2 model, pre-trained on a broad protein corpus, may reflect inherent biases from its training data distribution, which could influence the specific sequence features highlighted as important for ECM-relatedness. Furthermore, biases within our specific datasets, such as potential limitations in UniProt annotations used for model training or the characteristics of the POP patient proteomic samples (e.g., sample size, heterogeneity, technical detection limits), might affect the generalizability and robustness of the interpretability findings. While our analyses offer valuable initial insights into ECM protein characteristics, these inherent data and model biases should be considered when interpreting the precise sequence determinants and warrant further validation in diverse contexts.

Moreover, ESM-2 has inherent limitations when applied to ECM proteins, their high molecular weight, repetitive motifs, and extensive post-translational modifications may not be fully captured by sequence-based models alone, potentially affecting prediction accuracy for certain protein classes.

5.5.3 Future Directions

To further advance our understanding of POP and translate our findings into clinical applications, we propose several key directions for future research.

Initial efforts should focus on experimentally validating the newly identified POP-associated proteins, to determine their biological relevance and clinical potential. We propose a stepwise experimental pipeline combining multiple approaches: (i) immunohistochemical staining of pelvic support tissues from POP patients and healthy controls to assess spatial expression patterns; (ii) quantification of protein levels using Western blot (S. C. Taylor and Posch 2014) on tissue lysates or biological fluids; (iii) in vitro functional assays using mechanically stressed fibroblasts (Kessler et al. 2001), with knockdown or overexpression of target proteins, to evaluate their role in ECM remodeling; and (iv) in vivo validation using animal models of pelvic floor dysfunction to assess their impact on tissue integrity and disease progression. This multi-level approach will help establish the mechanistic roles of these proteins in POP pathogenesis and evaluate their suitability as biomarkers or therapeutic targets.

Moving beyond individual proteins, we can integrate our proteomic data with transcriptomic and metabolomic profiles from matched POP samples. This multi-omics approach will help us map the regulatory networks controlling ECM remodeling and identify the upstream factors driving the ECM changes we observed. Rather than merely addressing the downstream manifestations of ECM dysfunction, this systems biology perspective could uncover novel intervention points that address the root molecular causes of pelvic floor weakening.

We also can enhance our computational analyses using advanced interpretability methods that can extract biological meaning from our AI model's predictions. By analyzing how specific protein domains and structural motifs influence our model's classification decisions, we expect to uncover previously unrecognized protein interaction networks crucial for maintaining pelvic floor integrity. These computational insights will guide our experimental focus and may reveal unique binding interfaces for targeted drug development.

Also, we see considerable value in applying the EPOP framework to related pelvic floor disorders, particularly stress urinary incontinence. Our preliminary analyses suggest overlapping molecular mechanisms between these conditions. A systematic comparison

5 Translating Molecular Insights to Clinical Relevance

could identify core pathological processes common to multiple pelvic floor disorders, potentially leading to broad-spectrum therapeutic approaches that address several conditions simultaneously.

Beyond direct applications to closely related pelvic floor disorders, the EPOP framework, built on ESM-2, also holds strong potential for extension to other, more diverse ECM-related disorders, including fibrotic diseases (Wynn 2008), cardiovascular conditions involving ECM remodeling (Kong, Christia, and Frangogiannis 2014), and certain cancers (Pickup, Mouw, and Weaver 2014). However, applying the model to different disease contexts or populations requires careful adaptation. ECM structure and molecular mechanisms vary across diseases, and a model trained on POP data may not generalize directly. Likewise, factors such as genetic background, age, and ethnicity influence ECM composition and disease manifestation, underscoring the need for validation in diverse cohorts. For new applications, the classification head of EPOP should be fine-tuned using disease-specific, high-quality proteomic datasets. Integration of relevant biological features and thorough experimental validation would also be essential to ensure prediction accuracy and clinical relevance.

Longitudinal studies tracking changes in ECM protein expression throughout disease progression could identify early biomarkers for POP risk assessment. Such markers might enable preventive interventions before significant tissue deterioration occurs, particularly in high-risk populations with genetic predispositions or following traumatic childbirth.

The development of targeted therapeutic strategies represents the ultimate translational goal of this research. Approaches might include promoting the expression of downregulated collagens, modulating the activity of matrix metalloproteinases, or enhancing crosslinking of existing ECM components to improve tissue strength and resilience. The integration of these research directions could significantly advance our understanding of POP pathogenesis and lead to novel diagnostic and therapeutic approaches.

Despite these promising avenues, there are several barriers to clinical translation, including the current lack of routine ECM protein testing in clinical practice. However, as personalized medicine advances and the importance of the ECM in various diseases becomes more recognized, the integration of proteomic research in clinical settings could yield significant benefits for millions of women affected by this condition worldwide.

In conclusion, our study demonstrates the power of AI-driven approaches in uncovering novel insights into complex diseases like POP. By combining advanced machine

learning techniques with rigorous biological validation, we have not only deepened our understanding of POP pathogenesis but also paved the way for future research into ECM-related disorders. As we continue to refine and expand this approach, we anticipate significant advancements in the diagnosis, prognosis, and treatment of POP.

5.6 Conclusion

In this study, we have demonstrated the remarkable potential of AI-based methods, specifically the ESM-2 protein language model, in predicting POP-associated ECM proteins and elucidating their functional roles in disease pathogenesis. By leveraging transfer learning and fine-tuning the ESM-2 model on expertly curated datasets, we developed EPOP, a highly accurate and efficient model for identifying proteins with high ECM functional relevance from their amino acid sequences.

The application of EPOP to POP patient-derived proteomic data revealed a set of novel POP-associated ECM proteins, potentially playing crucial roles in ECM remodeling and metabolic dysregulation underlying POP pathogenesis. Functional enrichment analyses supported the biological significance of these predicted proteins by highlighting their involvement in key ECM-related processes and pathways.

A comparison of our predicted POP-associated ECM proteins with previous research findings provided strong evidence for their biological relevance. The substantial overlap with known associations reported in previous studies, as well as the identification of novel proteins implicated in other ECM-related disorders, underscores the value of our approach in uncovering new insights into POP pathogenesis.

Our study highlights the immense potential of AI-based methods in advancing our understanding of complex diseases and facilitating the discovery of novel disease-related proteins. The ability of deep learning models to learn rich representations from vast amounts of protein sequence data has revolutionized the field of protein function prediction. Our fine-tuning approach demonstrates the feasibility and effectiveness of adapting pre-trained language models to specific disease contexts, enabling the identification of novel disease-associated proteins and the exploration of their functional roles.

This promising methodology now needs to be translated into clinical practice, where its potential to enhance diagnostic and therapeutic strategies can be fully realized. Future work should focus on experimental validation of predicted proteins, integration with

other omics data, and development of targeted interventions based on these findings. As we continue to refine and expand this approach, we anticipate significant advancements in the diagnosis, prognosis, and treatment of POP, ultimately improving patient outcomes and quality of life.

5.7 Data and Material Availability

To promote transparency, reproducibility, and further research in this field, we have made the datasets, trained models, and code used in this study publicly available on GitHub. The repository can be accessed at: <https://github.com/Atobelin/EPOP>. The repository contains the following resources:

1. Datasets: The curated datasets of ECM and non-ECM proteins used for model training and validation.
2. Code: Python scripts for data preprocessing, model training, evaluation, and interpretation.
3. Model: The EPOP model, which can be directly loaded and applied to new datasets.

We encourage researchers to use, adapt, and build upon these resources for further investigations into POP and other ECM-related disorders. If you use the datasets, models, or code from this repository, please cite our paper. Please contact the corresponding author for any questions, suggestions, or collaborations.

Table 5.2: Cross-validation of Literature Can Support Proteins.

UniProt ID	Protein Name	Expression in POP
P12111	Collagen alpha-3(VI) chain	Different expression patterns (Yaqian Li et al. 2021)
P02458	Collagen alpha-1(II) chain	Different expression patterns (Yaqian Li et al. 2021)
Q99715	Collagen alpha-1(XII) chain	Extensive changes (Yaqian Li et al. 2021)
Q15582	Transforming growth factor-beta-induced protein ig-h3	Extensive changes (Yaqian Li et al. 2021)
P98160	Basement membrane-specific heparan sulfate proteoglycan core protein	Extensive changes (Yaqian Li et al. 2021)
Q05707	Collagen alpha-1(XIV) chain	Differentially expressed (Yaqian Li et al. 2021; X.-J. Li et al. 2019)
P22105	Tenascin-X	Differentially expressed (X.-J. Li et al. 2019)
P12109	Collagen alpha-1(VI) chain	Downregulated (Yaqian Li et al. 2021; Goepel et al. 2011)
P04004	Vitronectin	Downregulated (Goepel et al. 2011)
Q06828	Fibromodulin	Downregulated (X.-J. Li et al. 2019)
P02751	Fibronectin	Upregulated (Yaqian Li et al. 2021; X.-J. Li et al. 2019)
P24821	Tenascin	Upregulated (X.-J. Li et al. 2019; Goepel et al. 2011; Martina et al. 2010)
Q12805	EGF-containing fibulin-like	Significant enrichment (Emmerson and Gargett 2016; Ziwei Xie et al. 2025)
P08123	Collagen alpha-2(I) chain	No significant change (Yaqian Li et al. 2021; X.-J. Li et al. 2019; Goepel et al. 2011)
P08572	Collagen alpha-2(IV) chain	No significant change (Goepel et al. 2011)
P05997	Collagen alpha-2(V) chain	No significant change (Goepel et al. 2011)

6 Protein Design Enters the AI Era

This chapter is based on the manuscript: **Yanlin Mi***, Arpit Shukla, Mark Tangney, Sabin Tabirca and Venkata VB Yallapragada. *Protein Design Enters the AI Era: Foundations, Tools, and Emerging Paradigms*.

Manuscript currently under submission.

I made the primary contribution to this work through conceptualization, data curation, formal analysis, investigation, resource provision, and writing of both the original and revised drafts.

6.1 Abstract

AI has transformed protein engineering by leveraging DL, protein language models and knowledge graphs to decode relationships between sequence, structure, and function. Models like AlphaFold2 achieve near-experimental accuracy in structure prediction, while transformer-based language models facilitate de novo sequence design under functional constraints. AI enhances therapeutic protein engineering, enzyme catalysis and synthetic biology, accelerating the transition from in silico design to experimental validation. These advances have catalyzed breakthroughs across healthcare, industrial biotechnology and environmental applications, driving pharmaceutical partnerships exceeding USD 2.2 billion by late 2023 and propelling the protein engineering market toward a projected USD 20.86 billion valuation by 2034. Despite successes, challenges remain in model interpretability, training data biases and integration of AI with experimental pipelines. This review examines the AI methodologies shaping protein design, translational applications and emerging frontiers that will define the next phase of innovation in synthetic biology, personalized medicine and sustainable biotechnology.

6.2 Landscape of Proteins-AI

The global protein engineering market size accounted for USD 4.35 billion in 2024, grew to USD 5.09 billion in 2025, and is expected to be worth around USD 20.86 billion by 2034, poised to grow at a Compound Annual Growth Rate (CAGR) of 16.97% between 2024 and 2034 (Precedence Research 2024). This expansion is driven by increasing demands across pharmaceutical, biotechnology, and industrial sectors for precisely engineered proteins with enhanced or novel functions (Walker, V. V. Yallapragada, and Tangney 2021). Engineered proteins, which encompass both rationally modified natural proteins and de novo designed proteins, represent a cornerstone of modern biotechnology. While engineered proteins typically involve modifications to existing protein scaffolds to enhance specific properties, designer proteins represent a more ambitious frontier where entirely novel protein structures and functions are created from scratch. This distinction is particularly relevant as AI technologies increasingly enable both the optimization of natural proteins and the creation of unprecedented protein architectures.

Proteins are nature's molecular machines, orchestrating virtually every biological process through their remarkable diversity of functions. From catalyzing complex chemical reactions to transmitting cellular signals and providing structural support, proteins are fundamental to life itself (A. Johnson, Rohlfs, and Silverman 1999; Karplus and Kuriyan 2005; Branden and Tooze 2012). Their functionality derives from precisely folded three-dimensional structures, determined by specific sequences of amino acids. This intricate relationship between sequence, structure, and function has long fascinated scientists and engineers, driving efforts to design proteins with novel and enhanced capabilities (Q. Shi et al. 2023; Bepler and Berger 2021; Park, Metzger, and J. W. Thornton 2024).

The significance of protein design extends far beyond academic interest. In medicine, engineered proteins form the basis of numerous therapeutics, from insulin analogs to monoclonal antibodies. AI-driven companies like AbCellera successfully developed COVID-19 antibody treatment (bamlanivimab) and are accelerating this field (Tuccori et al. 2021). In industrial biotechnology, engineered enzymes designed by companies such as Zymergen and Ginkgo Bioworks catalyze processes from biofuel production to waste treatment (R. Saini et al. 2020). Recent achievements highlight the field's potential: researchers have created antibody nanocages that enhance therapeutic delivery (Divine et

6 Protein Design Enters the AI Era

al. 2021; S. A. Kim et al. 2023), developed novel protein assemblies with programmable geometries(Jie Zhu et al. 2021; Kortemme 2024), and engineered protein-based molecular machines with sophisticated functions(Y. A. Khan, White, and Brunger 2022).

The commercial landscape for AI in protein design is expanding at an unprecedented rate, underscored by a surge in high-value partnerships between AI-native biotechnology companies and established pharmaceutical giants. As shown in Table 6.1, these collaborations, often involving potential milestone payments exceeding a billion dollars, highlight the industry's conviction in AI's potential to break through longstanding bottlenecks in drug discovery (Dealmakers 2024; Nagra et al. 2023). The following table provides an overview of some of the most significant recent partnerships, moving beyond a simple market summary to analyze the specific AI applications being deployed and the core technical challenges these collaborations aim to solve. Analysis of these partnerships reveals several key trends. First, the financial commitments are substantial, indicating that AI-driven design is now a central pillar of pharmaceutical research and development strategy. Second, the scope of these deals is broad, targeting not only antibodies but also the historically challenging domain of small molecule design (as seen in the Isomorphic Labs collaborations). Third, the focus is increasingly on generative approaches; creating novel molecular structures rather than simply screening existing ones. However, the "Key Technical Challenge" column underscores a critical reality: while AI models can generate promising candidates, ensuring these digital designs translate into safe, effective, and manufacturable drugs remains the primary hurdle for the entire field.

Traditional approaches to protein design have relied heavily on rational design and directed evolution. Rational design attempts to predict beneficial mutations based on structural knowledge and biochemical principles, while directed evolution mimics natural selection in the laboratory to optimize protein properties. These methods, while powerful, face significant limitations. The complexity of protein folding, the vast sequence space to explore (20^{100} possible sequences for a modest 100-amino-acid protein), and the intricate relationships between sequence, structure, and function have historically made protein design a challenging endeavor(Korendovych and DeGrado 2020; Kuhlman and P. Bradley 2019a).

The emergence of AI has initiated a fundamental transformation in protein design methodology. This revolution began with improvements in protein structure prediction, exemplified by AlphaFold 2's breakthrough in achieving near-experimental accu-

Table 6.1: Overview of Recent Major AI-Driven Drug Discovery Partnerships

Date	AI Company	Pharma Partner	Therapeutic Focus	Value	AI Application	Key Challenge
Jan 2024	Isomorphic Labs	Eli Lilly	Multiple Small Molecules	\$1.7 B	Structure-based generative design of novel small molecule drugs.	Bridging the gap from <i>in silico</i> binding predictions to successful <i>in vivo</i> efficacy and safety.
Jan 2024	Isomorphic Labs	Novartis	Multiple Small-Molecules	\$1.2 B	Predicting protein structures and designing novel chemical entities.	Accurately predicting ADME/Tox properties computationally to reduce downstream failures.
Jan 2024	Generate Biomedicines	Amgen	Multiple Therapeutic Areas	\$1.9 B	Generative AI platform (Chroma) to create novel protein-based therapeutics.	Designing complex biologics with specific functions, stability, and low immunogenicity from scratch.
May 2024	Absci	Merck	Oncology	\$610 M	<i>De novo</i> design of therapeutic antibodies against specified targets.	Creating novel antibodies with optimal binding affinity and developability characteristics.
Nov 2023	BioMap	Sanofi	Multiple Therapeutic Areas	\$1 B	Modeling the immune system to discover and design novel biologics.	Mapping complex biological systems to identify novel targets and predict therapeutic effects.
Total Potential Milestone Payments					2,222 million dollars	

6 Protein Design Enters the AI Era

racy(Jumper, R. Evans, Pritzel, T. Green, Figurnov, Ronneberger, et al. 2021; Abramson et al. 2024). This advance has been complemented by developments in protein language models(Nijkamp et al. 2023; Ferruz, Schmidt, and Höcker 2022; Z. Lin et al. 2023) and DL approaches that can generate novel protein sequences with desired properties(Notin et al. 2024; Anishchenko et al. 2021). The impact of these technologies extends beyond mere computational improvements; they have fundamentally changed how we approach the protein design problem.

AI methods are starting to advance protein design through several transformative capabilities. Deep learning approaches enable rapid exploration of vast sequence spaces through efficient sampling algorithms that learn from existing protein structures and sequences. These methods allow direct optimization of protein properties through differentiable models that can be trained end-to-end. Additionally, AI facilitates the integration of multiple data modalities, including sequence, structure, and functional information, providing a more comprehensive approach to protein design. The technology also enables the automated design of complex protein architectures previously considered impossible to engineer(Notin et al. 2024; Jumper, R. Evans, Pritzel, T. Green, Figurnov, Ronneberger, et al. 2021; Abramson et al. 2024).

Consider the example of de novo protein design, where AI has enabled the creation of proteins with entirely novel folds. Traditional methods typically began with known protein structures and made incremental modifications. In contrast, AI approaches can generate completely new protein architectures by learning the fundamental principles of protein folding from existing structures. This capability was demonstrated in recent work where researchers used DL to create protein structures with unprecedented geometric complexity, including protein rings, cages, and lattices with precise specifications(Shunzhi Wang et al. 2024; Schmid et al. 2023; Jie Zhu et al. 2021; Suwardi et al. 2022).

In summary, the confluence of AI technologies with protein design has ushered in a new era of possibilities in protein engineering. From enhanced structure prediction capabilities to sophisticated sequence generation models, AI methods have started to fundamentally alter both the scope and efficiency of protein design. These advances have enabled unprecedented achievements in therapeutic development, industrial applications, and fundamental protein science. As shown in Fig 6.1, the following sections delve into the key AI technologies driving these advances, examining their technical implementations,

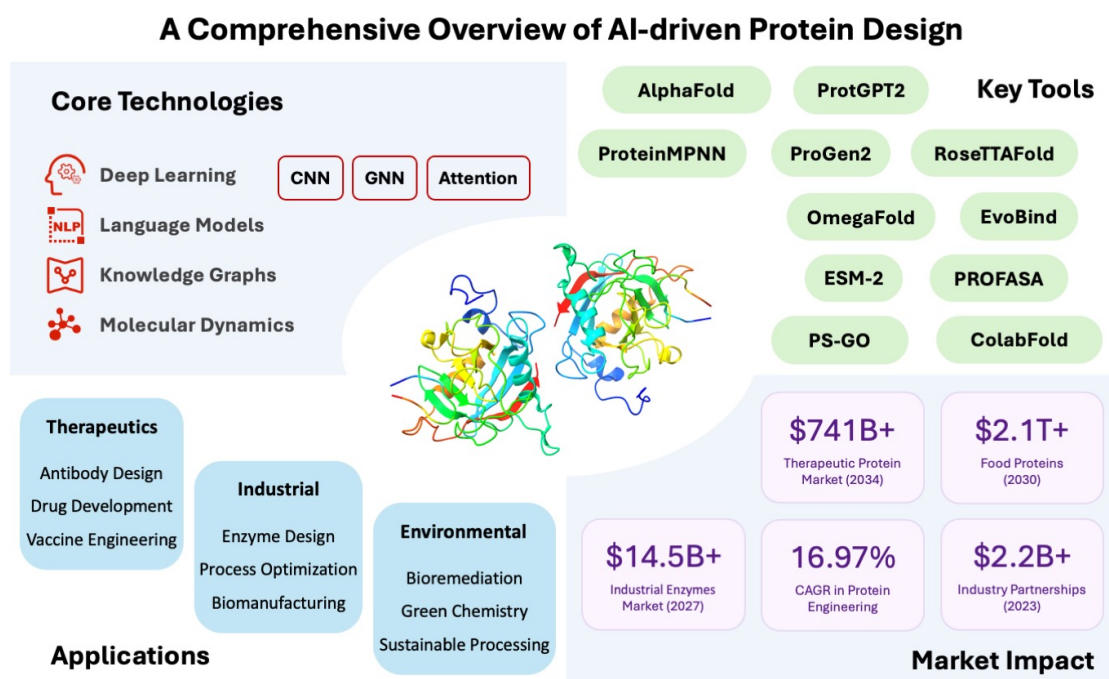


Figure 6.1: A comprehensive overview of AI-driven protein design. The diagram is organized into several key sections: on the top left, it lists core technologies including DL (with CNN, GNN, and Attention mechanisms), language models, knowledge graphs, and molecular dynamics; at the center, it features a colorful 3D protein structure illustration; the top right shows key tools including AlphaFold, RoseTTAFold, and ESM-2. The lower left section outlines three major application domains (therapeutics, industrial, and environmental) with their specific use cases. Additionally, the market impact is highlighted in the lower right.

practical applications, market impact and the challenges that lie ahead. We begin by exploring the core AI methodologies that form the foundation of modern protein design, including DL architectures, protein language models, and knowledge graph frameworks, before proceeding to discuss their practical implementations and impact across various domains.

6.3 Key AI Technologies in Protein Design

The transition from traditional computational methods to AI-driven approaches has been enabled by advances in four key areas: DL architectures, protein language models, knowledge graph technologies, and the integration of molecular dynamics simulations. Understanding these foundations is crucial for appreciating the current capabilities and future potential of AI in protein design. Each of these technological pillars contributes unique capabilities to the protein design process, from structure prediction and sequence optimization to dynamic behavior analysis and knowledge integration.

6.3.1 Deep Learning Architectures

The evolution of DL architectures in protein design has been marked by increasingly complex approaches to handling the complex hierarchical nature of protein structures (Pearce and Yang Zhang 2021; Jingxue Wang et al. 2018; W. Gao et al. 2020; Dauparas et al. 2022). The implementation of these architectures has revealed several key innovations in network design and training methodologies, particularly in the realm of CNN and their adaptations for protein structure analysis (Yuan Zhang et al. 2020).

Modern CNN architectures have evolved to incorporate dilated convolutions specifically designed for protein structure analysis (Yuan Zhang et al. 2020; Ferruz, Heinzinger, et al. 2023). As shown in Fig 6.2 (A), these networks employ a complex multi-scale approach, utilizing various dilation rates simultaneously to capture structural patterns at different spatial scales. A typical implementation employs dilation rates of 1, 2, 4, and 8, enabling the network to process both local and long-range structural patterns effectively. The architecture is further enhanced by residual connections between dilated convolutional layers, which maintain robust gradient flow during training while processing extended structural patterns (Torrìsi, Pollastri, and Le 2020; W. Gao et al. 2020; Jisna and Jayaraj 2021).

Advanced CNN architectures designed specifically for protein analysis incorporate multiple parallel convolutional streams, each dedicated to processing distinct structural features. One stream focuses on backbone angle predictions, while others simultaneously process side chain orientations and contact maps. These streams are complemented by specialized pooling operations that successfully reduce computational complexity while preserving critical structural information. The architecture's effectiveness is further

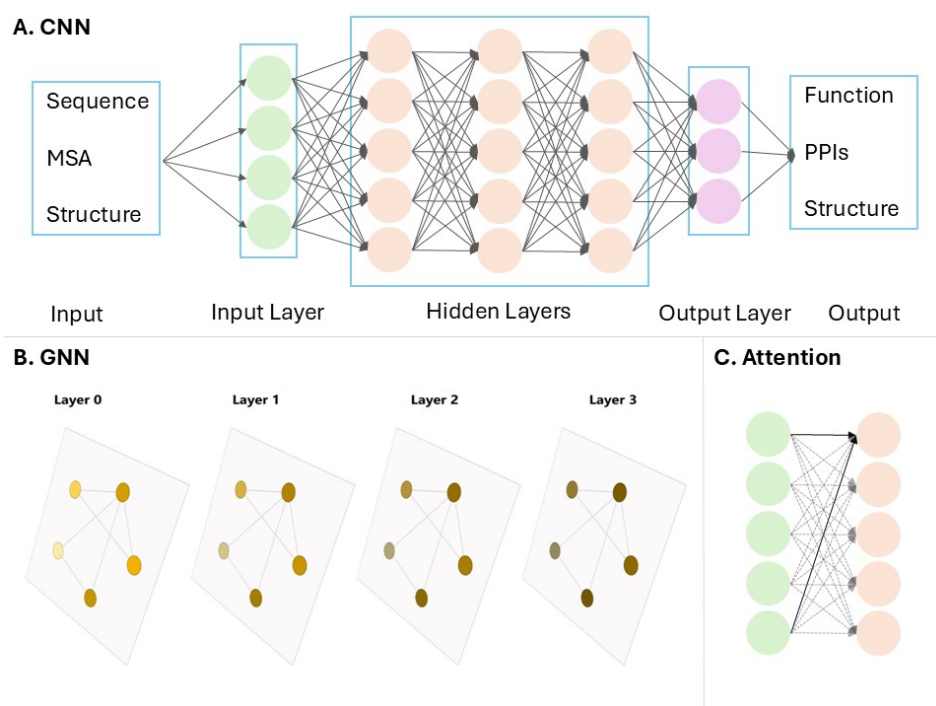


Figure 6.2: Three DL architectures are used in protein design. (A) The CNN section shows a typical feed-forward network architecture. The input layer receives sequence, multiple sequence alignment, and structure information, processes it through multiple hidden layers, and ultimately predicts protein function, protein-protein interactions, and structural information in the output layer. The network features dense connections for forward information propagation. (B) The GNN section demonstrates how protein graph representations are processed across multiple layers. From Layer 0 to Layer 3, each layer contains multiple nodes connected by lines to form a graph structure. The nodes' color gradually deepens from Layer 0 to Layer 3, representing the progression and update of information through the graph network. (C) The Attention section displays a simplified attention layer architecture. It shows five input nodes on the left and five output nodes on the right, connected by sparse lines representing attention weights, demonstrating how the model selectively attends to different input information.

enhanced by attention-gated convolutions, which dynamically adjust filter weights based on the local structural context of each protein region(Nazem et al. 2024; Arya and Sriparna Saha 2021; Zohora et al. 2019).

GNN in protein design have undergone significant evolution, now incorporating complex edge features and message-passing schemes (Réau et al. 2023; Z. Gao, Tan, and S. Z. Li 2022). Contemporary implementations utilize comprehensive edge features that encode both covalent and non-covalent interactions. These features capture the full spectrum of molecular interactions, including hydrogen bonds, hydrophobic interactions, and electrostatic forces. As shown in Fig 6.2 (B), the networks implement multiple rounds of message passing with learned update functions, allowing for sophisticated information propagation across the protein structure. The architecture is completed by hierarchical graph pooling operations that maintain the essential aspects of protein topology while reducing computational complexity (S. Zhang et al. 2023; Shuyu Wang et al. 2023).

A significant breakthrough in GNN architecture has been the development of equivariant GNN. These networks maintain geometric invariances crucial for protein structure prediction through the implementation of SO(3)-equivariant layers. These specialized layers process geometric features while preserving rotational invariance, a property that has proven essential for accurate structure prediction and validation (Dhakal, Gyawali, and J. Cheng 2023; Jing et al. 2021; F. Wu et al. 2023).

6.3.2 Diffusion-Based Generative Models

Recently, diffusion-based generative models have emerged as a powerful new paradigm for protein design, representing a significant leap in generating novel protein structures. These models operate by learning to reverse a noise-addition process. During training, they learn to denoise a structure, starting from a random cloud of atoms and gradually refining it into a physically coherent protein backbone. This approach is revolutionary because it is not constrained by existing protein templates or folds, enabling true de novo design of entirely novel architectures.

A prominent example is RFdiffusion, which has demonstrated the ability to generate a wide range of complex and functional protein topologies with high atomic accuracy. A key advantage of this framework is its ability to be conditioned. Designers can specify a functional motif, such as an enzyme active site or a binder interface, and RFdiffusion can generate a novel protein scaffold built precisely around it—a task known as motif-scaffolding. This capability has profound implications for designing bespoke enzymes,

6.3 Key AI Technologies in Protein Design

therapeutics, and protein-based materials (Watson et al. 2023). Complementary tools like EvoDiff apply the diffusion concept to generate new protein sequences, further expanding the toolkit for generative protein engineering (Alamdari et al. 2023). To provide a clearer understanding of how these different architectures contribute to specific protein design objectives, Table 6.2 summarizes their primary applications, representative tools, and inherent strengths and limitations. This comparative overview highlights the shift from discriminative models used for analysis to generative models that are now capable of creating novel proteins from scratch.

Table 6.2: Comparative Overview of Key Deep Learning Architectures in Protein Design

Model Architecture	Representative Tool(s)	Primary Protein Design Application(s)	Strengths	Limitations
GNN	ProteinMPNN	Designing sequences that fold into a predefined backbone structure	Excels at capturing the 3D local environment of residues for high-fidelity sequence recovery	Requires a given backbone; cannot generate novel folds
Language Models	ESM-2, Pro-Gen2	Generating novel sequences based on learned evolutionary patterns	Learns functional patterns from vast sequence data without needing structures	Indirect control over 3D structure; generated sequences may not fold correctly
Diffusion Models	RFdiffusion	Generating novel protein backbones from scratch, with or without functional constraints (motif scaffolding)	Generates entirely new folds not seen in nature; can be precisely conditioned	Computationally intensive; ensuring biological function remains a challenge

Recent innovations in attention mechanisms have substantially improved protein design capabilities (Z. Niu, Zhong, and H. Yu 2021). Advanced implementations now utilize axial attention mechanisms that separately process spatial and channel dimensions, allowing for more efficient and effective feature extraction. These systems are complemented by multi-head attention mechanisms, where different heads specialize in detecting various types of structural patterns. The effectiveness of these attention mechanisms is further enhanced by gradient-guided attention systems, which utilize backpropagated gradients to focus computational resources on functionally important regions of the protein structure (Lifan Chen et al. 2020; L. Wei et al. 2021).

These sophisticated attention mechanisms have been particularly revolutionary in capturing the complex dependencies in protein structures. As shown in Fig 6.2 (C), position-specific attention weights now incorporate both sequence and structural context,

6 Protein Design Enters the AI Era

enabling the model to learn subtle relationships between distant residues that may be spatially proximate in the folded structure. This has proven especially valuable for understanding allosteric effects and long-range interactions that influence protein function(T. Gong et al. 2023; Pang and B. Liu 2020; H. Wu et al. 2024).

The integration of structure-aware attention has led to significant improvements in modeling protein conformational changes. By incorporating distance matrices and contact maps into the attention computation, these mechanisms can better capture the physical constraints and dynamic nature of protein structures. This has enabled more accurate prediction of protein flexibility, conformational switches, and interaction interfaces(Can Chen et al. 2023; Shuangli Li et al. 2021).

Furthermore, recent implementations have introduced hierarchical attention schemes that operate at multiple structural levels - from local backbone geometry to domain-level organization. This multi-scale approach allows the model to simultaneously consider both fine-grained atomic interactions and larger-scale architectural features that determine protein stability and function. The attention weights at each level are dynamically adjusted based on the specific design task, whether it's optimizing binding interfaces, engineering allosteric regulation, or designing novel folds(Gu et al. 2023; Vig et al. 2020; Rives et al. 2021).

6.3.3 Protein Language Models

As shown in Fig 6.3, by analogy with the English language model, we can easily understand the Protein language model. Modern protein language models have achieved remarkable advances through sophisticated architectural innovations and training strategies(Chowdhury et al. 2022). The training process has evolved to incorporate multiple complementary objectives that enhance the models' understanding of protein structure and function. These objectives include structure-aware masked token prediction, which strategically masks portions of protein sequences to force the model to learn structural patterns. This is complemented by domain-adapted next sentence prediction tasks that help the model understand the relationships between different protein domains. Additionally, contrastive learning between sequence and structure representations has proven particularly effective in developing rich, informative protein embeddings (Ferruz and Höcker 2022; Bepler and Berger 2021).

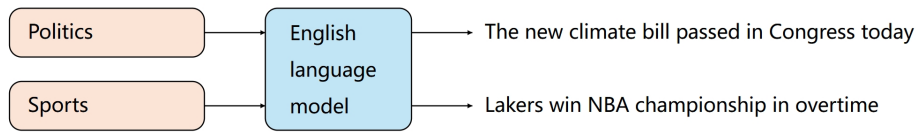
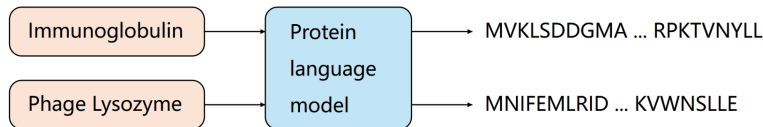
A. Control tag AI Model Generated English sentence**B. Control tag AI Model Generated protein sequence**

Figure 6.3: The conceptual parallel between natural language models and protein language models. (A) It demonstrates a conventional English language model that generates domain-specific sentences based on control tags (e.g., "Politics" or "Sports"). (B) It shows an analogous protein language model that generates amino acid sequences conditioned on protein family tags (e.g., "Immunoglobulin" or "Phage Lysozyme"). This comparison highlights how the fundamental principles of large language models can be adapted to biological sequence generation: just as language models learn to produce contextually appropriate text based on input prompts, protein language models learn to generate biochemically viable amino acid sequences based on protein family specifications. The architectural similarity between these models suggests that the underlying patterns in protein sequences can be captured and generated using similar computational frameworks as those used for natural language processing.

The training methodology has been refined to overcome computational limitations while maximizing model performance. Modern approaches implement gradient accumulation across multiple GPUs, enabling the use of larger effective batch sizes that improve training stability and model convergence. This is combined with mixed-precision training techniques that reduce memory requirements while maintaining numerical stability. The training process typically follows a curriculum learning approach, beginning with shorter sequences and progressively increasing sequence length as training proceeds, allowing the model to build up complexity gradually (Lv et al. 2024; Huo et al. 2024; Zeinalipour et al. 2024).

Architectural innovations have significantly enhanced model performance. Recent models incorporate relative positional encodings that more accurately capture local structural relationships within proteins. These encodings are complemented by sparse attention patterns that focus computational resources on evolutionarily related positions, improving the model's ability to capture meaningful sequence relationships. The architecture is further enhanced by specialized embedding layers that encode detailed amino acid chemical properties, providing the model with rich input representations that facilitate the learning of structure-function relationships (Chowdhury et al. 2022; Ferruz and Höcker 2022; Filipavicius et al. 2020; Xiao et al. 2021).

6.3.4 Knowledge Graph Technologies

The implementation of knowledge graph technologies in protein design represents a sophisticated approach to integrating and reasoning about complex biological information (A. Santos et al. 2022). Modern protein knowledge graphs have evolved to incorporate multiple layers of biological information through advanced construction techniques and reasoning mechanisms. The graph construction process has been revolutionized by automated extraction systems that utilize specialized biomedical natural language processing algorithms to mine protein-protein interactions from scientific literature. These systems achieve high precision through context-aware entity recognition and relationship extraction, significantly reducing the manual curation burden while maintaining data quality (Nicholson and Greene 2020; Chandak, K. Huang, and Zitnik 2023).

The integration of experimental data from multiple sources has become increasingly sophisticated, with modern knowledge graphs incorporating structural information from X-ray crystallography, nuclear magnetic resonance spectroscopy, and cryo-electron microscopy. This experimental data is further enriched by predicted structural features from AI models, creating a comprehensive representation of protein properties. The resulting graphs maintain provenance information for each data point, enabling researchers to assess the reliability of different information sources and make informed decisions during the design process (Mohamed, Nounu, and Nováček 2021; Gaudet et al. 2021).

The graph schema itself has evolved to capture the hierarchical nature of protein organization. Protein nodes are annotated with comprehensive attribute sets describing

sequence characteristics, structural features, and functional properties. These attributes are dynamically updated as new information becomes available, ensuring the knowledge graph remains current. Interaction edges carry detailed metadata including confidence scores, experimental validation methods, and relevant environmental conditions under which the interactions were observed. The schema also captures hierarchical relationships between proteins, domains, and motifs, enabling multi-scale analysis of protein properties(Lam et al. 2023; Bonner et al. 2022; Waagmeester et al. 2020).

Advanced reasoning mechanisms have been developed to exploit this rich knowledge representation. Path-based inference algorithms traverse the graph to predict novel protein-protein interactions, utilizing both direct and indirect relationships to generate predictions. These algorithms employ sophisticated path scoring mechanisms that account for the reliability and relevance of different relationship types. Subgraph matching algorithms have been optimized to identify functional modules within the larger protein interaction network, enabling the discovery of conserved structural and functional motifs that can be applied in protein design(Nováček et al. 2020; Mohamed, Nounu, and Nováček 2021; Unni et al. 2022).

Embedding-based reasoning has emerged as a particularly powerful approach, combining the advantages of knowledge graphs with modern machine learning techniques. These methods project proteins and their relationships into high-dimensional vector spaces while preserving their biological properties. The resulting embeddings enable rapid similarity searches and property predictions, significantly accelerating the protein design process. The embedding spaces have been shown to capture meaningful biological relationships, with proteins of similar function clustering together and geometric relationships between embeddings reflecting biological relationships(Sousa, S. Silva, and Pesquita 2024; Jiaoyan Chen et al. 2023).

6.3.5 Integration of Molecular Dynamics and AI

The convergence of molecular dynamics simulations and AI has led to transformative advances in protein design methodology(C.-H. Yu and Buehler 2020). Enhanced sampling techniques have emerged as a particular focus of innovation, with modern approaches achieving unprecedented efficiency in exploring protein conformational spaces. Neural network-based collective variables have revolutionized enhanced sampling by automati-

6 Protein Design Enters the AI Era

cally identifying relevant degrees of freedom for protein motion. These learned variables adapt dynamically during simulations, enabling efficient exploration of complex conformational landscapes that would be difficult to sample using traditional approaches(Y. I. Yang et al. 2019; Anishchenko et al. 2021).

Online learning of bias potentials demonstrates another significant advancement in the field. Modern systems continuously update their understanding of the protein's energy landscape during simulation, allowing for adaptive modification of sampling strategies. This approach has proven particularly effective for studying protein folding pathways and conformational transitions, providing insights that directly inform the design process. The integration of AI-guided adaptive sampling strategies has further enhanced simulation efficiency by focusing computational resources on promising regions of conformational space(K. K. Yang, Zachary Wu, and Arnold 2019; Friederich et al. 2021).

Force field development has been transformed by the integration of AI techniques. Machine learning potentials trained on high-level quantum mechanical calculations now provide unprecedented accuracy in modeling protein energetics. These potentials capture subtle electronic effects that are difficult to illustrate in traditional force fields, enabling more accurate prediction of protein properties. Neural network corrections to traditional force fields have emerged as a practical approach to improving accuracy while maintaining computational efficiency. These corrections learn to predict and compensate for systematic errors in classical force fields, resulting in more reliable simulations(X. He et al. 2020; Jiang et al. 2024).

The development of end-to-end differentiable force fields represents a particularly exciting advance. These force fields enable the gradient-based optimization of protein properties, allowing direct optimization of sequence and structure to achieve desired characteristics. The differentiable nature of these force fields facilitates integration with modern DL frameworks, enabling seamless combination of physical modeling with machine learning approaches. This integration has proven particularly valuable for designing proteins with specific dynamic properties, as it allows explicit optimization of conformational behavior(AlQuraishi 2019; Navarro, Majewski, and De Fabritiis 2023).

Table 6.3: Commercial AI Platforms for Protein Design

Platform	Core Technology	Commercial Applications	Key Limitations
AlphaFold2 (Zhenyu Yang et al. 2023; Agu and Obulose 2024; Ozaybi et al. 2024; V. Agarwal and McShan 2024)	Evolutionary attention	Insilico Medicine: Drug discovery for fibrosis targets; AbCellera: Antibody engineering for therapeutic development	Static structure prediction; No dynamics
RoseTTAFold (Dolgin 2024; P.-S. Huang et al. 2011; Baek et al. 2021)	Three-track architecture	Outpace Bio: Protein therapeutics design platform; Arzeda: Computational enzyme design for sustainable materials	High computational cost; Limited to single-chain proteins
ESM-2 (Z. Lin et al. 2023; Lingtao Chen et al. 2024; Chungyoun and Gray 2023)	Protein language model	Absci: AI-driven antibody design and optimization; Generate Biomedicines: De novo protein design via Chroma platform	Limited physical constraints; High experimental validation failure rate

6.4 AI-Driven Protein Design Tools

The evolution of AI tools for protein engineering is marked by a crucial transition from structure prediction to protein design. Foundational tools like AlphaFold2 solved the long-standing challenge of predicting a protein’s 3D structure from its amino acid sequence. While revolutionary, these models are fundamentally discriminative. The cutting edge of the field has since moved toward generative models, which address the inverse problem: given a desired structure, function, or property, how can we generate a novel protein sequence (and backbone) to match it? This shift from analysis to creation, exemplified by tools like ProteinMPNN and RFdiffusion, underpins the current era of true AI-driven protein design (Dauparas et al. 2022; Watson et al. 2023).

The landscape of AI-driven protein design tools has been transformed by several breakthrough platforms, each contributing unique capabilities and methodological innovations. As shown in Table 6.3, commercial AI platforms such as AlphaFold2, RoseTTAFold, and ESM-2 represent the cornerstone of current protein design technology, with widespread applications across pharmaceutical, biotechnology, and academic sectors. These tools symbolize the practical implementation of theoretical advances in AI technology, enabling unprecedented accuracy and efficiency in protein design and analysis.

6.4.1 AlphaFold 2 Architecture and Implementation

AlphaFold 2 indicates a revolutionary advancement in protein structure prediction, implemented through a sophisticated multi-component architecture(Jumper, R. Evans, Pritzel,

T. Green, Figurnov, Ronneberger, et al. 2021). As highlighted in Table 6.3, this platform has been successfully deployed in commercial applications at companies like Insilico Medicine and AbCellera for drug discovery and therapeutic development. The system's core consists of specialized Evoformer blocks that process multiple sequence alignments with remarkable efficiency. These blocks implement novel attention mechanisms that capture both sequence and structural relationships simultaneously, enabling the model to leverage evolutionary information effectively. The structure module incorporates equivariant attention mechanisms that maintain geometric invariances crucial for accurate structure prediction, while the template processing network integrates structural information from known protein structures to guide predictions (Skolnick et al. 2021).

Performance metrics for AlphaFold 2 demonstrate its exceptional capabilities. The system consistently achieves GDT-TS scores exceeding 90 for many protein targets, particularly impressive for single-domain proteins where TM-scores frequently exceed 0.9. The model's accuracy extends to challenging cases, including the prediction of disordered regions, which had traditionally been difficult to characterize computationally. This performance level has transformed structural biology, enabling reliable structure prediction for previously intractable protein targets (Stevens and Y. He 2022).

The implementation architecture of AlphaFold 2 incorporates several technical innovations that contribute to its exceptional performance. The attention mechanisms are carefully designed to process both local and global structural features simultaneously, while the template processing system can effectively utilize partial structural information when available. The model's training process incorporates sophisticated loss functions that balance different aspects of structural accuracy, ensuring comprehensive optimization of prediction quality (Jumper, R. Evans, Pritzel, T. Green, Figurnov, Tunyasuvunakool, et al. 2020).

6.4.2 RoseTTAFold Technical Implementation

RoseTTAFold introduces a novel three-track architecture that processes protein information at multiple levels of abstraction simultaneously. The system implements parallel processing streams for one-dimensional sequence data, two-dimensional distance maps, and three-dimensional coordinate information. This multi-track approach enables the model to capture different aspects of protein structure at various scales, facilitating more

accurate predictions and design capabilities (Mansoor et al. 2023). As shown in Table 6.3, RoseTTAFold has been successfully commercialized through platforms like Outpace Bio for protein therapeutics design and Arzeda for computational enzyme design for sustainable materials, demonstrating its versatility across different industrial applications.

The architecture's technical implementation includes specialized attention mechanisms for each track, enabling efficient information exchange between different representations. The gradient-based optimization of backbone coordinates allows for precise structural refinement, while the integration of evolutionary information helps guide the prediction process. This sophisticated implementation has enabled RoseTTAFold to achieve remarkable success in various applications, from complex structure prediction to protein-protein interface design and de novo protein creation (Rohith Krishna et al. 2024).

6.4.3 ESM-2 System Architecture

ESM-2 represents the state-of-the-art in protein language models, implementing a sophisticated 33-layer transformer architecture with approximately 650 million parameters. The system incorporates specialized positional encodings designed specifically for protein sequences, enabling better capture of local structural patterns and long-range interactions. The implementation includes multiple attention heads that process different aspects of sequence patterns, allowing the model to simultaneously capture various types of biological relationships (H. Zhang et al. 2020). As detailed in Table 6.3, ESM-2 technology has been leveraged commercially by companies such as Absci for AI-driven antibody design and Generate Biomedicines for de novo protein design through their Chroma platform, illustrating its powerful capabilities in therapeutic protein engineering.

The model demonstrates exceptional performance across multiple prediction tasks. Secondary structure prediction accuracy consistently exceeds 85%, while contact prediction precision surpasses 70% for many protein targets. Perhaps most impressively, the system achieves correlation coefficients above 0.7 in mutation effect prediction tasks without specific training on mutation data, demonstrating strong transfer learning capabilities (Zhidian Zhang et al. 2024).

6.4.4 Comprehensive Ecosystem of AI Tools

The field of protein design is driven by a combination of core platforms and specialized tools, each addressing different challenges in the design process. Table 6.3 shows the commercial AI platforms that currently dominate the market, which have demonstrated practical application in the pharmaceutical, biotechnology, and industrial sectors. From AlphaFold2 to ESM-2, these technologies have revolutionized structural prediction and sequence-function modeling possibilities through innovative architectural approaches.

Complementing these major business platforms is a range of tools that specialize in specific problems. These new tools complement the limitations of the base platform and focus on interface design, sequence optimization, and segmentation in industrial applications. This diverse technology ecology reflects the complexity of protein engineering and the need for a diverse approach to various design requirements. Table 6.4 provides a comprehensive overview of these professional tools, detailing their technical underpinnings, major areas of application, and practical examples that demonstrate their scientific and commercial value.

6.5 Applications and Impact

The integration of AI in protein design has facilitated transformative advances in fields ranging from therapeutic development and food engineering to industrial biotechnology (Holzinger, Keiblinger, et al. 2023). These advances, coupled with substantial market growth and commercial success, represent fundamental shifts in our ability to engineer proteins with precise specifications and novel functions. The global impact of these technologies is reflected in both their scientific achievements and their rapidly expanding market presence.

6.5.1 Therapeutic Protein Development

Table 6.4: Overview of Additional AI-Driven Protein Design Tools

Tool	Developer	Core Technology	Key Features	Primary Applications
ProteinMPNN (Dau- paras et al. 2022)	Baker Lab	Message passing neural networks	Multiple sequence design; Backbone conditioning; Fast sampling	Interface design; Sequence optimization

6.5 Applications and Impact

Tool	Developer	Core Technology	Key Features	Primary Applications
OmegaFold (Ruidong Wu et al. 2022)	HeliXon	Single-sequence structure prediction	No multiple sequence alignment requirement; Fast runtime; Attention mechanism	Rapid structure prediction; Drug discovery
ProGen2 (Nijkamp et al. 2023)	Salesforce	Autoregressive generation	Controlled synthesis; Family-specific design; Transfer learning	De novo protein design; Function engineering
EvoBind (Bryant and Elofsson 2022)	Baker Lab	Physics-guided DL	Monte Carlo sampling; Physics-based scoring; Fast prediction	Drug design; Protein-ligand interactions
ColabFold (Mirdita et al. 2022)	Steinegger Lab	AlphaFold integration	Cloud computing; MM-seqs2 search; User-friendly interface	Academic research; Educational use
ESM-IF1 (Bryant and Elofsson 2023)	Meta AI	Inverse folding	Structure conditioning; Fast sampling; Diverse outputs	Structure-based design; Sequence generation
trRosetta (Z. Du et al. 2021)	Baker Lab	Deep residual networks	Distance prediction; Orientation modeling; Multiple sequence alignment processing	Structure prediction; Contact prediction
ProtGPT2 (Ferruz, Schmidt, and Höcker 2022)	OpenBioML	GPT architecture	Zero-shot prediction; Sequence generation; Domain awareness	Novel protein generation; Function prediction
PROFASA (Mi et al. 2023b)	CRR Lab	AI Tool Integration	Alphafold 2; ESMfold; Visualization	Structure prediction; Structural analysis; Sequence Search
PS-GO (Mi et al. 2024)	CRR Lab	Natural language processing	Parameter Prediction; RL Sequence generation	Parameter Search

The impact of AI on therapeutic protein development has been particularly profound, revolutionizing both the design process and the range of achievable outcomes. The global protein therapeutics market size is calculated at USD 401.97 billion in 2025 and is forecasted to reach around USD 741.16 billion by 2034, accelerating at a CAGR of 7.04% from 2025 to 2034 (Precedence Research 2025). This growth is driven by the increasing adoption of AI technologies in drug discovery and development processes, with particular success in antibody engineering and enzyme replacement therapies.

A landmark achievement in this field was the development of antibody nanocages through AI-driven design (S. A. Kim et al. 2023). These structures demonstrate a new paradigm in therapeutic delivery, where antibodies are precisely arranged in geometric configurations to enhance their therapeutic efficacy. The design process integrated DL models for structure prediction with optimization algorithms that considered both geometric constraints and biological functionality. The resulting nanocages demonstrated

6 Protein Design Enters the AI Era

superior drug delivery properties compared to traditional antibody formulations, with improved tissue penetration and reduced off-target effects. Commercial implementation of this technology has attracted significant investment, with major pharmaceutical companies committing substantial resources to AI-driven antibody design platforms. As shown in Table 6.1, in the last quarter of 2023 alone, four major partnerships were formed with potential milestone payments totaling USD 2.2 billion (Nagra et al. 2023; Dealmakers 2024).

The development of protein-based vaccines expresses another area of significant impact, particularly highlighted during the COVID-19 pandemic. AI methods have accelerated vaccine development through multiple approaches, including the identification of viral antigenic epitopes, exploration of neutralizing antibodies, and optimization of vaccine design using reverse vaccinology techniques. For example, the Harvard T.H. Chan School of Public Health and the Human Vaccines Project launched the Human Immunomics Initiative, which utilizes state-of-the-art AI models to accelerate vaccine development by analyzing epidemiology, immune monitoring, and network biology data (Kaushik and U. Raj 2020). During the COVID-19 pandemic, AI platforms demonstrated their value in vaccine development through initiatives such as Magar et al.'s machine learning model for antibody exploration, which successfully identified eight potential antibodies against COVID-19 by analyzing over 1,900 previously reported antigen-antibody sequences. Companies across the pharmaceutical industry have recognized this potential, leading to increased adoption of AI technologies in vaccine development workflows (Savage 2021).

6.5.2 Food Protein Engineering

The food protein field has made significant progress through AI-driven protein design, particularly in enhancing protein functionality and developing alternative protein sources (L. Day, Cakebread, and Loveday 2022), (Grossmann and Weiss 2021). Driven by the growing demand for sustainable and nutritious protein solutions, the global food protein market is expected to reach over USD 2.1 trillion by 2030 (Karabulut, Goksen, and Khaneghah 2024). Facing the increasing protein demand and higher consumer expectations for food quality, traditional protein improvement methods can no longer meet the requirements.

AI technology has demonstrated significant advantages in food protein development in recent years. Amai Proteins from Israel developed a computational platform named Agile Integrated Computational Protein Design (AI-CPD). This method analyzes structural properties of extremophilic proteins to optimize monellin, a natural sweet protein originally identified in tropical berries. By integrating AI-driven amino acid sequence modification with structural prediction algorithms, the team engineered the thermostable sweet protein sweelin. This protein can reduce 40%-70% of sugar content without affecting taste. Compared to the original protein that denatures at 45 degrees, sweelin remains stable even at high temperatures. The product has been tested in various foods such as tomato sauce and chocolate, with professional tasters unable to detect differences. It has received GRAS certification and is now collaborating with multiple food manufacturers(Melton 2023).

In addition, Impossible Foods developed leghemoglobin by deeply analyzing the hemoglobin protein family in soybean roots and using machine learning to assist in screening the most suitable molecular structures. They successfully achieved large-scale production of this protein through microbial fermentation. When the protein denatures during heating, the heme iron group becomes more exposed, promoting a series of oxidation reactions that generate flavor compounds remarkably similar to real meat. This molecular mechanism-based innovation has been successfully applied in products like Beyond Burger, with professional food critics stating that its flavor performance is almost indistinguishable from real meat(McClements and Grossmann 2021).

6.5.3 Industrial Applications

The industrial sector has witnessed significant advances in biocatalysis through AI and machine learning-driven protein engineering. Over the last two decades, biocatalysis has secured its position as a standard approach in the chemist's toolbox, particularly in manufacturing active pharmaceutical ingredients (APIs), as well as fine and bulk chemicals. The integration of computational tools, especially machine learning algorithms, has revolutionized protein engineering for biocatalytic applications and accelerated the development timelines previously needed to optimize enzymes to more efficient variants(Markus et al. 2023). The demand and trade for industrial enzymes has been growing

6 Protein Design Enters the AI Era

steadily. The global enzyme market was valued at USD 8.63 billion in 2019 and is projected to reach USD 14.5 billion by 2027(Espina, Atalah, and Blamey 2021).

A key development has been the application of machine learning methods to enzyme engineering for industrial applications. While traditional enzyme engineering faced limitations in simultaneously improving multiple properties like activity, selectivity, stability, and solvent tolerance, machine learning approaches have enabled more efficient optimization of these multiple parameters. These advances have particularly benefited pharmaceutical manufacturing, where the combination of bioinformatics and machine learning algorithms offers new tools to potentially locate advanced enzyme engineering starting points and more direct routes to de novo and highly engineered high-performance biocatalysts(Markus et al. 2023).

The field of biocatalysis has seen remarkable advances in the design of enzymes for industrial synthesis. Recent collaborations between major companies demonstrate the practical impact of AI-driven protein engineering. For example, Unilever, working with Arzeda, developed new stain-fighting enzymes for cleaning and laundry products with 50% fewer ingredients while increasing stability, performance, and sustainability. What's particularly notable is the dramatically reduced development time - this enzyme development was achieved in just 18 months, which is five times faster than previously possible (Cooper 2023).

In another case, BASF created an AI system called Emollient Maestro that uses machine learning to identify optimal emollient mixes for cosmetic products based on various parameters like sensory properties, performance, and sustainability. This system leverages a huge database of previous experiments together with prediction models to accelerate the prototyping and design of new emollient products (Lahoti 2023).

6.5.4 Future Market Trajectory and Implementation Challenges

The protein engineering market shows a strong growth trajectory, with projections indicating a market value of USD 20.86 billion by 2034. This expansion reflects the field's transition from traditional methods to computational approaches, though currently over 95% of protein engineering still relies on screening-based methods (Walker, V. V. Yallapragada, and Tangney 2021).

Implementation challenges remain significant, primarily in production costs and computing resource requirements. While theoretical protein design space is vast - a single 100-amino-acid protein can have 20^{100} sequence variations compared to nature's 10^{12} explored proteins - practical limitations in production and testing continue to constrain commercial applications. Recent advances in cloud computing infrastructure and improved design tools suggest these barriers are beginning to lower. The convergence of more accessible computational resources with refined production processes indicates the field is positioned for accelerated growth and broader commercial adoption in the coming years (Walker, V. V. Yallapragada, and Tangney 2021).

As a leading user of AI protein design methods, the pharmaceutical industry provides an insightful case study of the potential and present drawbacks of these methods. Although agreements with potential milestone payments surpassing USD 2.2 billion are displayed in Table 6.1, there is still little real integration of AI technologies into pharmaceutical research and development workflows. Pharmaceutical companies are adopting AI protein design technologies cautiously, as evidenced by recent industry trends that show they mainly use these technologies through strategic partnerships rather than acquisitions (Soham Saha et al. 2024).

A significant variation (a "gap") exists between what AI and computer models anticipate will work in protein design and what works in lab tests. When scientists design and test proteins in lab settings or living organisms, the success rates are lower than what the computational models predicted. In addition, information asymmetry makes collaboration much more challenging because pharmaceutical businesses with large historical data must weigh knowledge sharing against intellectual property concerns. Furthermore, few companies have fully mastered the complex process of creating proteins, which necessitates the seamless fusion of computational power with pharmacological and biological domain knowledge.

6.6 Challenges and Future Directions

6.6.1 Technical Challenges

One fundamental challenge lies in the accurate prediction of protein dynamics and conformational flexibility (Schauperl and Denny 2022). While current methods excel

6 Protein Design Enters the AI Era

at predicting static structures, proteins are inherently dynamic molecules, and their function often depends on specific conformational changes. Developing AI models that can accurately capture and predict these dynamics represents a crucial next step in the field (Z. Smith et al. 2020). This challenge is particularly evident in cases involving allosteric regulation, where conformational changes at sites distant from the active site modulate protein function. Current AI methods struggle to capture these complex dynamic relationships, especially when multiple conformational states are involved in the protein's functional cycle.

The computational complexity of modeling protein dynamics presents additional challenges, particularly for large proteins or protein complexes. The vast configurational space that needs to be explored, combined with the long timescales of many biological processes, makes it computationally prohibitive to simulate protein dynamics comprehensively. This limitation is exacerbated when considering ensemble-based properties or when multiple dynamic processes occur simultaneously at different timescales. The development of efficient sampling methods that can adequately explore this vast conformational space while maintaining accuracy remains a critical challenge.

The integration of multiple scales of protein behavior presents another significant challenge. Proteins function across multiple temporal and spatial scales, from rapid atomic vibrations to slow conformational changes, and from local interactions to cellular-level organization. Current AI methods typically focus on one or two of these scales, but developing models that can seamlessly integrate across all relevant scales remains a significant challenge (Selvaraj, Chandra, and S. K. Singh 2021). This multi-scale integration becomes particularly complex when considering protein-protein interactions, where both local contact interfaces and global conformational changes play crucial roles. The challenge extends to understanding how atomic-level perturbations propagate through protein structures to affect macroscopic properties, a key consideration in protein engineering.

The accuracy of current AI models in predicting specific protein properties presents unique challenges. While models have achieved remarkable success in predicting protein structures, accurately predicting properties such as binding affinity, catalytic activity, and stability remains difficult. This challenge is compounded by the fact that these properties often depend on subtle, atomic-level details that current models may not capture effectively. The prediction of protein-protein interaction specificity and the

design of selective binding interfaces represent particularly challenging areas where current methods often fall short of desired accuracy levels.

Data quality and availability continue to present challenges, particularly for specialized applications. While the quantity of available protein sequence data has grown exponentially, high-quality structural and functional data remain relatively scarce. This limitation is particularly acute for certain classes of proteins, such as membrane proteins and intrinsically disordered proteins, where experimental characterization is especially challenging (Jincai Yang, C. Shen, and N. Huang 2020). The scarcity of high-quality training data for specific protein classes or functions can lead to biased or incomplete models that may not generalize well to novel design challenges. Furthermore, the available experimental data often contains inherent noise and uncertainties that can affect model training and validation.

Model interpretability expresses another significant technical challenge in AI-driven protein design. While DL models have demonstrated impressive predictive capabilities, understanding the basis for their predictions remains difficult. This lack of interpretability can make it challenging to identify potential failure modes or to improve model performance. The challenge is particularly acute when models make unexpected predictions, as it can be difficult to determine whether these predictions represent genuine insights or artifacts of the model's training data and architecture.

The validation of AI-generated protein designs presents additional technical challenges. While computational validation methods can provide initial assessments of design quality, experimental validation remains essential but resource-intensive. The development of more efficient validation protocols that can rapidly assess the success of designed proteins while providing meaningful feedback for model improvement represents an important technical challenge. This validation challenge becomes particularly significant when designing proteins for novel functions where established assessment protocols may not exist.

Despite these remarkable achievements, several significant challenges remain in AI-driven protein design. These challenges present opportunities for future development and highlight areas requiring focused research attention.

6.6.2 Future Directions

The future of AI-driven protein design holds immense promise, with several emerging trends likely to shape the field's development. These advances will likely come from both technological improvements and novel applications of existing methods.

Advanced AI Architectures: The next generation of AI architectures for protein design is likely to address current limitations while introducing new capabilities. One promising direction is the development of multi-scale transformers that can simultaneously process information at different levels of protein organization(Buehler 2022). These models would integrate atomic-level details with higher-order structural features, enabling more comprehensive protein design strategies. Current implementations often struggle with the simultaneous consideration of local interactions and global protein architecture, but emerging research suggests that novel attention mechanisms could bridge this gap, particularly in modeling long-range interactions that influence protein stability and function(Q. Zhao et al. 2022).

Significant advances are also expected in the area of generative models. Current approaches often struggle with the precise control of multiple protein properties simultaneously. Future models may incorporate improved conditioning mechanisms that allow designers to specify multiple desired properties with greater precision, such as stability, binding affinity, and catalytic activity(Angelo et al. 2023). This could enable the design of proteins with complex combinations of structural, functional, and dynamic properties, moving beyond the current limitations of single-property optimization.

The integration of physics-based knowledge with DL represents another crucial frontier. While current models learn primarily from data, future approaches may combine the strengths of physical modeling with the pattern-recognition capabilities of AI. This hybrid approach could be particularly valuable for designing proteins to perform novel functions where limited experimental data exists(Malbranke et al. 2023; Heo, Janson, and Feig 2021).

Experimental Integration: The future of protein design will likely see closer integration between computational and experimental methods(Brown et al. 2020). The development of high-throughput characterization methods, combined with AI-driven

experimental design, could create powerful feedback loops that accelerate the design process.

One promising direction is the development of active learning frameworks that can intelligently guide experimental efforts(Ren et al. 2021). These frameworks would identify the most informative protein variants to test, optimizing the use of experimental resources while maximizing knowledge gain. By incorporating experimental feedback in real-time, such systems could continuously refine their predictions and design strategies, leading to more accurate and reliable protein designs.

The integration of AI with automated experimental platforms expresses another exciting frontier. Recent advances in laboratory automation, combined with AI-driven design and analysis, could enable fully automated protein engineering pipelines. Such systems could dramatically accelerate the design-build-test cycle, enabling more rapid development of novel proteins.

6.7 Conclusion

The integration of AI in protein design has begun to fundamentally transformed our ability to engineer proteins with precise specifications and novel functions. This review has highlighted the remarkable progress achieved by applying DL, protein language models, and integrated design approaches. These advances have enabled unprecedented achievements across multiple domains, from therapeutic development to environmental applications.

The field continues to evolve rapidly, with new methods and applications emerging regularly. The challenges that remain, particularly in areas such as protein dynamics and multi-scale integration, depict opportunities for future development. The increasing integration of AI with experimental methods and the emergence of new application areas suggest that the impact of AI on protein design will continue to grow.

Looking forward, we anticipate that AI-driven protein design will play an increasingly central role in addressing major societal challenges, from healthcare to environmental protection. The combination of advancing AI technologies, improving experimental methods, and expanding application domains suggests that we are still in the early stages of realizing the full potential of this transformative approach to protein engineering.

6 Protein Design Enters the AI Era

As we move forward, continued advancement will require close collaboration between computational and experimental scientists, as well as engagement with emerging technologies such as quantum computing and automated experimentation. The future of protein design lies in successfully integrating these various approaches, working together to unlock new possibilities in protein engineering and its applications.

7 Protein Fragment and Structure Analysis Workstation

This chapter was published as: **Yanlin Mi**, Stefan-Bogdan Marcu, Sabin Tabirca and Venkata VB Yallapragada. (2023). *PROFASA - A Web-based Protein Fragment and Structure Analysis Workstation*. Frontiers in Bioengineering and Biotechnology, 11, p.1192094.

DOI:10.3389/fbioe.2023.1192094

Publisher: Frontiers

JCR Quartile: Q1

Impact Factor: 4.3

I made the primary contribution to this work through conceptualization, data curation, formal analysis, investigation, methodology development, resource provision, software creation, validation, visualization, and writing the original and final draft.

7.1 Abstract

In recent years, our understanding of protein structure and function has resulted in exponential growth in various applications such as food, diagnostics, therapeutics and novel materials. Today, there is an arsenal of in silico tools available for protein modeling (Cramer 2021), designing proteins (Walker, V. V. Yallapragada, and Tangney 2021) and analysis (V. V. Yallapragada, Walker, et al. 2020). However, most of these tools to-date require high technical expertise to access, install and effectively use them. Current in silico tools excel at providing cutting edge computational technology but fall short in usability due to a lack of user friendly interface and complicated working protocols. As the market need for protein based applications expands, it also raises the need for a one-stop platform for in silico tools. In this article, we introduce PROFASA, a

protein workstation, that aims to provide a unified platform for studying protein structure. PROFASA combines state of the art protein structure visualisation with cutting edge tools and algorithms for in silico protein analysis. Features such as real-time collaboration, note sharing, and shared projects make PROFASA a great tool for researchers and teaching professionals alike. The design, user interface, features, applications and ambitions of PROFASA are discussed in this manuscript.

7.2 Introduction

7.2.1 Myriad of Computational Tools for Tools

Understanding protein structure and function is a crucial aspect of life sciences. Commercially, synthetic proteins are poised to drive a major sector of synthetic biology (Walker, V. V. Yallapragada, and Tangney 2021). Protein engineering market is expected to reach USD 3,023.29 million by 2027 (Market Research Future 2023). The idea of using computational tools to enhance our scientific arsenal to study proteins is well documented (Walker, V. V. Yallapragada, and Tangney 2021). Highly significant and difficult problems (Kuhlman and P. Bradley 2019a) such as predicting the 3D structure of a protein from its amino acid sequence are now being solved with great accuracy by tools such as AlphaFold2, using advanced DL models. Entirely novel structures that never existed in nature can now be generated using de novo protein design (Gront et al. 2011). During COVID-19 pandemic, 3D protein visualisation is employed in the teaching of undergraduate medicinal chemistry courses to investigate drug-target interactions (Abdel-Halim 2020).

Protein engineering will become increasingly more efficient and precise in the future as artificial intelligence and DL progress. When it comes to protein structure prediction, we're living in exciting times, AlphaFold2 by Google's Deepmind has already been a blockbuster for the prediction of protein structures with high accuracy. Improvements are also expected in areas such as better DL-based algorithms for Multiple Sequence Alignments (MSA) generation; transformer-based approaches for protein structure prediction; DL-based approaches for multi-domain protein structure prediction and so on (Pakhrin et al. 2021). However, as in other fields, one obstacle to the wider use of DL in protein structure informatics is the black-box nature of DL models. In this context,

the development of Explainable Artificial Intelligence (XAI) approaches to improve the interpretability of protein structure predictions is an emerging trend in the area. Such as InterPretContactMap which was developed by Cheng's Lab (Adhikari, Hou, and J. Cheng 2018). It uses two attention mechanisms (sequence and regional) in the CNN framework to do contact map prediction and improves the contact map prediction results as well as provides some level of interpretability, providing some insights into the key fold-determining residues in the protein. Therefore, it is important to use advances in DL algorithms to fill the existing gap between protein sequence to protein structure, and XAI might become one of the methods to achieve the goal (Pakhrin et al. 2021). With de novo protein design methods, as the protein structure database expands, protein structures with novel functions can be generated through AI. All these necessitate a lot of mathematical power, thus, as with structure prediction, DL might propel the study of protein design and engineering to new heights. There are many freely available protein tools on the market. For sequence alignment, BLAST has long been a widely used tool (Sf 1997). For structure prediction, Rosetta and I-TASSER are the most popular online tools to process protein structure analysis (Leman et al. 2020; Jianyi Yang and Yang Zhang 2015), and the recent revolution AlphaFold2 is based on DL and millions of structures have already been generated by it (Cramer 2021). Apart from structure prediction, advancements can also be seen in protein visualisation and edutainment. Tools such as Chimera, Rasmol, Pymol and Csynth offer excellent features to highlight and visualise molecular structures in great detail (Pettersen et al. 2021; Fraley et al. 2022; Yuan, Chan, and Hu 2017; Todd et al. 2021). Recently, efforts have also been made to provide immersive viewing with virtual reality and augmented reality (V. V. Yallapragada, T. Xu, et al. 2021). In edutainment, Fold it by Rosetta is a fantastic resource for understanding the 3D mechanics of protein structures in a gamified environment (Curtis 2015). The above mentioned tools are only a snapshot of the rapidly expanding field of protein design. Although there has been extraordinary progress in the individual tools, a unified interface that can act as a one-stop-shop for all protein needs is required to bring the power of protein design to en masses and would expand the horizons of protein design.

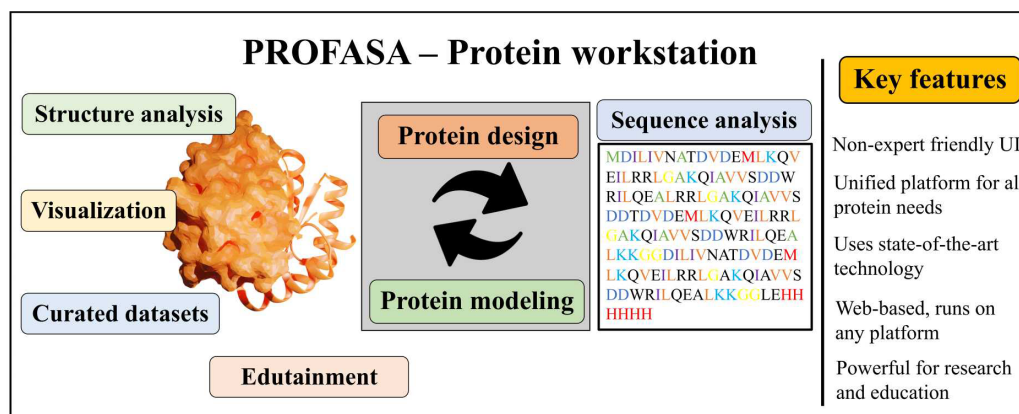


Figure 7.1: PROFASA, combines all the major computational tools and provides a single platform interface for computational protein works.

7.2.2 One tool to rule them all

PROFASA (Fig. 7.1) is an in-silico workstation for protein structure evaluation, parametric analysis, and protein visualisation. It provides (i) a unified interface for rapid conserved domain analysis of input sequences and displays the results using a molecular visualisation UI (ii) users can construct protein models with one click using the I-TASSER or the AlphaFold2, (iii) it can fold numerous input sequences, create new protein sequences, models, and calculate protein parameters, (iv) it uses tables and plots to display and compare all data, these results provide a guide to the expected protein performance prediction using Function2Form (F2F) plot (V. V. Yallapragada, Walker, et al. 2020).

PROFASA has fast model generation and structural analysis time and is faster than most tools on the market, as shown in the (Table 7.1). Users only have to submit a sequence to get parameters like accessibility, instability, hydrophobicity, isoelectricity, size, and Ramachadran plot score (Fig. 7.2). Drawing upon the precision of well-established bioinformatics tools, PROFASA provides accurate results across all its features. Its sequence alignment functionality uses the NCBI BLAST toolkit, which is known for its reliable sequence comparison capabilities. Structure prediction in PROFASA is handled through I-TASSER (CASP15 2023) and AlphaFold2v, both recognized as leaders in protein structure prediction, thus the accuracy in this aspect is assured. Furthermore,

Table 7.1: Model generation timetable

Uniprot ID	Length	Execution Duration			
		PROFASA with AlphaFold2 (with 6 parameters calculation)	PROFASA with I- TASSER (with 6 parameters calculation)	I-TASSER Server (with 1 parameter calculation)	AlphaFold2 Colab (with 1 parameter calculation)
P63212	57	38min	9h52min	12h32min	10min
P09210	100	40min	14h13min	21h05min	14min
P61626	128	40min	13h13min	21h51min	13min
O75469	135	44min	16h34min	18h34min	15min

calculations pertaining to protein superimposition and other protein parameters are performed based on standard, scientifically accepted formulas (see Materials and Methods section), ensuring the results generated are as precise as these foundational formulas. PROFASA has an excellent user interface, provides an interactive experience, and is simple to operate. When compared to most existing protein analysis tools, it is appropriate for both professionals and non-professionals due to its ease of use and data visualisation. This is a key feature in reducing the entry barrier for protein computation, analysis, and visualisation tools, allowing more people to participate in protein research and education (Fig. 7.3).

7.3 Materials and Methods

7.3.1 Architecture design

Serving as a one-stop workstation, we aim to provide the users with a platform for all the protein sequence and structure based analysis. The architecture design specifies the goal, key functions, activity elements, and important processes, as well as the nature of their interconnection (Gharajedaghi 2011). Each module of the PROFASA project can be split down based on the user requirements as shown in Fig. 7.4.

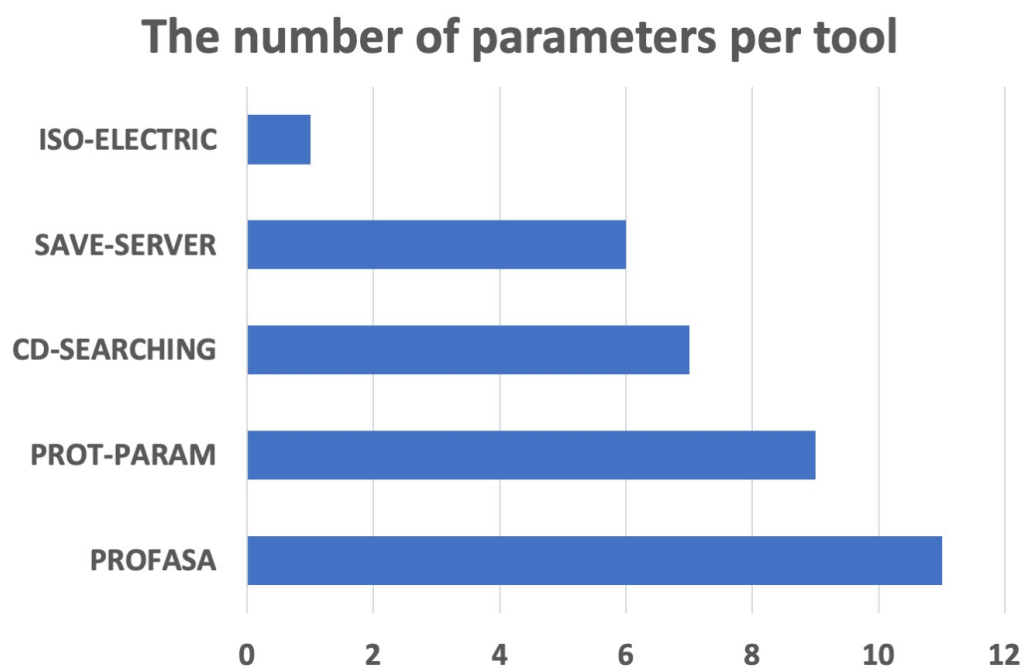


Figure 7.2: The number of parameters per tool

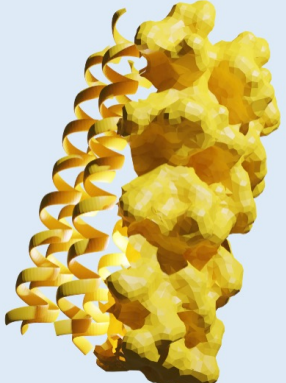
	Features	Details/Format
	Easy to use	Web-based, no download, cross-platform
	Structure build	I-TASSER and AlphaFold
	Two-way processing	Blast and Fold
	Result Format	Text, model, plot
	Visualization	Web3D/WebXR
	Structure format	Cartoon, Backbone, Ball+stick, Rocket, Ribbon, Rope
	Sequence operation	Positioning amino acids, split protein
	Model operation	Change color, change structure format, protein superimpose

Figure 7.3: PROFASA key features

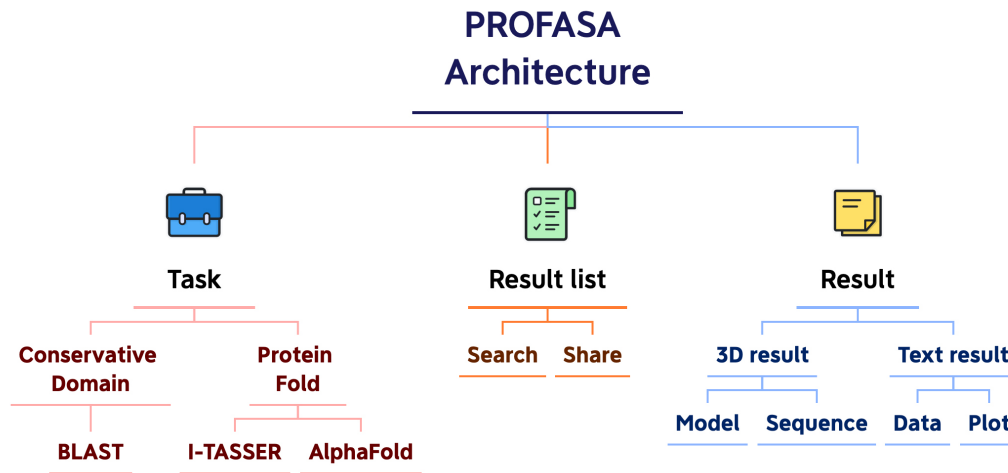


Figure 7.4: PROFASA architecture

The application architecture of PROFASA describes design and development patterns and techniques. When developing applications, a strong application architecture gives a roadmap and recommended practices to follow (Eder and Missikoff 2003). Applications in PROFASA are divided into presentation layer, logical layer, and computing layer. The front end is the display layer, which contains the Home page, Blast page, Fold page, Resources page, Result List page, Text-result page and 3D-result page. Node.js part of the backend is a logical layer, which is divided into account service, mail service, note service, sequence service, I-TASSER service, AlphaFold service, and parameter service. Finally, the command line program part is the computing layer, which is divided into RpsBlast, I-TASSER, AlphaFold2 and Ramachandran.

7.3.2 Front end design

PROFASA uses TypeScript as its primary programming language because it operates in a web environment. TypeScript is a JavaScript superset that adds types, interfaces, and other useful features to JavaScript (Bierman, Abadi, and Torgersen 2014). ReactJS, the most popular framework in the world, was chosen for the technical decision of the programming development framework. It's a JavaScript library for designing user interfaces using JSX quickly and effortlessly (TSX is used in PROFASA). UmiJS is used

to handle the react-router and react-state in this project. It is an enterprise-level React application framework created by Alibaba that includes comprehensive routing functions (basically an improved encapsulation of react-Router and react-router-DOM) and a set of state management techniques based on fixed hooks. This is a lot easier than Redux. PROFASA is a powerful and versatile work station with a wide range of complicated and comprehensive capabilities. It uses two high-quality JavaScript libraries: Three.js and ngl.js, to perform its main duties of visual rendering and interaction with protein 3D models. Ricardo Cabello, aka Mr.Doob, designed Three.js in 2010 (Danchilla and Danchilla 2012). It has a variety of draw modes and can fall back to a 2D rendering environment if WebGL isn't available. The amount of initial effort or "boilerplate" required is reduced with default settings (Danchilla and Danchilla 2012). The lighting, camera, camera controller, renderer, animation, and mouse interaction events in the scene are all built using Three.js. Because users may transition between multi-model and single-model scenes with a single click, PROFASA needs to provide varied mouse hover and click interactions in different scenes, the 3D scenes feature particularly complicated mouse interactions. The ngl.js is used to load and parse PDB files, which are available objects in Three.js. In addition, one of PROFASA's most important functions is the ability to superimpose, extract protein sequences and change the color and shape of selected fragments, which is also dependent on ngl.js. Multiple PDB files are loaded using ngl.js when users visit the 3D result page. Following successful loading, different colored 3D models will be rendered based on the quantity of models. When users change the style of the selected piece of the model, or the style of the entire model, it effectively destroys the old model and produces a new one using ngl.js. The ngl.js is also used by protein superimpose to determine comparable amino acids in two proteins and recalculate the locations of the proteins to finish the superimpose. The Root-mean-square Deviation (RMSD) is an essential statistic for determining the excellent and bad superimpose criterion for protein. The average distance between the atoms of the two stacked proteins is measured by the RMSD (B. Patel, V. Singh, and D. Patel 2019). The smaller the value, the closer the two overlaid proteins are in terms of exterior features and functions, and the more they converge. RMSD can be calculated using Formula 7.1. One of the first steps in calculating RMSD is to align the sequences. Sequence alignment is a method for matching protein sequences in order to find functional or structural similarities or differences between them (Chao, F. Tang, and L. Xu 2022). PROFASA

calculates it using a matrix technique (T. Wang et al. 2017). Sequence alignment determines which residues in two sequences are identical, then locates the alpha carbon atoms that correspond to these residues and obtains their position coordinates using the protein object, which is generated when parsing the PDB file and will update the position coordinates of the atoms in it after superimpose of the protein. Then, after having all of the atoms' location coordinates, Formula 7.2 could be used to get the RMSD. In addition, through the coordinates of these alpha carbon atoms, The Global Distance Test - Total Score (GDT_TS) can be calculated. When the two structures differ in detail, GDT_TS (CASP14 2023; CASP13 2023) is better than RMSD (CASP14 2023; CASP13 2023) in detecting superimpose similarity (Zemla 2003). GDT_TS can be calculated by Formula 7.2, where Global Distance Test - Percent Nativeness (GDT_PN) denotes percent of residues under distance cutoff $N \leq \text{\AA}$, as in Formula 7.3 shown.

$$RMSD(v, w) = \sqrt{\frac{\sum_{i=1}^n ((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2)}{n}} \quad (7.1)$$

$$GDT_TS = \frac{GDT_P1 + GDT_P2 + GDT_P4 + GDT_P8}{4} \quad (7.2)$$

$$GDT_PN = \frac{N[(v_x - w_x)^2 + (v_y - w_y)^2 + (v_z - w_z)^2]}{n} \times 100 \quad (7.3)$$

PROFASA uses Canvas to create a radar plot of F2F Result, allowing users to intuitively see and compare the data of each protein, as shown in Fig. 7.5. Two of the reasons for drawing manually rather than using an existing chart library, are the need for extensive flexibility and the need to keep the downloaded visuals consistent with what is displayed on the website. This is done with the Pixi.js library, which is the world's fastest WebGL-2D rendering tool. In addition, PROFASA also has a document download feature for each protein sequence. Users can individually copy the sequence, download the PDB file, download the diagram, or download the complete PDF document. To display and download PDF files, the React PDF Renderer library is used.

PROFASA has always been built with a wide variety of users in mind, including professionals and non-professionals. As a result, the visual focus is stronger, the interaction mode is friendlier, and the learning costs are reduced.

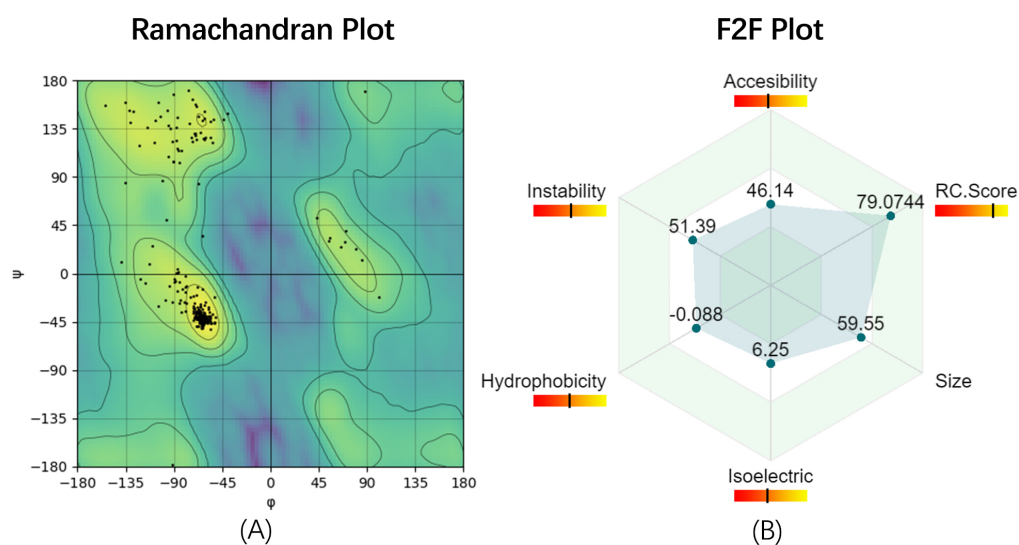


Figure 7.5: Ramachandran Plot and F2F Plot. (A) Ramachandran plot is used to describe whether the conformation of the protein is reasonable (Hooft, Sander, and Vriend 1997). (B) F2F plot shows six important protein parameters that will be used in protein predictive analysis in the future (V. V. Yallapragada, Walker, et al. 2020).

7.3.3 Back end design

The backend of PROFASA is written in JavaScript and runs in the Node.js environment, a runtime environment based on the Chrome V8 JavaScript Engine that runs outside the browser. For building Restful APIs, PROFASA uses KOA as a web server framework. KOA is a Node.js web framework with a powerful middleware system that can be used to handle request processing, CORS processing, token authentication, request interception, error handling, logging, and other operations elegantly. Furthermore, PROFASA uses Sequelize as an Object Relational Mapping (ORM) framework to build relationships between code data structures and database tables. Using Sequelize eliminates the need to write most of the SQL. It is very helpful for database migration and test environment setup. In version 2.0, the backend added the ability to take simple notes on the current model and the ability to share tasks between users. These features, although seemingly simple at the moment, are important changes in the long run when it comes to protein education and multi-person collaboration.

PROFASA uses Node.js not just to establish servers, listen to requests, and read and write databases, but also to run a variety of command-line scripts to perform complicated features. Open-source third-party programs and Python scripts produced by the developers themselves are examples of this. Command-line programs are run in the main Node.js program using the Node.js standard library's `child process.exec()` and `child process.execSync()` methods. The most of these command-line programs are C or C++ compiled programs, or C or C++ programs launched by Python or the Shell. This provides these programs a lot of power and performance. In other words, PROFASA's back end acts as a glue that holds together C or C++ programs that perform well and efficiently, but are difficult to install, difficult to use, and expensive to learn. For example, AlphaFold requires not only sophisticated hardware but also programming skills; The calculation process of I-TASSER is usually measured in days and cannot run continuously on PC; RpsBlast's command-line program is almost unreadable to non-experts. Despite the complexity of the back-end logic, users can access the results of all these fantastic programs through the PROFASA front-end interface, which is simple to use and comprehend. Among them, the RpsBlast program does conservative domain queries, which is one of PROFASA's fundamental features. RpsBlast searched according to Conserved Domain Database (CDD) and obtained a set of protein profiles

7 Protein Fragment and Structure Analysis Workstation

(Biotechnology Information (US) and C. Camacho 2008). However, in PROFASA, the RpsBproc program is used to parse and filter the RpsBlast results and perform ACD dataset searches against the results. The results are then parsed line by line, picking out useful information to display in the front end along with the RpsBproc results. As the most accurate and stable program for protein structure prediction in the non-neural learning area, the I-TASSER program is undoubtedly an important aspect of PROFASA. AlphaFold2, a tool developed by Google, is now the most advanced, powerful, accurate, and rapid neural learning protein structure prediction program on the market. Therefore, PROFASA added support for AlphaFold2.

Table 7.2: F2F plot parameters

F2F parameters	Scale range	Description	Generated
<i>Size</i>	kDa	Total weight of a protein	ProtParam Hosted by Expasy
<i>RC.Score</i>	0 to 100	Quality of a model	Saves Server
<i>Accessibility</i>	0 to 100	Active site solvent accessibility	PROFASA
<i>Instability</i>	0 to 100	Half-life of protein in vitro	ProtParam Hosted by Expasy
<i>Hydrophobicity</i>	-4.5 to +4.5	Surface hydrophobicity	ProtParam Hosted by Expasy
<i>Isoelectric</i>	0 to 14	Point at which molecule carries no net charge	Protein Isoelectric Point Calculator

The PROFASA server also supports calls to RCSB search, Protparam, isoelectric.org, and SAVES server, which are all well-known Protein APIs. PROFASA analyzes these API calls further and visualizes the intended outcomes, making it simple to learn about protein parameters and what they signify. As shown in Table 7.2, it lists the parameters in the F2F plot, some of which may be accessed using the APIs mentioned above. Among them, as shown in Formula 7.4, it can be used to calculate accessibility, which is defined as the relative accessible surface area (Marsh and Teichmann 2011). Accessible Surface Area (ASA) (Wikipedia 2023b) is determined using the Dictionary of Secondary Structure of Proteins (DSSP) tool, whereas MaxASA (Wikipedia 2023b) is calculated by

traversing all residues in PDB files and summing according to MaxASA for each residue, as shown in Table 7.3.

$$Accessibility = RSA = \frac{ASA}{MaxASA} = \frac{\Sigma ASA}{\Sigma_{residue} maxASA} \quad (7.4)$$

Table 7.3: MaxASA of amino acids

<i>Residue</i>	<i>MaxASA</i>	<i>Residue</i>	<i>MaxASA</i>
ALA	129.00	LEU	201.00
ARG	274.00	LYS	236.00
ASN	195.00	MET	224.00
ASP	193.00	PHE	240.00
CYS	167.00	PRO	159.00
GLU	223.00	SER	155.00
GLN	225.00	THR	172.00
GLY	104.00	TRP	285.00
HIS	224.00	TYR	263.00
ILE	197.00	VAL	174.00

Because PROFASA focuses on front-end interaction and back-end command line calls and results processing, the database was the easiest one of the project. All PROFASA database tables have three common fields: ID (Private Key), CreatedAt, and UpdatedAt. For the details, as shown in Fig. 7.6.

PROFASA: Users can try PROFASA using this link: <https://profasa.ucc.ie/#/>

7.4 Result

PROFASA (Fig. 7.7) is divided into four sections: Analysis, Superimpose, Structure Blast and the Fold. Each section takes a different input and provides outputs as discussed below.

7 Protein Fragment and Structure Analysis Workstation

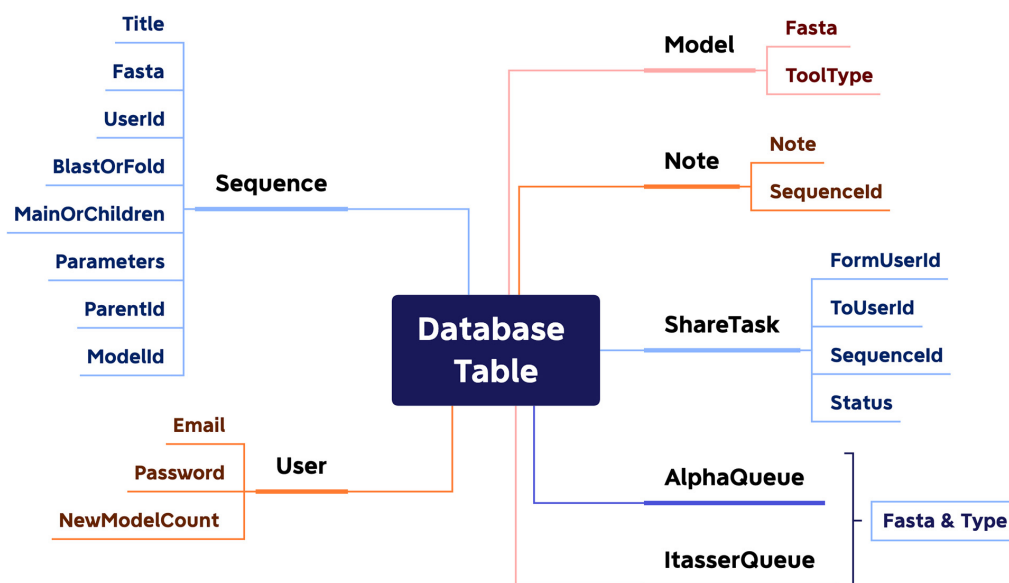


Figure 7.6: Database design

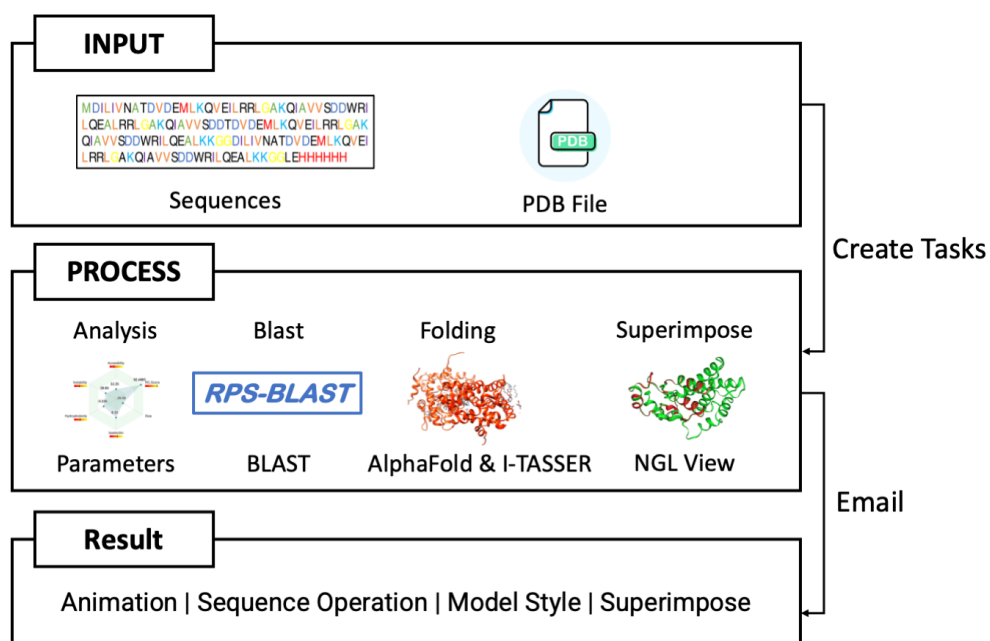


Figure 7.7: PROFASA flowchart

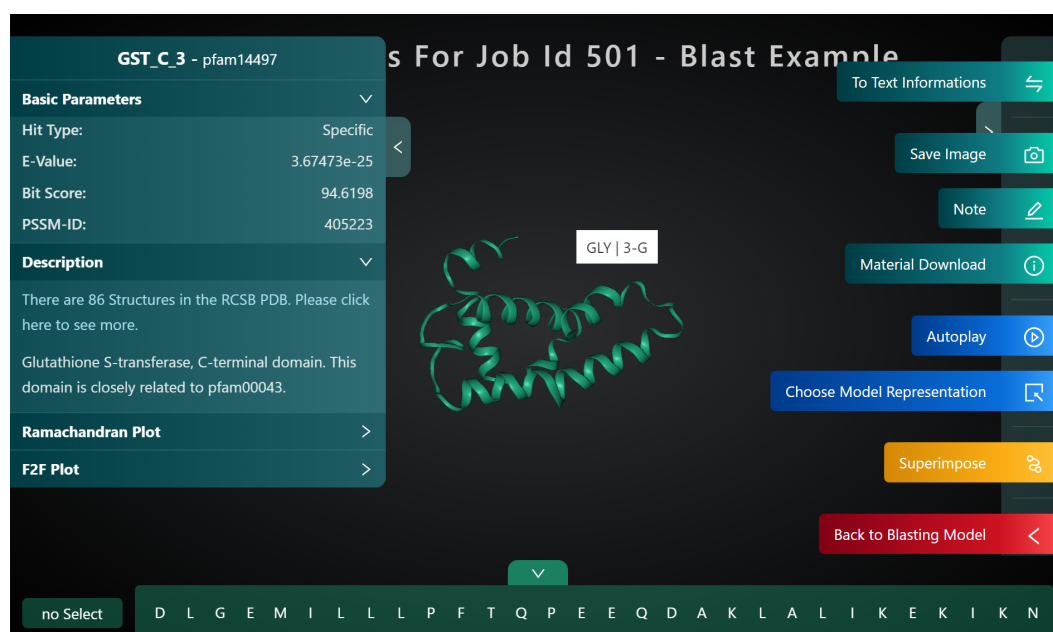


Figure 7.8: Analysis user interface

Analysis: To compute and understand various structural, sequence based and functional parameters of a protein (Fig. 7.8), the user would upload a PDB file and PROFASA would compute all the parameters (as listed previously) and generate a PDF file.

Superimpose: Superimposing two or more 3D structures is a powerful way to examine the differences in the structures. Biologists could use this for studying mutations, changes to functional hotspots, improving a novel fragment. PROFASA provides a 3D live image of the input PDB files, superimposed in the best configuration and calculates RMSD and GDT_TS to mathematically compare the difference between the structures (Fig. 7.9).

Structure blast: Structure blast (Fig. 7.10) is one of the unique features that we have developed for PROFASA. Traditional sequence based BLAST finds matches for the full and parts of a user defined sequence. PROFASA extends this a step further by giving a 3D model for the hits. These hits could either be modeled using I-TASSER or Alpha Fold for unknown sequences. As a result the user would be able to visualize how a large complex protein can have domains and parts from various smaller protein fragments existing in nature.

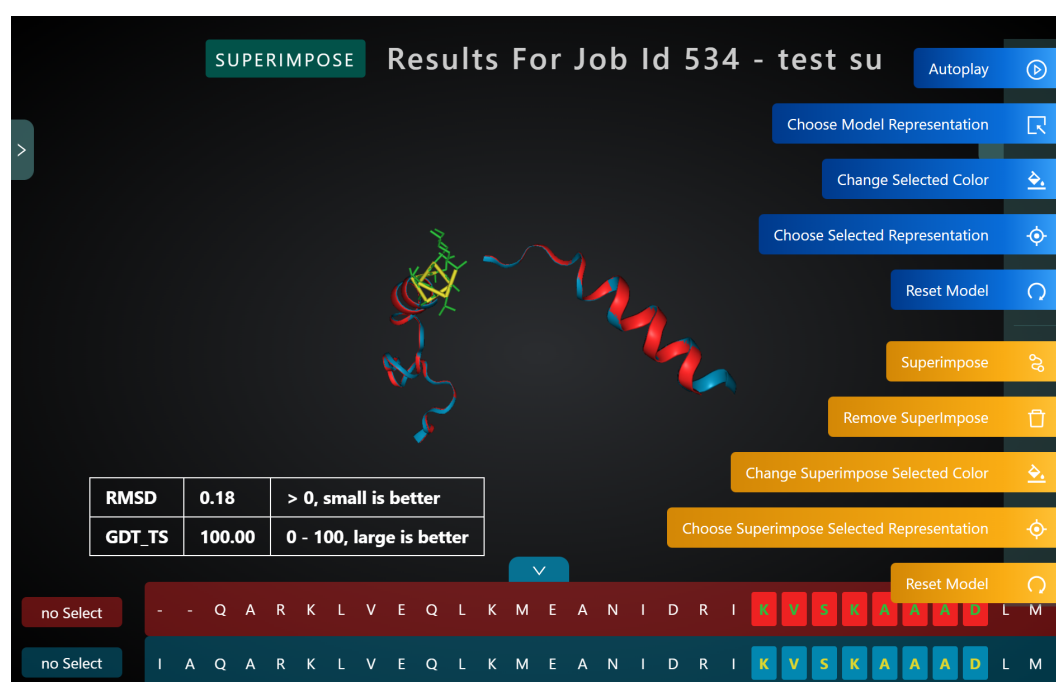


Figure 7.9: RMSD and GDT-TS calculations are used to compare the differences between two superimposed 3D structures

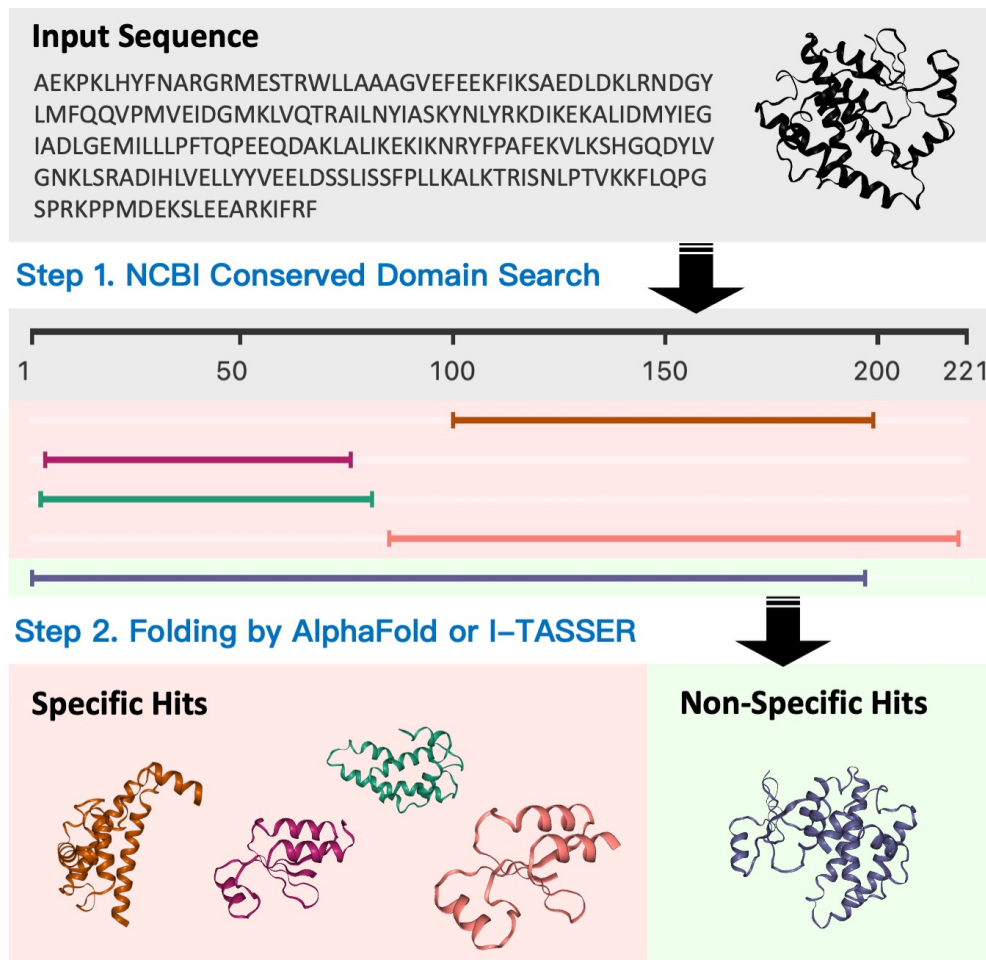


Figure 7.10: Structure blast

7 Protein Fragment and Structure Analysis Workstation

Fold: PROFASA provides a user-friendly UI for protein folding or modeling. The user can fold multiple sequences in parallel and choose between I-TASSER and AlphaFold. The output is detailed in the in-built 3D visualisation tool along with all the computed parameters.

Molecular visualisation: Having a clean, simple and interactive molecular visualisation embedded integrally into every result is a key for PROFASA's highly user friendly interface. Unlike other protein analysis tools and folding software, users can access their 3D structure and interact with them on the same platform. Our web-based molecular visualisation is an important step that turns PROFASA into a one-stop-shop tool for biologists.

Bio edutainment: The most important feature of PROFASA is the interaction with individual models, which sets it more characteristic than other protein analysis tools, as shown in Fig. 7.8. Hovering the mouse over the model reveals the amino acid at the present location as well as the sequence subscript. Additionally, the user could alter the color and style of the model area corresponding to a certain section of the sequence by selecting it by the sequence letter. Users also would be able to make notes on each structure and project. This level of interactivity makes PROFASA an excellent tool for teaching and bio edutainment.

7.5 Discussion and Outlook

7.5.1 As a single UI protein workstation

Protein structure modeling and analysis is a crucial aspect in computational biology and bioinformatics, as it allows researchers to study the structure and function of proteins. Protein structures are complex, and determining their 3D structures experimentally can be time-consuming and expensive. Thus, computational methods are often used as a complement to experimental approaches. However, protein structure modeling and analysis are relatively difficult processes that frequently transition between many platforms or applications, posing a steep learning curve even for experienced computational biologists. Many of these tools and platforms require significant computational skills, as well as programming. This makes it challenging for biologists without a computational background to get started in the field. To address this problem, PROFASA, a

one-stop protein structure modeling and analysis workstation, was created. The aim of PROFASA is to offer one platform for all protein sequence and structure analyses needs. It is worth noting that PROFASA has some important advantages over other existing tools, such as SAMSON Connect. For example, SAMSON Connect (SAMSON Connect 2023) is an excellent integrated molecular design platform. Unlike the existing tools, PROFASA offers a one-stop solution with integrated tools (expanding library), robust data management, and no need for downloads as a web-based platform. Its user-friendly interface and low learning curve make it accessible to researchers without extensive computational backgrounds. We believe PROFASA is set to become an indispensable tool for researchers and educators in their quest to unravel the mysteries of protein structure and function. See Fig. 7.11.

Protein Design: PROFASA contributes significantly to the development of protein design as a one-stop workstation for modeling and analysis of protein structures. It combines into one platform many protein modeling and analysis tools such as, protein homology analysis, protein structure modeling, structure and function based parametric analysis, and protein structure superimposition. This enables researchers to quickly find potential ideas with a high chance of success without switching platforms. Another key advantage of the one-stop workstation is its ability to improve the accuracy and consistency of predictions by lowering the chance of practical mistakes and human error. PROFASA also enhances data management by enabling the integration and sharing of data obtained from multiple sources, facilitating the expansion of computational protein design. This can lead to the development of more sophisticated prediction techniques, and a large dataset can be accumulated to support the subsequent incorporation of machine learning algorithms. PROFASA has a user-friendly interface. This makes it easy to acquire and analyse protein structures for non-experts.

Educators: PROFASA was created with its use cases in the education field in mind from the start. It features powerful interactive molecular structure modification and gamified interface for molecular structure viewing, including high-definition 3D model display. Such a gamified user interface offers an immersive and interactive learning experience in contrast to traditional passive teaching techniques like lectures and textbook reading, and it has been shown to be a successful tool for science teaching and learning (Sailer and Homner 2020). Further, PROFASA offers functionality for high-resolution note-taking on protein structures as well as peer-to-peer sharing of results from protein

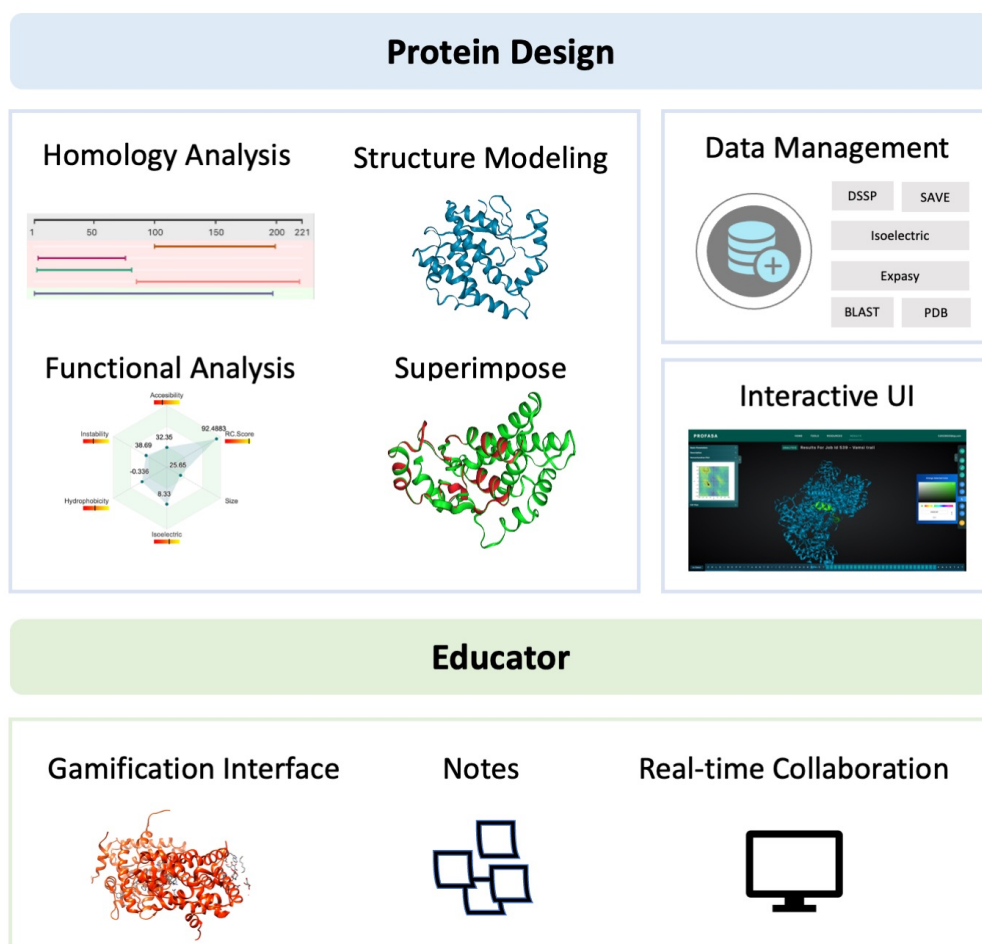


Figure 7.11: PROFASA functional classification

structure analyses. These features can meet the needs of both teachers who want to share teaching cases with students and students who want to submit learning outcomes to teachers. It meets the need for education to be enjoyable while simultaneously guaranteeing that instruction is effective and accurate. This makes it simpler to instruct students in computational biology, biochemistry, and biotechnology. Also, it helps to advance the training of researchers in the area.

7.5.2 Bottlenecks and challenges for PROFASA

While PROFASA presents an innovative and efficient solution, one major challenge of PROFASA arises from the fact that PROFASA uses multiple external tools to perform its analyses, and the accuracy of the final results depends heavily on the accuracy of these external tools. Another challenge is that these analyses are computationally intensive processes, which can be both time-consuming and expensive to maintain. Additionally, given that the tool aims to generate large amounts of data, there is a need to ensure that the data is stored safely and securely, particularly if it is to be released openly. A further challenge is the need to continually improve the speed and accuracy of the tool, in order to keep pace with the rapidly evolving field of computational biology and bioinformatics.

7.5.3 Ambitions and future outlook

PROFASA already is a powerful tool for researchers and educators. PROFASA 2.0 is currently being created with more advanced features and ambitions, see Fig. 7.12. PROFASA 2.0 would include an automated workflow system to enhance the user experience for researchers with less experience and to expand data accessibility. Automation of parameter optimization, prediction validation, mistake detection, and report production are some of the features included in this automated workflow system. In addition to making protein structure modeling and analysis much simpler, this will also enable timely self-checking, which will lower the data error rate and raise the system's degree of confidence.

Furthermore, PROFASA 2.0 would be upgraded with high-performance computational resources, like server clusters and load balancing and specialised GPU resources. This would allow capabilities for multi-scale modeling at various resolution levels, including coarse-grained, all-atomic, and explicit solvents. Additionally, PROFASA 2.0 will incor-

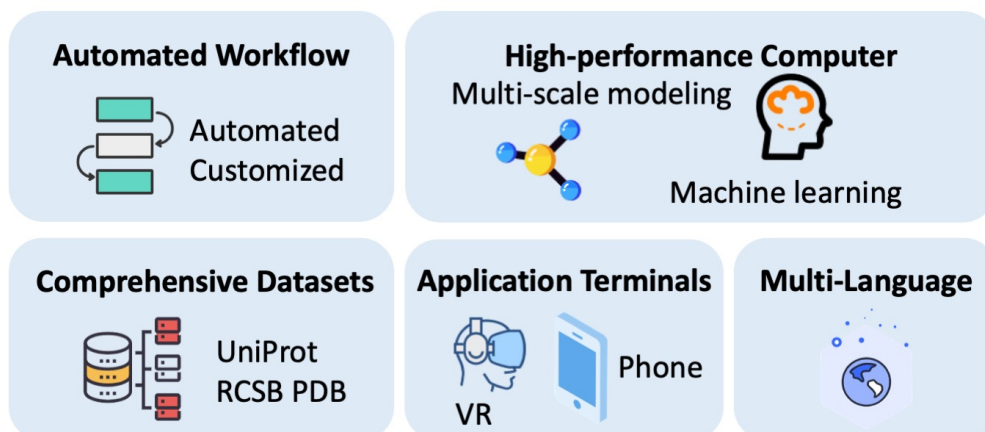


Figure 7.12: Ambitions and future outlook

porate more comprehensive datasets to enhance protein prediction speed and accuracy in the future. These databases may include experimental data, functional annotations, information on protein evolution, and known protein structures and properties. PROFASA 2.0 would also be accessible through virtual reality headsets that brings an immersive learning experience to the users. Finally, the PROFASA 2.0 interface would also be available in multiple languages to extend its potential user base and make it more accessible to researchers from other locations.

7.6 Data and Software Availability

PROFASA: Users can try PROFASA using this link: <https://profasa.ucc.ie/#/>

The dataset used in this study includes the following resources that are essential for running AlphaFold2: The Bacterial, Fungal, and Plant protein DataBank (BFD), UniProt, MGnify, PDB70, PDB_mmCIF, PDB_seqres, UniRef30, UniRef90. In addition, all the code used in this study was written using JavaScript (ECMAScript 6) and Python 3.9. The front-end code relied on several npm libraries, including Ant Design (v4.16.13), NGL Viewer (v3.0.5), Pixi.js (v5.3.5), React (v17.0.2), Swiper (v6.8.4), Three.js (v0.129.0), and Umi (v3.5.19). Furthermore, The back-end code used several libraries from the Node.js ecosystem, including Koa (v2.13.1), Sequelize (v6.6.5), jsonwebtoken (v8.5.1), and mysql2 (v2.3.0). The database used in this study was MySQL (v5.7.34). Additionally,

7.6 Data and Software Availability

the Python scripts used the BioPython (v1.78) library. The AlphaFold2 (v2.0.0) software and BLAST (v2.11.0) tool were also used as third-party software in this study. In addition to the libraries and software mentioned earlier, this study also relied on external websites to calculate several protein parameters. It includes the Protparam website and Isoelectric.org website. It's worth noting that these two websites are freely accessible and do not require any special permissions or licenses to use.

For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

8 PS-GO Parametric Protein Search Engine

This chapter was published as: **Yanlin Mi**, Stefan-Bogdan Marcu, Sabin Tabirca and Venkata VB Yallapragada. (2024). *PS-GO Parametric Protein Search Engine*. Computational and Structural Biotechnology Journal, 23, pp.1499-1509.

DOI:10.1016/j.csbj.2024.04.003

Publisher: Elsevier

JCR Quartile: Q1

Impact Factor: 4.4

I made the primary contribution to this work through conceptualization, data curation, formal analysis, investigation, methodology development, resource provision, software creation, validation, visualization, and writing the original and final draft.

8.1 Abstract

With the explosive growth of protein-related data, we are confronted with a critical scientific inquiry: How can we effectively retrieve, compare, and profoundly comprehend these protein structures to maximize the utilization of such data resources, PS-GO, a parametric protein search engine, has been specifically designed and developed to maximize the utilization of the rapidly growing volume of protein-related data. This innovative tool addresses the critical need for effective retrieval, comparison, and deep understanding of protein structures. By integrating computational biology, bioinformatics, and data science, PS-GO is capable of managing large-scale data and accurately predicting and comparing protein structures and functions.

The engine is built upon the concept of parametric protein design, a computer-aided method that adjusts and optimizes protein structures and sequences to achieve desired

biological functions and structural stability. PS-GO utilizes key parameters such as amino acid sequence, side chain angle, and solvent accessibility, which have a significant influence on protein structure and function. Additionally, PS-GO leverages computable parameters, derived computationally, which are crucial for understanding and predicting protein behavior.

The development of PS-GO underscores the potential of parametric protein design in a variety of applications, including enhancing enzyme activity, improving antibody affinity, and designing novel functional proteins. This advancement not only provides a robust theoretical foundation for the field of protein engineering and biotechnology but also offers practical guidelines for future progress in this domain.

8.2 Introduction

8.2.1 Topology of protein landscape

Within the field of life sciences, deciphering and understanding the intricacy of proteins remain substantial challenges. Proteins, as the fundamental functional units in biological systems, exhibit functionalities determined primarily by their structures. With the explosive growth of protein-related data, we are confronted with a critical scientific inquiry: **How can we effectively retrieve, compare, and profoundly comprehend these protein structures to maximize the utilization of such data resources?** The advancement of Internet technology offers possible resolution pathways. Take Google's search engine as an example; it spearheaded a significant transformation in the information retrieval field. Google's search engine, through its innovative PageRank algorithm, successfully tackled the problem of swiftly locating relevant information among a vast amount of web pages (Brin and L. Page 1998). PageRank algorithm, which is based on the hyperlink structure of web pages to assess their significance, enabled the effective ranking of web pages. Consequently, users can quickly obtain the most relevant search results. This structure-based processing approach significantly elevated the efficiency of information retrieval. Inspired by this strategy, we need to introduce similar methods in the field of biosciences to process and interpret the massive volume of protein data. The ideal solution should profoundly reveal the topological structure of proteins, accurately infer their functional properties, and perform precise searches and comparisons within large-scale

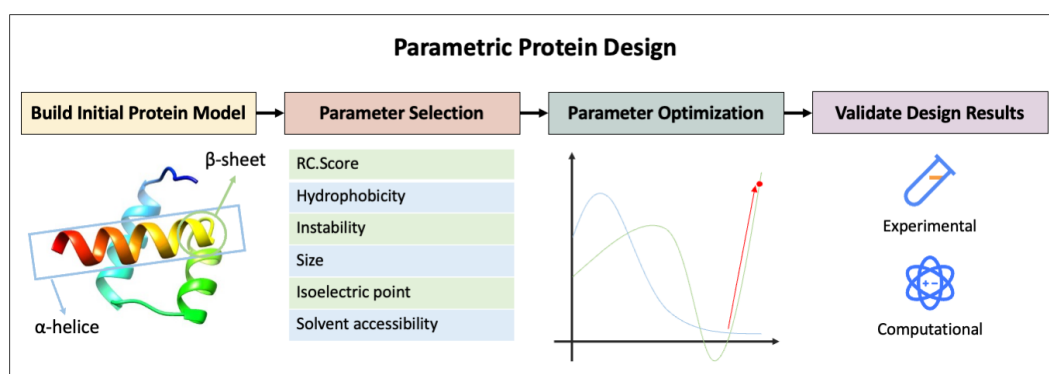


Figure 8.1: The steps in parametric protein design.

protein datasets. To achieve this goal, we require a tool that integrates computational biology, bioinformatics, and data science, capable of handling big data and accurately predicting and comparing the structure and function of proteins.

Protein topology, as the arrangement and spatial relationship of secondary structure elements, plays a crucial role in understanding key aspects such as protein folding, stability, function, and evolution, as shown in Figure 8.1. There is a close connection between protein structural topology and its biological function. A specific structural topology of the protein determines key functional features such as active sites, subunit interaction interfaces, and molecular recognition patterns. Studying the structural topological features of proteins can better predict protein function and reveal functional and structural similarities between different protein families (Murzin et al. 1995). Furthermore, studying protein structural topology can help to reveal the relationship between protein structure and function, and in turn, apply this knowledge to protein design. The goal of protein design is to create proteins with desired functions, stability, and solubility. During the design process, researchers often use computational methods to optimize protein sequences to achieve the desired structural topology.

However, due to the diversity of amino acid species and the endless possibilities of arrangement, the number of combinations of protein structures is enormous. In this vast search space, many combinations of amino acids do not produce stable protein structures or have biological functions. Therefore, understanding the quantitative differences between the available mathematical combinations and the actual possibilities is important for parametric protein searches to explore protein sequence space more

efficiently and discover proteins with the desired properties. In this regard, algorithms such as machine learning and metaheuristics can help to filter out protein sequences with potential biological functions, thereby reducing the complexity of the search space. **In practice, the combination of existing knowledge of protein structures, parametric models and efficient search strategies can be used to mine proteins with specific functions, stability and solubility with limited computational resources, providing strong support for protein design and optimization.**

8.2.2 Parametric protein design

Parametric protein design is a computer-aided protein design method, with the primary goal of adjusting and optimizing parameters of protein structures and sequences to achieve desired biological functions and structural stability (Kuhlman and P. Bradley 2019b). To achieve this goal, the right combination of parameters needs to be found to give the protein the desired function and properties. The key steps in parametric protein design are shown in Figure 8.1. The two most important steps in the design process are parameter selection and parameter optimization. Parameter selection requires the selection of parameters related to protein structure and function, such as amino acid sequence, side chain angle, and solvent accessibility, according to the design objectives (Rebecca F. Alford et al. 2017); parameter optimization can be classified according to the optimization strategy and search strategy. Optimization strategies mainly include energy minimization and probabilistic search. Energy minimization methods focus on optimizing the energy of protein structures to improve structural stability and function; while probabilistic search methods employ random sampling and acceptance or rejection criteria to search within the parameter space. Search strategies include genetic algorithms, Monte Carlo simulations, and more. As a heuristic search method, genetic algorithms have high search efficiency and flexibility; Monte Carlo simulations are suitable for handling complex energy landscapes and parameter relationships (Gainza, K. E. Roberts, and Donald 2012). Parametric protein design, and in particular the parameter selection and parameter optimization steps involved, is provided by theoretical underpinnings and practical guidance from protein structure topology studies.

The study of protein structure topology can provide a basis for parameter selection in the parametric protein design process. For example, it has been found that the interactions

between protein secondary structure elements and the mechanism of tertiary structure formation are crucial for protein function (Dill and MacCallum 2012). In addition, understanding the topological characteristics of protein structures helps to better grasp the impact of parameter tuning on protein structure and function. For example, protein structural topology studies have revealed certain general principles of structural stability, such as hydrogen bonding networks and hydrophobic cores (Mirny and Shakhnovich 1999). Furthermore, Protein structure topology studies have revealed the phenomenon of "protein folding topology" (Orengo, D. T. Jones, and J. M. Thornton 1994), which provides valuable clues for parametric protein design. By changing parameters such as amino acid sequence or side chain angle, proteins can be guided to form specific topological patterns to achieve function and stability (Kuhlman and P. Bradley 2019b).

Understanding the relationship between structure, parameters, and function is crucial in parametric protein design (Rebecca F. Alford et al. 2017). Parameters such as amino acid sequence, side chain angle, and solvent accessibility have a significant impact on protein structure and function (Gainza, K. E. Roberts, and Donald 2012). Firstly, the amino acid sequence determines the primary structure of a protein and indirectly influences its secondary, tertiary, and quaternary structure. By adjusting the amino acid sequence, properties such as the folding stability, functional activity, and affinity of a protein can be altered (Murzin et al. 1995). In addition, the side chain angle determines the spatial arrangement of the amino acid side chains during protein folding and affects the steric conformation and stability of the protein. Solvent accessibility affects the interaction of the protein with its surroundings, including binding to ligands, substrates, or other proteins. By optimizing the side chain angle and solvent accessibility, the affinity of a protein to a specific ligand can be improved, allowing for specific antibody design (Gainza, K. E. Roberts, and Donald 2012). Parametric protein design has played a key role in many successful examples of protein design, including enzyme activity improvement, antibody affinity enhancement, and the design of novel functional proteins (Rebecca F. Alford et al. 2017). These success cases provide ample evidence of the reliability and effectiveness of parametric protein design in achieving specific goals.

Computable parameters are a class of parameters related to protein structure and function that can be derived computationally. These parameters are essential for understanding and predicting the behavior of proteins, as they can directly affect the structure, stability, hydrophobicity, and other properties of proteins, as shown in Table 8.1.

Table 8.1: Computable parameters and description

RC Score: It is a parameter that describes the importance of amino acid residue interactions. The higher the score, the more critical the amino acid residues are in the protein structure. Calculating the RC score helps to optimize the stability and folding ability of the protein and is essential for improving thermal stability and resistance to protease degradation. Hydrophobicity: It is a parameter that measures the hydrophilicity of amino acid residues and is usually calculated using molecular dynamics simulations and computational chemistry. Optimizing the hydrophobicity of proteins enhances their solubility, stability, and interaction with other molecules. Instability: It refers to the half-life of a protein molecule in vitro and is closely related to factors such as amino acid sequence, structure, and hydrophobicity. Calculating the instability of amino acid residues helps to predict the stability and susceptibility to degradation of the whole protein molecule, which is of great importance for drug development and protein engineering. Size: Protein size refers to the total number of amino acid residues in its molecule. Size can have an impact on the folding rate, stability, and function of a protein. Calculating protein size can help predict its properties and function and provide a basis for optimizing protein structure. Isoelectric: The isoelectric point is the pH value of a protein molecule when the sum of its charges is zero. Calculating the isoelectric point helps to predict the charge state of a protein in different pH environments and further speculates on its role in the organism. In addition, calculating the isoelectric point also helps to optimize the charge distribution of proteins, thereby improving their stability and biological activity in various environments. Solvent accessibility: It describes the extent to which an amino acid residue or part of it in a protein molecule is exposed in solution. It is closely related to protein folding, stability, and molecular interactions. Calculating the solvent accessibility of amino acid residues helps to predict the stability and biological function of proteins while revealing the structural features of proteins and the role of amino acid residues in the structure.

In areas such as protein design, drug development, and bioengineering, the use of computable parameters helps to optimize protein function and enhance its applications (V. V. B. Yallapragada et al. 2020). Researchers have used computer simulations, data mining, and bioinformatics methods to calculate and predict these parameters, thereby making connections between protein sequence, structure, and function. This provides an innovative and powerful direction for the modification and optimization of proteins.

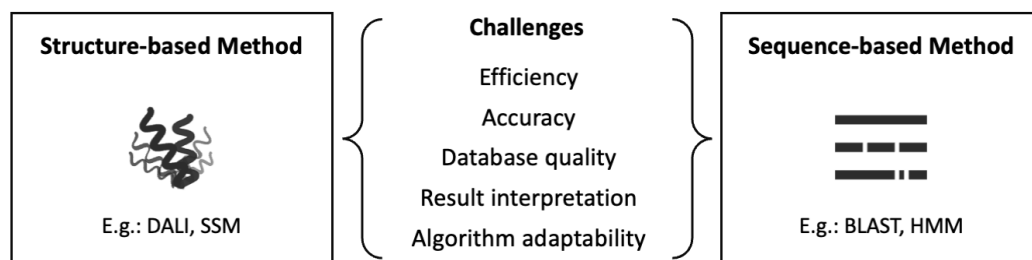


Figure 8.2: Issues and Challenges of Existing Protein Search Methods.

8.2.3 Challenges and Bottlenecks in Searching the Protein Space

Protein research heavily relies on extensive experimental data stored in various databases. The RCSB PDB is a widely-used repository for 3D protein structures determined by experimental methods (Berman et al. 2000). UniProt is another comprehensive resource that integrates protein sequences and functional annotations from several databases, including the manually annotated Swiss-Prot and the automatically annotated TrEMBL (“UniProt: the universal protein knowledgebase in 2021” 2021). Additionally, databases like CATH, SCOP, and Pfam classify and annotate proteins based on their structural, sequence, or functional characteristics (Orengo, Michie, et al. 1997; Murzin et al. 1995; Finn, Coghill, et al. 2016). These databases provide valuable information for understanding protein structure, function, and evolution, making them essential for protein research and design.

Protein searching typically employs two main approaches: structure-based and sequence-based methods. Structure-based methods, such as DALI (Holm and Laakso 2016) and TM-align (Yang Zhang and Skolnick 2005), compare the 3D structures of proteins to identify similarities, which is particularly useful for proteins with diverse sequences but similar structures. In contrast, sequence-based methods, like BLAST (J. Ye, McGinnis, and Madden 2006) and HMMER (Finn, Clements, and Eddy 2011), detect homology by comparing amino acid sequences and excel at identifying proteins with similar sequences.

However, as illustrated in Figure 8.2, these methods have limitations in terms of efficiency, accuracy, and adaptability (Rost 1999; Gligorijević et al. 2021; Schnoes et al. 2009; Pearson 2013). Structure-based methods are computationally intensive and slow, while sequence-based methods may struggle with proteins exhibiting significant

structural differences despite sequence similarity. The non-linear relationship between protein structure and sequence can lead to errors in search results, with structure-based methods potentially failing to identify functional similarities between proteins with similar structures but different sequences, and sequence-based methods failing to accurately detect structural similarities in proteins with similar sequences but different structures. Database quality, including incorrect annotations, redundancy, and incomplete data, can also negatively impact search results. Moreover, interpreting protein search results and comparing them with experimental data can be challenging and error-prone for large datasets. Existing search methods may not adapt well to various protein types and biological problems, particularly when targeting specific proteins or scenarios.

To address these challenges, researchers have developed various rapid protein structure search methods in recent years. For instance, 3D-AF-SURFER (Aderinwale et al. 2022) utilizes deep learning to learn 3D feature representations of protein surfaces, enabling fast and accurate structure alignment and similarity search. GraSR (C. Xia et al. 2022) introduces a graph representation learning approach, improving search efficiency and generalizability by converting protein structures into graphs and learning their embedded representations. DeepFold (J.-W. Lee et al. 2023) employs convolutional neural networks to extract features directly from 3D structures, accelerating the structure alignment process. MADOKA (Tilahun et al. 2022) designs a MapReduce-based distributed computing framework, achieving parallelization and scalability for large-scale protein structure searches. These methods have made significant contributions to accelerating searches and expanding functionality in the field of protein structure search.

Despite the progress made by existing methods, there are still limitations in terms of insufficient utilization of sequence information, limited search flexibility, and interpretability. These limitations highlight the need for innovative approaches that can effectively integrate sequence and structural information, support flexible search queries, and provide interpretable results.

Parametric protein search emerges as a promising alternative that tackles the challenges in protein search from a different perspective. Instead of solely relying on traditional structure or sequence-based methods, parametric protein search introduces a novel paradigm that leverages the power of parametric protein design. By carefully selecting and optimizing a set of key parameters that define protein structure and function, such as amino acid composition, hydrophobicity, and secondary structure propensities,

parametric protein search enables a more comprehensive and fine-grained representation of proteins.

This parametric approach offers several advantages over existing methods. It allows for the seamless integration of sequence and structural information, as the parameters can capture both the global topology and the local physicochemical properties of proteins. Furthermore, it provides a highly flexible and customizable search framework, where users can define their own parameter combinations and search criteria based on their specific research needs. Finally, the parametric representation facilitates the interpretability of search results, as the key parameters contributing to the similarity or difference between proteins can be easily identified and visualized.

8.2.4 PS-GO - Transforming Protein Exploration with a Breakthrough Search Method

PS-GO(Parametric Search, GO!) is an innovative parametric protein search engine that cleverly integrates protein structure, sequence information, and computable parameters. Our search engine aims to address the multiple limitations of existing protein search methods and provide a more powerful, efficient, and flexible tool for protein design research and applications.

A core feature of PS-GO is its unique search capability, which enables users to specify specific parameters and constraints to search for protein sequences and structures that meet their needs through a parametric protein design approach. This approach improves the ability to control protein structure and function with precision by calculating and tuning key parameters in the protein molecule. Specifically, PS-GO can handle a variety of complex protein sequences and structure parameters such as RC.Score, hydrophobicity, instability, size, isoelectric point, and solvent accessibility to meet different research needs. This precise search capability not only improves the accuracy of the search results but also contributes to a deeper understanding of the structure-function relationships of proteins, which is of great academic value for research in the field of protein design and engineering.

As shown in Figure 8.3, PS-GO (<https://psgo.ucc.ie/>) utilizes parallel computing and rapid algorithms and processes a vast array of protein structure parameters swiftly. Even when dealing with extensive protein data, PS-GO returns results that cater to users' needs

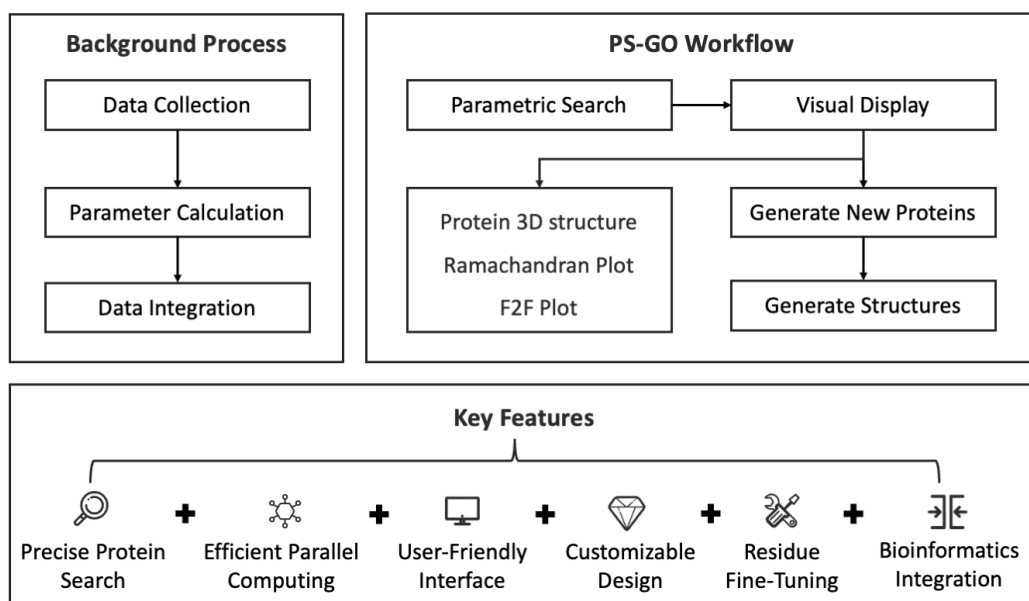


Figure 8.3: How PS-GO search works and its key features.

promptly. Significantly more efficient than traditional structure or sequence-based search methods, PS-GO allows researchers to acquire necessary protein structure information in less time, accelerating the experimental procedure.

Another highlight of PS-GO is the enhanced user experience. Its intuitive graphical interface, coupled with natural language processing technology, allows effortless input and interaction, eliminating the requirement of specialized biological or computer technology knowledge. Therefore, PS-GO's usability extends beyond researchers, catering to a broader user base including students, teachers, biologists, and drug researchers. This capacity helps proliferate protein design knowledge and techniques, extending their applicability.

PS-GO also provides customization of parameters and optimization conditions according to user needs, facilitating realistic protein design. Its high adaptability and customization allow its utility across diverse research requirements and application scenarios, offering researchers more innovative avenues and fostering the advancement and application of protein design technologies.

In conjunction with the Silver Surfer service, PS-GO enables users to freeze certain protein sequence residues while adjusting others to generate novel protein sequences.

This feature empowers researchers to probe deeper into protein sequence space, enhancing the diversity and adaptability of protein structures. This holds immense academic merit in investigating the interplay between protein sequence and structure, optimizing proteins' functional attributes, and creating opportunities for protein modification and novel drug development.

The integration of PS-GO with Protein Fragment And Structure Analysis (PROFASA) interlinks it with other bioinformatics tools and data resources, creating a comprehensive ecosystem for protein design and research (Mi et al. 2023a). Through PROFASA, PS-GO integrates with protein databases (like PDB, UniProt), sequence alignment tools (such as BLAST), and structure prediction software, establishing a one-stop platform for protein design and research, bolstering research and applications in related domains.

8.2.5 Parametric Search Strategy

PS-GO employs a novel parametric search strategy to efficiently explore the vast protein sequence and structure space. The key steps of this strategy are outlined below:

Data Collection and Processing: PS-GO collects protein data from various public databases, such as PDB and UniProt. The collected data is carefully curated and filtered based on sequence and structure quality, redundancy, and completeness. A series of preprocessing steps, including sequence alignment, structure validation, and data normalization, are performed to ensure the consistency and reliability of the data.

Parameter Calculation: For each protein in the database, PS-GO calculates a set of key parameters that characterize its physicochemical properties and structural features. These parameters include amino acid composition, hydrophobicity, instability index, molecular weight, isoelectric point, secondary structure content, solvent accessibility, and more. The calculation of these parameters is based on well-established algorithms and empirical formulas, such as the Kyte-Doolittle hydrophobicity scale (Kapcha and Rossky 2014) and the Chou-Fasman secondary structure prediction method (T. A. Kumar 2013).

Data Integration: The calculated parameters, along with the sequence and structure information, are integrated into a unified database. This database serves as the foundation for the subsequent parametric search and protein design tasks. The integration process

involves data mapping, indexing, and optimization to facilitate efficient data retrieval and analysis.

Parametric Search: Given a user-specified query, PS-GO performs a parametric search against the integrated database. The query can be a combination of various parameters, such as sequence similarity, structural motifs, and physicochemical properties. PS-GO employs advanced indexing and hashing techniques to quickly identify proteins that match the query criteria. The search results are ranked based on their relevance and similarity to the query, providing users with a prioritized list of candidate proteins.

Protein Design and Optimization: In addition to searching existing proteins, PS-GO also supports the design of novel proteins with desired properties. Users can specify the target parameters, such as stability, solubility, and function, and PS-GO will generate a set of candidate sequences that satisfy these requirements. The design process involves a combination of data-driven modeling, molecular simulations, and optimization algorithms. PS-GO iteratively refines the candidate sequences based on their predicted structures and functions, ultimately delivering a set of optimized protein designs for experimental validation.

By integrating these key steps, PS-GO provides a powerful and flexible platform for parametric protein search and design. The modular architecture of PS-GO allows for easy extension and customization, enabling researchers to adapt the system to their specific needs and incorporate new algorithms and data sources.

8.3 Methods

8.3.1 Parametric Search Algorithm

The core of PS-GO is an innovative parametric search algorithm that enables efficient and flexible retrieval of proteins based on user-specified multi-dimensional parameter conditions. The algorithm consists of the following key steps:

Parameter Extraction: Given a protein sequence or structure, PS-GO first extracts a set of parameters that characterize the physicochemical properties and structural features of the protein. These parameters include:

(1) Amino acid composition: The frequencies of 20 amino acids in the protein sequence are calculated using the ProtParam module of BioPython (Cock et al. 2009), resulting in a 20-dimensional composition vector.

(2) Isoelectric point: The `isoelectric_point()` function in the ProtParam module is used to calculate the isoelectric point of the protein based on the pKa values of amino acids.

(3) Molecular weight: The `molecular_weight()` function in the ProtParam module is used to calculate the molecular weight of the protein based on the molecular mass of amino acid residues.

(4) Hydrophobicity: We implemented the Kyte-Doolittle hydrophobicity scale algorithm to calculate the hydrophobicity score of each amino acid residue using a sliding window approach. The average score is then taken as the hydrophobicity index of the entire protein. Specifically, the following formula is used:

$$H = \frac{1}{n} \sum_{i=1}^n h_i \quad (8.1)$$

where H is the hydrophobicity index of the protein, n is the number of amino acid residues, and h_i is the hydrophobicity score of the i -th residue (obtained from the Kyte-Doolittle scale table).

(5) Instability index: The `instability_index()` function in the ProtParam module is used to calculate the instability index of the protein based on the dipeptide frequency and the Dipeptide instability weight values (DIWV) instability weight values (Guruprasad, Reddy, and Pandit 1990).

(6) Secondary structure: The DSSP program is used to predict the secondary structure (like α -helix, β -sheet, coil) of the protein, and the proportions of each secondary structure type are calculated (Zacharias and Knapp 2014). DSSP uses hydrogen-bonding patterns and geometrical criteria to reliably identify secondary structure elements.

(7) Solvent accessibility: The NACCESS program is used to calculate the relative solvent accessibility (RSA) of residues on the protein surface (S. J. Hubbard 1993). NACCESS employs the Lee-Richards algorithm to estimate the accessibility of residues by rolling a probe over the protein surface. We calculate the RSA value for each residue and then take the average as the solvent accessibility index of the entire protein.

The above parameter extraction process is implemented by combining existing bioinformatics tools (like DSSP, NACCESS) and self-developed algorithms (like Kyte-Doolittle

hydrophobicity calculation) to ensure the efficiency and accuracy of the calculation. The algorithms are implemented in Go, and the bioinformatics tools are invoked through system calls.

To speed up the parameter extraction process, we leverage Go's goroutine mechanism to perform concurrent calculations. Each protein is assigned to a goroutine, which independently carries out the parameter extraction tasks. Within each goroutine, we invoke Python scripts or custom Go scripts to compute specific parameters. This parallel computing approach greatly improves the efficiency of parameter extraction, especially when dealing with large-scale protein datasets.

Natural Language Search: In addition to range-based parameter search, PS-GO also supports natural language queries. Users can input natural language sentences describing the desired protein properties or functions, such as "Find proteins with high stability and low molecular weight". We utilize the GPT-4 model from OpenAI to convert the natural language query into structured parameter conditions through function calls (Achiam et al. 2023).

Specifically, we define a function named "parse_query" that takes a natural language query as input and outputs the parameter conditions in JSON format. For example, given the above query, GPT-4 will return the following JSON:

```
{
  "stability": "high",
  "molecular_weight": "low"
}
```

We then map these qualitative descriptions to specific numeric ranges. For instance, "high stability" corresponds to an instability index less than 40, and "low molecular weight" corresponds to a molecular weight less than 50 kDa. Finally, the converted parameter conditions are used for the subsequent search process.

Indexing and Hashing: To accelerate the retrieval process, PS-GO builds inverted indexes for the extracted parameter values. Each parameter value is mapped to a unique hash code, and the proteins are organized into corresponding hash buckets based on their parameter features. This indexing scheme allows for constant-time lookup of proteins that match a given parameter query condition.

8 PS-GO Parametric Protein Search Engine

Specifically, we adopt a multi-dimensional hashing approach that converts each parameter value into a 64-bit hash code. For numeric parameters (like molecular weight), we use a double hashing function to map the value to a hash bucket:

$$h_1(x) = (ax + b) \% p \quad (8.2)$$

$$h_2(x) = 1 + ((x + c) \% (p - 1)) \quad (8.3)$$

where x is the parameter value, p is a prime number (chosen as the smallest prime greater than the range of the parameter), and a, b, c are randomly selected constants. The double hashing function effectively reduces hash collisions and improves retrieval efficiency.

For categorical parameters (like secondary structure), we use the MurmurHash3 algorithm to map the category labels to hash codes (Dahlggaard, Knudsen, and Thorup 2017). MurmurHash3 is a non-cryptographic hash function known for its fast speed and good distribution properties.

During retrieval, the system calculates the hash codes corresponding to the user query and then performs matching in the corresponding hash buckets to filter out the proteins that satisfy the conditions.

Query Processing: When a user submits a query containing multiple parameter conditions, PS-GO first parses the query and converts it into a set of hash codes. Then, the algorithm retrieves proteins from the corresponding hash buckets and filters them to find the results that satisfy all the conditions.

We designed a set operation-based query processing method that treats multiple parameter conditions as sets and uses set operations (like intersection, union, difference) to combine the conditions. For example, given the query "(stability >0.8) AND (size <2000)", we first find the sets of proteins satisfying "stability >0.8" and "size <2000", denoted as S_1 and S_2 , respectively. We then calculate the intersection of S_1 and S_2 to obtain the final result.

This set-based query processing method fully leverages the inverted indexes and avoids linear scans over the entire database, thereby significantly improving query efficiency.

8.3.2 Protein Database Construction

The PS-GO database is constructed by integrating two major public data sources: PDB and UniProt.

Data Integration: We first download data files from the FTP servers of PDB and UniProt, respectively. For PDB, we download the mmCIF format structure files and FASTA format sequence files. For UniProt, we download the XML format annotation files and FASTA format sequence files.

We then use the BioPython library in Python to parse these files and extract the fields of interest, such as PDB ID, UniProt ID, amino acid sequence, secondary structure, and functional annotations. These fields are mapped to a relational database schema, creating tables for proteins, residues, atoms, and annotations.

During the integration process, we need to handle cross-references between PDB and UniProt to establish a mapping between the two databases. We utilize the "PDB cross-reference" field in UniProt, which provides the correspondence between UniProt entries and PDB structures. Through these cross-references, we can associate structural information with sequence and functional annotations.

We developed an automated data integration pipeline using the Luigi workflow framework to manage tasks such as data downloading, parsing, cleaning, and mapping. The pipeline runs periodically to ensure that the database stays in sync with PDB and UniProt.

Data Storage: PS-GO uses a MySQL relational database management system to store and manage the protein data. MySQL is a widely-used open-source database known for its high performance, scalability, and ease of use.

We designed a normalized database schema that includes the following main tables:

(1) Protein table: stores basic information about proteins, such as PDB ID, UniProt ID, sequence length, and species.

(2) Residue table: stores information about amino acid residues in proteins, such as residue number, residue name, and secondary structure type.

(3) Atom table: stores the atomic coordinate information of proteins, such as atom number, atom name, and x/y/z coordinates.

(4) Annotation table: stores functional annotation information for proteins, such as Gene Ontology (GO) terms, Enzyme Commission (EC) numbers, and Pfam domains.

8 PS-GO Parametric Protein Search Engine

(5) Parameter table: stores various parameters of proteins, such as molecular weight, isoelectric point, hydrophobicity, and instability index.

We utilize MySQL's indexing mechanisms (like B+ tree index, hash index) to speed up data retrieval. For frequently queried fields (like PDB ID, UniProt ID), we create single-column indexes. For fields that are often queried together (like residue number and atom name), we create composite indexes.

Additionally, we developed stored procedures and user-defined functions (UDFs) to perform certain computation tasks on the database server side, reducing the burden on the application server.

Data Access: PS-GO uses the database/sql package in Go to communicate with the MySQL database (Yellavula 2017). The database/sql package provides a generic interface for interacting with SQL databases, supporting multiple database drivers.

We define Go structs that correspond to the tables in the database. These structs, also known as "models", contain fields that map to the columns of the tables. We use the GORM library, which is an object-relational mapping (ORM) library for Go, to simplify database operations (Yellavula 2017). GORM allows us to define database models using Go structs and provides high-level APIs for querying and manipulating data.

For example, the following code defines the Protein and Residue models:

```
type Protein struct {
    ID          uint          `gorm:"primaryKey"`
    PdbID       string        `gorm:"index"`
    UniprotID   string        `gorm:"index"`
    Sequence    string
    Residues    []Residue
}

type Residue struct {
    ID          uint          `gorm:"primaryKey"`
    ProteinID   uint
    Number      int
    Name        string
    SSType      string
}
```

```
}
```

With these model definitions, we can easily perform database operations using GORM's methods, such as Create, Find, Update, and Delete. For instance, the following code demonstrates how to query a protein by its PDB ID and retrieve its residue information:

```
var protein Protein
db.Preload("Residues").First(&protein, "pdb_id = ?", "1ake")

for _, residue := range protein.Residues {
    fmt.Println(residue.Number, residue.Name, residue.SSType)
}
```

By leveraging GORM, we can express complex SQL queries in a more intuitive and object-oriented way, greatly improving development efficiency.

8.3.3 Silver Surfer

The integration with Silver Surfer service takes PS-GO's capabilities a step further. Should the initial search not yield any closely matching sequences, Silver Surfer enables the creation of novel protein sequences. Silver Surfer is a novel tool that applies a genetic algorithm for protein sequence generation with a fitness function derived from the query. The fitness function plays a crucial role in guiding the evolutionary process towards sequences that satisfy the user-specified target properties.

In Silver Surfer, the fitness function is designed to measure the similarity between the generated sequences and the target protein properties specified by the user. The fitness function calculates a weighted sum of the mean absolute errors (MAEs) between the predicted properties of the generated sequence and the user-defined target values. The MAE for each property is computed as follows:

$$\text{MAE}_i = \frac{1}{n} \sum_{j=1}^n |p_{ij} - t_{ij}| \quad (8.4)$$

8 PS-GO Parametric Protein Search Engine

where i is the index of the property, n is the number of residues in the sequence, p_{ij} is the predicted value of the i -th property for the j -th residue, and t_{ij} is the user-defined target value for the i -th property at the j -th residue.

The properties considered in the fitness function include various physicochemical and structural parameters, such as stability, hydrophobicity, molecular weight, and secondary structure composition. For each property, the user specifies a desired target value or range at each residue position. The fitness function then evaluates how well the generated sequence matches these target values by calculating the MAE for each property.

To account for the relative importance of different properties, the fitness function assigns a weight to each property based on user preferences or predefined settings. The weighted sum of MAEs is calculated as follows:

$$\text{Fitness} = \sum_{i=1}^m w_i \cdot \text{MAE}_i \quad (8.5)$$

where m is the number of properties, w_i is the weight assigned to the i -th property, and MAE_i is the mean absolute error for the i -th property.

The weights allow users to prioritize certain properties over others. For example, if stability is more important than hydrophobicity for a specific protein design task, the user can assign a higher weight to the stability property. The fitness function will then place more emphasis on minimizing the MAE for stability compared to other properties.

In addition to the weighted sum of MAEs, the fitness function also incorporates a penalty term for sequences that violate the "frozen" regions specified by the user. The frozen regions are specific portions of the sequence that should remain unchanged during the optimization process. If a generated sequence modifies any residue in the frozen regions, a large penalty value is added to the fitness score, effectively discouraging the selection of such sequences.

The final fitness score is calculated as follows:

$$\text{Fitness} = \sum_{i=1}^m w_i \cdot \text{MAE}_i + \lambda \cdot \text{FrozenPenalty} \quad (8.6)$$

where λ is a hyperparameter that controls the strength of the penalty, and FrozenPenalty is the number of residues in the frozen regions that have been modified.

A lower fitness score indicates a better fit between the generated sequence and the target properties. The genetic algorithm in Silver Surfer aims to minimize the fitness score by selecting and mutating sequences with lower scores.

During the evolutionary process, the genetic algorithm maintains a population of candidate sequences. In each generation, the sequences with the lowest fitness scores are selected for reproduction and mutation. The mutation operator introduces random changes to the sequences while respecting the frozen regions. The modified sequences are then evaluated using the fitness function, and the process continues for a specified number of generations or until a satisfactory solution is found.

By incorporating a comprehensive fitness function that considers multiple properties, assigns weights based on user preferences, and penalizes violations of frozen regions, Silver Surfer enables the generation of novel protein sequences that closely match the user-defined target properties. This approach allows users to fine-tune the desired characteristics of the generated sequences and explore a vast sequence space beyond the existing proteins in the database.

The genetic algorithm employed by Silver Surfer, as shown in Table 8.2, leverages the fitness function to guide the search process efficiently. The iterative process of selection, mutation, and evaluation allows Silver Surfer to continuously improve the generated sequences until a satisfactory solution is found or a predefined stopping criterion is met.

This cycle repeats until a stopping condition, such as reaching a maximum number of generations or achieving a fitness score below a specified threshold, is met. The use of the weighted MAE-based fitness function with frozen region constraints allows Silver Surfer to efficiently explore the vast sequence space and generate novel proteins that closely match the user-defined target properties while preserving critical regions of the sequence.

8.4 Results

8.4.1 Key Functionalities and User Interface of PS-GO

PS-GO offers a comprehensive suite of functionalities for protein search, analysis, and visualization, all accessible through a user-friendly web interface. The primary feature of PS-GO is its parametric search capability, which allows users to search for proteins

Table 8.2: The genetic algorithm employed by Silver Surfer

Initial Population: The process begins with a set of protein sequences, which can be randomly generated or seeded from existing sequences. Each sequence is represented as a string over a finite alphabet, in the case of protein sequences, this consists of 20 amino acids. All sequences constitute a continuous individual pool. Fitness Function: A fitness function is used to evaluate the quality of each sequence. This function uses a weighted sum of MAEs and a penalty for frozen region violations to measure how well the protein sequence performs. Based on the results of the fitness function evaluation, each sequence is assigned a fitness score. Selection and Mutation: In each iteration, the sequences with the lowest fitness scores are selected from the individual pool for reproduction and mutation. The selected sequences undergo mutation operations, generating a series of new sequences. During the mutation process, certain parts can be set as "frozen", remaining unchanged. Sequence Evaluation: The newly generated sequences are evaluated using the fitness function, and their fitness scores are calculated. Population Update: The new sequences are added to the population, replacing the sequences with higher fitness scores. The size of the population remains constant throughout the iterations. Iterative Process: The process continues to iterate, selecting the sequences with the lowest fitness scores for mutation and updating the population with the newly generated sequences. This process goes on until a sequence meeting the specified criteria is found or a predefined number of iterations is reached.
--

based on a combination of sequence, structure, and physicochemical properties. Users can specify search criteria using a flexible parameter selection tool, which includes options for amino acid composition, hydrophobicity, secondary structure propensities, and various geometric features.

In addition to parametric search, PS-GO also supports natural language search, enabling users to query proteins using plain text descriptions of desired properties or functions. The natural language search engine employs advanced NLP techniques to interpret user queries and retrieve relevant protein entries from the database.

Figure 8.4 showcases PS-GO's user-friendly search interface, which allows users to easily input their search criteria and view the resulting protein information.

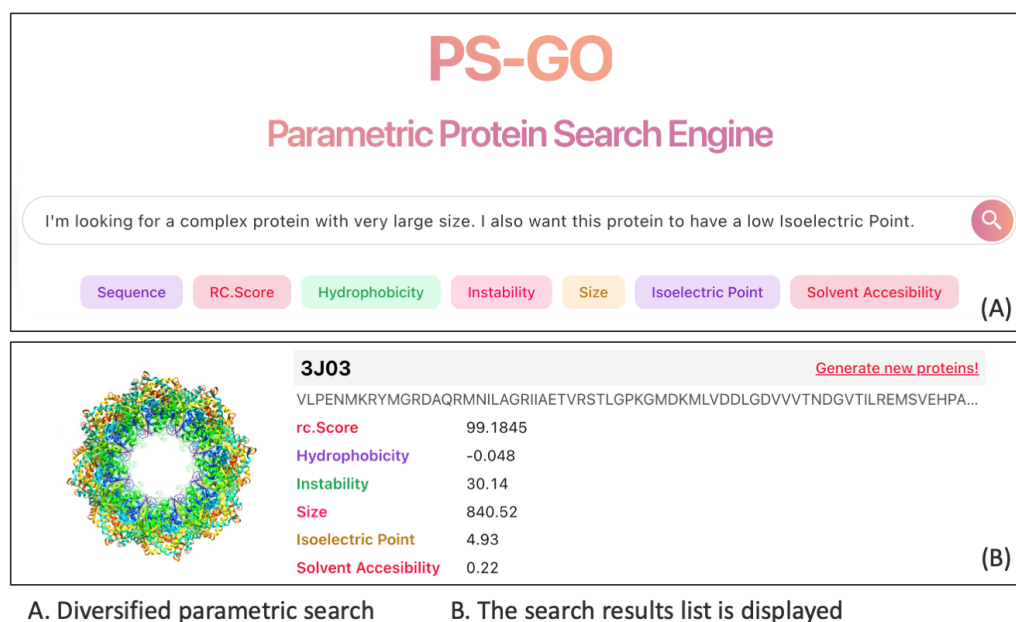


Figure 8.4: PS-GO protein search engine. A. Parametric search page, select the corresponding parameter and give the filter range to search. B. Search result list, showing the model picture, sequence, and parameter of the protein.

With PS-GO, users can visualize protein 3D structures directly from the search results page. The structure viewer provides various representation options, such as cartoon, surface, and ball-and-stick models. Users can interact with the displayed structure by rotating, zooming, and panning to examine it from different angles.

Furthermore, PS-GO integrates the PROFASA (Protein Fragment and Structure Analysis) tool for protein structure visualization and analysis. PROFASA is a web-based platform that offers a range of functionalities for exploring and understanding protein structures (Mi et al. 2023a).

PROFASA's analysis tools, accessible through PS-GO, enable users to investigate protein structural features in detail. These tools include secondary structure analysis, structural alignment, binding site prediction, and structural superposition.

Figure 8.5 illustrates an example of protein structure visualization and analysis using PS-GO and PROFASA.

PS-GO efficiently calculates a range of protein parameters, including RC.Score, hydrophobicity, instability, molecular size, isoelectric point, and solvent accessibility. These

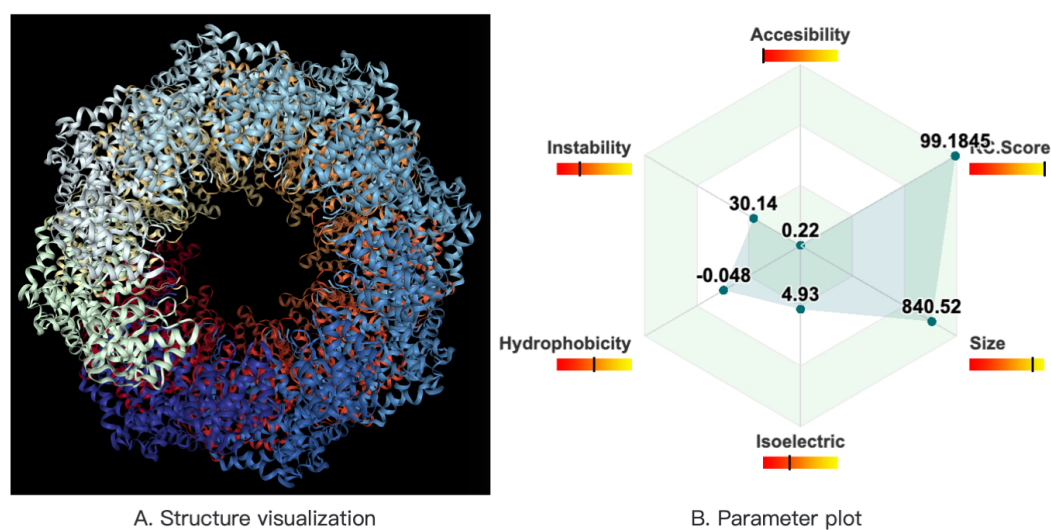


Figure 8.5: Structural visualization of protein models and graphical representation of parameters using PS-GO and PROFASA.

parameters offer valuable insights into the physicochemical properties and potential functions of proteins.

The calculated parameters are presented in an intuitive and visually appealing manner within PS-GO's interface (Figure 8.4). Users can easily view and analyze the parameter values, facilitating the interpretation of protein characteristics and the identification of proteins with desired properties.

8.4.2 Advantages and Potential Impact of PS-GO

PS-GO introduces a novel protein search framework based on parametric protein design principles. By integrating sequence, structure, and physicochemical property information, PS-GO provides users with a unique perspective to search and explore proteins. Different from existing methods that rely on sequence or structure alignment, PS-GO's parametric search approach allows users to find proteins that satisfy specific criteria based on a combination of various properties. This innovative search paradigm complements traditional methods by enabling the discovery of proteins that might be difficult to identify using sequence or structure similarity alone.

One of the major strengths of PS-GO is its flexibility and customizability. The parametric search engine allows users to define their own search criteria and parameter

combinations, tailoring the search process to specific research questions or application scenarios. This flexibility empowers researchers to explore protein relationships from different angles and to uncover novel insights that might be missed by fixed, predefined search methods.

Another key advantage of PS-GO is its interpretability. By providing detailed parameter profiles and visualizations, PS-GO helps users understand the key factors contributing to protein similarity and function. This interpretability is particularly valuable for tasks such as protein function prediction, rational protein design, and evolutionary analysis, where understanding the underlying mechanisms is crucial.

Table 8.3 provides a summary comparison of the features and characteristics of PS-GO and other representative protein search methods.

Table 8.3: Comparison of features and characteristics between PS-GO and other protein search methods

Method	Search Mechanism	Protein Representation	Flexibility	Interpretability
PS-GO	Parametric protein design-based	Integration of sequence, structure, and physicochemical properties	Allows user-defined parameter combinations and search criteria	Provides insights into key parameters contributing to similarity or difference
Dali (Holm and Laakso 2016)	Structure-based	3D coordinates of protein backbone	Limited to structural similarity	Focuses on global structural alignment
Foldseek (Kempen et al. 2022)	Structure-based	3D coordinates of protein backbone	Allows different distance measures and alignment modes	Provides local alignment information
PSI-BLAST (Bhagwat and Aravind 2008)	Sequence-based	Amino acid sequence	Allows iterative search and position-specific scoring matrices	Focuses on sequence similarity

The modular architecture of PS-GO also makes it a valuable platform for integrating new protein tools. As new methods for protein representation and comparison are devel-

oped, they can be easily incorporated into the PS-GO framework, allowing researchers to complement their work.

8.4.3 Performance Evaluation

To assess the performance of PS-GO in handling various tasks, we conducted benchmark tests using a script that iteratively executes parametric searches, natural language searches, and new sequence generation functions for 100 times. The benchmark script is available at github (<https://github.com/Atobelin/psgo-benchmarking>). In each iteration, we randomly generated or selected different parameter values, query statements, and target conditions to simulate the diversity of real-world usage scenarios. The test results are presented in Table 8.4.

Table 8.4: The average performance of PS-GO in handling different tasks

Task	Average Operation Time (s)
Parametric Search	0.84
Natural Language Search	1.62
New Sequence Generation	5.10
Parameter Calculation	6.4s per 50 items

For the parametric search task, PS-GO achieves an average response time of 0.84 seconds. This demonstrates that PS-GO is capable of processing complex multi-parameter queries within sub-second latency, exhibiting exceptional retrieval efficiency. Even under dynamic variations of parameter values, PS-GO maintains stable and rapid performance, which can be attributed to our optimized indexing structure and query algorithms.

In the natural language search task, PS-GO yields an average response time of 1.62 seconds. Although natural language queries involve more semantic understanding and matching computations, PS-GO still manages to return precise results within 2 seconds. This is primarily due to the advanced natural language processing techniques we employ, such as deep learning-based semantic representations and intelligent query expansion, enabling PS-GO to efficiently handle natural language queries.

Regarding the new sequence generation task, we conducted 50 iterations of testing, resulting in an average operation time of 5.10 seconds. Generating new sequences is a computationally intensive task that requires searching and optimizing within a vast sequence space, thus taking relatively longer time. Nevertheless, PS-GO completes

the generation of new sequences within a short period, owing to the heuristic search strategies and parallel computing techniques we adopt, which significantly accelerate the sequence generation process.

We did not include benchmark tests for the parameter calculation task. This is because parameter calculation is a one-time operation performed during database construction, mainly involving feature extraction and statistical analysis on large-scale protein data. Although computationally intensive, this process can be efficiently completed in the background through preprocessing and caching optimizations. Once the parameter calculation is finished, subsequent search and analysis tasks can directly utilize the pre-computed results without repetitive calculations. Therefore, assessing the performance of parameter calculation is less relevant to evaluating PS-GO's real-time responsiveness and user experience.

Overall, the benchmark test results demonstrate that PS-GO exhibits outstanding performance and stability in key tasks such as parametric search, natural language search, and new sequence generation. Even under dynamically changing query conditions, PS-GO consistently delivers efficient service, providing researchers with a fast, reliable, and user-friendly platform for protein analysis, which has the potential to greatly enhance the efficiency and depth of protein research.

8.5 Discussion and Outlook

Parametric protein design has emerged as a powerful technique for precisely manipulating protein structure and function by computationally adjusting key parameters in protein molecules (Korendovych and DeGrado 2020). This approach has gained significant traction across multiple fields, including drug discovery, biotechnology, and synthetic biology, owing to rapid advancements in computer science and experimental techniques (W. Wei et al. 2020). However, the full potential of parametric protein design is often hindered by the limitations of traditional protein structure databases, which face challenges in search efficiency and computational complexity.

One major hurdle is the time-consuming data processing and significant computational power required to search through vast protein databases (Rost 1999). Additionally, identifying proteins with substantial sequence differences using sequence-based methods or discerning the functional similarity of proteins with varying sequences using structure-

based methods can be complex and computationally intensive (Gligorijević et al. 2021). These challenges underscore the need for an enhanced, accurate, and adaptable search strategy to complement parametric protein design efforts.

Parametric protein search engines, such as PS-GO, offer a promising solution to these challenges by incorporating advanced algorithms, optimizing computational resources, and improving protein feature understanding. These search strategies aim to efficiently navigate the vast protein sequence and structure space, enabling researchers to quickly identify promising candidates for further optimization through parametric protein search.

By integrating parametric protein search with parametric protein design, researchers can unlock the untapped potential of these technologies and fuel research progress in related fields. The development of efficient protein search engines not only enhances the speed and efficiency of protein design but also opens up new avenues for groundbreaking discoveries in bioinformatics and computational biology.

In essence, the symbiotic relationship between parametric protein design and parametric protein search holds the key to overcoming the limitations of traditional protein structure databases and accelerating innovation in protein engineering. As these technologies continue to evolve and integrate, they promise to revolutionize our understanding of proteins and their applications in various domains.

PS-GO, as a search method based on parametric protein design, has significant advantages over traditional protein structure databases in the following areas:

- **Highly accurate search capabilities:** PS-GO allows users to specify specific parameters and constraints to meet different research needs through a parametric protein design approach. This precise search capability improves control over protein structure and function and is of great academic value to the field of protein design.
- **Optimized computing efficiency:** Using parallel computing and fast algorithms, PS-GO is able to process a large number of protein structure parameters in a short time. Compared to traditional search methods, PS-GO's computational efficiency is significantly improved, speeding up the experimental process.
- **User-friendly interface and customization:** PS-GO provides an intuitive graphical interface and natural language processing technology for easy user input and

Application Area	Outlook
Pharmaceuticals	 Deep learning technology  Cloud computing & distributed computing  Integration of bioinformatics tools & data resources
Protein Interaction Studies	
Bioengineering	
Biomaterials	

Figure 8.6: PS-GO's future development will focus on refining algorithms, expanding the database, and using advanced computing technologies to optimize protein design performance.

interaction. At the same time, it allows users to customize parameters and optimization conditions according to their needs, adapting to various research requirements and application scenarios.

- **Integration with other bioinformatics tools:** Combined with the Silver Surfer service, PS-GO supports users to freeze partial residues of protein sequences and integrates with PROFASA to form a complete protein design and research ecosystem with other bioinformatics tools and data resources, providing a one-stop protein design and research platform.

Despite its many significant advantages, PS-GO has a number of limitations in practical application. PS-GO has a relatively high demand for computational resources, which requires a certain level of computational power from the user. In particular, more accurate algorithm design and powerful computational support are required when dealing with complex protein structure parameters. The performance and accuracy of PS-GO are limited by the quality and quantity of the protein database. If specific types of proteins are missing from the database, or if there is incorrect or inaccurate information, the performance and accuracy of PS-GO may be compromised. The number of parameters that PS-GO can currently handle is still limited, which means that continued research and development of new algorithms is required to broaden the range of computable parameters.

As shown in Figure 8.6, in order to continuously improve the efficiency and accuracy of PS-GO, the future development direction can include the following aspects:

8 *PS-GO Parametric Protein Search Engine*

In order to improve the search accuracy of PS-GO, algorithms can be optimised for specific types of parameters in the future. This includes the introduction of more advanced machine learning and artificial intelligence techniques, such as deep learning, as a means of automatically learning and recognising complex patterns in proteins to improve the effectiveness and accuracy of parametric design. Finer parametric design methods can help users better control the structure and function of proteins to meet different research needs.

Meanwhile, PS-GO's database can be expanded by exchanging data with other bioinformatics databases or automatically extracting protein data from literature and the web through data mining and machine learning methods. This expansion can provide richer and more accurate protein search results. The expansion and management of the database can be done by applying advanced database management systems and data mining techniques, which can further improve the efficiency and accuracy of the search.

Since PS-GO needs to deal with a large amount of protein data and complex computational tasks, the combination of cloud computing and distributed computing technologies will greatly enhance its computational capability. Specifically, cloud computing provides elastic computing resources that can be dynamically adjusted to meet different computational needs. Distributed computing, on the other hand, can break down large-scale data processing tasks to multiple computers, greatly improving the efficiency and stability of computation.

In addition, PS-GO plans to initiate deeper collaborations with more computational biology tools to enhance its search capabilities and broaden the scope of its services. Among them, collaboration with protein-related tools will be especially critical. For example, with the help of more protein structure prediction tools, PS-GO can design new protein sequences while at the same time predicting and displaying the many possible 3D structures of the protein. Through integration with protein function annotation tools, PS-GO can further provide information on the possible biological functions of proteins and their roles in biological processes. Integration with protein interaction network analysis tools will help PS-GO to demonstrate the interaction of proteins with other molecules in the search results, helping researchers to understand the behavior and function of proteins in cells. Such cooperation and integration will not only provide users with more comprehensive information, but also improve the search accuracy of PS-GO.

The PS-GO parametric protein search engine, while not directly engaging in drug development, protein interaction studies, or bioengineering, serves as a pivotal tool in expediting and refining protein research across these disciplines. It significantly impacts the pharmaceutical field by enhancing the efficiency of identifying proteins for early-stage drug discovery, even though it doesn't discover drug molecules itself. In the bioengineering sector, PS-GO proves instrumental by enabling researchers to find proteins with specific functions and properties relevant to their work, albeit not designing proteins or directly supporting synthetic biology. For bioinformatics and structural biology, PS-GO boosts research progress in areas like protein folding, protein dynamics, and disease-related mutations, not through direct contributions to protein structure prediction or functional annotation, but by facilitating the rapid location of proteins of interest. Hence, PS-GO, with its ability to swiftly locate relevant proteins based on specified parameters, indirectly strengthens protein research across various domains.

In conclusion, PS-GO, as an innovative parametric protein search engine, is expected to play a significant role in various fields. In the biomedical field, PS-GO can accelerate the process of drug screening and design, helping researchers quickly find protein molecules with specific functions, thereby shortening the drug development cycle and reducing research and development costs. In the field of synthetic biology, PS-GO can assist in designing entirely new artificial protein molecules, creating novel functions and features that do not exist in nature, thus promoting the development of synthetic biology. In proteomics and structural biology research, PS-GO can improve the efficiency of protein database retrieval and homology analysis, deepening the understanding of the relationship between protein structure and function. Through continuous optimization and upgrading, PS-GO is expected to become an important research tool in the life sciences.

9 Generative Protein Sequence Design via Parametric Optimization: The Silver Surfer Approach

This chapter was published as: Stefan-Bogdan Marcu, **Yanlin Mi**, Venkata VB Yallapragada and Sabin Tabirca. (2024). *Silver Surfer: Navigating the Parametric Protein Space with Genetic Algorithms*. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 479-490).

DOI:10.1007/978-981-97-4677-4_39

Publisher: Springer Nature

I made the contribution to this work through data curation, resource collection, and writing the final draft.

9.1 Abstract

Designing and synthesising proteins with specific physicochemical properties pose significant challenges in biotechnology, environmental sciences, and pharmaceuticals. Recent advancements in machine learning have opened up possibilities but have often been constrained by their focus on limited facets of the multifaceted nature of proteins. This paper presents the Silver Surfer, a new platform utilising a customised genetic algorithm designed to explore the complex space of protein sequences. The Silver Surfer implementation offers a powerful, easily extensible tool to meet diverse scientific needs, operating based on four general-use protein properties: Instability Index, Monoisotopic Mass, Grand Average Hydrophobicity Score, and Isoelectric Point. By harnessing the

generative capabilities of genetic algorithms, the Silver Surfer project opens new horizons in protein design.

9.2 Introduction

Proteins are long chains of amino acids with roles in virtually all of life's processes. These polymers can be considered the building blocks of life, each performing with extreme efficiency, economy and precision in operation one or a few specifically encoded roles that, in their ensemble, define the individual organism (W. Ding, Nakai, and H. Gong 2022). Proteins play various roles in all of the major functions of a living organism, such as movement (myosin and dyneins), structure and support (keratin), digestion (Amylase and Pepsin), respiration (haemoglobin), etc. In order for proteins to fulfil their role, they need to get in direct contact with other proteins, molecules or structures (X. Du et al. 2016). From the perspective of this interaction, the protein can be split into two functional areas: the direct interaction site and the supporting framework. Membership in one of the groups is mainly dictated by the 3D structure that the protein arranges itself through the folding process.

9.2.1 Protein Engineering

Proteins manufactured in, extracted from, or semi-synthesised from biological sources represent eight of the top 10 best-selling drugs in 2018 (Narayanan et al. 2021) due to their high specificity and affinity combined with the lower toxicity and fewer side effects when compared to small molecules (Carter and Lazar 2018; Kesik-Brodacka 2018). This process is based on finding functional sequences in nature and manipulating them in order for the protein to express the desired properties. This technique has multiple drawbacks, such as the laborious and expensive process of finding proteins in nature and testing them for activity in the desired function. Another drawback is the dependence on naturally evolved proteins. Evolution is a slow process that has explored only a fraction of the possible proteins (W. Ding, Nakai, and H. Gong 2022; Vila 2020) and may be lagging in our fast-evolving and rapidly changing environment (W. Ding, Nakai, and H. Gong 2022). The objective of protein engineering is to alter or generate amino acid sequences such that the resulting protein has specific desired properties or functionality of value

for pharmaceutical, medical or technological applications (K. K. Yang, Zachary Wu, and Arnold 2019). The field of protein engineering focuses on improving the process of developing valuable proteins with specific desired physiochemical properties.

The generation of proteins with targeted functionality has the potential of revolutionising fields such as biotechnology, environmental sciences and pharmaceuticals through the creation of proteins with specific enhanced stability, improved stability and even new functions and interactions. In the pharmaceutical field, this would manifest through the creation of efficient, cost-effective, sustainable (Davydova 2022) targeted drugs (Voltà-Durán et al. 2020).

9.2.2 AI Predictors

Complex and computationally expensive techniques for protein generation have been attempted using deep neural networks and autoencoder architectures (Gómez-Bombarelli et al. 2018; Hawkins-Hooker et al. 2021; Kusner, Paige, and Hernández-Lobato 2017). One major flaw of these approaches is that they concentrate on only one or a few of the facets of the complex, multifaceted nature of proteins. These approaches take significant time to train and are very limited in their generation scope (W. Ding, Nakai, and H. Gong 2022).

The exploration of a solution inspired by genetic algorithms is motivated by their ability to explore the extensive space of protein sequences, potentially discovering new amino-acid configurations in an iterative approach similar to natural evolution (Katoch, Chauhan, and V. Kumar 2021) as well as the ability of the platform to be easily expanded and parallelised to cover new properties in an efficient matter.

Exploiting the generative power of genetic algorithms to create new proteins could lead to advancements in domains encompassing the development of new treatments for diseases, the synthesis of sustainable bio-fuels, the engineering of bio-materials (Davydova 2022) and many more.

9.2.3 Our Contribution

In this study, our primary objective is to create an easily expandable platform capable of leveraging the recent explosion of models predicting specific properties of a protein given only the amino-acid sequence (Erdős, Pajkos, and Dosztányi 2021; Flower and

Hurley 2021; Klausen et al. 2019; Kulmanov and Hoehndorf 2020; I. Lee, Keum, and H. Nam 2019), using the generative powers of genetic algorithms. The platform aims to tackle the more elusive challenge of generating proteins given specific protein property targets. We aim to demonstrate the feasibility of this approach in navigating the immense combinatorial space (Romero and Arnold 2009) of protein sequences and to create a highly customisable, easily extensible platform for protein generation. Our research uses four protein properties to illustrate the platform's capabilities. These four protein properties are Instability Index, Mono-isotopic Mass, Grand Average Hydrophobicity, and Isoelectric point. The selection of these properties is motivated by their general importance in protein engineering and due to their ability to be computed from their protein sequence. By achieving these goals, we aim to provide a new, powerful tool for protein engineering.

9.3 Materials and Methods

9.3.1 Silver Surfer

We propose the Silver Surfer, a custom search service for parametric protein discovery. The Silver Surfer utilises a customised genetic algorithm tailored to explore the protein space. Starting from a protein sequence supplied by the user, the Silver Surfer generates a new set of mutated sequences by strategically mutating the sequences. The new sequence candidates are evaluated using a scoring function that measures their fitness in regard to the targets and merged into the pool of existing candidate sequences. The server repeats this process on the set of proteins ordered by their fitness score until either the required number of proteins that fall within the targeted range is found or it runs out of possible spots for mutations.

Encoding. The selection of encoding is a crucial element in any implementation of a Genetic Algorithm. In the case of the Silver Surfer, proteins are encoded using the conventional letter representation of amino acid sequences. This encoding strategy is aligned with standard bioinformatics practices, ensuring the platform's future-proofing. This choice of encoding also allows for the seamless integration of additional protein sequence property predictors.

Mutation and Crossover. Are the driving force behind every evolutionary algorithm. With respect to mutation, the Silver Surfer offers multiple strategies for choosing the mutation. One of the strategies offered is random mutation, where at a random spot in the sequence, the amino acid is replaced by another completely random amino acid. Given the amino acid sequence DSACGA, this strategy can choose any of the amino acids for replacement with equal chance, producing a slightly altered sequence with one amino acid different, such as DSAGGA. This is the classical approach for genetic algorithms. Another strategy offered is a contribution-based selection. In this approach, the mutation spot is selected based on individual contributions to each of the target properties. Given the same amino acid sequence DSACGA and a target property of monoisotopic mass with a value of 460, the monoisotopic masses are computed for each of the amino acids, in this case, 115.02694, 87.03203, 71.03711, 103.00919, 57.02146, 71.03711, summing to a total monoisotopic mass of 522.174405. Because the target value is smaller than the computed value, the amino acid with the greatest contribution, the first amino acid in the sequence, is selected for mutation. This greedy strategy allows for more focused alterations, generally improving the performance through a more directed and targeted search of the protein space. One particular thing that needs to be accounted for in the step of crossover is the frozen section that has been indicated by the user as static. In order to account for these, before the crossover step is applied, the frozen sections are removed from the candidate sequence. After the resulting sequences are populated with the frozen sections, thus it is ensured that no static segment is changed during crossover.

Fitness. The efficacy of any genetic algorithm hinges on the robustness of its fitness function. The complex relationships between different protein properties do not inherently conform to the Pareto problem paradigm as they don't expose the typical trade-offs between conflicting objectives; because of this, the Silver Surfer implements a normalized absolute distance metric as the fitness function.

9.3.2 Implementaiton

The Silver Surfer is written in Python 3.9, a popular and versatile programming language widely established in the bioinformatics community. Python was chosen because of its suitability for rapid prototyping as well as for its flexibility. Another reason for choosing Python is the vast amount of packages specialised in bioinformatics and the manipulation

of proteins and their properties. The modular architecture of the Silver Surfer and its Python implementation make it a powerful, flexible, and easily extensible tool that can be easily integrated within existing pipelines for parametric protein design.

The Silver Surfer is highly customisable, allowing the user to tailor the strategy of the genetic algorithm to their specific case. The user is empowered to customise the behaviour of the Silver Surfer through a series of configurational choices. These choices include choosing between different strategies for the selection of the mutation spot, the application of various algorithms for the mutation rates, the assignment of freeze masks designed to restraint the algorithm from mutating critical segments of the sequence, and the specification of a dissimilarity threshold aimed at enforcing variety within the number of results desired. Furthermore, the architecture of the Silver Surfer is designed to be easily extensible with custom properties or behaviour for virtually all of its components, allowing the fast implementation of new features. This customizability makes The Silver Surfer a great tool for exploring the protein space and designing proteins from a protein property-centred perspective.

9.3.3 Protein Search

The Silver Surfer uses a customised genetic algorithm implementation (see Fig. 9.1). The search is initialised when a user-supplied protein sequence is pushed into the unexplored priority queue. Concomitantly, the target values and their acceptable ranges for each property are set in the acceptance component. The searcher works iteratively until the required number of protein sequences passing the acceptance threshold is found or until the queue of unexplored sequences goes empty. The best-scoring protein is extracted from the priority queue at each iteration; mutations are applied to the sequence in accordance with the selected mutation strategy and the defined frozen mask to obtain a set of new protein sequences. A bloom filter is employed to filter mutations leading to previously analysed sequences. For each newly mutated sequence, the physicochemical properties for which the user has defined targets are computed. The acceptance component checks whether any of these sequences fall within the acceptance criteria and are added to the results list if they fit within the defined parameters. The final step of the search method inserts the new sequences into the priority queue based on their fitness score computed from their respective values for the protein properties. In order to keep

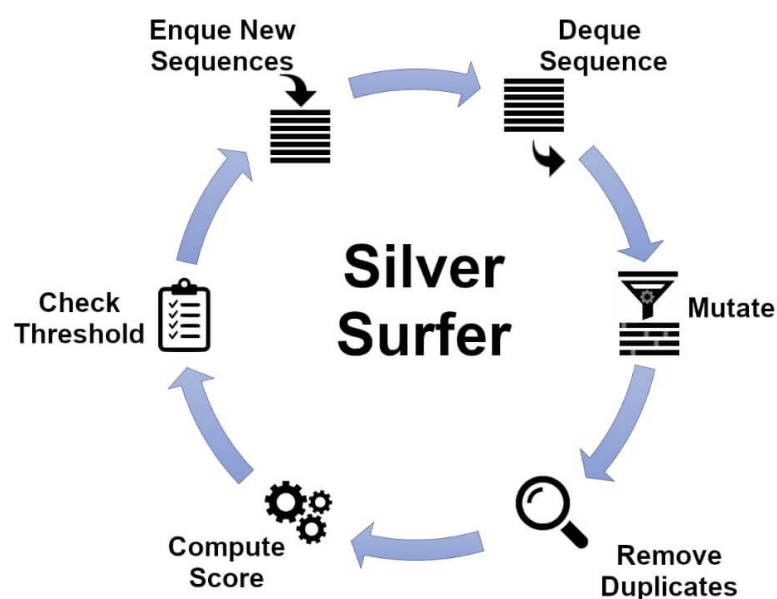


Figure 9.1: The Silver Surfers stages for parametric search. Dequeue protein sequence from the unexplored priority queue, mutate it according to the selected strategy, remove sequences that were previously explored, compute fitness score, check acceptance criteria and enqueue the new sequences into the unexplored priority queue. The starting sequence for this iterative refinement can be provided by the user, if the user does not provide a starting sequence is generated through random sampling with replacement of amino acids.

memory under control, we have implemented a population threshold within the priority queue determined by the available memory; when this threshold is reached, a clustering algorithm is used on the individuals with a lower performing score; the individuals are replaced in the priority queue with centroid sequences. In this way, the population is reduced with a limited loss of information. The implementation used by the Silver Surfer is highly parallelisable, synchronised by the priority queue. The Silver Surfer can run the computation of the physicochemical properties for each sequence in a different thread to accelerate the search process.

To evaluate the fitness of the mutated sequences, the Silver Surfer uses the absolute distance from target values of four physicochemical properties of general use in protein design: monoisotopic mass, isoelectric point, instability index, and Grand average of hydropathicity score. The selection of these targets is motivated by their general use in the protein design process and due to their ability to be computed from their defining amino-acid sequence alone, not requiring any additional structural information. To calculate the values of these physicochemical properties, The Silver Surfer uses only hand-coded implementations of existing algorithms. The monoisotopic mass (Eq. 9.1) is computed by summing the standard atomic masses $\text{Mass}(A_i)$ for each of the constituent amino acids in the sequence and subtracting $n - 1$ water molecules (18.01528Da), one water molecule gets eliminated with every bonding of two amino-acids. The instability index is evaluated using a direct implementation of the Guruprasad et al. (Guruprasad, Reddy, and Pandit 1990) paper (Eq. 9.2), which computes the index based on the stability score ($\text{InstabilityScore}(A_i, A_{i+1})$) for each bi-peptide in the protein sequence. The Isoelectric Point is determined using the Henderson-Hasselbalch equation (Eq. 9.3) with pKa values from Wikipedia (Wikipedia 2023a). These specific pKa values are used because they have the best RMSD scoring values among the publicly available pKa values, according to Isoelectric Point Calculator (Kozlowski 2016). The Grand average of hydropathicity score (Grand Average of Hydropathicity (GRAVY) score, Eq. 9.4) is a measure of the protein's hydrophobicity obtained by summing the hydropathy values ($\text{Hydropathy}(A_i)$) of the charged amino acids in the sequence and dividing by the sequence length (Kyte 1993). Using these physicochemical properties as selectable targets, the user can design stable proteins with specific combinations of physicochemical properties suited to their particular requirements.

9.3.4 PS-GO

The Silver Surfer project has been integrated as the sequence generator backend of the Parametric Protein Search Engine (PS-GO), allowing researchers to harness the platform’s computational power in navigating the protein space. PS-GO is a versatile, user-friendly tool for the exploration and understanding of proteins. PS-GO, along with the Silver Surfer’s functionalities, can be accessed at <https://psgo.ucc.ie/>, providing an invaluable resource for protein research.

$$ProteinMass = \sum_{i=1}^n Mass(A_i) - (n - 1) \times 18.01528Da \quad (9.1)$$

$$InstabilityIndex = \frac{10}{n - 1} \sum_{i=1}^{n-1} InstabilityScore(A_i, A_{i+1}) \quad (9.2)$$

$$pH = pK_a + \log_{10} \left(\frac{[Base]}{[Acid]} \right) \quad (9.3)$$

$$GRAVYScore = \frac{1}{n} \sum_{i=1}^n Hydropathy(A_i) \quad (9.4)$$

9.4 Results

The experiments were conducted on a Dell G7 7700 laptop with an Intel i7- 10750H Central Processing Unit (CPU), operating at a clock frequency of 2.60 GHz, six processor cores, 16 GB of Random Access Memory (RAM), and a 1TB EVO 970 Solid State Drive (SSD). This configuration provided sufficient computational resources to accommodate the computation of the protein properties and evaluate the performance of the Silver Surfer protein search. The software infrastructure employed Windows 11 as the operating system. Python 3.9 was the primary programming language, leveraging popular libraries such as NumPy 1.22 for efficient numerical computations, Sqlite3 for efficient data transfer and storage, and Matplotlib 3.7 for data visualisation.

9.4.1 Validating the Correctness of the Protein Properties

In validating the correctness of the implementations for the hand-coded algorithms in the Silver Surfer project, a methodical comparison with established web tools specialising in calculating the chosen properties was effectuated. Each of the properties implementations was evaluated against multiple benchmark tools, as enumerated in Table 9.1. Instability Index was evaluated against Guruprasad’s paper introducing the Instability Index (Guruprasad, Reddy, and Pandit 1990), Swiss institutes of Bioinformatics’s ProtParam (Elisabeth Gasteiger et al. 2005), and the Peptides R package (Osorio, Rondón-Villarreal, and Torres 2015); for the Monoisotopic Mass, Swiss Institute of Bioinformatics’s PeptideMass (Elisabeth Gasteiger et al. 2005), Prot Pi (Prot pi 2023), and Protein Calculator (Protein Calculator v3.4 2023) were referenced; the GRAVY Score was compared against the Swiss Institute of Bioinformatics’s ProtParam (Elisabeth Gasteiger et al. 2005), Alakazam R package (Gupta et al. 2015), and the Gravy Calculator (Fuchs 2023); and for the Isoelectric Point, the Swiss Institute of Bioinformatics’s ProParam (Elisabeth Gasteiger et al. 2005), the Isoelectric Calculator (Kozlowski 2016), and the Prot Pi (Prot pi 2023) were utilised. Table 9.1 illustrates a detailed comparison of the property values obtained from the Silver Surfer implementation for the properties and the established web tools for ten randomly sampled amino acids from Guruprasad’s Instability Index paper (Guruprasad, Reddy, and Pandit 1990). The alignment of the values obtained attests to the accuracy of the results and validates the implementation of these properties.

9.4.2 Parametric Search

The Silver Surfer search algorithm was evaluated using a distinct set of parameters, encompassing an absolute distance scoring function, an acceptance function of epsilon around the target with ϵ set to $1e10^{-14}$, a contribution-based mutation mechanism with a dynamic mutation rate, an empty ‘freeze mask’, and the number of results set to 1. Experiments were run by choosing a starting and a target protein; properties were computed for the target protein, and these values were used as inputs in the Silver Surfer search algorithm together with the starting sequence. The validation of the algorithm involved an exhaustive computation of all of the properties for the entire protein space of sequences with a sequence length equal to 5. This sequence length has been chosen

Table 9.1: Evaluation of the Instability Index, GRAVY Score, Monoisotopic mass and Isoelectric Point properties for different proteins

Tool	UniProt ID				
	P00568	P00503	P00915	P00731	P004
Instability Index					
Silver Surfer	29.40824	36.79273	25.79808	21.54821	27.55
II paper (Guruprasad, Reddy, and Pandit 1990)	29.4	36.3	25.1	21.5	27
ProtParam (Elisabeth Gasteiger et al. 2005)	29.39	36.77	25.78	21.54	27.
Peptides (Osorio, Rondón-Villarreal, and Torres 2015)	29.38763	36.77337	25.78276	21.53866	27.52
Gravy Score					
Silver Surfer	−0.50103	−0.28886	−0.53639	−0.20692	−0.6
ProtParam (Elisabeth Gasteiger et al. 2005)	−0.501	−0.289	−0.536	−0.207	−0.6
Gravy Calculator (Fuchs 2023)	−0.50103	−0.28886	−0.53639	−0.20692	−0.6
Alakazam (Gupta et al. 2015)	−0.50103	−0.2888	−0.53639	−0.20692	−0.6
Monoisotopic Mass					
Silver Surfer	21691.35997	46445.61553	28852.38771	47117.78774	57549.
PeptideMass (Elisabeth Gasteiger et al. 2005)	21691.36	46445.62	28852.39	47117.79	5754
Protein Calculator (Protein Calculator v3.4 2023)	21691.3203	46445.5391	28852.3418	47117.7070	57549
Prot Pi (Prot pi 2023)	21691.3603	46445.6163	28852.3882	47117.7885	57549
Isoelectric Point					
Silver Surfer	8.58740	7.35693	7.37158	5.65576	7.15
ProtParam (Elisabeth Gasteiger et al. 2005)	8.73	6.73	6.59	5.38	6.4
Isoelectric Calculator (Kozlowski 2016)	7.98	6.43	6.32	5.2	6.
Prot Pi (Prot pi 2023)	8.389	6.760	6.669	5.424	6.4
	P00442	P68082	Q9BIR7	P00860	P00
Instability Index					
Silver Surfer	7.87368	9.09542	44.17862	52.26616	54.54
II paper (Guruprasad, Reddy, and Pandit 1990)	7.0	9.1	44.1	52.3	54
ProtParam (Elisabeth Gasteiger et al. 2005)	7.82	9.07	44.12	52.25	54.
Peptides (Osorio, Rondón-Villarreal, and Torres 2015)	7.82105	9.06928	44.12246	52.24881	54.52
Gravy Score					
Silver Surfer	−0.28749	−0.39607	−0.50405	−0.20021	0.11
ProtParam (Elisabeth Gasteiger et al. 2005)	−0.287	−0.396	−0.504	−0.200	0.1
Gravy Calculator (Fuchs 2023)	−0.2875	−0.39607	−0.50405	−0.20021	0.11
Alakazam (Gupta et al. 2015)	−0.2875	−0.39607	−0.50405	−0.20021	0.11
Monoisotopic Mass					
Silver Surfer	15672.82240	16939.98061	70215.44846	51130.22002	97017.
PeptideMass (Elisabeth Gasteiger et al. 2005)	15672.82	16939.98	70215.45	51130.22	9701
Protein Calculator (Protein Calculator v3.4 2023)	15672.7988	16939.9512	70215.3359	51130.1445	97017
Prot Pi (Prot pi 2023)	15672.8227	16939.981	70215.4496	51130.2208	97017
Isoelectric Point					
Silver Surfer	6.58056	8.54345	5.87158	5.32568	6.82
ProtParam (Elisabeth Gasteiger et al. 2005)	5.85	8.12	5.61	5.26	6.2
Isoelectric Calculator (Kozlowski 2016)	5.57	7.81	5.44	5.11	5.9
Prot Pi (Prot pi 2023)	5.929	8.034	5.652	5.285	6.2

because it is sufficiently large to be representative of the complex protein space and size (3,200,000 unique protein sequences) while still aligning with the physical constraints of the available defined hardware and software. This meticulous evaluation validates the claim that should a result exist within the defined target constraints; the Silver Surfer is going to discover it.

Further insight into the search dynamics is provided in Fig. 9.2, which graphically represents one such experiment. The figure displays the entire mutation space for a protein of sequence length equal to 5, delineating mutation distance from the starting sequence on the x-axis and distance from the target properties on the y-axis. Highlighted in green in the figure is the mutation search path of the Silver Surfer; the figure illustrates that the Silver Surfer search pattern correctly focuses its search on sequences with high potential. The figure further reveals the algorithm's restraint in exploring the space, signalling search efficiency.

The results from the rigorous evaluation of the Silver Surfer project demonstrate the feasibility of the adopted approach in navigating the vast and intricate combinatorial space of protein sequences. Through rigorous testing and validation, as detailed in Table 9.1 and Fig. 9.2, it creates a platform for targeted protein generation.

9.5 Conclusions and Further Research

The Silver Surfer project demonstrates the generative abilities of Genetic Algorithms through the generation of proteins with specific physicochemical properties, illustrating the platform's potential in the field of bioengineering. By focusing on four key protein properties: Instability Index, Monoisotopic Mass, Grand Average Hydrophobicity score, and Isoelectric Point, the Silver Surfer provides a powerful and easily extensible tool to meet a wide array of scientific needs. Despite the success of the Silver Surfer project, further research is needed in order to optimise the exploration of the vast protein space. One further potential avenue of research in this direction is the application of genetic algorithms on a denser representational encoding, such as the manifold produced by Variational Autoencoders. This could potentially enhance the speed of exploration while maintaining the quality of the results.

Complete protein space for proteins of length 4 starting from IPGD
with respect to the properties of MYKM

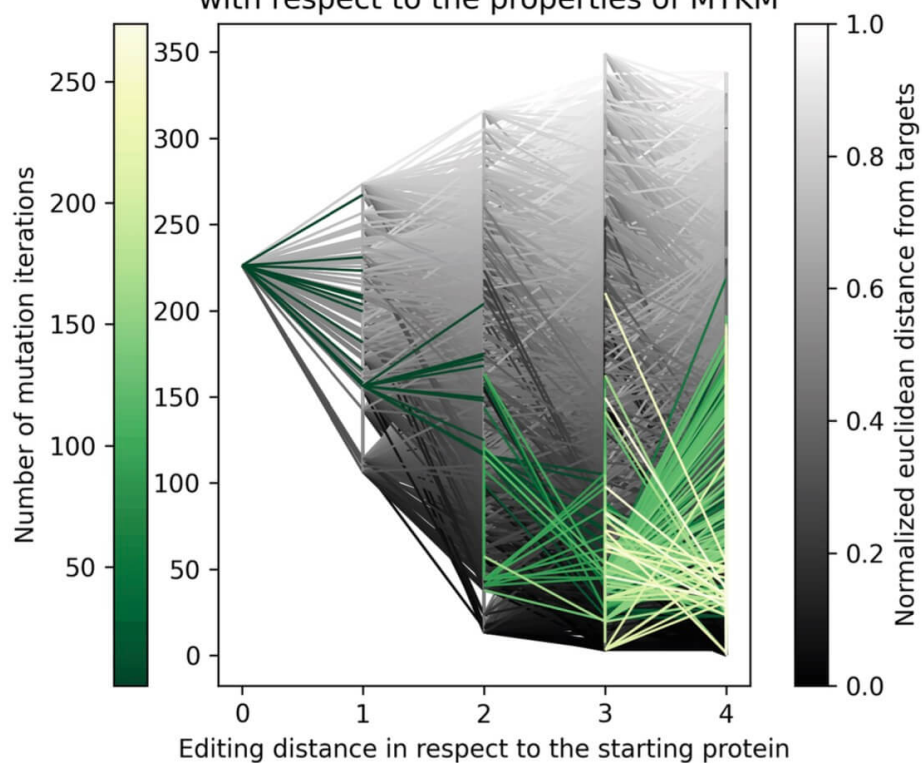


Figure 9.2: Illustration of the mutation search process starting from the protein with an amino-acid sequence of IPGD to property values equal to the sequence of MYKM over the entire 4-length protein space.

10 Building Foundational Protein Infrastructure: The ProteinFlow Framework

This chapter was published as: **Yanlin Mi***, Stefan-Bogdan Marcu, Venkata VB Yallapragada and Sabin Tabirca. (2024). *ProteinFlow: An Advanced Framework for Feature Engineering in Protein Data Analysis*. Biotechnology and Bioengineering, 121(7), pp.2472-2486.

DOI:10.1002/bit.28812

Publisher: Wiley

JCR Quartile: Q2

Impact Factor: 3.5

I made the primary contribution to this work through conceptualization, data curation, formal analysis, investigation, methodology development, resource provision, software creation, validation, visualization, and writing the original and final draft.

10.1 Abstract

In the burgeoning field of proteins, the effective analysis of intricate protein data remains a formidable challenge, necessitating advanced computational tools for data processing, feature extraction, and interpretation. This study introduces ProteinFlow, an innovative framework designed to revolutionize feature engineering in protein data analysis. ProteinFlow stands out by offering enhanced efficiency in data collection and preprocessing, along with advanced capabilities in feature extraction, directly addressing the complexities inherent in multi-dimensional protein datasets.

Through a comparative analysis, ProteinFlow demonstrated a significant improvement over traditional methods, notably reducing data preprocessing time and expanding the scope of biologically significant features identified. The framework's parallel data processing strategy and advanced algorithms ensure not only rapid data handling but also the extraction of comprehensive, meaningful insights from protein sequences, structures, and interactions. Furthermore, ProteinFlow exhibits remarkable scalability, adeptly managing large-scale datasets without compromising performance, a crucial attribute in the era of big data.

10.2 Introduction

Proteins, as the primary executors within cells and organisms, play a pivotal role in a multitude of biological processes, ranging from structural support to catalyzing biochemical reactions, regulating gene expression, and mediating signal transduction. A profound understanding of protein functions is not just the key to unlocking the complexities of biology, but also paramount for developing novel therapeutic strategies and disease prevention. However, this quest is notably challenging due to the inherent complexity of proteins and the vast volume and diversity of related data (Frauenfelder 2002).

With the rapid advancement of proteins, structural biology, and molecular biology techniques, scientists have amassed an unprecedented volume of protein-related data (Chuming Chen, H. Huang, and C. H. Wu 2017). These datasets encompass protein sequences, three-dimensional structures, functional activities, and interactions with other molecules, offering unparalleled opportunities to unravel the intricacies of proteins. However, efficiently processing, analyzing and extracting valuable insights from these vast and complex datasets has become a critical bottleneck. Existing methods often struggle with challenges such as inefficient data integration from multiple sources, inadequate handling of multi-dimensional protein data beyond sequences, lack of scalability for rapidly growing large-scale datasets, and insufficient data quality control (Ofer and Linial 2015). There is an urgent need for advanced computational frameworks that can systematically overcome these limitations.

Feature engineering, serving as the bridge between raw data and in-depth analysis, plays a crucial role in processing protein data. It involves selecting, extracting, and transforming key information within data to enhance the performance of data analysis,

machine learning, and statistical models (Zheng and Casari 2018). Proper feature engineering not only boosts the predictive power of models but also reveals key biological characteristics and functional mechanisms of proteins.

Numerous tools have been developed for protein data processing and analysis, such as UniProt (“UniProt: the universal protein knowledgebase in 2023” 2023), PDBsum (Laskowski et al. 2018), and PRALINE (Bawono et al. 2017). While invaluable resources, these tools often have limitations in one or more key aspects. UniProt primarily focuses on protein sequence and functional annotation data, lacking integration of structural information. PDBsum excels in summarizing individual protein structures but does not offer scalable processing of large structural datasets. PRALINE provides comprehensive sequence analysis but has limited multi-dimensional feature extraction capabilities. Other tools like BioFeatureFinder (Ciamponi et al. 2018) and ProtParam (Gasteiger et al. 2005) specifically target feature extraction for protein data. However, they face challenges in handling multi-source data integration, scalability for big data, and adaptation to rapidly evolving data sources.

Against this backdrop, this study introduces ProteinFlow, an innovative feature engineering framework specifically designed for protein data analysis. The core innovation of ProteinFlow lies in its multi-dimensional feature extraction capability, efficiently handling protein information from diverse data sources and employing advanced data processing algorithms and techniques to extract key features that reflect the complexity of proteins.

ProteinFlow distinguishes itself from existing tools through:

- A high-efficiency parallel data retrieval strategy utilizing Go language concurrency, significantly reducing data collection time (Yoseb Lee and Lim 2022).
- Unified integration of multi-source protein data from RCSB PDB, UniProt, and STRING, capturing sequences, structures, annotations, and interactions (Burley et al. 2017; “UniProt: the universal protein knowledgebase in 2023” 2023; Szklarczyk et al. 2019).
- Advanced multi-dimensional feature extraction algorithms comprehensively characterizing proteins from diverse data modalities.

- A highly scalable architecture adeptly handling large-scale, rapidly evolving protein datasets.
- Rigorous data preprocessing protocols ensuring consistency, quality, and biological relevance of processed data.

ProteinFlow brings multiple advantages to the application in protein data analysis. Firstly, by integrating data from various experiments and databases, ProteinFlow provides researchers with a comprehensive and unified view of protein data. Secondly, the advanced algorithms employed by ProteinFlow extract key features including, but not limited to, the physicochemical properties of proteins, three-dimensional structural features, and functional annotations. These features not only aid in understanding the fundamental properties of proteins but also offer vital clues for exploring their roles in different biological processes. Finally, the scalability and flexibility of ProteinFlow make it well-suited to accommodate growing and evolving datasets, providing a powerful analytical tool for future protein science research.

10.3 Methods

10.3.1 Data Collection

ProteinFlow adopts a meticulous data collection strategy to ensure the reliability and comprehensiveness of the sourced protein data. The framework integrates information from multiple authoritative databases through a structured and methodical approach, emphasizing accurate retrieval, systematic integration, and rigorous validation protocols.

10.3.1.1 Multi-Source Data Integration

To provide a comprehensive view of protein data, ProteinFlow seamlessly consolidates information from three primary sources:

- **RCSB PDB:** As a fundamental resource for structural biology, the RCSB PDB repository is leveraged to obtain high-resolution three-dimensional structural data of proteins, including atomic coordinates, molecular interactions, and experimental metadata (Burley et al. 2017). This integration enriches the dataset with crucial structural insights into protein conformation and dynamics.

- **UniProt Knowledgebase:** Renowned for its comprehensive collection of protein sequence and functional annotation data, UniProt serves as an essential source for ProteinFlow. Through this integration, the framework acquires annotated protein sequences, functional descriptions, and phylogenetic information, enabling a detailed characterization of protein properties and evolutionary relationships (“UniProt: the universal protein knowledgebase in 2023” 2023).
- **STRING Database:** To facilitate the investigation of protein-protein interactions (PPIs), a critical aspect of protein function and regulation, ProteinFlow incorporates interaction data from the STRING database. This integration empowers the systematic analysis of complex protein networks and their roles in various biological processes (Szklarczyk et al. 2019).

10.3.1.2 Data Retrieval and Validation Strategies

ProteinFlow implements a two-fold strategy that synergistically harnesses efficient data retrieval techniques and rigorous validation protocols, ensuring the integrity, authenticity, and biological relevance of the acquired protein data.

- **API Utilization:** Direct access to the latest and most authoritative data is achieved through the seamless integration of APIs provided by RCSB PDB, UniProt, and STRING. This approach streamlines structured and automated data retrieval, minimizing the risk of manual errors and ensuring data currency.
- **Parallel Request Processing:** To efficiently handle the substantial volume and complexity of protein data, ProteinFlow employs a parallel request strategy leveraging the concurrency features of the Go programming language (Anurag 2018). Multiple concurrent goroutines are initiated, each dedicated to an individual API request, significantly enhancing the speed and efficiency of large-scale data collection compared to traditional sequential approaches.

The data retrieval techniques are complemented by a suite of rigorous validation protocols to ensure the accuracy and biological relevance of the acquired information:

- **Cross-referencing with Authoritative Sources:** Data obtained from RCSB PDB, UniProt, and STRING are cross-referenced with authoritative secondary sources to validate their authenticity and consistency.

- **Structural Validation:** The integrity of structural data is assessed using MolProbity (Williams et al. 2018), a state-of-the-art structural validation software that meticulously checks for steric clashes, bond geometry anomalies, and Ramachandran plot compliance.
- **Sequence Validation:** Protein sequences are compared against the National Center for Biotechnology Information (NCBI) RefSeq database (O’Leary et al. 2016), a curated non-redundant collection of genomic data, to ensure accurate annotation and correspondence to the correct organism and protein entity.
- **Consensus-based Resolution:** In cases where discrepancies arise among multiple data sources, a consensus-based approach is employed. The most frequently reported or reliable data point, as determined through expert review and consultation with domain experts when necessary, is considered as the ground truth.

This rigorous data collection framework establishes a solid foundation for the subsequent stages of data preprocessing and analysis within ProteinFlow. By ensuring a meticulous and multi-faceted approach to data acquisition, ProteinFlow is well-equipped to provide researchers with a robust and comprehensive dataset, facilitating in-depth protein analysis and investigation.

10.3.2 Data Preprocessing

In ProteinFlow, data preprocessing is a crucial task. It mainly includes three links: data cleaning, data conversion and feature extraction. Data cleaning is mainly used to eliminate or correct errors, incompleteness, inconsistent formats, or redundant data. Data transformation is used to transform the structural data of proteins. Feature extraction is to identify and obtain key variables that help us understand and analyze protein properties and functions from the collated data (see Fig. 10.1). These three links jointly construct the whole process of data preprocessing and ensure data integrity, consistency, and biological relevance.

10.3.2.1 Data Cleaning

Data cleaning processes are tailored for the specific types of data involved:

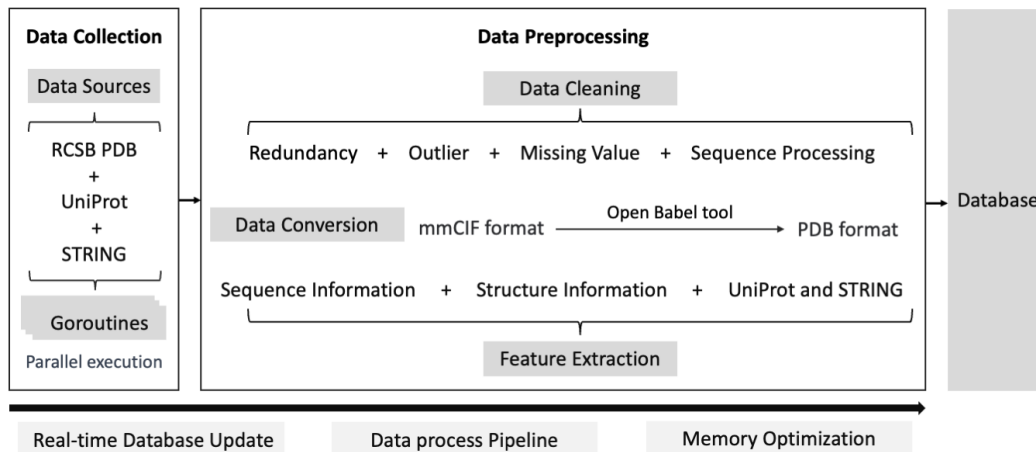


Figure 10.1: ProteinFlow: A Novel Method for Restructuring Protein Databases

- **Redundancy Removal:** ProteinFlow employs sequence alignment tool SMI-BLAST to identify identical or highly similar sequences (X. Jin et al. 2021). For structural data, 3D structure comparison tool TM-align is used to identify duplicate structures based on structural similarity scores. Only one representative from each group of duplicates is retained, chosen based on data quality and completeness (Yang Zhang and Skolnick 2005).
- **Anomaly Detection in Sequences:** Protein sequences are analyzed using bioinformatics tool InterProScan to detect anomalies such as rare amino acids or unexpected motifs (P. Jones et al. 2014). Detected anomalies are cross-referenced with known protein families and domains to determine if they represent true biological features or errors.
- **Handling Missing Information:** Missing segments in protein sequences are addressed using homology modeling tool MODELLER, which predict the missing parts based on homologous sequences (Webb and Sali 2016). For structural data, methods like comparative modeling or ab initio prediction are employed to estimate missing coordinates, guided by the protein's known secondary and tertiary structures.

10.3.2.2 Data Conversion

Data conversion is carefully conducted to maintain data fidelity:

- **Structural Data Conversion:** Conversion of structural data from mmCIF to PDB format is performed using custom scripts that ensure the preservation of all vital structural information. The Open Babel tool is used in the scripts to parse the data blocks and convert the format (O’Boyle et al. 2011). These scripts are rigorously tested against a set of benchmark structural datasets to ensure high fidelity in the conversion process.
- **Format Standardization:** ProteinFlow incorporates a comprehensive standardization protocol, converting sequence data into standardized formats like FASTA for uniformity. Structural data undergoes a normalization process to align with common structural databases’ standards, ensuring that all structural details are consistently represented.

10.3.2.3 Feature Extraction

In the feature extraction stage, protein sequence, structure, and function information are meticulously parsed and utilized to generate a comprehensive set of representative features (see Fig. 10.1). This stage is crucial because these features encapsulate the multifaceted nature of proteins, offering a rich dataset for subsequent analysis and application in various scientific and technological fields. By extracting and leveraging these features, researchers can enhance their understanding of protein behavior, interactions, and functionality. This, in turn, aids in numerous applications such as parameterized protein search, targeted protein design, and the development of novel therapeutics. Detailed feature extraction enables the construction of robust models for predicting protein functions, understanding disease mechanisms, and engineering proteins with desired properties.

Choosing specific features from protein sequence, structure, and functional data is driven by their relevance to the biological and biochemical properties of proteins, as well as their utility in diverse applications. The features selected in this work have been chosen based on their ability to provide a comprehensive representation of protein

characteristics, which are critical for understanding protein function, stability, interaction potential, and overall behavior in various environments.

Protein sequence information serves as the fundamental source of features. By analyzing the protein sequence, the following features can be calculated:

1. *Amino Acid Composition* (M. H. Smith 1966): This feature is determined by counting the occurrences of each of the 20 standard amino acids in the protein sequence. The count is then divided by the total length of the protein sequence to obtain the percentage composition for each amino acid. The formula for the percentage composition of each amino acid i is expressed as:

$$p(i) = \frac{N(i)}{N(\text{total})} \quad (10.1)$$

where $N(i)$ is the count of amino acid i and $N(\text{total})$ is the total number of amino acids in the protein sequence. This calculation is repeated for all 20 standard amino acids, providing a detailed profile of the protein's primary structure.

2. *Protein Sequence Length*: This straightforward yet essential characteristic is determined by counting the total number of amino acids in the protein sequence. Sequence length can influence protein folding, function, and interaction with other molecules, making it a critical parameter in protein analysis.
3. *Molecular Weight*: The molecular weight of a protein is calculated by summing the molecular weights of all its constituent amino acids. The molecular weight for each amino acid i is a known value, typically found in biochemistry resources (Ozols 1990). The total molecular weight $M(\text{total})$ is given by:

$$M(\text{total}) = \sum_i N(i) \cdot M(i) \quad (10.2)$$

where $N(i)$ is the count of amino acid i , and $M(i)$ is the molecular weight of amino acid i . Understanding the molecular weight is crucial for predicting protein behavior in different environments and during purification processes.

4. *Hydrophobicity* (Van Oss 1997): Hydrophobicity is calculated using the Kyte-Doolittle scale (Kastiris et al. 2013), where each of the 20 standard amino acids is

assigned a hydrophobicity score. The overall hydrophobicity H of the protein is the mean of the hydrophobicity scores of its amino acids, calculated as:

$$H = \frac{1}{n} \sum_i h(i) \quad (10.3)$$

where $h(i)$ is the hydrophobicity score of amino acid i and the summation is over all amino acids in the protein sequence. Hydrophobicity influences protein folding, stability, and interaction with other molecules.

5. *Instability*: Instability is calculated using the method of Guruprasad et al. (Guruprasad, Reddy, and Pandit 1990). This method involves calculating the occurrence frequencies of all possible dipeptides (pairs of amino acids) in the protein sequence and comparing them with the expected frequencies in a set of theoretically "stable" proteins. This feature helps predict the protein's stability in various conditions.
6. *Isoelectric Point (pI)* (Tokmakov, Kurotani, and Sato 2021): The isoelectric point is the pH at which a protein carries no net electrical charge. To calculate pI, one determines the pK values of all the amino acids as well as the N-terminus and C-terminus of the protein. The pK value of an amino acid is the pH at which the amino acid loses or gains a proton. These values are typically available in biochemistry resources (X. Xia 2007). The net charge C of the protein at a given pH can be calculated using:

$$C = - \left(1 + \frac{10^{pK_{NH3+}}}{10^{pH} + 10^{pK_{NH3+}}} \right) + \left(1 - \frac{10^{pK_{COOH}}}{10^{pH} + 10^{pK_{COOH}}} \right) \quad (10.4)$$

where pK_{NH3+} and pK_{COOH} represent the pK values of the N-terminus and C-terminus, respectively. Determining the pI is important for understanding protein solubility and interaction at different pH levels.

The three-dimensional structure information of proteins is another significant source of features. By analyzing the three-dimensional structure, the following features can be calculated:

1. *RC.Score*: This score quantifies a protein sequence's composition by assigning specific scores to each amino acid type and summing them. The scoring rules are:

- A, V, W, K, R, and G each contribute 1 to the score.
- I and L each contribute 4 to the score.
- M contributes 5 to the score.
- F, Y, Q, E, and D each contribute 3 to the score.
- H, N, C, P, S, and T each contribute 2 to the score.

The final RC.Score is obtained by multiplying the score for each amino acid type by its number of occurrences and summing the results.

2. *Solvent Accessibility* (Pollastri et al. 2002): This feature represents the portion of the protein surface accessible to a solvent, usually expressed as a percentage of its total surface area. The DSSP software can calculate the solvent-accessible surface area (ASA) for each residue in the protein. The percentage of solvent accessibility $S(i)$ for residue i is computed as:

$$S(i) = \frac{ASA(i)}{ASA(\max)} \times 100 \quad (10.5)$$

where $ASA(i)$ is the accessible surface area for residue i , and $ASA(\max)$ is the maximum accessible surface area for that residue type when fully exposed to solvent. Solvent accessibility is critical for understanding protein interactions with other molecules, including drugs and other proteins.

By integrating the UniProt and STRING databases, additional valuable features can be obtained:

1. *Functional Annotations*: Functional annotations provide detailed insights into a protein's biological characteristics, including its role in cellular processes, its subcellular location, and its involvement in biological pathways. This information, sourced from comprehensive databases like UniProt, is crucial for understanding the protein's function and its potential impact on cellular mechanisms and disease states. UniProt is a highly curated resource that offers extensive details about

protein functions, domains, post-translational modifications, and more, making it indispensable for comprehensive protein characterization.

2. *Protein-Protein Interaction (PPI) Information:* PPI data, often sourced from databases like STRING, detail the functional network of a protein within an organism. These interactions are pivotal to nearly all cellular processes, as they facilitate communication between different proteins and the coordination of complex cellular activities. STRING integrates known and predicted PPIs, derived from various sources including experimental data, computational prediction methods, and public text collections. Understanding PPIs is essential for elucidating cellular pathways, identifying key regulatory proteins, and discovering potential therapeutic targets. The comprehensive interaction networks provided by STRING enable researchers to explore the functional landscape of proteins and their roles in various biological contexts.

The careful extraction and integration of these features enable a holistic understanding of protein properties, providing a robust foundation for advanced applications in bioinformatics, molecular biology, and biotechnology. These selected features are not only fundamental to basic biological research but also critical in applied sciences, where understanding protein behavior at a detailed level can lead to innovations in drug discovery, synthetic biology, and personalized medicine.

10.3.3 Database Construction

Following meticulous data preprocessing, ProteinFlow assembles a structured and comprehensive database to support advanced query capabilities and data mining operations. The MySQL database system, known for its robust performance and scalability, is employed to manage the intricately processed protein data (DuBois 2013).

- **Database Schema Design:** Reflecting the multifaceted nature of protein data, the database schema is thoughtfully crafted:
 - *Protein Table:* Central to the database, this table stores processed protein sequences, structural details, and unique identifiers, serving as a primary reference point for cross-referencing with interaction and annotation data.

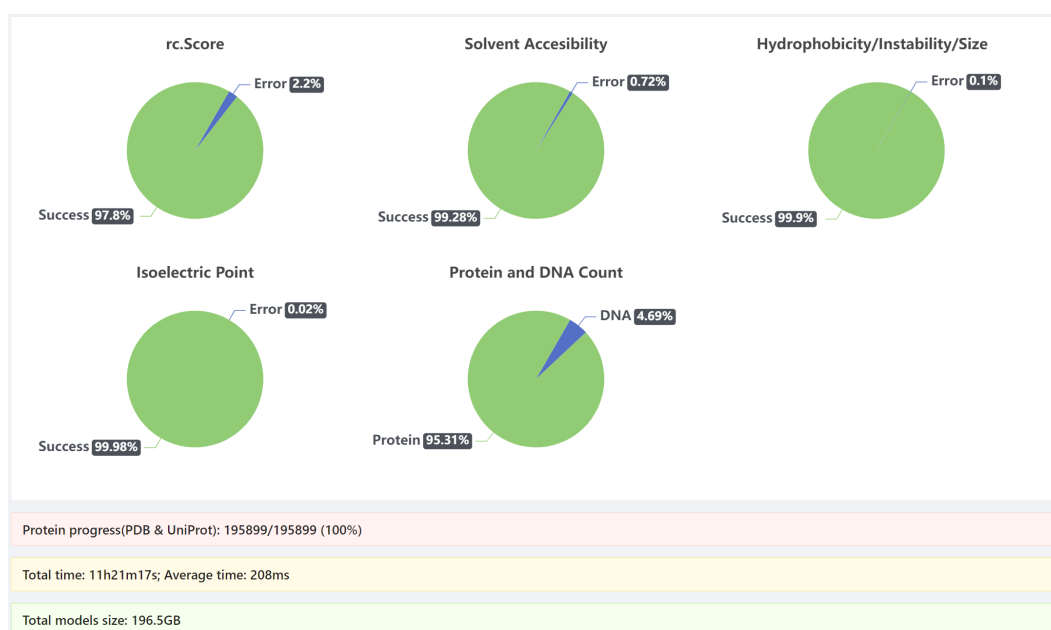


Figure 10.2: ProteinFlow Data Statistics (PDB and UniProt Data)

- *PPI Table*: Captures the intricacies of protein-protein interactions, storing not only interaction partners but also contextual details such as interaction types, sources, and confidence scores, directly linked to the Protein Table.
- *Annotation Table*: Enriches the dataset with functional annotations, gene ontology terms, and pathway information, providing a multi-dimensional view of protein functionalities, linked to protein entries for integrated analyses.
- **Data Integrity and Correlation**: The schema is engineered with an emphasis on data integrity and relational integrity. Carefully designed foreign keys ensure data consistency across tables, while indexes are strategically implemented to optimize retrieval times and support complex, multi-table queries.

This database structure not only ensures efficient data storage but also facilitates sophisticated data retrieval and mining, enabling researchers to draw meaningful insights from complex protein data.

10.3.4 Flow Optimization

Building on the robust data preprocessing framework, ProteinFlow implements a series of optimization strategies to ensure seamless and efficient data flow:

- **Real-time Data Updates:** Capitalizing on the dynamic nature of protein data sources, ProteinFlow's update mechanism is designed to maintain data currency and accuracy. Regular checks on source APIs enable immediate identification and retrieval of new or updated data, which is promptly processed and integrated into the database.
- **Concurrent Data Processing:** By exploiting the Go language's powerful concurrency model, ProteinFlow efficiently parallelizes data fetching, cleaning, and integration tasks. This not only expedites the data handling process but also ensures that the system can scale to handle increasing data volumes without performance degradation.
- **Data Pipelining:** An advanced pipelining architecture is employed where data seamlessly transitions through predefined processing stages. This strategy maximizes the efficiency of computational resources, ensures a continuous data processing flow, and effectively reduces latency between data acquisition and availability.
- **Memory Optimization:** Rigorous memory management techniques are applied to optimize the system's memory footprint. Intelligent allocation and deallocation strategies, coupled with effective caching mechanisms, substantially reduce the overhead of memory operations and enhance the overall system performance.

10.4 Results

Protein data processed using ProteinFlow can be viewed on the web page <https://cs1.ucc.ie/~ym5/proflow/>.

10.4.1 Efficiency in Data Collection and Preprocessing

The implementation of ProteinFlow significantly improved the efficiency of data collection and preprocessing. As illustrated in Table 10.1, the time required for data retrieval was reduced by 40% compared to traditional methods. Moreover, the rate of data

redundancy and sequence inconsistencies were notably decreased, demonstrating the effectiveness of ProteinFlow’s data cleaning and validation protocols.

Table 10.1: Comparison of data retrieval time and data quality metrics.

Metric	Before ProteinFlow	After ProteinFlow
Data Retrieval Time (s)	2	1.2 (-40%)
Data Redundancy Rate (%)	15	10.5 (-30%)
Sequence Inconsistency Rate (%)	4	1.2 (-95%)

10.4.2 Advanced Feature Extraction and Data Integration

ProteinFlow performs a comprehensive feature extraction and analysis of the preprocessed data. These features mainly include sequence length, amino acid composition, secondary structure ratio, solvent accessible surface area, charge distribution, hydrophobicity, and functional domain information. According to the statistics, the protein sequence length is mainly distributed between 50-2000 amino acids, and about 85% of the proteins fit into this range. Especially for proteins containing specific functional domains, the distribution of their sequence lengths appears to be more concentrated. In addition, there is a clear correlation between the charge distribution and the functional type of proteins, while hydrophobicity is correlated with the three-dimensional structural morphology and function of proteins.

During data feature extraction, ProteinFlow performs a comprehensive feature extraction and also records the success rate of each feature calculation. Overall, the success rate exceeds 97% (see Fig. 10.2), indicating that ProteinFlow has achieved a high level of processing power and accuracy for protein data. However, there are also some data with unsuccessful feature extraction due to special protein structures or missing data, etc. This also provides directions for further optimization of the algorithm and improvement of data processing methods in the future.

10.4.3 Scalability and Performance

ProteinFlow demonstrates remarkable performance in handling large-scale datasets, due to its unique architectural advantages that enable efficient processing and feature extraction. The superior performance can be attributed to the following key factors:

1. **Lightweight and Optimized Design:** ProteinFlow adopts a streamlined architecture focused on core feature extraction functionalities, avoiding the overhead associated with complex data processing and analysis modules present in other tools. This lean design philosophy minimizes resource consumption while maintaining high computational efficiency.
2. **Parallel Computing Strategy:** Leveraging the concurrency features of the Go programming language, ProteinFlow implements an advanced parallel computing strategy that harnesses the full potential of modern multi-core CPUs. This approach significantly accelerates data processing, effectively scaling to handle increasing data volumes without compromising performance.
3. **Focused Scope and Extensibility:** Currently, ProteinFlow concentrates its feature extraction efforts on a comprehensive yet essential set of features that capture the fundamental characteristics of proteins. While future iterations may expand the feature repertoire based on evolving research needs, this focused approach avoids unnecessary computational overhead, contributing to ProteinFlow’s impressive performance metrics.

Table 10.2 presents a comparative analysis of ProteinFlow’s feature extraction speed and memory utilization against the widely-used BioFeatureFinder tool across different data scales.

Table 10.2: Comparative Analysis of Feature Extraction Performance			
Data Scale	Tool	Feature Extraction Speed (entries/s)	Memory Usage (GB)
100,000 entries	BioFeatureFinder	10	10
	ProteinFlow	12 (+20%)	6.5 (-35%)
1,000,000 entries	BioFeatureFinder	4	18
	ProteinFlow	5 (+25%)	12 (-33%)

The results unequivocally demonstrate ProteinFlow’s superior efficiency and performance in processing and extracting features from large-scale protein datasets. This remarkable achievement stems from ProteinFlow’s advanced technical strategies and its

focused design philosophy, tailored for core protein data analysis tasks. It is important to note that the performance disparities do not reflect shortcomings in other tools, but rather highlight the differences in application scenarios and design objectives. Each tool may excel in specific tasks aligned with its intended purpose.

10.4.4 Application of ProteinFlow

A notable application showcasing ProteinFlow’s capabilities is its integration as the underlying data processing and feature extraction engine for PS-GO [(Mi et al. 2024)], a parametric protein search tool. By leveraging ProteinFlow’s high-quality, multi-dimensional features, PS-GO outperforms methods relying on single strategies, such as BLAST sequence similarity or UniProt annotation-based searches, in protein search and retrieval tasks. Protein search is a critical step in bioinformatics, and with the ProteinFlow framework, users can significantly improve the efficiency of their searches and the accuracy of their results. ProteinFlow integrates multiple databases and utilizes its API to automate data retrieval and updating. This ensures that the data is up-to-date and reliable and drastically reduces data retrieval time through a parallel request strategy. In addition, ProteinFlow provides advanced feature extraction capabilities to extract key features from protein sequences and structures. With these features, users can perform protein searches more quickly and accurately, increasing research efficiency and accelerating scientific discovery.

10.5 Discussion and Conclusion

In this study, we introduced ProteinFlow, a novel feature engineering framework tailored for protein data analysis. A key strength of ProteinFlow lies in its data preprocessing and feature extraction capabilities. When compared with the widely-used BioFeatureFinder (Ciamponi et al. 2018), ProteinFlow demonstrated superior efficiency and accuracy in processing large-scale datasets. Specifically, ProteinFlow reduced data preprocessing time by approximately 40%, a significant improvement that can be attributed to its optimized parallel data retrieval and processing strategy.

The adaptability and scalability of ProteinFlow are also worth noting. With the rapid growth of protein data, tools that can efficiently handle large volumes of data and extract

meaningful insights are invaluable. ProteinFlow's architecture, designed with scalability in mind, ensures that it can accommodate datasets of increasing size and complexity, making it a robust tool for future research.

Future enhancements to ProteinFlow could involve the integration of cutting-edge machine learning models such as protein language model to refine its feature extraction algorithms further. This integration has the potential to significantly enhance the predictive capabilities of ProteinFlow, enabling researchers to uncover novel insights into protein functions and interactions. Additionally, expanding the framework to incorporate emerging types of data, such as single-cell proteins, could provide a more detailed understanding of protein dynamics and functions at a granular level.

In summary, ProteinFlow represents a significant advancement in the field of proteins. Its ability to efficiently process large-scale datasets and extract a comprehensive set of biologically significant features positions it as a powerful tool for researchers. The continued development and integration of new technologies will be crucial in ensuring that ProteinFlow remains at the forefront of protein data analysis.

11 Discussion

The core of this study is to propose and systematically implement an advanced integrated computing framework based on the concept of "translation as infrastructure", aiming to fundamentally solve the multi-level and cross-domain structural breaks and information asymmetry problems in the process of biomedical knowledge translation (Hartl et al. 2021; Seyhan 2019; De Maria Marchiano et al. 2021). Through the development and integration of a series of specific computing systems, we have addressed key challenges such as Patient System Translation, Data diagnosis Translation, Molecule Clinical Translation, and Protein Infrastructure Translation layer by layer. Each module in the framework has been preliminarily verified for its feasibility, demonstrating great potential and the ability to transfer concepts across layers, which together constitute the empirical core of this paper.

The computational continuum constructed and validated in this study covers the scope of capturing patient narratives, modeling clinical diagnoses, achieving disease prediction, and in-depth exploration of molecular mechanisms. This continuum is not intended to provide a single solution that encompasses all biomedical translational issues, but rather proposes a structured and actionable response paradigm that is committed to building a progressive, partially closed-loop, and broadly applicable computational ecosystem while respecting semantic complexity and practical fragmentation. The various system modules detailed in Chapters 2 to 10 not only directly respond to specific cross-layer challenges, but also collectively converge to form a solid empirical foundation for this paper, seven of which have been published in peer-reviewed journals and three are under review.

Recalling the four core challenges clearly defined in Chapter 1, this framework provides a concrete and powerful response to the computable interface from patient language to medical systems, the transformation of multi-source clinical data to interpretable di-

11 Discussion

agnostic modeling, the extension of molecular-scale discoveries to clinical practice validation, and the precise docking of computational tools with biomedical semantics.

At the Patient System Translation layer, the development of MSaaVS (Chapter 2) marks a milestone in the automated structured modeling of patients' subjective narrative information. It realizes the initial automated conversion of patients' natural language expressions into standardized intent labels that are operable by the medical system, greatly improving the efficiency of docking patient semantics with medical logic (Koleck et al. 2019; Gonzalez-Hernandez et al. 2017). On this basis, PHPlace (Chapter 3) deepens this concept into the field of POP health management. Through dynamic interactive questionnaires and innovative 2D animation modules, it not only optimizes the structuring of patient information, but also enhances the communication efficiency between doctors and patients and the patients' understanding of symptoms and treatment plans through intuitive visualization (Alsos, Das, and Svanæs 2012; Shabir and Kenzie 2025; Parmar and I. K. Verma 2024). The contribution of this level is that it has initially solved the problem of objectification of subjective experience, and it demonstrates the great potential of multimodal computational semantic modeling in improving the intuitiveness of medical communication and promoting personalized medical services.

The focus of the Data Diagnosis Translation layer is to extract interpretable diagnostic insights from heterogeneous clinical data, especially in scenarios where data is scarce. Strengthening Diagnostic Translation (Chapter 4) effectively improves the generalization and robustness of the model in a small sample environment by introducing a semantic consistency constraint mechanism while ensuring data authenticity (J. Zhou et al. 2021; Y. Zhou et al. 2024).

At the Molecular Clinical Translation layer, we are committed to bridging the gap between molecular data and clinical practical applications. The EPOP (Chapter 5) uses a protein language model to fine-tune POP-related proteomics data, achieving an interpretable mapping from protein sequences to disease-associated functional labels. This allows molecular discoveries to no longer remain at the sequence level, but to be transformed into information that is of practical clinical significance. As a theoretical supplement to this layer, Protein Design Enters the AI Era (Chapter 6) critically examines the current focus of AI protein design research, emphasizing that future research needs to more deeply integrate biological semantics and application scenario relevance, providing

important theoretical guidance for the transformation of molecular insights into clinical practice.

The Protein Infrastructure Translation layer builds a basic platform to support the transformation of protein representation into reusable, generative, and task-oriented computing assets. PROFASA (Chapter 7) provides an intuitive interactive visualization environment that simplifies the protein fragment analysis and property evaluation process, improving the efficiency of researchers' understanding of molecular mechanisms. PS-GO (Chapter 8), as an efficient parameterized protein search engine, realizes accurate retrieval based on user-defined biophysical properties, greatly improving the flexibility of function-driven protein structure discovery. On this basis, Silver Surfer (Chapter 9) realizes the automatic generation of new protein sequences through innovative genetic algorithms, and can perform targeted optimization under multiple biophysical property constraints. Finally, ProteinFlow (Chapter 10), as a comprehensive large-scale protein feature engineering platform, provides standardized, highly scalable core data and feature buses for upper-level protein computing, design, and diagnostic applications. These three systems together build the core of the basic computing layer of this study, providing solid support for a variety of protein-related tasks.

The power of this framework lies in its strict adherence to and deep practice of core design principles, which are applied throughout the design and implementation of systems at all levels:

- **Semantic continuity** is not a simple transfer of information, but rather ensuring that the integrity and interpretability of medical or biological meanings are preserved when knowledge is transformed between different levels of abstraction. MSaaVS (Chapter 2) not only formalizes patient narratives, but also strives to capture their original intent, and EPOP (Chapter 5) associates abstract protein sequences with specific disease functions to ensure that molecular-level findings can be understood and applied by clinicians. This extreme pursuit of semantic fidelity is the cornerstone of this framework for achieving effective knowledge transformation.
- The **modular structure** is reflected in the high degree of independence and plug-gability of each system. Each module, such as MSaaVS (Chapter 2) and PHPlace (Chapter 3), can run independently and carry specific translation functions. This

not only greatly improves the development efficiency and maintainability of the system, but also provides the possibility for flexible combination, expansion, and replacement according to new scientific research or clinical needs in the future, ensuring the adaptability and vitality of the framework.

- **Interoperability** is the key to connecting modules and achieving knowledge flow. By developing a unified data exchange protocol and a shared semantic representation paradigm, we have initially achieved smooth information transfer across systems. As a unified feature platform for protein data, ProteinFlow (Chapter 10) can provide standardized input for all protein-related computing modules in the upper layer, and the conceptual alignment of MSaaVS (Chapter 2) and PHPlace (Chapter 3) in intent classification also lays the foundation for the seamless transition of patient information to diagnostic models. Although fully automated end-to-end interoperability is still under exploration, existing practices have proven its feasibility.
- **Interpretability and operability** are the embodiment of the deep integration of this framework with clinical practice. The 2D animation and interactive design of PHPlace (Chapter 3) significantly improve patients' understanding and participation in complex medical information. In protein design, PROFASA (Chapter 7), PS-GO (Chapter 8), and Silver Surfer (Chapter 9) allow researchers to accurately query and generate based on biophysical properties, making the abstract molecular design process controllable and understandable, thereby ensuring the biological rationality and engineering practicality of the final design.

From a holistic perspective, the system constructed in this paper does not solve a problem at a certain level in isolation, but forms a "cognition-computation-operation" continuum. Among them, "cognition" represents our deep understanding of the nature and challenges of the problem; "computation" refers to how we model these problems and design an operational system architecture; and "operation" is reflected in the deployment, use, and potential feedback mechanisms of these systems in real data. However, in the current implementation, this continuum has not yet been completely closed. For example, the data flow between systems has not yet been fully automatically connected, and the reasoning path of the semantic layer lacks unified specifications. However,

this structured design concept has initially proved its value. It emphasizes that when facing the problem of multi-level transformation of biomedical knowledge, we advocate a systematic and combined solution rather than relying on a single general solution. This way of thinking also directly responds to the philosophical discussion of "translation" at the beginning of Chapter 1: translation is not just the replacement of symbols, but also the reconstruction of knowledge structure. In contemporary biomedical practice, a significant gap exists between patient expression, doctor judgment, model prediction, and biological mechanisms. This study aims to bridge these critical gaps through a scalable and stable approach using the powerful method of computational translation.

11.1 Limitations

Although the computational translation framework proposed in this study has made significant progress and is comprehensive, its inherent limitations also define the scope of the current work and provide clear guidance for future research directions.

11.1.1 Semantic alignment

Although various systems have achieved preliminary semantic structuring and mapping within specific modalities (such as natural language, structured data, and molecular sequences), a truly unified and deeply integrated cross-modal semantic understanding framework remains an unsolved problem (Duan et al. 2024; H. Zhou et al. 2024; Rajendran et al. 2024). Current mapping mechanisms rely heavily on manually defined rules or statistical associations, and it is difficult to fully capture and transform the implicit, context-related deep semantics between different modalities. For example, the psychological, cultural, and socioeconomic background information contained in the patient's subjective language, although crucial to clinical decision-making, is not yet fully and accurately captured and conveyed by current symbolization or question-answering systems. Future work needs to break through the limitations of this reliance on surface associations and move towards more reasoning-capable and generalized semantic understanding.

11.1.2 Causal inference capabilities

Although many systems in this framework excel in prediction and association discovery, most are still based on correlation rather than causal modeling (S.-H. Lin and Ikram 2020; Kleinberg and Hripcsak 2011). When identifying disease-related proteins, the EPOP (Chapter 5) system uses advanced protein language models. However, its internal "black box" characteristics make it difficult for us to provide a clear mechanistic explanation for "why a protein is identified as relevant." This limitation in causal explanation capabilities restricts the model's upgrade from pure association prediction to a more insightful understanding of mechanisms, thereby affecting its ability to provide deeper and more reliable insights into clinical decision-making and basic scientific research.

11.1.3 Automated linkage between systems

Although this framework emphasizes modular design and conceptual interoperability, fully automated data flow and task scheduling between systems have not yet been achieved (T. Ma and A. Zhang 2016; Tiwari and Sekhar 2007; Abdi et al. 2024; Banimfreg 2023). Currently, data usually needs to be manually exported, converted, and imported between different modules, or follow a preset, relatively fixed pipeline. This limits the adaptability and dynamics of the framework as a general platform, making it difficult to automatically orchestrate complex translation tasks according to real-time needs. For example, automatically triggering subsequent molecular target analysis from diagnostic results, or seamlessly starting the protein design process from molecular predictions, these complex end-to-end closed loops are still mainly at the architectural design level, rather than fully automated operation practices.

11.1.4 Domain generalization and adaptability

Although the proposed hierarchical architecture has a certain degree of universality, the current system implementation is mainly focused on two specific areas: female pelvic floor health management and protein engineering. Directly applying this framework and its specific modules to other biomedical fields (such as oncology, neuroscience, or infectious diseases) will face significant challenges and require a lot of customized development and verification work for new data types, unique semantic systems, and task

requirements, which shows that its current universality still needs to be further expanded (Kouw and Loog 2018; Singhal et al. 2023).

11.1.5 Non-computational factors

This study focuses on the problem of biomedical knowledge translation solved by computational modeling and structured methods. However, in real-world clinical settings, the success of translation is also affected by many non-computational factors, including but not limited to the medical system, ethical norms, patient compliance, social and cultural background, and the trust relationship between doctors and patients (Holden et al. 2013; De Miranda, Doggett, and J. T. Evans 2005). This framework does not deeply explore or integrate the impact of these non-computational factors on knowledge translation, which to some extent limits the complexity and challenges it may face in actual application scenarios.

11.2 Future Work

Although the "translation as infrastructure" framework proposed in this study has laid a solid foundation, its future development still holds great potential, especially in addressing current limitations and enhancing the intelligence and universality of the framework. The following will explore the key directions of future research in more depth from multiple dimensions.

In terms of semantic alignment and knowledge reasoning, although we have achieved preliminary semantic structuring in specific modalities, it remains imperative to build a truly unified, cross-modal, and reasoning-capable biomedical semantic understanding framework. Future work should focus on developing a multi-level, dynamically evolving biomedical knowledge graph. Through large-scale corpus learning, weakly supervised semantic extraction, and expert collaborative verification, a semantic network that can dynamically capture the complex relationships between biomedical concepts (including causality, hierarchy, attributes) is constructed. This requires breaking through the boundaries of traditional symbolism and connectionism and exploring cutting-edge technologies such as neural symbolic reasoning, so that the model can understand semantic ambiguity and contextual dependence while performing precise logical reasoning

11 Discussion

(Nallabantu and Guruprakash 2024). For example, by introducing a large language model enhanced by a knowledge graph, we can train the model to understand better the deep medical intent contained in the patient's colloquial narrative, and seamlessly connect it with structured clinical data and molecular-level knowledge, thereby achieving more accurate disease diagnosis assistance and personalized treatment plan recommendations. In addition, studying how to explain complex computational results in a natural language form that is easy for patients or clinicians to understand through reverse semantic mapping is also a key step in improving the operability of the framework.

In response to the challenges of causal inference and explainable artificial intelligence, future research should fundamentally enhance the framework's ability to transition from association prediction to mechanistic insight. This requires us to be no longer satisfied with statistical correlation, but to introduce a systematic causal reasoning framework. Specifically, we can explore the construction of structural causal models at different translation levels, and identify and quantify the causal relationship between variables from massive observational data by combining domain expert knowledge and advanced causal discovery algorithms. For example, at the data diagnosis translation level, we can use causal graph analysis to not only identify factors associated with disease risk, but also reveal their causal chain, thereby guiding more effective intervention measures. At the molecular clinical translation level, by introducing causal representation learning, even black box models based on deep learning, such as the EPOP (Chapter 5), can learn feature representations with causal invariance at a higher level of abstraction, thereby explaining how specific molecular features mediate the occurrence and development of diseases, rather than just providing prediction results (Schölkopf 2022; Pearl 2022). In addition, the development of interactive explainable artificial intelligence tools that allow clinicians and researchers to explore the model reasoning process from global model interpretation to local decision attribution, and to evaluate the robustness and potential impact of decisions by simulating counterfactual scenarios, will be an important way to improve model trust and clinical adoption.

In terms of building automation, future work should focus on integrating the current decentralized, modular systems into a unified, highly automated computing infrastructure that is schedulable on demand and supports multi-task collaboration. This requires the introduction of advanced cloud-native and microservice architecture concepts. Specifically, each independent system is decoupled into independent, independently deployable and

scalable microservices, and managed using containerization technology. The core lies in the development of an intelligent workflow orchestration engine that can automatically select the most appropriate microservice based on the complex translation tasks defined by the user, and intelligently manage data flow, resource allocation, and task dependencies. For example, an end-to-end treatment plan design request can begin with the capture of patient symptoms, automatically flow through the diagnostic model, and then trigger molecular target analysis, ultimately recommending and generating candidate protein sequences with specific functions. To achieve this goal, it is also necessary to establish a unified, high-throughput, and semantically labeled data bus to ensure that data can be efficiently and seamlessly exchanged and understood between different microservices. At the same time, the introduction of event-driven architecture enables services to communicate and collaborate through asynchronous messaging mechanisms, thereby improving the system's elasticity and responsiveness, and ultimately building a platform that can support on-demand calls in multiple scenarios, such as clinical decision-making, drug discovery, and basic research (P. Raj, Vanga, and Chaudhary 2022; Nivedhaa 2024).

Finally, in terms of domain generalization and the expansion of clinical validation paths, future research should make the framework more universal, allowing it to be efficiently applied to a broader range of biomedical problems and gradually promote its validation in real clinical environments. This means not only applying the framework to new disease areas, but more importantly, systematically studying the capture, representation, and translation challenges of data modalities unique to these new fields (such as medical imaging genomics, genomics, metabolomics, electronic health records, wearable device data, etc.). For these new data types, it is necessary to develop corresponding pre-processing, multimodal fusion, and cross-layer translation modules. At the same time, by introducing technologies such as transfer learning, meta-learning, and federated learning, the knowledge acquired in existing fields can be leveraged to accelerate the adaptation and validation of the framework in new fields, thereby reducing data privacy barriers. More importantly, it is necessary to establish close cooperation with medical institutions, conduct multicenter, prospective clinical pilot projects, and rigorously validate the clinical systems implemented within the framework. This validation should quantify their actual value in improving diagnostic accuracy, optimizing treatment plans, enhancing patient prognosis, and reducing medical costs. This kind of translational validation from

11 Discussion

the laboratory to the bedside will be a key step for "translation as infrastructure" to move from theory to practice and from prototype to widespread application.

In summary, this study focuses on the computational translation of "molecule to clinic" as its primary objective, builds a system that spans multiple levels, and verifies it in practice across various challenging scenarios. This work not only provides specific and operational technical solutions but also fundamentally demonstrates the feasibility and significant value of "translation as infrastructure" as a new research paradigm. Although the road to a fully automated and universally applicable translation infrastructure is still ongoing, the structural foundation laid by this paper, along with the demonstrated cross-layer concept transfer capabilities, has achieved initial and encouraging results, paving a solid and promising path for future research and continued evolution in the field of biomedical knowledge translation.

12 Conclusion

This study focuses on "From Molecules to Medicine". It systematically explores how to build a structured, computable, and scalable computing infrastructure in the multi-source, multi-layer, and multi-task medical translation process. After building and experimenting with many underlying systems, we proposed the concept of "Translational Computing Infrastructure" as a medium to connect cognitive structures and operating systems, in response to the fractures and mismatches in the flow of biomedical knowledge.

The contributions of this study are mainly reflected in the following aspects. First, at the theoretical level, this study identified four key challenges in the transformation of biomedical knowledge, corresponding to patient-system interface, data-diagnostic modeling, molecular-clinical mechanism, and tool-semantic docking. This four-layer structure not only reveals the complexity of the transformation problem, but also provides a structured guide for subsequent system design. Moreover, at the method level, we proposed a research paradigm of "structure-driven, cross-layer collaboration". This paradigm emphasizes that under the premise of respecting semantic complexity, the distributed modeling of complex medical processes is achieved through data interfaces, feature abstraction and system integration with computable modules as units. In addition, at the system implementation level, we have built a complete platform system covering "medical dialogue-diagnostic reasoning-disease prediction-molecular verification-protein exploration". These systems not only show good performance in experiments, but also preliminarily demonstrate the ability of data migration and method migration across task boundaries. Finally, in terms of method promotion, the PS-GO system integrates many of the above achievements and verifies the possibility of synergy between feature extraction, visual search and directional generation in actual protein exploration tasks. It demonstrates that when the basic feature framework is general enough, researchers can move from single-point modeling to a property-oriented, goal-driven protein design path.

12 Conclusion

Although this study has made a systematic exploration on the path of "computational translation", we are also clearly aware that it is still in its early stages of development. In particular, the following aspects still need to be further deepened: there is still a lack of a unified modeling framework at the semantic level, especially from natural language expression to structured computational graphs, a stable and highly versatile semantic bridging mechanism needs to be built; causal reasoning and explanation transparency are still relatively weak, and most systems are still driven by correlation, lacking clear mechanism paths and cross-modal causal links; automated scheduling and data circulation have not yet been achieved between modules, and further system integration work is still needed to achieve a truly "unified platform".

Future research can continue to advance around the above shortcomings, especially in the areas of semantic modeling and causal structure. In addition, we also plan to expand the framework to more application scenarios, such as chronic disease management, precision medicine, and automated scientific discovery platforms, to promote it from feasibility verification to general infrastructure.

In conclusion, the "translational computing infrastructure" proposed in this thesis is both a conceptual paradigm shift and an integrated expression of a series of implemented systems. It does not attempt to solve all problems with one model, but is committed to building a future-oriented technical foundation that enables multidimensional knowledge to be efficiently translated, integrated, and activated. We believe that this "structure-driven, semantically aligned, model combination" approach will provide a solid foundation for future biomedical artificial intelligence and breed a truly intelligent biomedical ecosystem in the process of continuous evolution.

Appendices

A Clinical Research Ethics Committee of the Cork Teaching Hospitals

This appendix includes the official letter of ethical approval issued by the CREC of the Cork Teaching Hospitals, corresponding to the PHPlace platform described in Chapter 3.

COISTE EITICE UM THAIGHDE CLINICIÚIL
Clinical Research Ethics Committee of the Cork Teaching Hospitals

Tel: +353-21-4901901
Email: crec@ucc.ie

University College Cork
Lancaster Hall
6 Little Hanover Street
Cork
Ireland

CREC Review Reference Number: ECM 4 (r) 24/10/2023
& ECM 4 (cc) 05/12/2023

Date: 29th November 2023

Professor Barry O'Reilly
Consultant Obstetrician and Gynaecologist
Cork University Maternity Hospital
Wilton
Cork

Study Title: Revolutionizing Urogynecology Care: Implementation and Assessment of the 'UroGynius' Mobile Application and Website.

Dear Professor O'Reilly

The following documents have been approved:

Study Protocol	Version 2
Information Leaflet	Version 2
Screenshots	
Data Protection Officer Approval	Dated 22 nd November 2023.

Full approval is now granted to carry out the above study. The date of this letter is the date of authorization of the study.

Please keep a copy of this signed approval letter in your study master file for audit purposes. The study must be carried out in accordance with General Data Protection Regulation and Health Research Regulation 2018/2021.

You should note that ethical approval will lapse if you do not adhere to the following conditions:

1. Submission of an Annual Progress Report/Annual Renewal Survey (due annually from the date of this approval letter). **We would encourage you to keep note of this date as the CREC will not issue a reminder.**
2. Report unexpected adverse events, serious adverse events or any event that may affect ethical acceptability of the study
3. Submit any change to study documentation (minor or major) to CREC for review and approval. Amendments must be submitted on an amendment application form and revised study documents must clearly highlight the changes and contain a new version number and date. Amendments cannot be implemented without written approval from CREC.
4. Notify CREC of discontinuation of the study
5. Submit an End of Trial Declaration Form and Final Study Report/Study Synopsis when the study has been completed.

Yours sincerely



Professor David Kerins
Chairman
Clinical Research Ethics Committee
of the Cork Teaching Hospitals

Figure A.1: Ethical approval letter from the CREC of the Cork Teaching Hospitals for the PHPlace platform.

B PROFASA User Guide

This appendix provides a user manual for the PROFASA (Chapter 8) platform.

1. Registration and Login

To access the features of PROFASA, you must first create an account. Navigate to the top right corner of the homepage and click on the **SIGN UP** button. You will be redirected to a registration page. Here, you will be required to fill in your email address and a password. After filling in these details, click on the **Register** button at the bottom of the page. An email will be sent to the provided address for account confirmation. Click on the link in the email to verify your account.

To log in, click on the **SIGN IN** button located at the top right corner of the homepage. Enter the email address and password you used during registration and click **Sign In**.

2. Using the BLAST Feature

The BLAST feature allows you to find known proteins that are similar to your protein sequence. From the homepage, click on the **BLAST** button. You will be directed to the BLAST page where you can input your protein sequence in the provided field. After entering your sequence, click **Submit**. The system will automatically perform a BLAST analysis and display the results on a new page. The results include the names, sequences, and similarity scores of known proteins that are similar to your sequence.

3. Using the FOLD Feature

The FOLD feature enables you to predict the three-dimensional structure of your protein sequence. From the homepage, click on the **FOLD** button. You will be directed to the

FOLD page where you can input your protein sequence in the provided field. After entering your sequence, click **Submit**. The system will automatically perform a protein folding prediction and display the results on a new page. The results include images and parameters of the predicted protein structure.

4. Using the ANALYSIS Feature

The ANALYSIS feature allows you to analyze various properties of your protein sequence. From the homepage, click on the **ANALYSIS** button. You will be directed to the ANALYSIS page where you can upload your PDB file in the provided field. After uploading your file, click **Submit**. The system will automatically perform a protein analysis and display the results on a new page. The results include the length, molecular weight, isoelectric point, solubility, and other properties of your protein sequence.

5. Using the SUPERIMPOSE Feature

The SUPERIMPOSE feature enables you to compare two or more protein structures. From the homepage, click on the **SUPERIMPOSE** button. You will be directed to the SUPERIMPOSE page where you can input the protein sequences you want to compare in the provided fields. After entering your sequences, click **Submit**. The system will automatically perform a protein superimposition and display the results on a new page. The results include images and parameters of the superimposed protein structures.

6. Viewing and Managing Your Tasks

From the homepage, click on the **My Results** button. You will be directed to the **My Results** page where you can view a list of all the tasks you have created. Each task has a unique ID, name, type, sequence length, model generation progress, and submission time. You can click on the name of each task to view its details.

You can also share tasks with collaborators. Simply click on the **Share** button next to each task and enter the email address of your collaborator. They will then be able to view your task.

7. Viewing and Downloading Protein Model Information

After a task is completed, you can view and download basic information about the protein model on the **Text Result Page**. This information includes the sequence, RSCB PDB, Ramachandran Plot, F2F Plot, and more.

Once all protein models for the task have been generated, you can switch to the **3D Result Page** through the button in the upper right corner.

8. Manipulating Protein Models on the 3D Result Page

On the **3D Result Page**, you can view the three-dimensional structure of the protein model. The model can be manipulated by rotating and zooming in/out to view the structure from different angles.

In addition to viewing the basic information of the protein model, you can also change the **Representation** and **Color** of the entire model or part of the model. This is done by clicking on the **Representation** and **Color** buttons on the page and selecting the desired options.

The **SUPERIMPOSE** function is also available on this page. It can be used to rotate and translate a structure to match or align it with another structure or structures. The system then uses RMSD and GDT_TS to evaluate the quality of the superimposition.

Support

This concludes the detailed user guide for the PROFASA platform. If you encounter any issues while using the platform, you can reach out to the support team at protein.blast.mi@gmail.com.

C PS-GO User Manual

This appendix provides a user manual for the PS-GO (Chapter 9) platform.

Objective

Search for a complex protein using natural language, read its parameters, modify some amino acids, and observe the changes in parameters.

Steps

1. In the search bar, enter a description of the protein you are looking for. For example: "I'm looking for a complex protein with very large size. I also want this protein to have a low Isoelectric Point." Click the "Search" button. Alternatively, you can click on the parameter buttons below the search bar to perform a more precise search based on specific criteria. As shown in Figure C.1



Figure C.1: PS-GO User Manual Step 1

2. A list of search results will be displayed. As shown in Figure C.2

C PS-GO User Manual

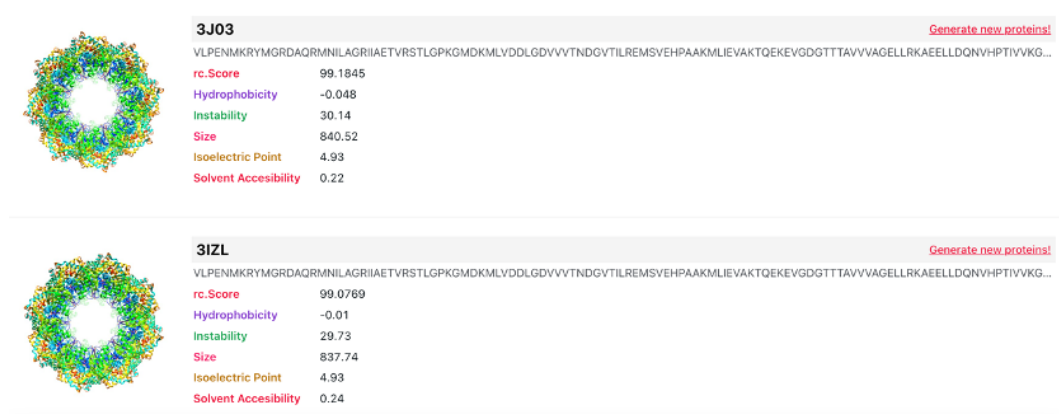


Figure C.2: PS-GO User Manual Step 2

3. Click on the first search result to enter the protein details page, observe and take note of the protein's parameters. As shown in Figure C.3
4. Return to the search results page and click on the "Generate new proteins!" button.
5. After a short wait, you will see the parameters of the protein with a few amino acid mutations. As shown in Figure C.4
6. You can also choose to freeze certain amino acids, preventing them from being mutated. As shown in Figure C.5
7. Continue exploring different proteins and their mutated versions to explore protein parameter optimization and understand how changes in the amino acid sequence affect the protein's properties.

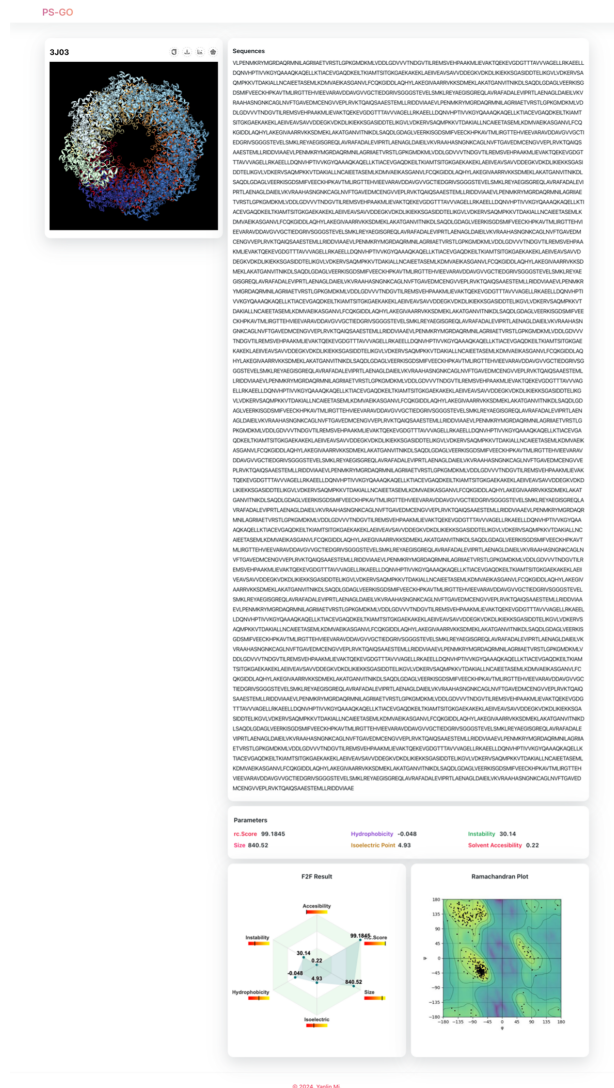


Figure C.3: PS-GO User Manual Step 3

Instability: 32.868520875762655
Size: 84984.98146500139
Hydrophobicity: -0.0892311608961304
Isoelectric: 5.16064453125

Figure C.4: PS-GO User Manual Step 5

Generate New Proteins

Sequence

VL PENMK RYMGRDAQR MNILAGRIIAETVRSTLGP KGMDKMLVDDLGDVVVTNDGVTILREMSVEHPAAKMLIEVAKTQEKEVG DTTTAVVVAGELLRKAEELLDQNVHPTIVVKG YQAAQKAQEL
⌵

Generate

Sequence Freeze

Freeze
Unfreeze
Reset
Freeze

V	L	P	E	N	M	K	R	Y	M	G	R	D	A	Q	R	M	N	I	L	A	G	R	I	I	A	E	T	V	R	S	T	L	G	P	K	G	M	D	K	M	L	V	D	D	L	G	D	V	V	V	T	N	D	G	V	T
I	L	R	E	M	S	V	E	H	P	A	A	K	M	L	I	E	V	A	K	T	Q	E	K	E	V	G	D	G	T	T	T	A	V	V	V	A	G	E	L	L	R	K	A	E	E	L	D	Q	N	V	H	P	T	I	V	
V	K	G	Y	Q	A	A	A	Q	K	A	Q	E	L	L	K	T	I	A	C	E	V	G	A	Q	D	K	E	I	L	T	K	I	A	M	T	S	I	T	G	K	G	A	E	K	A	K	E	K	L	A	E	I	I	V	E	A
V	S	A	V	D	D	E	G	K	V	D	K	D	L	I	K	I	E	K	K	S	G	A	S	I	D	D	T	E	L	I	K	G	V	L	V	D	K	E	R	V	S	A	Q	M	P	K	K	V	T	D	A	K	I	A	L	
L	N	C	A	I	E	E	T	A	S	E	M	L	K	D	M	V	A	E	I	K	A	S	G	A	N	V	L	F	C	Q	K	G	I	D	D	L	A	Q	H	Y	L	A	K	E	G	I	V	A	A	R	R	V	K	K	S	D
M	E	K	L	A	K	A	T	G	A	N	V	I	T	N	I	K	D	L	S	A	Q	D	L	G	D	A	G	L	V	E	E	R	K	I	S	G	D	S	M	I	F	V	E	E	C	K	H	P	K	A	V	T	M	L	I	R
G	T	T	E	H	V	I	E	E	V	A	R	A	V	D	D	A	V	G	V	G	C	T	I	E	D	G	R	I	V	S	G	G	G	S	T	E	V	E	L	S	M	K	L	R	E	Y	A	E	G	I	S	G	R	E	Q	
L	A	V	R	A	F	A	D	A	L	E	V	I	P	R	T	L	A	E	N	A	G	L	D	A	I	E	I	L	V	K	V	R	A	A	H	A	S	N	G	N	K	C	A	G	L	N	V	F	T	G	A	V	E	D	M	C
E	N	G	V	V	E	P	L	R	V	K	T	Q	A	I	Q	S	A	A	E	S	T	E	M	L	L	R	I	D	D	V	I	A	A	E	V	L	P	E	N	M	K	R	Y	M	G	R	D	A	Q	R	M	N	I	L	A	G
R	I	I	A	E	T	V	R	S	T	L	G	P	K	G	M	D	K	M	L	V	D	D	L	G	D	V	V	V	T	N	D	G	V	T	I	L	R	E	M	S	V	E	H	P	A	A	K	M	L	I	E	V	A	K	T	Q
E	K	E	V	G	D	G	T	T	T	A	V	V	V	A	G	E	L	L	R	K	A	E	E	L	D	Q	N	V	H	P	T	I	V	V	K	G	Y	Q	A	A	A	Q	K	A	Q	E	L	L	K	T	I	A	C	E	V	
G	A	Q	D	K	E	I	L	T	K	I	A	M	T	S	I	T	G	K	G	A	E	K	A	K	E	K	L	A	E	I	I	V	E	A	V	S	A	V	V	D	D	E	G	K	V	D	K	D	L	I	K	I	E	K	S	

Figure C.5: PS-GO User Manual Step 6

D Other Published Works during PhD

During the PhD period, I also participated in a range of research works beyond the core systems discussed in Chapters 2 to 11. This appendix lists the publications:

Stefan-Bogdan Marcu, **Yanlin Mi**, Mark Tangney, and Sabin Tabirca. (2025). *ChemFlow: Bridging Computational Tools and Bioinformatics Through LLM-Driven Workflow Automation*. In Proceedings of the International Conference on Modelling and Development of Intelligent Systems (MDIS 2024), Communications in Computer and Information Science, vol 2486, pp. 123–137.

DOI:10.1007/978-3-031-87386-7_9

Publisher: Springer, Cham

This paper presents ChemFlow, a proof-of-concept platform that uses large language models and prompt engineering to automatically generate and execute bioinformatics workflows from natural language descriptions. My role involved contributing to conceptual discussions and feedback during manuscript preparation.

Reut Rotem, Daniel Galvin, Yair Daykan, **Yanlin Mi**, Sabin Tabirca, and Barry A. O'Reilly. (2024). *Revolutionizing urogynecology: Machine learning application with patient-centric technology*. European Journal of Obstetrics & Gynecology and Reproductive Biology, 300: 49–53.

DOI:10.1016/j.ejogrb.2024.07.009

Publisher: Elsevier

JCR Quartile: Q3

Impact Factor: 2.1

This review introduces an AI-driven perspective on pelvic floor dysfunction and its management, summarizing advances in diagnostic imaging, digital health platforms,

D Other Published Works during PhD

and ML enhanced surgical decision-making. My contribution involved system-level insight from the PHPlace project, bridging applications with real-world urogynecology workflows.

Isabel Ronan, **Yanlin Mi**, Venkata Vamsi Bharadwaj Yallapragada, Cárthach Ó Nuanáin, and Sabin Tabirca. (2024). *A Machine Learning Approach to 3D Protein Structure Sonification*. International Journal of Music Science, Technology and Art (IJMSTA), 6(1): 9–18.

DOI:10.48293/IJMSTA-106

Publisher: Music Academy “Studio Musica”

This work explored how features of protein 3D structures can be translated into sound using ML, enabling auditory analysis of structural properties. I contributed by providing bioinformatics and protein structure expertise throughout the project, and was actively involved in reviewing and refining the manuscript during its preparation and submission.

Stefan-Bogdan Marcu, **Yanlin Mi**, Venkata V.B. Yallapragada, Mark Tangney, and Sabin Tabirca. (2023). *Generating Jigsaw Puzzles and an AI Powered Solver*. In Proceedings of the International Conference on Modelling and Development of Intelligent Systems (MDIS 2022), Communications in Computer and Information Science, vol 1761, pp. 148–160.

DOI:10.1007/978-3-031-27034-5_10

Publisher: Springer Nature Switzerland

This paper presents a system combining rule-based puzzle generation with deep reinforcement learning for solving, using vision and edge-matching strategies. My role involved contributing to early-stage conceptual discussions and feedback during manuscript preparation.

References

- Abdel-Halim, Heba (2020). “Distant learning challenges and solutions: Incorporation of 3D protein visualisation in an undergraduate pharmacy medicinal chemistry course: Innovation in teaching delivery or learning technology”. In: *Pharm. Educ* 20, pp. 17–18 (Cited on page 138).
- Abdi, Ghomareza et al. (2024). “Unveiling the dynamic role of bioinformatics in automation for efficient and accurate data processing and interpretation”. In: *Advances in Bioinformatics*. Springer, pp. 279–319 (Cited on page 226).
- Abdi Reyhan, Masoumeh et al. (2021). “Development of a core data set for pelvic floor disorder patients registry”. In: *LUTS: Lower Urinary Tract Symptoms* 13.1, pp. 144–153 (Cited on page 64).
- Abgrall, Gwénolé et al. (2024). “Should AI models be explainable to clinicians?” In: *Critical Care* 28.1, p. 301 (Cited on page 3).
- Abramson, Josh et al. (2024). “Accurate structure prediction of biomolecular interactions with AlphaFold 3”. In: *Nature*, pp. 1–3 (Cited on page 112).
- Achiam, Josh et al. (2023). “Gpt-4 technical report”. In: *arXiv preprint arXiv:2303.08774* (Cited on page 173).
- Aderinwale, Tunde et al. (2022). “Real-time structure search and structure classification for AlphaFold protein models”. In: *Communications biology* 5.1, p. 316 (Cited on page 167).
- Adhikari, Badri, Jie Hou, and Jianlin Cheng (2018). “DNCON2: improved protein contact prediction using two-level deep convolutional neural networks”. In: *Bioinformatics* 34.9, pp. 1466–1472 (Cited on page 139).
- Agarwal, Vinayak and Andrew C McShan (2024). “The power and pitfalls of AlphaFold2 for structure prediction beyond rigid globular proteins”. In: *Nature Chemical Biology* 20.8, pp. 950–959 (Cited on page 123).

References

- Agu, Peter Chinedu and Chidiebere Nwiboko Obulose (2024). “Piquing artificial intelligence towards drug discovery: Tools, techniques, and applications”. In: *Drug Development Research* 85.2, e22159 (Cited on page 123).
- Ahmed, Awais et al. (2023). “Harnessing big data analytics for healthcare: A comprehensive review of frameworks, implications, applications, and impacts”. In: *IEEE Access* 11, pp. 112891–112928 (Cited on page 7).
- Alamdari, Sarah et al. (2023). “Protein generation with evolutionary diffusion: sequence is all you need”. In: *BioRxiv*, pp. 2023–09 (Cited on page 117).
- Alarab, May et al. (2014). “Expression of extracellular matrix-remodeling proteins is altered in vaginal tissue of premenopausal women with severe pelvic organ prolapse”. In: *Reproductive Sciences* 21.6, pp. 704–715 (Cited on page 100).
- Alford, Rebecca F. et al. (June 2017). “The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design”. en. In: *Journal of Chemical Theory and Computation* 13.6, pp. 3031–3048. ISSN: 1549-9618, 1549-9626. DOI: 10.1021/acs.jctc.7b00125. URL: <https://pubs.acs.org/doi/10.1021/acs.jctc.7b00125> (visited on 05/07/2023) (Cited on pages 163, 164).
- Ali, Abdulrazzaq Qasem et al. (2019). “A systematic mapping study on the customization solutions of software as a service applications”. In: *IEEE Access* 7, pp. 88196–88217 (Cited on page 29).
- Allouche, David et al. (2014). “Computational protein design as an optimization problem”. In: *Artificial Intelligence* 212, pp. 59–79 (Cited on page 10).
- Alperin, Marianna et al. (2016). “Pregnancy-induced adaptations in intramuscular extracellular matrix of rat pelvic floor muscles”. In: *American journal of obstetrics and gynecology* 215.2, 210–e1 (Cited on page 75).
- AlQuraishi, Mohammed (2019). “End-to-end differentiable learning of protein structure”. In: *Cell systems* 8.4, pp. 292–301 (Cited on page 122).
- Alsos, Ole Andreas, Anita Das, and Dag Svanæs (2012). “Mobile health IT: The effect of user interface and form factor on doctor–patient communication”. In: *International journal of medical informatics* 81.1, pp. 12–28 (Cited on page 222).
- Angelo, Jaqueline S et al. (2023). “Multi-and many-objective optimization: present and future in de novo drug design”. In: *Frontiers in Chemistry* 11, p. 1288626 (Cited on page 134).

- Anishchenko, Ivan et al. (2021). “De novo protein design by deep network hallucination”. In: *Nature* 600.7889, pp. 547–552 (Cited on pages 112, 122).
- Anurag, VN Nikhil (2018). *Distributed computing with Go: practical concurrency and parallelism for Go applications*. Packt Publishing Ltd (Cited on page 207).
- Arya, Nikhilanand and Sriparna Saha (2021). “Multi-modal advanced deep learning architectures for breast cancer survival prediction”. In: *Knowledge-Based Systems* 221, p. 106965 (Cited on page 115).
- Baek, Minkyung et al. (2021). “Accurate prediction of protein structures and interactions using a three-track neural network”. In: *Science* 373.6557, pp. 871–876 (Cited on page 123).
- Bandara, Kasun et al. (2021). “Improving the accuracy of global forecasting models using time series data augmentation”. In: *Pattern Recognition* 120, p. 108148 (Cited on page 66).
- Banimfreg, Bayan H (2023). “A comprehensive review and conceptual framework for cloud computing adoption in bioinformatics”. In: *Healthcare Analytics* 3, p. 100190 (Cited on page 226).
- Bawono, Punto et al. (2017). “Multiple sequence alignment”. In: *Bioinformatics: volume I: data, sequence analysis, and evolution*, pp. 167–189 (Cited on page 205).
- Bella, Jordi et al. (1994). “Crystal and molecular structure of a collagen-like peptide at 1.9 Å resolution”. In: *Science* 266.5182, pp. 75–81 (Cited on page 96).
- Bennett, Karen (2024). “Epistemic translation: towards an ecology of knowledges”. In: *Perspectives*, pp. 1–16 (Cited on pages 1–3).
- Bepler, Tristan and Bonnie Berger (2021). “Learning the protein language: Evolution, structure, and function”. In: *Cell systems* 12.6, pp. 654–669 (Cited on pages 109, 118).
- Berman, Helen M et al. (2000). “The protein data bank”. In: *Nucleic acids research* 28.1. Publisher: Oxford University Press, pp. 235–242. ISSN: 0305-1048 (Cited on page 166).
- Bhagwat, Medha and L Aravind (2008). “Psi-blast tutorial”. In: *Comparative genomics*, pp. 177–186 (Cited on page 183).
- Bierman, Gavin, Martín Abadi, and Mads Torgersen (2014). “Understanding typescript”. In: *ECOOP 2014—Object-Oriented Programming: 28th European Conference, Upp-*

References

- sala, Sweden, July 28–August 1, 2014. *Proceedings* 28. Springer, pp. 257–281 (Cited on page 143).
- Bini, Stefano A (2018). “Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care?” In: *The Journal of arthroplasty* 33.8, pp. 2358–2361 (Cited on page 62).
- Biotechnology Information (US), National Center for and Christiam Camacho (2008). *BLAST (r) command line applications user manual*. National Center for Biotechnology Information (US) Bethesda, MD (Cited on page 148).
- Boadu, Frimpong and Jianlin Cheng (2024). “Improving protein function prediction by learning and integrating representations of protein sequences and function labels”. In: *bioRxiv* (Cited on page 76).
- Boniolo, Fabio et al. (2021). “Artificial intelligence in early drug discovery enabling precision medicine”. In: *Expert Opinion on Drug Discovery* 16.9, pp. 991–1007 (Cited on page 74).
- Bonner, Stephen et al. (2022). “A review of biomedical datasets relating to drug discovery: a knowledge graph perspective”. In: *Briefings in Bioinformatics* 23.6, bbac404 (Cited on page 121).
- Bradley, Catherine S et al. (2010). “The questionnaire for urinary incontinence diagnosis (QUID): validity and responsiveness to change in women undergoing non-surgical therapies for treatment of stress predominant urinary incontinence”. In: *Neurourology and urodynamics* 29.5, pp. 727–734 (Cited on page 52).
- Branden, Carl Ivar and John Tooze (2012). *Introduction to protein structure*. Garland Science (Cited on page 109).
- Brin, Sergey and Lawrence Page (1998). “The anatomy of a large-scale hypertextual web search engine”. In: *Computer networks and ISDN systems* 30.1-7, pp. 107–117 (Cited on page 161).
- Brodsky, Barbara et al. (2008). “Triple-helical peptides: An approach to collagen conformation, stability, and self-association”. In: *Biopolymers: Original Research on Biomolecules* 89.5, pp. 345–353 (Cited on page 98).
- Brown, Nathan et al. (2020). “Artificial intelligence in chemistry and drug design”. In: *Journal of Computer-Aided Molecular Design* 34, pp. 709–715 (Cited on page 134).
- Bryant, Patrick and Arne Elofsson (2022). “EvoBind: in silico directed evolution of peptide binders with AlphaFold”. In: *bioRxiv*, pp. 2022–07 (Cited on page 127).

- (2023). “Peptide binder design with inverse folding and protein structure prediction”. In: *Communications Chemistry* 6.1, p. 229 (Cited on page 127).
- Buehler, Markus J (2022). “Multiscale modeling at the interface of molecular mechanics and natural language through attention neural networks”. In: *Accounts of Chemical Research* 55.23, pp. 3387–3403 (Cited on page 134).
- Burley, Stephen K et al. (2017). “Protein Data Bank (PDB): the single global macromolecular structure archive”. In: *Protein crystallography: methods and protocols*, pp. 627–641 (Cited on pages 205, 206).
- Cahan, Eli M et al. (2019). “Putting the data before the algorithm in big data addressing personalized healthcare”. In: *NPJ Digital Medicine* 2.1, p. 78 (Cited on page 6).
- Cai, Qiong et al. (2019). “A survey on multimodal data-driven smart healthcare systems: approaches and applications”. In: *IEEE Access* 7, pp. 133583–133599 (Cited on page 7).
- Camacho, Luís, Georgios Douzas, and Fernando Bacao (2022). “Geometric SMOTE for regression”. In: *Expert Systems with Applications* 193, p. 116387 (Cited on page 71).
- Cannataro, Mario et al. (2022). *Artificial intelligence in bioinformatics: from omics analysis to deep learning and network mining*. Elsevier (Cited on pages 4, 11).
- Caro, Matthias C et al. (2022). “Generalization in quantum machine learning from few training data”. In: *Nature communications* 13.1, p. 4919 (Cited on page 71).
- Carter, Paul J and Greg A Lazar (2018). “Next generation antibody drugs: pursuit of the ‘high-hanging fruit’”. In: *Nature Reviews Drug Discovery* 17.3, pp. 197–223 (Cited on page 191).
- CASP13 (2023). *LCS and GDT score descriptions*. https://predictioncenter.org/casp13/doc/LCS_GDT_README. Accessed on February 27, 2023 (Cited on page 145).
- CASP14 (2023). *Help on GDT and other scores*. <https://predictioncenter.org/casp14/doc/help.html>. Accessed on February 27, 2023 (Cited on page 145).
- CASP15 (2023). *Official website*. <https://predictioncenter.org/casp15/index.cgi>. Accessed on February 27, 2023 (Cited on page 140).
- Chandak, Payal, Kexin Huang, and Marinka Zitnik (2023). “Building a knowledge graph to enable precision medicine”. In: *Scientific Data* 10.1, p. 67 (Cited on page 120).

References

- Chao, Jiannan, Furong Tang, and Lei Xu (2022). “Developments in algorithms for sequence alignment: A review”. In: *Biomolecules* 12.4, p. 546 (Cited on page 144).
- Chen, Can et al. (2023). “Structure-aware protein self-supervised learning”. In: *Bioinformatics* 39.4, btad189 (Cited on page 118).
- Chen, Chongyang et al. (2023). “Applications of multi-omics analysis in human diseases”. In: *MedComm* 4.4, e315 (Cited on page 9).
- Chen, Chuming, Hongzhan Huang, and Cathy H Wu (2017). “Protein bioinformatics databases and resources”. In: *Protein bioinformatics: from protein modifications and networks to proteomics*, pp. 3–39 (Cited on page 204).
- Chen, Jiaoyan et al. (2023). “Knowledge graphs for the life sciences: Recent developments, challenges and opportunities”. In: *arXiv preprint arXiv:2309.17255* (Cited on page 121).
- Chen, Lifan et al. (2020). “TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments”. In: *Bioinformatics* 36.16, pp. 4406–4414 (Cited on page 117).
- Chen, Lingtao et al. (2024). “AI-driven deep learning techniques in protein structure prediction”. In: *International journal of molecular sciences* 25.15, p. 8426 (Cited on page 123).
- Chen, Qingfeng (2024). “Computational Linguistics and Biological Sequences in Artificial Intelligence”. In: *Association Analysis Techniques and Applications in Bioinformatics*. Springer, pp. 123–158 (Cited on pages 4, 11).
- Chen, Richard J et al. (2021). “Synthetic data in machine learning for medicine and healthcare”. In: *Nature Biomedical Engineering* 5.6, pp. 493–497 (Cited on page 64).
- Chen, Zhewei, Linyue Zhou, and Wenwen Yu (2021). “ADASYN- Random forest based intrusion detection model”. In: *Proceedings of the 2021 4th International Conference on Signal Processing and Machine Learning*, pp. 152–159 (Cited on page 71).
- Chen, Zhidong et al. (2023). “Accelerating therapeutic protein design with computational approaches toward the clinical stage”. In: *Computational and Structural Biotechnology Journal* 21, pp. 2909–2926 (Cited on page 10).
- Cheng, Zhangyu, Chengming Zou, and Jianwei Dong (2019). “Outlier detection using isolation forest and local outlier factor”. In: *Proceedings of the conference on research in adaptive and convergent systems*, pp. 161–168 (Cited on page 65).

- Chowdhury, Ratul et al. (2022). “Single-sequence protein structure prediction using a language model and deep learning”. In: *Nature Biotechnology* 40.11, pp. 1617–1623 (Cited on pages 118, 120).
- Chungyoun, Michael F and Jeffrey J Gray (2023). “AI models for protein design are driving antibody engineering”. In: *Current Opinion in Biomedical Engineering* 28, p. 100473 (Cited on page 123).
- Ciamponi, Felipe E et al. (2018). “BioFeatureFinder: Flexible, unbiased analysis of biological characteristics associated with genomic regions”. In: *bioRxiv*, p. 279612 (Cited on pages 205, 219).
- Cioletti, Alessandra G et al. (2025). “Deep Learning in Bioinformatics: Principles and Applications”. In: *Effective Techniques for Bioinformatic Exploration*. IGI Global, pp. 137–168 (Cited on page 76).
- Cock, Peter JA et al. (2009). “Biopython: freely available Python tools for computational molecular biology and bioinformatics”. In: *Bioinformatics* 25.11, p. 1422 (Cited on page 172).
- Cohen, I Glenn and Michelle M Mello (2018). “HIPAA and protecting health information in the 21st century”. In: *Jama* 320.3, pp. 231–232 (Cited on page 32).
- Collins, Sarah and Christina Lewicky-Gaupp (2022). “Pelvic organ prolapse”. In: *Gastroenterology Clinics* 51.1, pp. 177–193 (Cited on page 75).
- Consolo, Letteria et al. (2023). “Barriers and facilitators of electronic patient-reported outcome measures (e-PROMs) for patients in home palliative cancer care: a qualitative study of healthcare professionals’ perceptions”. In: *BMC Palliative Care* 22.1, p. 111 (Cited on page 7).
- Consortium, Gene Ontology (2004). “The Gene Ontology (GO) database and informatics resource”. In: *Nucleic acids research* 32.suppl_1, pp. D258–D261 (Cited on page 79).
- Cooper, Robert G (2023). “The artificial intelligence revolution in new-product development”. In: *IEEE Engineering Management Review* (Cited on page 130).
- Correa-Aguila, Raidel, Niuxia Alonso-Pupo, and Erix W Hernández-Rodríguez (2022). “Multi-omics data integration approaches for precision oncology”. In: *Molecular Omics* 18.6, pp. 469–479 (Cited on page 9).
- Côté, Catherine Isadora (2024). “A critical and systematic literature review of epistemic justice applied to healthcare: recommendations for a patient partnership approach”. In: *Medicine, Health Care and Philosophy* 27.3, pp. 455–477 (Cited on page 6).

References

- Couto, Maria Raquel Lopes de and Breno Miranda (2022). “Towards improving automation support for internationalization and localization testing”. In: *Simpósio Brasileiro de Qualidade de Software (SBQS)*. SBC, pp. 9–14 (Cited on page 56).
- Cramer, Patrick (2021). “AlphaFold2 and the future of structural biology”. In: *Nature structural & molecular biology* 28.9, pp. 704–705 (Cited on pages 137, 139).
- Curtis, Vickie (2015). “Motivation to participate in an online citizen science game: A study of Foldit”. In: *Science Communication* 37.6, pp. 723–746 (Cited on page 139).
- Cusumano, Michael (2010). “Cloud computing and SaaS as new computing platforms”. In: *Communications of the ACM* 53.4, pp. 27–29 (Cited on page 28).
- Dahlgaard, Søren, Mathias Knudsen, and Mikkel Thorup (2017). “Practical hash functions for similarity estimation and dimensionality reduction”. In: *Advances in Neural Information Processing Systems* 30 (Cited on page 174).
- Dai, Xinyue and Yu Chen (2023). “Computational biomaterials: computational simulations for biomedicine”. In: *Advanced Materials* 35.7, p. 2204798 (Cited on page 10).
- Danchilla, Brian and Brian Danchilla (2012). “Three.js framework”. In: *Beginning WebGL for HTML5*, pp. 173–203 (Cited on page 144).
- Darvish, Diana M (2022). “Collagen fibril formation in vitro: From origin to opportunities”. In: *Materials Today Bio* 15, p. 100322 (Cited on page 100).
- Dauparas, Justas et al. (2022). “Robust deep learning–based protein sequence design using ProteinMPNN”. In: *Science* 378.6615, pp. 49–56 (Cited on pages 114, 123, 126).
- Davydova, Elena K (2022). “Protein engineering: advances in phage display for basic science and medical research”. In: *Biochemistry (Moscow)* 87.Suppl 1, S146–S167 (Cited on page 192).
- Day, Li, Julie A Cakebread, and Simon M Loveday (2022). “Food proteins from animals and plants: Differences in the nutritional and functional properties”. In: *Trends in Food Science & Technology* 119, pp. 428–442 (Cited on page 128).
- De Maria Marchiano, Ruggero et al. (2021). “Translational research in the era of precision medicine: where we are and where we will go”. In: *Journal of personalized medicine* 11.3, p. 216 (Cited on page 221).
- De Miranda, Michael A, Mark Doggett, and Jane T Evans (2005). “Medical technology: Contexts and content in science and technology”. In: (Cited on page 227).

- Dealmakers, Biopharma (2024). “Generative AI platforms drive drug discovery dealmaking”. In: *Biopharma Deal* 10 (Cited on pages 110, 128).
- DeGroat, William et al. (2024). “Multimodal AI/ML for discovering novel biomarkers and predicting disease using multi-omics profiles of patients with cardiovascular diseases”. In: *Scientific Reports* 14.1, p. 26503 (Cited on page 74).
- Deng, Zhi-Min et al. (2021). “Advances in molecular mechanisms of pelvic organ prolapse”. In: *Experimental and Therapeutic Medicine* 22.3, pp. 1–7 (Cited on page 75).
- Desomer, Anja et al. (2018). “Use of patient-reported outcome and experience measures in patient care and policy”. In: (Cited on page 7).
- Dhakal, Ashwin, Rajan Gyawali, and Jianlin Cheng (2023). “Predicting protein-ligand binding structure using E (n) Equivariant graph neural networks”. In: *bioRxiv*, pp. 2023–08 (Cited on page 116).
- Dill, K. A. and J. L. MacCallum (Nov. 2012). “The Protein-Folding Problem, 50 Years On”. en. In: *Science* 338.6110, pp. 1042–1046. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1219021. URL: <https://www.sciencemag.org/lookup/doi/10.1126/science.1219021> (visited on 05/07/2023) (Cited on page 164).
- Ding, Wenze, Kenta Nakai, and Haipeng Gong (2022). “Protein design via deep learning”. In: *Briefings in bioinformatics* 23.3, bbac102 (Cited on pages 191, 192).
- Divine, Robby et al. (2021). “Designed proteins assemble antibodies into modular nanocages”. In: *Science* 372.6537, eabd9994 (Cited on page 109).
- Dolgin, Elie (2024). “Protein Innovators Win Chemistry Nobel”. In: (Cited on page 123).
- Dowling, Paul et al. (2021). “The dystrophin node as integrator of cytoskeletal organization, lateral force transmission, fiber stability and cellular signaling in skeletal muscle”. In: *Proteomes* 9.1, p. 9 (Cited on page 95).
- Drolet, Brian C and Nancy M Lorenzi (2011). “Translational research: understanding the continuum from bench to bedside”. In: *Translational Research* 157.1, pp. 1–5 (Cited on pages 1, 2).
- Du, Xing et al. (2016). “Insights into protein–ligand interactions: mechanisms, models, and methods”. In: *International journal of molecular sciences* 17.2, p. 144 (Cited on page 191).

References

- Du, Zongyang et al. (2021). “The trRosetta server for fast and accurate protein structure prediction”. In: *Nature protocols* 16.12, pp. 5634–5651 (Cited on page 127).
- Duan, Junwei et al. (2024). “Deep learning based multimodal biomedical data fusion: An overview and comparative review”. In: *Information Fusion*, p. 102536 (Cited on page 225).
- Duarte, Yorley et al. (2019). “Integration of target discovery, drug discovery and drug delivery: a review on computational strategies”. In: *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology* 11.4, e1554 (Cited on page 10).
- DuBois, Paul (2013). *MySQL*. Addison-Wesley (Cited on page 214).
- Durán, Juan Manuel and Karin Rolanda Jongsma (2021). “Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI”. In: *Journal of medical ethics* 47.5, pp. 329–335 (Cited on page 8).
- Eder, Johann and Michele Missikoff (2003). *Advanced Information Systems Engineering: 15th International Conference, CAiSE 2003, Klagenfurt, Austria, June 16-18, 2003, Proceedings*. Vol. 2681. Springer (Cited on page 143).
- Elfanagely, Omar et al. (2021). “Machine learning and surgical outcomes prediction: a systematic review”. In: *Journal of Surgical Research* 264, pp. 346–361 (Cited on page 62).
- Emmerson, Stuart J and Caroline E Gargett (2016). “Endometrial mesenchymal stem cells as a cell based therapy for pelvic organ prolapse”. In: *World journal of stem cells* 8.5, p. 202 (Cited on page 107).
- Erdős, Gábor, Mátyás Pajkos, and Zsuzsanna Dosztányi (2021). “IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation”. In: *Nucleic acids research* 49.W1, W297–W303 (Cited on page 192).
- Espina, Giannina, Joaquín Atalah, and Jenny M Blamey (2021). “Extremophilic oxidoreductases for the industry: five successful examples with promising projections”. In: *Frontiers in Bioengineering and Biotechnology* 9, p. 710035 (Cited on page 130).
- Fatton, Brigitte et al. (2020). “Pelvic organ prolapse and sexual function”. In: *Nature Reviews Urology* 17.7, pp. 373–390 (Cited on page 75).
- Ferruz, Noelia, Michael Heinzinger, et al. (2023). “From sequence to function through structure: Deep learning for protein design”. In: *Computational and Structural Biotechnology Journal* 21, pp. 238–250 (Cited on page 114).

- Ferruz, Noelia and Birte Höcker (2022). “Controllable protein design with language models”. In: *Nature Machine Intelligence* 4.6, pp. 521–532 (Cited on pages 118, 120).
- Ferruz, Noelia, Steffen Schmidt, and Birte Höcker (2022). “ProtGPT2 is a deep unsupervised language model for protein design”. In: *Nature communications* 13.1, p. 4348 (Cited on pages 112, 127).
- Filipavicius, Modestas et al. (2020). “Pre-training protein language models with label-agnostic binding pairs enhances performance in downstream tasks”. In: *arXiv preprint arXiv:2012.03084* (Cited on page 120).
- Finn, Robert D, Jody Clements, and Sean R Eddy (2011). “HMMER web server: interactive sequence similarity searching”. In: *Nucleic acids research* 39.suppl_2, W29–W37 (Cited on page 166).
- Finn, Robert D, Penelope Coghill, et al. (2016). “The Pfam protein families database: towards a more sustainable future”. In: *Nucleic acids research* 44.D1. Publisher: Oxford University Press, pp. D279–D285. ISSN: 1362-4962 (Cited on page 166).
- Flier, Jeffrey S and Joseph Loscalzo (2017). “Categorizing biomedical research: the basics of translation”. In: *The FASEB Journal* 31.8, p. 3210 (Cited on page 4).
- Flower, Thomas G and James H Hurley (2021). “Crystallographic molecular replacement using an in silico-generated search model of SARS-CoV-2 ORF8”. In: *Protein Science* 30.4, pp. 728–734 (Cited on page 192).
- Forsman, Maria and Nikodemus Solitander (2003). “Knowledge transfer in clusters and networks: an interdisciplinary conceptual analysis”. In: *Journal of International Business Studies* 3, pp. 1–23 (Cited on pages 1, 2).
- Fraley, Amber D et al. (2022). “Exploring protein structure with RasMol, a free online tool”. In: *The American Biology Teacher* 84.6, pp. 369–373 (Cited on page 139).
- Frauenfelder, Hans (2002). “Proteins: paradigms of complexity”. In: *Proceedings of the National Academy of Sciences* 99.suppl_1, pp. 2479–2480 (Cited on page 204).
- Friederich, Pascal et al. (2021). “Machine-learned potentials for next-generation matter simulations”. In: *Nature Materials* 20.6, pp. 750–761 (Cited on page 122).
- Fu, Limin et al. (2012). “CD-HIT: accelerated for clustering the next-generation sequencing data”. In: *Bioinformatics* 28.23, pp. 3150–3152 (Cited on page 79).
- Fuchs, S. (2023). *Gravy calculator*. <http://gravy-calculator.de/>. Accessed on March 14, 2023 (Cited on pages 199, 200).

References

- Fujiwara, Toshihiro et al. (2016). “Extracellular superoxide dismutase deficiency impairs wound healing in advanced age by reducing neovascularization and fibroblast function”. In: *Experimental dermatology* 25.3, pp. 206–211 (Cited on page 95).
- Gackenheimer, Cory (2015). *Introduction to React*. Apress (Cited on page 48).
- Gainza, Pablo, Kyle E Roberts, and Bruce R Donald (2012). “Protein design using continuous rotamers”. In: *PLoS computational biology* 8.1. Publisher: Public Library of Science San Francisco, USA, e1002335. ISSN: 1553-734X (Cited on pages 163, 164).
- Gao, Wenhao et al. (2020). “Deep learning in protein structural modeling and design”. In: *Patterns* 1.9 (Cited on page 114).
- Gao, Zhangyang, Cheng Tan, and Stan Z Li (2022). “Alphadesign: A graph protein design method and benchmark on alphafolddb”. In: *arXiv preprint arXiv:2202.01079* (Cited on page 116).
- Gasteiger, E et al. (2005). *ExPASy ProtParam* (Cited on page 205).
- Gasteiger, Elisabeth et al. (2005). “Protein identification and analysis tools on the ExPASy server”. In: *The proteomics protocols handbook*, pp. 571–607 (Cited on pages 199, 200).
- Gaudelet, Thomas et al. (2021). “Utilizing graph machine learning within drug discovery and development”. In: *Briefings in bioinformatics* 22.6, bbab159 (Cited on page 120).
- Gharajedaghi, Jamshid (2011). *Systems thinking: Managing chaos and complexity: A platform for designing business architecture*. Elsevier (Cited on page 141).
- Gligorijević, Vladimir et al. (2021). “Structure-based protein function prediction using graph convolutional networks”. In: *Nature communications* 12.1, p. 3168 (Cited on pages 166, 186).
- Goepel, Christian et al. (2011). “Changes of glycoprotein and collagen immunolocalization in the uterine artery wall of postmenopausal women with and without pelvic organ prolapse”. In: *Acta Histochemica* 113.3, pp. 375–381 (Cited on page 107).
- Gómez-Bombarelli, Rafael et al. (2018). “Automatic chemical design using a data-driven continuous representation of molecules”. In: *ACS central science* 4.2, pp. 268–276 (Cited on page 192).
- Gong, Runqi and Zhijun Xia (2019). “Collagen changes in pelvic support tissues in women with pelvic organ prolapse”. In: *European Journal of Obstetrics & Gynecology and Reproductive Biology* 234, pp. 185–189 (Cited on pages 75, 100).

- Gong, Tiansu et al. (2023). “SASA-Net: a spatial-aware self-attention mechanism for building protein 3D structure directly from inter-residue distances”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 20.6, pp. 3482–3488 (Cited on page 118).
- Gonzales, Gerard Bryan et al. (2022). “The role of albumin and the extracellular matrix on the pathophysiology of oedema formation in severe malnutrition”. In: *EBioMedicine* 79 (Cited on page 95).
- Gonzalez-Hernandez, Graciela et al. (2017). “Capturing the patient’s perspective: a review of advances in natural language processing of health-related text”. In: *Yearbook of medical informatics* 26.01, pp. 214–227 (Cited on page 222).
- Grimes, Cara L et al. (2020). “Correlation of electronic (web-based and smartphone) administration of measures of pelvic floor dysfunction: a randomized controlled trial”. In: *Urogynecology* 26.6, pp. 396–400 (Cited on page 46).
- Gront, Dominik et al. (2011). “Generalized fragment picking in Rosetta: design, protocols and applications”. In: *PloS one* 6.8, e23294 (Cited on page 138).
- Grossmann, Lutz and Jochen Weiss (2021). “Alternative protein sources as technofunctional food ingredients”. In: *Annual Review of Food Science and Technology* 12.1, pp. 93–117 (Cited on page 128).
- Gu, Zhonghui et al. (2023). “Hierarchical graph transformer with contrastive learning for protein function prediction”. In: *Bioinformatics* 39.7, btad410 (Cited on page 118).
- Guadagno, Noemi Antonella and Cinzia Progida (2019). “Rab GTPases: switching to human diseases”. In: *Cells* 8.8, p. 909 (Cited on page 95).
- Gubert, Luis Claudio, Cristiano André da Costa, and Rodrigo da Rosa Righi (2020). “Context awareness in healthcare: a systematic literature review”. In: *Universal Access in the Information Society* 19.2, pp. 245–259 (Cited on page 8).
- Guler, Zeliha and Jan Paul Roovers (2022). “Role of fibroblasts and myofibroblasts on the pathogenesis and treatment of pelvic organ prolapse”. In: *Biomolecules* 12.1, p. 94 (Cited on pages 75, 100).
- Gupta, Namita T et al. (2015). “Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data”. In: *Bioinformatics* 31.20, pp. 3356–3358 (Cited on pages 199, 200).
- Guruprasad, Kunchur, BV Bhasker Reddy, and Madhusudan W Pandit (1990). “Correlation between stability of a protein and its dipeptide composition: a novel approach

References

- for predicting in vivo stability of a protein from its primary sequence”. In: *Protein Engineering, Design and Selection* 4.2, pp. 155–161 (Cited on pages 172, 197, 199, 200, 212).
- Gutt, Ernst-August (2014). *Translation and relevance: Cognition and context*. Routledge (Cited on page 1).
- Haleem, Abid et al. (2021). “Telemedicine for healthcare: Capabilities, features, barriers, and applications”. In: *Sensors international* 2, p. 100117 (Cited on page 27).
- Hallgren, Jeppe et al. (2022). “DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks”. In: *BioRxiv*, pp. 2022–04 (Cited on page 79).
- Hartl, Dominik et al. (2021). “Translational precision medicine: an industry perspective”. In: *Journal of translational medicine* 19.1, p. 245 (Cited on page 221).
- Hawkins-Hooker, Alex et al. (2021). “Generating functional protein variants with variational autoencoders”. In: *PLoS computational biology* 17.2, e1008736 (Cited on page 192).
- He, Xibing et al. (2020). “A fast and high-quality charge model for the next generation general AMBER force field”. In: *The Journal of Chemical Physics* 153.11 (Cited on page 122).
- Heath, Michele and Tracy Porter (2019). “Sensemaking through a storytelling lens: Physician perspectives of health information exchange”. In: *Qualitative Research in Organizations and Management: An International Journal* 14.4, pp. 428–443 (Cited on page 6).
- Hebbbar, Shripad, Harshita Pandey, and Arun Chawla (2015). “Understanding King’s Health Questionnaire (KHQ) in assessment of female urinary incontinence”. In: *Int J Res Med Sci* 3.3, pp. 531–538 (Cited on page 52).
- Heo, Lim, Giacomo Janson, and Michael Feig (2021). “Physics-based protein structure refinement in the era of artificial intelligence”. In: *Proteins: Structure, Function, and Bioinformatics* 89.12, pp. 1870–1887 (Cited on page 134).
- Holden, Richard J et al. (2013). “SEIPS 2.0: a human factors framework for studying and improving the work of healthcare professionals and patients”. In: *Ergonomics* 56.11, pp. 1669–1686 (Cited on page 227).
- Holm, Liisa and Laura M Laakso (2016). “Dali server update”. In: *Nucleic acids research* 44.W1, W351–W355 (Cited on pages 166, 183).

- Holzinger, Andreas, Katharina Keiblinger, et al. (2023). “AI for life: Trends in artificial intelligence for biotechnology”. In: *New Biotechnology* 74, pp. 16–24 (Cited on page 126).
- Holzinger, Andreas, Bernd Malle, et al. (2021). “Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI”. In: *Information Fusion* 71, pp. 28–37 (Cited on page 8).
- Hooft, Rob WW, Chris Sander, and Gerrit Vriend (1997). “Objectively judging the quality of a protein structure from a Ramachandran plot”. In: *Bioinformatics* 13.4, pp. 425–430 (Cited on page 146).
- Hossain, MD Imran et al. (2025). “Explainable AI for medical data: current methods, limitations, and future directions”. In: *ACM Computing Surveys* 57.6, pp. 1–46 (Cited on page 8).
- Huang, Guangming et al. (2024). “From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality?” In: *Computational and Structural Biotechnology Journal* (Cited on page 8).
- Huang, Po-Ssu et al. (2011). “RosettaRemodel: a generalized framework for flexible backbone protein design”. In: *PloS one* 6.8, e24109 (Cited on page 123).
- Hubbard, Simon J (1993). “naccess”. In: *Dept Biohem Mol Biol UCL* (Cited on page 172).
- Huo, Mingjia et al. (2024). “Multi-Modal Large Language Model Enables Protein Function Prediction”. In: *bioRxiv*, pp. 2024–08 (Cited on page 119).
- Hynes, Richard O and Alexandra Naba (2012). “Overview of the matrisome—an inventory of extracellular matrix constituents and functions”. In: *Cold Spring Harbor perspectives in biology* 4.1, a004903 (Cited on pages 91, 94).
- Iglesia, Cheryl B and Katelyn R Smithling (2017). “Pelvic organ prolapse”. In: *American family physician* 96.3, pp. 179–185 (Cited on pages 46, 62).
- Jackson-Boeters, Linda, Weiyan Wen, and Douglas W Hamilton (2009). “Periostin localizes to cells in normal skin, but is associated with the extracellular matrix during wound repair”. In: *Journal of cell communication and signaling* 3.2, pp. 125–133 (Cited on page 94).
- James, Gilad (2023). *Introduction to Google Translate*. Gilad James Mystery School (Cited on page 3).

References

- Jelovsek, J Eric, Kevin Chagin, et al. (2018). “Models for predicting recurrence, complications, and health status in women after pelvic organ prolapse surgery”. In: *Obstetrics & Gynecology* 132.2, pp. 298–309 (Cited on page 63).
- Jelovsek, J Eric, Christopher Maher, and Matthew D Barber (2007). “Pelvic organ prolapse”. In: *The Lancet* 369.9566, pp. 1027–1038 (Cited on pages 45, 46).
- Jiang, Mingrui et al. (2024). “Binding of carbon monoxide to hemoglobin in an oxygen environment: force field development for molecular dynamics”. In: *Journal of Chemical Theory and Computation* (Cited on page 122).
- Jin, Hao et al. (2019). “A review of secure and privacy-preserving medical data sharing”. In: *IEEE Access* 7, pp. 61656–61669 (Cited on page 28).
- Jin, Hongyuan et al. (2021). “Rab GTPases: central coordinators of membrane trafficking in cancer”. In: *Frontiers in Cell and Developmental Biology* 9, p. 648384 (Cited on page 95).
- Jin, Xiaopeng et al. (2021). “SMI-BLAST: a novel supervised search framework based on PSI-BLAST for protein remote homology detection”. In: *Bioinformatics* 37.7, pp. 913–920 (Cited on page 209).
- Jing, Bowen et al. (2021). “Equivariant graph neural networks for 3d macromolecular structure”. In: *arXiv preprint arXiv:2106.03843* (Cited on page 116).
- Jisna, VA and PB Jayaraj (2021). “Protein structure prediction: conventional and deep learning perspectives”. In: *The protein journal* 40.4, pp. 522–544 (Cited on page 114).
- Johnson, AM, EM Rohlf, and LM Silverman (1999). “Protein”. In: *Tietz textbook of clinical chemistry. 3rd ed. WB Saunders, Philadelphia*, pp. 521–525 (Cited on page 109).
- Johnson, Stephen B et al. (2008). “An electronic health record based on structured narrative”. In: *Journal of the American Medical Informatics Association* 15.1, pp. 54–64 (Cited on page 6).
- Jones, Philip et al. (2014). “InterProScan 5: genome-scale protein function classification”. In: *Bioinformatics* 30.9, pp. 1236–1240 (Cited on page 209).
- Jordan, Michelle E et al. (2009). “The role of conversation in health care interventions: enabling sensemaking and learning”. In: *Implementation Science* 4, pp. 1–13 (Cited on page 6).

- Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, et al. (2021). “Highly accurate protein structure prediction with AlphaFold”. In: *nature* 596.7873, pp. 583–589 (Cited on pages 74, 112, 123).
- Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Kathryn Tunyasuvunakool, et al. (2020). “AlphaFold 2”. In: *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction* (Cited on page 124).
- Kanehisa, Minoru et al. (2025). “KEGG: biological systems database as a model of the real world”. In: *Nucleic acids research* 53.D1, pp. D672–D677 (Cited on page 84).
- Kapcha, Lauren H and Peter J Rossky (2014). “A simple atomic-level hydrophobicity scale reveals protein interfacial structure”. In: *Journal of molecular biology* 426.2, pp. 484–498 (Cited on page 170).
- Karabulut, Gulsah, Gulden Goksen, and Amin Mousavi Khaneghah (2024). “Plant-based protein modification strategies towards challenges”. In: *Journal of Agriculture and Food Research*, p. 101017 (Cited on page 128).
- Karamanos, Nikos K et al. (2021). “A guide to the composition and functions of the extracellular matrix”. In: *The FEBS journal* 288.24, pp. 6850–6912 (Cited on page 75).
- Karplus, Martin and John Kuriyan (2005). “Molecular dynamics and protein function”. In: *Proceedings of the National Academy of Sciences* 102.19, pp. 6679–6685 (Cited on page 109).
- Karras, Tero et al. (2020). “Training generative adversarial networks with limited data”. In: *Advances in neural information processing systems* 33, pp. 12104–12114 (Cited on page 71).
- Kastritis, Panagiotis L et al. (2013). “Solvated protein–protein docking using Kyte–Doolittle-based water preferences”. In: *Proteins: Structure, Function, and Bioinformatics* 81.3, pp. 510–518 (Cited on page 211).
- Katoch, Sourabh, Sumit Singh Chauhan, and Vijay Kumar (2021). “A review on genetic algorithm: past, present, and future”. In: *Multimedia tools and applications* 80, pp. 8091–8126 (Cited on page 192).
- Kaur, Jagpreet, Shweta Lamba, and Preeti Saini (2021). “Advanced encryption standard: attacks and current research trends”. In: *2021 international conference on advance computing and innovative technologies in engineering (ICACITE)*. IEEE, pp. 112–116 (Cited on page 32).

References

- Kaushik, Aman Chandra and Utkarsh Raj (2020). “AI-driven drug discovery: A boon against COVID-19?” In: *Ai Open* 1, pp. 1–4 (Cited on page 128).
- Kay, Misha, Jonathan Santos, Marina Takane, et al. (2011). “mHealth: New horizons for health through mobile technologies”. In: *World Health Organization* 64.7, pp. 66–71 (Cited on page 45).
- Kempen, Michel van et al. (2022). “Foldseek: fast and accurate protein structure search”. In: *Biorxiv*, pp. 2022–02 (Cited on page 183).
- Kerosuo, Hannele (2010). “Lost in translation: a patient-centred experience of unintegrated care”. In: *International Journal of Public Sector Management* 23.4, pp. 372–380 (Cited on pages 6, 7).
- Kesik-Brodacka, Malgorzata (2018). “Progress in biopharmaceutical development”. In: *Biotechnology and applied biochemistry* 65.3, pp. 306–322 (Cited on page 191).
- Kessler, Daniela et al. (2001). “Fibroblasts in mechanically stressed collagen lattices assume a “synthetic” phenotype”. In: *Journal of Biological Chemistry* 276.39, pp. 36575–36585 (Cited on page 103).
- Khan, Yousuf A, K Ian White, and Axel T Brunger (2022). “The AAA+ superfamily: a review of the structural and mechanistic principles of these molecular machines”. In: *Critical reviews in biochemistry and molecular biology* 57.2, pp. 156–187 (Cited on page 110).
- Kim, Seong A et al. (2023). “Protein-based nanocages for vaccine development”. In: *Journal of Controlled Release* 353, pp. 767–791 (Cited on pages 110, 127).
- Kirkley, Sonny E and Jamie R Kirkley (2005). “Creating next generation blended learning environments using mixed reality, video games and simulations”. In: *TechTrends* 49.3, pp. 42–53 (Cited on page 31).
- Klausen, Michael Schantz et al. (2019). “NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning”. In: *Proteins: Structure, Function, and Bioinformatics* 87.6, pp. 520–527 (Cited on page 193).
- Kleinberg, Samantha and George Hripcsak (2011). “A review of causal inference for biomedical informatics”. In: *Journal of biomedical informatics* 44.6, pp. 1102–1112 (Cited on page 226).
- Kline, Adrienne et al. (2022). “Multimodal machine learning in precision health: A scoping review”. In: *npj Digital Medicine* 5.1, p. 171 (Cited on page 8).

- Klovning, Atle et al. (2009). “Comparison of two questionnaires for assessing the severity of urinary incontinence: the ICIQ-UI SF versus the incontinence severity index”. In: *Neurourology and Urodynamics: Official Journal of the International Continence Society* 28.5, pp. 411–415 (Cited on page 52).
- Kluiwers, Kirsten B et al. (2023). “Molecular landscape of pelvic organ prolapse provides insights into disease etiology”. In: *International Journal of Molecular Sciences* 24.7, p. 6087 (Cited on page 75).
- Koleck, Theresa A et al. (2019). “Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review”. In: *Journal of the American Medical Informatics Association* 26.4, pp. 364–379 (Cited on page 222).
- Kong, Ping, Panagiota Christia, and Nikolaos G Frangogiannis (2014). “The pathogenesis of cardiac fibrosis”. In: *Cellular and molecular life sciences* 71.4, pp. 549–574 (Cited on page 104).
- Korendovych, Ivan V and William F DeGrado (2020). “De novo protein design, a retrospective”. In: *Quarterly reviews of biophysics* 53, e3 (Cited on pages 110, 185).
- Kortemme, Tanja (2024). “De novo protein design—From new structures to programmable functions”. In: *Cell* 187.3, pp. 526–544 (Cited on page 110).
- Kouba, Petr et al. (2023). “Machine learning-guided protein engineering”. In: *ACS catalysis* 13.21, pp. 13863–13895 (Cited on page 76).
- Kouw, Wouter M and Marco Loog (2018). “An introduction to domain adaptation and transfer learning”. In: *arXiv preprint arXiv:1812.11806* (Cited on page 227).
- Kozlova, Irina and Vladimir Sytnyk (2024). “Cell adhesion molecules as modulators of the epidermal growth factor receptor”. In: *Cells* 13.22, p. 1919 (Cited on page 95).
- Kozlowski, Lukasz P (2016). “IPC—isoelectric point calculator”. In: *Biology direct* 11, pp. 1–16 (Cited on pages 197, 199, 200).
- Krishna, Rajesh, Hans Guenter Schaefer, and Ole J Bjerrum (2007). “Effective integration of systems biology, biomarkers, biosimulation and modelling in streamlining drug development”. In: *European journal of pharmaceutical sciences* 31.1, pp. 62–67 (Cited on page 2).
- Krishna, Rohith et al. (2024). “Generalized biomolecular modeling and design with RoseTTAFold All-Atom”. In: *Science* 384.6693, ead12528 (Cited on page 125).

References

- Kuhlman, Brian and Philip Bradley (2019a). “Advances in protein structure prediction and design”. In: *Nature reviews molecular cell biology* 20.11, pp. 681–697 (Cited on pages 110, 138).
- (2019b). “Advances in protein structure prediction and design”. In: *Nature reviews molecular cell biology* 20.11, pp. 681–697 (Cited on pages 163, 164).
- Kuivaniemi, Helena and Gerard Tromp (2019). “Type III collagen (COL3A1): Gene and protein structure, tissue distribution, and associated diseases”. In: *Gene* 707, pp. 151–171 (Cited on page 100).
- Kuivaniemi, Helena, Gerard Tromp, and Darwin J Prockop (1991). “Mutations in collagen genes: causes of rare and some common diseases in humans”. In: *The FASEB journal* 5.7, pp. 2052–2060 (Cited on page 98).
- Kulmanov, Maxat and Robert Hoehndorf (2020). “DeepGOPlus: improved protein function prediction from sequence”. In: *Bioinformatics* 36.2, pp. 422–429 (Cited on page 193).
- Kumar, T Ashok (2013). “CFSSP: Chou and Fasman secondary structure prediction server”. In: *Wide Spectrum* 1.9, pp. 15–19 (Cited on page 170).
- Kusner, Matt J, Brooks Paige, and José Miguel Hernández-Lobato (2017). “Grammar variational autoencoder”. In: *International conference on machine learning*. PMLR, pp. 1945–1954 (Cited on page 192).
- Kyte, J (1993). “A simple method for displaying the hydropathic character of a protein.” In: *J. Mol. Biol.* 268, pp. 10558–10563 (Cited on page 197).
- Lahoti, Ronith (2023). “Charting New Horizons: The Expanding Role of AI in Healthcare and Protein Engineering”. In: *Available at SSRN* 4982612 (Cited on page 130).
- Lam, Hoang Thanh et al. (2023). “Otter-Knowledge: benchmarks of multimodal knowledge graph representation learning from different sources for drug discovery”. In: *arXiv preprint arXiv:2306.12802* (Cited on page 121).
- Laskowski, Roman A et al. (2018). “PDBsum: Structural summaries of PDB entries”. In: *Protein science* 27.1, pp. 129–134 (Cited on page 205).
- Lee, Ingoo, Jongsoo Keum, and Hojung Nam (2019). “DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences”. In: *PLoS computational biology* 15.6, e1007129 (Cited on page 193).

- Lee, Jae-Won et al. (2023). “DeepFold: enhancing protein structure prediction through optimized loss functions, improved template features, and re-optimized energy function”. In: *Bioinformatics* 39.12, btad712 (Cited on page 167).
- Lee, Yoseb and Young-Han Lim (2022). “Concurrency processing comparison of large data list using GO language”. In: *The Journal of the Convergence on Culture Technology* 8.2, pp. 361–366 (Cited on page 205).
- Leman, Julia Koehler et al. (2020). “Macromolecular modeling and design in Rosetta: recent methods and frameworks”. In: *Nature methods* 17.7, pp. 665–680 (Cited on page 139).
- Li, Irene et al. (2022). “Neural natural language processing for unstructured data in electronic health records: a review”. In: *Computer Science Review* 46, p. 100511 (Cited on page 7).
- Li, Shuangli et al. (2021). “Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity”. In: *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 975–985 (Cited on page 118).
- Li, Xiang-Juan et al. (2019). “Proteomics of uterosacral ligament connective tissue from women with and without pelvic organ prolapse”. In: *PROTEOMICS–Clinical Applications* 13.4, p. 1800086 (Cited on pages 80, 107).
- Li, Yaqian et al. (2021). “Single-cell transcriptome profiling of the vaginal wall in women with severe anterior vaginal prolapse”. In: *Nature Communications* 12.1, p. 87 (Cited on page 107).
- Li, Yu et al. (2021). “Structural, functional and molecular pathogenesis of pelvic organ prolapse in patient and Loxl1 deficient mice”. In: *Aging (Albany NY)* 13.24, p. 25886 (Cited on page 75).
- Liao, Lizhi et al. (2022). “An empirical study of the impact of hyperparameter tuning and model optimization on the performance properties of deep neural networks”. In: *ACM Transactions on Software Engineering and Methodology (TOSEM)* 31.3, pp. 1–40 (Cited on page 66).
- Lin, Asiyah Yu et al. (2023). “Improving the quality and utility of electronic health record data through ontologies”. In: *Standards* 3.3, pp. 316–340 (Cited on page 8).

References

- Lin, Sheng-Hsuan and Mohammad Arfan Ikram (2020). “On the relationship of machine learning with causal inference”. In: *European journal of epidemiology* 35, pp. 183–185 (Cited on page 226).
- Lin, Wei-Chao and Chih-Fong Tsai (2020). “Missing value imputation: a review and analysis of the literature (2006–2017)”. In: *Artificial Intelligence Review* 53, pp. 1487–1509 (Cited on page 65).
- Lin, Zeming et al. (2023). “Evolutionary-scale prediction of atomic-level protein structure with a language model”. In: *Science* 379.6637, pp. 1123–1130 (Cited on pages 74, 75, 80, 101, 112, 123).
- Liu, Pingyang et al. (2021). “Barriers and facilitators to interdisciplinary communication during consultations: a qualitative study”. In: *BMJ open* 11.9, e046111 (Cited on page 3).
- Liverman, Catharyn T et al. (2013). “The CTSA program at NIH: Opportunities for advancing clinical and translational research”. In: (Cited on page 4).
- Lo, Felix Tun Han (2024). “The paradoxical transparency of opaque machine learning”. In: *AI & SOCIETY* 39.3, pp. 1397–1409 (Cited on page 7).
- Lord, Laura and Nicola Gale (2014). “Subjective experience or objective process: understanding the gap between values and practice for involving patients in designing patient-centred care”. In: *Journal of health organization and management* 28.6, pp. 714–730 (Cited on page 2).
- Loshchilov, Ilya and Frank Hutter (2017). “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (Cited on page 81).
- Lv, Liuzhenghao et al. (2024). “Prollama: A protein large language model for multi-task protein language processing”. In: *arXiv preprint arXiv:2402.16445* (Cited on page 119).
- Ma, Tianle and Aidong Zhang (2016). “Omics informatics: from scattered individual software tools to integrated workflow management systems”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 14.4, pp. 926–946 (Cited on page 226).
- Maharana, Kiran, Surajit Mondal, and Bhushankumar Nemade (2022). “A review: Data pre-processing and data augmentation techniques”. In: *Global Transitions Proceedings* 3.1, pp. 91–99 (Cited on page 71).

- Malbranke, Cyril et al. (2023). “Machine learning for evolutionary-based and physics-inspired protein design: Current and future synergies”. In: *Current Opinion in Structural Biology* 80, p. 102571 (Cited on page 134).
- Mansoor, Sanaa et al. (2023). “Zero-shot mutation effect prediction on protein stability and function using RoseTTAFold”. In: *Protein Science* 32.11, e4780 (Cited on page 125).
- Marcos, Enrique and Daniel-Adriano Silva (2018). “Essentials of de novo protein design: Methods and applications”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 8.6, e1374 (Cited on page 10).
- Marcos-Zambrano, Laura Judith et al. (2021). “Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment”. In: *Frontiers in microbiology* 12, p. 634511 (Cited on page 74).
- Marcu, Radu Dragos et al. (2020). “Oxidative stress: a possible trigger for pelvic organ prolapse”. In: *Journal of Immunology Research* 2020 (Cited on page 75).
- Market Research Future (2023). *Protein Engineering Market Research Report*. <https://www.marketresearchfuture.com/reports/protein-engineering-market-691>. Accessed in 2023 (Cited on page 138).
- Markus, Braun et al. (2023). “Accelerating biocatalysis discovery with machine learning: a paradigm shift in enzyme engineering, discovery, and design”. In: *ACS catalysis* 13.21, pp. 14454–14469 (Cited on pages 129, 130).
- Marsh, Joseph A and Sarah A Teichmann (2011). “Relative solvent accessible surface area predicts protein conformational changes upon binding”. In: *Structure* 19.6, pp. 859–867 (Cited on page 148).
- Martina, Enrico et al. (2010). “Tenascin-W is a specific marker of glioma-associated blood vessels and stimulates angiogenesis in vitro”. In: *The FASEB Journal* 24.3, p. 778 (Cited on page 107).
- McClements, David Julian and Lutz Grossmann (2021). “The science of plant-based foods: Constructing next-generation meat, fish, milk, and egg analogs”. In: *Comprehensive Reviews in Food Science and Food Safety* 20.4, pp. 4049–4100 (Cited on page 129).

References

- Meier, Joshua et al. (2021). “Language models enable zero-shot prediction of the effects of mutations on protein function”. In: *Advances in neural information processing systems* 34, pp. 29287–29303 (Cited on page 101).
- Meijerink, Aukje M, Reinier H Van Rijssel, and Paul JQ Van Der Linden (2013). “Tissue composition of the vaginal wall in women with pelvic organ prolapse”. In: *Gynecologic and Obstetric Investigation* 75.1, pp. 21–27 (Cited on page 75).
- Melton, Lisa (2023). “So sweet, no spike”. In: *Nat Biotechnol* 41, pp. 883–885 (Cited on page 129).
- Memon, Anjum et al. (2016). “Perceived barriers to accessing mental health services among black and minority ethnic (BME) communities: a qualitative study in South-east England”. In: *BMJ open* 6.11, e012337 (Cited on page 7).
- Mi, Yanlin et al. (2023a). “PROFASA-A Web-based Protein Fragment and Structure Analysis Workstation”. In: *Frontiers in Bioengineering and Biotechnology* 11, p. 1192094. DOI: 10.3389/fbioe.2023.1192094 (Cited on pages 170, 181).
- (2023b). “PROFASA—a web-based protein fragment and structure analysis workstation”. In: *Frontiers in Bioengineering and Biotechnology* 11, p. 1192094 (Cited on page 127).
- (2024). “PS-GO parametric protein search engine”. In: *Computational and Structural Biotechnology Journal* 23, pp. 1499–1509 (Cited on pages 127, 219).
- Milacic, Marija et al. (2024). “The reactome pathway knowledgebase 2024”. In: *Nucleic acids research* 52.D1, pp. D672–D678 (Cited on page 84).
- Milani, Richard V et al. (2017). “Improving hypertension control and patient engagement using digital tools”. In: *The American journal of medicine* 130.1, pp. 14–20 (Cited on page 59).
- Mirdita, Milot et al. (2022). “ColabFold: making protein folding accessible to all”. In: *Nature methods* 19.6, pp. 679–682 (Cited on page 127).
- Mirny, Leonid A and Eugene I Shakhnovich (1999). “Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function”. In: *Journal of molecular biology* 291.1. Publisher: Elsevier, pp. 177–196. ISSN: 0022-2836 (Cited on page 164).
- Mirza, Zeenat et al. (2023). “Identification of novel diagnostic and prognostic gene signature biomarkers for breast cancer using artificial intelligence and machine

- learning assisted transcriptomics analysis”. In: *Cancers* 15.12, p. 3237 (Cited on page 74).
- Misra, Biswapriya B et al. (2019). “Integrated omics: tools, advances and future approaches”. In: *Journal of molecular endocrinology* 62.1, R21–R45 (Cited on page 9).
- Mohamed, Sameh K, Aayah Nounu, and Vít Nováček (2021). “Biological applications of knowledge graph embedding models”. In: *Briefings in bioinformatics* 22.2, pp. 1679–1693 (Cited on pages 120, 121).
- Moya, Juan A Berrios (2024). “Addressing the gaps in early dementia detection: A path towards enhanced diagnostic models through machine learning”. In: *arXiv preprint arXiv:2409.03147* (Cited on page 7).
- Murzin, Alexey G et al. (1995). “SCOP: a structural classification of proteins database for the investigation of sequences and structures”. In: *Journal of molecular biology* 247.4, pp. 536–540 (Cited on pages 162, 164, 166).
- Naba, Alexandra (2023). “Ten years of extracellular matrix proteomics: accomplishments, challenges, and future perspectives”. In: *Molecular & Cellular Proteomics* 22.4 (Cited on page 75).
- Nagarajan, Radhakrishnan et al. (2015). “Social network analysis to assess the impact of the CTSA on biomedical research grant collaboration”. In: *Clinical and translational science* 8.2, pp. 150–154 (Cited on page 4).
- Nagra, Navraj S et al. (2023). “Understanding the company landscape in AI-driven biopharma R&D”. In: *Biopharma Deal* 10 (Cited on pages 110, 128).
- Nallabantu, Sri Charan and J Guruprakash (2024). “AI and the next medical revolution: deep learning’s uncharted healthcare promise”. In: (Cited on page 228).
- Narayanan, Harini et al. (2021). “Machine learning for biologics: opportunities for protein engineering, developability, and formulation”. In: *Trends in pharmacological sciences* 42.3, pp. 151–165 (Cited on page 191).
- Navarro, Carles, Maciej Majewski, and Gianni De Fabritiis (2023). “Top-down machine learning of coarse-grained protein force fields”. In: *Journal of Chemical Theory and Computation* 19.21, pp. 7518–7526 (Cited on page 122).
- Naylor, David et al. (2014). “The cost of the” s” in https”. In: *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*, pp. 133–140 (Cited on page 48).

References

- Nazem, Fatemeh et al. (2024). “Deep Attention Network for Identifying Ligand-Protein Binding Sites”. In: *Journal of Computational Science*, p. 102368 (Cited on page 115).
- Nicholson, David N and Casey S Greene (2020). “Constructing knowledge graphs and their biomedical applications”. In: *Computational and structural biotechnology journal* 18, pp. 1414–1428 (Cited on page 120).
- Nijkamp, Erik et al. (2023). “Progen2: exploring the boundaries of protein language models”. In: *Cell systems* 14.11, pp. 968–978 (Cited on pages 112, 127).
- Nimni, Marcel E and Robert D Harkness (2018). “Molecular structure and functions of collagen”. In: *Collagen*. CRC Press, pp. 1–78 (Cited on page 100).
- Nittari, Giulio et al. (2020). “Telemedicine practice: review of the current ethical and legal challenges”. In: *Telemedicine and e-Health* 26.12, pp. 1427–1437 (Cited on page 27).
- Niu, Zhaoyang, Guoqiang Zhong, and Hui Yu (2021). “A review on the attention mechanism of deep learning”. In: *Neurocomputing* 452, pp. 48–62 (Cited on page 117).
- Nivedhaa, N (2024). “Software architecture evolution: Patterns, trends, and best practices”. In: *Journal ID* 1664, p. 1400 (Cited on page 229).
- Notin, Pascal et al. (2024). “Machine learning for functional protein design”. In: *Nature biotechnology* 42.2, pp. 216–228 (Cited on page 112).
- Nováček, Vít et al. (2020). “Accurate prediction of kinase-substrate networks using knowledge graphs”. In: *PLoS computational biology* 16.12, e1007578 (Cited on page 121).
- O’Boyle, Noel M et al. (2011). “Open Babel: An open chemical toolbox”. In: *Journal of cheminformatics* 3.1, pp. 1–14 (Cited on page 210).
- O’Leary, Nuala A et al. (2016). “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”. In: *Nucleic acids research* 44.D1, pp. D733–D745 (Cited on page 208).
- Obermeyer, Ziad and Ezekiel J Emanuel (2016). “Predicting the future—big data, machine learning, and clinical medicine”. In: *The New England journal of medicine* 375.13, p. 1216 (Cited on page 28).
- Ofer, Dan and Michal Linial (2015). “ProFET: Feature engineering captures high-level protein functions”. In: *Bioinformatics* 31.21, pp. 3429–3436 (Cited on page 204).

- Ogunjobi, Taiwo Temitope et al. (2024). “Bioinformatics applications in chronic diseases: A comprehensive review of genomic, transcriptomics, proteomic, metabolomics, and machine learning approaches”. In: *Medinformatics* (Cited on page 74).
- Omboni, Stefano et al. (2022). “The worldwide impact of telemedicine during COVID-19: current evidence and recommendations for the future”. In: *Connected health* 1, p. 7 (Cited on page 27).
- Orengo, Christine A, David T Jones, and Janet M Thornton (1994). “Protein superfamilies and domain superfolds”. In: *Nature* 372.6507. Publisher: Nature Publishing Group UK London, pp. 631–634. ISSN: 0028-0836 (Cited on page 164).
- Orengo, Christine A, Alex D Michie, et al. (1997). “CATH—a hierarchic classification of protein domain structures”. In: *Structure* 5.8, pp. 1093–1109 (Cited on page 166).
- Orgel, Joseph PRO et al. (2006). “Microfibrillar structure of type I collagen in situ”. In: *Proceedings of the National Academy of Sciences* 103.24, pp. 9001–9005 (Cited on page 98).
- Osorio, Daniel, Paola Rondón-Villarreal, and Rodrigo Torres (2015). “Peptides: a package for data mining of antimicrobial peptides”. In: (Cited on pages 199, 200).
- Ozaybi, Mazen Qassem Bohais et al. (2024). “The Role of Artificial Intelligence in Drug Discovery and Development”. In: *Egyptian Journal of Chemistry* 67.13, pp. 1541–1547 (Cited on page 123).
- Ozols, Juris (1990). “[44] Amino acid analysis”. In: *Methods in enzymology*. Vol. 182. Elsevier, pp. 587–601 (Cited on page 211).
- Paananen, Jussi and Vittorio Fortino (2020). “An omics perspective on drug target discovery platforms”. In: *Briefings in bioinformatics* 21.6, pp. 1937–1953 (Cited on page 9).
- Page, M, Jane Hogg, and Doreen E Ashhurst (1986). “The effects of mechanical stability on the macromolecules of the connective tissue matrices produced during fracture healing. I. The collagens”. In: *The Histochemical Journal* 18.5, pp. 251–265 (Cited on page 91).
- Pakhrin, Subash C et al. (2021). “Deep learning-based advances in protein structure prediction”. In: *International journal of molecular sciences* 22.11, p. 5553 (Cited on pages 138, 139).
- Pan, Yi, Xiujuan Lei, and Yuchen Zhang (2022). “Association predictions of genomics, proteinomics, transcriptomics, microbiome, metabolomics, pathomics, radiomics,

References

- drug, symptoms, environment factor, and disease networks: A comprehensive approach". In: *Medicinal research reviews* 42.1, pp. 441–461 (Cited on page 9).
- Pang, Yihe and Bin Liu (2020). "SelfAT-Fold: protein fold recognition based on residue-based and motif-based self-attention networks". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.3, pp. 1861–1869 (Cited on page 118).
- Park, Yeonwoo, Brian PH Metzger, and Joseph W Thornton (2024). "The simplicity of protein sequence-function relationships". In: *Nature Communications* 15.1, p. 7953 (Cited on page 109).
- Parmar, Kashish and Indresh Kumar Verma (2024). "A Comprehensive Approach to Enhancing Doctor-Patient Interaction: Bridging the Gap for Better Healthcare." In: *International Journal of Online & Biomedical Engineering* 20.13 (Cited on page 222).
- Patel, Bhumi, Vijai Singh, and Dhaval Patel (2019). "Structural bioinformatics". In: *Essentials of Bioinformatics, Volume I: Understanding Bioinformatics: Genes to Proteins*, pp. 169–199 (Cited on page 144).
- Pattison, Amanda M et al. (2016). "Guanylyl cyclase C signaling axis and colon cancer prevention". In: *World journal of gastroenterology* 22.36, p. 8070 (Cited on page 95).
- Pearce, Robin and Yang Zhang (2021). "Deep learning techniques have significantly impacted protein structure prediction and protein design". In: *Current opinion in structural biology* 68, pp. 194–207 (Cited on page 114).
- Pearl, Judea (2022). "Causal Inference: history, perspectives, adventures, and unification (an interview with Judea Pearl)". In: *Observational Studies* 8.2, pp. 23–36 (Cited on page 228).
- Pearson, William R (2013). "An introduction to sequence similarity ("homology") searching". In: *Current protocols in bioinformatics* 42.1, pp. 3–1 (Cited on page 166).
- Pedersen, Anne Reff (2016). "The role of patient narratives in healthcare innovation: supporting translation and meaning making". In: *Journal of health organization and management* 30.2, pp. 244–257 (Cited on page 6).
- Pejler, Gunnar et al. (2010). "Mast cell proteases: multifaceted regulators of inflammatory disease". In: *Blood, The Journal of the American Society of Hematology* 115.24, pp. 4981–4990 (Cited on page 95).

- Perez-Riverol, Yasset et al. (2022). “The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences”. In: *Nucleic acids research* 50.D1, pp. D543–D552 (Cited on page 80).
- Pettersen, Eric F et al. (2021). “UCSF ChimeraX: Structure visualization for researchers, educators, and developers”. In: *Protein science* 30.1, pp. 70–82 (Cited on page 139).
- Pickup, Michael W, Janna K Mouw, and Valerie M Weaver (2014). “The extracellular matrix modulates the hallmarks of cancer”. In: *EMBO reports* 15.12, pp. 1243–1253 (Cited on page 104).
- Pile, James C (2006). “Evaluating postoperative fever”. In: *Cleveland Clinic journal of medicine* 73.1, S62–6 (Cited on page 8).
- Plebani, Mario (2016). “Harmonization in laboratory medicine: requests, samples, measurements and reports”. In: *Critical Reviews in Clinical Laboratory Sciences* 53.3, pp. 184–196 (Cited on page 2).
- Pollastri, Gianluca et al. (2002). “Prediction of coordination number and relative solvent accessibility in proteins”. In: *Proteins: Structure, Function, and Bioinformatics* 47.2, pp. 142–153 (Cited on page 213).
- Precedence Research (2024). *Protein Engineering Market Size, Share, and Trends 2024 to 2034*. <https://www.precedenceresearch.com/protein-engineering-market>. Accessed on March 10, 2025 (Cited on page 109).
- (2025). *Protein Therapeutics Market Size, Share, and Trends 2025 to 2034*. <https://www.precedenceresearch.com/protein-therapeutics-market>. Accessed on March 10, 2025 (Cited on page 127).
- Prianichnikov, Nikita et al. (2020). “MaxQuant software for ion mobility enhanced shotgun proteomics”. In: *Molecular & Cellular Proteomics* 19.6, pp. 1058–1069 (Cited on page 80).
- Price, W Nicholson and I Glenn Cohen (2019). “Privacy in the age of medical big data”. In: *Nature medicine* 25.1, pp. 37–43 (Cited on page 28).
- Prot pi (2023). *Prot pi—protein tool*. <https://www.protpi.ch/Calculator/ProteinTool>. Accessed 14 Mar 2023 (Cited on pages 199, 200).
- Protein Calculator v3.4 (2023). *Protein Calculator v3.4*. <https://protpi.sourceforge.net/>. Accessed 14 Mar 2023 (Cited on pages 199, 200).

References

- Quinn, Martha et al. (2019). “Electronic health records, communication, and data sharing: challenges and opportunities for improving the diagnostic process”. In: *Diagnosis* 6.3, pp. 241–248 (Cited on page 2).
- Quinn, Thomas P et al. (2022). “The three ghosts of medical AI: Can the black-box present deliver?” In: *Artificial intelligence in medicine* 124, p. 102158 (Cited on page 8).
- Raikar, Gokuldas (Vedant) Sarvesh, Amisha Sarvesh Raikar, and Sandesh Narayan Somnache (2023). “Advancements in artificial intelligence and machine learning in revolutionising biomarker discovery”. In: *Brazilian Journal of Pharmaceutical Sciences* 59, e23146 (Cited on page 74).
- Raj, Pethuru, Skylab Vanga, and Akshita Chaudhary (2022). *Cloud-Native Computing: How to design, develop, and secure microservices and event-driven applications*. John Wiley & Sons (Cited on page 229).
- Rajendran, Suraj et al. (2024). “Learning across diverse biomedical data modalities and cohorts: Challenges and opportunities for innovation”. In: *Patterns* 5.2 (Cited on page 225).
- Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane (2019). “Machine learning in medicine”. In: *New England Journal of Medicine* 380.14, pp. 1347–1358 (Cited on page 28).
- Raju, Rubin and Brian J Linder (2021). “Evaluation and management of pelvic organ prolapse”. In: *Mayo Clinic Proceedings*. Vol. 96. 12. Elsevier, pp. 3122–3129 (Cited on page 46).
- Raju, VN Ganapathi et al. (2020). “Study the influence of normalization/transformation process on the accuracy of supervised classification”. In: *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE, pp. 729–735 (Cited on page 65).
- Ramshaw, John AM, Naina K Shah, and Barbara Brodsky (1998). “Gly-XY tripeptide frequencies in collagen: a context for host–guest triple-helical peptides”. In: *Journal of structural biology* 122.1-2, pp. 86–91 (Cited on page 98).
- Réau, Manon et al. (2023). “DeepRank-GNN: a graph neural network framework to learn patterns in protein–protein interfaces”. In: *Bioinformatics* 39.1, btac759 (Cited on page 116).

- Remeseiro, Beatriz and Veronica Bolon-Canedo (2019). “A review of feature selection methods in medical applications”. In: *Computers in biology and medicine* 112, p. 103375 (Cited on page 65).
- Ren, Pengzhen et al. (2021). “A survey of deep active learning”. In: *ACM computing surveys (CSUR)* 54.9, pp. 1–40 (Cited on page 135).
- Reyero Lobo, Paula et al. (2023). “Semantic Web technologies and bias in artificial intelligence: A systematic literature review”. In: *Semantic Web* 14.4, pp. 745–770 (Cited on page 7).
- Ricoeur, Paul (2007). *On translation*. Routledge (Cited on page 1).
- Rives, Alexander (2025). “Language Models at the Scale of Evolution”. PhD thesis. New York University (Cited on page 76).
- Rives, Alexander et al. (2021). “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences”. In: *Proceedings of the National Academy of Sciences* 118.15, e2016239118 (Cited on pages 74, 118).
- Rogers, Rebecca G et al. (2003). “A short form of the pelvic organ prolapse/urinary incontinence sexual questionnaire (PISQ-12)”. In: *International Urogynecology Journal* 14, pp. 164–168 (Cited on page 52).
- Romero, Philip A and Frances H Arnold (2009). “Exploring protein fitness landscapes by directed evolution”. In: *Nature reviews Molecular cell biology* 10.12, pp. 866–876 (Cited on page 193).
- Rosati, Diletta et al. (2024). “Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: a review”. In: *Computational and structural biotechnology journal* (Cited on page 9).
- Rossert, Jerome and Benoit de Crombrughe (2002). “Type I collagen: structure, synthesis, and regulation”. In: *Principles of bone biology*. Elsevier, pp. 189–XVIII (Cited on page 100).
- Rost, Burkhard (1999). “Twilight zone of protein sequence alignments”. In: *Protein engineering* 12.2, pp. 85–94 (Cited on pages 166, 185).
- Rowland, Simon P et al. (2020). “What is the clinical value of mHealth for patients?” In: *NPJ digital medicine* 3.1, p. 4 (Cited on page 45).
- Saha, Soham et al. (2024). “Strategic Management for Innovation in Life Sciences: A Comprehensive Overview”. In: *Innovation in Life Sciences: The Digital Revolution*, pp. 191–212 (Cited on page 131).

References

- Sailer, Michael and Lisa Homner (2020). “The gamification of learning: A meta-analysis”. In: *Educational psychology review* 32.1, pp. 77–112 (Cited on page 155).
- Saini, Rahul et al. (2020). “Advanced biofuel production and road to commercialization: An insight into bioconversion potential of *Rhodospiridium* sp.” In: *Biomass and bioenergy* 132, p. 105439 (Cited on page 109).
- SAMSON Connect (2023). *SAMSON Connect Official website*. <https://www.samson-connect.net>. Accessed on February 27, 2023 (Cited on page 155).
- Santos, Alberto et al. (2022). “A knowledge graph to interpret clinical proteomics data”. In: *Nature biotechnology* 40.5, pp. 692–702 (Cited on page 120).
- Sarwar, Tabinda et al. (2022). “The secondary use of electronic health records for data mining: data characteristics and challenges”. In: *ACM Computing Surveys (CSUR)* 55.2, pp. 1–40 (Cited on page 7).
- Savage, Neil (2021). “Tapping into the drug discovery potential of AI”. In: *Nature. com*. <https://www.nature.com/articles/d43747-021-00045-7> (Cited on page 128).
- Schauperl, Michael and Rajiah Aldrin Denny (2022). “AI-based protein structure prediction in drug discovery: impacts and challenges”. In: *Journal of Chemical Information and Modeling* 62.13, pp. 3142–3156 (Cited on page 131).
- Schmid, Sakshi Yadav et al. (2023). “Mechanisms of Biomolecular Self-Assembly Investigated Through In Situ Observations of Structures and Dynamics”. In: *Angewandte Chemie International Edition* 62.48, e202309725 (Cited on page 112).
- Schmirler, Robert, Michael Heinzinger, and Burkhard Rost (2023). “Fine-tuning protein language models boosts predictions across diverse tasks”. In: *bioRxiv*, pp. 2023–12 (Cited on page 76).
- Schnoes, Alexandra M et al. (2009). “Annotation error in public databases: misannotation of molecular function in enzyme superfamilies”. In: *PLoS computational biology* 5.12, e1000605 (Cited on page 166).
- Schölkopf, Bernhard (2022). “Causality for machine learning”. In: *Probabilistic and causal inference: The works of Judea Pearl*, pp. 765–804 (Cited on page 228).
- Schwalbe, Nina and Brian Wahl (2020). “Artificial intelligence and the future of global health”. In: *The Lancet* 395.10236, pp. 1579–1586 (Cited on page 28).
- Selvaraj, Chandrabose, Ishwar Chandra, and Sanjeev Kumar Singh (2021). “Artificial intelligence and machine learning approaches for drug design: challenges and oppor-

- tunities for the pharmaceutical industries”. In: *Molecular diversity*, pp. 1–21 (Cited on page 132).
- Sen, Ilker et al. (2006). “The impact of urinary incontinence on female sexual function”. In: *Advances in therapy* 23, pp. 999–1008 (Cited on page 52).
- Seyhan, Attila A (2019). “Lost in translation: the valley of death across preclinical and clinical divide—identification of problems and overcoming obstacles”. In: *Translational Medicine Communications* 4.1, pp. 1–19 (Cited on page 221).
- Sf, Altschul (1997). “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”. In: *Nucleic acids research* 25, pp. 3389–3402 (Cited on page 139).
- Shabir, Ghulam and Florence Kenzie (2025). “Enhancing Patient-Doctor Communication with Digital Tools: A Path to Better Health Literacy and Outcomes”. In: (Cited on page 222).
- Shao, Xinhao et al. (2020). “MatrisomeDB: the ECM-protein knowledge database”. In: *Nucleic acids research* 48.D1, pp. D1136–D1144 (Cited on pages 91, 94).
- Shi, Qiangqiang et al. (2023). “Engineering precise sequence-defined polymers for advanced functions”. In: *Progress in Polymer Science* 141, p. 101677 (Cited on page 109).
- Shorten, Connor and Taghi M Khoshgoftaar (2019). “A survey on image data augmentation for deep learning”. In: *Journal of big data* 6.1, pp. 1–48 (Cited on page 64).
- Shoulders, Matthew D and Ronald T Raines (2009). “Collagen structure and stability”. In: *Annual review of biochemistry* 78, pp. 929–958 (Cited on page 96).
- Singh, Drishtant, Vikrant Rai, and Devendra K Agrawal (2023). “Regulation of collagen I and collagen III in tissue injury and regeneration”. In: *Cardiology and cardiovascular medicine* 7.1, p. 5 (Cited on page 91).
- Singhal, Peeyush et al. (2023). “Domain adaptation: challenges, methods, datasets, and applications”. In: *IEEE access* 11, pp. 6973–7020 (Cited on page 227).
- Skolnick, Jeffrey et al. (2021). “AlphaFold 2: why it works and its implications for understanding the relationships of protein sequence, structure, and function”. In: *Journal of chemical information and modeling* 61.10, pp. 4827–4831 (Cited on page 124).
- Smith, Malcolm H (1966). “The amino acid composition of proteins”. In: *Journal of Theoretical Biology* 13, pp. 261–282 (Cited on page 211).

References

- Smith, Zachary et al. (2020). “Discovering protein conformational flexibility through artificial-intelligence-aided molecular dynamics”. In: *The Journal of Physical Chemistry B* 124.38, pp. 8221–8229 (Cited on page 132).
- Söderberg, Marie Westergren (2008). *Studies on the extracellular matrix of the dysfunctional pelvic floor*. Karolinska Institutet (Sweden) (Cited on page 75).
- Soenksen, Luis R et al. (2022). “Integrated multimodal artificial intelligence framework for healthcare applications”. In: *NPJ digital medicine* 5.1, p. 149 (Cited on page 8).
- Solano-Flores, Guillermo, Eduardo Backhoff, and Luis Ángel Contreras-Niño (2009). “Theory of test translation error”. In: *International Journal of Testing* 9.2, pp. 78–91 (Cited on page 2).
- Sorushanova, Anna et al. (2021). “The influence of animal species, gender and tissue on the structural, biophysical, biochemical and biological properties of collagen sponges”. In: *Journal of Materials Science: Materials in Medicine* 32, pp. 1–12 (Cited on page 100).
- Sousa, Rita T, Sara Silva, and Catia Pesquita (2024). “Explaining protein–protein interactions with knowledge graph-based semantic similarity”. In: *Computers in Biology and Medicine* 170, p. 108076 (Cited on page 121).
- Srikrishna, Sushma, Dudley Robinson, and Linda Cardozo (2010). “Validation of the Patient Global Impression of Improvement (PGI-I) for urogenital prolapse”. In: *International urogynecology journal* 21, pp. 523–528 (Cited on page 52).
- Stausberg, Jürgen et al. (2008). “Reliability of diagnoses coding with ICD-10”. In: *International journal of medical informatics* 77.1, pp. 50–57 (Cited on page 6).
- Stevens, Amy O and Yi He (2022). “Benchmarking the accuracy of AlphaFold 2 in loop structure prediction”. In: *Biomolecules* 12.7, p. 985 (Cited on page 124).
- Sun, Bo (2021). “The mechanics of fibrillar collagen extracellular matrix”. In: *Cell Reports Physical Science* 2.8 (Cited on page 100).
- Suwardi, Ady et al. (2022). “Machine learning-driven biomaterials evolution”. In: *Advanced Materials* 34.1, p. 2102703 (Cited on page 112).
- Szklarczyk, Damian et al. (2019). “STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets”. In: *Nucleic acids research* 47.D1, pp. D607–D613 (Cited on pages 205, 207).

- Tanabe, Lorraine and W John Wilbur (2002). “Tagging gene and protein names in biomedical text”. In: *Bioinformatics* 18.8, pp. 1124–1132 (Cited on page 9).
- Tashildar, Aakanksha et al. (2020). “Application development using flutter”. In: *International Research Journal of Modernization in Engineering Technology and Science* 2.8, pp. 1262–1266 (Cited on page 48).
- Tatsukawa, Hideki and Kiyotaka Hitomi (2021). “Role of transglutaminase 2 in cell death, survival, and fibrosis”. In: *Cells* 10.7, p. 1842 (Cited on page 95).
- Taylor, Sean C and Anton Posch (2014). “The design of a quantitative western blot experiment”. In: *BioMed research international* 2014.1, p. 361590 (Cited on page 103).
- Teleman, PIA et al. (2011). “Validation of the Swedish short forms of the pelvic floor impact questionnaire (PFIQ-7), pelvic floor distress inventory (PFDI-20) and pelvic organ prolapse/urinary incontinence sexual questionnaire (PISQ-12)”. In: *Acta obstetrica et gynecologica Scandinavica* 90.5, pp. 483–487 (Cited on page 52).
- Terui, Sachiko (2017). “Conceptualizing the pathways and processes between language barriers and health disparities: review, synthesis, and extension”. In: *Journal of Immigrant and Minority Health* 19, pp. 215–224 (Cited on page 7).
- Teufel, Felix et al. (2022). “SignalP 6.0 predicts all five types of signal peptides using protein language models”. In: *Nature biotechnology* 40.7, pp. 1023–1025 (Cited on page 79).
- Thirunavukarasu, Arun James et al. (2023). “Large language models in medicine”. In: *Nature medicine*, pp. 1–11 (Cited on page 41).
- Thorn, Alan (2015). *Unity Animation Essentials*. Packt Publishing Ltd (Cited on page 49).
- Tian, Zhao et al. (2024). “The difference in extracellular matrix metabolism in women with and without pelvic organ prolapse: A systematic review and meta-analysis”. In: *BJOG: An International Journal of Obstetrics & Gynaecology* (Cited on page 75).
- Tilahun, Shimeles et al. (2022). “Prestorage high CO₂ and 1-MCP treatment reduce chilling injury, prolong storability, and maintain sensory qualities and antioxidant activities of ”Madoka” peach fruit”. In: *Frontiers in Nutrition* 9, p. 903352 (Cited on page 167).
- Tinwell, Angela (2014). *The uncanny valley in games and animation*. CRC press (Cited on page 31).

References

- Tiwari, Abhishek and Arvind KT Sekhar (2007). “Workflow based framework for life science informatics”. In: *Computational biology and chemistry* 31.5-6, pp. 305–319 (Cited on page 226).
- Todd, Stephen et al. (2021). “CSynth: an interactive modelling and visualization tool for 3D chromatin structure”. In: *Bioinformatics* 37.7, pp. 951–955 (Cited on page 139).
- Tokmakov, Alexander A, Atsushi Kurotani, and Ken-Ichi Sato (2021). “Protein pI and intracellular localization”. In: *Frontiers in Molecular Biosciences* 8, p. 775736 (Cited on page 212).
- Tong, Li et al. (2023). “Integrating multi-omics data with EHR for precision medicine using advanced artificial intelligence”. In: *IEEE Reviews in Biomedical Engineering* 17, pp. 80–97 (Cited on page 9).
- Torrìsi, Mirko, Gianluca Pollastri, and Quan Le (2020). “Deep learning methods in protein structure prediction”. In: *Computational and Structural Biotechnology Journal* 18, pp. 1301–1310 (Cited on page 114).
- Tsang, Siny, Colin F Royse, and Abdullah Sulieman Terkawi (2017). “Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine”. In: *Saudi journal of anaesthesia* 11.Suppl 1, S80 (Cited on page 30).
- Tuccori, Marco et al. (2021). “An overview of the preclinical discovery and development of bamlanivimab for the treatment of novel coronavirus infection (COVID-19): reasons for limited clinical use and lessons for the future”. In: *Expert Opinion on Drug Discovery* 16.12, pp. 1403–1414 (Cited on page 109).
- “UniProt: the universal protein knowledgebase in 2021” (2021). In: *Nucleic acids research* 49.D1, pp. D480–D489 (Cited on page 166).
- “UniProt: the universal protein knowledgebase in 2023” (2023). In: *Nucleic acids research* 51.D1, pp. D523–D531 (Cited on pages 205, 207).
- “UniProt: the universal protein knowledgebase in 2025” (2025). In: *Nucleic acids research* 53.D1, pp. D609–D617 (Cited on pages 77, 91, 94).
- Unni, Deepak R et al. (2022). “Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science”. In: *Clinical and translational science* 15.8, pp. 1848–1855 (Cited on page 121).
- Van Oss, Carel Jan (1997). “Hydrophobicity and hydrophilicity of biosurfaces”. In: *Current opinion in colloid & interface science* 2.5, pp. 503–512 (Cited on page 211).

- Vig, Jesse et al. (2020). “Bertology meets biology: Interpreting attention in protein language models”. In: *arXiv preprint arXiv:2006.15222* (Cited on page 118).
- Vila, Jorge A (2020). “About the protein space vastness”. In: *The Protein Journal* 39.5, pp. 472–475 (Cited on page 191).
- Vitorino, Rui (2024). “Transforming clinical research: the power of high-throughput omics integration”. In: *Proteomes* 12.3, p. 25 (Cited on page 9).
- Vodovotz, Yoram and Gary An (2014). *Translational systems biology: concepts and practice for the future of biomedical research*. Elsevier (Cited on page 4).
- Voltà-Durán, Eric et al. (2020). “Controlling self-assembling and tumor cell-targeting of protein-only nanoparticles through modular protein engineering”. In: *Science China Materials* 63, pp. 147–156 (Cited on page 192).
- Waagmeester, Andra et al. (2020). “Wikidata as a knowledge graph for the life sciences”. In: *Elife* 9, e52614 (Cited on page 121).
- Walker, Sidney P, Venkata VB Yallapragada, and Mark Tangney (2021). “Arming yourself for the in silico protein design revolution”. In: *Trends in Biotechnology* 39.7, pp. 651–664 (Cited on pages 109, 130, 131, 137, 138).
- Wang, Jingxue et al. (2018). “Computational protein design with deep learning neural networks”. In: *Scientific reports* 8.1, pp. 1–9 (Cited on page 114).
- Wang, Shunzhi et al. (2024). “Bond-centric modular design of protein assemblies”. In: *bioRxiv*, pp. 2024–10 (Cited on page 112).
- Wang, Shuyu et al. (2023). “ProS-GNN: predicting effects of mutations on protein stability using graph neural networks”. In: *Computational Biology and Chemistry* 107, p. 107952 (Cited on page 116).
- Wang, Ting et al. (2017). “7 × 7 RMSD matrix: A new method for quantitative comparison of the transmembrane domain structures in the G-protein coupled receptors”. In: *Journal of structural biology* 199.2, pp. 87–101 (Cited on page 145).
- Wang, X et al. (2002). “Effects of Collagen Unwinding and Cleavage on the Mechanical Integrity of the Collagen Network in Bone”. In: *bone* 4, p. 11 (Cited on page 91).
- Watson, Joseph L et al. (2023). “De novo design of protein structure and function with RFdiffusion”. In: *Nature* 620.7976, pp. 1089–1100 (Cited on pages 117, 123).
- Webb, Benjamin and Andrej Sali (2016). “Comparative protein structure modeling using MODELLER”. In: *Current protocols in bioinformatics* 54.1, pp. 5–6 (Cited on page 209).

References

- Wei, Lesong et al. (2021). “ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism”. In: *Briefings in Bioinformatics* 22.5, bbab041 (Cited on page 117).
- Wei, Wenxiu et al. (2020). “Fsp3: A new parameter for drug-likeness”. In: *Drug Discovery Today* 25.10, pp. 1839–1845 (Cited on page 185).
- Weintraub, Adi Y, Hannah Gliner, and Naama Marcus-Braun (2019). “Narrative review of the epidemiology, diagnosis and pathophysiology of pelvic organ prolapse”. In: *International braz j urol* 46, pp. 5–14 (Cited on page 46).
- Whisstock, James C and Arthur M Lesk (2003). “Prediction of protein function from protein sequence and structure”. In: *Quarterly reviews of biophysics* 36.3, pp. 307–340 (Cited on page 10).
- Wikipedia (2023a). *Protein pKa calculations*. Accessed 14 Mar 2023. URL: https://en.wikipedia.org/wiki/Protein_pKa_calculations#cite_note-Hass2015-1 (Cited on page 197).
- (2023b). *Relative accessible surface area*. https://en.m.wikipedia.org/wiki/Relative_accessible_surface_area. Accessed on February 27, 2023 (Cited on page 148).
- Wilkins, Maggie F and Jennifer M Wu (2016). “Epidemiology of pelvic organ prolapse”. In: *Current Obstetrics and Gynecology Reports* 5, pp. 119–123 (Cited on page 45).
- Williams, Christopher J et al. (2018). “MolProbity: More and better reference data for improved all-atom structure validation”. In: *Protein Science* 27.1, pp. 293–315 (Cited on page 208).
- Wu, Fang et al. (2023). “When geometric deep learning meets pretrained protein language models”. In: *bioRxiv*, pp. 2023–01 (Cited on page 116).
- Wu, Hongjie et al. (2024). “AttentionMGT-DTA: A multi-modal drug-target affinity prediction using graph transformer and attention mechanism”. In: *Neural Networks* 169, pp. 623–636 (Cited on page 118).
- Wu, Ruidong et al. (2022). “High-resolution de novo structure prediction from primary sequence”. In: *BioRxiv*, pp. 2022–07 (Cited on page 127).
- Wu, Ruo et al. (2022). “Promising therapeutic approaches of utrophin replacing dystrophin in the treatment of Duchenne muscular dystrophy”. In: *Fundamental Research* 2.6, pp. 885–893 (Cited on page 95).

- Wynn, Thomas A (2008). “Cellular and molecular mechanisms of fibrosis”. In: *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 214.2, pp. 199–210 (Cited on page 104).
- Xia, Chunqiu et al. (2022). “Fast protein structure comparison through effective representation learning with contrastive graph neural networks”. In: *PLoS computational biology* 18.3, e1009986 (Cited on page 167).
- Xia, Xuhua (2007). “Protein isoelectric point”. In: *Bioinformatics and the cell: Modern computational approaches in genomics, proteomics and transcriptomics*, pp. 207–219 (Cited on page 212).
- Xiao, Yijia et al. (2021). “Modeling protein using large-scale pretrain language model”. In: *arXiv preprint arXiv:2108.07435* (Cited on page 120).
- Xie, Qizhe et al. (2020). “Unsupervised data augmentation for consistency training”. In: *Advances in neural information processing systems* 33, pp. 6256–6268 (Cited on page 66).
- Xie, Zhuorui et al. (2021). “Gene set knowledge discovery with Enrichr”. In: *Current protocols* 1.3, e90 (Cited on page 84).
- Xie, Ziwei et al. (2025). “Identification of potential drug targets for pelvic organ prolapse using a proteome-wide Mendelian randomization approach”. In: *Scientific Reports* 15.1, p. 8291 (Cited on page 107).
- Xu, Hanhui and Kyle Michael James Shuttleworth (2024). “Medical artificial intelligence and the black box problem: a view based on the ethical principle of “do no harm””. In: *Intelligent Medicine* 4.1, pp. 52–57 (Cited on page 3).
- Xu, Xi et al. (2024). “A comprehensive review on synergy of multi-modal data and ai technologies in medical diagnosis”. In: *Bioengineering* 11.3, p. 219 (Cited on page 7).
- Xue, Lina et al. (2020). “Learning relations of knowledge transfer (KT) and knowledge integration (KI) of doctoral students during online interdisciplinary training: an exploratory study”. In: *Higher education research & development* 39.6, pp. 1290–1307 (Cited on pages 1, 2).
- Yallapragada, Venkata V. B. et al. (Mar. 2020). “Function2Form Bridge—Toward synthetic protein holistic performance prediction”. en. In: *Proteins: Structure, Function, and Bioinformatics* 88.3, pp. 462–475. ISSN: 0887-3585, 1097-0134. DOI:

References

- 10.1002/prot.25825. URL: <https://onlinelibrary.wiley.com/doi/10.1002/prot.25825> (visited on 09/27/2022) (Cited on page 165).
- Yallapragada, Venkata VB, Sidney P Walker, et al. (2020). “Function2Form Bridge—Toward synthetic protein holistic performance prediction”. In: *Proteins: Structure, Function, and Bioinformatics* 88.3, pp. 462–475 (Cited on pages 137, 140, 146).
- Yallapragada, Venkata VB, Tianshu Xu, et al. (2021). “Pepblock builder VR—an open-source tool for gaming-based bio-edutainment in interactive protein design”. In: *Frontiers in Bioengineering and Biotechnology* 9, p. 674211 (Cited on page 139).
- Yang, Jianyi and Yang Zhang (2015). “I-TASSER server: new development for protein structure and function predictions”. In: *Nucleic acids research* 43.W1, W174–W181 (Cited on page 139).
- Yang, Jincai, Cheng Shen, and Niu Huang (2020). “Predicting or pretending: artificial intelligence for protein-ligand interactions lack of sufficiently large and unbiased datasets”. In: *Frontiers in pharmacology* 11, p. 69 (Cited on page 133).
- Yang, Kevin K, Zachary Wu, and Frances H Arnold (2019). “Machine-learning-guided directed evolution for protein engineering”. In: *Nature methods* 16.8, pp. 687–694 (Cited on pages 122, 192).
- Yang, Yi Isaac et al. (2019). “Enhanced sampling in molecular dynamics”. In: *The Journal of chemical physics* 151.7 (Cited on page 122).
- Yang, Zhenyu et al. (2023). “AlphaFold2 and its applications in the fields of biology and medicine”. In: *Signal Transduction and Targeted Therapy* 8.1, p. 115 (Cited on page 123).
- Yang, Zongbao et al. (2023). “Interpretable Disease Prediction via Path Reasoning over medical knowledge graphs and admission history”. In: *Knowledge-Based Systems* 281, p. 111082 (Cited on page 8).
- Ye, Jian, Scott McGinnis, and Thomas L Madden (2006). “BLAST: improvements for better sequence analysis”. In: *Nucleic acids research* 34.suppl_2, W6–W9 (Cited on page 166).
- Yellavula, Naren (2017). *Building RESTful Web services with Go: Learn how to build powerful RESTful APIs with Golang that scale gracefully*. Packt Publishing Ltd (Cited on page 176).

- Yu, Chi-Hua and Markus J Buehler (2020). “Sonification based de novo protein design using artificial intelligence, structure prediction, and analysis using molecular modeling”. In: *APL bioengineering* 4.1 (Cited on page 121).
- Yuan, Shuguang, HC Stephen Chan, and Zhenquan Hu (2017). “Using PyMOL as a platform for computational drug design”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 7.2, e1298 (Cited on page 139).
- Zacharias, Jan and Ernst-Walter Knapp (2014). “Protein secondary structure classification revisited: processing DSSP information with PSSC”. In: *Journal of chemical information and modeling* 54.7, pp. 2166–2179 (Cited on page 172).
- Zeadally, Sherali and Oladayo Bello (2021). “Harnessing the power of Internet of Things based connectivity to improve healthcare”. In: *Internet of Things* 14, p. 100074 (Cited on page 41).
- Zeadally, Sherali, Farhan Siddiqui, et al. (2020). “Smart healthcare: Challenges and potential solutions using internet of things (IoT) and big data analytics”. In: *PSU research review* 4.2, pp. 149–168 (Cited on page 27).
- Zeinalipour, Kamyar et al. (2024). “Design Proteins Using Large Language Models: Enhancements and Comparative Analyses”. In: *arXiv preprint arXiv:2408.06396* (Cited on page 119).
- Zemla, Adam (2003). “LGA: a method for finding 3D similarities in protein structures”. In: *Nucleic acids research* 31.13, pp. 3370–3374 (Cited on page 145).
- Zhang, Chenyue et al. (2024). “The Historical Evolution and Significance of Multiple Sequence Alignment in Molecular Structure and Function Prediction”. In: *Biomolecules* 14.12, p. 1531 (Cited on page 76).
- Zhang, He et al. (2020). “Description and climate simulation performance of CAS-ESM version 2”. In: *Journal of Advances in Modeling Earth Systems* 12.12, e2020MS002210 (Cited on page 125).
- Zhang, Shuke et al. (2023). “SS-GNN: a simple-structured graph neural network for affinity prediction”. In: *ACS omega* 8.25, pp. 22496–22507 (Cited on page 116).
- Zhang, Yang and Jeffrey Skolnick (2005). “TM-align: a protein structure alignment algorithm based on the TM-score”. In: *Nucleic acids research* 33.7, pp. 2302–2309 (Cited on pages 166, 209).
- Zhang, Yuan et al. (2020). “Prodconn-protein design using a convolutional neural network”. In: *Biophysical Journal* 118.3, 43a–44a (Cited on page 114).

References

- Zhang, Zhidian et al. (2024). “Protein language models learn evolutionary statistics of interacting sequence motifs”. In: *bioRxiv*, pp. 2024–01 (Cited on page 125).
- Zhang, Zhiqiang et al. (2019). “Deep learning in omics: a survey and guideline”. In: *Briefings in functional genomics* 18.1, pp. 41–57 (Cited on page 9).
- Zhao, Qichang et al. (2022). “HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism”. In: *Bioinformatics* 38.3, pp. 655–662 (Cited on page 134).
- Zhao, Xinran O et al. (2022). “Mast cell chymase regulates extracellular matrix remodeling-related events in primary human small airway epithelial cells”. In: *Journal of Allergy and Clinical Immunology* 150.6, pp. 1534–1544 (Cited on page 95).
- Zheng, Alice and Amanda Casari (2018). *Feature engineering for machine learning: principles and techniques for data scientists.* ” O’Reilly Media, Inc.” (Cited on page 205).
- Zhou, Huajun et al. (2024). “Multimodal data integration for precision oncology: Challenges and future directions”. In: *arXiv preprint arXiv:2406.19611* (Cited on page 225).
- Zhou, Jing et al. (2021). “Flipda: Effective and robust data augmentation for few-shot learning”. In: *arXiv preprint arXiv:2108.06332* (Cited on page 222).
- Zhou, Yue et al. (2024). “A survey on data augmentation in large model era”. In: *arXiv preprint arXiv:2401.15422* (Cited on page 222).
- Zhu, Jie et al. (2021). “Protein assembly by design”. In: *Chemical reviews* 121.22, pp. 13701–13796 (Cited on pages 110, 112).
- Zohora, Fatema Tuz et al. (2019). “DeepIso: a deep learning model for peptide feature detection from LC-MS map”. In: *Scientific reports* 9.1, p. 17168 (Cited on page 115).