

Dissecting Biomolecular Interactions Using Deep Learning Approaches

Zur Erlangung des akademischen Grades einer
Doktorin der Ingenieurwissenschaften (DR. -ING.)
von der KIT-Fakultät für Chemieingenieurwesen und Verfahrenstechnik
des Karlsruher Instituts für Technologie (KIT)

genehmigte
DISSERTATION

von
M. Sc. Bahar Dadfar
aus Shiraz, Iran

Tag der mündlichen Prüfung: 27.11.2024

Erstgutachter: Prof. Dr.-Ing Matthias Franzreb

Zweitgutachter: Prof. Dr. Jörg Lahann

Dedication

This dissertation is dedicated to my family, whose steadfast love, support, and encouragement have been my guiding light throughout this journey. To my parents, for instilling in me the value of education and perseverance, and for always believing in my dreams.

To my brother, Amir Sina, and my sister, Faranak, who stood by me during the toughest times, offering their unwavering encouragement and companionship.

This work is for all those who continue to pursue knowledge and growth, with the hope that it may contribute in some small way to the advancement of science and understanding.

Acknowledgements

The completion of this PhD dissertation would not have been possible without the support, guidance, and encouragement of many individuals, to whom I am deeply grateful.

First and foremost, I would like to express my heartfelt gratitude to my advisor, Prof. Dr. Jörg Lahann, for giving me the opportunity to join his lab and become a part of his academic family. I am deeply grateful for the opportunity to work with him, as he continually pushed me beyond my limits, especially during moments of doubt. He taught me to think creatively and approach new projects with fresh perspectives. His humility, sharp intellect, and scientific expertise have greatly contributed to my growth as a scientist.

Throughout my Ph.D. journey, I have been fortunate to have the guidance of two advisors. I would especially like to extend my gratitude to Prof. Dr. Matthias Franzreb, who welcomed me into his academic group and treated me as part of his scholarly family. His humility and approachability made working with him an incredibly rewarding experience, and he was always available whenever I needed guidance or support.

A special thanks to my family—my parents, for their unconditional love, endless support, and belief in me. To my siblings, your encouragement during difficult moments meant the world to me, and I am incredibly grateful for your presence in my life. I would like to sincerely thank my uncle Parviz for his constant support and encouragement, which have been a great source of inspiration throughout my studies. His kindness and belief in me have meant more than words can express.

I have also been fortunate to be surrounded by an incredible group of friends and colleagues, whose support, humor, and companionship helped brighten many challenging days. A special thank you to Tahereh, Safoura, Gözde, and Martina for your friendship and

invaluable advice throughout this journey. I am deeply grateful to my friends from Iran, Sahar and Hossein, for their unwavering support in every possible way. I feel truly blessed to have you both in my life.

I would like to express my gratitude to Bianca Posselt, Dr. Erik Strandberg, and Dr. Stefan Heißler for their assistance with the CD spectroscopy. I also want to acknowledge my student, Cristian Haret, for his dedicated contributions to this work. Additionally, I extend my thanks to my colleagues and lab members, both from the German and US sides, for making this journey memorable and for their valuable feedback and insightful discussions.

Lastly, special thanks go out to Dr. Angela Weiss, Astrid Biedermann, Karin Wölk, and Stefanie Sellheim-Ret for helping with all administrative matters at the Institute of Functional Interfaces (IFG).

Abstract

Immunoglobulins play a vital role in both biological processes and biotechnological applications, particularly due to their centrality in immune responses and therapeutic drug design. Understanding how these proteins interact is crucial for advancing immunology and developing targeted therapies. Despite significant progress in molecular biology, accurately predicting and manipulating protein-protein interactions (PPIs) remains a challenge, hindering breakthroughs in areas like drug discovery and diagnostics. However, the advent of advanced deep-learning techniques presents new opportunities to overcome these challenges by enabling rapid and precise predictions of PPIs. This study is motivated by the potential of these computational tools to enhance our understanding of immunoglobulin interactions, thereby contributing to innovation in biotechnology and therapeutic development. In this study, we have adapted a widely accessible Convolutional Neural Network (CNN) to achieve species-specific classification of various immunoglobulin G (IgG) complexes. We prepared droplets of different immunoglobulins mixed with the B-cell superantigen (SAg), recombinant staphylococcal Protein A, and deposited them onto hydrophobic polymer substrates. These protein stains were then imaged using polarized light microscopy (PLM).

Our extensive analysis based on 23,745 images revealed that the pre-trained CNN, InceptionV3, not only successfully categorized IgGs from four different species, but also predicted their relative binding strength to Protein A. Averaged over 36 binding pairs, we observed (i) an overall accuracy of 81.4%, (ii) the highest prediction accuracies for human IgG, the antibody with the highest binding affinity for Protein A, and that (iii) the classification accuracy regarding the various IgG/Protein A ratios generally correlates with the binding strength of the protein-protein-complex as determined via Circular Dichroism spectroscopy (CD). Furthermore, the CNN, initially trained with IgG/Protein A stain images, was tested with

a new set of images using a different superantigen, recombinant Protein G. Remarkably, despite the unfamiliar superantigen, the CNN correctly classified the binding strength of human IgG and Protein G, achieving a 94% accuracy across various molar binding ratios as the most similar IgG complex exist in the training dataset.

Additionally, graph theory analysis was employed to augment the image-based approach with a parameter-based neural network strategy. This innovative approach was inspired by the observation of structural crystal patterns in different protein samples. Graph theory, known for its versatile applications across various scientific disciplines, was employed to convert images into graphs. Using the StructuralGT python package developed at the University of Michigan, we extracted a set of meaningful features from these graphs to serve as input datasets. This method was compared to traditional convolutional neural network results.

The study found that by using the graph-derived features as an input dataset for a designed neural network, the required time for training was significantly reduced (about 3 times) compared to image-based classification, while still maintaining high accuracy levels in the optimized scheme.

The findings underscore the potential of combining graph theory with deep learning for protein interaction analysis. Appropriately graph-based features can be used to predict protein-protein interactions beyond the initial training dataset. This approach is simplified by the processing of numerical data and enables classification on non-GPU-dependent systems, reducing both computational cost and training time. The results suggest a promising method for biological graph-like image classification and protein interaction strength prediction, which is beneficial for protein engineering, understanding self-aggregation, and maintaining protein stability in complex environments.

Kurzzusammenfassung

Immunglobuline spielen sowohl bei biologischen Prozessen als auch bei biotechnologischen Anwendungen eine wichtige Rolle, insbesondere aufgrund ihrer zentralen Bedeutung für Immunreaktionen und die Entwicklung therapeutischer Arzneimittel. Um die Immunologie voranzubringen und gezielte Therapien zu entwickeln, ist es entscheidend zu verstehen, wie diese Proteine interagieren. Trotz bedeutender Fortschritte in der Molekularbiologie bleibt die genaue Vorhersage und Beeinflussung von Protein-Protein-Interaktionen (PPIs) eine Herausforderung, die den Durchbruch in Bereichen wie der Arzneimittellentdeckung und Diagnostik behindert. Das Aufkommen fortschrittlicher Deep-Learning-Techniken eröffnet jedoch neue Möglichkeiten zur Überwindung dieser Herausforderungen, indem es schnelle und präzise Vorhersagen von PPIs ermöglicht. Diese Studie ist motiviert durch das Potenzial dieser computergestützten Werkzeuge, unser Verständnis von Immunglobulin-Interaktionen zu verbessern und dadurch zu Innovationen in der Biotechnologie und therapeutischen Entwicklung beizutragen. In dieser Studie haben wir ein weithin zugängliches neuronales Faltungsnetzwerk (Convolutional Neural Network, CNN) angepasst, um eine speziesspezifische Klassifizierung verschiedener Immunglobulin G (IgG) Komplexe zu erreichen. Wir haben Tröpfchen verschiedener Immunglobuline gemischt mit dem B-Zell-Superantigen (SAg), rekombinantem Staphylokokkenprotein A, hergestellt und auf hydrophobe Polymersubstrate aufgebracht. Diese Proteinfärbungen wurden dann mit Hilfe der Polarisationslichtmikroskopie (PLM) abgebildet.

Unsere umfassende Analyse auf der Grundlage von 23.745 Bildern ergab, dass das vortrainierte CNN, InceptionV3, nicht nur IgGs aus vier verschiedenen Spezies erfolgreich kategorisierte, sondern auch ihre relative Bindungsstärke an Protein A vorhersagte. Im Durchschnitt von 36 Bindungspaaren beobachteten wir (i) eine Gesamtgenauigkeit von 81.4%,

(ii) die höchste Vorhersagegenauigkeit für humanes IgG, den Antikörper mit der höchsten Bindungsaffinität für Protein A, und dass (iii) die Klassifizierungsgenauigkeit für die verschiedenen IgG/Protein A-Verhältnisse im Allgemeinen mit der Bindungsstärke des Protein-Protein-Komplexes korreliert, die mittels Circular dichroismus-Spektroskopie (CD) bestimmt wurde. Darüber hinaus wurde das CNN, das ursprünglich mit IgG/Protein A-Farbbildern trainiert wurde, mit einem neuen Satz von Bildern getestet, bei denen ein anderes Superantigen, rekombinantes Protein G, verwendet wurde. Bemerkenswerterweise klassifizierte das CNN trotz des ungewohnten Superantigens die Bindungsstärke von humanem IgG und Protein G korrekt und erreichte eine Genauigkeit von 94 % über verschiedene molare Bindungsverhältnisse hinweg, da der ähnlichste IgG-Komplex im Trainingsdatensatz vorhanden war.

Darüber hinaus wurde eine graphentheoretische Analyse eingesetzt, um den bildbasierten Ansatz mit einer parameterbasierten neuronalen Netzwerkstrategie zu ergänzen. Dieser innovative Ansatz wurde durch die Beobachtung von strukturellen Kristallmustern in verschiedenen Proteinproben inspiriert. Die Graphentheorie, die für ihre vielseitigen Anwendungen in verschiedenen wissenschaftlichen Disziplinen bekannt ist, wurde eingesetzt, um Bilder in Graphen umzuwandeln. Mit Hilfe des an der Universität von Michigan entwickelten Python-Pakets StructuralGT extrahierten wir aus diesen Graphen eine Reihe aussagekräftiger Merkmale, die als Eingabedatensätze dienten. Diese Methode wurde mit den Ergebnissen herkömmlicher neuronaler Faltungsnetzwerke verglichen.

Die Studie ergab, dass durch die Verwendung der aus den Graphen abgeleiteten Merkmale als Eingabedatensatz für ein entwickeltes neuronales Netz die erforderliche Trainingszeit im Vergleich zur bildbasierten Klassifizierung erheblich (etwa um das Dreifache) reduziert werden konnte, wobei die Genauigkeit im optimierten Schema dennoch hoch blieb.

Die Ergebnisse unterstreichen das Potenzial der Kombination von Graphentheorie und

Deep Learning für die Proteininteraktionsanalyse. Geeignete graphbasierte Merkmale können zur Vorhersage von Protein-Protein-Interaktionen über den ursprünglichen Trainingsdatensatz hinaus verwendet werden. Dieser Ansatz wird durch die Verarbeitung numerischer Daten vereinfacht und ermöglicht die Klassifizierung auf nicht-GPU-abhängigen Systemen, was sowohl die Rechenkosten als auch die Trainingszeit reduziert. Die Ergebnisse deuten auf eine vielversprechende Methode zur Klassifizierung von biologischen Graphen und zur Vorhersage der Stärke von Proteininteraktionen hin, die für das Protein-Engineering, das Verständnis der Selbstaggregation und die Aufrechterhaltung der Proteinstabilität in komplexen Umgebungen von Nutzen ist.

List of Abbreviations

AI	Artificial Intelligence
BCR	B-Cell Receptor
CAR	Chimeric Antigen Receptor
CD	Circular Dichroism
CDR	Complementary Determining Region
CH	Heavy chain constant domain
CL	Light chain constant domain
CNN	Convolutional Neural Network
Co-IP	Co-Immunoprecipitation
CVD	Chemical Vapor Deposition
ELISA	Enzyme-linked Immunosorbent Assay
ELU	Exponential Linear Unit
Fab	Fragment Antigen-binding Region
Fc	Fragment Crystallizable Region
FcγR	Fc gamma receptor
FRET	Fluorescence Resonance Energy Transfer
GNN	Graph Neural Network
Grad-CAM	Gradient-weighted Class Activation Mapping
GT	Graph Theory

HP	High Performance
HSA	Human Serum Albumin
IgG	Immunoglobulin G
IHC	Immunohistochemistry
IP	Immunoprecipitation
KNN	K-Nearest Neighbors
mAb	Monoclonal Antibody
MIA	Multi-Image Alignment
MOCN	Metal–organic covalent network
MRMR	Minimum Redundancy Maximum Relevance
ML	Machine Learning
MS	Mass Spectrometry
NMR	Nuclear Magnetic Resonance (NMR)
pAb	Polyclonal Antibody
PLM	Polarized Light Microscopy
PpL	Peptostreptococcal Protein L
PPI	Protein-Protein Interaction
PPX	Poly-(<i>p</i> -Xylylene)
ReLU	Rectified Linear Unit
SAg	Superantigen
ScAb	Single-chain Antibody

SEB	Surface Energy Balance
SEM	Scanning Electron Microscopy
SGD	Stochastic Gradient Descent
SPEA	Streptococcal Pyrogenic Exotoxin A
SpA	Staphylococcal Protein A
SpG	Streptococcal Protein G
SVM	Support Vector Machine
TAP	Tandem Affinity Purification
TCR	Target Specific Region
ToF-SIMS	Time-of-Flight Secondary Ion Mass Spectrometry
t-SNE	t-Distributed Stochastic Neighbor Embedding
UV	Ultraviolet
VH	Heavy chain variable domain
VL	Light chain variable domain
WB	Western Blotting
WI	Wiener Index
WHO	World Health Organization
Y2H	Yeast Two Hybrid

List of Symbols

Latin Symbols

A	Absorbance value
A	Adjacency matrix
b	Path length [cm]
c	Protein concentration [mol/L]
C_B	Betweenness centrality
C_C	Closeness centrality
d	Network diameter
e	Number of edges
E_{glob}	Global efficiency
F_v	Variable region of the antibody
HCl	Hydrogen chloride
i	Pertaining to node i
Inf	Infinity
k	Graph degree
K_a	Association constant
K_d	Dissociation constant
L(i,j)	Shortest path between nodes i and j
M(i,j)	Minimum No. of edges that need to be removed to disconnect nodes i and j.

n	Number of nodes
Na₂HPO₄	Sodium hydrogen phosphate
NaH₂PO₄	Sodium dihydrogen phosphate
NaN	Not a number
n_Y	Number of Tyrosin residues
n_W	Number of Tryptophan residues
n_C	Number of Cystein residues
pCp	[2.2]paracyclophane
r	Assortativity coefficient
T_i	Number of connected tripels on node i

Greek Symbols

ε₂₈₀	Molar extinction coefficient [Lmol ⁻¹ cm ⁻¹]
δ	Clustering coefficient
κ	Nodal connectivity
λ	Largest eigenvalue of A
ρ	Graph density
σ(u,v)	Shortest path between nodes u and v
σ(u,v i)	Number of shortest paths that pass through node i

List of Contents

Dedication	I
Acknowledgements	II
Abstract.....	IV
Kurzzusammenfassung.....	VI
List of Abbreviations	IX
List of Symbols	XII
1. Introduction	1
2. Background.....	4
2.1. Protein-protein interactions and their importance in biological processes	4
2.2. Immune system	6
2.3. Introduction of superantigens (Protein A and Protein G)	10
2.3.1. Diagnostic potential of superantigens.....	13
2.3.2. Therapeutic potential of superantigens.....	13
2.4. Species-specific differences of IgGs.....	14
2.4.1. Bovine IgG	15
2.4.2. Goat IgG	16
2.4.3. Human IgG.....	16
2.4.4. Rabbit IgG	17
2.5. IgG:Protein A/G interactions	17
2.6. Existing methods for protein-protein interaction analysis – their limitations.....	20
2.6.1. Yeast two-hybrid system	20
2.6.2. Co-immunoprecipitation.....	21
2.6.3. Fluorescence resonance energy transfer	21
2.6.4. Mass spectrometry.....	22
2.6.5. Tandem affinity purification.....	22
2.6.6. Affinity chromatography	22
2.6.7. Protein arrays.....	23
2.6.8. Fragment complementation	24
2.6.9. Phage display.....	24
2.6.10. X-ray crystallography	25
2.6.11. Nuclear magnetic resonance spectroscopy	25
2.6.12. Circular dichroism spectroscopy	26
2.7. Drying droplets	28

2.8. Chemical vapor deposition polymerization	30
2.9. Artificial intelligence in biological science and biomolecular interactions	32
2.10. Convolutional neural network and image classification	37
2.10.1. Image classification with convolutional networks	39
2.10.2. The CNN framework.....	41
2.10.3. Hyper-parameter and parameter tuning.....	42
2.10.4. t-distributed stochastic neighbor embedding (t-SNE)	44
2.10.5. Gradient-weighted class activation mapping (Grad-CAM).....	45
2.11. Graph theory introduction.....	46
3. Materials and Methods.....	50
3.1. Chemicals.....	50
3.2. Instrumentation	51
3.3. Software	51
3.4. Buffer system and protein samples preparation.....	52
3.5. Exact protein's concentration measurement	53
3.6. CD spectroscopy measurement.....	54
3.7. High – Performance Protein A SpinTrap column.....	54
3.8. Substrate preparation	56
3.9. Chemical Vapor Deposition (CVD) polymerization	56
3.10. Droplet dispensing	58
3.11. Polarized light microscopy imaging	59
3.12. ToF-SIMS	60
3.13. SEM imaging	60
3.14. Convolutional Neural Network.....	61
3.15. Grad-Cam image analysis.....	63
3.16. t-SNE clustering plot.....	63
3.17. Graph theory analysis	64
3.18. Neural network design	66
4. Results and Discussion.....	68
4.1. Traditional protein-protein interaction analysis.....	69
4.1.1. High-Performance Protein A SpinTrap column.....	69
4.1.2. CD spectroscopy analysis	71
4.2. Classification of interaction strength for IgG:Protein A complexes using neural networks....	74
4.2.1. Image classification using a pretrained CNN.....	78
4.2.1.1. Classification of single proteins	78
4.2.1.2. Classification on transformed test dataset.....	79
4.2.1.3. Classification of IgG:Protein A complexes with various molar ratios.....	82

4.2.1.4. Image classification robustness.....	87
4.2.2. Feature classification based on graph theory analysis	90
4.2.2.1. Neural network design	94
4.2.2.2. Classification of IgG:Protein A complexes.....	96
4.2.3. Neural network optimization.....	100
4.2.3.1. Input dataset size reduction.....	100
4.2.3.2. Feature selection algorithm	103
4.2.3.3. Feature classification robustness.....	107
4.2.4. Image classification vs. feature classification	108
5. Conclusion and Outlook.....	111
List of Tables	114
List of Figures.....	115
Appendix A.....	119
Appendix B.....	121
References.....	127

1. Introduction

In biological systems, proteins play a vital role due to their varied functions and modes of interactions. Understanding the complexities of protein interactions is essential, because even subtle changes can have a profound effect on their biological function and stability.^[1] An arsenal of methods for studying protein-protein interactions exists, such as tandem affinity purification, affinity chromatography, co-immunoprecipitation, protein arrays, fragment complementation, phage display, X-ray crystallography, and NMR spectroscopy.^[2–4] However, most protein interactions are transient, characterized by small contact zones and moderate conformational changes.^[5] Thus, it becomes more challenging to identify weak and non-specific interactions between proteins.

A crucial element of biological research is the humoral immune response, which is conserved among most species and involves the production of polyclonal antibodies. These antibodies are known for their structural and functional variability, which makes them highly useful for both research and diagnostic purposes.^[6–8] Immunoglobulin G (IgG) antibodies play a vital role in the immune system, as they are responsible for recognizing and attaching to specific target proteins or antigens, a process that is critical for triggering immune responses and ensuring overall health.^[9] The primary antibodies present in serum are IgG proteins, which are categorized into four subclasses: IgG1 (66%), IgG2 (23%), IgG3 (7%), and IgG4 (4%).^[10] Additionally, the species origin of IgG molecules—whether human, rabbit, bovine, or goat—contributes to their unique properties, leading to differences in antigen-binding affinities.^[11–13]

An important class of B-cell superantigens (SAGs) interacts with IgG molecules in an unusual way by binding to the antigen-binding fragment (Fab) region of the IgGs outside the complementarity-determining region.^[12,14] As a result of its ability to sequester antibodies via

their crystallizable fragment (Fc) domain, SAGs have been demonstrated to inhibit opsonophagocytosis by orienting the antibodies at an incorrect angle.^[15] SAGs have proven to be useful in exploring fundamental questions in immunobiology, such as mechanisms for cell activation, tolerance, and autoimmunity.

In recent years, the incorporation of Machine Learning (ML) and Artificial Intelligence (AI) into biological research has paved the way for novel approaches to predicting and analyzing protein-protein interactions. These advanced technologies enable the development of predictive and correlative models, which have the potential to transform contemporary biology.^[16,17] Our previous study investigated the development of simple and accurate method for predicting single amino acid mismatches in proteins using deep learning approaches.^[18] We demonstrated that important information about primary and secondary peptide structures can be deduced from the stains left by drying droplets.^[18] Deep-learning neural networks were presented based on polarized light microscopy images obtained from the drying droplet deposits of a variety of amyloid-beta peptides to assess complex stain patterns. A sessile droplet of water or another volatile solvent, which contains non-volatile solutes or colloidal particles, normally leaves a non-uniform stain or deposit on a substrate when it dries. This stain pattern is highly characteristic of processes involving heat, momentum, and mass transport within a droplet during evaporation.^[19–22] Several physical processes influence the pattern of deposited aqueous droplets on a characterized surface, referred to as coffee rings, including superhydrophobicity, contact-line motion, Marangoni flow, surface-tension-driven flows, thermal-hydrodynamic instability, or liquid spreading. Following the drying of complex fluid droplets, one important aspect of droplet spreading and evaporation is the resulting patterns.^[23–26]

In this study, we build from our previous work with ML-based image classification^[18] for species-specific typification of immunoglobulin complexes using stain patterns. To analyze the images, a trained neural network is used to stratify polyclonal immunoglobulin G from

various species (human, rabbit, goat, and bovine) based on their interactions with a common binding partner, Protein A. Recombinant Protein A is a B-cell superantigen with high binding affinity to human IgG and is widely used in various biological fields for the isolation and purification of human IgG.^[27] For this reason, IgG was chosen to screen different protein-protein interaction levels obtained with different IgG sources and different molar ratios ^[28] in order to predict the binding strengths using the classified different protein-protein interaction levels.

Furthermore, to address the need for optimization in both the training time and computational costs associated with the employed neural network, an innovative approach utilizing graph theory analysis was implemented.^[29–31] This concept originated from meticulous observations of the structural crystal patterns present in various protein samples. Graph theory is widely used in various scientific and engineering disciplines due to its powerful ability to model complex relationships and interactions. This novel approach was then rigorously compared to the conventional results obtained from convolutional neural networks (CNNs), highlighting the potential advantages and effectiveness of using graph-based feature extraction for neural network training and classification tasks.

StructuralGT is a Python package designed for processing images through the lens of graph theory analysis. This tool developed at the University of Michigan ^[32] provides a robust framework for converting images into graph representations, enabling the extraction and utilization of structural features for various analytical purposes. This highlighted the effectiveness of using graph theory for image analysis, particularly in biological and protein interaction studies.

2. Background

2.1. Protein-protein interactions and their importance in biological processes

Proteins, second only to water in abundance, are pivotal molecules in biology. They encompass an expansive spectrum of around 100,000 diverse types, each governing or modulating nearly every vital chemical reaction essential for our existence. Their individuality stems from the unique arrangement of amino acids in a typically 300-unit polymeric sequence. Post-synthesis, proteins undergo a crucial transformation, adopting intricately folded, compact structures imperative for their functionality. Despite the complexity of many protein structures, folding is generally remarkably efficient, demonstrating the effectiveness of evolutionary mechanisms. However, under specific conditions, natural proteins may exhibit behaviors akin to unprocessed polymers, underscoring the delicate balance between biological function and environmental factors.^[33]

Proteins serve as essential components in nearly all biological and biotechnological processes, exhibiting varying degrees of affinity and specificity. Their functionalities are governed by intricate regulatory networks of transient protein-protein interactions (PPIs). Omics technologies, including genomics, transcriptomics, and proteomics, have significantly contributed to our understanding of PPIs in biological systems, revealing their pivotal role in orchestrating the assembly of multi-protein complexes essential for various biological functions. Alterations in PPIs are implicated in numerous diseases, such as cancer, inflammation, autoimmune disorders, diabetes, osteoporosis, neurodegenerative conditions, and viral/bacterial infections, thus emerging as a promising frontier in drug development. Biologics, such as monoclonal antibodies or recombinant versions of ligand proteins or soluble

receptor regions, constitute a key therapeutic approach targeting these interactions, including well-established examples like hormone-receptor or protease-inhibitor complexes.^[34–40]

The term "interaction," beyond direct physical binding, encompasses various indirect associations like complex co-membership, regulatory links, and genetic interactions, collectively referred to as "functional associations". However, in literature and databases, "interaction" and "functional association" are often used interchangeably. The forces governing protein folding (e.g., hydrophobicity, hydrogen bonding, electrostatic interactions, van der Waals forces) coincide with those governing protein-protein interactions. Predicting and studying PPIs entail employing a diverse array of techniques developed over decades, encompassing in vitro and in vivo assays. Despite notable progress, these approaches face limitations such as false-positives/negatives, challenges in obtaining protein crystal structures, and detecting transient PPIs. Moreover, parameters influencing interactions include protein concentration, ionic strength, redox potential, pH value, dissociation constants, and oligomeric states. Experimental methods used to discern interactions also influence their detectability. To circumvent these limitations, novel approaches have gained prominence, aiding in the investigation of PPIs.^[34,38,41–47]

A significant challenge hindering unbiased evaluation and objective comparison of different methods is the uncertainty surrounding the definition of the negative class. Assigning proteins to the "non-interacting" category is inherently tentative due to the incomplete knowledge of all potential interactions and their partners.^[41,48]

The ongoing reduction in the cost of high-throughput experiments, coupled with advancements in novel prediction methods, has led to the generation of extensive datasets of protein-protein interactions. This capability to furnish relatively comprehensive and dependable collections of PPIs has spurred the creation of numerous databases, each with distinct objectives and strengths, aiming to aggregate and consolidate the available data.^[45]

2.2. Immune system

In 1890, von Behring and Kitasato documented the presence of a substance in the blood capable of neutralizing diphtheria toxin.^[49] The subsequent year saw the introduction of the term "Antikörper," or antibodies, in research illustrating the agent's capacity to differentiate between two immune substances. Consequently, the substance triggering antibody production was termed "Antisomatogen," equal to "Immunkörperbildner," or the agent that induces antibodies. The term "antigen" originates from this phrase. Therefore, an antibody and its antigen form a classic tautological relationship.^[50] In 1939, Tiselius and Kabat employed electrophoresis to fractionate immunized serum, segregating it into albumin, alpha-globulin, beta-globulin, and gamma-globulin components.^[51] Exposure of the serum to the antigen led to the depletion of the gamma-globulin fraction. This process resulted in the terms "gamma globulin," "immunoglobulin" (Ig), and IgG.

Immunoglobulins, or antibodies, represent crucial elements of the animal immune system, present in both serum and tissue fluids. They play a central role in the adaptive, or humoral, immune response. Naive B cells express a diverse array of antibodies, each in the form of a B cell receptor (BCR). Upon encountering antigens, B cells with receptors specific to the antigen proliferate, a process known as clonal selection. Some antibodies remain bound to the surface of B cells, serving as receptors for specific antigens, while others circulate freely in the blood or lymph, acting as effectors of humoral immunity. Therefore, immunoglobulins serve dual roles: firstly, as cell-surface receptors for antigens, enabling cell signaling and activation, and secondly, as soluble effector molecules capable of individually binding to and neutralizing antigens from a distance.^[50]

Subsequently, "sizing" columns were utilized to further categorize immunoglobulins based on their molecular weights into "heavy" (IgM), "regular" (IgA, IgE, IgD, IgG), and

"light" (light chain dimers) classes. Among the five different classes of antibodies, IgGs stand out as the most abundant.^[50,52]

Structurally, IgGs are glycoproteins characterized by a common Y-shaped structure. Each antibody consists of two identical light chains (L) and two identical heavy chains (H), linked by disulfide covalent bonds and non-covalent interactions. These chains comprise constant (CL and CH) and variable domains (VL and VH), forming two antigen-binding fragments (Fab) and a constant region responsible for the antibody's effector functions and biodistribution (Fc), connected via a flexible hinge region (**Figure 2-1**). Although the overall mass of the protein is 150 kDa, the shape of IgG can change to a T-shaped form in solution, providing more flexibility to the hinge region, allowing the angle between the two Fab regions to extend up to 180°. The interaction between antigens and antibodies is facilitated by electrostatic and hydrophobic forces. The Fc region of the antibody mediates many effector functions through its affinity for Fcγ receptors, while antigen binding primarily occurs at the tips of the Fab regions. The Fab consists of variable and constant domains of a heavy chain, forming a disulfide-stabilized heterodimer with the variable and constant domain of a light chain. The variability in antibody repertoires is mainly found in the N-terminal variable domains, particularly in hypervariable loops known as complementarity determining regions (CDRs). These CDRs, contributed by both heavy and light chains, define the binding interface of the antibody, with CH3 and CL playing a dominant role in binding interactions and antigenic determination (**Figure 2-1**).^[7,53–57]

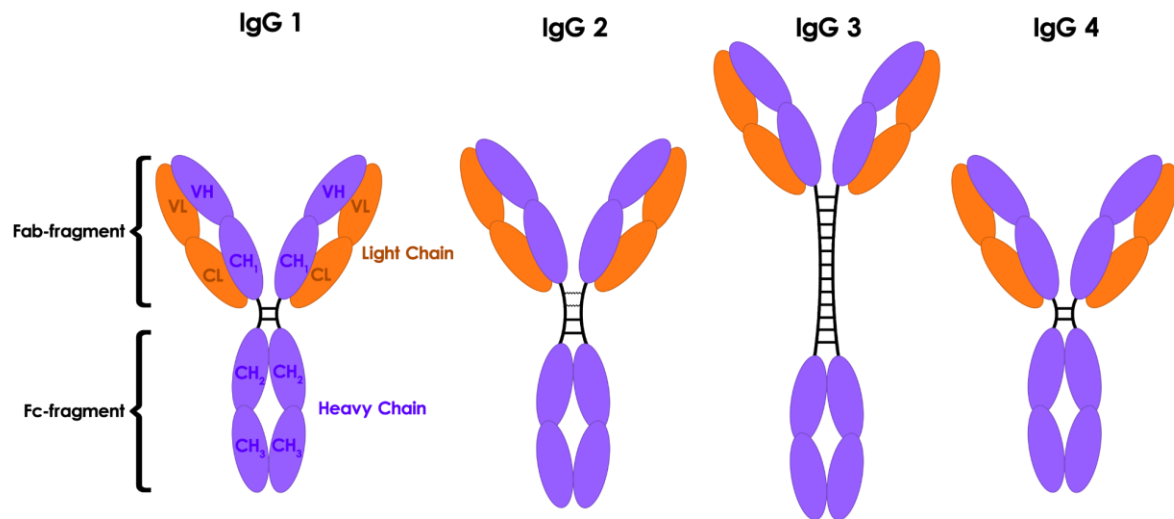


Figure 2-1: Linear model of structural variation of the different four subtypes of human IgG. An IgG molecule consists of two heavy and two light chains connected by disulfide bonds. The light chain has variable (VL) and constant (CL) regions, while the heavy chain includes one variable (VH) and three constant (CH1, CH2, CH3) regions. The Fc region mediates immune system functions, while the Fab region, containing the VL and VH domains, is responsible for antigen recognition and binding. Human IgG has four subclasses—IgG1, IgG2, IgG3, and IgG4—which differ in the size of their hinge regions and the number and arrangement of the interchain disulfide bonds linking their heavy chains.

In human sera, IgG1 stands out as the most prevalent antibody, succeeded by IgG2, IgG3, and IgG4, respectively. Despite being over 90% identical in terms of amino acid sequence, each IgG subclass possesses distinct characteristics concerning structure, antigen binding, immune complex formation, complement activation, Fc gamma receptor (FcγR) triggering, half-life, and placental transport (see **Figure 2-1**). IgG1, IgG3, and, to some extent, IgG4 typically develop against protein antigens, whereas IgG2 predominantly targets repetitive T cell-independent polysaccharide structures found on encapsulated bacteria. IgG3 often initiates the immune response, followed by the dominance of IgG1 responses later on.^[9]

The activation of the humoral immune system commences when antibodies identify an antigen, initiating effector functions via interactions with Fc engaging molecules. The strength of this interaction between the IgG-Fc domain and these Fc engaging molecules, and consequently the potential efficacy of their effector functions, is influenced by various factors such as the IgG subclass, allotype, and glycosylation pattern.^[9]

In biolayer interferometry experiments, the optimized design demonstrates binding to IgG with a dissociation constant (K_d) of approximately 4 nM at pH 8.2. However, at pH 5.5, the binding affinity weakens significantly, being approximately 500-fold less potent. The protein exhibits remarkable stability and heat resistance, alongside high expression levels in bacteria. Its pH-dependent binding behavior facilitates precise control over IgG affinity purification and diagnostic devices.^[58]

Antibodies with exceptional specificity and affinity find extensive applications in diagnostics and therapeutics. Among the most prevalent forms utilized are polyclonal antibodies and monoclonal antibodies, alongside various antibody fragments. These fragments include single-chain variable fragments (scFv), fragment antigen-binding (Fab) fragments, and single-chain antibodies (scAb).^[57]

The term "polyclonal antibodies" (pAbs), in contrast to "monoclonal antibodies" (mAbs), can indeed be ambiguous and lead to confusion. Polyclonal antibodies may be generated against various targets: the full-length protein, large protein fragments, or small peptides. Furthermore, there are distinctions between antiserum, Protein A/G-purified, and antigen affinity-purified pAbs. Generally, pAbs represent a collection of antibodies raised against multiple epitopes, potentially targeting multiple proteins. When an antibody is raised against an entire protein, it's understandable how different parts of the protein can generate a range of specificities and affinities. This diversity in pAbs can be advantageous for certain applications such as immune precipitation (IP) and Western blotting (WB), where cross-reactivity can be readily identified by differences in molecular weight (unless there's cross-reactivity with proteins of similar molecular weight). In immunohistochemistry (IHC), pAbs can also prove useful, provided there's no cross-reaction with other proteins present in the tissue sections of interest. Antibody specificity can be evaluated through several methods, such as comparing endogenous protein expression levels to those in knock-down experiments,

contrasting uninduced cells with cells exhibiting elevated expression after induction, or analyzing tissues where the protein of interest is anticipated to be localized within a specific compartment or cell type.^[59] Indeed, antibodies raised against a protein fragment exhibit heightened specificity when the amino acid sequence of the fragment is unique in the proteome. This uniqueness combines the advantages of pAbs with the distinctiveness of the antigen. While such antibodies may not match the monospecific characteristics of mAbs, they can complement them in various assays. For instance, in sandwich-type enzyme-linked immunosorbent assays (ELISA) and immune precipitation (IP), one antibody can serve as the capturer while the other acts as the reporter, facilitating enhanced detection sensitivity and specificity.^[59]

2.3. Introduction of superantigens (Protein A and Protein G)

Superantigens represent unconventional antigens as they induce a response by binding outside the complementary determining regions (CDRs) of their target immune receptor macromolecules, including antibodies or B-cell receptors.^[12] The concept of proteins fitting this definition of superantigens was initially introduced in the early 1990s.^[60] B cells, or B lymphocytes, are key components of the adaptive immune system, responsible for producing antibodies that mediate humoral immunity. After maturation, B cells express B cell receptors (BCRs) specific to antigens, enabling them to recognize and bind pathogens. Upon activation, B cells can differentiate into plasma cells that secrete antibodies like IgG, IgA, and IgM, or into memory B cells for long-term immune protection. In addition, B-cells adapt their antibody responses to different pathogens, and IgG, the most abundant antibody isotype, is crucial for neutralizing toxins and pathogens while engaging Fc receptors and complement pathways for enhanced pathogen clearance.^[61,62] Superantigens can bind to the Fab fragment of IgG, leading

to polyclonal activation and eventual depletion of B-cells rather than directly triggering apoptosis via B-cell receptor (BCR) hyperactivation. Although all B cells express BCRs on their surfaces, the activation caused by B-cell superantigens can lead to a broad and uncontrolled immune response. Additionally, B-cell superantigens are often utilized for their ability to bind the Fc region of antibodies, making them useful as affinity resins for antibody purification, but their superantigenic activity primarily involves dysregulating B-cell function through Fab binding. ^[12]

Staphylococcal Protein A (SpA), Streptococcal Protein G (SpG), and Peptostreptococcal Protein L (PpL) represent B-cell superantigens situated on the bacterial cell wall. SpA was characterized as a superantigen in 1995 due to its observed impact on B-cells. ^[63] However, SpA was initially isolated in 1940 and identified in 1964 due to its ability to bind to the Fc region. ^[64,65] It is a 42 kDa protein organized into five homologous domains (E-D-A-B-C), each adopting a three α -helix bundle fold. ^[66,67] These domains are interconnected by conserved, flexible linkers (**Figure 2-2**). Native SpA also contains region X, a 12x 8-residue repeat sequence responsible for binding to peptidoglycan. All five domains (A-E) can bind both Fc and Fab fragments. ^[68] The binding affinity for specific immunoglobulins varies depending on the isotype and species origin. In humans, SpA exhibits strong binding to IgG1, IgG2, IgG4, and weak binding to IgA1, IgA2, and IgM. ^[12,69]

Notably, SpA binds to the Fab region without interfering with the antibody's antigen-binding site, acting as a B-cell superantigen that induces B-cell proliferation and depletion while inhibiting complement fixation through the classical pathway when bound to IgG. ^[70]

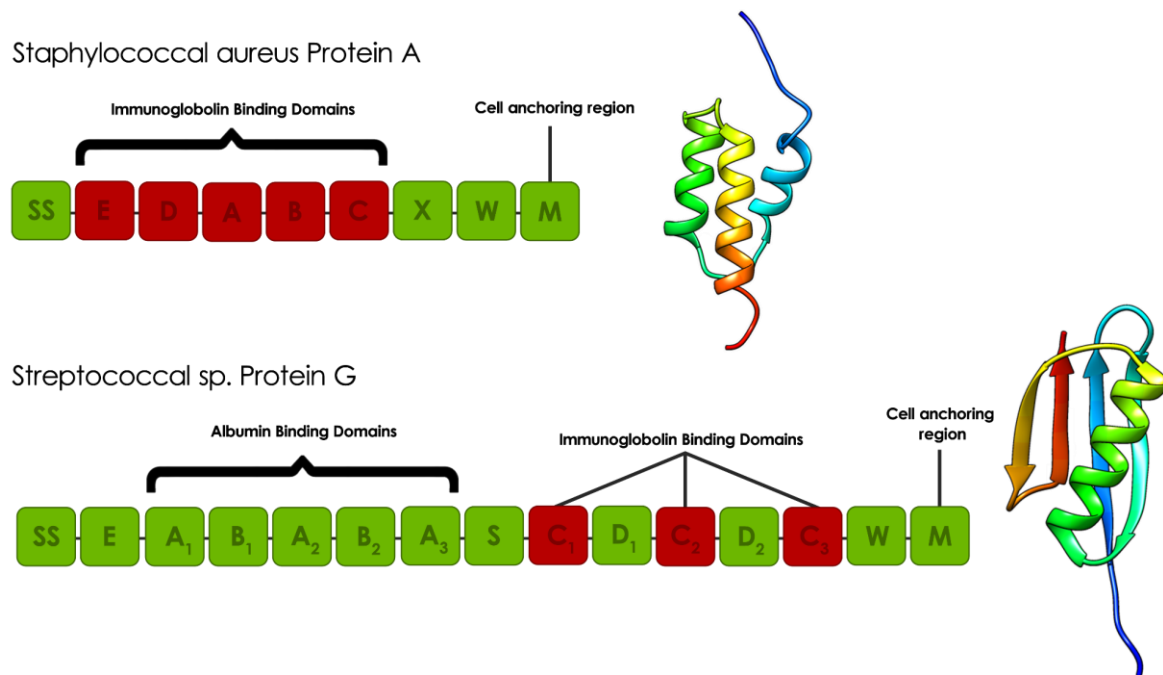


Figure 2-2: Schematic diagrams of SpA, and SpG domain structures. A) Left panel: The structure of individual SpA domains, which include the S (sorting peptide), Domains E-D-A-B-C, and Regions X and M. Right panel: The SpA Domain C (PDB code: 1BDD) shows that each immunoglobulin-binding domain in SpA is made up of three α -helices. (B) Left panel: The structure of individual SpG domains, which consist of the S (sorting peptide), Region E, Albumin Binding Domains A1-A2-A3, immunoglobulin-binding domains B1-B2/C1-C2-C3, and Region W. Right panel: The SpG Domain B1 (PDB code: 3GB1) reveals that each immunoglobulin-binding domain in SpG comprises one α -helix and four anti-parallel β -strands.^[12]

SpG was initially identified in 1984 by Björck and Kronvall ^[71] and subsequently characterized as a B-cell superantigen. The sequence of SpG varies depending on the Streptococcus strain of origin. SpG from group C Streptococcus sp. contains two immunoglobulin-binding domains (B1-B2), whereas group G has three (C1-C2-C3) (**Figure 2-2**).^[72–74] Between each immunoglobulin-binding domain are 'spacers', known as D domains. All SpG immunoglobulin-binding domains can bind both the Fc and Fab fragments.^[68] SpG has emerged as an alternative to SpA in antibody manufacturing due to its capability to bind to some antibody isotypes not recognized by SpA. It can strongly bind to all four human IgG subclasses (IgG1, IgG2, IgG3, and IgG4).^[12]

Superantigens have found utility in various applications, spanning clinical and industrial domains particularly in immunotherapy, diagnostics, and biotechnological

processes.^[75–77] A comprehensive understanding of the biochemistry underlying the superantigen-antibody interfaces serves as a valuable resource for the development of innovative biotechnological and pharmaceutical applications.^[12] In the following, some areas of application in which superantigens already play a decisive role are presented.

2.3.1. Diagnostic potential of superantigens

Superantigens recognized by IgGs facilitate the detection of *Staphylococcus aureus* in disease states.^[78,79] Engineering superantigens to target specific regions of T cell receptors (TCRs) or antibody V-region families or isotypes could be utilized for developing diagnostic kits, enabling the quantification of disease-associated proteins such as IgE in allergies.^[80] Their ability to bind antibodies selectively also facilitates the development of point-of-care testing devices, particularly relevant during events like the COVID-19 pandemic.^[81,82] These superantigen-based diagnostics can be integrated with colorimetric, user-friendly devices, such as mobile spectrophotometers. With the increasing association of antibody VH families with certain diseases (e.g., VH5 in nickel allergy^[83]), superantigens capable of distinguishing antibody VH families hold significant potential in diagnostic kit development.^[12]

2.3.2. Therapeutic potential of superantigens

The involvement of superantigens in sepsis, a major cause of mortality as identified by the WHO, underscores their significance as a key factor in the development of toxic shock syndrome.^[84] Certain short peptide regions (~40 residues) from SpA and Streptococcal Pyrogenic Exotoxin A (SPEA) have been pinpointed as contributors to vasodilation, suggesting their potential application in the development of antihypertensive drugs.^[85] Superantigens show promise as targets for an anti-*Staphylococcus aureus* vaccine. Despite limited success in vaccine development, studies have shown that anti-SpA antibodies can enhance opsonophagocytic clearance of *S. aureus*.^[12,86,87]

Superantigens have exhibited potential in cancer treatment by synergizing with antibodies to recruit T-cells.^[88] Surface Energy Balance (SEB)'s capability to hyper-stimulate and proliferate Chimeric Antigen Receptor (CAR) T-cells has resulted in a more robust antitumor response when utilized in combination.^[89]

2.4. Species-specific differences of IgGs

Specific antibodies are generated by immunizing animals with the antigen of interest. The resulting antibodies, known as polyclonal, exhibit heterogeneity in their binding patterns against the antigen due to the diverse range of antibodies produced during the immunization process. On the other hand, monoclonal antibodies, which are specific for a precise epitope on the antigen, are produced by isolating antibody-secreting lymphocytes from the immunized animal, fusing them with a myeloma cell line, and isolating a clone that produces identical antibodies. Both polyclonal and monoclonal antibodies are indispensable in research, playing crucial roles in various applications as mentioned in section 2.2.^[90] Immunoglobulin G (IgG), the most abundant antibody class in mammals, is essential for immune defense by recognizing and neutralizing pathogens like bacteria and viruses. Produced by plasma cells, IgG facilitates critical immune processes, including opsonization, complement activation, and the promotion of long-term immunity.^[50,91] Bacterial IgG receptors demonstrate interaction not only with human IgG but also with immunoglobulins from various mammalian species. Investigations into Protein A reactivity have uncovered that this reactivity within a species may be limited to specific IgG subclasses.^[92] In the following, IgG from different species and their affinity to Protein A are introduced.

2.4.1. Bovine IgG

Bovine immunoglobulins have been the subject of study for as long as those of other mammals. However, detailed immunochemical analyses of bovine immunoglobulins have historically lagged behind those of their counterparts in species such as humans and mice. Several factors contribute to this discrepancy. Firstly, the absence or infrequent occurrence of paraproteins in cattle diminishes the urgency of studying bovine immunoglobulins in certain contexts, as paraproteins are often indicators of diseases in humans. Secondly, medical scientists typically associate bovine research with veterinary medicine rather than human health, resulting in comparatively less attention and resources allocated to the study of bovine immunoglobulins. Lastly, the lack of substantial support for basic molecular biological research on the immune systems of animals of veterinary importance, including cattle, has hindered the progress of detailed immunochemical analyses of bovine immunoglobulins.^[92,93]

The IgG immunoglobulins from bovine serum can be divided into two subclasses, IgG1 and IgG2.^[94] In comparison with other species, bovine IgG has a relatively low reactivity with Protein A. Additionally, there is uncertainty regarding which bovine IgG classes and subclasses are reactive. Sloan and Butler (1978) demonstrated that neither bovine IgG1 nor IgG2 precipitated with Protein A.^[95] Conversely, Goudswaard et al. (1978) reported that 98% of IgG2 and 26% of IgG1 bound to Protein A-Sepharose.^[96] Additionally, studies indicate that the Protein A reactivity with bovine immunoglobulins may vary depending on the pH value.^[97,98] It has been indicated that at pH 8.0, Protein A only binds to immunoglobulins of the IgG2 subclass.^[97]

However, studies investigating the reactivity of Protein A with bovine IgG classes and subclasses have produced conflicting findings. Some reports suggest that Protein A exclusively binds to IgG2, while others propose that a portion of antibody molecules within the IgG1 subclass can also bind. Conversely, certain studies indicate that Protein A exhibits poor

reactivity with bovine IgG2 and no detectable binding with IgG1.^[99]

2.4.2. Goat IgG

Goats have been extensively utilized for generating large volumes of specific antisera, which find applications in diagnostic assays and various immunological research endeavors. Consequently, significant attention has been devoted to quantifying these goat antibodies when combined with specific antigens.^[100] Nevertheless, similar to bovine IgG the reactivity of goat immunoglobulins with Protein A has been the subject of thorough investigation, yielding conflicting results. Protein A has been reported to exhibit poor binding to free goat IgG. Richman et al. (1982) found that only 1 out of 9 goat sera tested reacted with Protein A at neutral pH.^[101] Conversely, Goudswaard et al. (1978) and Delacroix and Vaerman (1979) reported that Protein A exclusively binds to goat immunoglobulins of the IgG subclass.^[96,102] In contrast, Duhamel et al. (1980) reported that both goat IgG1 and IgG2 bind to immobilized Protein A at pH 7.1, with each subclass selectively eluted (i.e., IgG1 at pH 6.7 and IgG2 at pH 5.8).^[100,103]

2.4.3. Human IgG

Immunoglobulin G (IgG) is the most abundant antibody class in human serum, accounting for approximately 75-80% of total serum immunoglobulins. Structurally as mentioned in section 2.2, IgG consists of two identical heavy chains and two identical light chains, forming a "Y" shape. The antigen-binding region, located at the variable domains (Fab), is responsible for recognizing specific antigens, while the constant region (Fc) interacts with immune effector cells. The Fc region of IgG has a high affinity for Protein A, a bacterial cell wall protein from *Staphylococcus aureus*, which specifically binds to the Fc region of IgG1, IgG2, and IgG4 subclasses. This affinity is widely exploited in biochemical techniques like antibody purification and immunoprecipitation. Additionally, IgG is capable of interacting with

the neonatal Fc receptor (FcRn), which extends its half-life in circulation by protecting it from degradation.^[50,104]

2.4.4. Rabbit IgG

In contrast to human IgG, normal rabbit IgG does not precipitate with Protein A; instead, it forms soluble complexes.^[105] Approximately 15% of the total IgG in normal rabbit serum has been observed to react with Protein A.^[105] Also, the non-linear Scatchard plots obtained for rabbit and guinea-pig IgG revealed the existence of low-affinity sites for SpA, with affinity constants of approximately 10^6 LM^{-1} . These values were augmented to the aforementioned levels through multiple bindings occurring between two sites on SpA and two sites on the identical IgG molecule.^[106]

2.5. IgG:Protein A/G interactions

As mentioned, SpA stands as one of the earliest identified immunoglobulin-binding molecules and has undergone extensive study over recent decades.^[70,107–109] Leveraging its affinity for immunoglobulins, SpA has found wide-ranging utility as a tool in antibody detection and purification, evolving into one of the most employed affinity purification systems.^[27] Remarkably, one molecule of Protein A can bind to two molecules of IgG.^[96,110,111]

Soluble SpA interacts with many, if not all, mammalian IgGs, often leading to the sequestration of IgGs from the serum pool by forming insoluble complexes. However, IgGs from other hosts form soluble complexes with varying stability. It has been proposed that these differences in reactivity stem from variations in the amino acid sequences of the IgG H-chains, although the exact mechanisms behind this phenomenon remain unclear.^[70]

SpA holds significance in qualitative and quantitative immunology owing to its specific

binding to the Fc portion of immunoglobulins across various mammalian species, including humans presented in **Table 2-1**.^[108,112]

Table 2-1: IgG binding of SpA

Species	Subclasses	Protein A
Human	IgG1	++
	IgG2	++
	IgG3	-
	IgG4	++
	IgA	variable
	IgD	-
	IgM	variable
Rabbit	No distinction	++
Guinea pig	IgG1	++
	IgG2	++
Bovine		+
Mouse	IgG1	+
	IgG2a	++
	IgG2b	+
	IgG3	+
	IgGM	variable
Chicken	IgY	-
strong binding ++, medium interaction +, weak or no interaction -		

SpA's robust binding to various mammalian antibodies with a corresponding number of residues from the CH2 and CH3 domains of the Fc region (**Figure 2-3**).^[13,52]

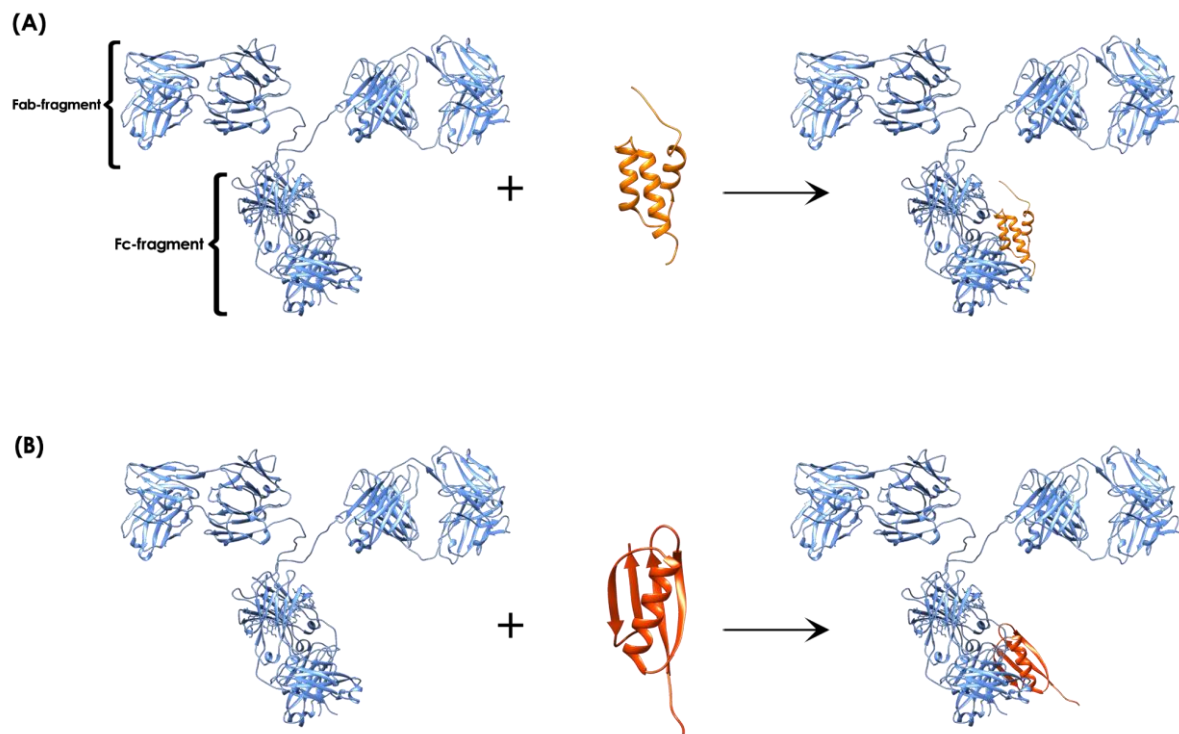


Figure 2-3: Crystal structure of IgG in interaction with A) Protein A, and B) Protein G. Blue ribbons present IgG (PDB code: 1IGT) while orange and red ribbons indicate SpA (PDB code: 1BDD) and SpG (PDB code: 3GB1) ligands, respectively. Two heavy chains in the IgG Fc fragment can be in complex with SpA and SpG.

The composition of the IgG-Protein A complex formed is contingent upon the ratio of immunoglobulins to Protein A in the reaction mixture. The most effective complement-fixing complexes exhibit a molecular composition of $[\text{IgG}]_4[\text{Protein A}]_2$ and demonstrate characteristics akin to IgM in terms of binding and activating complement.^[107]

Another ligand used for IgG purification, derived from a bacterial source, is SpG. Although Protein A and Protein G are widely employed in antibody purification, their stability as ligands poses a limitation on their utility. Protein A demonstrates greater stability than Protein G due to its more rigid and robust structure, which enhances its thermal and chemical resistance. This superior stability makes Protein A more suitable for industrial purification processes, limiting the use of Protein G in such applications. Nonetheless, Protein G remains the preferred choice for serum and human IgG subclass III purification due to Protein A's

reduced affinity for this subclass.^[27]

While Protein A is generally favored for binding with antibodies from rabbits, pigs, dogs, and cats, Protein G exhibits a higher binding capacity for a wider array of mouse and human IgG subclasses. However, it's worth noting that Protein G also contains an albumin binding site, which interacts with one of the major constituents of serum proteins. Therefore, for antibody purification purposes, a recombinant form of Protein G that lacks the albumin binding site is often preferred to avoid nonspecific binding to serum albumin, which can interfere with the purity and specificity of the antibody isolation process.^[52,113] This suggests that SpG could potentially replace SpA in immunoassays, particularly those involving IgG with weak binding to SpA, such as sheep IgG, certain mouse monoclonals, and human IgG3.^[72,114–116]

2.6. Existing methods for protein-protein interaction analysis – their limitations

Understanding protein-protein interactions is crucial for elucidating cellular functions and mechanisms. Various methodologies have been developed to detect and analyze these interactions, each offering unique strengths and facing specific challenges. In the following sections, some essential methods are presented.

2.6.1. Yeast two-hybrid system

The yeast two-hybrid (Y2H) system is a commonly employed genetic technique for detecting binary protein-protein interactions. This system functions by reassembling a functional transcription factor when the two target proteins interact within the yeast cell nucleus.^[117,118] Due to its high throughput and relatively low cost, the Y2H method is well-

suited for large-scale mapping of protein interactions. However, Y2H is prone to non-specific interactions, leading to a high rate of false positives that require further validation. Additionally, interactions are detected within the yeast nucleus, which may not accurately reflect interactions in other cellular environments or in higher eukaryotes. The system also struggles to detect interactions involving membrane-bound or insoluble proteins, limiting its applicability.

2.6.2. Co-immunoprecipitation

Co-immunoprecipitation (Co-IP) is an antibody-based technique used to identify physical interactions between proteins in their native cellular context.^[35,119] It involves the use of specific antibodies to capture protein complexes from cell lysates. Using specific antibodies ensures precise detection of protein interactions and allows these interactions to be observed under native cellular conditions, maintaining their physiological relevance. However, the success of Co-IP depends on the availability and quality of specific antibodies, which can be a significant limitation. Co-IP is labor-intensive and not suitable for high-throughput screening, and it may not efficiently capture weak or transient interactions, leading to potential underestimation of interaction networks. Non-specific interactions can also lead to false positives, necessitating rigorous controls.

2.6.3. Fluorescence resonance energy transfer

Fluorescence resonance energy transfer (FRET) is a biophysical method that detects the transfer of energy between two fluorescently labeled molecules, which signifies their proximity and interaction.^[120,121] Offering high spatial and temporal resolution, FRET enables researchers to observe interactions in real-time within living cells, making it especially suitable for studying interactions in their native cellular environments. However, the technique requires genetic fusion of fluorescent tags to the proteins of interest, which can potentially interfere with

their natural functions. FRET is limited to interactions occurring within a certain distance range (typically 1-10 nm), potentially missing interactions outside this range, and high background fluorescence can complicate data interpretation and reduce sensitivity.

2.6.4. Mass spectrometry

Mass spectrometry (MS) is a highly effective analytical technique used to identify and quantify proteins within complex mixtures, delivering in-depth insights into PPI networks.^[122,123] MS can accurately and specifically detect a broad spectrum of protein interactions, enabling a thorough analysis of both strong and weak interactions. However, mass spectrometry requires extensive sample preparation, which can result in sample loss and variability. Additionally, the equipment is costly, and the data analysis process is complex, demanding specialized expertise. Interpreting quantitative interaction data can also be difficult due to the wide dynamic range of protein expression levels.

2.6.5. Tandem affinity purification

Tandem affinity purification (TAP) is a highly effective method for isolating protein complexes under native conditions, using a dual-tag system that allows sequential purification steps to minimize contaminants.^[124,125] TAP is advantageous because it provides high specificity and yields cleaner protein complexes, reducing the rate of false positives compared to single-step purifications. The method is particularly useful for identifying stable and robust protein complexes. However, TAP is less effective at detecting transient interactions due to the stringent washing steps involved. Additionally, the fusion of tags may potentially interfere with the natural function or localization of the proteins of interest.^[126,127]

2.6.6. Affinity chromatography

Affinity chromatography is a method utilized to purify proteins by taking advantage of

specific interactions between a protein and a ligand that is attached to a chromatographic matrix. This technique provides high specificity and efficiency in isolating proteins based on their binding characteristics. It is especially effective for purifying proteins with known binding partners, including antibodies, enzymes, and receptor-ligand pairs.^[128,129] The primary advantages of affinity chromatography are its capacity to achieve high purity and yield in a single step, as well as its applicability to a diverse range of proteins. However, this method necessitates prior knowledge of the protein's binding characteristics and the availability of an appropriate ligand. Furthermore, the interaction between the protein and the ligand must be sufficiently strong to endure washing steps while still enabling efficient elution of the target protein, which can sometimes be a difficult balance to maintain.^[130]

2.6.7. Protein arrays

Protein arrays are high-throughput tools utilized for examining protein-protein interactions, protein-DNA interactions, and various biochemical activities on a broad scale. This approach entails immobilizing numerous proteins to a solid surface, like a glass slide or microarray chip, and testing them with different target molecules to observe interactions.^[131,132] The primary advantages of protein arrays include their ability to analyze thousands of interactions simultaneously, their suitability for studying post-translational modifications, and their use in diagnostics and biomarker discovery. However, protein arrays face several limitations, such as the challenge of maintaining protein functionality and stability upon immobilization, potential non-specific binding leading to false positives, and the requirement for high-quality, purified proteins. Additionally, the surface chemistry of the array and the orientation of the immobilized proteins can significantly impact the results, necessitating careful optimization of experimental conditions.^[133,134]

2.6.8. Fragment complementation

Fragment complementation is a technique used to study protein-protein interactions by splitting a reporter protein into two non-functional fragments. These fragments are fused to the proteins of interest, and interaction between the target proteins brings the fragments into proximity, allowing them to reconstitute the functional reporter protein. Common reporters used in fragment complementation assays include enzymes like β -galactosidase, fluorescent proteins, and luciferases.^[135,136] The primary advantages of fragment complementation include its ability to detect interactions in living cells, monitor dynamic interactions in real-time, and study interactions within specific cellular compartments. However, the technique has limitations such as the potential for steric hindrance or altered protein function due to the fusion of fragments, the need for proper orientation and proximity of the fragments for complementation to occur, and the possibility of non-specific reconstitution leading to false positives. Despite these challenges, fragment complementation remains a powerful tool for investigating protein interactions and functional relationships in a native cellular context.^[137,138]

2.6.9. Phage display

Phage display is a powerful technique for studying protein-protein interactions and identifying peptide ligands with high affinity for target proteins. In phage display, a library of peptide or protein fragments is genetically fused to the coat proteins of bacteriophages. The phages then display these peptide or protein fragments on their surface while maintaining the genetic information encoding them. By incubating the phage library with immobilized target proteins, specific binding interactions can be selected. After several rounds of selection and amplification, individual phage clones that bind strongly to the target protein are isolated and characterized.^[139,140] Phage display is advantageous because it allows for the screening of large

peptide libraries (up to billions of variants), enabling the identification of high-affinity binders and epitope mapping. However, phage display has limitations such as the potential for non-specific binding, the need for robust selection conditions to ensure specificity, and the requirement for careful optimization of experimental protocols to minimize false positives.^[141,142]

2.6.10. X-ray crystallography

X-ray crystallography is a powerful method used to determine the three-dimensional structure of biological macromolecules, including proteins and nucleic acids, with atomic resolution. This technique involves crystallizing a purified sample of the molecule under controlled conditions. Once crystallized, the crystals are exposed to an X-ray beam, causing the X-rays to scatter off the atoms in the crystal lattice. The resulting diffraction pattern, which is captured by a detector, provides valuable information about the arrangement and intensities of the scattered X-rays. Through complex mathematical analysis and computational methods, known as crystallographic refinement, a detailed electron density map of the molecule can be reconstructed.^[143,144] This map reveals the precise atomic coordinates of the protein or nucleic acid within the crystal, providing insights into its structure and function. X-ray crystallography is invaluable for drug discovery, protein engineering, and understanding molecular interactions. However, it requires high-quality crystals, which can be challenging to obtain for some proteins, as well as sophisticated instrumentation and expertise for data collection and analysis.^[145]

2.6.11. Nuclear magnetic resonance spectroscopy

Nuclear Magnetic Resonance (NMR) spectroscopy is a powerful technique for studying protein-protein interactions in solution at atomic resolution. In NMR spectroscopy, proteins are isotopically labeled with ^{15}N and/or ^{13}C to enhance sensitivity and spectral resolution.

Interactions between proteins can be detected by monitoring changes in chemical shifts, line widths, or intensity of NMR signals upon complex formation. NMR provides information on the structural dynamics, binding affinity, and stoichiometry of protein complexes. It is particularly valuable for studying transient or weak interactions that may not be amenable to other structural techniques.^[146] NMR spectroscopy can also reveal details about the conformational changes induced upon binding and can distinguish between different binding modes.^[147] However, NMR spectroscopy requires relatively high concentrations of purified proteins and is sensitive to sample quality and conditions such as pH value and temperature. Additionally, data interpretation can be complex and time-consuming, requiring specialized expertise in both NMR instrumentation and protein chemistry.^[148]

2.6.12. Circular dichroism spectroscopy

Circular dichroism (CD) spectroscopy is a valuable technique for studying protein-protein interactions in solution, particularly for observing changes in protein secondary structure. CD in the far-ultraviolet (UV) region (178–260 nm) arises from the absorption of amides in the protein backbone and is highly sensitive to protein conformation. This makes CD well-suited to detect conformational changes when proteins interact.^[149] Additionally, because CD is a quantitative method, changes in the spectra are directly proportional to the concentrations of protein-protein complexes formed, allowing for the estimation of binding constants. Researchers can monitor both "intrinsic" CD signals, which reflect the conformational transitions of the protein backbone and amino acid side chains, and "extrinsic" CD signals, which arise from ligands or prosthetic groups bound to the protein. Moreover, CD spectroscopy can also provide insights into the thermodynamics of protein folding during interaction, which further aids in calculating binding affinities. For example, changes in ellipticity at specific wavelengths (e.g., negative bands at 222 and 208 nm for α -helices or at

218 nm for β -structures) offer a direct readout of the secondary structure's alterations during protein-protein interactions.^[149]

The method is particularly useful for understanding the structure-function relationship of biological macromolecules in their native conformation within solution. CD spectra of folded proteins reflect their conformational features, making it ideal for monitoring structural changes caused by denaturation (via heat, chemical agents, or mutations) and during complex formation. It can also be used to explore the stability of proteins and their interactions in dynamic systems.^[150]

Despite its utility, CD spectroscopy has limitations when applied to protein-protein interaction studies. One of the main drawbacks is its lack of specificity. While CD effectively detects changes in secondary structure (e.g., α -helices and β -sheets), it provides little information on the specific binding interface or residues involved in the interaction, making it challenging to pinpoint the precise interaction sites.^[151] Additionally, when dealing with large protein complexes, the signals from different structural elements can overlap, making it difficult to interpret conformational changes accurately. This overlap can weaken the CD signal or convolute the spectra.^[152] Furthermore, CD is less sensitive to small or subtle structural changes that might occur during protein-protein interactions, reducing its ability to detect minor conformational adjustments.^[153] These limitations highlight the need to complement CD with other structural techniques for a more comprehensive understanding of protein-protein interactions.

Having reviewed various traditional methods for analyzing protein-protein interactions, exploring alternative approaches that can offer new insights or complement existing techniques is important. One such approach involves the analysis of protein interactions through the drying

of protein droplets on surfaces, capturing the patterns they deposited, and using machine learning approaches for classification. This method offers distinct advantages in examining protein behavior in a more controlled, simplified environment, providing valuable information about protein-protein interaction strengths that might otherwise be challenging to capture in solution-based methods. This transition from conventional solution-based protein-protein interaction techniques to a deep learning-driven approach using dried droplets broadens the toolkit for investigating complex protein interactions in diverse environmental contexts.

2.7. Drying droplets

The study of droplet evaporation has seen remarkable growth since the 1980s, driven by its significance across various domains such as inkjet printing, paints, polymers, nanotechnology, and medical diagnostics.^[154] Early theoretical contributions, notably Maxwell's work in 1877, initially proposed that drop evaporation was diffusion-controlled. However, it was later understood that both heat and mass transfer mechanisms play crucial roles in this process. A major breakthrough came in 1997 with Deegan et al.'s research, which explained the "coffee stain" effect—a phenomenon where droplets containing colloids leave a characteristic ring-shaped deposit due to differential evaporation rates and particle migration.

[22,155–161]

The evaporation of sessile droplets involves a complex interplay of factors including heat, momentum, and mass transfer. The dynamics of these factors, such as the behavior of the three-phase contact line where the droplet meets the substrate, significantly influence the final deposition pattern. Interactions between the deposited materials, the substrate, and the surrounding air further affect the spatial distribution and morphology of the residue. Understanding these dynamics is crucial for controlling deposition patterns, which can vary

based on solution composition, substrate condition, and evaporation rate. Sessile droplets have practical applications in surface energy measurements and other fields where understanding surface interactions is essential. The insights gained from studying droplet evaporation are valuable for optimizing processes and materials in these applications.^[162]

Chemical Vapor Deposition (CVD) technology, known for its ability to produce pinhole-free surfaces, shares similarities with droplet drying in terms of requiring precise control over deposition parameters. CVD's capability to maintain uniform coatings and consistent film composition highlights the importance of controlled conditions, which parallels the need for accuracy in droplet evaporation studies and explained in the following.

Research has identified various deposition patterns resulting from droplet evaporation, including the well-known coffee ring effect. This pattern emerges due to contact line pinning and differential evaporation, with additional influences from convection and Marangoni effects. These patterns have been extensively investigated in contexts such as protein assays and microelectronics.^[162]

In recent developments, pattern recognition techniques have been employed to analyze complex deposition patterns automatically. This advancement has expanded the scope of research into areas like biotechnology and medical diagnostics, where detailed analysis of droplet patterns can enhance disease diagnosis and forensic analysis. The application of artificial intelligence in these studies, especially in the context of protein interactions, reflects broader trends in leveraging advanced technologies to gain deeper insights into complex systems.

Overall, the study of droplet evaporation combines insights from fluid dynamics, surface science, and material science. It continues to evolve, driven by its relevance across multiple scientific and industrial fields. This interdisciplinary approach enhances our

understanding of droplet behavior and its applications, paving the way for innovations in material science and diagnostics. These multifaceted approaches underscore the diverse and interconnected phenomena at play, from fluid dynamics and thermal effects to surface interactions and material deposition patterns, which are summarized in **Figure 2-4**.^[21,22,163]

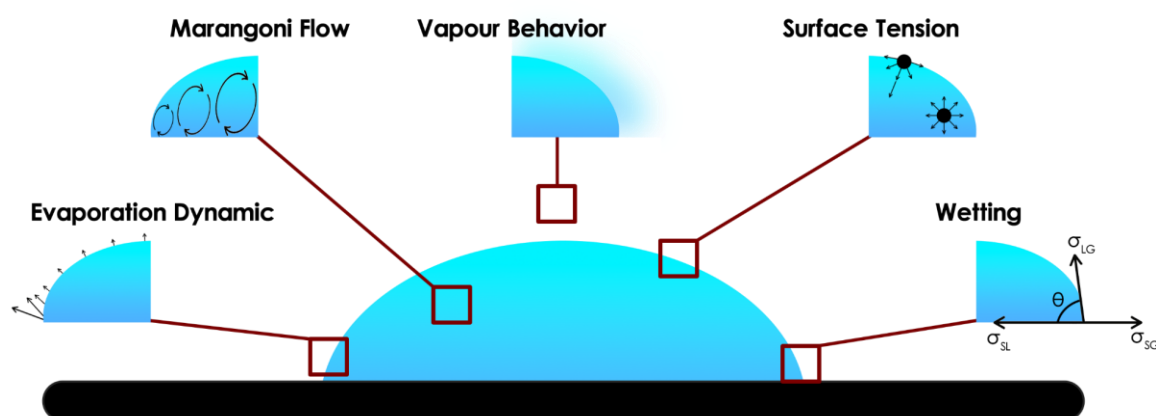


Figure 2-4: Key aspects of droplet wetting and evaporation. Physical mechanisms occur as a single-component droplet evaporates on a solid substrate. Critical factors such as marangoni flow within the droplet, evaporation dynamics at the droplet edge, vapor behavior in the surrounding air, surface tension at the liquid-air interface, and wetting phenomena on the solid substrate, characterized by contact angle and surface energy balances (σ_{LG} : liquid-gas tension, σ_{SL} : solid-liquid tension, σ_{SG} : solid-gas tension). Each of these mechanisms plays a role in defining the droplet's behavior during evaporation.

2.8. Chemical vapor deposition polymerization

Chemical Vapor Deposition (CVD) polymerization is a sophisticated technique for synthesizing organic polymer thin films on substrates using vapor-phase reactants. This method is applicable to a range of polymers, including dielectric, semiconducting, electrically conducting, and ionically conducting types. CVD is advantageous due to its solvent-free operation, which ensures high-purity films free from pinhole defects, and allows for precise control over film thickness and composition, often achieving ultrathin layers less than 10 nanometers thick ^[164–168]

One of the main benefits of CVD is its capability to produce insoluble conductive

polymers and those enriched with organic functional groups, which are crucial for producing durable and chemically resistant coatings used in electronic circuits and corrosion protection. Furthermore, CVD's excellent conformality ensures even coverage over complex and uneven surfaces, such as textiles and micro- or nanostructured devices. The technique also provides environmental and operational benefits, including fewer solvent-related problems and reduced substrate damage, thanks to its lower operating temperatures compared to conventional methods.^[164,169–171]

Historically, CVD has been employed for depositing inorganic thin films in microelectronics and optoelectronics, with early research focused on enhancing deposit uniformity and quality. Recently, the application of CVD has expanded to include organic polymer thin films, demonstrating its adaptability. Innovations such as metal–organic covalent network (MOCN) thin films have showcased CVD's versatility in producing specialized materials, like gas-separation membranes.^[164]

The primary advantage of CVD polymerization is its capacity to integrate bio-based materials into thin film fabrication, offering a more sustainable method for surface modification. Notably, the deposition process can be carried out at room temperature or even lower, which is critical for maintaining the integrity of temperature-sensitive bio-based substrates. This stands in contrast to traditional high-temperature deposition methods, which are frequently inappropriate for sensitive materials. The integration of high-temperature monomer activation followed by low-temperature deposition allows for the creation of high-quality polymer coatings without damaging the structure of bio-based materials. As a result, CVD polymerization emerges as a versatile and environmentally friendly technique suitable for various applications.^[172,173]

CVD's integration into industrial applications underscores its importance in surface engineering and device fabrication. Its ability to uniformly coat large areas and adhere to

various substrates without specific modifications improves process efficiency and reduces fabrication complexity. This capability extends to specialized applications, including biomedical devices and energy storage systems, where customized polymer properties, such as anti-fouling or icephobic surfaces, are beneficial.^[174–180]

In summary, CVD polymerization is a refined and sustainable technique for producing conformal, defect-free organic thin films with adjustable properties. Its precise coating capabilities provide a solid foundation for investigating droplet drying processes, ensuring high accuracy and minimal effects from surface defects, which is crucial for optimizing applications involving droplet behavior.

After utilizing CVD to coat surfaces with precise, uniform thin films, these treated surfaces serve as an ideal platform for studying protein droplet deposition. The controlled properties of the coated substrates allow for a more systematic investigation of the behavior of proteins during evaporation and deposition. However, analyzing the complex patterns left behind by the drying droplets requires advanced computational tools. This is where artificial intelligence, particularly deep learning, becomes invaluable. By employing AI-based approaches, we can classify and interpret the intricate patterns formed on the CVD-treated surfaces, enabling a deeper understanding of protein interactions and deposition behaviors that would be challenging to discern through conventional methods.

2.9. Artificial intelligence in biological science and biomolecular interactions

Artificial Intelligence (AI) is a branch of computer science refers to the development of systems that can perform tasks typically requiring human intelligence, such as decision-making, pattern recognition, problem-solving, and language processing. As one of the most transformative fields in technology, AI has applications across various domains, including

healthcare, finance, manufacturing, and scientific research. A major component of AI is machine learning (ML), a subset where algorithms learn from data and improve performance over time without being explicitly programmed. ML has revolutionized many areas by allowing computers to process large datasets, uncover hidden patterns, and make predictions. This can be done through supervised learning (using labeled datasets), unsupervised learning (finding patterns in unlabeled data), or reinforcement learning (training systems based on rewards and penalties). In addition to ML, AI encompasses other fields such as deep learning (DL), which leverages neural networks with multiple layers to model complex data patterns, especially in image, sound, and language recognition tasks. Deep learning has been particularly impactful in the field of pattern recognition, such as in medical imaging and protein interaction studies, where AI techniques have been used to analyze complex biological data, offering new insights that were previously unattainable with traditional methods.^[181–184] **Figure 2-5** represents the diagram of AI.

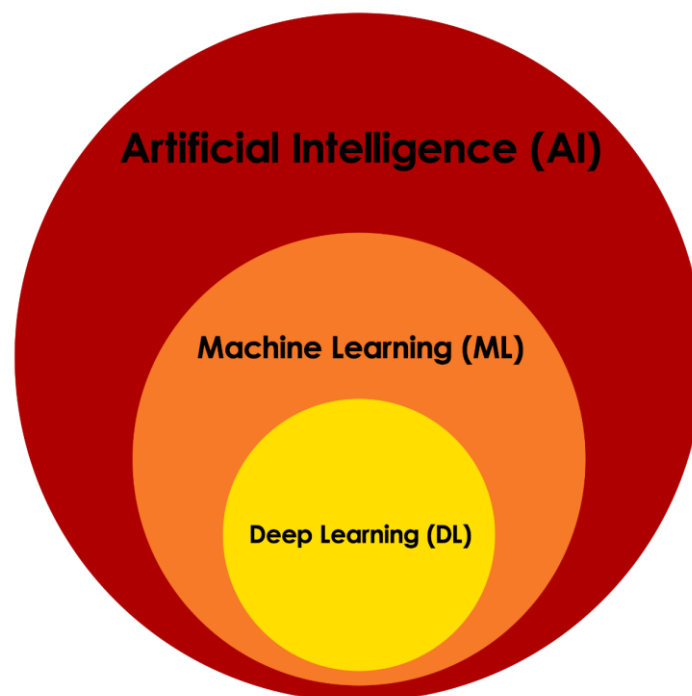


Figure 2-5: High-level AI diagram. This diagram illustrates the hierarchical relationship between Artificial Intelligence, its subset Machine Learning, and the further specialization of Deep Learning, highlighting their interconnected roles in developing intelligent systems and solving complex problems.

ML has transformed the field of bioinformatics, especially in the analysis and prediction of biomolecular interactions. By utilizing extensive datasets and advanced algorithms, researchers can elucidate complex biological processes that are often difficult to understand with conventional experimental techniques. Increasingly, machine learning algorithms are being employed to predict interactions among biomolecules, including proteins, nucleic acids, and small molecules. These biomolecular interactions are fundamental to numerous biological activities, such as signaling pathways, metabolic processes, and the development of pharmaceuticals. For example, various supervised learning methods, including Support Vector Machines (SVM), Random Forests, and neural networks, are employed to analyze and categorize protein-protein interactions. They achieve this by leveraging features obtained from both sequence and structural information. These methodologies can reliably identify binding partners and estimate interaction affinities, contributing to a better understanding of biological systems.^[185–189]

The applications of ML in biomolecular interactions are extensive. In the realm of drug discovery, ML models have been created to predict the binding affinities of small molecules to target proteins, which notably decreases the time and costs linked with conventional screening methods. For example, deep learning techniques have been utilized to interpret molecular dynamics simulations, facilitating the PPI networks. Additionally, natural language processing methods have been adapted to extract relevant information from scientific literature about biomolecular interactions, thereby enhancing the datasets used for training ML models.^[190–192]

AI is transforming biological research by leveraging large datasets and advanced algorithms to drive discoveries across a range of fields. In biological research, the volume of data generated by genomics, proteomics, and other high-throughput methods has created a need for computational tools that can process, analyze, and derive meaningful insights. AI techniques, particularly machine learning (ML) and deep learning (DL), excel at identifying

patterns within these complex datasets, which can significantly enhance understanding of biological systems and facilitate predictions.^[193]

One of the key areas where AI has had a profound impact is in biological diagnostics, where AI-powered tools are used to analyze medical images, detect anomalies, and predict disease outcomes. For example, DL-based methods can interpret microscopy images, identifying cellular and tissue-level patterns with high accuracy. A notable technique is convolutional neural networks (CNNs), which are widely used for imaging-based diagnostics, enabling automated analysis of histopathology images to detect diseases like cancer.^[194] Moreover, AI techniques such as radiomics analyze imaging data beyond visual inspection, extracting quantitative features from medical images to improve diagnostic precision and predict treatment outcomes.^[195] These advanced imaging methods are also applied beyond medical diagnostics, extending into ecological studies, where AI helps monitor biodiversity and track environmental changes through image analysis.

AI-driven diagnostic tools have transformed disease diagnosis by providing high accuracy, speed, and cost-effectiveness in detecting diseases from medical images, biological markers, and patient data. In medical imaging, CNNs have demonstrated superior performance in interpreting medical images for diagnosing conditions such as cancer, neurological disorders, and cardiovascular diseases.^[196] For biomarker discovery, AI algorithms identify potential biomarkers for various diseases by analyzing omics data, leading to early and accurate diagnosis.^[197] Predictive analytics enable AI models to predict disease progression and patient outcomes based on electronic health records and genetic data, aiding in personalized treatment plans.^[198,199]

Despite the considerable advancements AI has brought to biological science and biomolecular interactions, several challenges and areas for future development remain. Achieving reliable predictions depends heavily on the availability of high-quality datasets. A

major obstacle lies in integrating diverse biological data from multiple sources, which requires the development of standardized formats and robust data integration frameworks. Furthermore, the interpretability of machine learning models is essential for their successful application in biological research, as understanding how these models make decisions can lead to valuable biological insights. This interpretability is particularly important in clinical settings, where trust in AI-driven decisions is paramount. Additionally, addressing ethical concerns related to data privacy, informed consent, and algorithmic biases is crucial for the responsible implementation of AI in biological sciences. Lastly, fostering interdisciplinary collaboration between biologists, computer scientists, and clinicians is essential to driving AI innovations in the field.

[200–202]

As AI continues to advance, various specialized techniques have emerged to enhance its capabilities, particularly in the field of image and pattern recognition. One of the most impactful methodologies within machine learning is Convolutional Neural Networks (CNNs), which have revolutionized how we approach visual data analysis. CNNs are built to automatically and adaptively learn spatial feature hierarchies from images, making them highly efficient for tasks such as object detection, image classification, and medical imaging diagnostics. By leveraging techniques like local connectivity and weight sharing, CNNs can process large-scale visual data effectively while maintaining computational efficiency. This has made them a cornerstone of modern AI applications, particularly in areas where image and video data play a crucial role. Transitioning from the broader scope of AI, we now delve into the specifics of CNNs, exploring their architecture, functioning, and applications in various domains.

2.10. Convolutional neural network and image classification

The past few years have witnessed remarkable progress in the field of visual recognition, largely driven by the success of deep convolutional neural networks (CNNs). While CNNs have been around for a while, their effectiveness was limited by factors such as the size of training sets and network architectures. However, breakthroughs such as the work by Krizhevsky et al.^[203] in training large networks with millions of parameters on datasets like ImageNet have propelled CNNs to the forefront of visual recognition tasks.^[204]

Traditionally, CNNs have been used for classification tasks, where the output is a single class label for an entire image. However, in many applications, especially in biomedical image processing, pixel-level localization is necessary. To address this, researchers like Cirosan et al.^[205] have developed methods to predict the class label for each pixel by training networks on local regions (patches) around each pixel.^[204,205]

Supervised learning is the dominant type of machine learning, where models are trained on data that includes labels to reduce prediction errors. In this framework, the model modifies its internal weights based on the gradient of a loss function, which assesses the difference between its predictions and the actual values. Stochastic gradient descent (SGD) is a commonly utilized optimization method for training CNNs, where weight adjustments are made based on gradients obtained from small batches of the training dataset.^[184]

After training the CNN, the performance of the system is evaluated on a separate set of examples called a test set to assess its generalization capability — its ability to make accurate predictions on new, unseen data. ^[1]

CNNs are specifically designed to process data that come in the form of multiple arrays, such as images with multiple color channels or audio spectrograms. There are four key concepts

that underpin CNNs: local connections, shared weights, pooling, and the use of many layers.

- **Local Connections:** CNNs take advantage of the principle that adjacent pixels in an image or nearby elements in a sequence tend to have strong correlations. In a convolutional layer, each neuron is connected solely to a localized area of the input volume, enabling it to concentrate on extracting features from small, spatially relevant regions.
 - **Shared Weights:** In CNNs, the same set of weights (or filters) is applied across the entire input volume. This parameter sharing ensures that the network can learn to detect the same feature regardless of its location in the input. It also greatly reduces the number of parameters in the network, making it more efficient to train.
 - **Pooling:** The pooling layer serves to decrease the dimensionality of feature maps while preserving the most significant information. A widely used pooling technique is max pooling, which retains the maximum value from a local area (such as a 2×2 patch) of the feature map. This process helps achieve spatial invariance, making the model more robust to minor shifts and distortions in the input data.
 - **Many Layers:** CNNs are typically composed of several layers, including convolution, non-linearity (such as ReLU activation), and pooling layers, arranged in a stacked configuration. This layered architecture enables the network to learn progressively more complex and abstract features by creating hierarchical representations of the input data. Deeper networks can identify more sophisticated patterns within the data. The convolutional layer's function is to recognize local combinations of features from the preceding layer, while the pooling layer aggregates semantically similar features through coarse-graining of their positions. This design allows the network to recognize motifs and objects, even with variations in their appearance and positioning within the
-

input.

In the following sections, we will outline the key steps involved in image classification. By customizing CNN architectures to analyze images of protein droplets or crystallization patterns, we can gain deeper insights into protein behavior, interactions, and functions.

2.10.1. Image classification with convolutional networks

Significant progress in image classification has been achieved through the emergence of CNNs, which are now a fundamental component of modern computer vision applications. CNNs are specifically built to automatically and adaptively learn the spatial hierarchies of features within images, allowing them to effectively recognize and classify objects with high accuracy. The architecture of CNNs typically consists of multiple layers, including convolutional layers, pooling layers, and fully connected layers, which collectively facilitate the extraction of intricate patterns and features from input images. The breakthrough performance of CNNs was notably demonstrated by AlexNet, which achieved unprecedented results in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012, showcasing the potential of deep learning in image recognition tasks. Since then, several advanced architectures, such as VGGNet, GoogLeNet, and ResNet, have been developed, further enhancing classification accuracy and efficiency.^[203,206–208] These innovations have not only improved the performance of image classification systems but have also paved the way for their applications in various fields, including healthcare, autonomous vehicles, and security systems.

As image classification techniques continue to evolve, researchers have developed increasingly sophisticated architectures to enhance the efficiency and accuracy of deep learning models. Among these advancements, the Inception module stands out for its innovative approach to feature extraction (**Figure 2-6**). By allowing the network to process multiple

convolutional filter sizes in parallel, the Inception module effectively captures a diverse range of features, making it particularly adept at recognizing complex patterns within images.^[207] This ability to learn representations at various scales significantly improves the model's overall performance while maintaining computational efficiency.

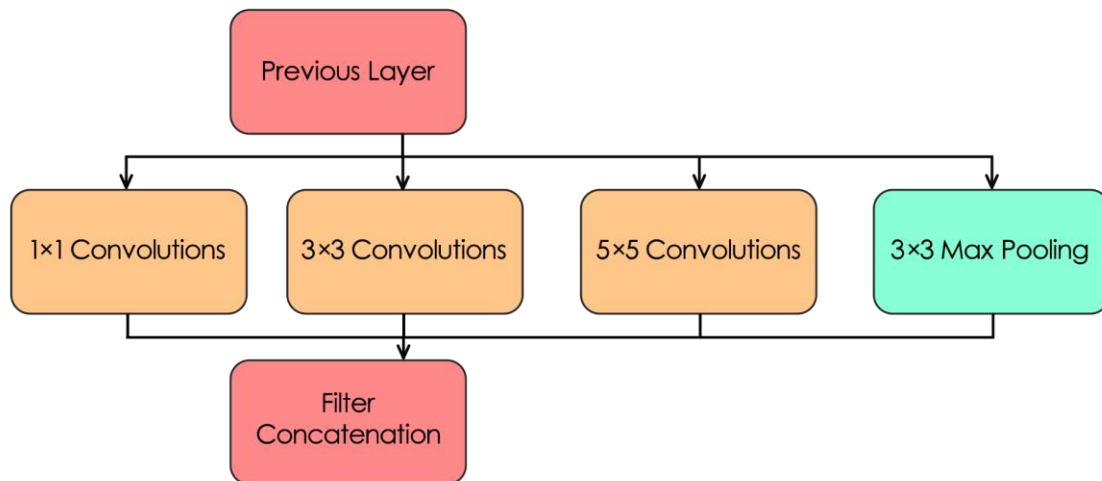


Figure 2-6: Inception module. Its parallel convolutional paths with different filter sizes, enabling the network to extract multi-scale features efficiently for improved image classification performance.

The original Inception architecture, known as GoogLeNet, introduced this concept, but it has since evolved into more sophisticated versions, including Inception V3. Building on these principles, Inception V3 further refines the architecture by incorporating factorized convolutions, batch normalization, and other optimization techniques, enabling deeper networks that deliver high accuracy with reduced resource requirements. This progression not only underscores the importance of the Inception module in advancing image classification but also highlights the potential of Inception V3 in tackling a wide array of visual recognition tasks, including those in biological and medical fields.^[207–209]

Additionally, the introduction of mini-batch stochastic gradient descent (SGD) algorithms has revolutionized the training of deep learning models. SGD, particularly with

large mini-batch sizes, enhances computational efficiency but may compromise generalization. Specialized training procedures have been developed to mitigate this issue.^[210,211]

2.10.2. The CNN framework

As previously discussed, CNNs are exceptional tools for various tasks, including image classification, object detection, and segmentation. Their design, inspired by the human visual system, features layers that are dedicated to extracting hierarchical features from the input data.^[17]

The fundamental architecture of a CNN generally comprises several essential layers that collaborate to automatically extract and learn hierarchical features from input images. Key components of a CNN include convolutional layers, pooling layers, and fully connected layers. Among these, the convolutional layer serves as the core building block of the CNN. It applies a series of convolutional filters (or kernels) to the input image, enabling the network to capture local patterns and features such as edges, textures, and shapes. Each filter slides over the image, performing an element-wise multiplication and summation, producing a feature map that highlights the presence of specific features at various spatial locations. This approach enables the network to learn spatial hierarchies of features across multiple layers. After these layers, activation functions such as ReLU or tanh are applied, introducing non-linearity that is crucial for capturing complex relationships within the data. Pooling layers are then used to decrease the spatial dimensions of the feature maps, effectively downsampling the information. This reduction helps decrease the computational load and prevents overfitting. The most common pooling operation is max pooling, which selects the maximum value from each region of the feature map, retaining the most important information while discarding less relevant details. After multiple convolutional and pooling layers, the output is usually flattened and fed into one or more fully connected layers. These layers integrate the features obtained from the earlier

layers and facilitate the final classification decision. The last layer employs an activation function, such as softmax, to generate probability scores for each class in a classification task.^[17,203,208,212] In **Figure 2-7**, the basic architecture of a CNN is shown.

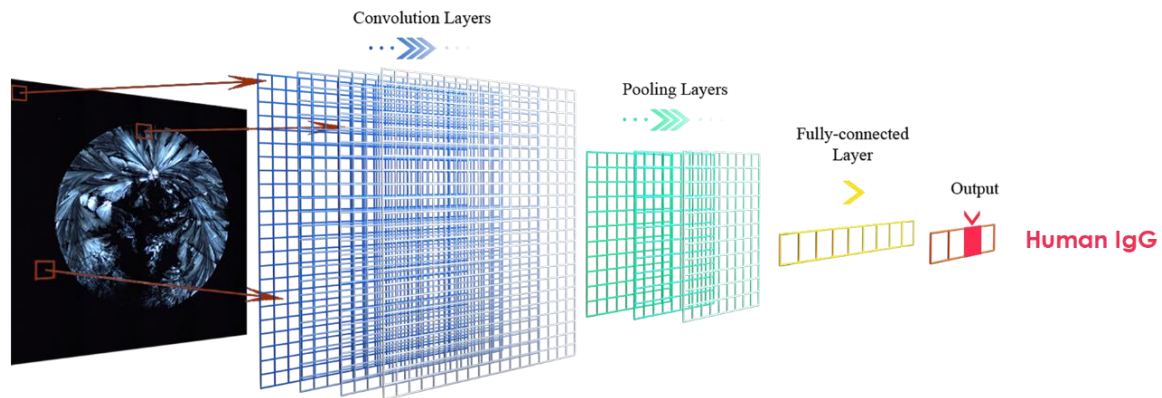


Figure 2-7: Basic architecture of CNN. Illustration of the fundamental structure of a CNN, highlighting its key components: convolutional layers for feature extraction, pooling layers for dimensionality reduction, and fully connected layers for final classification. The flow of data through these layers demonstrates how the network processes and learns hierarchical features from input images, enabling effective image recognition and classification tasks.

The combination of these layers enables CNNs to learn complex feature representations directly from the data, significantly improving performance in image classification and other visual recognition tasks.

Training a CNN involves iterative processes of forward propagation, parameter initialization, loss evaluation, and backpropagation. During forward propagation, data travels through the network to generate predictions. Parameters, such as kernel weights and biases, are randomly initialized, and a loss function assesses the prediction errors. Following this, backpropagation modifies the parameters based on gradients to minimize the loss.^[17]

2.10.3. Hyper-parameter and parameter tuning

Parameters and hyperparameters play crucial roles in determining the performance and behavior of neural network models. Parameters are internal variables of the model that are learned from the data during training, such as kernel weights in convolutional layers. They are

initialized but not set by the user and directly affect the model's performance. On the other hand, hyperparameters are external to the model and set by the user before training. They include parameters like learning rates, number of iterations, and network architecture-related choices such as the number of layers. Tuning hyperparameters involves adjusting them to optimize the model's performance on a specific task. When tuning a neural network, weights from previously trained models can be transferred to a new network, except for the last fully connected layer, which often requires retraining. Determining the appropriate number of parameters and hyperparameters depends on the specific application. The number of parameters is closely tied to the complexity of the neural network architecture and has a significant effect on its accuracy. An excessive number of parameters can result in overfitting, where the model becomes adept at memorizing the training data instead of generalizing to new, unseen examples. Thus, finding a balance between the number of parameters and hyperparameters is essential for creating neural network models that can effectively generalize across various datasets.^[17]

However, challenges persist. CNNs often require large labeled datasets for training, posing hurdles in data-scarce domains like medical imaging. Additionally, training deep CNNs demands substantial computational resources and is prone to issues like overfitting and convergence.^[213–218]

Building on the understanding of hyperparameters and their critical role in optimizing deep learning models, we can explore advanced visualization techniques that enhance our interpretation of model performance and data distributions. One such technique is t-distributed Stochastic Neighbor Embedding (t-SNE), a powerful dimensionality reduction method that effectively visualizes high-dimensional data in a lower-dimensional space. By mapping high-dimensional feature representations learned by the model into two or three dimensions, t-SNE plots allow researchers to visually assess the clustering and relationships among data points,

providing valuable insights into the underlying structure of the data.

In addition to t-SNE, Gradient-weighted Class Activation Mapping (Grad-CAM) serves as another important visualization tool, particularly for understanding the decision-making processes of CNNs. Grad-CAM generates heatmaps that highlight the regions of an input image that significantly contribute to the model's predictions. By integrating t-SNE and Grad-CAM, these methods provide a robust approach to analyzing and interpreting the intricate behaviors of neural networks, ultimately enhancing model performance and offering valuable insights into the data. The subsequent sections will detail these two techniques.

2.10.4. t-distributed stochastic neighbor embedding (t-SNE)

Visualization of high-dimensional data presents a challenge across various domains, where datasets can vary widely in dimensionality. For instance, cell nuclei relevant to breast cancer may be described by around 30 variables, while pixel intensity vectors representing images or word-count vectors representing documents often have thousands of dimensions. Over the years, several techniques have been proposed to visualize such high-dimensional data, including iconographic displays like Chernoff faces, pixel-based techniques, and graph-based methods. However, these techniques typically focus on displaying more than two data dimensions and rely on human interpretation for data understanding, limiting their applicability to real-world datasets containing thousands of high-dimensional datapoints.^[219]

t-distributed Stochastic Neighbor Embedding (t-SNE) addresses this challenge with a novel approach to visualization. It differs from Stochastic Neighbor Embedding (SNE) in two key ways: first, it utilizes a symmetrized version of the SNE cost function with simpler gradients, initially introduced by Cook et al. (2007);^[220] second, it employs a Student-t distribution instead of a Gaussian to compute similarity between points in the low-dimensional space. By using a heavy-tailed distribution in the low-dimensional space, t-SNE aims to

mitigate both the crowding problem and the optimization challenges encountered in SNE. This unique approach enhances the visualization of high-dimensional data, providing insights into complex datasets with thousands of dimensions.^[219]

2.10.5. Gradient-weighted class activation mapping (Grad-CAM)

Gradient-weighted Class Activation Mapping (Grad-CAM) is a powerful visualization technique designed to improve the interpretability of convolutional neural networks (CNNs) in image classification tasks. As deep learning models have achieved remarkable success in various computer vision applications, the complexity of these models often leads to a lack of transparency in their decision-making processes. Grad-CAM addresses this challenge by providing visual explanations that highlight the regions of an image that contribute most significantly to the model's predictions.^[221]

The main innovation of Grad-CAM is its capability to produce localization maps that are specific to different classes. By leveraging the gradients of the predicted class score in relation to the feature maps from the final convolutional layer, Grad-CAM effectively captures the spatial information that impacts the network's decision-making process. This process allows for the creation of heatmaps that indicate which areas of the input image were most relevant for the prediction, effectively illustrating how the model interprets visual data.^[221] One of the standout features of Grad-CAM is its versatility; it can be applied to various architectures and tasks beyond simple image classification. For instance, Grad-CAM can also be used to visualize important regions in images used for caption generation or in visual question answering systems. The technique enhances the user's understanding of the underlying mechanics of CNNs, making it easier to identify potential flaws in model behavior or to gain insights into specific features that the model prioritizes.^[221]

Having utilized CNNs to analyze image-based data in my project, it became necessary

to explore more sophisticated optimization methods to improve model performance. While CNNs excel at feature extraction from structured data like images, optimizing their architecture and performance often requires advanced techniques. To address this, graph theory-based neural networks into the optimization process is incorporated. Graph theory offers a flexible framework for modeling complex relationships between data points, making it highly effective for optimizing neural networks by capturing non-linear dependencies and improving efficiency. This shift from CNNs to methods based on graph theory offers a more robust and scalable strategy for optimizing networks, thereby improving both the overall accuracy and interpretability of the model.

2.11. Graph theory introduction

Neural network architectures are expanding rapidly, reaching sizes with thousands to billions of parameters. Efforts are directed towards extracting high-level design insights from architectures that are automatically discovered, or implementing specific architectural designs to attain enhanced accuracy, performance, and reduced energy or memory usage. Architectural designs can be understood as directed acyclic graphs with or without labels, subject to certain restrictions.^[222]

Although the relationship between network analysis and graph theory is acknowledged, it is often not thoroughly examined. Graph theory is a branch of mathematics made up of interconnected tautologies, which creates a clear framework for understanding its history. The concept of graphs existed prior to the formalization of graph theory, initially utilized for memory aids, and later incorporated into the gradual evolution of graph theory over the past two centuries. The first application of graph theory to network analysis emerged in 1953, leading to the union of graph theory and network analysis, a notable attempt to formalize

customary social arrangements and configurations.^[223]

Graph theory methods for neural connectivity patterns rely on connection matrices derived from cortico-cortical pathways databases, individual neuron studies, and computational neuroscience models. Deep learning has become central in artificial intelligence and machine learning, showing superior performance across various domains. However, applying traditional deep learning architectures to graphs presents challenges due to their irregular structures, heterogeneity, scale, and interdisciplinary nature.^[223]

To address these challenges, significant efforts have been made, resulting in a diverse literature of related papers and methods. Graph neural networks (GNNs) have emerged as effective tools for analyzing graph-structured data, adapting deep representation learning approaches from Euclidean to non-Euclidean domains. Challenges in learning on graphs include modeling temporal graphs, incorporating edge features, and dealing with the theoretical and practical obstacles in graph model generalization.^[223]

Extending CNNs to graphs requires altering fundamental operations to operate on other geometric objects, enabling applications in diverse fields like traffic prediction and molecular engineering. The popularity of deep learning on graphs is evident from numerous recent surveys, although some fundamental theoretical and practical challenges remain unaddressed.^[224]

As the potential of graph-based deep learning continues to unfold, it is essential to explore its practical implications in real-world applications. In particular, the integration of graph theory has proven instrumental in elucidating complex biological interactions and materials design.

In biological contexts, one of the most significant applications is in the analysis of PPI networks. In these networks, proteins are represented as nodes, while interactions between

them are depicted as edges. This structure allows for the identification of crucial proteins that play significant roles in cellular processes. For instance, Barabási and Oltvai (2004) emphasized that understanding the topology of PPI networks can unveil insights into the cellular machinery, potentially guiding drug discovery efforts by highlighting essential targets for therapeutic intervention.^[225]

Another key area is the gene regulatory networks, where nodes represent genes and edges denote regulatory interactions, such as activation or inhibition. The study by Shen-Orr et al. (2002) illustrates how the topological features of these networks can reveal the regulatory mechanisms governing gene expression. This understanding is essential for clarifying developmental processes and pinpointing dysregulation in diseases like cancer. By utilizing graph-theoretic measures, researchers can investigate how disruptions within these networks influence gene interactions, ultimately enhancing the overall comprehension of genetic regulation.^[226]

In addition to PPI and gene regulatory networks, graph theory is extensively used in the modeling of metabolic networks. In these networks, metabolites serve as nodes, while enzymatic reactions are represented as directed edges. Karp et al. (2002) highlighted that using graph theory can significantly enhance the analysis of metabolic pathways, allowing scientists to identify bottlenecks and optimize metabolic efficiency. This is particularly beneficial in biotechnology, where the manipulation of metabolic networks can lead to improved yield in bio-manufacturing processes.^[227]

Graph theory also finds application in neuroscience, specifically in the analysis of brain connectivity networks. In this context, brain regions are represented as nodes, and the connections between them—whether functional or structural—are represented as edges. Sporns (2011) discusses how the application of graph-theoretic concepts can illuminate brain organization, connectivity patterns, and how these networks are altered in neurological

disorders, such as Alzheimer's disease. This approach facilitates a deeper understanding of the brain's complex architecture and its relationship to cognitive function and behavior.^[228]

In the field of materials science, graph theory is employed to describe the arrangement of atoms or molecules in materials. By representing atomic structures as graphs, researchers can derive important insights into the material's properties and behaviors. The structural analysis using graph theory aids in predicting how materials will perform under different conditions, thereby informing the design of new materials with desired characteristics.^[229]

In summary, the incorporation of graph theory into various scientific fields has significantly enriched our understanding of complex systems, facilitating the analysis of intricate relationships and interactions. The examples discussed—ranging from protein-protein interaction networks in biology to the characterization of materials at the atomic level—demonstrate the versatility and power of graph-based approaches in uncovering insights that traditional methods may overlook. The subsequent chapter will delve into the specific methodologies employed in applying these graph theory concepts to underscore the transformative potential of integrating such theoretical frameworks into a practical application.

3. Materials and Methods

3.1. Chemicals

The chemicals were used without further purification and were of analytical grade. For all experiments, Milli-Q water was used, which was purified with a MilliQ-Plus System from Merck Millipore. **Table 3-1** lists all used chemicals, materials, and lab supplies, which were additionally obtained to the ones from VWR.

Table 3-1: Used chemicals and materials

Proteins	Company
IgG from human serum $M_W = 150$ KDa	Sigma Aldrich, Taufkirchen, Germany
IgG from rabbit serum $M_W = 150$ KDa	Sigma Aldrich, Taufkirchen, Germany
IgG from goat serum $M_W = 144$ KDa	Sigma Aldrich, Taufkirchen, Germany
IgG from bovine serum $M_W = 160$ KDa	Sigma Aldrich, Taufkirchen, Germany
Human Serum Albumin (HSA) $M_W = 66.4$ KDa	Sigma Aldrich, Taufkirchen, Germany
Recombinant Protein A $M_W = 36$ KDa	Abcam, Cambridge, UK
Recombinant Protein G $M_W = 31$ KDa	Abcam, Cambridge, UK

Chemicals	Company
Na_2HPO_4 $M_W = 142$ g/mol	Merck Chemicals, Darmstadt, Germany
NaH_2PO_4 $M_W = 138$ g/mol	Merck Chemicals, Darmstadt, Germany
Glycine	Merck Chemicals, Darmstadt, Germany
HCl	Merck Chemicals, Darmstadt, Germany
Methanol	VWR, Darmstadt, Germany
Acetone	VWR, Darmstadt, Germany
[2.2] paracyclophane	Curtiss-Wright Surface Technologies, Galway, Ireland

Materials	Company
Glass plates (120 mm × 80 mm)	Optrovision, München, Germany
Microtubes	Sigma Aldrich, Taufkirchen, Germany
syringe filter 0.2 µm	Sartorius Stedim Biotech, Göttingen, Germany
Cuvette 1 mm layer thickness	Helma, Müllheim, Germany
HP Protein A SpinTrap column assay	Cytiva Europe, Freiburg, Germany

3.2. Instrumentation

In **Table 3-2** all instruments are listed, which were used to conduct this work. This includes the entire image capturing processes and also required analysis.

Table 3-2: List of used instruments

Instrument	Company
Plasma Cleaner	PIE Scientific, San Francisco, USA
SB3 tube rotator	Stuart, Stone, UK
Centrifuge 5804	Eppendorf, Hamburg, Germany
NanoDrop One	Thermo Scientific, Darmstadt, Germany
Jasco-1500 CD Spectrophotometer	JASCO, Pfungstadt, Deutschland GmbH
BX-53F	Olympus, Tokyo, Japan
EpMotion 5070	Eppendorf, Hamburg, Germany
TS10	Eppendorf, Hamburg, Germany
ICH 750	Mommert, Schwabach, Germany
ToF-SIMS	ION-TOF, Münster, Germany
LEO 1530 Gemini	Zeiss, Jena, Germany

3.3. Software

Table 3-3 lists all software, which was used for deep learning and analyzing the data.

Table 3-3: List of used software

Software	Company
MATLAB	MathWorks, Massachusetts, USA
PyCharm	JetBrains, Prague, Czech Republic
Origin	OriginLab Corporation, Massachusetts, USA
Photoshop	Adobe, California, USA

3.4. Buffer system and protein samples preparation

The sodium phosphate buffer at pH 8.1 was selected due to its prevalent use in IgG purification processes, where the interaction of human IgG with Protein A is reported to be maximal. IgG samples from various species and human serum albumin (HSA) were sourced from Sigma-Aldrich (Sigma-Aldrich Chemie GmbH, Taufkirchen, Germany). Recombinant Protein A and Protein G were obtained from Abcam (Abcam plc, Cambridge, UK). All lyophilized proteins were reconstituted in a 100 mM sodium phosphate buffer to achieve a final concentration of 0.3 mg/mL. This buffer consisted of 94.7 mM Na₂HPO₄ and 5.3 mM NaH₂PO₄ (Merck Chemicals GmbH), prepared using ultrapure water from the Mili-Q Plus system (Millipore, Schwalbach, Germany). The solution was filtered twice through a 0.2 µm syringe filter (Sartorius Stedim Biotech GmbH, Göttingen, Germany). For optimal dissolution, each IgG was dissolved in the buffer using an SB3 tube rotator (Stuart, Stone, UK) at 10 rpm for 2 hours at room temperature. Due to the presence of insoluble aggregates in the IgG solutions, centrifugation was performed (Centrifuge 5804, Eppendorf, Hamburg, Germany) at 4000 rpm for 4 minutes. The supernatant was then separated, and the exact protein concentration was determined using a Nanodrop micro-volume spectrophotometer at 280 nm (NanoDrop One, Thermo Scientific, Darmstadt, Germany), utilizing the molecular weight and extinction coefficient of each protein. Protein A, Protein G, and HSA, which dissolved readily in aqueous solutions, had their concentrations directly determined after 30 minutes of mixing using the

Nanodrop device. Following the preparation of individual protein solutions (stock samples), various protein-protein solutions were prepared with defined molar ratios, maintaining a total mass concentration of 0.3 mg/mL. These solutions were mixed using an SB3 tube rotator (Stuart, Stone, UK) at 10 rpm for one hour to maximize interaction between each antibody and antigen pair. The samples were aliquoted after mixing and stored at -20°C. Among the molar ratios tested were 0:1, 1:0, 1:2, 1:1, 2:1, 3:1, 4:1, and 5:1 (antibody:antigen), with 1:0 and 0:1 representing solutions containing only the antibody or antigen, respectively.

3.5. Exact protein's concentration measurement

To determine the concentration of the aqueous protein solutions, their absorbance at 280 nm was measured using a Nanodrop micro-volume spectrophotometer (NanoDrop One, Thermo Scientific, Darmstadt, Germany). Given the known molecular weight and extinction coefficient of each protein, the exact concentration of each stock protein sample was calculated using the following equations. In equation (1), c is the protein concentration (mol/L), A is the absorbance value, b is the path length (cm), and ϵ is the molar extinction coefficient (L/mol.cm).

$$c = \frac{A}{\epsilon b} \quad (3-1)$$

$$\epsilon_{280} = (n_W \times 5500) + (n_Y \times 1490) + (n_C \times 125) \quad (3-2)$$

The molar extinction coefficient at 280 nm can be approximated by the weighted sum of the 280 nm molar absorption coefficients of the three constituent amino acids, as described in equation (2), where ϵ_{280} is the molar extinction coefficient at 280 nm, n_W is the number of Tryptophan residues, n_Y is the number of Tyrosine residues, and n_C the number of Cysteine residues.

This approach was particularly necessary for IgG, as it forms insoluble aggregates. After centrifuging the solution and separating the supernatant, these aggregates were removed. For subsequent preparation of protein-protein samples, the precise concentration of each protein was essential. The Nanodrop device directly calculates the exact concentration of the protein solutions.

3.6. CD spectroscopy measurement

Far-UV CD spectroscopy analysis was performed at 20°C using a J-1500 Spectrophotometer (JASCO, Pfungstadt, Deutschland GmbH). Measurements were taken in quartz glass cuvettes with a path length of 1 mm (Helma GmbH & Co. KG, Müllheim, Germany) over a wavelength range of 260 to 180 nm, with data intervals of 0.5 nm. Each sample and its respective quartz glass baseline underwent two repeat scans at a scan rate of 100 nm/min. Additionally, the spectrum of a protein-free buffer was recorded. The proteins were at the same concentration as in the protein stock solutions, dissolved in 20 mM sodium phosphate buffer (pH=8.1). After smoothing the spectra using Origin 2023 (OriginLab Corporation, Massachusetts, USA) with adjacent-averaging method, a final comparison was made.

3.7. High – Performance Protein A SpinTrap column

For affinity percentage measurements, single protein solutions (IgG from various species and HSA) were prepared similarly to section 3-4, but with 20 mM sodium phosphate buffer at the same pH value. Samples were prepared at different molar concentrations ranging from 0.5 μ M to 5 μ M.

All prepared single protein solutions with these concentrations (0.5 - 5 μ M) were used in the HP Protein A SpinTrap column assay (Cytiva Europe GmbH, Freiburg, Germany) to assess the binding affinity of IgG from various sources at different concentrations. This procedure was performed according to the assay protocol. Prior to injecting the protein samples into the Protein A sepharose columns, their exact concentrations were determined using the Nanodrop device, based on their molecular weights and extinction coefficients.

A total of 500 μ L of each protein solution was injected into the Protein A Sepharose column. After injection, mixing and incubation were performed for 1 hour at room temperature. The columns were then centrifuged (Centrifuge 5430 R, Eppendorf, Hamburg, Germany) at 2000 rpm for 90 seconds and washed twice with 500 μ L of binding buffer (20 mM sodium phosphate, pH 8.1). The concentration of the washed solutions, representing the unbound protein, was determined using the Nanodrop spectrophotometer.

To elute the bound proteins from the columns, 500 μ L of glycine-HCl buffer (1 M, pH=2.9) was used. Elution was performed twice, and the concentration of the proteins in these solutions, representing the bound protein, was determined using the Nanodrop device. After each washing or elution step, centrifugation was performed at 2000 rpm for 90 seconds to collect the solution.

By applying a mass balance to the initial mass of the input proteins, the mass of interacting proteins (eluted solution), and non-interacting proteins (washed solution), the percentage of relative binding of each protein (IgG from different sources and HSA) to Protein A was calculated for different protein concentrations. **Figure 3-1** illustrates the scheme of the SpinTrap column.



Figure 3-1: The scheme of Protein A HP SpinTrap column. This column utilizes Protein A Sepharose for the purification of antibodies. The column is designed to facilitate the selective binding of immunoglobulin G (IgG) antibodies through the Protein A ligands, allowing for efficient separation from other proteins in a sample. This figure illustrates different parts of the column.

3.8. Substrate preparation

Glass plates matching the exact dimensions of 96-well plates (120 mm × 80 mm) were used which were custom-made with specifications of extra white float, clear, and uncoated with a thickness of 1.0 ± 0.05 mm (Optrovision, München, Germany). Prior to the coating process, the glass plates were cleaned using a Plasma Cleaner (PIE Scientific, San Francisco, USA) with dry air to remove surface contaminants, alter the surface energy, and enhance bonding strength. This cleaning process involved applying a power of 75 W for a 5-minute emission.

3.9. Chemical Vapor Deposition (CVD) polymerization

Following the meticulous cleaning process, the glass plates underwent a coating procedure involving the deposition of poly(*p*-xylylene) (PPX) using CVD polymerization, following a well-established method as previously detailed.^[230,231] The precursor material, [2.2]paracyclophane, sourced from Curtiss-Wright Surface Technologies (Galway, Ireland),

was subjected to sublimation under vacuum conditions. Through controlled pyrolysis, the precursor was transformed into quinomethane, which then spontaneously polymerized as it condensed onto the glass surface. This process takes place in a CVD system, which typically comprises four main components (**Figure 3-2, A**): a gas inlet for introducing vaporized precursors (such as monomers) into the deposition chamber; sublimation and pyrolysis zones where the precursor is transformed into reactive monomers; and a deposition chamber where these precursors undergo chemical activation, enabling polymerization reactions on or near the substrate surface.^[177] To facilitate the polymerization process, a constant flow of argon gas, maintained at 20 standard cubic centimeters per minute (sccm), was utilized as a sweep gas. The sublimation temperature of the precursor material was carefully regulated within the range of 100-110 °C, followed by pyrolysis conducted at a temperature of 660 °C. Throughout the entire coating process, the pressure was meticulously controlled, maintaining a stable coating pressure of 0.15 millibars.

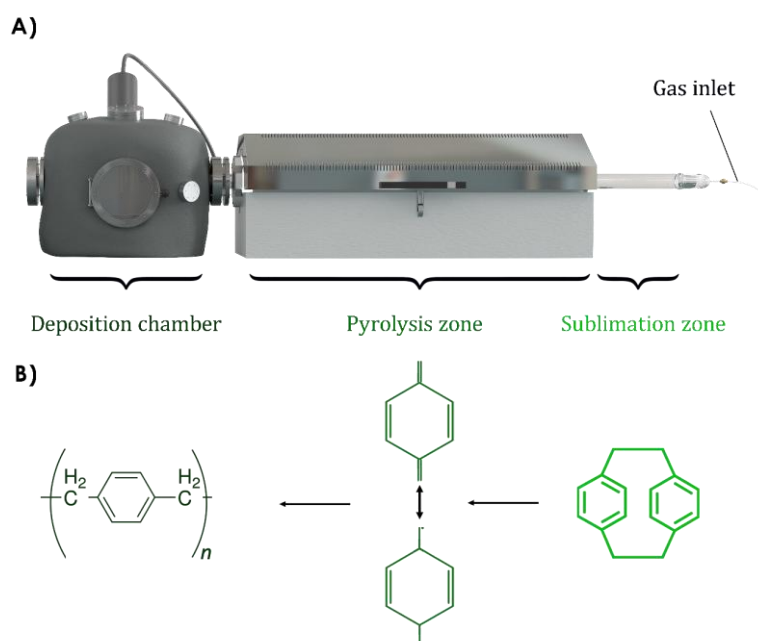


Figure 3-2: Chemical vapor deposition (CVD) polymerization. A) The CVD reactor generally comprises four primary components: the gas inlet, sublimation zone, pyrolysis zone, and deposition chamber. B) The polymerization process converting [2.2]paracyclophane into PPX under the specified conditions within a CVD system. Adapted from ^[177]

3.10. Droplet dispensing

Each protein sample solution maintained a consistent concentration of 0.3 mg/ml, meticulously prepared in a 100 mM sodium phosphate buffer with a pH of 8.1. The precise deposition of a predefined array of droplets onto the glass slide was executed using an automated 96-well microplate pipetting device, the EpMotion 5070, manufactured by Eppendorf AG, Hamburg, Germany. This device was coupled with a 1-channel dispenser (TS10, Eppendorf AG, Hamburg, Germany).

To ensure stringent control over environmental conditions during the deposition process, the pipetting system was strategically positioned within a climate chamber (ICH 750, Mommert GmbH - Co. KG, Schwabach, Germany). Environmental parameters were extensively regulated to maintain a temperature of $23^{\circ}\text{C} \pm 0.5^{\circ}\text{C}$ and a relative humidity of $40\% \pm 5\%$.

Each droplet was precisely dispensed at a controlled speed of 3 mm/s, with a uniform volume of 2 μL . The pipetting system was programmed to dispense a total of 96 droplets per glass plate, systematically arranged in the form of 12 columns and 8 rows. This ensured the creation of a well-defined array of droplets, facilitating subsequent experimental procedures with precision and efficiency.

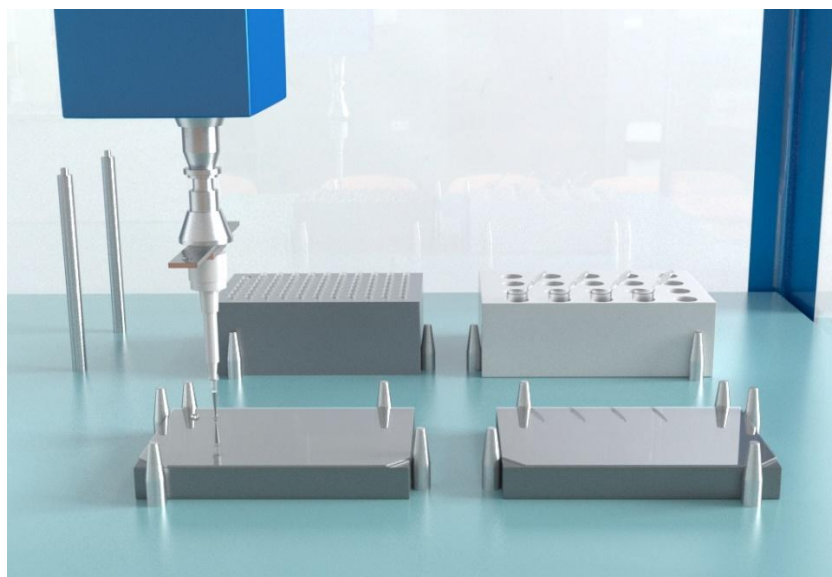


Figure 3-3: Automated droplet dispenser. It designed for precise and controlled dispensing of liquid droplets onto various substrates. The system features an array of dispensing nozzles, a programmable control interface, and a platform for substrate positioning.

3.11. Polarized light microscopy imaging

Following a drying period of at least 6 hours, the deposition patterns of the dried droplets were meticulously examined using a polarizing optical microscope (Olympus BX-53F, Tokyo, Japan). This microscope was equipped with an automated stage to facilitate precise and consistent imaging across multiple samples.

Capturing images of the deposition patterns involved employing a standardized light intensity and utilizing a 10x objective lens to ensure optimal resolution and clarity. To create comprehensive images of the entire deposition area, a stitching process was employed, utilizing the multi-image alignment (MIA) algorithm embedded within the CellSens software suite, developed by Olympus, Tokyo, Japan.

The resulting images were captured in *.jpg* format and boasted a square dimension of 8013×8013 pixels to capture more details. To optimize processing efficiency without compromising image quality, these images were subsequently resized to a final dimension of

2003×2003 pixels, enabling rapid importation into the network for training purposes. This meticulous approach ensured that the data acquired for analysis were both comprehensive and readily accessible for subsequent computational analysis and modeling.

3.12. ToF-SIMS

Time-of-flight secondary-ion mass spectrometry (ToF-SIMS) was performed using a ToF-SIMS instrument (ION-TOF GmbH, Münster, Germany) equipped with a Bi cluster liquid metal primary-ion source and a non-linear time-of-flight analyzer. For spectrometry, short primary-ion pulses (<1 ns) of the Bi source was operated in the “bunched” mode providing Bi¹⁺ ion pulses at 25 keV energy and a lateral resolution of 5 μm. As the droplets were larger than the maximum deflection range of the primary-ion gun of 500 × 500 μm², the images were obtained using the manipulator stage scan mode. Negative polarity spectra were calibrated on the C⁻, CH⁻, and CH²⁻ peaks. Spectrometry was performed in static SIMS mode by limiting the primary-ion dose to <10¹¹ ions cm⁻². Charge compensation was necessary because of the glass substrate so that an electron flood gun providing electrons of 20 eV was applied and the secondary-ion reflectron tuned accordingly.

3.13. SEM imaging

The morphology of the deposited proteins and salt within the dried droplets underwent thorough analysis using scanning electron microscopy (SEM) (LEO 1530 Gemini, Zeiss, Germany). Prior to SEM imaging, a delicate yet uniform layer of gold was meticulously sputtered onto the samples. This gold coating served the crucial function of minimizing surface charging, ensuring the accurate and high-resolution imaging of the sample surfaces with the

applying voltage of 20 kV.

This preparation process was imperative to obtain clear and precise images of the deposited proteins and salt, enabling comprehensive analysis of their morphology and spatial distribution within the dried droplets. By leveraging the capabilities of SEM in conjunction with gold sputtering, the structural characteristics of the deposited components could be elucidated with exceptional detail and clarity, facilitating deeper insights into the underlying processes governing droplet deposition and drying behavior.

3.14. Convolutional Neural Network

All raw images underwent preprocessing and training in MATLAB (Release 2023a, Math Works Inc.) to prepare them for further analysis and utilization in training a convolutional neural network (CNN). Initially, these images were resized to dimensions compatible with the input layer of the CNN and converted to grayscale mode to simplify processing and eliminate the color effect of the images. This step was done using an additional function applied to the training dataset preparation.

For training purposes, the InceptionV3 model, a pre-trained CNN renowned for its accuracy and efficiency, was employed. InceptionV3 boasts an image input size of 299×299 pixels and comprises a total of 315 layers, organized into five inception modules. Its suitability for transfer learning, particularly in scenarios with limited datasets, renders it an ideal choice.

Adopting a transfer-learning approach, the pre-trained InceptionV3 network, equipped with a rich set of image features, was fine-tuned using the relatively small set of new images acquired in this study. During transfer learning, the final classification layer was removed from the network, and retraining was conducted using the new dataset. Fine-tuning involved adjusting the parameters across all layers, utilizing a global learning rate of 0.001, a minimum

batch size of 64 images, and a maximum of 80 training epochs.

To minimize the risk of overfitting and enhance the generalization capabilities of the network, image augmentation techniques were applied. This process involved randomly performing horizontal and vertical flips on each image with a probability of 50%. At least 400 images per class were used for training, with 10% of these images randomly selected for validation during the training phase. Additionally, a distinct set of 100 images per class were reserved for testing after network training. Ensuring data integrity, there was no overlap among the training, validation, and testing datasets.

To evaluate the network's performance, both total accuracy metrics and confusion charts were generated for testing datasets. Furthermore, to assess the network's ability to generalize to unseen data, an entirely new image set was introduced to the trained network for classification, enabling comparisons between the given images and those encountered during training. This comprehensive evaluation process ensured robustness and reliability in the CNN's classification capabilities across diverse datasets and scenarios. The architecture of InceptionV3 is illustrated in **Figure 3-4**.

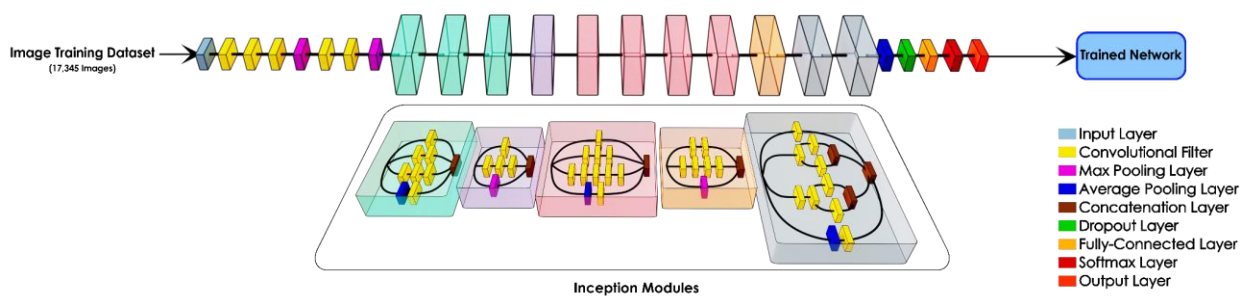


Figure 3-4: The architecture of pre-trained InceptionV3. The InceptionV3 architecture is composed of multiple Inception modules, which allow for efficient extraction of multi-scale features by applying convolutional filters of varying sizes. Key components of the network include convolutional layers, max-pooling layers, and fully connected layers, culminating in a softmax classifier for output prediction. Adapted from ^[232]

3.15. Grad-Cam image analysis

The Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm was employed as a visualization tool to elucidate the regions within the image that exert the greatest influence on the classification decision made by the convolutional neural network (CNN). By leveraging Grad-CAM, insights into the critical areas of the image that drive the network's classification outcomes were gleaned. This technique was also employed using MATLAB (Release 2023a, Math Works Inc.).

3.16. t-SNE clustering plot

The t-distributed stochastic neighbor embedding (t-SNE) algorithm, renowned for its effectiveness in visualizing high-dimensional data, was employed to analyze the "SoftMax" layer of the trained neural network. This layer is typically used as the final layer for classification tasks. It converts the raw output (logits) from the network into a probability distribution across multiple classes. This application aimed to illustrate the network's proficiency in clustering various levels of protein-protein interactions. Leveraging the capabilities of the MATLAB Machine Learning Package, t-SNE was implemented to transform the complex high-dimensional data into a lower-dimensional representation while preserving the intrinsic relationships between data points. By mapping the SoftMax layer outputs onto a two-dimensional space, the t-SNE algorithm enabled the visualization of distinct clusters corresponding to different levels of protein-protein interactions, providing valuable insights into the network's classification performance and the underlying patterns within the data.

3.17. Graph theory analysis

To optimize the convolutional neural network (CNN) within the constraints of training time and computational resources, a novel approach was adopted involving the conversion of images to graphs. This approach was motivated by the intricate, graph-like structures observed in the captured images, particularly regarding the changing morphology of salt crystals under various protein mixtures and interactions. Given the potential for enhanced feature extraction and simplified analysis offered by graph-based representations, the StructuralGT Python package, originally developed at the University of Michigan ^[32,233], was adapted for this project's specific requirements using the PyCharm platform.

The StructuralGT package utilizes graph theory principles to transform images into graphs, facilitating the extraction of significant features related to the underlying structures represented in the images. This adaptable tool offers users various input settings, enabling them to fine-tune parameters such as global threshold value, and gamma adjust to achieve the most accurate representation of the image in the resulting graph. Given the complexity of the observed patterns, five distinct input settings were adjusted across a range of gamma adjustment and global threshold values to capture the most intricate details present in the images.

Upon conversion of images to graphs, a comprehensive table of extracted features was generated (**Table 3-4**), encompassing 15 meaningful features out of a total of 20 for each input setting (the red-colored features are the non-meaningful features of this study). Consequently, a total of 75 features were extracted for each individual image. Additionally, through mathematical operations performed on these extracted features, novel features were derived and incorporated for training purposes. All these features are described in Appendix A.

Table 3-4: Features extracted from individual image using graph theory analysis. The red-colored features represent the non-meaningful features in this study.

Unweighted GT parameters		Weighted GT parameters
1	Number of nodes	Weighted Average Degree
2	Number of edges	Length-weighted Wiener Index (Inf)
3	Average degree	Max flow between periphery (NaN)
4	Network diameter (NaN)	Weighted Assortativity Coefficient
5	Graph density	Width-Weighted Average Betweenness Centrality
6	Global efficiency	Length-Weighted Average Closeness Centrality
7	Wiener Index (Inf)	Width-Weighted Average Eigenvector Centrality
8	Average clustering coefficient	
9	Average nodal connectivity (NaN)	
10	Assortativity coefficient	
11	Average betweenness centrality	
12	Average closeness centrality	
13	Average eigenvector centrality	

In the image-to-graph conversion process, the choice of input parameters significantly influenced the resulting binary images obtained from the original images. The gamma adjustment and global threshold values played pivotal roles in shaping the characteristics of these binary images. Specifically, lower gamma adjustments and higher global threshold values tended to emphasize the most prominent features of the patterns, resulting in graph representations where these dominant aspects were more pronounced. Conversely, increasing the gamma adjustment and reducing the global threshold values facilitated the inclusion of finer details from the patterns into the resulting graphs.

By systematically varying these input parameters across the five selected settings, a comprehensive range of binary images capturing different levels of detail and complexity within the patterns was generated. This approach ensured that the resulting set of features derived from the graph representations encompassed a broad spectrum of information, from

the macroscopic to the microscopic aspects of the observed patterns. Consequently, the dataset prepared from these diverse input settings provided a rich foundation for analyzing the patterns with the utmost level of detail and granularity, facilitating more comprehensive insights into the underlying structures and dynamics. **Table 3-5**, shows the selected input parameters of the StructuralGT tool.

Table 3-5: Selected input parameters for graph theory analysis of given images

Gamma adjust	0.9	1.0	1.3	1.5	2.5
Global Threshold value	140	127	100	70	80

Once all the necessary features were collected from the images, a comprehensive table of features was generated, wherein each row corresponded to an image label and each column represented an extracted feature. This table effectively served as the input dataset for feature training using the custom-designed neural network. By organizing the data in this structured format, the neural network could efficiently learn and extract meaningful patterns and relationships from the input features to make accurate predictions or classifications.

3.18. Neural network design

In the process of training a neural network using a table of numerical features for classification, the MATLAB programming software was utilized. The neural network was configured with specific parameters to ensure efficient and effective training.

The training procedure involved several key steps:

- **Neural Network Configuration:** The neural network was configured with a mini-batch size of 128 and a maximum number of epochs set to 80. The ADAM optimizer was employed as the solver algorithm for optimizing the network's weights and biases.

- **Activation Function:** The activation function used in the hidden layers of the neural network was the Exponential Linear Unit (ELU). This choice was based on empirical evidence suggesting that ELU outperformed Rectified Linear Unit (ReLU) in terms of accuracy during training.
- **Validation Set:** To assess the performance of the trained model and prevent overfitting, 10% of the input dataset was randomly selected to serve as the validation set. This validation set was disjoint from the training dataset to ensure unbiased evaluation.
- **Testing Set:** For evaluating the performance of the trained model, the same distinct converted images used in image classification were employed. This consistency ensured an accurate comparison of the model's performance across different tasks.
- **Feature Selection:** The Maximum Relevance-Minimum Redundancy (MRMR) algorithm was applied to identify the most effective features for training. Each feature was assigned an importance score, and features with scores below the selected threshold of 0.1 were removed from the dataset to reduce redundancy and noise.

Additionally, to expedite the feature extraction process and minimize computational overhead, a reduced input dataset was created by randomly selecting 10% of the initial input dataset. The training procedure was then repeated using the same architecture on this reduced dataset.

By following this systematic approach, the neural network could be trained effectively using the most relevant features, leading to improved performance and efficiency in classification tasks.

4. Results and Discussion

In this section, we present the findings from our study, which focused on classifying the strength of protein-protein interactions using deep learning approach informed by PLM imaging techniques. The results are organized to emphasize key discoveries and insights obtained from the image classification of deposited patterns of IgG:Protein A complexes, along with the analysis of their CD spectroscopy measurements. Furthermore, through the application of graph theory analysis, we successfully optimized both the computational cost and processing time required for the neural network.

The results are divided into several subsections, each focusing on distinct aspects of the research. We begin by detailing traditional protein-protein interaction methods, such as High-Performance SpinTrap Protein A column, to assess IgG binding to Protein A immobilized on sepharose, providing insight into protein binding behavior across varying concentrations. This is followed by an investigation of CD spectroscopy analysis using the same protein samples as in our proposed approach, to validate the results obtained by the CNN. Next, we explore the image classification approach using CNN and its optimization. In the subsequent section, we propose the design of a graph theory-based neural network as an alternative approach. By utilizing graph theory analysis, this approach seeks to optimize feature extraction, reduce computational complexity, and decrease the training time of CNNs. These findings underscore the relevance of this method in the context of current literature and contribute to ongoing discussions within the field.

By critically analyzing and interpreting these results, we aim to lay the groundwork for future research and practical applications, ultimately contributing to the field of protein-protein interaction studies.

4.1. Traditional protein-protein interaction analysis

4.1.1. High-Performance Protein A SpinTrap column

In this section, the results obtained from High-Performance Protein A SpinTrap columns are analyzed. The binding capacity of the used columns was more than 10 mg/ml of human IgG, which is well above the amounts applied in these experiments (ranging from 0.5 to 5 μ M). These purification columns are widely utilized for isolating specific antibodies from serum or solution samples, making the specificity and binding affinity of Protein A crucial. To investigate the effect of protein concentration on Protein A binding affinity, IgG concentrations ranging from 0.5 to 5 μ M were tested. IgG solutions were prepared, and their initial concentrations measured using absorbance at 280 nm. Following the SpinTrap protocol, unbound IgG was washed, and bound IgG was eluted. Mass balance calculations were performed on both the bound and unbound fractions, with results consistently aligning with the initial input concentrations. These results are shown in **Figure 4-1**, which represents the bound fraction of IgG to Protein A sepharose. The average binding percentage was calculated by the ratio of the mass of bound protein to the initial input protein. Regarding the polyclonal IgGs used in this study, the binding percentage to Protein A sepharose represents the average binding percentages across all IgG subclasses. As the Protein A sepharose binding capacity is large enough to quantitatively bind all "binding" IgG subtypes, the binding fraction would remain consistent regardless of the applied concentration.

Figure 4-1 illustrates the average binding percentage of IgG from different species to Protein A sepharose, along with HSA as a negative control, at various molar concentrations. The results demonstrate a high affinity of IgG from human and rabbit serum, a moderate affinity for bovine IgG, and a weak affinity of goat IgG to Protein A. Furthermore, HSA

exhibited no propensity to bind to Protein A sepharose. The calculated average binding percentage for IgG from human, rabbit, bovine, goat, and HSA were determined as 98%, 95%, 60%, 22%, and 0% respectively. These findings suggest that within the selected concentration range, binding was not significantly influenced by IgG concentration. Instead, it was likely determined by the relative abundance of binding subtypes within the polyclonal IgG mixtures used.

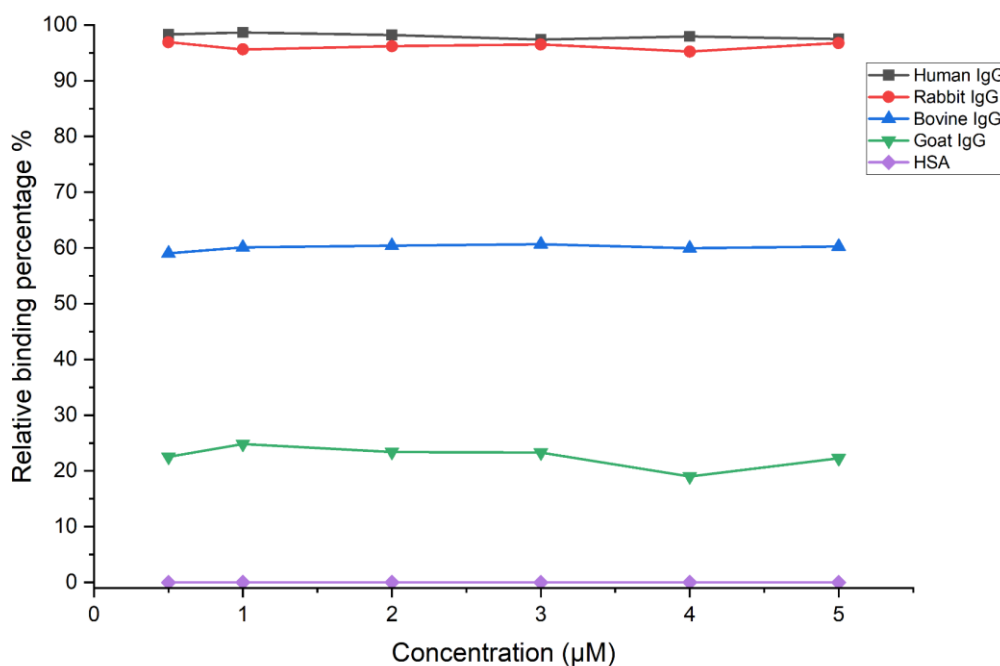


Figure 4-1: Relative average binding percentages of IgG from different species (human, rabbit, bovine, goat) to Protein A sepharose at varying concentrations (0.5 to 5 μ M) in sodium phosphate buffer, pH=8.1. The graph demonstrates high binding affinity for human and rabbit IgG, moderate affinity for bovine IgG, and low affinity for goat IgG. HSA showed no significant binding. These results suggest that binding percentages are primarily influenced by the presence of IgG subtypes capable of binding to Protein A, rather than the total concentration of IgG applied.

The immobilization of Protein A on sepharose resin enhances the availability of its binding sites for target proteins, such as human IgG, facilitating efficient binding. However, it's important to note that the Protein A used in Protein A sepharose columns may vary from the recombinant Protein A utilized in this project, potentially resulting in some variation in results.

4.1.2. CD spectroscopy analysis

The content discussed in this section of the chapter has been published and can be accessed via <https://doi.org/10.1002/ssr.202400204>

CD spectroscopy is a widely used analytical technique in studying protein-protein interactions, offering insights into the secondary structure and conformational changes of proteins. This technique is particularly useful for assessing alterations in alpha-helices and beta-sheets, which are often indicative of binding events or structural rearrangements during interactions.

In this study, CD spectroscopy was employed to evaluate the structural characteristics of IgG antibodies from various species (human, rabbit, bovine, goat) in a mixture with Protein A at different molar ratios, similar to the ratios used in the CNN image classification analysis in the following sections. By analyzing these samples, we can assess how the interaction between IgG and Protein A affects their secondary structures, providing a deeper understanding of species-specific differences in binding affinity and structural responses.

Figure 4-2 illustrate the relative binding strengths of human IgG to recombinant Protein A, derived from CD spectroscopy measurements. In these measurements, CD spectra of individual proteins were acquired and aggregated to generate the spectra of their respective mixtures. Subsequently, these aggregated spectra were compared to the CD spectra of the actual protein mixture solution. Notably, in cases of robust protein interaction, discrepancies between the CD spectra of the mixture and the aggregated spectra of the individual proteins emerge, attributed to alterations in their secondary structure resulting from protein-protein interactions.^[149,234]

By analyzing the CD intensity at 217 nm ^[149,234], the percentage of relative change in secondary structure can be quantified. The CD spectroscopy measurements for rabbit IgG,

bovine IgG, goat IgG, and HSA mixed with Protein A can be found in Appendix B (Figures B-1 to B-4). The maximum detected changes for Human Serum Albumin (HSA), goat IgG, bovine IgG, rabbit IgG, and human IgG in their respective mixtures with Protein A were determined as 6%, 9%, 16%, 16%, and 81% respectively. This delineates human IgG as exhibiting a strong affinity for Protein A, while rabbit IgG and bovine IgG demonstrate weaker to moderate interaction strengths. Conversely, negligible interactions were observed for goat IgG and HSA (Figures B-1 to B-4). These findings align with previous literature reports ^[28,97,103,105,107,235,236], reinforcing the consistency and reliability of the results. Furthermore, the similarity in secondary structure among IgG complexes across different molar ratios was also observed.

Figure 4-3 visually represents the diverse binding affinities of these proteins to Protein A through a color map, providing a comprehensive overview of the relative strengths of protein-protein interactions.

The data depicted in **Figures 4-3** underscore the differential binding affinities of various proteins to Protein A, shedding light on the nuanced interactions within IgG complexes across different molar ratios.

Human IgG:Protein A

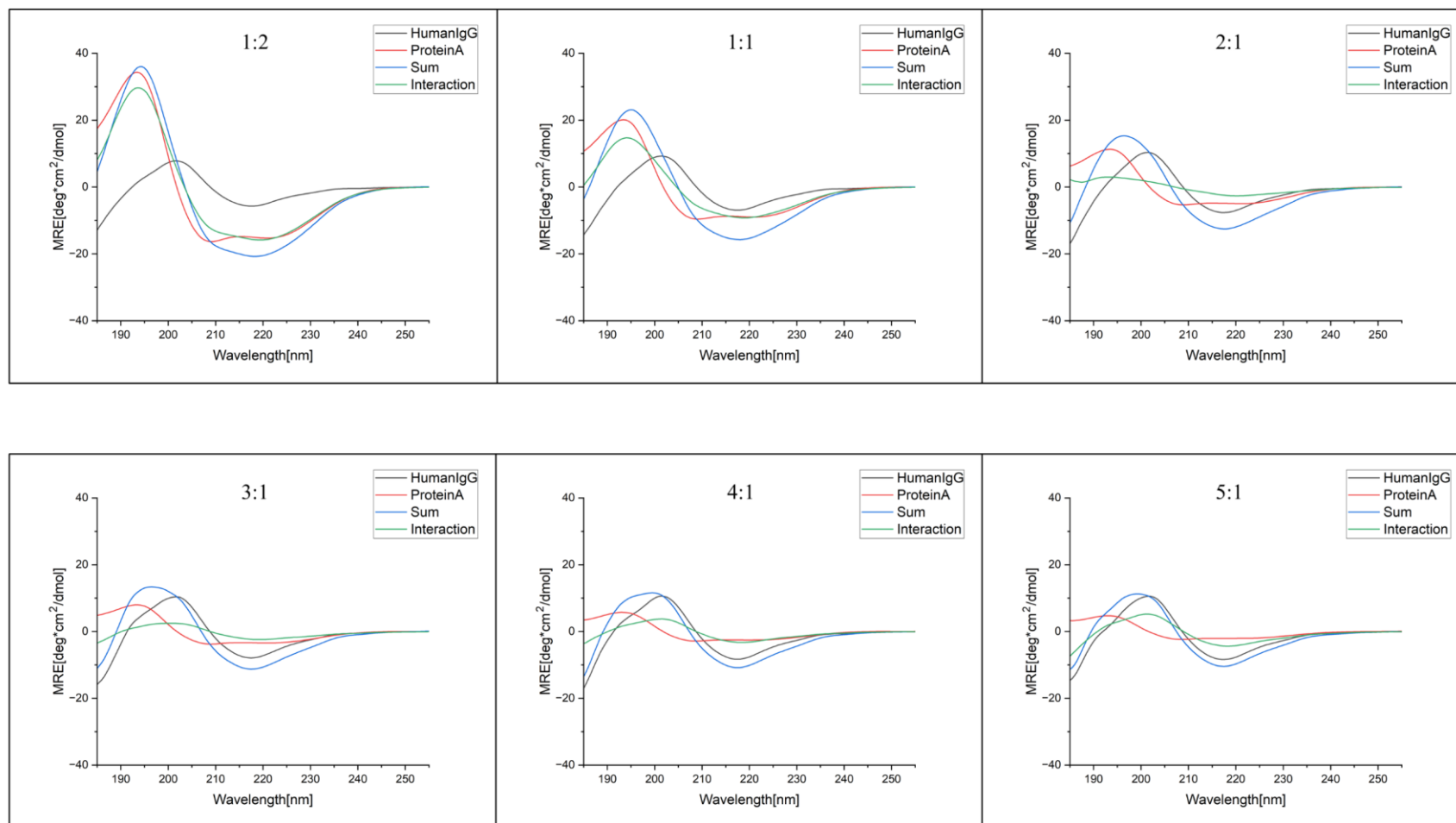


Figure 4-2: CD spectroscopy measurements results for human IgG:Protein A in different molar ratios with the constant total mass concentration. The deviations between summation and interaction spectra, particularly observed at 217 nm, indicate alterations in the secondary structure of the protein mixture via interaction.

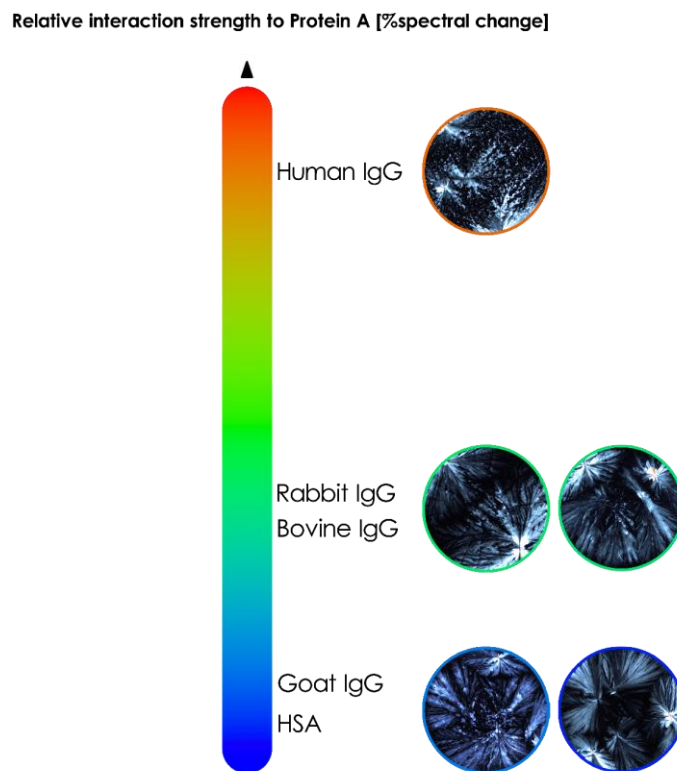


Figure 4-3: Illustration of relative binding affinity of IgG from different species to Protein A based on CD spectroscopy measurement.

4.2. Classification of interaction strength for IgG:Protein A complexes using neural networks

The content discussed in this section, as well as subsections 4.2.1.3 and 4.2.1.4 of the chapter, has been published and can be accessed via <https://doi.org/10.1002/ssstr.202400204>

To investigate the typification of IgG complexes, their deposition patterns were analyzed using Polarized Light Microscopy (PLM). Protein solutions were prepared in 0.1 M sodium phosphate buffer with a pH of 8.1, a commonly utilized medium for IgG purification.^[237,238] Despite maintaining a constant total mass of all protein samples at 0.3 mg/ml, various molar ratios were employed for screening protein-protein interactions. To ensure uniform droplet radii (2 mm), a hydrophobic substrate was employed. Utilizing

chemical vapor deposition (CVD) polymerization, poly(p-xylylene) coatings were applied to glass surfaces, a critical step for achieving consistent and reproducible droplet deposition over broad areas. Circular droplets were deposited and allowed to dry under controlled humidity and temperature conditions for a minimum of six hours (**Figure 4-4, A-B**).

After drying, PLM imaging was used to capture images of the deposition patterns, maintaining consistent parameters such as resolution, magnification, light intensity, gain, and exposure time (**Figure 4-4, C**). To reveal the chemical composition of the drying patterns, Time-of-Flight Secondary Ion Mass Spectrometry (ToF-SIMS) was utilized. Signal mapping highlighted proteins (represented by CNO^- ions) and buffer components (represented by PO_4^{2-} ions), indicating the tendency for solution components to segregate into distinct agglomerates of either protein or buffer components upon drying (**Figure 4-4, E**). This observation was further corroborated by scanning electron microscopy (SEM), which revealed the presence of crystalline structures in the central region, indicative of high salt content (**Figure 4-4, D**).

The investigation extended to IgG from four different species and human serum albumin (HSA) as a negative control.^[92,106,111] Different molar ratios were employed to assess varying levels of protein-protein interactions. **Figure 4-5** summarizes the stain patterns obtained from the IgGs, HSA, and Protein A, as well as from various protein-protein combinations. Notably, patterns of complexes involving IgGs with low affinity for Protein A, such as Goat IgG, closely resembled IgG patterns in the absence of Protein A. This trend was similarly observed in the case of HSA-Protein A mixtures, highlighting the visual similarity between patterns of complexes and their respective components.

After generating a sufficient number of images depicting the deposited protein patterns, the dataset will be prepared for the subsequent image classification tasks using a CNN. This curated dataset will serve as the foundation for training and evaluating the CNN model, allowing it to learn and differentiate between various protein pattern classes.

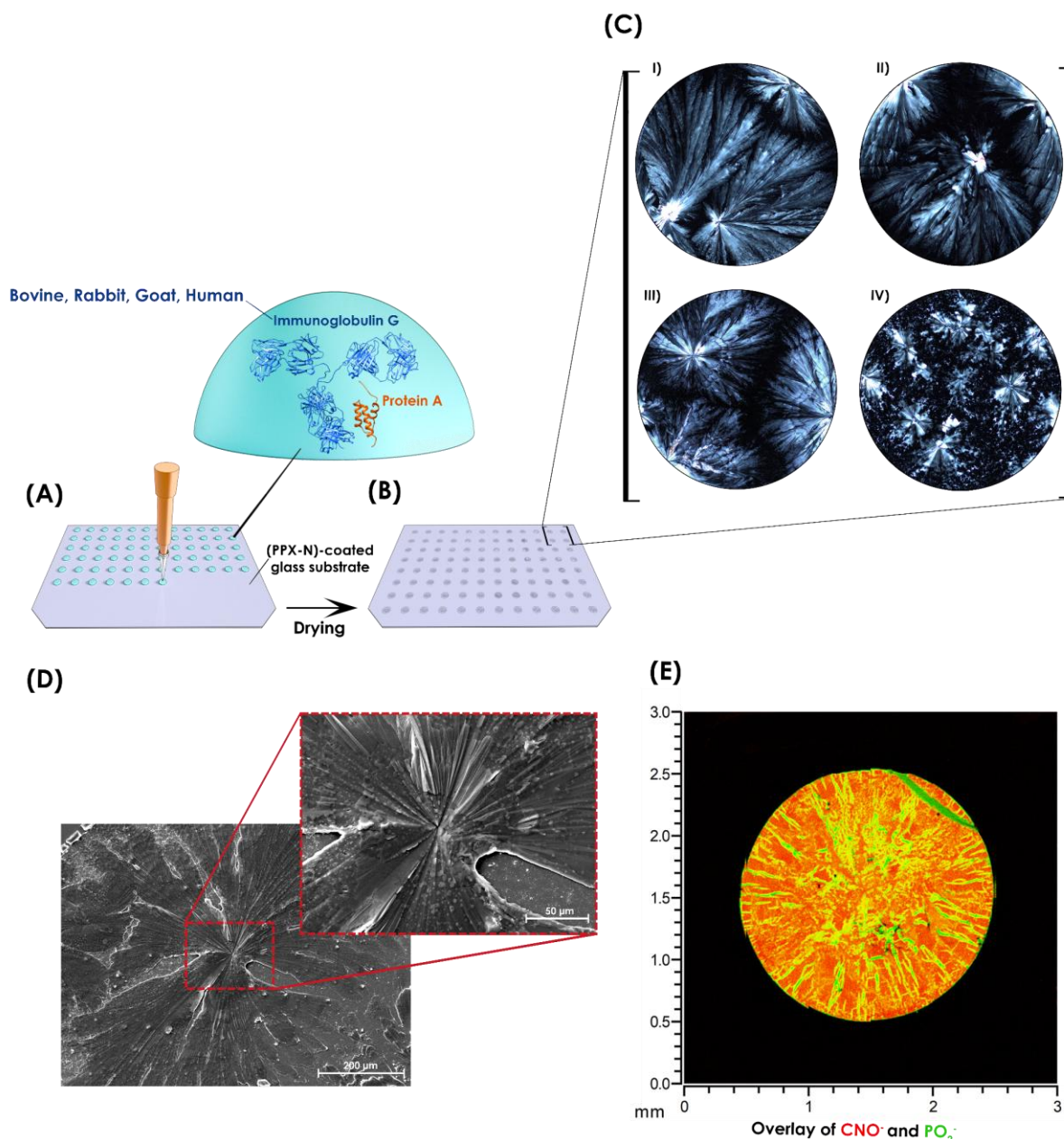


Figure 4-4: Formation of protein stains using controlled droplet deposition and drying process.

A) Cleaned glass substrates were coated with poly(*p*-xylylene) via CVD polymerization in order to obtain reliable hydrophobic surface conditions to ensure a reproducible surface interaction. An automated pipetting system was used for dispensing several protein sample droplets (2 μl) containing different molar ratios of IgG from different species and Protein A. B) Dispensed droplets were dried under controlled environmental conditions ($T = 25^\circ\text{C}$, relative humidity = 40 %). C) PLM imaging was used to collect all the deposited proteins' patterns under the same conditions to prepare sufficient images for each category for image classification with CNN implementation, IgG from (I) bovine, (II) rabbit, (III) goat and (IV) human serum interacting with Protein A. D) SEM image analysis of human IgG stains. E) ToF-SIMS analysis of the deposition pattern of a complex of Goat IgG with Protein A: RGB overlay image of the distribution map of PO_2^- ions (green) and CNO^- ions (red). Adapted from ^[232]

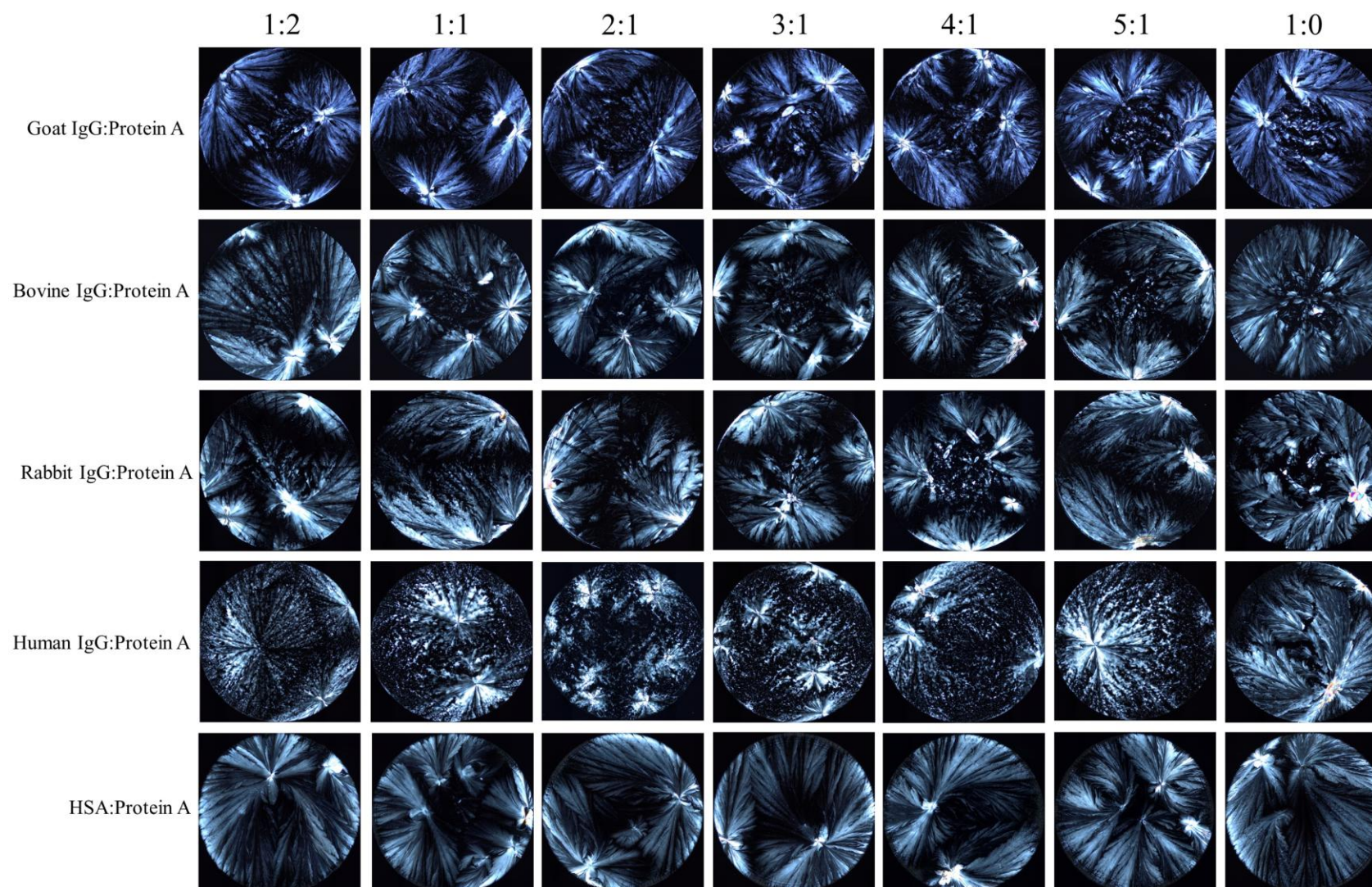


Figure 4-5: The deposited patterns of IgG from different species and Protein A with various molar ratios. The ratio 1:0 is regarding the single IgG/HSA pattern. Adapted from ^[232]

4.2.1. Image classification using a pretrained CNN

In this section, the results of image classification using images obtained from the deposited dried patterns of protein mixtures are presented. The convolutional neural network utilized for this task was InceptionV3, featuring 315 hidden layers, chosen for its reasonable training time compared to other high-performance CNNs such as NasNetLarge.^[18]

The process involved introducing all relevant labeled image classes into the CNN for both training and validation. Following training and validation, groups of distinct images as test dataset were subjected to classification using the trained CNN. The outcome was a confusion matrix, where the Y-axis represents the labels of true classes and the X-axis represents the labels of predicted image classes. Correctly predicted images are depicted along the diagonal in bluish color, while incorrect predictions are displayed in reddish color elsewhere in the matrix. These misclassifications can be examined to understand the reasons behind the incorrect decisions made by the CNN. Additionally, the percentage accuracy for each individual class is shown on the right side of the confusion matrix, with the total accuracy representing the average of these single-class prediction accuracies.

For training, at least 400 images were prepared for each class,^[184,239,240] with 10% randomly selected for validation. Additionally, 100 distinct images for each class were allocated for the testing dataset, ensuring no overlap between training, validation, and testing datasets.

4.2.1.1. Classification of single proteins

Initially, single protein images were used as the input dataset, with patterns of IgG from different species compared using the CNN. These polyclonal IgGs share similar molecular weights and secondary structures, with only minor differences observed in their patterns. (as shown in **Figure 4-5**, last column). Notably, InceptionV3 distinguished between these patterns

effectively, as demonstrated in **Figure 4-6**. The overall accuracy of this confusion matrix was determined to be 98.75%, indicating the reliable performance of InceptionV3 in categorizing similar patterns with subtle structural differences.

True Class	Bovine-IgG	100				100.0%	
	Goat-IgG	1	98	1		98.0%	2.0%
	Human-IgG			98	2	98.0%	2.0%
	Rabbit-IgG			1	99	99.0%	1.0%
		Bovine-IgG	Goat-IgG	Human-IgG	Rabbit-IgG		
		Predicted Class					

Figure 4-6: Confusion matrix of IgG from different species. The number of true positives (correctly classified) classes are placed on the diagonal with a bluish color, while other cells represent the misclassifications. The overall accuracy of the prediction is 98.75%, reflecting the proportion of correctly predicted instances out of the total instances.

4.2.1.2. Classification on transformed test dataset

In the initial performance evaluation of the trained InceptionV3, we conducted a rigorous examination by subjecting the model to testing with transformed images. This comprehensive assessment aimed to scrutinize the robustness of the CNN's classification capabilities in the face of image transformations, specifically horizontal and vertical flips, as well as combined horizontal and vertical flips.

The dataset utilized for this evaluation consisted of distinct image classes, prominently featuring human IgG and Protein A. Within this dataset were single proteins (human IgG and Protein A) and various molar ratios resulting from their mixing. Notably, human IgG's strong affinity for binding to Protein A is well-documented, with an established optimum binding ratio of 2:1.^[107,241] This ratio, validated through CD spectroscopy measurements, manifests the most significant deviation in secondary structure, signifying heightened interaction within the

solution.

Initially, the CNN underwent testing using the original test dataset, yielding an impressive total accuracy of 92.5% shown in the obtained confusion matrix in **Figure 4-7A**. Strikingly, the 2:1 binding ratio displayed the highest-class prediction accuracy at 98%, indicative of the CNN's adeptness in discerning the optimal binding configuration.

Subsequently, we embarked on a meticulous evaluation of the CNN's performance under transformed image conditions. Employing the same test dataset, we subjected the images to horizontal, vertical, and combined horizontal-vertical flips. The resulting confusion matrices revealed total prediction accuracies of 93%, 93.4%, and 93% respectively (**Figure 4-7, B-D**). Noteworthy is the consistent accuracy in predicting single proteins across all transformations, reaffirming the CNN's robustness in distinguishing between individual proteins. Moreover, the 2:1 binding ratio consistently exhibited the highest prediction accuracy among different molar ratios, underscoring the CNN's reliability in identifying optimal binding configurations despite image transformations.

In essence, this comprehensive evaluation underscores the resilience and efficacy of the trained InceptionV3 CNN model in accurately classifying transformed images, thereby enhancing our confidence in its suitability for complex image classification tasks.

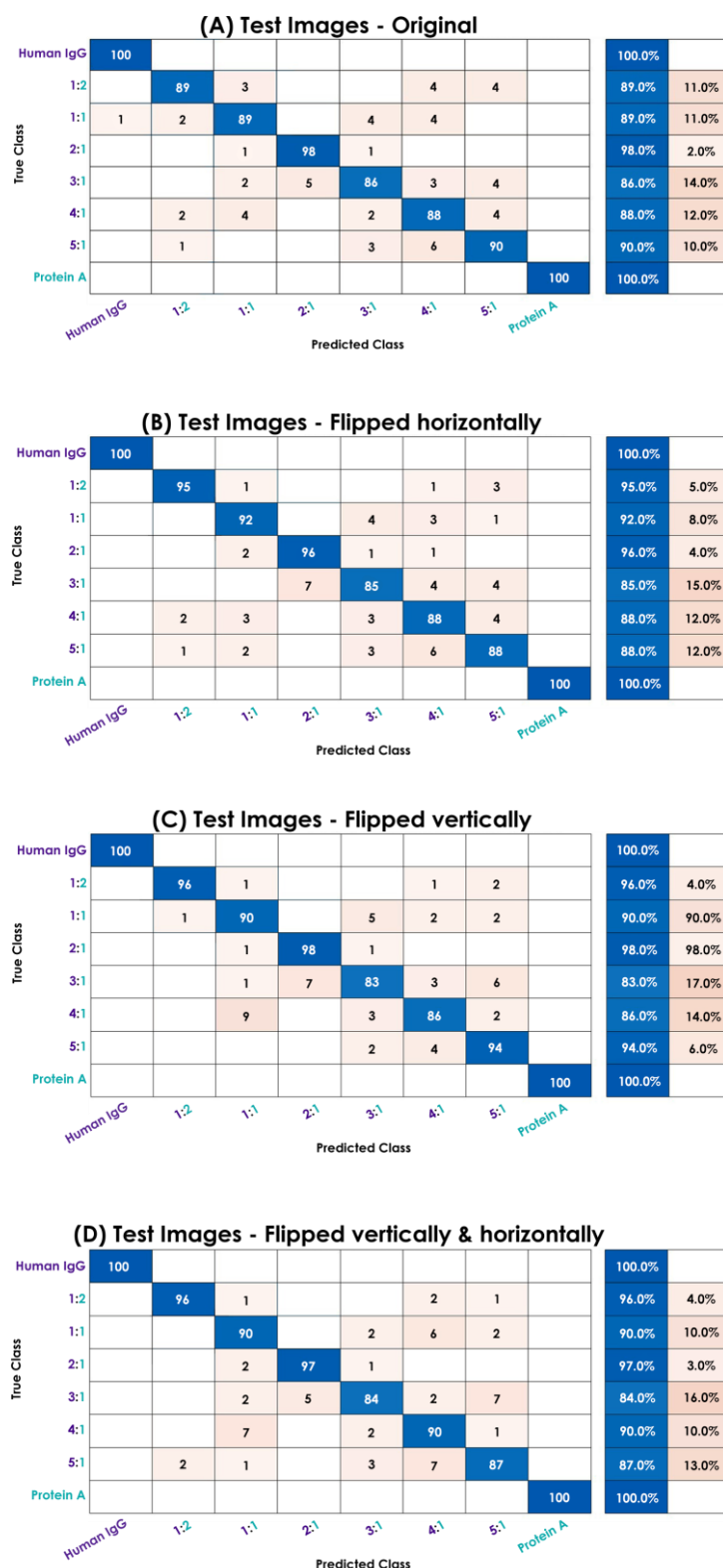


Figure 4-7: Confusion matrix of human IgG:Protein A image classes across different molar ratios, evaluated using the original and transformed test datasets. A) Original dataset with a prediction accuracy of 92.5%. B) Horizontally flipped dataset with 93% accuracy. C) Vertically flipped dataset achieving 93.4% accuracy. D) Dataset flipped both horizontally and vertically with 93% accuracy.

4.2.1.3. Classification of IgG:Protein A complexes with various molar ratios

In this section, we focus on the classification of IgG:Protein A complexes formed at various molar ratios using IgG antibodies from different species based on the PLM imaging technique. Accurate classification of these complexes provides insight into the binding characteristics and behavior of antibodies across species, which is crucial for understanding protein-protein interactions. To achieve this, we employ the InceptionV3 convolutional neural network, a powerful deep learning architecture known for its ability to efficiently handle complex image classification tasks mentioned previously. By applying this architecture, we aim to accurately differentiate IgG:Protein A interactions based on PLM imaging of the patterns they deposited on the surface, enhancing our understanding of antibody behavior across species. The sample patterns are shown in **Figure 4-5**.

In **Figure 4-8**, we present the confusion matrices detailing the prediction outcomes of Protein A immunocomplexes involving IgGs from four distinct species: human, rabbit, bovine, and goat. These IgGs exhibit a spectrum of binding affinities to Protein A, ranging from weak to strong, as elucidated in **Figure 4-3**. The assessment encompasses a range of molar ratios, spanning from 0.5:1 to 5:1 (IgG:Protein A).

The comprehensive confusion matrix compiled from 36 distinct immunoprotein complexes provides insight into the predictive capabilities of the CNN concerning the binding affinity of the protein complexes (as illustrated in **Figure 4-9**). For a more species-specific examination of the four immunoglobulin complexes, **Figure 4-8** offers a breakdown of the comprehensive confusion matrix. Additionally, to facilitate clustering analysis of the trained network, t-distributed stochastic neighbor embedding (t-SNE) plot was generated for each set.^[219]

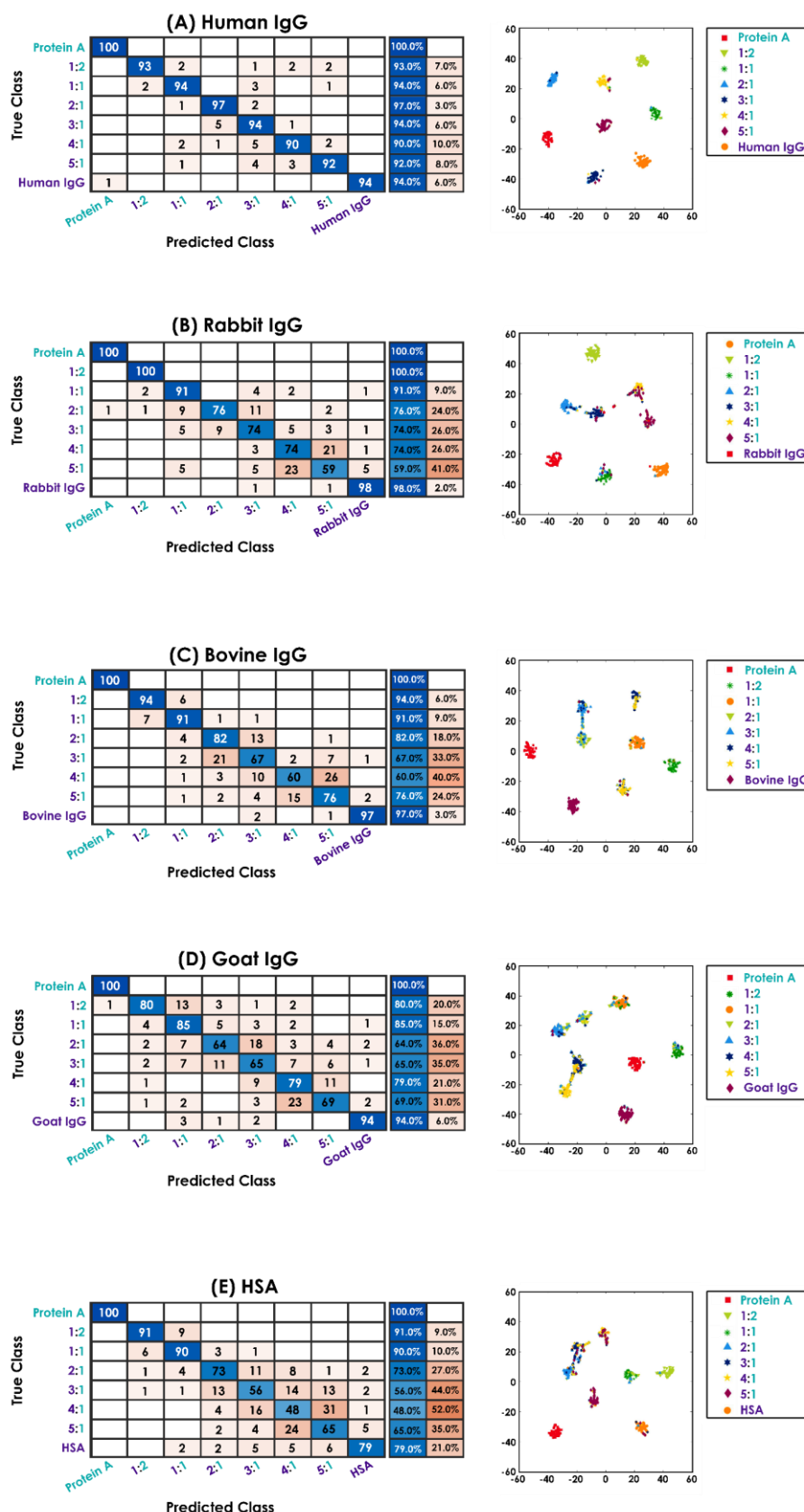


Figure 4-8: Confusion matrix of different protein-protein interactions. A) human IgG, B) rabbit IgG, C) bovine IgG, D) goat IgG, E) HSA. The confusion chart was obtained for a test set of 36 categories divided into smaller charts with t-SNE plot analysis for each protein complex. Adapted from [232]

Each t-SNE plot was applied to the "Softmax" layer of the CNN, yielding a 4-D array comprising the x- and y-spatial dimensions of the images, channels of the images, and batch dimension, respectively. By visualizing the output of this layer, high-dimensional data were effectively evaluated in the t-SNE plot, thereby aiding in the classification of distinct image classes. Notably, the t-SNE plots demonstrated robust clustering, indicative of the CNN's efficacy in distinguishing between different protein-protein interaction levels.

Figure 4-8A showcases the confusion matrix pertaining to human IgG and Protein A image classes. With a reported high affinity for Protein A, human IgG is expected to exhibit a pronounced interaction,^[13,27] as verified by CD spectroscopy (**Figure 4-2**). The CNN distinguished the different molar ratios of human IgG complexes with an overall accuracy of 93.4 %. Misclassifications with rabbit IgG complexes were observed (**Figure 4-9**), which can be attributed to the similarities in molecular weight and secondary structure of human IgG and rabbit IgG. The confusion matrix confirms the CNN's capability to discern the optimal binding ratio of 2:1 (IgG:Protein A), achieving a prediction accuracy of 97%. This aligns with the experimentally determined optimal binding ratio, as corroborated by previous studies.^[107,241] The t-SNE plot also indicates very good clustering.

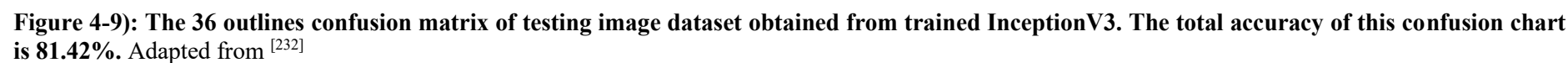
Conversely, **Figure 4-8B** depicts the confusion chart for rabbit IgG:Protein A image classes. CD spectroscopy data revealed a weaker to medium affinity of rabbit IgG for Protein A (**Figure B-1**), with negligible changes observed in the secondary structure compared to human IgG complexes.^[105] Despite this, the CNN exhibited some misclassifications between human and rabbit IgG images, resulting in an accuracy of 81.7% which revealed several misclassifications between different molar ratios (**Figure B-1**). The t-SNE plot provides further insight into the clustering behavior, highlighting areas of potential confusion.

Continuing to **Figure 4-8C**, we delve into the confusion matrix of bovine IgG:Protein A image classes. Analogous to rabbit IgG, bovine IgG exhibits weak to medium binding affinity

for Protein A, resulting in an accuracy of 81%. Misclassifications primarily stem from similarities in staining patterns, compounded by variations in molar ratios. These similarities may be attributed to small changes in the secondary structure of the protein solution upon complexation, a tendency that was also observed by CD spectroscopy (**Figure B-2**) and several literatures.^[97,99,242] These nuances are reflected in the t-SNE plot, albeit with discernible clustering among single protein and high Protein A content samples. Additionally, for single bovine IgG and for the molar ratio that the Protein A content is higher than bovine IgG (0.5:1), clustering worked well with an accuracy of 97 % and 94 %, respectively. Furthermore, the trained InceptionV3 did not misclassify bovine IgG with IgGs from the three other species (**Figure 4-9**).

Moving to **Figure 4-8D**, the confusion chart of goat IgG:Protein A image classes is presented. Characterized by weak interactions with Protein A,^[100,107] goat IgG exhibits a high error rate in predicting different molar ratios, resulting in a local accuracy of 76.6%. The t-SNE plot reflects this uncertainty, displaying limited clustering of protein pattern classes across various molar ratios.

Lastly, **Figure 4-8E** showcases the confusion matrix of HSA:Protein A image classification. As a control protein lacking propensity for Protein A binding, HSA yielded the lowest prediction accuracy among the tested complexes. Notably, no evidence of HSA-Protein A complex formation was observed across all molar ratios, resulting in a local accuracy of 71.7%. The t-SNE plot corroborates these findings, highlighting the disparate clustering of HSA:Protein A images.



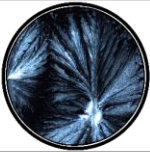

The total accuracy of the comprehensive confusion matrix is determined to be 81.4%, as depicted in **Figure 4-9**. It is imperative to note that the training of the CNN necessitated significant computational resources, amounting to 26951 minutes (approximately 19 days) using an NVIDIA TITAN RTX 24G GPU. This underscores the computational complexity inherent in training neural networks for image classification tasks, emphasizing the importance of optimizing computational resources for efficient model development and training.

4.2.1.4. Image classification robustness

In order to evaluate the performance of the trained network with IgG from different sources and Protein A (**Figure 4-9**), we introduced the patterns of Protein G and its interaction with IgG from human serum as unknowns to observe how the network classifies them across various interaction categories and molar ratios. Both Protein G and Protein A serve as superantigens for human IgG binding, sharing similar functionality and structural properties, as outlined in **Table 4-1**. Despite Protein G possessing fewer binding sites for human IgG compared to Protein A, its affinity for IgGs is notably higher,^[112,243] as evidenced by CD spectroscopy analysis showing similar interaction levels between human IgG:Protein G and human IgG:Protein A (**Figure B-5**).

Table 4-1: Characteristics of the applied recombinant Protein A and Protein G. Adapted from

[232]

Properties Ligand	MW (kDa)	Secondary Structure	No. of Active sites for binding IgG	Binding Constant for human IgG (M ⁻¹) * 10 ⁻⁹	Relative Percentage bond to human IgG	Deposited pattern
Protein A	36	Three alpha-helices	5	44.1	81%	
Protein G	31	one alpha-helix packed onto a four-stranded β-sheet	2	67.4	83%	

In **Figure 4-10**, we present the confusion chart generated from test images of human IgG:Protein G classes. Interestingly, although Protein G patterns exhibit visual distinctions from those of Protein A (**Figure 4-11**), the trained network adeptly classified 83% of them into the Protein A image class (depicted with purple brackets in the last row). Additionally, 13% of Protein G images were predicted as HSA:Protein A with a molar ratio of 1:2, attributed to the aforementioned structural similarities between these proteins. Further classification efforts on four different image datasets of mixtures of IgGs with Protein G yielded similar results, with the predominant prediction (94%) aligning with human IgG:Protein A image classes (displayed with purple brackets).

True Class	HlgG-ProteinG-1-1	HlgG-ProteinG-1-2	HlgG-ProteinG-2-1	HlgG-ProteinG-3-1	HlgG-ProteinG-4-1	HlgG-ProteinG-5-1	ProteinG	BlgG-ProteinA-1-1	BlgG-ProteinA-1-2	BlgG-ProteinA-2-1	BlgG-ProteinA-3-1	BlgG-ProteinA-4-1	BlgG-ProteinA-5-1	BovineIgG	GlgG-ProteinA-1-1	GlgG-ProteinA-1-2	GlgG-ProteinA-2-1	GlgG-ProteinA-3-1	GlgG-ProteinA-4-1	GlgG-ProteinA-5-1	GoatIgG	HlgG	HlgG-ProteinA-1-1	HlgG-ProteinA-1-2	HlgG-ProteinA-2-1	HlgG-ProteinA-3-1	HlgG-ProteinA-4-1	HlgG-ProteinA-5-1	ProteinA	HSA	HSA-ProteinA-1-1	HSA-ProteinA-1-2	HSA-ProteinA-2-1	HSA-ProteinA-3-1	HSA-ProteinA-4-1	RlgG-ProteinA-1-1	RlgG-ProteinA-1-2	RlgG-ProteinA-2-1	RlgG-ProteinA-3-1	RlgG-ProteinA-4-1	RlgG-ProteinA-5-1	RabbitIgG						
HlgG-ProteinG-1-1						1																10	11	2	58		1	4																	1	1		11
HlgG-ProteinG-1-2		2																				30	1	97																								
HlgG-ProteinG-2-1	1																					11	15		94																							
HlgG-ProteinG-3-1																						51	2	38	2																							
HlgG-ProteinG-4-1		2																				77	4	8	3	2	3																					
HlgG-ProteinG-5-1		2																				66	7	9	11	2	5																					
ProteinG																						1																										
BlgG-ProteinA-1-1																																																
BlgG-ProteinA-1-2																																																
BlgG-ProteinA-2-1																																																
BlgG-ProteinA-3-1																																																
BlgG-ProteinA-4-1																																																
BlgG-ProteinA-5-1																																																
BovineIgG																																																
GlgG-ProteinA-1-1																																																
GlgG-ProteinA-1-2																																																
GlgG-ProteinA-2-1																																																
GlgG-ProteinA-3-1																																																
GlgG-ProteinA-4-1																																																
GlgG-ProteinA-5-1																																																
GoatIgG																																																
HlgG																																																
HlgG-ProteinA-1-1																																																
HlgG-ProteinA-1-2																																																
HlgG-ProteinA-2-1																																																
HlgG-ProteinA-3-1																																																
HlgG-ProteinA-4-1																																																
HlgG-ProteinA-5-1																																																
ProteinA																																																
HSA																																																
HSA-ProteinA-1-1																																																
HSA-ProteinA-1-2																																																
HSA-ProteinA-2-1																																																
HSA-ProteinA-3-1																																																
HSA-ProteinA-4-1																																																
RlgG-ProteinA-1-1																																																
RlgG-ProteinA-1-2																																																
RlgG-ProteinA-2-1																																																
RlgG-ProteinA-3-1																																																
RlgG-ProteinA-4-1																																																
RlgG-ProteinA-5-1																																																
RabbitIgG																																																

Figure 4-10: Performance of a pre-trained network using human IgG:Protein G patterns. Confusion chart of human IgG:Protein G image classification as an unknown sample (not trained). Adapted from ^[232]

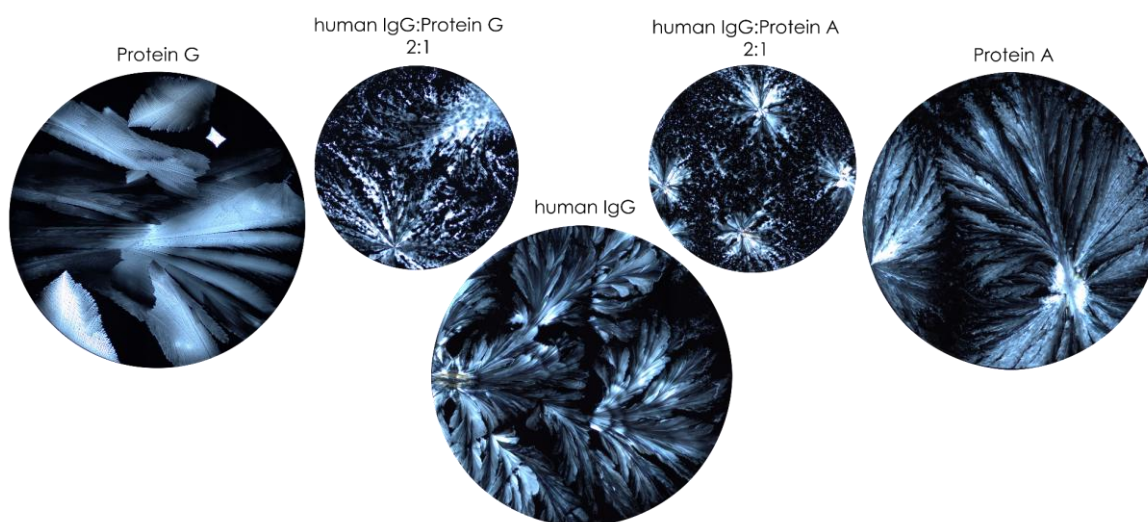


Figure 4-11: Examples of obtained patterns of Protein A, Protein G, human IgG, human IgG:Protein G with a molar ratio of 2:1 in comparison to Protein A interaction with human IgG pattern with the same molar ratio.

Figure 4-11 provides examples of patterns of human IgG:Protein G with a molar ratio of 2:1, alongside human IgG and human IgG:Protein A, illustrating the striking similarities between the Protein G:human IgG complex patterns and those of the Protein A: human IgG complexes compared to single Protein G and Protein A images.

To delve deeper into the network's decision-making process, we performed Gradient-weighted Class Activation Mapping (Grad-CAM) analysis.^[221] Grad-CAM elucidates the rationale behind the network's classification decisions by highlighting regions within images that heavily influence classification outcomes. As showcased in **Figure 4-12**, the Grad-CAM analysis of single images of human IgG-Protein G reveals that regions with warmer colors in the heatmap correspond to areas with higher prediction scores. Notably, these crucial regions predominantly reside within the patterns' central areas, reaffirming their significance in the classification process, while background elements play a lesser role, as anticipated.

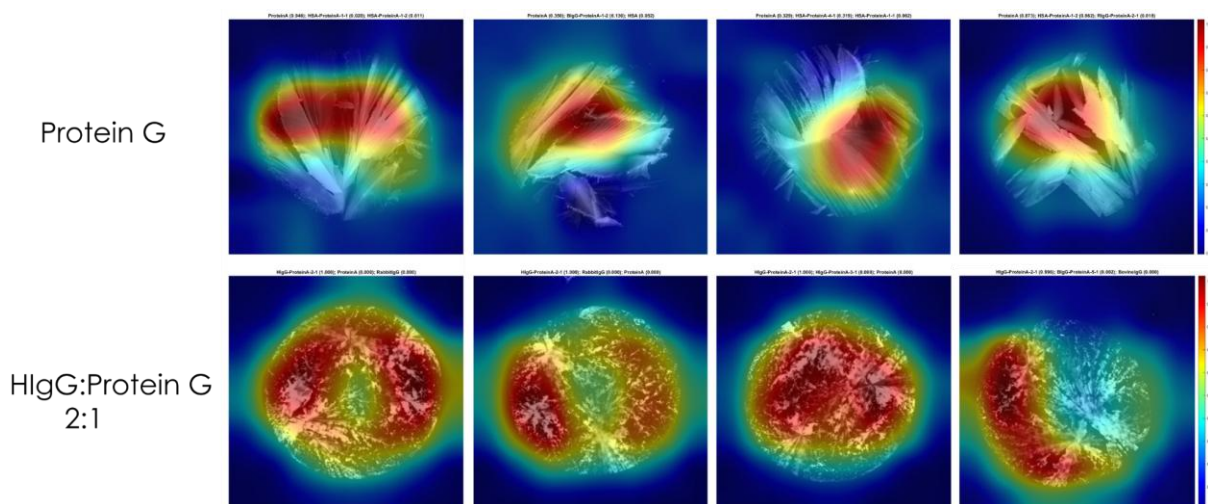


Figure 4-12: Grad-Cam image analysis. The heatmap overlay shows areas of importance that contributed to the model's decision, offering insights into how the CNN interprets IgG:Protein A complex patterns.

4.2.2. Feature classification based on graph theory analysis

This section marks a shift from image classification to feature classification, focusing on analyzing the extracted features from protein patterns using neural networks. While the previous section employed convolutional neural networks to classify images based on protein-protein interaction patterns, this section adopts a different perspective by analyzing numerical representations of these patterns through graph theory. By concentrating on the key features from the dataset, the aim is to improve classification efficiency while minimizing computational cost. This section provides a comprehensive overview of the feature classification approach and demonstrates its application to the IgG:Protein A complexes examined earlier.

The optimization strategy employed to reduce the training time and computational cost involved leveraging graph theory, a powerful mathematical framework for modeling and analyzing complex systems. As described previously (refer to sections 3.17), we utilized the StructuralGT Python package to convert the images used for classification into graphs, enabling us to extract meaningful features from these graphical representations.

Rather than directly utilizing the graphs themselves, we focused on extracting pertinent features from them. From a total of 20 features generated by the StructuralGT package, we identified 15 features that were deemed meaningful for our classification task (refer to section 3.17, **Table 3-4**). Consequently, for each image, 15 numerical features using graph theory analysis were computed, resulting in a comprehensive dataset of image features. Additionally, by utilizing 5 distinct input parameters in the StructuralGT tool (refer to section 3.17, **Table 3-5**), as depicted in **Figure 4-13**, a total of 75 features per image (15×5) were generated. This approach enhanced the feature representation, capturing the intricate details of the patterns, as discussed in section 3.17.

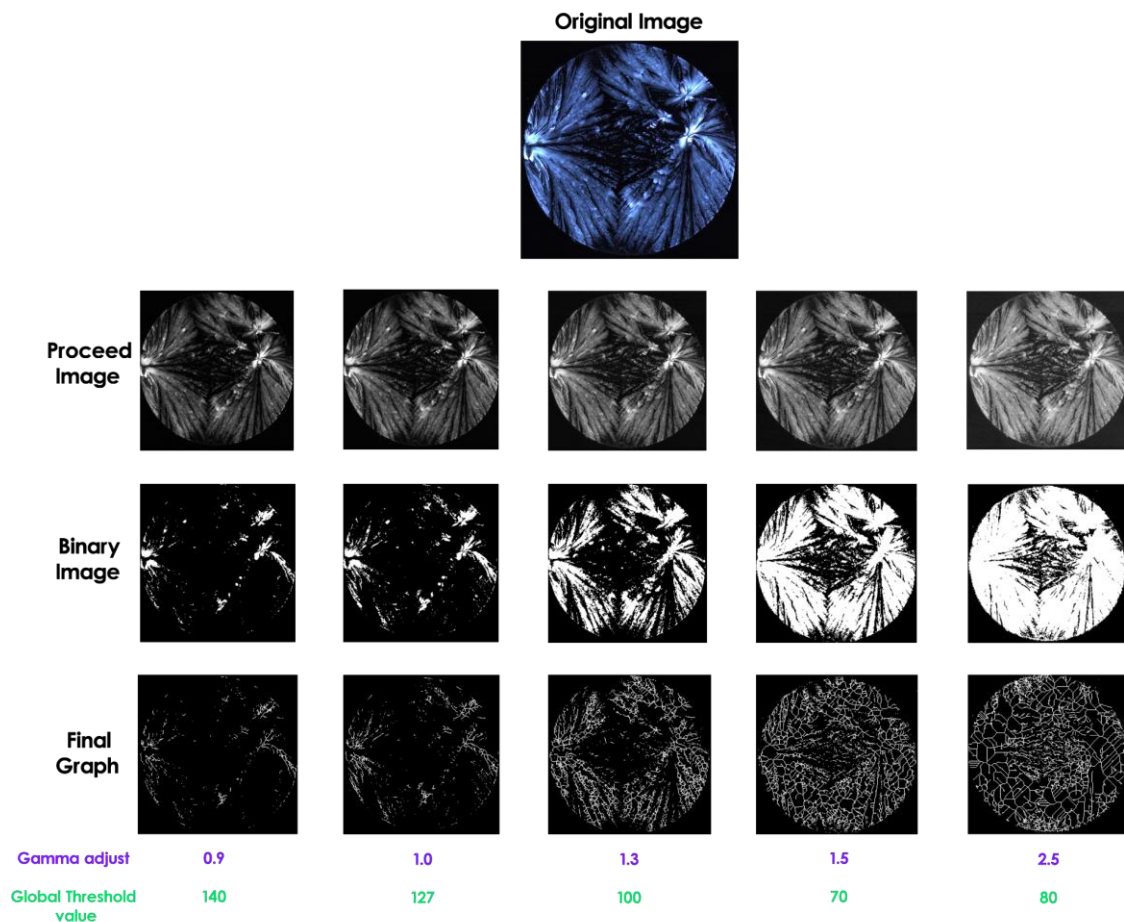


Figure 4-13: Image-to-graph conversion using graph theory analysis. Applying five different gamma adjust and global threshold values as input parameters to capture the most intricate details of the given images.

To enhance the network's recognition capabilities, we also derived new features from the graph theory features by applying simple mathematical operations. These operations served to augment the dataset with additional information that could potentially aid in improving classification accuracy. Some of the operations included:

- Summation: Aggregating feature values to capture overall structural characteristics.
- Difference: Calculating the disparity between feature values to highlight variations.
- Ratio: Determining the proportion between feature values to discern relative magnitudes.
- Product: Multiplying feature values to capture combined effects.

In this study, only summation and multiplication (product) operations were employed. These two operations were chosen for their simplicity and effectiveness in combining existing features, allowing the network to capture interactions between them without introducing overly complex transformations. Summation captures the cumulative relationships between different features, while multiplication helps model interactions and correlations. By focusing on these operations, we sought to provide the network with richer, yet interpretable, data that could contribute to improving classification accuracy without increasing computational complexity. By incorporating these derived features alongside the original graph theory features, we aimed to enrich the dataset with comprehensive information that could enhance the network's ability to discern patterns and make accurate classifications. This approach not only streamlined the training process but also optimized computational resources by reducing the complexity of the input data while retaining relevant information essential for classification tasks. For each set of input parameters, the 6 new features were computed as described below:

$$\mathbf{sumNodesEdges} = \text{Average Clustering Coefficient} \times (\text{Number of Nodes} + \text{Number of Edges}) \quad (4-1)$$

$$\mathbf{sumAssortativity} = \text{Average Clustering Coefficient} \times (\text{Assortativity Coefficient} + \text{Weighted Assortativity Coefficient}) \quad (4-2)$$

$$\mathbf{sumDegree} = \text{Average Clustering Coefficient} \times (\text{Average Degree} + \text{Weighted Average Degree}) \quad (4-3)$$

$$\mathbf{sumBetweennessCentrality} = \text{Average Clustering Coefficient} \times (\text{Average Betweenness Centrality} + \text{Width Weighted Average Betweenness Centrality}) \quad (4-4)$$

$$\mathbf{sumClosenessCentrality} = \text{Average Clustering Coefficient} \times (\text{Average Closeness Centrality} + \text{Length Weighted Average Closeness Centrality}) \quad (4-5)$$

$$\text{sumEigenvectorCentrality} = \text{Average Clustering Coefficient} \times (\text{Average Eigen Vector Centrality} + \text{Width Weighted Average Betweenness Centrality})$$

(4-6)

The "Average Clustering Coefficient" in graph theory measures the fraction of all possible triangles (three-node subgraphs) in a graph, averaged over all nodes. In protein pattern analysis, this feature is crucial as it highlights structural variations between different protein complexes, providing a quantifiable measure of the overall connectivity and efficiency of the graph. This directly correlates with the structural integrity and interaction patterns of proteins.

To improve classification accuracy based on the graph theory, one approach applied and involved multiplying the "Average Clustering Coefficient" by the sum of weighted and non-weighted features, creating composite metrics that encapsulate more structural information. This method combines the efficiency of the graph with specific structural attributes, providing a more robust indicator for classification. Another derived feature involved summing the number of nodes and the number of edges within the graph. This aggregated feature can reveal more about the graph's complexity and density than considering nodes and edges separately. By combining the number of nodes and edges, the feature captures both the size and connectivity of the graph, offering a more comprehensive view of its structure.

In practice, these enhanced features were integrated into the input dataset for neural network training. This allows the classification model to access richer, more informative data, enabling it to distinguish between different protein patterns more effectively. The inclusion of the "Average Clustering Coefficient" and its derived features has several impacts. Firstly, it improves the model's ability to identify subtle structural differences between protein patterns, leading to higher classification accuracy. Secondly, with more informative features, the neural network can converge faster during training, potentially reducing the required time and computational resources. Lastly, features like the "Average Clustering Coefficient" provide

meaningful insights into the structural properties of different protein stains, aiding in the interpretability of the classification results. All the features and their descriptions are available in Appendix A.

4.2.2.1. Neural network design

To develop a neural network tailored for training and predicting the features extracted from protein patterns, a simple yet effective architecture was carefully crafted. Given that the input dataset comprised a table of numerical values, a straightforward neural network architecture with a modest number of layers was deemed suitable for the classification task.

The architecture commenced with a feature input layer tasked with ingesting the tabular data containing 17,345 rows and $((15+6) \times 5) = 105$ columns, representing the extracted features. Subsequently, a batch normalization layer was introduced to normalize the input data, followed by an activation layer to introduce non-linearity and facilitate feature transformation. Additionally, a pivotal early fully connected layer was strategically incorporated to aid in the categorization process based on the extracted features. This layer, with a size of 10 times the number of classes, played a crucial role in enabling the network to discern patterns and make informed classification decisions.

To refine the classification process and mitigate the influence of unwanted weights, a tandem of batch normalization and activation layers was employed. These layers worked synergistically to cleanse the network of any extraneous information that could hinder accurate classification.

Furthermore, a dropout layer was subsequently introduced to prevent overfitting and enhance the network's generalization capabilities. Following this, a second fully connected layer, sized according to the number of classes, was integrated into the architecture to further refine the classification process.

Finally, the architecture culminated with a softmax layer, responsible for converting the network's raw output into probability scores corresponding to each class, and a classification layer for assigning the final class label.

Notably, this architecture diverges from conventional neural networks by incorporating an early fully connected layer, which proved instrumental in facilitating effective training and classification. Comparative analyses with other classification techniques, including Support Vector Machine (SVM), decision trees, K-Nearest Neighbor (KNN), and simple regression models, underscored the superior accuracy achieved by the designed neural network.^[184,244–246]

Figure 4-14 illustrates the architectural layout of the designed neural network, showcasing its distinctive configuration optimized for accurate classification of protein patterns based on extracted features from graph theory analysis tool.

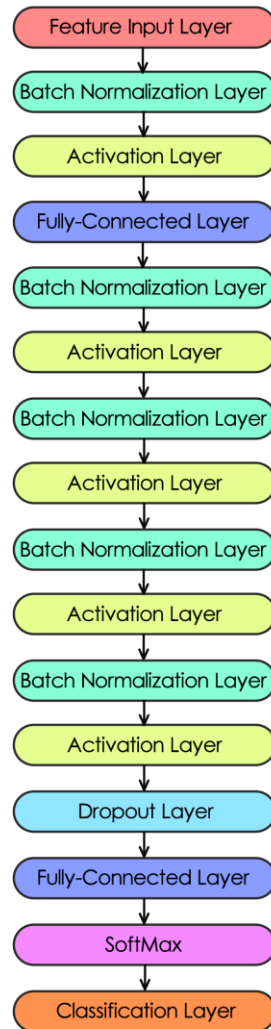


Figure 4-14: The architecture of the neural network designed for feature classification.

4.2.2.2. Classification of IgG:Protein A complexes

After developing a custom-designed neural network, the next step was to apply it to the task of classifying the extracted features derived from protein patterns. Each feature set was assigned the same labels as those used in the image classification task conducted with InceptionV3. Following the same procedural steps as image classification (refer to section 4.2.1.3), the network was trained and tested using a designated test dataset.

Figure 4-15 illustrates the resulting confusion matrix obtained from the feature classification process. Notably, the total accuracy achieved was 58.25%, which marked a

decrease compared to the accuracy attained in image classification (81.42%). This decline in accuracy can be attributed to the network's tendency to identify similarities in the strength of interactions, particularly evident in cases such as rabbit IgG:Protein A and bovine IgG:Protein A.

Further analysis revealed that rabbit IgG and bovine IgG both exhibited weak to medium strength interactions with Protein A, as determined by CD spectroscopy measurements (**Figures B-1 and B-2**). While image classification excelled in identifying the source of IgG, it struggled to discern differences in the strength of interaction between different IgG species. Conversely, feature classification enabled more nuanced monitoring of these similarities in interaction strength.

The local accuracies of each IgG species image class followed a consistent trend, with higher accuracies associated with stronger interaction strengths. Specifically, accuracies of 81.86%, 56.28%, 54.14%, 50.43%, and 44.86% were achieved for IgG from human, rabbit, bovine, goat, and HSA, respectively, in interaction with recombinant Protein A. Remarkably, the optimum binding ratio of 2:1 for human IgG:Protein A yielded the highest accuracy of 98% among other molar ratios.

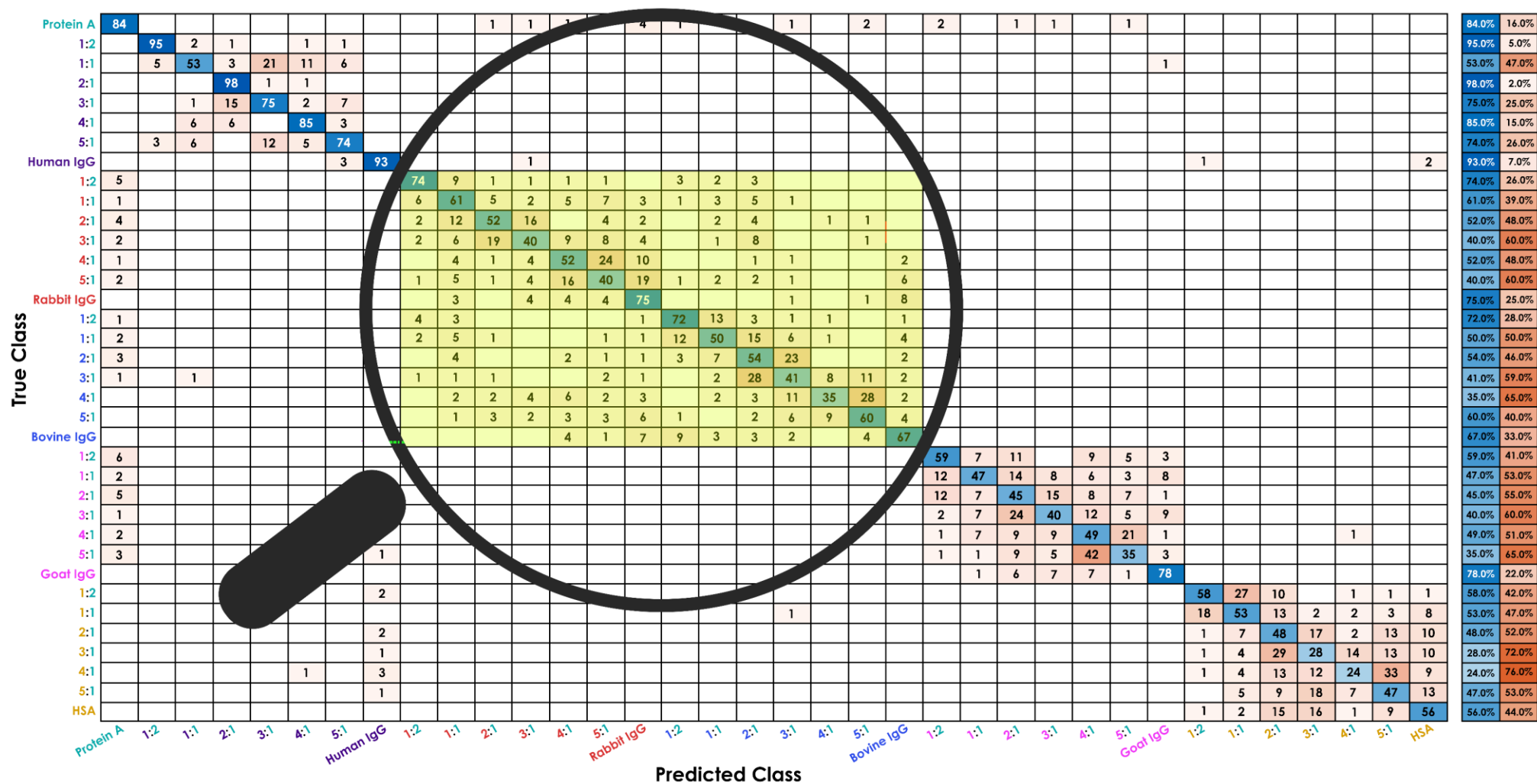


Figure 4-15: Feature classification confusion matrix. The total accuracy of this confusion chart is 58.25%.

Transparent yellow box under the magnifier in the confusion matrix highlight misclassifications stemming from similarities in interaction strength between IgG from rabbit and bovine serum and Protein A. Additionally, misclassifications were more prevalent in different molar ratios within each IgG species class, again reflecting similarities in interaction strength confirmed by CD spectroscopy analysis (**Figures B-1 to B-4**).

It is important to note that in terms of the time and computational resources needed for training and classification tasks, training the feature classification neural network took about 7 minutes, regardless of whether a GPU was used. Furthermore, the network excelled in recognizing interaction strength over molar ratios or protein sources. To leverage this capability, three main labels were chosen based on relative interaction strength: **Weak** (goat IgG:Protein A and HSA:Protein A), **Medium** (rabbit IgG:Protein A and bovine IgG:Protein A), and **Strong** (human IgG:Protein A). The single protein images were excluded from this revised input dataset. Upon re-labeling the features and retraining the network with the same parameters, a confusion matrix was obtained, as depicted in **Figure 4-16**. Impressively, a total accuracy of 99.83% was achieved in just 6 minutes of training. These results underscore the high performance of using extracted features derived from graph theory algorithms in predicting protein-protein interaction strength.

True Class	Strong	600			100.0%	
	Medium	1	1198	1	99.8%	0.2%
	Weak		3	1197	99.8%	0.2%
		Strong	Medium	Weak		
		Predicted Class				

Figure 4-16: Confusion matrix of different protein-protein interaction strengths using feature classification (re-labeled). The obtained total accuracy of prediction is 99.83%

4.2.3. Neural network optimization

In the previous feature classification section, while the training time was negligible (6 minutes) compared to the image classification (about 19 days), a key bottleneck was identified in the feature extraction process, which required significant time despite being less than that of the image classification procedure. To optimize the designed neural network, efforts focused on reducing the time required for initial feature extraction and identifying the most effective features for training.

4.2.3.1. Input dataset size reduction

Given the high accuracy achieved in the previous feature classification (99.83%), a subset comprising 10% of the input labels was randomly selected and separated for training. Consequently, instead of working with a table of 13,806 rows (excluding single protein classes) and 105 columns, a reduced dataset of 1,380 rows and 105 columns was introduced to the neural network as the input dataset. This reduction aimed to address potential training inefficiencies arising from unnecessary data and to streamline the feature extraction process. This input size reduction was also applied on test dataset.

In this revised approach, random patterns of different protein mixtures were utilized for

training, without consideration for the protein source or molar ratio, as long as the total mass concentration remained constant. This approach aimed to generalize the network's learning beyond specific protein sources or ratios, facilitating broader applicability.

The resulting confusion matrix, depicted in **Figure 4-17 A**, achieved a total accuracy of 99.33% with a remarkably reduced training time of only about 1 minute. This significant improvement in both the required time for feature extraction and the training time, while maintaining a high accuracy of prediction, underscores the effectiveness of the optimization strategy.

Since the input dataset size was reduced for feature classification, a similar approach was applied to image classification to enable an appropriate comparison between the two methods. The same procedure involved re-labeling the image classes and randomly reducing the input image dataset (both training and testing) to 10% of the original size. InceptionV3, a pre-trained CNN, was utilized for this reduced image classification task. The training process for this reduced dataset took approximately 8 hours, resulting in a total prediction accuracy of 99.00%. The obtained confusion matrix can be found in **Figure 4-17 B**.

A) Feature Classification with reduced dataset size

True Class	Strong	65			100.0%	
	Medium		126	1	99.2%	0.8%
	Weak		1	107	99.1%	0.9%
		Strong	Medium	Weak		
		Predicted Class				

B) Image Classification with reduced dataset size

True Class	Strong	60			100.0%	
	Medium	1	119		99.2%	0.8%
	Weak		2	118	98.3%	1.7%
		Strong	Medium	Weak		
		Predicted Class				

Figure 4-17: Confusion matrix of reduced-size input dataset for protein-protein interaction strengths. Randomly 10% of the input data were selected. A) Feature classification using the designed neural network with the prediction accuracy of 99.33% B) Image classification using InceptionV3, a pre-trained CNN with the prediction accuracy of 99.00%.

By implementing these changes, we can make a direct comparison between feature classification and image classification in terms of training time and accuracy. The reduction in dataset size significantly improved the efficiency of the feature classification process, achieving similar high accuracy in a fraction of the time required for image classification. This comparison highlights the potential advantages of feature extraction and classification, especially when computational resources and time are limited. This leads to the conclusion that using features extracted from graphs, obtained from graph-like images through graph theory analysis, can be successfully employed as input datasets for classification problems. Processing

a table of numerical values is not only easier than processing images through neural networks, but also allows classification to be performed on simple, non-GPU dependent computing systems, reducing the required training time.

4.2.3.2. Feature selection algorithm

Further optimization can involve selecting the most effective features that contribute to the final decision. Several feature selection algorithms are available, among which Maximum Relevance – Minimum Redundancy (MRMR) is considered highly effective.^[247–250] MRMR assigns each input feature a predictor importance score, indicating its effectiveness in the classification process. The importance score are non-negative values. In this study, features with an importance score threshold of 0.1 or higher were selected, while the others were removed before being introduced to the neural network. Consequently, 22 out of 105 features were retained. The names of the selected features (highlighted in the green transparent box) and their importance scores are shown in **Figure 4-18**.

In **Figure 4-18**, the “Average Degree” graph theory feature, derived from the input parameters using moderate gamma adjust of 1.3 and global threshold value of 100 (refer to section 3.17, **Table 3-5**), achieved the highest selection score. This indicates that this feature plays a critical role in distinguishing between different protein-protein interaction patterns. The "Average Degree" of a graph is a fundamental measure in graph theory that provides insights into the connectivity of the network. It is defined as the average number of edges (connections) per node in the graph. Mathematically, it is computed as the sum of the degrees of all the nodes in the graph, divided by the total number of nodes. The average degree reflects how well-connected or sparse the network is, and it can reveal the structural properties of the system (protein-deposited patterns). A high feature selection score suggests that this parameter carries significant discriminative power in the classification task, making it a key contributor to improving model accuracy.

Notably, three of the newly calculated features, derived through simple mathematical operations on the extracted features (refer to section 4.2.2), were included in the final selection of features. This inclusion underscores the potential of basic mathematical manipulations to enhance feature representation, demonstrating that even straightforward transformations can significantly contribute to the model's ability to distinguish between different classes. Such findings suggest that leveraging mathematical operations can enrich the dataset, thereby improving the classification performance of neural networks and offering a more nuanced understanding of the underlying data.

The training was conducted using the reduced feature set (input size reduced previously), and the prediction on the test dataset followed. The obtained confusion matrix is displayed in **Figure 4-19**, with a total prediction accuracy of 98.67% and a training time of about 1 minute. These results confirm that feature selection algorithms can enhance the performance of feature classification problems. Algorithms like MRMR are preferable due to their efficiency in memory consumption, required time, performance, and the explainability of the results.

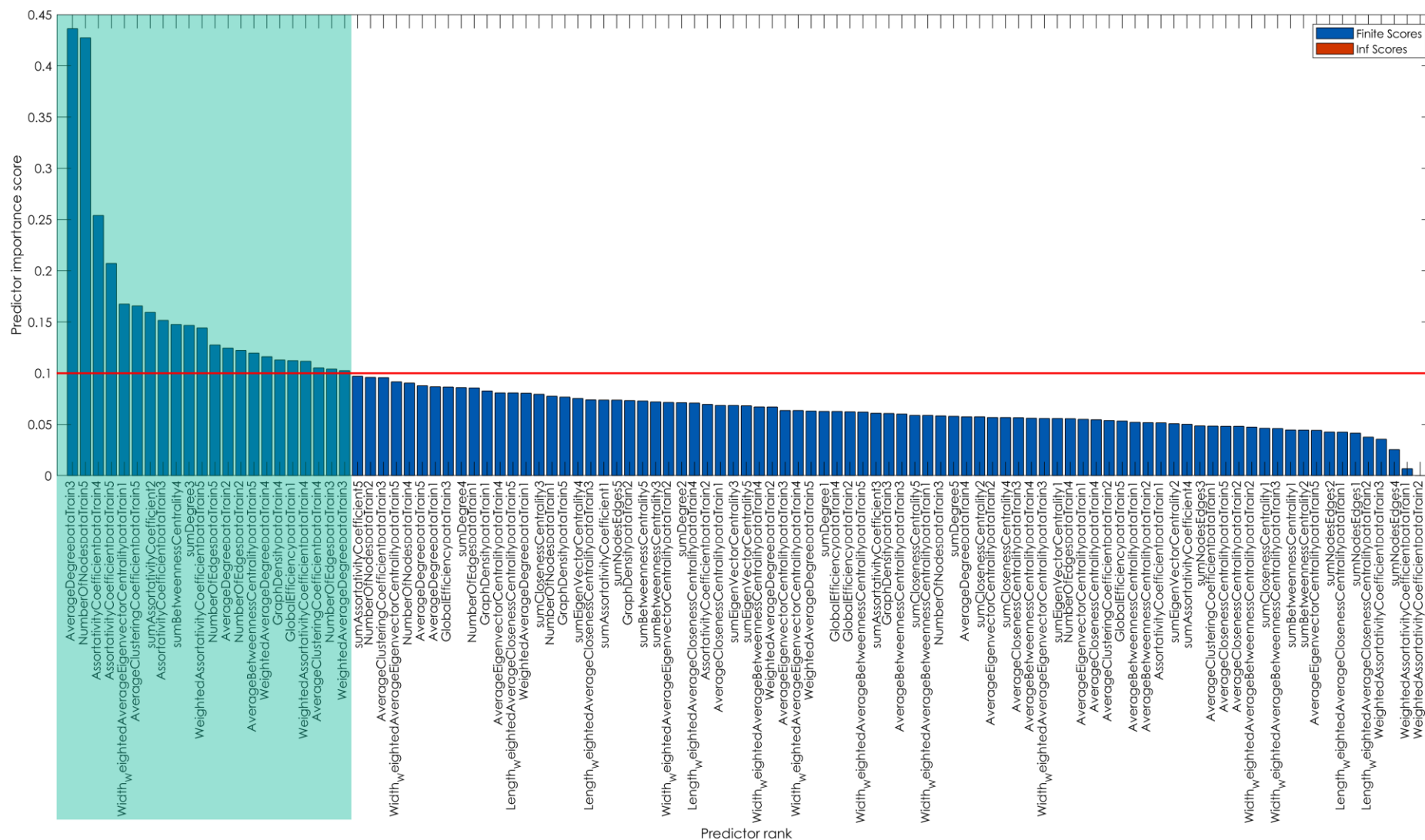


Figure 4-18: Feature selection scores. The given scores to the extracted/computed features using MRMR feature selection algorithm. The green box shows the selected features with the threshold score of 0.1 and higher.

True Class	Strong	65			100.0%	
	Medium	1	119		99.2%	0.8%
	Weak		3	112	97.4%	2.6%
		Strong	Medium	Weak		
		Predicted Class				

Figure 4-19: Confusion matrix for protein-protein interaction strength using feature classification with input size reduction and MRMR feature selection algorithm with the prediction accuracy of 98.67%.

Feature selection algorithms are crucial in machine learning and data science because they determine which input features (or variables) are most relevant for making accurate predictions while minimizing noise and redundancy. The ability to perform feature selection offers several important advantages, and the implications of using such techniques are broad. First, by reducing the number of features to those most relevant, feature selection simplifies the model, leading to faster training times and reduced computational costs. This is particularly important in large datasets, where too many features can cause models to overfit or become computationally expensive.^[251–253]

The broader implications of feature selection go beyond efficiency. From a scientific perspective, feature selection helps reveal which variables are most significant to the problem being studied. This can lead to new insights and a deeper understanding of the underlying processes. For instance, in biological or medical research, identifying crucial biomarkers from a vast array of potential features can help prioritize the most significant factors associated with a disease or condition, ultimately leading to the development of new treatments or therapies. Feature selection goes beyond merely creating more efficient models; it also involves uncovering important patterns and relationships within the data that may otherwise remain

hidden.^[247,249,251–253]

4.2.3.3. Feature classification robustness

To evaluate the performance of the trained neural network in predicting unknown samples and check its robustness, the images of human IgG:Protein G complexes (the same data that was used to check the robustness of image classification) were converted to graphs, and the necessary features were extracted. The same procedure used for test dataset preparation was followed. Given that human IgG exhibits a strong interaction with Protein G, all the images were labeled as "Strong" and introduced to the trained network to observe its classification accuracy for these protein-protein interaction classes.

The resulting confusion chart for the unknown sample prediction is presented in **Figure 4-20**. The trained neural network accurately categorized 90.07% of the human IgG:Protein G images as exhibiting a "Strong" strength of interaction. This classification is consistent with CD spectroscopy measurements, as shown in **Figure B-5**. These results highlight the neural network's effectiveness in predicting the interaction strength of previously unseen samples based on features extracted from biological graph-like images.

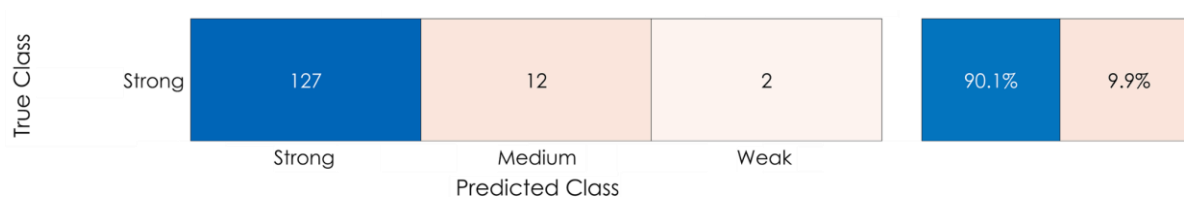


Figure 4-20: Confusion matrix for feature classification of re-labeling features for human IgG:Protein G class as unknown sample (not-trained)

Consequently, feature classification and the resulting trained network offer a simple, reliable, and efficient method for the classification and screening of biological graph-like images. This approach not only demonstrates high prediction accuracy but also provides a

practical solution for the analysis of unknown samples. The ability to accurately predict interaction strengths using graph theory-derived features underscores the potential of this method for broader applications in protein interaction studies and other biological classification tasks. This methodology can significantly streamline the process of identifying and categorizing protein interactions, thereby enhancing our understanding and facilitating further research in the field of proteomics and related areas.

4.2.4. Image classification vs. feature classification

To facilitate a comprehensive comparison between image classification and feature classification, **Table 4-2** summarizes the key parameters affecting the final performance of each classification method. This table outlines the required training time and the computational costs associated with each method, both in their original schemes and after the optimizations described in previous sections. Notably, feature classification demonstrates a significant reduction in training time and computational cost compared to image classification, maintaining a constant ratio of reduction even after optimization. This reduction is particularly important for classification purposes, as feature classification achieves more than two times faster training in total (dataset preparation + training) without relying on GPU-dependent computing systems.

Table 4-2: Comparison of Image Classification and Feature Classification

Parameter	Image Classification	Feature Classification
Input dataset preparation time (Original)	2400 minutes	11563 minutes (~8 days)
Input dataset preparation time (Optimized)	240 minutes	289 minutes
Training time (Original)	26951 minutes (~18 days)	7 minutes
Training time (Optimized)	480 minutes (8 hours)	1 minute
Total required time (Original)	29351 minutes (~21 days)	11570 minutes (~8 days)
Total required time (Optimized)	720 minutes (12 hours)	290 minutes (~5 hours)
Accuracy (Original)	81.42%	58.25%
Accuracy (Optimized)	99.00%	98.67%
Computational Cost	High (GPU dependent)	Low (Non-GPU dependent)
Training Data Size Reduction	Yes (10% of original size)	Yes (10% of original size)
Performance Consistency	High with GPU	High without GPU

- Training Time

Original Scheme: The original training time for image classification is significantly higher than feature classification. Image classification required approximately 29351 minutes (about 20 days) compared to 11570 minutes (about 8 days) for feature classification.

Optimized Scheme: After optimization, the training time for image classification is reduced to about 480 minutes (8 hours), while feature classification training time is reduced to just 1 minute. The total time needed for image classification is 12 hours, whereas feature classification only takes 5 hours.

- Accuracy:

Original Scheme: The accuracy of the original image classification method (81.42%) is

higher than that of the original feature classification method (58.25%).

Optimized Scheme: With optimization, both methods achieve high accuracy levels, with image classification reaching 99.00% and feature classification achieving 98.67%.

- Computational Cost: Image classification is computationally intensive and heavily dependent on GPU resources, whereas feature classification significantly reduces computational costs and does not depend on GPUs, making it more accessible and feasible for various applications.
- Training Data Size Reduction: Both methods benefit from reducing the training data size to 10% of the original dataset, maintaining a high accuracy with a smaller dataset, which highlights the efficiency of the optimization process.

Performance Consistency: Feature classification maintains high performance without relying on GPU resources, ensuring consistent results across various computing environments. Image classification, on the other hand, requires GPU support to achieve optimal performance.

5. Conclusion and Outlook

Image classification using deep learning approaches is poised to find a wide range of applications in different fields of biology and medicine. Images of stain patterns of biomacromolecules and their mixtures contain previously untapped levels of structural and functional information. In this study, we have demonstrated strong correlations between structural differences in protein-protein complexes and the resulting characteristic changes in their respective deposition patterns. By employing a machine learning algorithm, we stratified protein-protein complexes based on the interactions of immunoglobulin G (IgG) from various species with the superantigen recombinant Protein A with the total classification accuracy of 81.42% over different molar ratios. The CD measurements also confirmed the results for protein-protein interaction analysis based on their secondary structure alteration in interaction for the same protein solutions. The pre-trained InceptionV3 model not only successfully distinguished different IgG:Protein A combinations but also predicted their binding propensities. Remarkably, for unknown samples, the pre-trained InceptionV3 accurately predicted their relative binding strengths with 94% in accuracy, underscoring the robustness of this approach and its potential for effective translation to unknown protein affinities. Our findings suggest that this methodology could contribute significantly to the development of precise, straightforward, and unbiased methods for predicting protein-protein interactions.

Although, the used pretrained CNN (InceptionV3) demonstrated high accuracy in image classification tasks, one of the main concerns of employing CNNs is the requirement of GPU sources. However, the required time for training is also considerable (about 20 days). To address these issues, innovative optimization strategies will be implemented to enhance efficiency and reduce time and computational costs.

Feature classification emerges as a highly efficient and practical alternative to image classification, particularly when computational resources and training time are critical constraints. The significant reduction in training time and computational cost, combined with high prediction accuracy, underscores the potential of feature classification for broader applications in protein interaction analysis and other biological classification tasks. By using features extracted from graphs, which were derived from protein pattern images using graph theory analysis, we were able to maintain high prediction accuracy while dramatically reducing the required computational resources. This method, implemented using the StructuralGT python package developed at the University of Michigan, transforms images into graphs and extracts meaningful features for classification. One key feature, the "Average Clustering Coefficient," represents the efficiency of clustering within the graph and helps differentiate structural differences between protein patterns.

The graph-based feature classification method using the designed neural network demonstrated comparable accuracy to traditional CNNs while significantly reducing training time and computational cost. In the optimized scheme for both methods, image classification achieved an accuracy of 99%, while feature classification yielded a prediction accuracy of 98.67% with more than two times reduction in the required time for classification task, without dependency on GPU sources. This highlights the effectiveness of using graph theory for image analysis, particularly in biological and protein interaction studies. The StructuralGT package introduces an innovative method for image processing by applying graph theory to extract structural features, optimizing neural network training, and offering a flexible, efficient alternative to conventional image-based analysis techniques.

In conclusion, the application of graph theory in feature classification presents a promising avenue for future research and practical applications. This method not only optimizes the computational efficiency of neural network training but also extends the utility

of image classification techniques to a broader range of scientific and engineering domains. However, several important future directions remain to further build upon this research.

A key next step is to expand the dataset, incorporating a broader range of protein patterns or biomolecular interactions to improve model robustness and generalization. A larger, high-quality dataset will enable the model to handle more complex and diverse interactions, strengthening its applicability to real-world scenarios. In parallel, optimizing the neural network architecture is crucial. While the current model shows promising results, exploring more advanced architectures—such as hybrid models or transformers—could yield improvements in both accuracy and efficiency, especially when dealing with larger datasets.

The use of graph theory for feature classification provides a unique and effective approach, but further refinement of the feature extraction process is necessary. Future research could explore additional graph metrics and mathematical operations beyond summation and multiplication to derive more insightful features. Optimizing the computational aspects of the model, and increasing its suitability for real-time applications or resource-constrained environments should also be a priority.

Finally, ensuring model interpretability is essential, particularly for scientific and clinical applications. Future work could focus on further optimization and application of this technique, potentially enhancing its utility in various scientific fields such as proteomics, bioinformatics, and systems biology. The integration of graph theory with deep learning methods represents a powerful toolset for advancing our understanding and capability in protein interaction analysis and beyond.

List of Tables

Table 2-1: IgG binding of SpA.....	18
Table 3-1: Used chemicals and materials.....	50
Table 3-2: List of used instruments	51
Table 3-3: List of used software.....	52
Table 3-4: Features extracted from individual image using graph theory analysis. The red-colored features represent the non-meaningful features in this study.	65
Table 3-5: Selected input parameters for graph theory analysis of given images	66
Table 4-1: Characteristics of the applied recombinant Protein A and Protein G. Adapted from ^[232]	87
Table 4-2: Comparison of Image Classification and Feature Classification	109
Table A-1: GT parameters description ^[30,32,233]	119

List of Figures

- Figure 2-1: Linear model of structural variation of the different four subtypes of human IgG.** An IgG molecule consists of two heavy and two light chains connected by disulfide bonds. The light chain has variable (VL) and constant (CL) regions, while the heavy chain includes one variable (VH) and three constant (CH1, CH2, CH3) regions. The Fc region mediates immune system functions, while the Fab region, containing the VL and VH domains, is responsible for antigen recognition and binding. Human IgG has four subclasses—IgG1, IgG2, IgG3, and IgG4—which differ in the size of their hinge regions and the number and arrangement of the interchain disulfide bonds linking their heavy chains.8
- Figure 2-2: Schematic diagrams of SpA, and SpG domain structures.** A) Left panel: The structure of individual SpA domains, which include the S (sorting peptide), Domains E-D-A-B-C, and Regions X and M. Right panel: The SpA Domain C (PDB code: 1BDD) shows that each immunoglobulin-binding domain in SpA is made up of three α -helices. (B) Left panel: The structure of individual SpG domains, which consist of the S (sorting peptide), Region E, Albumin Binding Domains A1-A2-A3, immunoglobulin-binding domains B1-B2/C1-C2-C3, and Region W. Right panel: The SpG Domain B1 (PDB code: 3GB1) reveals that each immunoglobulin-binding domain in SpG comprises one α -helix and four anti-parallel β -strands.^[12]12
- Figure 2-3: Crystal structure of IgG in interaction with A) Protein A, and B) Protein G.** Blue ribbons present IgG (PDB code: 1IGT) while orange and red ribbons indicate SpA (PDB code: 1BDD) and SpG (PDB code: 3GB1) ligands, respectively. Two heavy chains in the IgG Fc fragment can be in complex with SpA and SpG.19
- Figure 2-4: Key aspects of droplet wetting and evaporation.** Physical mechanisms occur as a single-component droplet evaporates on a solid substrate. Critical factors such as marangoni flow within the droplet, evaporation dynamics at the droplet edge, vapor behavior in the surrounding air, surface tension at the liquid-air interface, and wetting phenomena on the solid substrate, characterized by contact angle and surface energy balances (σ_{LG} : liquid-gas tension, σ_{SL} : solid-liquid tension, σ_{SG} : solid-gas tension). Each of these mechanisms plays a role in defining the droplet's behavior during evaporation.30
- Figure 2-5: High-level AI diagram.** This diagram illustrates the hierarchical relationship between Artificial Intelligence, its subset Machine Learning, and the further specialization of Deep Learning, highlighting their interconnected roles in developing intelligent systems and solving complex problems.33
- Figure 2-6: Inception module.** Its parallel convolutional paths with different filter sizes, enabling the network to extract multi-scale features efficiently for improved image classification performance.40
- Figure 2-7: Basic architecture of CNN.** Illustration of the fundamental structure of a CNN, highlighting its key components: convolutional layers for feature extraction, pooling layers for dimensionality reduction, and fully connected layers for final classification. The flow of data through these layers demonstrates how the network processes and learns hierarchical features from input images, enabling effective image recognition and classification tasks.42

- Figure 3-1: The scheme of Protein A HP SpinTrap column.** This column utilizes Protein A Sepharose for the purification of antibodies. The column is designed to facilitate the selective binding of immunoglobulin G (IgG) antibodies through the Protein A ligands, allowing for efficient separation from other proteins in a sample. This figure illustrates different parts of the column.56
- Figure 3-2: Chemical vapor deposition (CVD) polymerization.** A) The CVD reactor generally comprises four primary components: the gas inlet, sublimation zone, pyrolysis zone, and deposition chamber. B) The polymerization process converting [2.2]paracyclophane into PPX under the specified conditions within a CVD system. Adapted from ^[177]57
- Figure 3-3: Automated droplet dispenser.** It designed for precise and controlled dispensing of liquid droplets onto various substrates. The system features an array of dispensing nozzles, a programmable control interface, and a platform for substrate positioning.59
- Figure 3-4: The architecture of pre-trained InceptionV3.** The InceptionV3 architecture is composed of multiple Inception modules, which allow for efficient extraction of multi-scale features by applying convolutional filters of varying sizes. Key components of the network include convolutional layers, max-pooling layers, and fully connected layers, culminating in a softmax classifier for output prediction. Adapted from ^[232]62
- Figure 4-1: Relative average binding percentages of IgG from different species (human, rabbit, bovine, goat) to Protein A sepharose at varying concentrations (0.5 to 5 μ M) in sodium phosphate buffer, pH=8.1.** The graph demonstrates high binding affinity for human and rabbit IgG, moderate affinity for bovine IgG, and low affinity for goat IgG. HSA showed no significant binding. These results suggest that binding percentages are primarily influenced by the presence of IgG subtypes capable of binding to Protein A, rather than the total concentration of IgG applied.70
- Figure 4-2: CD spectroscopy measurements results for human IgG:Protein A in different molar ratios with the constant total mass concentration.** The deviations between summation and interaction spectra, particularly observed at 217 nm, indicate alterations in the secondary structure of the protein mixture via interaction.73
- Figure 4-3: Illustration of relative binding affinity of IgG from different species to Protein A based on CD spectroscopy measurement.**74
- Figure 4-4: Formation of protein stains using controlled droplet deposition and drying process.** A) Cleaned glass substrates were coated with poly(*p*-xylylene) via CVD polymerization in order to obtain reliable hydrophobic surface conditions to ensure a reproducible surface interaction. An automated pipetting system was used for dispensing several protein sample droplets (2 μ l) containing different molar ratios of IgG from different species and Protein A. B) Dispensed droplets were dried under controlled environmental conditions (T = 25 °C, relative humidity = 40 %). C) PLM imaging was used to collect all the deposited proteins' patterns under the same conditions to prepare sufficient images for each category for image classification with CNN implementation, IgG from (I) bovine, (II) rabbit, (III) goat and (IV) human serum interacting with Protein A. D) SEM image analysis of human IgG stains. E) ToF-SIMS analysis of the deposition pattern of a complex of Goat IgG with Protein A: RGB overlay image of the distribution map of PO₂⁻ ions (green) and CNO⁻ ions (red). Adapted from ^[232]76
- Figure 4-5: The deposited patterns of IgG from different species and Protein A with various molar ratios.** The ratio 1:0 is regarding the single IgG/HSA pattern. Adapted from ^[232]77

- Figure 4-6: Confusion matrix of IgG from different species.** The number of true positives (correctly classified) classes are placed on the diagonal with a bluish color, while other cells represent the misclassifications. The overall accuracy of the prediction is 98.75%, reflecting the proportion of correctly predicted instances out of the total instances.79
- Figure 4-7: Confusion matrix of human IgG:Protein A image classes across different molar ratios, evaluated using the original and transformed test datasets.** A) Original dataset with a prediction accuracy of 92.5%. B) Horizontally flipped dataset with 93% accuracy. C) Vertically flipped dataset achieving 93.4% accuracy. D) Dataset flipped both horizontally and vertically with 93% accuracy.81
- Figure 4-8: Confusion matrix of different protein-protein interactions.** A) human IgG, B) rabbit IgG, C) bovine IgG, D) goat IgG, E) HSA. The confusion chart was obtained for a test set of 36 categories divided into smaller charts with t-SNE plot analysis for each protein complex. Adapted from ^[232]83
- Figure 4-9: The 36 outlines confusion matrix of testing image dataset obtained from trained InceptionV3.** The total accuracy of this confusion chart is 81.42%. Adapted from ^[232]86
- Figure 4-10: Performance of a pre-trained network using human IgG:Protein G patterns.** Confusion chart of human IgG:Protein G image classification as an unknown sample (not trained). Adapted from ^[232]88
- Figure 4-11: Examples of obtained patterns of Protein A, Protein G, human IgG, human IgG:Protein G with a molar ratio of 2:1 in comparison to Protein A interaction with human IgG pattern with the same molar ratio.**.....88
- Figure 4-12: Grad-Cam image analysis.** The heatmap overlay shows areas of importance that contributed to the model's decision, offering insights into how the CNN interprets IgG:Protein A complex patterns.....89
- Figure 4-13: Image-to-graph conversion using graph theory analysis.** Applying five different gamma adjust and global threshold values as input parameters to capture the most intricate details of the given images.91
- Figure 4-14: The architecture of the neural network designed for feature classification.**96
- Figure 4-15: Feature classification confusion matrix. The total accuracy of this confusion chart is 58.25%.**98
- Figure 4-16: Confusion matrix of different protein-protein interaction strengths using feature classification (re-labeled).** The obtained total accuracy of prediction is 99.83% 100
- Figure 4-17: Confusion matrix of reduced-size input dataset for protein-protein interaction strengths.** Randomly 10% of the input data were selected. A) Feature classification using the designed neural network with the prediction accuracy of 99.33% B) Image classification using InceptionV3, a pre-trained CNN with the prediction accuracy of 99.00%. 102
- Figure 4-18: Feature selection scores.** The given scores to the extracted/computed features using MRMR feature selection algorithm. The green box shows the selected features with the threshold score of 0.1 and higher. 105

Figure 4-19: Confusion matrix for protein-protein interaction strength using feature classification with input size reduction and MRMR feature selection algorithm with the prediction accuracy of 98.67%.	106
Figure 4-20: Confusion matrix for feature classification of re-labeling features for human IgG:Protein G class as unknown sample (not-trained).....	107
Figure B-1: CD spectroscopy measurements results for rabbit IgG:Protein A in different molar ratios with the constant total mass concentration	122
Figure B-2: CD spectroscopy measurements results for bovine IgG:Protein A in different molar ratios with the constant total mass concentration	123
Figure B-3: CD spectroscopy measurements results for goat IgG:Protein A in different molar ratios with the constant total mass concentration	124
Figure B-4: CD spectroscopy measurements results for HSA:Protein A in different molar ratios with the constant total mass concentration	125
Figure B-5: CD spectroscopy measurements results for human IgG:Protein G in different molar ratios with the constant total mass concentration.	126

Appendix A

In the **Table A-1**, Graph Theory parameters and their descriptions are presented.

Table A-1: GT parameters description ^[30,32,233]

Parameter	Formula ($n = \#$ of nodes, $e = \#$ of edges, $i =$ pertaining to node i)	Description
Degree	$k_i = e_i$ $\bar{k} = \frac{\sum k_i}{n}$	The degree of a node is the number of edges connected to that particular node.
Graph Density	$\rho = \frac{2e}{[n(n-1)]}$	The density of a graph is the fraction of edges that exist compared to all possible edges in a complete graph.
Network Diameter	d	The diameter is the maximum edges that will ever need to be traversed to get anywhere else in the graph. Also referred to as the maximum eccentricity, or the longest-shortest path of the graph.
Global Efficiency	$E_{glob} = \frac{E(G)}{(E(G^{ideal}))}$ $E(G) = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{L(i,j)}$ <p>$L(i,j)$ is the shortest path between nodes i and j. G^{ideal} - the graph with all possible edges present.</p>	The efficiency is the reciprocal distance between a pair of nodes. The average efficiency of $G(n,e)$, is the average across all pairs of nodes in the graph. The global efficiency is the average efficiency compared to the efficiency of a fully connected graph, though for unweighted graphs, the average efficiency is equal to the global efficiency.
Wiener Index	$WI = \sum_{i \neq j} d(i,j)$	The Wiener index is the sum of the shortest distance between all pairs of nodes in the graph.
Average Clustering Coefficient	$\delta_i = \frac{2T_i}{[k_i(k_i-1)]}$ $\Delta = \frac{\sum \delta_i}{n}$	The clustering coefficient is the fraction of neighbors of a node that are directly connected to each other as well (forming a triangle). T_i is the number of connected triples (visually triangles) on node i .
Average Nodal Connectivity	$\bar{\kappa} = \frac{2 \sum_{i \neq j} M(i,j)}{n(n-1)}$ <p>$M(i,j)$ is the minimum number of edges that need to be removed to disconnect nodes i and j</p>	The nodal connectivity, $M(i,j)$ is the minimum number of edges that would need to be removed to disconnect nodes i and j . The maximum value is the lower value between k_i and k_j , since disconnecting a node from all of its neighbors will necessarily disconnect it from any other node. The average nodal connectivity is the connectivity value averaged over all n choose 2 pairs of nodes.
Assortativity Coefficient	$r = \frac{\sum_{ij} ij(e_{ij} - q_i q_j)}{\sigma_q^2}$ <p>j, k are the excess degree of the vertices</p> $q_j = \frac{(j+1)p_{j+1}}{\bar{\kappa}} = \sum_i e_{ij}$	The assortativity coefficient, r , measures similarity of connections by node degree. This value approaches 1 if nodes with the same degree are directly connected to each other, and approaches -1 if nodes are all connected to nodes with different degree. A value near 0 indicates random orientation.

Average Betweenness Centrality	$\overline{C_B} = \frac{1}{n} \sum C_B(i)$ $C_B(i) = \sum_{u,v} \frac{\sigma(u, v i)}{\sigma(u, v)}$	The betweenness centrality of node i is a measure of how frequently the shortest path between other nodes u and v pass through node i . $\sigma(u, v)$ represents the number of shortest paths that exist between nodes u and v , with the term $\sigma(u, v i)$ representing the number of those paths that pass through node i .
Average Closeness Centrality	$\overline{C_C} = \frac{1}{n} \sum C_C(i)$ $C_C(i) = \frac{n-1}{\sum_{j=1}^{n-1} L(i, j)}$	The closeness centrality of node i is the reciprocal of the average shortest distance from node i to all other nodes.
Eigenvector Centrality	$Ax = \lambda x$	The eigenvector centrality of node i is the i -th element of vector x that solves the eigenvector equation, where A is the adjacency matrix of the graph. There exists a solution such that all elements of x are positive if λ is the largest eigenvalue of A . This is a measure of influence that node i has on the network.
Width-Weighted Average Betweenness Centrality	$\overline{C_B} = \frac{1}{n} \sum C_B(i)$ $C_B(i) = \sum_{u,v} \frac{\sigma(u, v i)}{\sigma(u, v)}$ <p><i>Each edge is weighted based on the width of the fiber (taken as pixel width along perpendicular bisector) in order to determine shortest path between nodes u and v. Therefore thinner fibers are considered shorter.</i></p>	Weighting the betweenness centrality no longer treats all edges as equal. Stress and percolation will be limited by the thinnest fibers in a network, which is considered here.
Length-Weighted Average Closeness Centrality	$\overline{C_C} = \frac{1}{n} \sum C_C(i)$ $C_C(i) = \frac{n-1}{\sum_{j=1}^{n-1} L(i, j)}$ <p><i>Each edge is weighted based on the length of the fiber (taken as pixel length of the edge) in order to determine the shortest path between nodes i and j.</i></p>	Weighting the closeness centrality no longer treats all edges as equal. Closeness is now determined by measured lengths instead of number of edges separating two nodes.

Analysis of weighted graphs should be only considered with images on the same physical scale and instrument magnification. In this case, an additional attribute, i.e. *weight*, is associated with each edge or node. If the user utilizes edge weights, the weighted and unweighted GT parameters are displayed separately.

Appendix B

In the following tables (**Table B-1** to **B-4**), the CD spectroscopy measurements of IgG from rabbit, bovine, and goat, as well as HSA, in a mixture with recombinant Protein A in various molar ratios are presented, respectively. For these experiments, the CD spectra of each individual protein were recorded and combined to generate the spectra corresponding to their mixtures at various molar ratios (shown as blue spectra). These aggregated spectra were then compared to the CD spectra of the actual protein mixture solution (represented as green spectra). Notably, in cases of strong protein interactions, differences emerged between the mixture's CD spectra and the aggregated spectra of the individual proteins, which were attributed to changes in their secondary structure due to protein-protein interactions.

Table B-5 presents the CD spectroscopy of the human IgG:Protein G complexes, which is known to exhibit strong interactions.

Rabbit IgG:Protein A

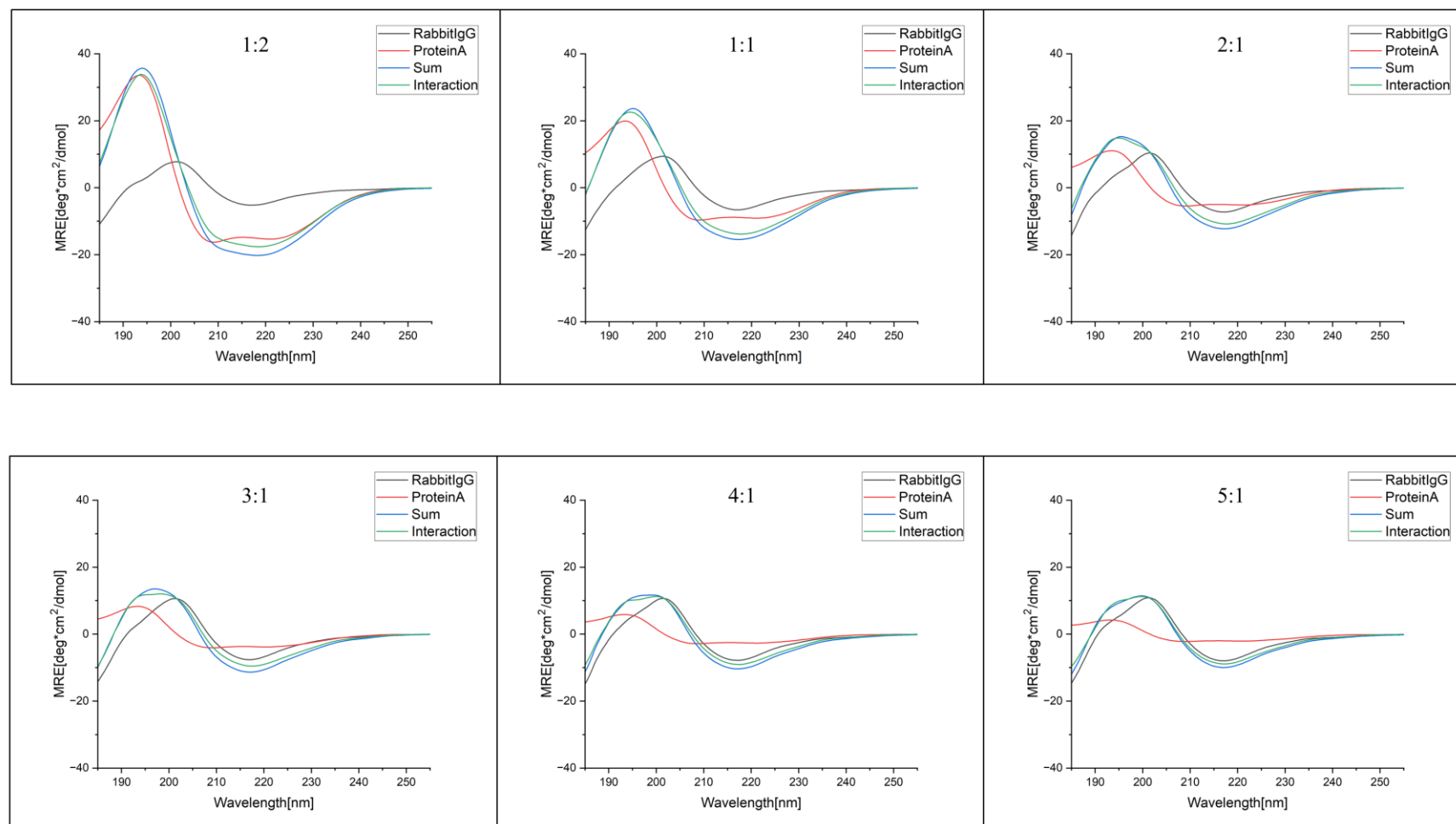


Figure B-1: CD spectroscopy measurements results for rabbit IgG:Protein A in different molar ratios with the constant total mass concentration

Bovine IgG:Protein A

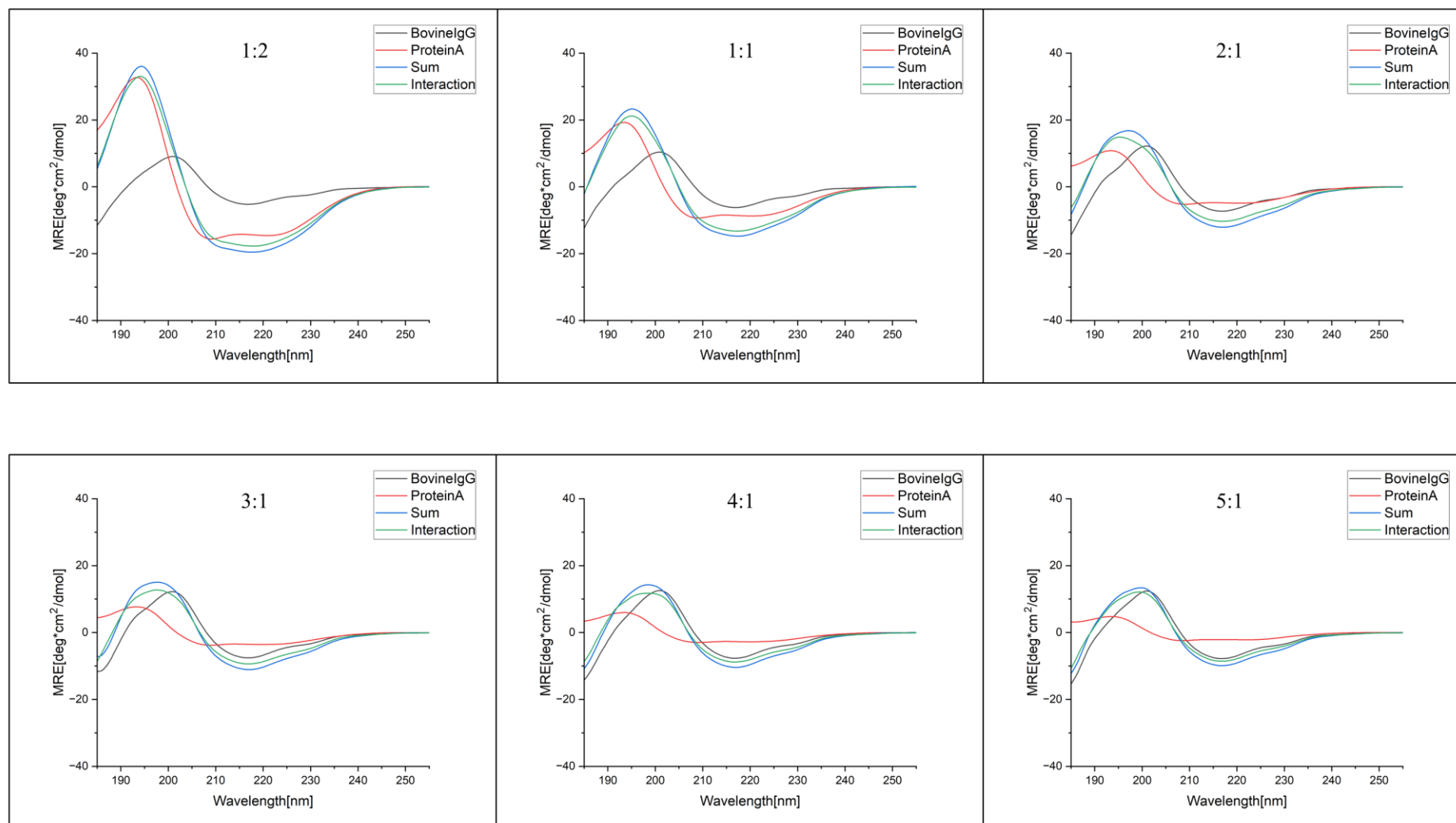


Figure B-2: CD spectroscopy measurements results for bovine IgG:Protein A in different molar ratios with the constant total mass concentration

Goat IgG:Protein A

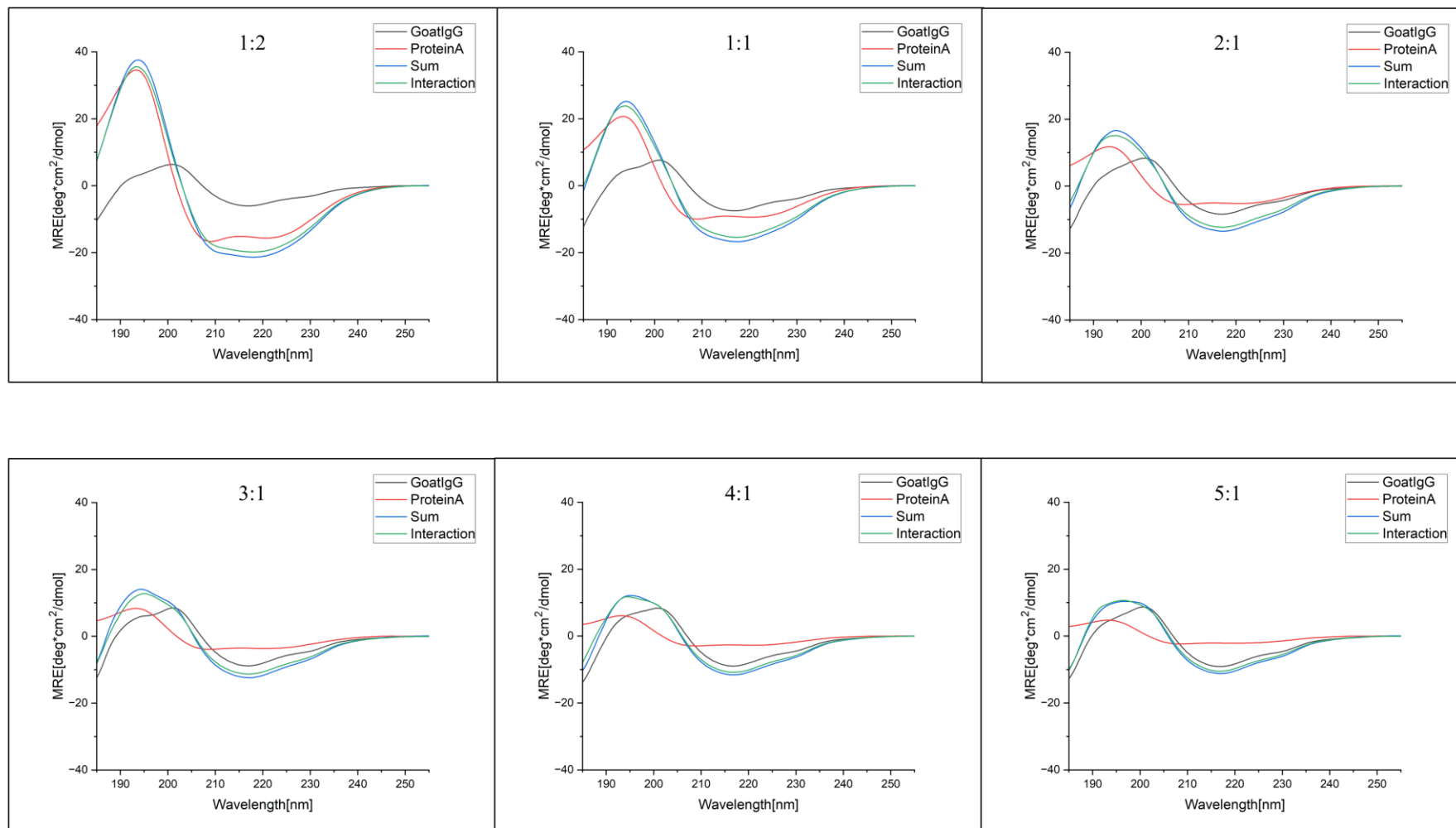


Figure B-3: CD spectroscopy measurements results for goat IgG:Protein A in different molar ratios with the constant total mass concentration

HSA:Protein A

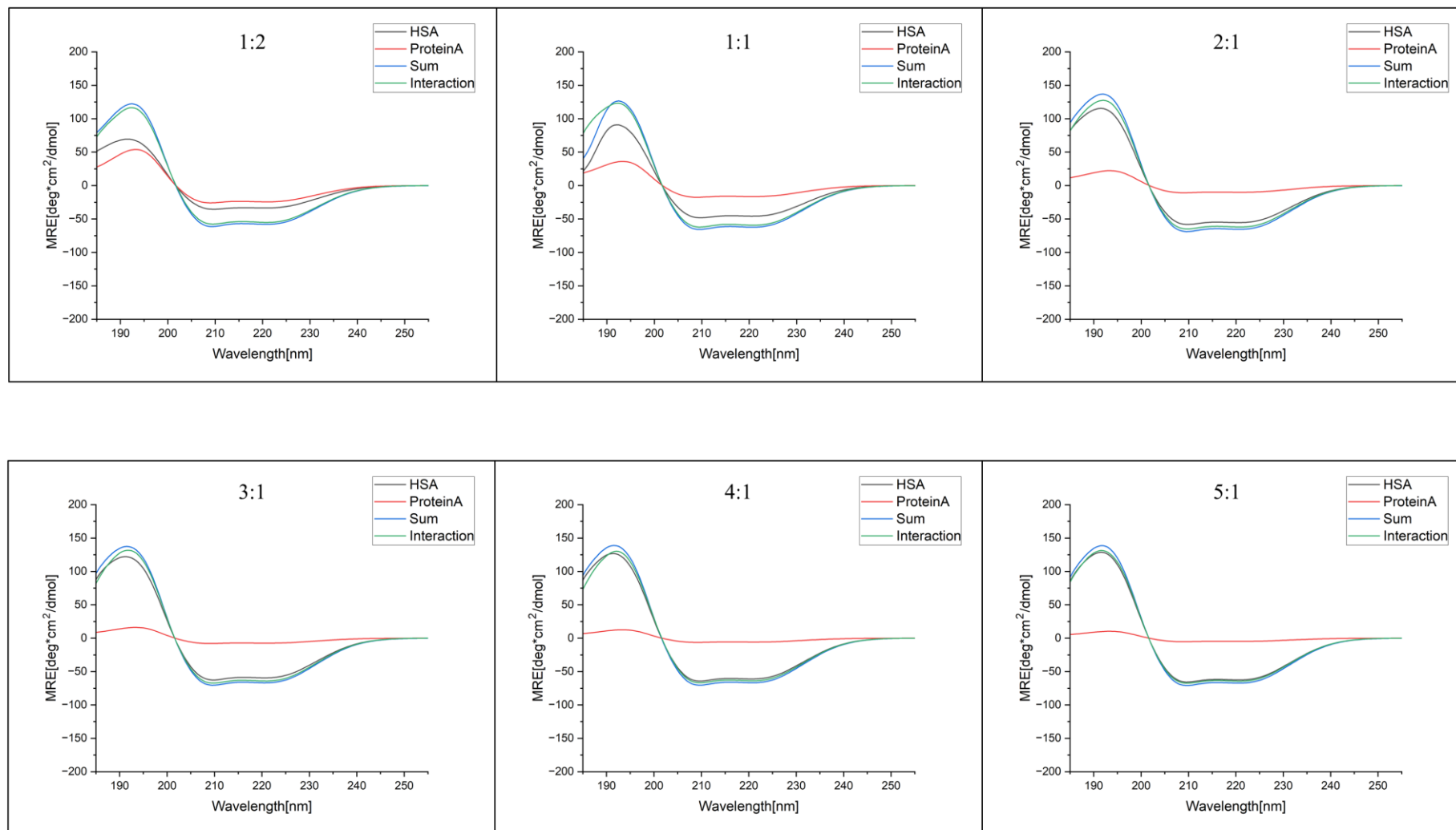


Figure B-4: CD spectroscopy measurements results for HSA:Protein A in different molar ratios with the constant total mass concentration

Human IgG:Protein G

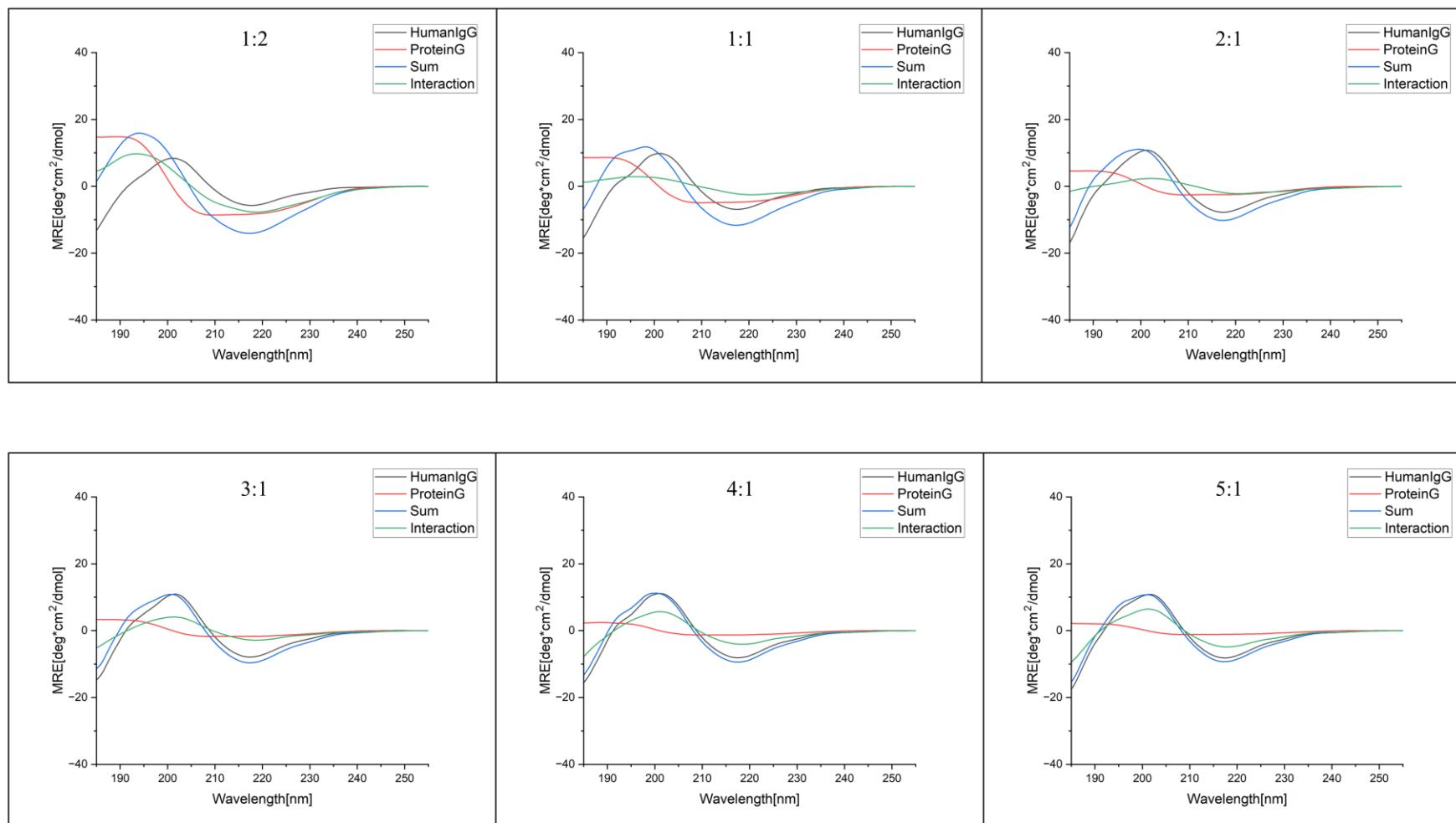


Figure B-5: CD spectroscopy measurements results for human IgG:Protein G in different molar ratios with the constant total mass concentration.

References

- [1] O. Keskin, A. Gursay, B. Ma, R. Nussinov, *Chem Rev* 2008, *108*, 1225.
 - [2] J. A. Miernyk, J. J. Thelen, *The Plant Journal* 2008, *53*, 597.
 - [3] T. Berggård, S. Linse, P. James, *Proteomics* 2007, *7*, 2833.
 - [4] V. S. Rao, K. Srinivas, G. N. Sujini, G. N. Kumar, *Int J Proteomics* 2014, *2014*.
 - [5] F. A. Aprile, P. Sormanni, M. Podpolny, S. Chhangur, L.-M. Needham, F. S. Ruggeri, M. Perni, R. Limbocker, G. T. Heller, T. Sneideris, *Proceedings of the National Academy of Sciences* 2020, *117*, 13509.
 - [6] C. Haußner, J. Lach, J. Eichler, *Curr Opin Chem Biol* 2017, *40*, 72.
 - [7] A. C. A. Roque, C. S. O. Silva, M. Â. Taipa, *J Chromatogr A* 2007, *1160*, 44.
 - [8] M. Caffrey, *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 2001, *1536*, 116.
 - [9] S. W. de Taeye, T. Rispens, G. Vidarsson, *Antibodies* 2019, *8*, 30.
 - [10] X. Chen, O. Schneewind, D. Missiakas, *Proceedings of the National Academy of Sciences* 2022, *119*, e2114478119.
 - [11] C. Aybay, *Immunol Lett* 2003, *85*, 231.
 - [12] A. M. Deacy, S. K.-E. Gan, J. P. Derrick, *Front Immunol* 2021, *12*, 731845.
 - [13] M. Eliasson, R. Andersson, A. Olsson, H. Wigzell, M. Uhlén, *J Immunol* 1989, *142*, 575.
 - [14] A. L. Anderson, R. Sporici, J. Lambris, D. LaRosa, A. I. Levinson, *Infect Immun* 2006, *74*, 1196.
 - [15] P. G. Fox, F. Schiavetti, R. Rappuoli, R. M. McLoughlin, F. Bagnoli, *mBio* 2021, *12*, 10.
 - [16] S. Hassoun, F. Jefferson, X. Shi, B. Stucky, J. Wang, E. Rosa, *Integr Comp Biol* 2021, *61*, 2267.
 - [17] J. M. Vaz, S. Balaji, *Mol Divers* 2021, *25*, 1569.
 - [18] A. Jeihanipour, J. Lahann, *Advanced Materials* 2022, *34*, 2110404.
 - [19] R. D. Deegan, O. Bakajin, T. F. Dupont, G. Huber, S. R. Nagel, T. A. Witten, *Nature* 1997, *389*, 827.
 - [20] R. G. Larson, *Angewandte Chemie International Edition* 2012, *51*, 2546.
 - [21] D. Brutin, V. Starov, *Chem Soc Rev* 2018, *47*, 558.
-

-
- [22] R. G. Larson, *AIChE Journal* 2014, 60, 1538.
- [23] A. Pal, A. Gope, G. Iannacchione, *Biomolecules* 2021, 11, 231.
- [24] R. G. G. Larson, M. A. López, D. W. Lim, J. Lahann, *MRS Online Proceedings Library (OPL)* 2010, 1273, 1273.
- [25] D. Brutin, *Colloids Surf A Physicochem Eng Asp* 2013, 429, 112.
- [26] H. Hu, R. G. Larson, *J Phys Chem B* 2006, 110, 7090.
- [27] S. Hober, K. Nord, M. Linhult, *Journal of Chromatography B* 2007, 848, 40.
- [28] E. V Sidorin, T. F. Solov'Eva, *Biochemistry (Moscow)* 2011, 76, 295.
- [29] Z. Zhang, P. Cui, W. Zhu, *IEEE Trans Knowl Data Eng* 2020, 34, 249.
- [30] D. A. Vecchio, S. H. Mahler, M. D. Hammig, N. A. Kotov, *ACS Nano* 2021, 15, 12847.
- [31] O. Sporns, *Neuroscience databases: A practical guide* 2003, 171.
- [32] A. Kadar, W. Wu, A. Emre, S. Glotzer, N. Kotov, *Bulletin of the American Physical Society* 2024.
- [33] C. M. Dobson, in *Semin Cell Dev Biol*, Elsevier, 2004, pp. 3–16.
- [34] W. E. Stites, *Chem Rev* 1997, 97, 1233.
- [35] E. M. Phizicky, S. Fields, *Protein-Protein Interactions: Methods for Detection and Analysis*, 1995.
- [36] I. M. A. Nooren, J. M. Thornton, *J Mol Biol* 2003, 325, 991.
- [37] S. Marchesseau, J. C. Mani, P. Martineau, F. Roquet, J. L. Cuq, M. Pugniere, *J Dairy Sci* 2002, 85, 2711.
- [38] Z. Nikolovska-Coleska, *Protein-Protein Interactions: Methods and Applications* 2015, 109.
- [39] L. L. Blazer, R. R. Neubig, *Neuropsychopharmacology* 2009, 34, 126.
- [40] J. A. Wells, C. L. McClendon, *Nature* 2007, 450, 1001.
- [41] W. Cai, H. Hong, *Protein-Protein Interactions: Computational and Experimental Tools*, BoD–Books On Demand, 2012.
- [42] J. M. Howell, T. L. Winstone, J. R. Coorssen, R. J. Turner, *Proteomics* 2006, 6, 2050.
- [43] H. Watanabe, H. Matsumaru, A. Ooishi, Y. Feng, T. Odahara, K. Suto, S. Honda, *Journal of biological chemistry* 2009, 284, 12373.
- [44] A. Velazquez-Campoy, S. A. Leavitt, E. Freire, *protein-protein interactions: methods and applications* 2015, 183.
- [45] D. Szklarczyk, L. J. Jensen, *Protein-Protein Interactions: Methods and Applications* 2015, 39.
-

-
- [46] I. A. Taylor, K. Rittinger, J. F. Eccleston, *Protein-Protein Interactions: Methods and Applications* 2015, 205.
- [47] T. Ehrenberger, L. C. Cantley, M. B. Yaffe, *Protein-Protein Interactions: Methods and Applications* 2015, 57.
- [48] S. Fletcher, A. D. Hamilton, *New Journal of Chemistry* 2007, 31, 623.
- [49] E. von Behring, S. Kitasato, *Deutsche Medizinische Wochenschrift* 1890, 16, 1145.
- [50] H. W. Schroeder Jr, L. Cavacini, *Journal of allergy and clinical immunology* 2010, 125, S41.
- [51] A. Tiselius, E. A. Kabat, *J Exp Med* 1939, 69, 119.
- [52] W. Choe, T. A. Durgannavar, S. J. Chung, *Materials* 2016, 9, 994.
- [53] A. Dalhoff, *Antimicrob Agents Chemother* 2018, 62, 10.
- [54] R. L. Kelly, Determinants of Antibody Specificity, Massachusetts Institute of Technology, 2017.
- [55] R. A. Dwek, A. C. Lellouch, M. R. Wormald, *J Anat* 1995, 187, 279.
- [56] R. Nezlin, *Immunol Lett* 2010, 132, 1.
- [57] H. Ma, C. Ó'Fágáin, R. O'Kennedy, *Biochimie* 2020, 177, 213.
- [58] E.-M. Strauch, S. J. Fleishman, D. Baker, *Proceedings of the National Academy of Sciences* 2014, 111, 675.
- [59] J. L. A. Voskuil, *F1000Res* 2014, 3.
- [60] V. Pascual, J. D. Capra, *Current Biology* 1991, 1, 315.
- [61] A. K. Abbas, A. H. Lichtman, S. Pillai, *Basic Immunology: Functions and Disorders of the Immune System, 6e: Sae-E-Book*, Elsevier India, 2019.
- [62] K. Murphy, C. Weaver, *Janeway's Immunobiology*, Garland Science, 2016.
- [63] A. I. Levinson, L. Kozlowski, Y. Zheng, L. Wheatley, *J Clin Immunol* 1995, 15, S26.
- [64] W. F. Verwey, *J Exp Med* 1940, 71, 635.
- [65] P. E. R. OEDING, A. GROV, B. Myklestad, *Acta Pathologica Microbiologica Scandinavica* 1964, 62, 117.
- [66] J. K. Myers, T. G. Oas, *Nat Struct Biol* 2001, 8, 552.
- [67] L. N. Deis, C. W. Pemble, Y. Qi, A. Hagarman, D. C. Richardson, J. S. Richardson, T. G. Oas, *Structure* 2014, 22, 1467.
- [68] T. Moks, L. ABRAHMSÉN, B. NILSSON, U. HELLMAN, J. SJÖQUIST, M. UHLÉN, *Eur J Biochem* 1986, 156, 637.
- [69] L. Jendeberg, P. Nilsson, A. Larsson, P. Denker, M. Uhlen, B. Nilsson, P.-Å. Nygren, *J Immunol Methods* 1997, 201, 25.
-

-
- [70] K. L. Atkins, J. D. Burman, E. S. Chamberlain, J. E. Cooper, B. Poutrel, S. Bagby, A. T. A. Jenkins, E. J. Feil, J. M. H. Van Den Elsen, *Mol Immunol* 2008, 45, 1600.
- [71] L. Björck, G. Kronvall, *J Immunol* 1984, 133, 969.
- [72] B. Guss, M. Eliasson, A. Olsson, M. Uhlen, A. K. Frej, H. Jörnvall, J. I. Flock, M. Lindberg, *EMBO J* 1986, 5, 1567.
- [73] U. Sjöbring, L. Björck, W. Kastern, *Journal of biological chemistry* 1991, 266, 399.
- [74] J. Nilvebrant, S. Hober, *Comput Struct Biotechnol J* 2013, 6, e201303009.
- [75] S. Hober, K. Nord, M. Linhult, *Journal of Chromatography B* 2007, 848, 40.
- [76] J. Fraser, V. Arcus, P. Kong, E. Baker, T. Proft, *Mol Med Today* 2000, 6, 125.
- [77] A. Sundstedt, M. Celander, M. W. Öhman, G. Forsberg, G. Hedlund, *Int Immunopharmacol* 2009, 9, 1063.
- [78] K. Urmann, P. Reich, J.-G. Walter, D. Beckmann, E. Segal, T. Scheper, *J Biotechnol* 2017, 257, 171.
- [79] H. Yue, Y. Zhou, P. Wang, X. Wang, Z. Wang, L. Wang, Z. Fu, *Talanta* 2016, 153, 401.
- [80] W.-L. Ling, Y.-L. Ng, A. Wipat, D. P. Lane, S. K.-E. Gan, *J Immunol Methods* 2020, 476, 112683.
- [81] K. M. Ng, C.-F. Wong, A. X. Liang, Y.-H. Liew, J. Y. Yeo, W.-H. Lua, X.-J. Qian, S. K.-E. Gan, *Scientific Phone Apps and Mobile Devices* 2019, 5.
- [82] J.-J. Poh, W.-L. Wu, N. W.-J. Goh, S. M.-X. Tan, S. K.-E. Gan, *Sens Actuators A Phys* 2021, 325, 112698.
- [83] W.-H. Lua, C. T.-T. Su, J. Y. Yeo, J.-J. Poh, W.-L. Ling, S.-X. Phua, S. K.-E. Gan, *Journal of Allergy and Clinical Immunology* 2019, 144, 514.
- [84] W. Salgado-Pabón, L. Breshears, A. R. Spaulding, J. A. Merriman, C. S. Stach, A. R. Horswill, M. L. Peterson, P. M. Schlievert, *mBio* 2013, 4, 10.
- [85] S. S. Bashraheel, A. D. AlQahtani, F. B. Rashidi, H. Al-Sulaiti, A. Domling, N. N. Orie, S. K. Goda, *Biomedicine & Pharmacotherapy* 2019, 115, 108905.
- [86] Y. Yang, M. Qian, S. Yi, S. Liu, B. Li, R. Yu, Q. Guo, X. Zhang, C. Yu, J. Li, *PLoS One* 2016, 11, e0149460.
- [87] A. K. Varshney, G. A. Kuzmicheva, J. Lin, K. M. Sunley, R. A. Bowling Jr, T.-Y. Kwan, H. R. Mays, A. Rambhadran, Y. Zhang, R. L. Martin, *PLoS One* 2018, 13, e0190537.
- [88] G. Forsberg, L. Ohlsson, T. Brodin, P. Björk, P. A. Lando, D. Shaw, P. L. Stern, M. Dohlsten, *Br J Cancer* 2001, 85, 129.
- [89] B. von Scheidt, M. Wang, A. J. Oliver, J. D. Chan, M. K. Jana, A. I. Ali, F. Clow, J. D. Fraser, K. M. Quinn, P. K. Darcy, *Proceedings of the National Academy of Sciences* 2019, 116, 25229.
-

-
- [90] D. Kanmert, Structure and Interactions of Human IgG-Fc, Linköping University, The Institute of Technology, 2011.
- [91] J. Ca, <http://www.garlandscience.com> 2001.
- [92] E. B. Myhre, G. Kronvall, *Comp Immunol Microbiol Infect Dis* 1981, 4, 317.
- [93] J. E. Butler, *Vet Immunol Immunopathol* 1983, 4, 43.
- [94] J. E. Butler, *J Dairy Sci* 1969, 52, 1895.
- [95] G. J. Sloan, J. E. Butler, *Am J Vet Res* 1978, 39, 935.
- [96] J. Goudswaard, J. A. Van der Donk, A. Noordzij, R. H. Van Dam, J. Vaerman, *Scand J Immunol* 1978, 8, 21.
- [97] M. J. P. Lawman, S. Joiner, D. R. Gauntlett, M. D. P. Boyle, *Comp Immunol Microbiol Infect Dis* 1985, 8, 1.
- [98] J. E. Butler, L. Peterson, P. L. McGivern, *Mol Immunol* 1980, 17, 757.
- [99] W. A. Wallner, M. J. P. Lawman, M. D. P. Boyle, *Appl Microbiol Biotechnol* 1987, 27, 168.
- [100] M. D. P. Boyle, W. A. Wallner, G. O. Von Mering, K. J. Reis, M. J. P. Lawman, *Mol Immunol* 1985, 22, 1115.
- [101] D. D. Richman, P. H. Cleveland, M. N. Oxman, K. M. Johnson, *J Immunol* 1982, 128, 2300.
- [102] D. Delacroix, J. P. Vaerman, *Mol Immunol* 1979, 16, 837.
- [103] R. C. Duhamel, E. Meezan, K. Brendel, *Mol Immunol* 1980, 17, 29.
- [104] G. Vidarsson, G. Dekkers, T. Rispens, *Front Immunol* 2014, 5, 520.
- [105] C. Endresen, *Acta Pathologica Microbiologica Scandinavica Section C Immunology* 1978, 86, 211.
- [106] R. Lindmark, K. Thorén-Tolling, J. Sjöquist, *J Immunol Methods* 1983, 62, 1.
- [107] M. D. P. Boyle, K. J. Reis, *Bio/technology* 1987, 5, 697.
- [108] M. Tashiro, G. T. Montelione, *Curr Opin Struct Biol* 1995, 5, 471.
- [109] E. VAN LOGHEM, B. FRANGIONE, B. RECHT, E. C. FRANKLIN, *Scand J Immunol* 1982, 15, 275.
- [110] J. SJÖDAHL, *Eur J Biochem* 1977, 73, 343.
- [111] J. J. Langone, *Biochem Biophys Res Commun* 1980, 94, 473.
- [112] B. Nilsson, T. Moks, B. Jansson, L. Abrahmsen, A. Elmblad, E. Holmgren, C. Henrichson, T. A. Jones, M. Uhlen, *Protein Engineering, Design and Selection* 1987, 1, 107.
- [113] B. Akerström, L. Björck, *Journal of Biological Chemistry* 1986, 261, 10240.
-

-
- [114] M. Eliasson, A. Olsson, E. Palmcrantz, K. Wiberg, M. Inganäs, B. Guss, M. Lindberg, M. Uhlen, *Journal of Biological Chemistry* 1988, 263, 4323.
- [115] J. B. Pilcher, V. C. W. Tsang, W. Zhou, C. M. Black, C. Sidman, *J Immunol Methods* 1991, 136, 279.
- [116] P. Nygren, M. Eliasson, L. Abrahmsén, M. Uhlén, E. Palmcrantz, *Journal of Molecular Recognition* 1988, 1, 69.
- [117] S. Fields, O. Song, *Nature* 1989, 340, 245.
- [118] U. Stelzl, E. E. Wanker, *Curr Opin Chem Biol* 2006, 10, 551.
- [119] E. D. Harlow, D. Lane, 1988.
- [120] T. Forster, *Ann Phys* 1948, 2, 55.
- [121] E. A. Jares-Erijman, T. M. Jovin, *Nat Biotechnol* 2003, 21, 1387.
- [122] R. Aebersold, M. Mann, *Nature* 2003, 422, 198.
- [123] A.-C. Gingras, M. Gstaiger, B. Raught, R. Aebersold, *Nat Rev Mol Cell Biol* 2007, 8, 645.
- [124] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, B. Séraphin, *Nat Biotechnol* 1999, 17, 1030.
- [125] O. Puig, F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, B. Séraphin, *Methods* 2001, 24, 218.
- [126] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, *Nature* 2002, 415, 180.
- [127] I. M. Cheeseman, A. Desai, *Science's STKE* 2005, 2005, p11.
- [128] J. Porath, J. A. N. Carlsson, I. Olsson, G. Belfrage, *Nature* 1975, 258, 598.
- [129] M. Wilchek, E. A. Bayer, *Anal Biochem* 1988, 171, 1.
- [130] D. S. Hage, *Clin Chem* 1999, 45, 593.
- [131] H. Zhu, M. Snyder, *Curr Opin Chem Biol* 2001, 5, 40.
- [132] G. MacBeath, S. L. Schreiber, *Science (1979)* 2000, 289, 1760.
- [133] F. X. R. Sutandy, J. Qian, C. Chen, H. Zhu, *Curr Protoc Protein Sci* 2013, 72, 21.
- [134] P. F. Predki, *Curr Opin Chem Biol* 2004, 8, 8.
- [135] I. Ghosh, A. D. Hamilton, L. Regan, *J Am Chem Soc* 2000, 122, 5658.
- [136] I. Remy, S. W. Michnick, *Biotechniques* 2007, 42, 137.
- [137] J. N. Pelletier, F.-X. Campbell-Valois, S. W. Michnick, *Proceedings of the National Academy of Sciences* 1998, 95, 12141.
- [138] M. L. MacDonald, J. Lamerdin, S. Owens, B. H. Keon, G. K. Bilter, Z. Shang, Z. Huang, H. Yu, J. Dias, T. Minami, *Nat Chem Biol* 2006, 2, 329.
-

-
- [139] G. P. Smith, *Science (1979)* 1985, 228, 1315.
- [140] C. F. Barbas, (*No Title*) 2001.
- [141] S. S. Sidhu, S. Koide, *Curr Opin Struct Biol* 2007, 17, 481.
- [142] H. R. Hoogenboom, P. Chames, *Immunol Today* 2000, 21, 371.
- [143] J. Drenth, *Principles of Protein X-Ray Crystallography*, Springer Science & Business Media, 2007.
- [144] G. Rhodes, *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*, Elsevier, 2010.
- [145] W. A. Hendrickson, *Science (1979)* 1991, 254, 51.
- [146] J. Cavanagh, *Protein NMR Spectroscopy: Principles and Practice*, Academic Press, 1996.
- [147] G. M. Clore, J. Iwahara, *Chem Rev* 2009, 109, 4108.
- [148] E. R. P. Zuiderweg, *Biochemistry* 2002, 41, 1.
- [149] N. J. Greenfield, *Protein-Protein Interactions: Methods and Applications* 2015, 239.
- [150] M. A. Haque, P. Kaur, A. Islam, M. I. Hassan, in *Advances in Protein Molecular and Structural Biology Methods*, Elsevier, 2022, pp. 213–224.
- [151] S. M. Kelly, N. C. Price, *Curr Protein Pept Sci* 2000, 1, 349.
- [152] N. J. Greenfield, *Nat Protoc* 2006, 1, 2876.
- [153] R. W. Woody, *Methods Enzymol* 1995, 246, 34.
- [154] Y. Choi, J. Han, C. Kim, *Korean Journal of Chemical Engineering* 2011, 28, 2130.
- [155] K. Sefiane, *Adv Colloid Interface Sci* 2014, 206, 372.
- [156] H. Y. Erbil, *Adv Colloid Interface Sci* 2012, 170, 67.
- [157] S. S. Sazhin, *Prog Energy Combust Sci* 2006, 32, 162.
- [158] R. G. Picknett, R. Bexon, *J Colloid Interface Sci* 1977, 61, 336.
- [159] S. Tonini, G. E. Cossali, *International Journal of Thermal Sciences* 2012, 57, 45.
- [160] R. Holyst, *TCE: The Chemical Engineer* 2008.
- [161] K. Sefiane, *Adv Colloid Interface Sci* 2014, 206, 372.
- [162] D. Kaya, V. A. Belyi, M. Muthukumar, *J Chem Phys* 2010, 133.
- [163] N. Kim, Z. Li, C. Hurth, F. Zenhausern, S.-F. Chang, D. Attinger, *Analytical methods* 2012, 4, 50.
- [164] M. Wang, X. Wang, P. Moni, A. Liu, D. H. Kim, W. J. Jo, H. Sojoudi, K. K. Gleason, *Advanced Materials* 2017, 29, 1604606.
-

-
- [165] M. Ramanathan, S. B. Darling, *Prog Polym Sci* 2011, 36, 793.
- [166] A. Khlyustova, Y. Cheng, R. Yang, *J Mater Chem B* 2020, 8, 6588.
- [167] W. J. Lau, A. F. Ismail, N. Misdan, M. A. Kassim, *Desalination* 2012, 287, 190.
- [168] P. T. Hammond, *Materials Today* 2012, 15, 196.
- [169] P. T. Hammond, *AIChE Journal* 2011, 57, 2928.
- [170] A. D. Price, A. P. R. Johnston, G. K. Such, F. Caruso, *Modern Techniques for Nano-and Microreactors/-reactions* 2010, 155.
- [171] F. Xue, Z. Liu, Y. Su, K. Varahramyan, *Microelectron Eng* 2006, 83, 298.
- [172] D. Klemm, B. Heublein, H. Fink, A. Bohn, *Angewandte chemie international edition* 2005, 44, 3358.
- [173] J. Lahann, *Polym Int* 2006, 55, 1361.
- [174] M. M. Byranvand, F. Behboodi-Sadabad, A. A. Eliwi, V. Trouillet, A. Welle, S. Ternes, I. M. Hossain, M. R. Khan, J. A. Schwenzer, A. Farooq, *J Mater Chem A Mater* 2020, 8, 20122.
- [175] H.-Y. Chen, J. Lahann, *Langmuir* 2011, 27, 34.
- [176] C. A. D. Dion, J. R. Tavares, *Powder Technol* 2013, 239, 484.
- [177] T. M. Hafshejani, X. Zhong, J. Kim, B. Dadfar, J. Lahann, *Organic Materials* 2023, 5, 98.
- [178] N. Winterton, *Clean Technol Environ Policy* 2021, 23, 2499.
- [179] R. S. Varma, 2016.
- [180] H. M. Marvaniya, K. N. Modi, D. J. Sen, *Int. J. Drug Dev. Res* 2011, 3, 42.
- [181] S. J. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, Pearson, 2016.
- [182] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, 2016.
- [183] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *Nature* 2016, 529, 484.
- [184] Y. LeCun, Y. Bengio, G. Hinton, *Nature* 2015, 521, 436.
- [185] A. Valencia, F. Pazos, *Structural Bioinformatics* 2003, 44, 409.
- [186] L. Hu, X. Wang, Y.-A. Huang, P. Hu, Z.-H. You, *Brief Bioinform* 2021, 22, bbab036.
- [187] Y. Wang, Z. You, L. Li, Z. Chen, *Front Comput Sci* 2020, 14, 1.
- [188] A. Dhakal, C. McKay, J. J. Tanner, J. Cheng, *Brief Bioinform* 2022, 23, bbab476.
- [189] C. Shi, J. Chen, X. Kang, G. Zhao, X. Lao, H. Zheng, *Protein Pept Lett* 2020, 27, 359.
- [190] E. H. Houssein, R. E. Mohamed, A. A. Ali, *IEEE Access* 2021, 9, 140628.
-

-
- [191] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, *Nat Rev Drug Discov* 2019, 18, 463.
 - [192] A. Ahmed, B. Mam, R. Sowdhamini, *Bioinform Biol Insights* 2021, 15, 11779322211030364.
 - [193] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, *J R Soc Interface* 2018, 15, 20170387.
 - [194] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, *Nature* 2017, 542, 115.
 - [195] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, *Nat Commun* 2014, 5, 4006.
 - [196] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, *Nature* 2017, 542, 115.
 - [197] M. W. Libbrecht, W. S. Noble, *Nat Rev Genet* 2015, 16, 321.
 - [198] A. Rajkomar, J. Dean, I. Kohane, *N Engl J Med* 2019, 380, 2589.
 - [199] K. W. Johnson, J. Torres Soto, B. S. Glicksberg, K. Shameer, R. Miotto, M. Ali, E. Ashley, J. T. Dudley, *J Am Coll Cardiol* 2018, 71, 2668.
 - [200] E. Tjoa, C. Guan, *IEEE Trans Neural Netw Learn Syst* 2020, 32, 4793.
 - [201] D. S. Char, N. H. Shah, D. Magnus, *N Engl J Med* 2018, 378, 981.
 - [202] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, *Wiley Interdiscip Rev Data Min Knowl Discov* 2019, 9, e1312.
 - [203] A. Krizhevsky, I. Sutskever, G. E. Hinton, *Adv Neural Inf Process Syst* 2012, 25.
 - [204] O. Ronneberger, P. Fischer, T. Brox, in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III* 18, Springer, 2015, pp. 234–241.
 - [205] D. Ciresan, A. Giusti, L. Gambardella, J. Schmidhuber, *Adv Neural Inf Process Syst* 2012, 25.
 - [206] K. Simonyan, A. Zisserman, *arXiv preprint arXiv:1409.1556* 2014.
 - [207] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.
 - [208] K. He, X. Zhang, S. Ren, J. Sun, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
 - [209] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.
-

-
- [210] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, K. He, *arXiv preprint arXiv:1706.02677* 2017.
 - [211] N. S. Keskar, R. Socher, *arXiv preprint arXiv:1712.07628* 2017.
 - [212] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *Proceedings of the IEEE* 1998, 86, 2278.
 - [213] M. Li, T. Zhang, Y. Chen, A. J. Smola, in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 661–670.
 - [214] S. Aggarwal, S. Gupta, R. Kannan, R. Ahuja, D. Gupta, S. Juneja, S. B. Belhaouari, *IEEE Access* 2022, 10, 83591.
 - [215] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, J. Liang, *IEEE Trans Med Imaging* 2016, 35, 1299.
 - [216] B. Liu, W. Shen, P. Li, X. Zhu, in *2019 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2019, pp. 1–8.
 - [217] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, *Med Image Anal* 2017, 42, 60.
 - [218] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.
 - [219] L. Van der Maaten, G. Hinton, *Journal of machine learning research* 2008, 9.
 - [220] J. Cook, I. Sutskever, A. Mnih, G. Hinton, in *Artificial Intelligence and Statistics*, PMLR, 2007, pp. 67–74.
 - [221] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, *arXiv preprint arXiv:1611.07450* 2016.
 - [222] J. Stier, M. Granitzer, *Software Impacts* 2022, 11, 100193.
 - [223] J. A. Barnes, F. Harary, *Soc Networks* 1983, 5, 235.
 - [224] S. Zhang, H. Tong, J. Xu, R. Maciejewski, *Comput Soc Netw* 2019, 6, 1.
 - [225] A.-L. Barabasi, Z. N. Oltvai, *Nat Rev Genet* 2004, 5, 101.
 - [226] S. S. Shen-Orr, R. Milo, S. Mangan, U. Alon, *Nat Genet* 2002, 31, 64.
 - [227] P. D. Karp, S. Paley, P. Romero, *Bioinformatics* 2002, 18, S225.
 - [228] O. Sporns, *Ann N Y Acad Sci* 2011, 1224, 109.
 - [229] J. Wang, H. Gao, Y. Han, C. Ding, S. Pan, Y. Wang, Q. Jia, H.-T. Wang, D. Xing, J. Sun, *Natl Sci Rev* 2023, 10, nwad128.
 - [230] X. Deng, C. Friedmann, J. Lahann, 2011.
 - [231] H.-Y. Chen, Y. Elkasabi, J. Lahann, *J Am Chem Soc* 2006, 128, 374.
 - [232] B. Dadfar, S. Vaez, C. Haret, M. Koenig, T. Mohammadi Hafshejani, M. Franzreb, J. Lahann, *Small Struct* n.d., n/a, 2400204.
-

-
- [233] H. Zhang, D. Vecchio, A. Emre, S. Rahmani, C. Cheng, J. Zhu, A. C. Misra, J. Lahann, N. A. Kotov, *MRS Bull* 2021, 46, 576.
- [234] S. Sechi, P. P. Roller, J. Willette-Brown, J. P. Kinet, *Journal of Biological Chemistry* 1996, 271, 19256.
- [235] N. L. Brown, S. P. Bottomley, M. D. Scawen, M. G. Gore, *Mol Biotechnol* 1998, 10, 9.
- [236] B. Akerström, E. Nielsen, L. Björck, *Journal of Biological Chemistry* 1987, 262, 13388.
- [237] P. L. Ey, S. J. Prowse, C. R. Jenkin, *Immunochemistry* 1978, 15, 429.
- [238] S. Ghose, M. Allen, B. Hubbard, C. Brooks, S. M. Cramer, *Biotechnol Bioeng* 2005, 92, 665.
- [239] L. Hamadeh, S. Imran, M. Bencsik, G. R. Sharpe, M. A. Johnson, D. J. Fairhurst, *Sci Rep* 2020, 10, 3313.
- [240] C. Angermueller, T. Pärnamaa, L. Parts, O. Stegle, *Mol Syst Biol* 2016, 12, 878.
- [241] D. Lancet, D. Isenman, J. Sjödaahl, J. Sjöquist, I. Pecht, *Biochem Biophys Res Commun* 1978, 85, 608.
- [242] Y. Kanamaru, S. Nagaoka, Y. Kuzuya, *Anim. Sci. Technol.(Jpn.)* 1992, 63, 385.
- [243] K. Saha, F. Bender, E. Gizeli, *Anal Chem* 2003, 75, 835.
- [244] Y. Bengio, A. Courville, P. Vincent, *IEEE Trans Pattern Anal Mach Intell* 2013, 35, 1798.
- [245] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, M. Data, in *Data Mining*, Elsevier Amsterdam, The Netherlands, 2005, pp. 403–413.
- [246] W. Rawat, Z. Wang, *Neural Comput* 2017, 29, 2352.
- [247] H. Liu, H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*, Springer Science & Business Media, 2012.
- [248] Z. A. Zhao, H. Liu, *Spectral Feature Selection for Data Mining*, Taylor & Francis, 2012.
- [249] C. Ding, H. Peng, *J Bioinform Comput Biol* 2005, 3, 185.
- [250] H. Liu, L. Yu, *IEEE Trans Knowl Data Eng* 2005, 17, 491.
- [251] Y. Saeys, I. Inza, P. Larranaga, *bioinformatics* 2007, 23, 2507.
- [252] I. Guyon, A. Elisseeff, *Journal of machine learning research* 2003, 3, 1157.
- [253] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, *Knowl Inf Syst* 2013, 34, 483.
-