UNIVERSITÀ
DI SIENA
1240

Department of
Biotechnology, Chemistry and Pharmacy

Doctorate program in
CHEMICAL AND PHARMACEUTICAL SCIENCES Cicle 38°
Coordinator: Prof. Maurizio Taddei

Rational design and data-driven strategies to optimize bioconjugation processes in antibody-drug conjugates

Supervisor:
Prof. Andrea Tafi

Co-Supervisors:
Prof. Elena Petricci

Candidate:
Lorenzo Angiolini
Mat. 131482

Academic Year 2025-2026

Abstract

The main focus of this thesis is to address the challenges in developing efficient and cost-effective drug delivery systems. Among the most promising approaches are antibody-drug conjugates (ADCs), which combine cytotoxic or bioactive agents with monoclonal antibodies (mAbs) to achieve targeted therapies. However, bioconjugation processes can lead to variable outcomes depending on the choice of mAb, amino acid residues, and linker-payload (LP) systems, making the design of effective ADCs a complex task. In this work, a machine learning (ML) framework capable of predicting bioconjugation outcomes is presented, thereby guiding the selection of optimal mAb-LP combinations and reaction conditions. Specifically, the eXtremeGradientBoosting (XGBoost) algorithm is used to model and predict the drug-to-antibody ratio (DAR) in ADC synthesis. The proposed approach demonstrates high predictive accuracy, achieving R^2 scores of 0.85 and 0.91 for lysine- and cysteine-conjugated datasets, respectively. By integrating ML algorithms into the design and optimization of ADC bioconjugation processes, this study provides a data-driven strategy to streamline ADC development and improve the efficiency of targeted drug delivery systems.

In addition, during my period at The Scripps Research Institute, La Jolla (CA). I contributed to a project aimed at improving the existing reactive docking methodology, designed to model and predict reactions between small molecules and biological macromolecules. In this work, pseudo-atoms (PAs) were introduced on the ligand warhead to encode the geometry and spatial orientation necessary for covalent bond formation, enabling the prediction of the optimal near-attack conformation (NAC). Here, I present the preliminary results obtained from reactive docking using PAs on two cysteine-reactive warheads (chloroacetamides and acrylamides), in predicting the correct reactive residues and the optimal geometric approach for covalent bond formation.

Contents

List of Abbreviations **7**

I Machine learning for predicting the drug-to-antibody ratio in the synthesis of antibody-drug conjugates **13**

1 Introduction **14**

1.1 Machine learning 15

1.1.1 Machine Learning (ML): a brief overview 16

1.2 ML in Scientific Research and Drug Discovery 20

1.3 Antibody-drug conjugates 26

1.3.1 Antibodies 34

1.3.2 Linkers 35

1.3.2.1 Cleavable linkers 37

1.3.2.2 Non-cleavable linkers 41

1.3.3 Payloads 42

1.3.3.1 Microtubule inhibitors 43

1.3.3.2 DNA damaging agents 44

1.3.3.3 Topoisomerase inhibitors 47

1.3.4 Beyond oncology antibody-drug conjugates (ADCs) 47

1.3.4.1 Anti-inflammatory ADCs 48

1.3.4.2 Antibody-antibiotic conjugates 48

1.3.4.3 Immunosuppressive ADCs 49
1.3.5 Conventional bioconjugation techniques and drug-to-antibody ratio (DAR) 49
1.3.6 Conjugation innovations 54
1.3.7 ADCs Characterization 58
1.3.7.1 UV/Vis spectrometry 58
1.3.7.2 Mass spectrometry (MS) methods 59
1.3.7.3 Chromatographic methods 60
1.3.8 Future perspectives 62
1.3.9 Aim of this research project 62

2 Results and discussion **64**
2.1 Data acquisition and preprocessing 65
2.1.1 Label encoding 66
2.2 Feature selection 70
2.2.1 Feature scaling 71
2.3 Training phase 71
2.3.1 Model evaluation: Classifiers 72
2.3.2 Model evaluation: Regressors 76
2.3.3 Hyperparameter tuning 84
2.4 Experimental validation 85
2.4.1 Case study 90
2.4.2 MALDI spectra 93

3 Conclusions **96**

4 Experimental section **98**
4.0.1 Conjugation to Lys residues 99
4.0.1.1 General procedure 99
4.0.1.2 In-situ procedure 100

4.0.1.3 Pre-functionalization procedure 100
4.0.2 Conjugation to Cys residues 100
4.0.2.1 Two steps protocol 100
4.0.2.2 Onepot protocol 101
4.0.3 MALDI-TOF analysis of bioconjugates to estimate DAR 101

II Development of a pseudo-atom approach to optimize reactive docking of covalent inhibitors 103

5 Introduction 104
5.1 Targeted covalent inhibitors 104
5.2 Reactive docking 106
5.3 Aim of the project 107

6 Results and discussion 109
6.1 Ligands preparation 109
6.2 Receptors preparation 110
6.3 Reactive docking 110
6.4 Parameters selection 112
6.5 Calibration and success rate 112
6.6 Virtual screening 114

7 Conclusions 116

Appendix A 117
7.1 Data Preprocessing Scripts 117
7.2 Model Training and Evaluation Scripts 119
7.3 Hyperparameter Optimization Scripts 123
7.4 Pseudo-atom coordinates calculation 124
7.5 Categorical features encoding 126

References **127**

Acknowledgements **146**

List of Abbreviations

AACs antibody-antibiotic conjugates.

ADCC antibody-dependent cellular cytotoxicity.

ADCP antibody-dependent cellular phagocytosis.

ADCs antibody-drug conjugates.

ADMET absorption-distribution-metabolism-excretion-toxicity.

AI Artificial Intelligence.

ALL acute lymphoblastic leukemia.

AML acute myeloid leukemia.

ANNs ArtificialNeural Networks.

BCMA B-cell maturation antigen.

BCP-ALL B-cell precursor acute lymphoblastic leukemia.

BTK Bruton's tyrosine kinase.

BVC Bevacizumab.

CASP14 14th Critical Assessment of protein Structure Prediction.

CDC complement-dependent cytotoxicity.

CML chronic myelogenous leukemia.

cryo-EM cryo-electron microscopy.

Cs corticosteroids.

CTX cetuximab.

DAR drug-to-antibody ratio.

DBM dibromo-maleimide.

DL Deep Learning.

DLBCL diffuse large B-cell lymphoma.

DMSO dimethylsulfoxide.

DNA Deoxyribonucleic Acid.

DPR diaminopropionic acid.

DTs Decision trees.

DTT dithiothreitol.

EDC-HCl 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride.

EGFR epidermal growth factor receptor.

EMA European Medicines Agency.

ESI Electrospray ionization.

Fab fragment antigen-binding.

Fc fragment crystallizable.

FcRs Fc receptors.

FDA Food and Drug Administration.

FGE formylglycine-generating enzyme.

fGly formylglycine.

FRα folate receptor alpha.

GAs Genetic algorithms.

GCR glucocorticoid receptor.

GNNs graph neural networks.

GSH glutathione.

GSK GlaxoSmithKline.

HCL hairy cell leukemia.

HER2 human epidermal growth factor receptor.

HIC hydrophobic interaction chromatography.

HIPS Hydrazino-iso-Pictet-Spengler.

HL Hodgkin's lymphoma.

HPC High Performance Computing.

HPLC high-performance liquid chromatography.

HR hormone receptor.

IgG immunoglobulin G.

IL-6 interleukin-6.

k-NN K-nearest neighbor.

LBCL large B-cell lymphoma.

LDA linear discriminant analysis.

LGA Lamarckian Genetic Algorithm.

LP linker-payload.

mAb monoclonal antibody.

MAE mean absolute error.

MALDI-TOF MS Matrix-assisted laser desorption/ionization-time-of-flight.

MAPK mitogen-activated protein kinase.

MC Maleimidocaproyl.

ML Machine Learning.

MOE-type molecular operating environment.

MolLogP partition coefficient.

MS Mass spectrometry.

MSE mean squared error.

NAC near-attack conformations.

NHS N-hydroxysuccinimide.

NK natutal killer.

NMR nuclear magnetic resonance.

NSCLC non-small cell lung cancer.

NUDT7 Nudix Hydrolase 7.

PABC p-aminobenzyl carbamate.

PAs Psuedo-Atoms.

PBD Pyrrolobenzodiazepine dimers.

PBS phosphate buffered saline.

PD pyridazinedione.

PE38 Pseudomonas exotoxin.

PEG polyethylene glycol.

PEOE Partial Equalization of Orbital Electronegativities.

Phe-Lys phenylalanine-lysine.

PI3K/AKT phosphatidylinositol 3-kinase.

PPI protein-protein interaction.

PSSMs position-specific scoring matrices.

QSAR Quantitative structure-activity relationships.

R/R MM elapsed or refractory multiple myeloma.

R^2 coefficient of determination.

RA rheumatoid arthritis.

RFs Random Forests.

RNA Ribonucleic Acid.

RNNs Recurrent Neural Networks.

RPLC reversed-phase liquid chromatography.

sALCL systemic anaplastic large cell lymphoma.

SD standard deviation.

SEC size exclusion chromatography.

SHAP SHapley Additive exPlanations.

SMILES Simplified Molecular Input Line Entry System.

SPAAC train-promoted azide-alkyne cycloadditions.

SPS Spatial Score.

sulfo-NHS N-hydroxysulfosuccinimide.

SVMs Support Vector Machines.

T-DM1 trastuzumab emtansine.

TCEP tris(2-carboxyethyl)phosphine.

TCIs targeted covalent inhibitors.

TCZ-ALD tocilizumab-alendronate.

TOP1 topoisomerase 1.

TOPcc Topoisomerase cleavage complexes.

TROP2 Trophoblast cell surface antigen 2.

TRX trastuzumab.

UV/Vis ultrviolet/visibile spectroscopy.

Val-Ala valine-alanine.

Val-Cit valine-citrulline.

VS Virtual screening.

VSA van der Waals surface area.

XAI Explainable AI.

Part I

Machine learning for predicting the drug-to-antibody ratio in the synthesis of antibody-drug conjugates

Chapter 1

Introduction

Drug development is a critical component of contemporary medicine, yet it remains a very complex, expensive, and time-consuming procedure. There is a less than 10% chance of success from initial discovery to regulatory approval, and the average development of a single new therapeutic treatment takes more than ten years and costs more than several billion dollars.[1] The enormous biological complexity of disease pathways, the challenge of finding safe and effective therapeutic candidates, and the limitations of conventional experimental and computational methods are the main causes of this high turnover rate.[1] Simultaneously, the need for novel treatments keeps rising due to the increasing number of infectious diseases, antibiotic resistance, and the growing prevalence of long-term illnesses like diabetes, cancer, and neurological disorders. Recent advances in Artificial Intelligence (AI) and ML are transforming the way biomedical data are analyzed, interpreted, and applied to pharmaceutical research.[2] ML algorithms have the ability to learn intricate patterns from high-dimensional information, which sets them apart from traditional rule-based computational techniques and opens up new possibilities for target identification, molecular design, and drug repurposing.[3] The disruptive potential of these methods is demonstrated by innovations like AI-enabled protein structure prediction, predictive models of pharmacokinetic and toxicity profiles, and deep generative models

for de novo molecule production. The potential of incorporating AI and ML into drug development processes is demonstrated by early achievements, such as AI-designed drugs progressing to clinical trials. However, critical challenges remain, including data scarcity, model interpretability, and the need for seamless integration with experimental validation.[4] Against this backdrop, this thesis explores how ML can be advanced and tailored to address specific bottlenecks in drug discovery, with the overarching goal of developing computational tools that enhance efficiency, reliability, and scientific insight in the design of novel therapeutics.

1.1 Machine learning

AI is a scientific discipline that aims to develop machines capable of performing tasks related to human perception. Although the term "artificial intelligence" was officially introduced in 1956 by computer scientist John McCarthy, other researchers had already made significant contributions to this field in previous years. As early as 1943, Warren McCulloch and Walter Pitt proposed the first artificial neuron to the scientific world[5], and already in the 1950s, the first working prototypes of neural networks were created.[6,7]. However, it was the research of Alan Turing that increased the interest of the public. Turing, attempting to explain how and to what extent computers could actually simulate human behaviour, devised a test - the Turing test -[8] to be able to give a measure of the thinking ability of machines. Initially, computers were more adept at solving intellectually challenging problems that were relatively straightforward for them, such as those that could be described using a list of mathematical rules[9]. However, the AI challenge later shifted to solving tasks that were easy for humans to perform but difficult to formally describe, as they were related to perceptual skills developed over hundreds of thousands of years through an evolutionary process. The goal was to create machines capable of acquiring knowledge independently through experience. The concept of a hierarchical

structure of concepts enabled the machine to learn complex notions by building them upon the simplest ones. For this reason, a stratified form of thinking was necessary to grasp complex concepts, as exemplified by Deep Learning (DL)[10], which draws inspiration from the way biological neural networks in the human brain process information. DL is merely a subcategory of a broader family of AI methods known as ML.

1.1.1 ML: a brief overview

ML emerged from the notion that computers can acquire skills through experience, enabling them to perform specific tasks. At its core, ML employs a series of algorithms that, starting from basic principles, learn to make specific decisions or execute actions over time. Computers are provided with a set of data (training set), which is systematically examined to extract information, akin to human learning. Based on the manner in which the machine learns data and information, four distinct learning methods can be distinguished.

- **Supervised learning**, as described by Cunningham *et al.*[11], involves training data that are labeled with a target, or an "expected result." This allows the system to acquire experience during the training phase and subsequently use that knowledge to solve problems that share similar fundamental concepts.

- **Unsupervised learning**[12], involves training a model without labeled data. The model's task is to identify relationships between the data points, utilizing no prior knowledge about the data itself.

- **Semi-supervised learning**, as proposed by Zhu *et al.*[13], involves training a model with a partially labeled dataset. This approach is particularly useful in scenarios where the knowledge about the data is incomplete or where the collection and sampling phase of labeled data is prohibitively expensive.

- **Reinforcement learning**, as described by Sutton and Barto[14], is a training method where the training set is unlabeled. Instead, an example is provided with a positive or negative result. This result serves as feedback for the algorithm, enabling it to assess whether the provided solution effectively addresses a problem. Essentially, it's the computerized version of human learning through trial and error.

The ML process consists of six components regardless of the algorithm adopted, as shown in Figure 1.

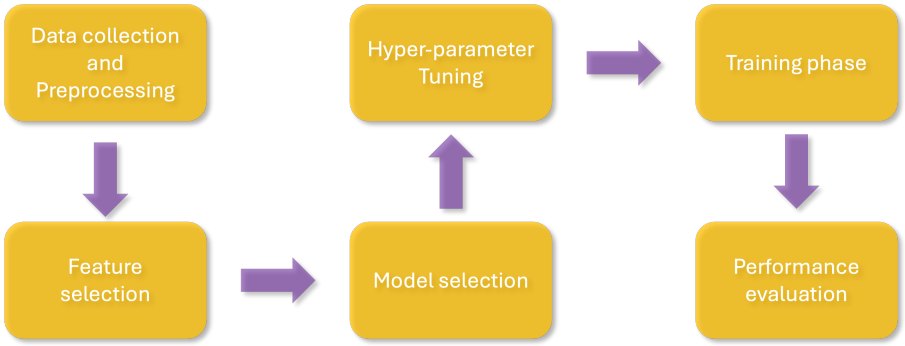


Figure 1.1: Six main components of the ML process.

Data collection and pre-processing involve preparing data in a format suitable for input to the algorithm. Unstructured, sparse, and often redundant data require cleaning and pre-processing into a structured format. Since ML models require numeric input and output variables, categorical data must be encoded using various strategies. Standardizing the dataset also helps avoid bias in the outcome. The pre-processed data may contain numerous features, not all of which are relevant to the learning process.[15] Feature selection reduces the number of input variables to minimize computational costs and potentially enhance model performance. However, not all ML algorithms are suitable for all problems; certain algorithms are better suited to specific classes of problems. Choosing the appropriate ML algorithm is crucial for achieving optimal results.[16,17] Once the most suitable model is identified, strategies must be employed to determine the optimal values for its parameters. This process, known as hyperparameter tuning, involves searching for the ideal model

architecture. Different ML models may require varying constraints, weights, or learning rates to effectively generalize to diverse data patterns.[18] The ultimate objective of any ML model is to acquire knowledge from examples in a manner that enables it to generalize the learned information to novel instances it has not encountered before. To achieve this, the model is trained on a subset of the total dataset and subsequently tested against unseen data to assess its learning progress using performance metrics such as accuracy, precision, and recall.[19] Furthermore, there are many tasks that an ML tool can perform:

- **Classification:** involves dividing input data into two or more classes and training a learning system to produce a model capable of assigning a class to each input.
- **Regression:** is conceptually similar to classification, except that the output belongs to a continuous domain instead of a discrete one.
- **Clustering:** involves dividing a set of input data into groups without prior knowledge. Unlike classification, neither the number nor the type of the classes (target) is known.

ML encompasses a diverse range of algorithms, starting with Decision trees (DTs), genetic and boosting algorithms, and metric techniques like the K-nearest neighbor (k-NN) algorithm, Support Vector Machines (SVMs), statistical methods, Bayesian networks, and ArtificialNeural Networks (ANNs).[20,21] Selecting the most appropriate ML approach for a particular problem is crucial for developing a robust and reliable model. Factors such as the problem statement and desired output type play a crucial role in this decision. The well-known "no free lunch" theorem for supervised ML states that all optimization algorithms perform equally well when their performance is averaged across all possible problems.[22] This implies that no single ML algorithm is universally the best-performing for all problems.

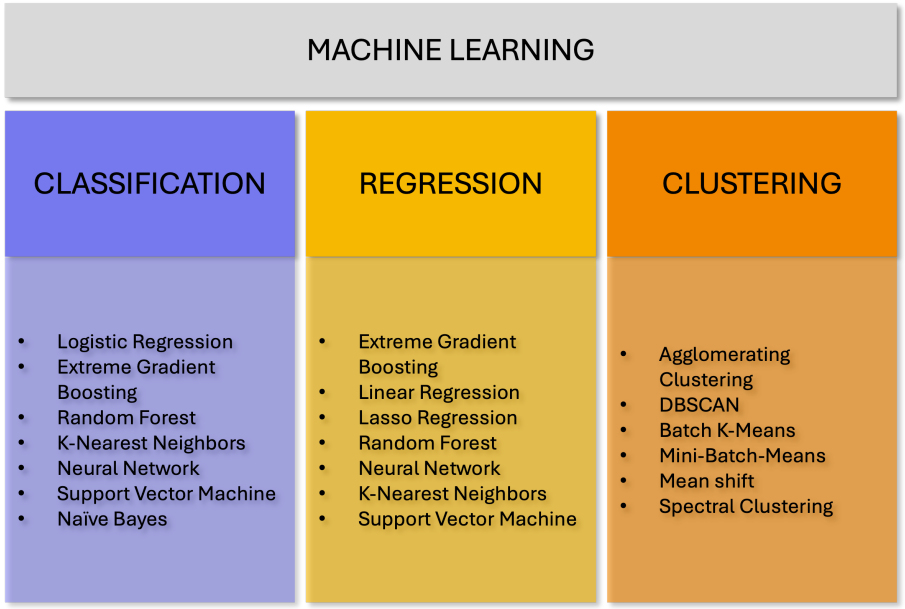


Figure 1.2: Overview of fundamental ML paradigms. The figure illustrates three primary categories of ML algorithms.

Despite a growing body of work demonstrating the diverse applications of ML, several enduring challenges remain, especially regarding choosing the optimal ML architecture and addressing issues of data, expertise, and trust. First, there is no widely accepted methodology for systematically selecting the "best-fit" ML architecture; this remains an open area for future research. Second, there is a skills gap: developers with deep ML knowledge are relatively rare, and many data scientists, though familiar with modeling and statistical methods, often lack strong software engineering skills, which are essential for production-quality ML systems.[23] Third, ML models typically require large volumes of high-quality data; in many domains, especially healthcare, such data are difficult to collect, are fragmented, may be inaccessible due to regulatory or institutional barriers, and when available are sometimes unreliable, any of which can degrade model performance, or worse, lead to harmful outcomes. Fourth, in sensitive application domains such as medicine or autonomous driving, explainability of ML decision-support systems is not optional: stakeholders (clinicians, patients, regulators) must understand how and why

decisions are made. Without explainability, even if models achieve superior accuracy, their adoption is likely to face ethical, legal, and social resistance.[24] Explainable AI (XAI) methods are thus increasingly regarded as a key frontier, especially in healthcare, for enabling transparency, fostering trust, and ensuring accountability.[25] Finally, domains like speech recognition, rare disease detection, and drug discovery continue to pose significant challenges for current ML and DL techniques. Looking forward, ML is expected to serve as a driver of changing innovation; however, for its potential to positively impact society and quality of life, these technical, ethical, and practical challenges must be addressed.

1.2 ML in Scientific Research and Drug Discovery

The process of finding and developing new drugs is difficult, time-consuming, and resource-intensive; it often takes over 10 years and billions of dollars to get a single chemical from first screening to market approval. Failure rates are still high, especially in late-stage clinical studies where compounds frequently fail because of unexpected safety issues or lack of efficacy. In this regard, AI and ML have become revolutionary instruments that can increase the chances of success, lower expenses, and speed up discovery.[3] The integration of ML techniques at various phases of the drug discovery process is growing. ML techniques are employed in early-stage discovery to evaluate large chemical libraries and forecast bioactivity, absorption-distribution-metabolism-excretion-toxicity (ADMET) traits, and physical features. Before expensive wet-lab investigations, these predictive models assist in ranking promising candidates. By capturing the nonlinear correlations between chemical structures and biological activity, DL in particular has made it possible to improve Virtual screening (VS), de novo drug creation, and molecular docking predictions.[26] These applications are part of a broader set of ML tasks, which include classification, regression, clustering, and data presentation for easy interpretation. Such approaches are not only cheaper and faster than traditional methods but also capable

of analyzing massive amounts of biological data in seconds, enabling the generation of predictive models that can inform decision-making throughout drug discovery.

- **Artificial neural networks (ANNs)**[5] are widely used thanks to the ability to solve complex real-world problems. ANNs are robust tools that can manage big data, identifying key components, thus providing a greater understanding of the biological system being modeled. A neural network consists of an oriented graph formed by nodes (organized in layers) connected by arcs. Each arc is associated with a weight, while nodes are equipped with activation functions that elaborate the inputs to produce the neuron output.[27] Supervised neural network learning is based on a feedback process, called back-propagation, in which the output of the network is compared with the output it was meant to produce, and the difference between the outputs is used to modify the weights of the connections between the neurons in the network. Moreover, they are usually resistant to noise and errors present in training data. In particular, Recurrent Neural Networks (RNNs) are a powerful and robust type of neural architecture, provided with feedback connections, that produce internal loops. Such loops induce a recursive dynamic within the networks and thus introduce delayed activation dependencies across the processing elements. In doing so, RNNs develop a kind of memory that makes them particularly tailored to process sequential data, such as text, Deoxyribonucleic Acid (DNA), proteins, etc.[28,29]

- **DTs**[30] are structures resembling trees, where each internal node represents a decision on an attribute, each leaf node represents a feature label, and each branch represents the value of that feature. A DT classifies instances by sorting them from the root to some leaf nodes based on feature values. The primary objective of DTs is to arrange different nodes based on valid data, as each node in a DT addresses an item in an occurrence to be sorted, and each branch addresses a value that the hub can accept. When using a DT, the focus is on determining which attribute is the most effective

classifier at each node level. Statistical measures such as information gain, Gini index, Chi-square, and entropy are calculated for each node to quantify its value. DTs have been successfully applied in practical applications for protein function prediction, protein-protein interaction (PPI) analysis, and cancer classification.[31]

- **Support Vector Machines (SVMs)**[32] are supervised learning tools that can be used for both classification and regression problems. They maximize the margin between two classes to ensure that the trained model generalizes well to test data. Each data item is first plotted as a point in an n-dimensional space, and the model classifies the data into different classes by finding a hyperplane that separates them. SVMs are relatively simple and flexible, making them suitable for addressing a variety of problems. They offer generally good predictive performance with a lower risk of overfitting, which is why they have been widely applied to many areas of bioinformatics, including protein function prediction, protease functional site recognition, transcription initiation site prediction, and gene expression data classification.[33]

- **Genetic algorithms (GAs)**[34] have gained popularity in science-related research due to their simplicity. These heuristic techniques, inspired by the mechanics of natural selection, use a finite series of standard steps to find solutions to problems.[35]

- **Ensemble learning**,[36] a widely used technique, combines multiple learning algorithms to enhance overall prediction accuracy and mitigate overfitting. Among the numerous ensemble methods applied to biological data, bagging[37], boosting[38], and Random Forests (RFs) stand out as the most popular. RFs are particularly useful for classification and regression tasks. They employ a bagging approach to generate a collection of DTs using a random subset of data.[39] The outputs from all these decisions are then combined to form the final predictive model. RFs have proven effective in various bioinformatic applications, including gene expression classifica-

tion, mass spectrum protein expression analysis, biomarker discovery, and statistical genetics.

Moreover, the emergence of DL has further enhanced the significance of ML in scientific disciplines. DL's ability to autonomously perform feature engineering sets it apart. It scans the data to identify relevant features and combines them, enabling faster learning processes, both parallel and distributed, without explicit instructions. Let's see in detail some of the biological fields of application of ML.

- **Omics**: a field of study encompassing various disciplines, aims to characterize and quantify biological molecule pools. This comprehensive approach enables scientists to unravel the intricate structure, functions, and dynamics of an organism. Bioinformatics has seen a surge in the use of individual omics and integrated profiles of multiple omics, such as the genome, epigenome, transcriptome, proteome, metabolome, antibodyome, and others. This is due to the advent of next-generation sequencing technology, which enables the acquisition of vast amounts of omics data. AI algorithms, capable of analyzing sequences, position-specific scoring matrices (PSSMs), and biological, physicochemical, and structural properties, are often employed as inputs. These algorithms enhance the interpretability of omics data. Omics technologies have the potential to revolutionize medicine by shifting from traditional symptom-based diagnosis and treatment towards individualized disease prevention and early diagnostics, as highlighted by Chen *et al.* (2013). ML methods have been applied to various genomics problems, including whole genome sequencing (Venter *et al.*, 2001), gene and Ribonucleic Acid (RNA) structure identification[40], gene identification[41], and gene-gene interaction identification[42] in genetic diseases. GAs have been utilized for DNA fragment assembly.[43] The introduction of ML into proteomics has been exemplified most prominently by AlphaFold, the DL framework developed by DeepMind for protein structure prediction. Historically, high-

resolution structural determination relied on experimental approaches such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, or cryo-electron microscopy (cryo-EM), each of which is laborious, time-intensive, and limited in scalability. AlphaFold2's performance in the 14th Critical Assessment of protein Structure Prediction (CASP14) demonstrated that deep neural networks can predict protein three-dimensional conformations from primary amino acid sequences with accuracy comparable to experimental methods.[44] The subsequent release of the AlphaFold Protein Structure Database[45] has provided predicted structures for hundreds of millions of proteins, enabling large-scale structural proteomics, functional annotation, and hypothesis generation in drug discovery. Nevertheless, significant challenges remain in applying ML to proteomics: current models often neglect the conformational dynamics of proteins, protein–protein and protein–ligand interactions, and the impact of post-translational modifications, which are central to biological function. Moreover, training data biases, arising from over-representation of structurally tractable proteins in experimental datasets, limit the generalizability of predictions to more complex or disordered proteins. These limitations underscore the need for integrative approaches that combine ML-driven predictions with experimental and systems-level data to fully harness ML for proteomic and therapeutic applications.

- **Drug Discovery**: Over the past decade, drug discovery and development have been radically transformed by advances in AI. Recent implementations extend beyond VS, retrosynthesis, and de novo design to include graph neural networks (GNNs), transformers, generative models, and self-/multi-task learning, which are now routinely applied across many stages of the discovery pipeline. For example, AI is used for VS and ligand and structure-based property predictions; for retrosynthesis and reaction prediction; and for generating new protein folds or novel small

molecules with optimized properties.[2] Quantitative structure-activity relationships (QSAR) methods have also evolved: in addition to classical tools like SVMs and ANNs, newer DL variants, GNNs, and transfer learning approaches are being used to improve predictions of biological activity, ADMET properties, and off-target effects.[1] In drug development, AI helps to address critical bottlenecks. Clinical trials remain long, costly, and risky: late-stage failure not only wastes the investment in the trial but also the preclinical work, often amounting to losses in the order of hundreds of millions to over a billion dollars for each failed drug. Key failure points include poor patient selection and recruitment, suboptimal trial design, and insufficient infrastructure to deal with the complexity of multi-site and multi-parameter trials. AI methods now play active roles in optimizing patient stratification, predicting which patients are likely to respond, improving inclusion/exclusion criteria, and simulating trial designs to reduce risks.[46] Recent work has also been propelled by better access to large and high-quality datasets (for example, AlphaFold's predicted structures), advances in computational power, and methodological improvements in model interpretability, uncertainty quantification, and hybrid methods integrating physics-based simulations with ML.[1]

While AI models must be robust and capable of accurately processing data, their performance largely depends on the quality of the input data. When biological datasets are "dirty," containing noise, experimental errors, misannotations, or inconsistencies from non-standard methods, the predictive accuracy of classifiers significantly drops. To address this issue, ML approaches should be designed to handle imperfect datasets, minimizing overfitting while still extracting meaningful patterns. Ensuring high data quality requires careful curation and expert oversight at both the input and output stages to maintain the reliability of results.

As ML in scientific research progresses, random practices such as arbitrary data

splitting, unstructured parameter tuning, or simplistic handling of missing values should be replaced with principled, statistically sound methods. These approaches are crucial for producing results that are interpretable, reproducible, and applicable to real-world biomedical problems.

ML is rapidly transforming the landscape of biomedical research, providing powerful methods to analyze complex datasets, extract hidden patterns, and generate predictive models that support decision-making throughout the drug discovery pipeline. By addressing challenges such as data heterogeneity, noise, and the integration of multimodal biological information, ML establishes a foundation for more accurate target identification, drug design, and patient stratification. These advancements are particularly pertinent in the era of precision medicine, where the overarching goal is to deliver the most appropriate therapy to the most suitable patient at the most opportune time. Among emerging therapeutic strategies, ADCs outline the long-standing vision of Paul Ehrlich's "magic bullet": the idea of a therapeutic agent that selectively targets diseased cells while leaving healthy tissue unharmed.[47] ADCs realize this principle by combining the high specificity of monoclonal antibody (mAb)s with the potent cytotoxicity of small-molecule drugs, enabling precise delivery of therapeutics to pathological cells. This targeted approach not only enhances efficacy but also reduces systemic toxicity, addressing one of the central challenges in oncology and beyond.

1.3 Antibody-drug conjugates

As of 2025, 15 ADCs have received approval from the U.S. Food and Drug Administration (FDA), marking a significant milestone in targeted therapy development.[48] Most of these ADCs are indicated for the treatment of hematologic malignancies and solid tumors, including breast, lung, bladder, and lymphoid cancers, where conventional chemotherapies often fall short due to systemic toxicity and limited specificity. Their clinical success

highlights the therapeutic potential of coupling the precision of monoclonal antibodies with the potency of small-molecule cytotoxins.[49]

- **Mylotarg**® developed by Pfizer, Gemtuzumab ozogamicin is the world's first ADC approved for marketing. It is used for the treatment of CD33-positive acute myeloid leukemia (AML) in adults and children over 2 years old. It is composed of a humanized mAb targeting the CD33 antigen on AML cells, a potent cytotoxic agent, N-acetyl-γ-calicheamicin, and a cleavable hydrazone linker, which releases the drug inside cancer cells.[50]

- **Adcetris**® Brentuximab vedotin was developed by Seagen Inc. in collaboration with Takeda Pharmaceutical Company. It is the second ADC approved for cancer treatment, targeting CD30-positive lymphomas such as Hodgkin's lymphoma (HL), systemic anaplastic large cell lymphoma (sALCL), and relapsed/refractory diffuse large B-cell lymphoma (DLBCL). It is composed of a chimeric mAb (cAC10) targeting CD30, a protease-cleavable dipeptide linker (mc-VC-PABC), and MMAE as payload, a microtubule-disrupting agent.[51]

- **Kadcyla**® ado-trastuzumab emtansine, also known as T-DM1, developed by Genentech (a member of the Roche Group) in collaboration with ImmunoGen, is an ADC approved for the treatment of HER2-positive breast cancer. It is composed of Trastuzumab, a humanized anti-HER2 IgG1 mAb, a non-cleavable thioether linker, and the cytotoxic agent DM1 (a maytansine derivative).[52]

- **Besponsa**® Inotuzumab ozogamicin is an ADC designed to treat CD22-positive B-cell precursor acute lymphoblastic leukemia (BCP-ALL), both in adults and pediatric patients with relapsed or refractory disease. It is formed by a humanized IgG4 mAb targeting CD22, an acid-labile hydrazone linker, and a DNA-damaging agent as payload.[53]

- **Lumoxiti**® Moxetumomab pasudotox was developed by AstraZeneca and represents the first immunotoxin approved specifically for the treatment of relapsed or refractory hairy cell leukemia (HCL) in adults. It is a fusion protein-based therapeutic that targets CD22, a cell surface marker consistently expressed on malignant B-cells in HCL. It is composed of a recombinant single-chain variable fragment (scFv) antibody targeting CD22, a 38 kDa truncated fragment of Pseudomonas exotoxin A (Pseudomonas exotoxin (PE38)), and a linker mc-VC-PABC. Moxetumomab pasudotox was voluntarily withdrawn from the U.S. market for commercial reasons.[54]

- **Polivy**® Polatuzumab vedotin was developed by Genentech (a member of the Roche Group). It is an ADC specifically designed to target CD79b, a B-cell-specific antigen highly expressed in most B-cell non-Hodgkin lymphomas, including DLBCL. Polivy® represents a major advancement in the treatment of relapsed or refractory DLBCL, particularly for patients who are not eligible for stem cells transplant. It is formed by a humanized mAb targeting CD79b, a protease-cleavable linker (mc-VC-PABC), and MMAE as payload.[55]

- **Padcev**® Enfortumab vedotin is the first ADC approved for the treatment of urothelial cancer, specifically targeting Nectin-4, a cell adhesion molecule highly expressed in over 97% of urothelial carcinomas and several other solid tumors. Padcev® offers a novel therapeutic approach for patients with locally advanced or metastatic urothelial cancer, especially those who have progressed following platinum-based chemotherapy and immune checkpoint inhibitors. Enfortumab vedotin is composed of a fully human mAb targeting Nectin-4, an mc-VC-PABC linker, and MMAE as payload.[55]

- **Enhertu**® Trastuzumab deruxtecan, is a next-generation HER2-targeted ADC co-developed by AstraZeneca and Daiichi Sankyo. Representing a major milestone in

ADC innovation, Enhertu®is indicated for a wide range of HER2-expressing cancers, including breast cancer, gastric cancer, non-small cell lung cancer (NSCLC), and other HER2-positive solid tumors. It is formed by Trastuzumab, a cleavable tetrapeptide linker, and a potent topoisomerase 1 (TOP1) inhibitor payload (a derivative of exatecan).[55]

- **Trodelvy**® Sacituzumab govitecan is the first FDA-approved ADC targeting Trop-2, a transmembrane glycoprotein highly expressed in a variety of epithelial tumors. It is currently approved for the treatment of adult patients with metastatic triple-negative breast cancer (mTNBC) and HR+/HER2- metastatic breast cancer who have received prior systemic therapies. It is composed of a humanized mAb targeting Trop-2, a cleavable, pH-sensitive CL2A linker, and SN-38, a potent TOP1 inhibitor and active metabolite of irinotecan.[55]

- **Blenrep**® Belantamab mafodotin was developed by GlaxoSmithKline (GSK) and is the first ADC approved globally to target B-cell maturation antigen (BCMA). It was granted accelerated approval by the FDA in August 2020 for the treatment of adult patients with relapsed or refractory multiple myeloma (R/R MM) who had received at least four prior therapies, including a proteasome inhibitor, an immunomodulatory agent, and a CD38 mAb. Belantamab mafodotin is composed of a humanized IgG1 mAb targeting BCMA, a non-cleavable Maleimidocaproyl (MC) linker, and MMAF, a potent microtubule inhibitor. It was voluntarily withdrawn from the U.S. market in December 2022.[55]

- **Zynlonta**® Loncastuximab tesirine is an innovative ADC developed by ADC Therapeutics. It is the first and only CD19-targeting ADC approved specifically for the treatment of relapsed or refractory large B-cell lymphoma (LBCL). The drug was granted accelerated approval by the U.S. FDA in April 2021 and received conditional approval from the European Medicines Agency (EMA) in December 2022. It is

composed of a humanized mAb directed against CD19, a protease-cleavable valine-alanine (Val-Ala) dipeptide linker, and SG3199, a synthetic pyrrolobenzodiazepine (Pyrrolobenzodiazepine dimers (PBD)) dimer cytotoxic payload.[55]

- **Tivdak**® Tisotumab vedotin is the first FDA-approved ADC specifically for adult patients with recurrent or metastatic cervical cancer whose disease has progressed on or after chemotherapy. This breakthrough therapy leverages a unique mechanism of action by targeting tissue factor (TF), a protein highly overexpressed in cervical cancer cells compared to normal tissue. It is formed by a fully human mAb that targets tissue factor (TF/CD142), an mc-VC-PABC linker, and MMAE as payload.[56]

- **Elahere**® Mirvetuximab soravtansine is a first-in-class ADC approved for the treatment of adult patients with folate receptor alpha (FRα)-positive, platinum-resistant epithelial ovarian, fallopian tube, or primary peritoneal cancer who have received one to three prior systemic treatment regimens. It is composed of a humanized anti-FRα mAb (IgG1 subtype, M9346A), a cleavable linker, and DM4, a maytansinoid microtubule inhibitor.[57]

- **Datroway**® Datopotamab deruxtecan is an innovative ADC co-developed by Daiichi Sankyo and AstraZeneca. It is designed to target Trophoblast cell surface antigen 2 (TROP2)-expressing tumors, including hormone receptor (HR)-positive, HER2-negative breast cancer and EGFR-mutant NSCLC. It is composed of a humanized IgG1 mAb targeting TROP2, a tetrapeptide-based cleavable linker, and a topoisomerase I inhibitor payload (DXd, a derivative of exatecan).[58]

- **Emrelis**® In May 2025. AbbVie announced that telisotuzumab vedotin-tllv has been granted accelerated approval by the U.S. FDA for the treatment of adult patients with locally advanced or metastatic, non-squamous NSCLC with high c-Met protein overexpression who have received a prior systemic therapy. It is formed

by a humanized mAb targeting c-Met, a receptor tyrosine kinase overexpressed in various solid tumors, including NSCLC. A protease-cleavable linker and MMAE as payload.[59]

ADCs combine the targeting properties of the mAb moiety with the potency of cytotoxic agents thanks to their three key components: a mAb, a linker, and a payload. The mAb is responsible for the high selectivity of ADCs, targeting only a specific antigen overexpressed on the tumor cell. The linker should resist premature degradation, avoid off-target effects, and fulfill a controlled release once ADCs reach the tumor microenvironment. Payloads are selected for their high potency, which is necessary to effectively kill cancer cells at the low concentrations delivered by ADCs.[60] From this point of view, the linker is not just a connection between the antibody and the payload, but plays a crucial role in the efficacy of the final ADC. A poorly designed linker can cause side effects in healthy tissue. Therefore, it must be stable in blood at physiological pH and, at the same time, easily and rapidly cleaved once inside the target cells. This way, the cytotoxic drug is active only when released, while it remains inactive, just like a prodrug, when connected to the mAb. Designing a successful new ADC is a complex process that involves various disciplines, including organic chemistry, biochemistry, and medicine.[61]

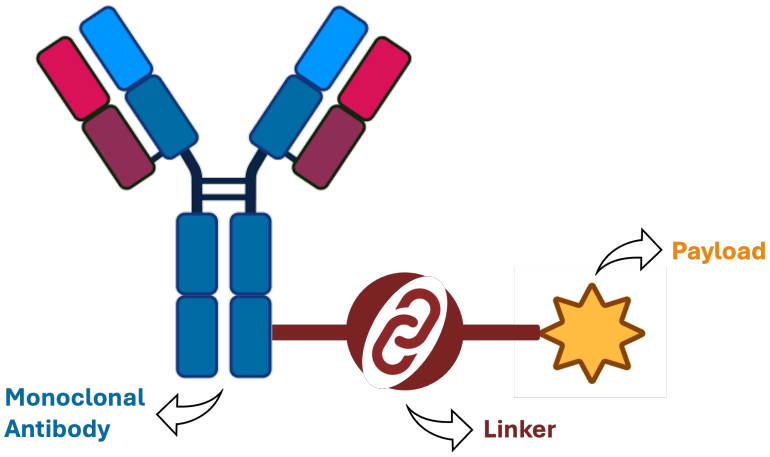


Figure 1.3: Generic structure of ADCs.

For ADCs to achieve therapeutic efficacy, they must be designed to undergo efficient cellular internalization, ensuring that the cytotoxic payload is released within the target cell. This process begins with the binding of the mAb moiety to its specific antigen on the cell surface, which triggers receptor-mediated endocytosis. The antigen-antibody complex is internalized into endocytic vesicles that mature into endosomes and subsequently fuse with lysosomes. Within these intracellular compartments, acidic pH, high protease activity, and a reducing environment promote the degradation of the mAb backbone and cleavage of the linker, culminating in the controlled release of the cytotoxic payload into the cytoplasm.[62] The design of the linker is central to this process: acid-labile linkers exploit the acidic lysosomal milieu, protease-cleavable linkers (e.g., valine-citrulline (Val-Cit)) rely on enzymatic activity, and disulfide linkers are sensitive to the reducing intracellular environment.[63] The choice of linker chemistry must balance stability in circulation with efficient intracellular release, as premature cleavage can result in systemic toxicity, whereas inefficient cleavage reduces potency. Once released, the drug engages its intracellular target, commonly DNA or microtubules, leading to selective tumor cell death.[64] Thus, the sequential steps of antigen recognition, internalization, intracellular trafficking, and

context-dependent linker cleavage together define the pharmacological precision of ADCs and remain a critical focus in their rational design. As represented in Figure 1.4 ADCs exert their therapeutic effects through a multifaceted mechanism that extends beyond simple targeted drug delivery, leveraging both the highly potent cytotoxic payload and the intrinsic properties of the antibody component.

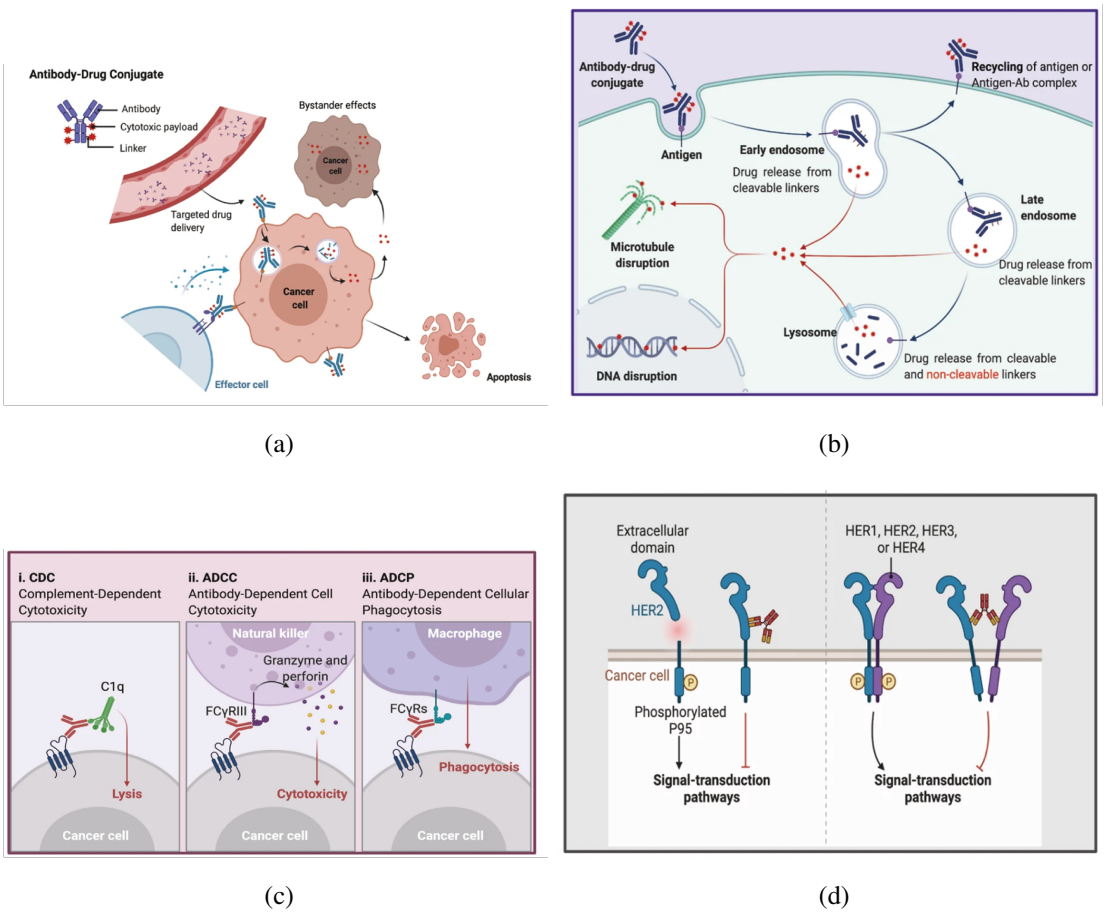


Figure 1.4: Multifaceted mechanism of action of ADCs.[65] (a,b) main core mechanism of action of ADCs; (c) mAb component engages with immune effector cells to elicit antitumor immunity, including complement-dependent cytotoxicity (CDC), antibody-dependent cellular cytotoxicity (ADCC), antibody-dependent cellular phagocytosis (ADCP) effects; (d) mAb retains its activity profile to interfere with target functions, dampen downstream signaling to inhibit tumor growth.

Specifically, the primary mechanism involves the ADC binding to a specific target antigen on the tumor cell surface, followed by internalization and subsequent release of

the cytotoxic payload within the cell, leading to cell death.[66] Beyond this core function, ADCs can exert antitumor effects through mechanisms inherent to their antibody component. These include ADCC, ADCP, and CDC.[67,68] In these processes, the fragment antigen-binding (Fab) region of the antibody binds specifically to antigens on the surface of tumor or virus-infected cells, while the fragment crystallizable (Fc) region interacts with Fc receptors (FcRs) on immune effector cells such as natutal killer (NK) cells and macrophages. This engagement recruits the immune system to directly eliminate target cells. Moreover, the antibody moiety of certain ADCs can modulate signaling pathways by blocking receptor ligand interactions. For example, trastuzumab, the antibody component of trastuzumab emtansine (T-DM1), binds to the extracellular domain of the type 2 human epidermal growth factor receptor (HER2) and prevents heterodimerization with other members of the HER family (HER1, HER3, and HER4). This blockade disrupts downstream signaling through key pathways such as phosphatidylinositol 3-kinase (PI3K/AKT) and mitogen-activated protein kinase (MAPK), thereby inhibiting cell survival and proliferation and ultimately promoting apoptosis.[69] These dual mechanisms, payload delivery and antibody-mediated immune or signaling effects, highlight the multifaceted therapeutic potential of ADCs and contribute to their clinical efficacy.

1.3.1 Antibodies

The discovery of the technology for generating mAbs by Georges Köhler and César Milstein in the mid-1970s marked a turning point in biomedical research.[70] This breakthrough initiated an intensive effort to develop novel mAbs, facilitating the identification of new antigens for both diagnostic and therapeutic applications. In the early 1980s, mAbs were primarily employed as molecular markers for tumor localization and disease detection, while by the mid-1980s, their potential as therapeutic agents began to be realized.[71] Over decades, mAbs have evolved into indispensable tools across biomedical disciplines, serv-

ing crucial roles in research, diagnostics, and therapy. Their clinical utility now spans a wide range of diseases, including cancer, inflammatory, antimicrobial, and autoimmune disorders, and other conditions such as migraine, underscoring their versatility and central role in modern medicine.[72–74]

Thanks to their combination of distinctive properties, mAbs represent the ideal targeting for ADCs. mAbs are capable of high antigen specificity and affinity, which minimizes off-target binding, and their Fc domains contribute favorable pharmacokinetics (long serum half-life) and reduced immunogenicity when humanized or fully human. Among the various immunoglobulin G (IgG) subclasses, IgG1 is most commonly employed in clinically approved ADCs due to its balanced effector activity and structural robustness. A notable example is Trastuzumab, an IgG1 targeting HER2 and being the antibody component of Kadcyla®and Enhertu®.

1.3.2 Linkers

The design of linkers is one of the most challenging aspects of ADC development. As we discussed earlier, linkers connect the payload to the mAb. Despite their seemingly simple role as connectors, linkers have a profound impact on ADC pharmacokinetic and pharmacodynamic properties, as well as the therapeutic window.[75] The linker must be sufficiently stable in plasma to allow ADC molecules to travel through the bloodstream and reach the tumor location without cleaving too soon. Premature release of the toxic payload and undesirable harm to healthy cells that are not the target are caused by linker instability, which can have negative consequences and systemic toxicity. For every combination of antigen, target tumor type, and payload, it is crucial to find ADC linkers with the best linker stability.[76] As the ADC is internalized into the target tumor cell, the linker must simultaneously have the capacity to cleave quickly and release the free drug. Hydrophobicity is another characteristic that should be taken into account while designing the

linker. Aggregation of ADC molecules is frequently facilitated by hydrophobic linkers in conjunction with hydrophobic payloads. For instance, using a multi-loading, hydrophobic dipeptide linker, King and colleagues saw non-covalent dimerization of the mAb BR96 conjugated with doxorubicin.[77]

The search for therapeutically beneficial ADCs is hampered by these molecules; aggregated proteins are often quickly sequestered in the liver and removed by the reticuloendothelial system, which can lead to hepatotoxicity.[78] Furthermore, it is probable that aggregated proteins will act as immunogenic substances, causing an undesirable immune response while in the bloodstream. Hydrophilic linkers with negatively charged pyrophosphate diester groups, polyethylene glycol (polyethylene glycol (PEG)) groups, or sulfonate groups can be used to solve this issue.[79–81]

Among the various types of linkers, a canonical distinction is made between "cleavable" linkers, which possess a specific group that becomes sensitive to a certain stimulus within the tumor microenvironment, and "uncleavable" linkers, which necessitate the complete degradation of lysosomes before the payload can be released (Table 1.1).

Linker type	**Cleavage mechanism**	**Chemistry of linker**
Cleavable	Chemically cleavable	pH sensitive
		Reduction sensitive
	Enzymatically cleavable	Peptide-based
		β-glucuronide based
		Phosphate-based
Non-cleavable	Not applicable	Thioether
		Maleimido caproyl

Table 1.1: Table to test captions and labels.

1.3.2.1 Cleavable linkers

Cleavable linkers represent the major class of ADCs linkers.[75] The purpose of cleavable linkers is to be cut up by certain lysosomal enzymes or by an environmental difference between the external and intracellular environments (e.g., pH, redox potential, etc.). This class of linkers is often constructed to release parental payload molecules following bond cleavage. By using known pharmacological properties of the free payload, researchers can estimate the conjugated payload's lethal potential using traceless drug release methods.[82]

- **pH-sensitive linkers:** cleavable linkers with an acidic group, such as the hydrazone group, are cleaved at the low pH value of the lysosome, releasing the drug. This strategy utilizes an environment with a lower pH, such as the endosome (pH = 5–6) and lysosome (pH = 4.8) compartments, compared to the cytoplasm (pH = 7.4).[83] Gemtuzumab ozogamicin and Inotuzumab ozogamicin are pertinent examples of

ADCs that involve a hydrazone linker.[84] Inotuzumab ozogamicin showed high stability in both human plasma and serum with a rate of hydrolysis of 1.5–2%/day over 4 days.[85]

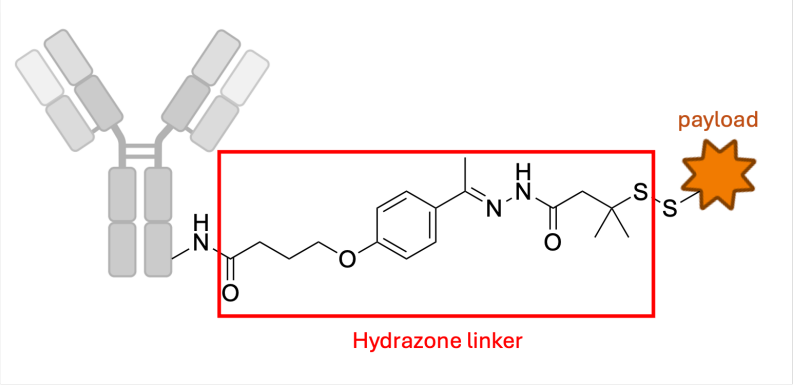


Figure 1.5: Structure of hydrazone linker, currently involved in FDA-approved ADCs Mylotarg® and Besponsa®.

Related work from our laboratory has contributed to this area, expanding the pH-sensitive linkers landscape in ADCs. In that study, our research group reported the synthesis and evaluation of novel orthoester-based linkers with specific release kinetics, providing valuable insights into linker design strategies.[86]

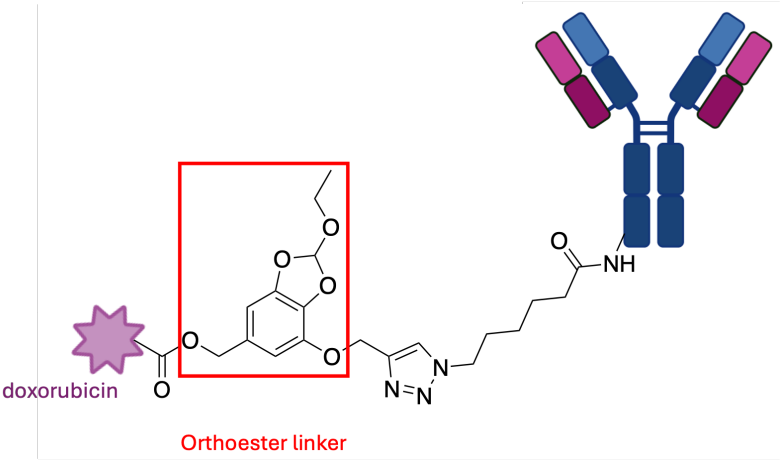


Figure 1.6: Structure of pH-sensitive orthoester linker.

- **Reduction sensitive linkers:** typically contain disulfide bonds, which remain stable

under the relatively oxidizing conditions of systemic circulation but are selectively cleaved in the reducing milieu of the cytoplasm or lysosomes of tumor cells, where high concentrations of glutathione (GSH) and other thiol-containing molecules are present.[65] The stability of disulfide linkers can be fine-tuned by modifying steric hindrance around the disulfide bond or introducing electron-withdrawing substituents to adjust cleavage rates.[87] The combination of disulfide linkers with the previously described hydrazone has led to important therapeutic uses in the creation of Pfizer's Mylotarg®and Besponsa®.

- **Peptide-based linkers:** the most popular linkers in ADC design, sometimes referred to as lysosomal protease-sensitive linkers. Examples of these are Val-Cit, phenylalanine-lysine (Phe-Lys), and Val-Ala dipeptide linkers. This tactic makes use of intracellular proteases, like Cathepsin B, which identify and break a dipeptide link, causing the cytotoxic drugs to be released.[88] Peptide-based linkers exhibit higher systemic stability and rapid enzymatic release of the payload in the target cell as a result of inappropriate pH conditions and serum protease inhibitors.[89] FDA-approved Brentuximab vedotin is a good example of an ADC in which a Val-Cit linker has been explored with marked clinical success.

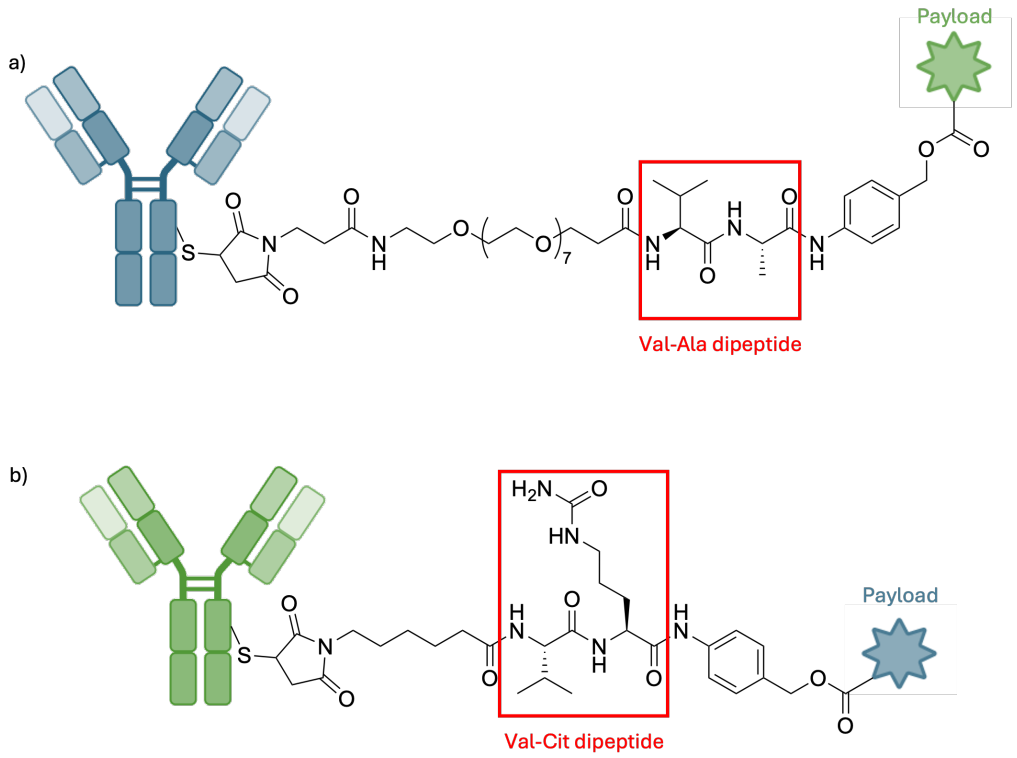


Figure 1.7: Example structures of two commercially available peptide linkers. a) Val-Ala-PAB linker used in FDA-approved ADC Zynlonta®, b) Val-Cit-PAB linker used in FDA-approved ADC Adcetris®.

- **β-glucuronide-based linkers:** a class of enzymatically cleavable linkers that exploit the elevated activity of β-glucuronidase in the lysosomes of tumor cells or within the tumor microenvironment.[90] β-Glucuronide linkers have demonstrated particular utility for highly potent payloads that require precise delivery, and their design can be tuned to optimize stability in plasma while ensuring efficient cleavage under intracellular conditions.[91]

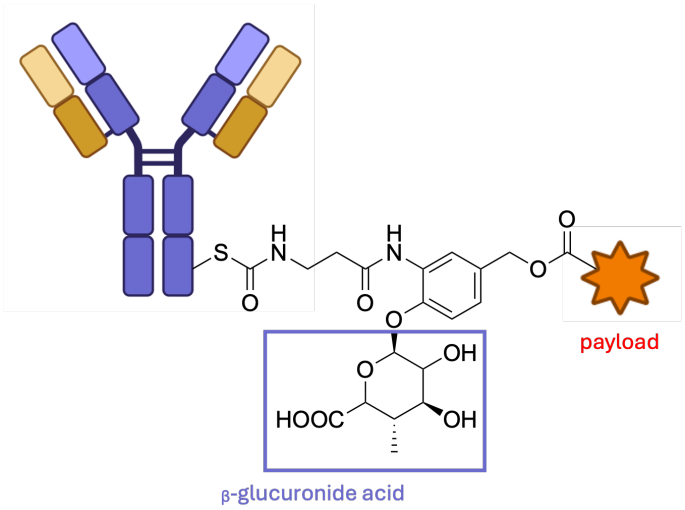


Figure 1.8: Structure of an ADC containing β-glucuronic acid.

- **Phosphate-based linkers:** belonging to another crucial class of enzyme-cleavable linkers targeted by enzymes expressed exclusively in the lysosomal compartment. These linkers are specifically targeted by pyrophosphatase and acid phosphatase enzymes, which catalyze the hydrolysis of pyrophosphates and terminal monophosphates into their respective alcohols. A phosphate-bridged Cathepsin B-sensitive linker was found to be effective in delivering glucocorticoids to tumor cells by Kern and colleagues.[92] They created an aqueous, soluble phosphate drug compound with the Val-Cit cleavable linker by employing the p-aminobenzyl carbamate (PABC) spacer molecule. PABC self-eliminates upon proteolytic cleavage of the dipeptide linker (Val-Cit), allowing phosphatase to hydrolyze the terminal phosphate and liberate the payload.

1.3.2.2 Non-cleavable linkers

MC and thioether are the two categories of non-cleavable linkers. In contrast to their cleavable counterparts, they possess stable linkages that confer enhanced plasma stability and impede proteolytic cleavage. In fact, monomethyl auristatin F (MMAF) drug con-

jugates with non-reducible thioether linkers were found to be more stable than Val-Cit conjugates, and they also preserved their potency.[93] ADCs with this kind of linker rely on the antibody's full lysosomal enzymatic breakdown in order to release payloads during internalization, which causes the linker to detach simultaneously.[94]

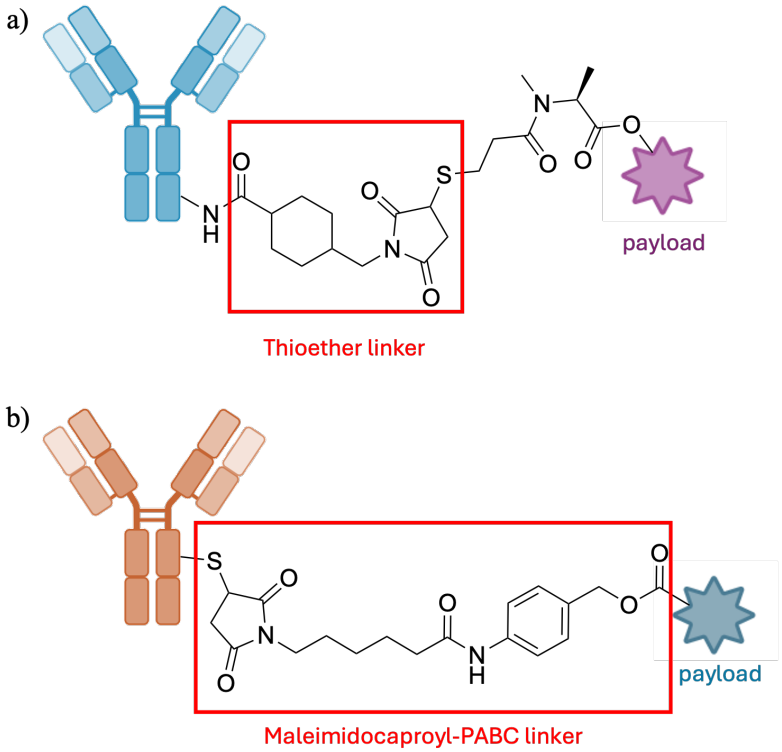


Figure 1.9: Examples of non-cleavable linkers: a) SMCC (N-succinimidyl-4-(maleimidomethyl) cyclohexane-1-carboxylate) linker. b) MC-PABC linker.

1.3.3 Payloads

Although any cytotoxic compound could potentially serve as an ADC payload, only a select subset of molecules possesses the requisite characteristics. Effective ADC payloads must combine high potency with functional groups suitable for stable and site-specific conjugation to the linker, ensuring both efficacy and safety.[95] Cytotoxic agents are highly potent compounds designed to eliminate cancer cells by interfering with critical cellular processes. They typically act either by disrupting microtubule assembly, thereby inhibiting

mitosis, or by binding to the minor groove of DNA, leading to double-strand DNA cleavage and subsequent apoptosis or cell death.[96] For therapeutic applications, cytotoxins must exhibit exceptional plasma stability and potent in vitro activity, with IC_{50} values in the subnanomolar range, since only 1–2% of administered ADCs effectively reach the tumor site.[97] Among the various cytotoxic scaffolds investigated, three main classes of small cytotoxic molecules are currently used in the development of anticancer ADCs (Table 1.2).

Category	**Structure**	**Mechanism of action**
Microtubule inhibitors	MMAE (auristatins)	Inhibits microtubule polymerization
	MMAF (auristatins)	Inhibits microtubule polymerization
	DM1 (maytansinoids)	Inhibits microtubule assembly
	DM4 (maytansinoids)	Inhibits microtubule assembly
DNA damaging agents	Calicheamicin	DNA double-strand breaks via radical generation
	PBD	DNA minor groove crosslinking
	Duocarmycin	DNA alkylation
Topoisomerase inhibitors	SN-38	Topoisomerase I inhibitor
	DXd	Topoisomerase I inhibitor

Table 1.2: Representative small molecular cytotoxic payloads.

1.3.3.1 Microtubule inhibitors

Two major classes of small molecules are part of this category:

- **Auristatins** Monomethyl auristatin E (MMAE) and MMAF, synthetic analogues

of dolastatin 10, are the most significant drugs in this class. Auristatins cause metaphase cell cycle arrest by binding to the tubulin vinca alkaloid binding domain. In comparison to dolastatin 10, the insertion of the alcohol group in MMAE, the carboxylic group in MMAF, and the secondary amine in both compounds improves their hydrophilicity and provides a better pharmacokinetic profile.[98]

- **Maytansinoids** DM1 and DM4 derived from Maytansine, a naturally existing benzoansamacrolide that was extracted from the bark of the African plant Maytenus ovatus. These substances attach to tubulin and prevent the formation of microtubules.[97] The -SH group (more hindered in DM4), which permits conjugation with the linker and regulates the clearance of the conjugates, is the primary distinction between DM1 and DM4 in comparison to maytansine.

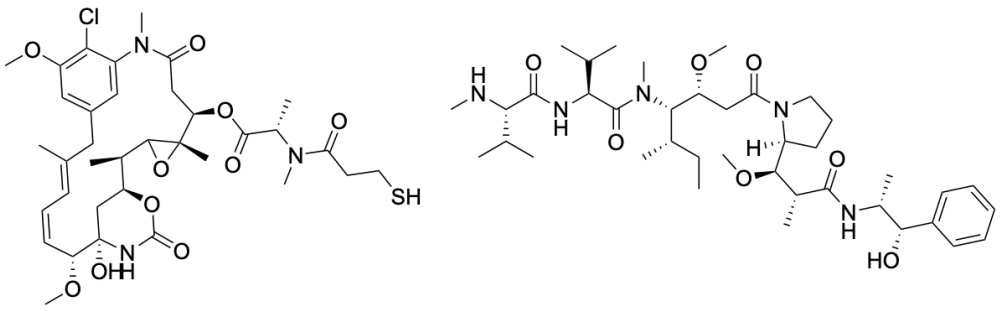


Figure 1.10: Structure of two microtubule inhibitors currently involved in approved ADCs, Kadcyla®(DM1) and Adcetris®(MMAE)

1.3.3.2 DNA damaging agents

DNA-damaging agents are a class of cytotoxic payloads that cause the nucleic acid strands to scission, alkylate, intercalate, or cross-link as a result of DNA binding in the double-helix minor groove. Camptothecin, anthracycline agents, calicheamicin, pyrrolobenzo-

diazepine, and duocarmycin payloads are representative examples of this class. Each of these agents can be designed as either a mono-alkylator or a bis-alkylator. DNA-damaging agents can fall into roughly four mechanistic categories:

- **DNA double-strand breakers** Calicheamicins represent a class of enediyne-containing DNA-cleaving antitumor agents originally discovered by the Lederle Laboratories (American Cyanamid Co.). Functionally similar to anthracyclines, calicheamicins diffuse into the cell nucleus once inside, target and bind to the DNA minor groove, and site-specifically induce double-strand DNA breaks, which ultimately cause rapid cell death via apoptosis. In fact, calicheamicins form reactive diradical species that eventually cause DNA strand cleavage at various locations.[99,100]

- **DNA alkylators** Duocarmycins are potent cytotoxic drugs that alkylate DNA minor grooves and exhibit their high potency by forming DNA adducts. A five-base pair AT-rich sequence that more easily fits the core pyrroloindole subunit is the one that duocarmycins favor. The way duocarmycins work is that they attach to the minor groove of DNA, attack the cyclopropane molecule there to produce a DNA adduct, and then alkylate the adenine at the N3 position. Induced irreversible DNA alkylation impairs the structural integrity and architecture of DNA, causes DNA cleavage, and ultimately results in apoptosis.[99]

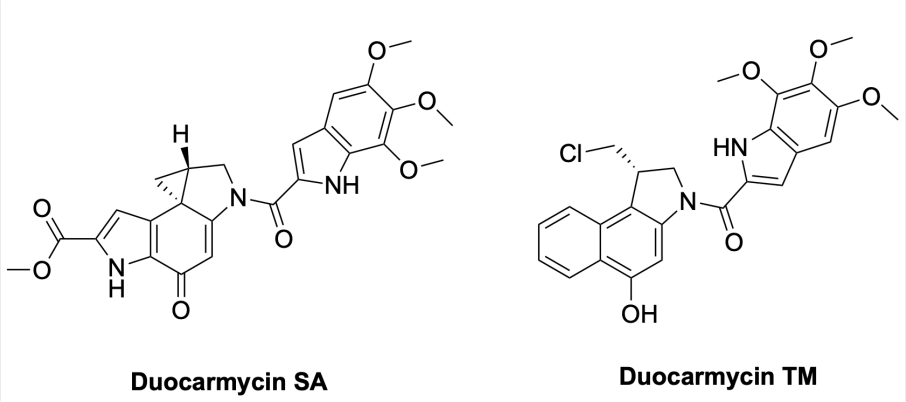


Figure 1.11: Example of two duocarmycin structures.

- **DNA intercalators** Anthracyclines are the most well-known family of chemicals in this area. Actinomycete-derived doxorubicin is an antimitotic anticancer agent that is 14-hydroxylated daunorubicin. It is one of the most effective antitumor chemotherapeutic agents that is frequently used in clinical settings and has been used to treat a variety of solid and non-solid tumors.[101]

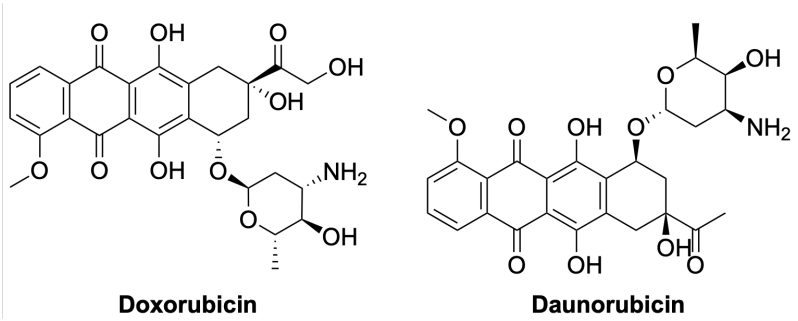


Figure 1.12: The structure of Doxorubicin and Daunorubicin.

- **DNA cross-linkers** Pyrrolobenzodiazepines were first isolated in 1965 by Berger *et al.*[102] PBD dimers cross-link DNA by binding in the minor groove and reacting with interstrand guanine residues at GATC or related sequences.[103] They are also unable to bind to single-stranded DNA (or RNA), representing extremely selective in the requirement of a minor groove structure for covalent binding to duplex or hairpin DNA.[104] Thanks to their selectivity and potency, numerous ADCs based on PBDs are now being studied.

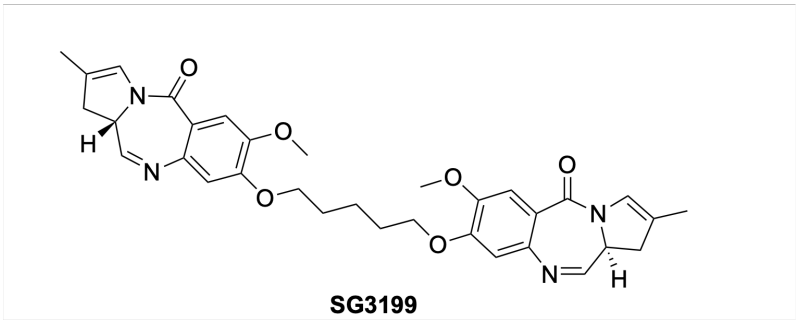
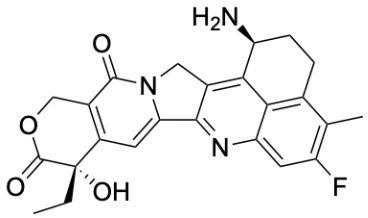


Figure 1.13: SG3199, the released warhead component of the ADC payload Tesirine (SG3249).

Compared with microtubule-targeting agents, this category of cytotoxic payload can kill target cells at any point in their life cycle.

1.3.3.3 Topoisomerase inhibitors

Camptothecin is a TOP1 inhibitor produced by nature.[105] TOP1 inhibitors function by attaching themselves to Topoisomerase cleavage complexes (TOPcc), more precisely to TOP1cc. Camptothecin is categorized as an "interfacial inhibitor" due to its preferential binding at the TOP1–DNA interface. The TOP1cc offers camptothecin a particular binding site at the intersection of TOP1 and DNA, since these types of inhibitors bind to the interface of macromolecules. Camptothecin and its derivatives have been demonstrated to be highly selective for these macromolecular complexes, with TOP1cc being their exclusive target. Consequently, it has been maintained that camptothecin is the perfect pharmacological agent because its therapeutic effect is determined by its selectivity rather than its potency.[106] In recent years, topoisomerase inhibitors have been widely studied, and several ADCs currently in clinical studies involve this type of cytotoxic agent.



Exatecan (DX-8951)

Figure 1.14: DX-8951, the released warhead component of the ADC payload Deruxtecan (MC-GGFG-DXD).

1.3.4 Beyond oncology ADCs

Beyond their established role in oncology, ADCs are increasingly recognized as a versatile platform for the treatment of non-cancer diseases. Numerous drug candidates have

historically failed in development because of limited target specificity and dose-limiting toxicities. The ADC strategy offers a compelling avenue to revisit these molecules, coupling them to selective antibodies to enhance tissue targeting, expand their therapeutic window, and reduce off-target effects.[107]

1.3.4.1 Anti-inflammatory ADCs

Given the unmet need for more potent and selective therapies in autoimmune disorders such as rheumatoid arthritis (RA), ADC-based strategies have recently gained attention as a means to enhance therapeutic efficacy beyond that achievable with mAbs alone, which are sometimes insufficient to treat certain diseases like RA due to their lower activity compared to available cytotoxic compounds.[108] Recently, researchers have investigated the conjugation of corticosteroids (Cs) with mAb. Buttgereit et al. developed a new anti-glucocorticoid receptor (GCR) ADC (ABBV-3373), which is a conjugation of an anti-TNF antibody (adalimumab) and a GCR modulator with the help of Maleimide-Gly-Ala-Ala linker.[109] Another example of anti-inflammatory ADCs is tocilizumab-alendronate (TCZ-ALD), a novel biotherapeutic targeting interleukin-6 (IL-6), crucial for the inflammation and immune response in RA. As explained in their study, Lee *et al.* demonstrated the enhanced activity of the bioconjugate in the treatment of RA.[110]

1.3.4.2 Antibody-antibiotic conjugates

ADCs technology has been repurposed for the treatment of intracellular bacterial infections. A notable example is an antibody-antibiotic conjugates (AACs) developed to target *Staphylococcus aureus*. This construct employs a protease-sensitive linker that is specifically cleaved within the phagolysosome, enabling site-specific release of the antibiotic at the site of infection. In preclinical models, the AAC demonstrated superior efficacy to vancomycin in treating bacteremia.[111]

1.3.4.3 Immunosuppressive ADCs

Similar approaches have been explored in the development of non-hormonal immunosuppressive therapies. Dasatinib, a tyrosine kinase inhibitor initially developed for oncology, is clinically used for the treatment of chronic myelogenous leukemia (CML) and certain forms of acute lymphoblastic leukemia (ALL).[112,113] Although dasatinib exhibits potent immunosuppressive activity, its clinical utility in immune modulation is limited by dose-dependent systemic toxicities, including neutropenia and myelosuppression. To mitigate these adverse effects, researchers have engineered a dasatinib-based ADC targeting CXCR4, a receptor abundantly expressed on T cells, B cells, and monocytes.[114] This targeted delivery strategy effectively restricted drug exposure to immune cells, thereby reducing off-target toxicity and expanding the therapeutic window of dasatinib for potential use in immune-mediated disorders.

These examples demonstrate the versatility of ADCs as a delivery platform that extends beyond the oncology domain. Underscoring the potential therapeutic applications in infectious diseases and chronic inflammatory disorders.

1.3.5 Conventional bioconjugation techniques and DAR

Critical factors in determining the performance of ADCs are bioconjugation sites and types. Bioconjugation strategies influence not only the stability of the conjugate but also its pharmacokinetic profile, efficacy, and safety. Conventional conjugation approaches often modify naturally occurring amino acid residues, such as lysines and cysteines, resulting in heterogeneous mixtures with varying DAR, which indicates the average value of linker-payload (LP) covalently linked to the antibody. While these stochastic methods have facilitated the development of several clinically approved ADCs, the resulting heterogeneity can impact the therapeutic index by altering clearance rates and potency. In contrast, site-specific conjugation techniques, such as engineered cysteine residues,

enzymatic modifications, or non-natural amino acid incorporation, enable more uniform ADCs with defined DAR values.[115] This uniformity improves consistency in drug loading and therapeutic outcomes. The DAR itself is a crucial parameter. Higher DAR values may enhance potency but can compromise solubility, stability, and safety, while lower DAR values may improve tolerability at the expense of efficacy. Therefore, optimizing both the conjugation method and DAR is essential to strike a balance between efficacy, pharmacokinetics, and safety.[116] In ADC development, traditional chemical conjugation approaches utilizing naturally occurring amino acid residues, specifically lysine and cysteine, have been widely employed. These methods exhibit distinct characteristics that affect heterogeneity, DAR, and manufacturing reproducibility.

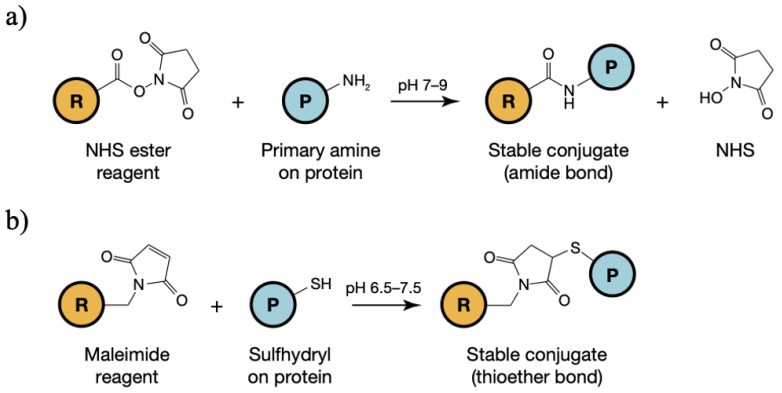


Figure 1.15: Overview of traditional conjugation technology, a) native Lys conjugation: N-hydroxysuccinimide (NHS) ester reacts with primary amines on the antibody to form stable amide bonds at pH 7–9. b) interchainbreak native Cys conjugation: maleimide reacts with thiol groups exposed by reducing disulfide bonds in the antibody, forming stable thioether bonds at pH 6.5–7.5.[117]

Lysine-based conjugation involves the reaction between the ϵ-amino group in lysine residues of an antibody and activated ester groups, such as NHS esters. This reaction results in the formation of stable amide bonds (Figure 1.5a).[118] This approach has been successfully employed in several of the first FDA-approved ADCs, including Mylotarg®, Kadcyla®, and Besponsa®. One advantage of lysine conjugation is its high reaction efficiency, which produces low-molecular-weight by-products that can be readily purified

by tangential flow filtration. However, a disadvantage is the lack of positional uniformity among lysine residues in antibodies. A typical IgG1 antibody contains approximately 80 lysine residues, with only 10 of these being solvent-exposed and primarily available for conjugation.[119] Consequently, this results in a heterogeneous mixture of related species.[120] This high degree of heterogeneity can impact efficacy, safety, and stability. Consequently, lysine-based conjugations pose significant challenges ranging from heterogeneity-related stability issues and therapeutic index to a chemistry, manufacturing, and control perspective.[121] Cysteine-based conjugation involves the selective reduction of interchain disulfide bonds within the antibody to generate free thiol groups, which subsequently react with maleimide-containing linkers.[122]

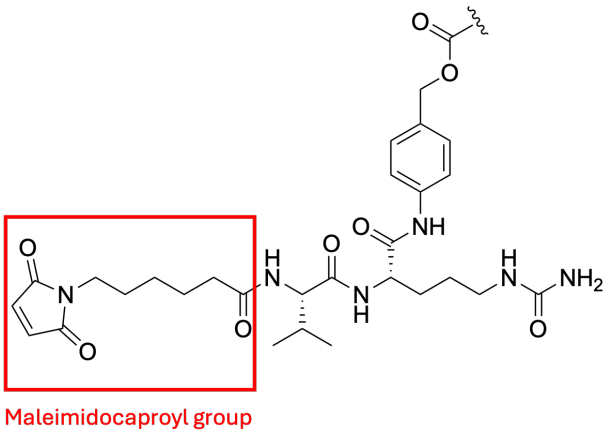


Figure 1.16: Example of a commercially available linker bearing a maleimide moiety.

Compared to lysine-based conjugation, this approach typically yields ADCs with a narrower DAR distribution and reduced heterogeneity, thereby enhancing manufacturing reproducibility. Moreover, cysteine conjugates are more suitable for analysis by high-performance liquid chromatography (HPLC), allowing straightforward assessment of DAR.[123] Nonetheless, the use of reducing agents such as tris(2-carboxyethyl)phosphine (TCEP) can promote antibody aggregation, necessitating careful optimization of process conditions. Another limitation arises from payload loss through the retro-Michael re-

action, which destabilizes the thiol-maleimide linkage; this contrasts with lysine-based conjugation, where payload detachment primarily results from the inherent heterogeneity of conjugation sites.[124] It has long been known that hydrolysis of the five-membered rings of maleimides and their associated thiol adducts significantly changes their chemistry.[125] Ring-opening hydrolysis of maleimides and resultant thiosuccinimide adducts attenuates both thiol addition and elimination.[126] As a result, there is now interest in identifying protein attachment sites that promote spontaneous thiosuccinimide hydrolysis, figuring out how to cause hydrolysis after conjugates have formed, and creating substitute tactics that completely avoid the use of maleimide/thioether conjugation.[127,128] Senter *et al.* present a clever modification of standard maleimide-thiol chemistry, introducing an N-alkylmaleimide functionality (via a diaminopropionic acid (DPR) linker) that promotes the self-hydrolysis of the resulting thiosuccinimide ring under physiological conditions.[129] This intramolecular catalysis differs from conventional maleimide adducts, which are susceptible to retro-Michael cleavage and payload loss in plasma. The hydrolyzed form generated by this intramolecular catalysis is no longer prone to elimination, thus locking in the conjugate more stably. In animal studies, ADCs built with self-hydrolyzing maleimides exhibited improved pharmacokinetics, enhanced antitumor efficacy, and reduced off-target toxicity (e.g., neutropenia) compared to their non-hydrolyzing counterparts. Another possibility to overcome challenges derived from the use of classic maleimide compounds is native disulfide rebridging between the interchain disulfides, which eliminates the dissociation of the ADC and results in homogeneous ADCs without the need for antibody re-engineering.

The classical maleimide reaction, widely used for cysteine conjugation in ADCs development, is often susceptible to instability due to retro-Michael deconjugation and thiol exchange. To overcome these limitations, alternative methods like disulfide rebridging have been explored. In a recent study, Behrens *et al.* proposed a strategy utilizing dibromomaleimide linkers to re-establish reduced disulfide bonds simultaneously with drug

payload delivery.[130] This approach resulted in more homogeneous and stable ADCs. After reduction of the four interchain disulfide bonds, the dibromo-maleimide (DBM) linker rebridges the cysteines to yield mainly DAR = 4 ADC species, improving homogeneity relative to traditional maleimide conjugation. In head-to-head comparisons, the DBM-based ADCs show enhanced pharmacokinetics, better efficacy, and reduced toxicity *in vivo* versus analogous heterogeneous ADCs built with conventional linkers. Crucially, this method retains antigen binding and Fc function while offering an accessible route to more uniform ADCs from unmodified antibodies, making it a valuable addition to the toolkit of antibody bioconjugation methods.

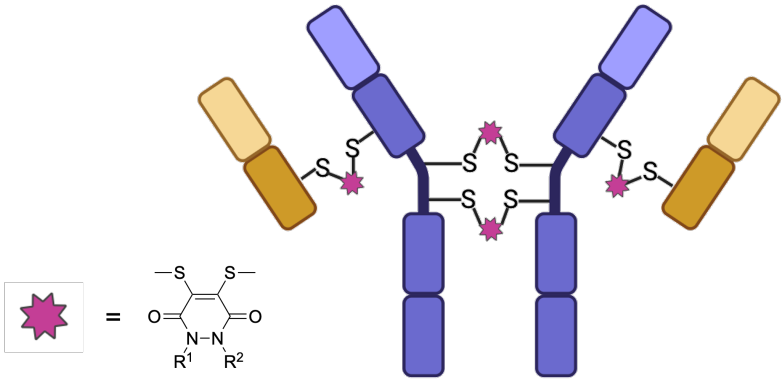


Figure 1.17: Example of a disulfide rebridging conjugation using PD as LP system. Created with BioRender.com

In 2017, Robinson *et al.* introduced the use of pyridazinedione (PD) reagents for disulfide rebridging in antibodies, specifically applying this strategy to trastuzumab to attach the cytotoxic payload MMAE.[131] By reducing the native inter-chain disulfide bonds and then re-bridging them with PD linkers, they obtain highly homogeneous ADCs with a DAR of 4 as the predominant species. Compared to classical maleimide conjugation, the PD-based constructs are more stable in serum, less susceptible to thiol exchange, and preserve antibody integrity. In the 2014 Bioconjugate Chemistry paper by Badescu *et al.*, the authors present a disulfide rebridging strategy for constructing homogeneous ADCs.

Their methodology employs a bis-reactive linker that reacts with both thiols obtained after reduction of an interchain disulfide.[132] They report an overall 78% conversion to the ADC species with a DAR of 4 and no residual unconjugated antibody, and demonstrate that the conjugates retain antigen binding, show serum stability, and exhibit potent, antigen-selective cytotoxicity in vitro and *in vivo*. Moreover, a novel ADC composed of an anti-EphA5 antibody conjugate to MMAE through this technology is currently in Phase 1/1b clinical trials.[133]

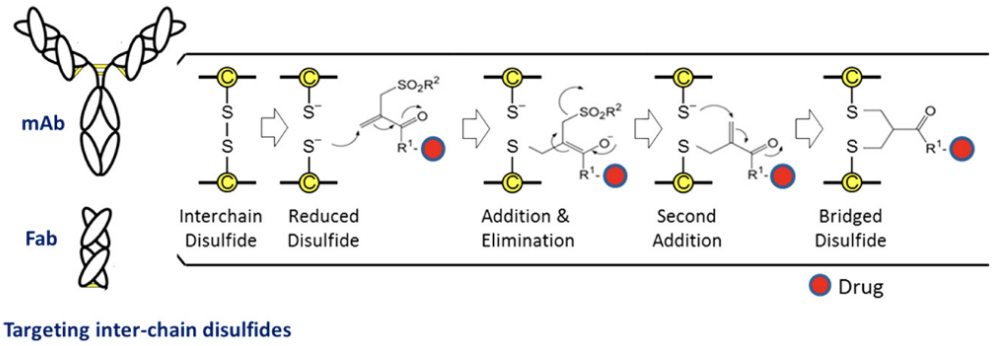


Figure 1.18: Michael addition and elimination reactions are used in a series of bis-alkylating conjugation reactions to conjugate a payload to a disulfide bridge.[132]

1.3.6 Conjugation innovations

Recent advances have driven the development of site-specific conjugation technologies, which enable precise control over DAR and improve ADC homogeneity.[134] These innovations enhance ADC stability, reduce off-target effects, and improve overall therapeutic performance. An early and significant example of site-specific antibody engineering to enhance ADCs therapeutic index is the THIOMAB^{TM} technology, developed by Genentech.[135] This method enables precision conjugation at specified places with determined stoichiometry by genetically introducing reactive cysteine residues into the constant region of the antibody Fab domain. THIOMAB^{TM} provides the benefit of preserving the structural and functional integrity of the antibody while generating a more homogenous ADC

in contrast to traditional conjugation techniques that depend on lysine or cysteine residues. There are multiple phases in the conjugation process used to produce THIOMAB^{TM} ADCs. To repair interchain disulfide bonds and maintain the designed cysteine residues in a reactive state, the inserted cysteine residues in the Fab domain are first activated through a reduction step and then oxidized. Following the production of antibodies, endogenous thiols like glutathione frequently cap the designed cysteines, resulting in mixed disulfides. A reducing agent like TCEP is utilized to selectively reduce these mixed disulfides without breaking native disulfide connections, exposing the free thiol groups needed for conjugation. Finally, only the thiol groups of the engineered cysteine residues react through Michael addition with maleimide-containing compounds under pH conditions of 6.5 - 7.5, enabling highly specific and controlled conjugation without affecting the structural integrity of the antibody. This approach eliminates the stochastic conjugation observed in traditional cysteine-based methods and ensures uniform ADC production.[136]

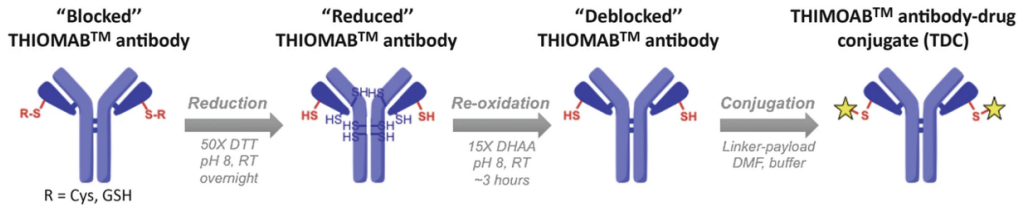


Figure 1.19: A method for "deblocking" a THIOMAB^{TM} antibody so that it can be conjugated to a thiol-reactive linker-drug moiety later on.[136]

The term "click chemistry," which was first used by Sharpless and associates, refers to a group of reactions that result in straightforward, quick, and high-yielding molecules.[137] Click chemistry has been used in the study of biological systems in recent years. The phrase "bioorthogonal chemistry" was therefore created. Bioorthogonal chemistry demands that the reactions produce harmless chemicals that are stable *in vivo* and do not disrupt endogenous biological processes, in addition to the click chemistry requirements. In the field of ADCs, click chemistry has emerged as a powerful site-selective conjugation

technique. By exploiting reactions such as azide-alkyne cycloaddition or train-promoted azide-alkyne cycloadditions (SPAAC), click chemistry enables site-specific attachment of payloads under mild, physiologically compatible conditions.[138] Building on this principle, Synaffix developed the GlycoConnect®technology, which harnesses the conserved N-glycan sites of IgG antibodies as accessible and uniform anchoring points. In this strategy, antibody glycans are enzymatically remodeled to install azide handles, which can then be coupled to drug linkers using click chemistry. The result is a homogeneous ADC population with a well-controlled DAR = 2, improved stability, and preserved antibody functionality.[139]

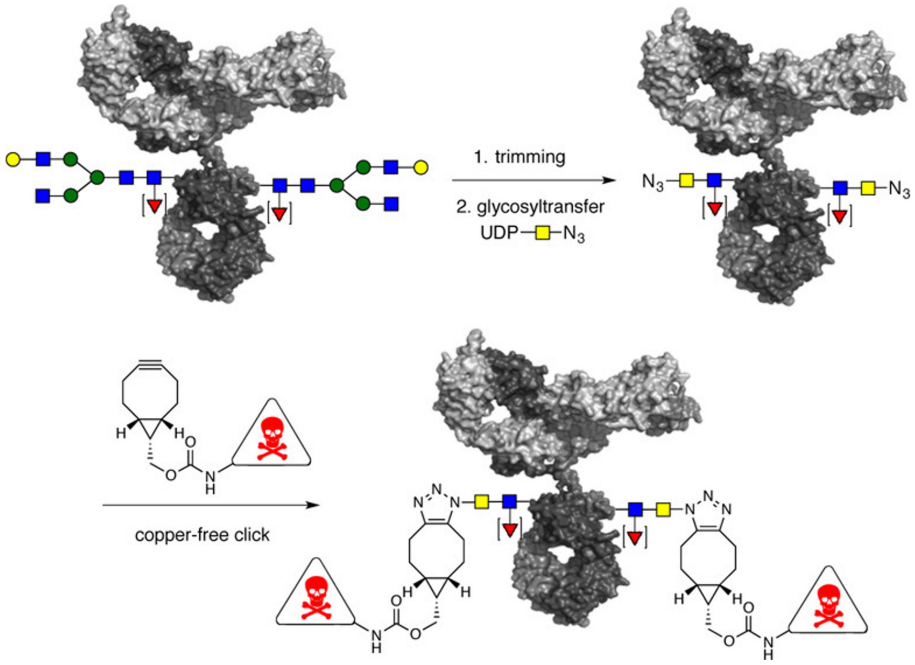


Figure 1.20: Synaffix chemoenzymatic protocol for the click chemistry conjugation of potent cytotoxic payloads, resulting in a DAR = 2 ADCs species.[139]

Pointing to lowering the DAR to improve homogeneity in ADCs, Sander *et al.* extend Synaffix's GlycoConnect platform to create DAR 1 ADCs without the need for antibody reengineering. The authors demonstrate that by adapting GlycoConnect®chemistry, remodeling glycans to install clickable handles followed by payload conjugation, they can

reliably produce homogeneous DAR 1 constructs across a variety of payload chemistries. In comparative spheroid penetration studies, the DAR 1 ADC showed significantly better penetration into tumor spheroids than its DAR 2 counterpart at an equivalent payload dose, underscoring the potential of reduced DAR formats for ultrapotent ADCs.[140] SMARTag®is a next-generation site-specific conjugation technology for ADCs, developed by Redwood Bioscience. By tackling the issues of heterogeneity and instability that come with traditional stochastic conjugation, this platform makes it possible to produce homogeneous and stable ADCs. SMARTag®technology improves the safety and therapeutic efficacy of ADCs by carefully regulating the DAR and guaranteeing a strong pharmacokinetic profile.[141] The SMARTag®technique adds a special conjugation site to the antibody by using the formylglycine-generating enzyme (FGE). The antibody has a genetically encoded six-amino-acid sequence (LCxPxR), which is then transformed into formylglycine (fGly) by FGE. This modification adds an aldehyde functional group at a specific location, enabling precise and regulated drug attachment without changing the general structure or antigen-binding capabilities of the antibody. SMARTag uses Hydrazino-iso-Pictet-Spengler (HIPS) chemistry for payload conjugation, which helps the aldehyde-tagged antibody and a hydrazino-functionalized PL establish a stable carbon-carbon bond. By lowering premature linker cleavage in circulation, this technique greatly increases ADC stability when compared to traditional conjugation techniques like hydrazone or oxime-based connections. Because the carbon-carbon bond is so stable, the medication stays conjugated until it reaches the target cells, reducing off-target toxicity and increasing therapeutic efficacy.[142] The capacity of SMARTag®technology to achieve high site-specificity of the PL and, consequently, specific DAR configurations (DAR 2 or DAR 4) is a significant advantage. This exact conjugation method improves pharmacokinetics and the therapeutic response by lowering batch-to-batch variability. Because of the conjugation site's position, the technique can support a wide variety of PLs, such as peptides, nucleic acids, maytansinoids, and MMAE, making it a flexible platform for ADCs development.

The site-specific AJICAP technology was developed by Ajinomoto to facilitate the manufacture of ADCs without the need for genetic alterations.[143] This technique targets conserved residues in the Fc region and enables the direct use of native IgG antibodies. By conjugating PLs at these specific places, AJICAP makes it easier to create ADCs with consistent DAR, which enhances batch consistency and pharmacokinetic characteristics. The AJICAP procedure utilizes an Fc-affinity peptide reagent that specifically binds to lysine residues (Lys248 and Lys288). This peptide portion is then cleaved by hydroxylamine, revealing the functional groups essential for site-specific PL conjugation.[144] A further implementation, AJICAP-M, refines the process by employing a traceless site-specific peptide reagent that allows a bioorthogonal conjugation through click chemistry.[145]

1.3.7 ADCs Characterization

ADC characterisation is essential for assessing the end product's characteristics. MS, chromatography, and ultrviolet/visibile spectroscopy (UV/Vis) spectroscopy are three widely used methods to evaluate ADCs properties like DAR, site of conjugation, and aggregation rate.

1.3.7.1 UV/Vis spectrometry

In the early development of ADCs, the simplest technique to analyze these complex bioconjugates relies on the UV/Vis spectroscopy. Using the absorbances, path lengths, extinction coefficients, and concentrations of the various components, a sequence of simultaneous equations can be written by measuring the absorbances of a multicomponent system at a number of wavelengths equal to or greater than n. It is possible to solve the simultaneous equations for the concentration of each component in the sample if the extinction coefficients and path length are known.[146] The average DAR in an ADC sample can be calculated using this technique. It requires that:

1. the drug possesses a UV/Vis chromophore
2. the drug and antibody have different absorption maxima in their UV/Vis spectra
3. the loaded drug does not affect the mAb ability to absorb light in the ADC sample and *vice versa*.

If these conditions are satisfied, the ADC sample can be regarded as a two-component mixture, and the relative concentrations of the drug and antibody can be ascertained using the Beer-Lambert law.[147] Therefore, it is possible to compute the average DAR.

1.3.7.2 MS methods

MS has emerged as an indispensable analytical tool in the characterization of ADCs, owing to its exceptional sensitivity, resolution, and ability to provide detailed molecular information. MS-based techniques enable comprehensive evaluation of DAR, conjugation site distribution, linker stability, and payload integrity throughout manufacturing and storage. Both intact mass analysis and middle-down or bottom-up MS approaches could be applied for ADCs characterization.

- **Matrix-assisted laser desorption/ionization-time-of-flight (MALDI-TOF MS) MS:** MALDI-TOF MS MS is a method that offers high mass accuracy. It has been successfully used for molecular weight measurements of proteins.[148–151] MALDI-TOF MS is one of the mass spectrometric methods that requires minimal sample preparation and does not necessitate a chromatographic step. Consequently, it is widely used for the rapid analysis of relatively complex samples, often delivering results within a few minutes. However, very high hydrophobic payloads and high DAR ADCs tend to reduce ionization, making the analysis more complex.[120,152] In addition, MALDI-TOF MS was successfully used in several works conducted

in our research group, highlighting the robustness of this method in bioconjugate analysis.[153,154]

- **ESI-MS:** Electrospray ionization (ESI) MS is a soft ionization technique in which the sample is first dissolved in a solvent (usually water mixed with organic solvent and a small amount of acid or base) and introduced through a fine, charged needle. A fine mist of charged droplets is produced by applying a high voltage, usually 3-5 kV. The solvent slowly evaporates as these droplets go through a heated capillary in the direction of the mass spectrometer, resulting in the droplets getting smaller and the charge density rising. The droplets eventually experience Coulombic fission, which releases gas-phase ions, which frequently have multiple charges in the case of big molecules. After that, the mass-to-charge (m/z) ratios of these ions are determined in the mass analyzer. Because ESI-MS produces multiply charged ions, the detection of large biomolecules like intact antibodies or ADCs can be challenging, due to mass range and resolving power exceeding.[155,156] ESI-MS methods are generally coupled with chromatography techniques to enhance the performance of the analysis and to get more detailed information about the analyzed sample.

1.3.7.3 Chromatographic methods

Chromatographic techniques play a crucial role in the characterization and quality assessment of ADCs, providing information on drug distribution, aggregation, and product heterogeneity. Among these, hydrophobic interaction chromatography (HIC), size exclusion chromatography (SEC), and reversed-phase liquid chromatography (RPLC) are three of the most widely employed methods.

- **HIC:** It separates ADC species based on differences in hydrophobicity arising from variations in drug loading; more conjugated antibodies typically exhibit greater hydrophobicity and elute later, allowing determination of the DAR distribution. The

primary benefit of HIC over other LC techniques is that it is non-denaturing, meaning that the original form of the protein should be preserved. Additionally, the separated proteins can be collected for additional activity assays. Under their physiological-like conformation, HIC may distinguish between the various DARs of natural ADCs, including DAR0, DAR2, DAR4, DAR6, and DAR8 of an IgG1-type ADC.[157,158] However, a major limitation of this technique lies in its strong dependence on payload hydrophobicity. ADCs bearing hydrophilic or moderately hydrophobic drugs often display poor chromatographic resolution, leading to inaccurate DAR determination or co-elution of multiple species.

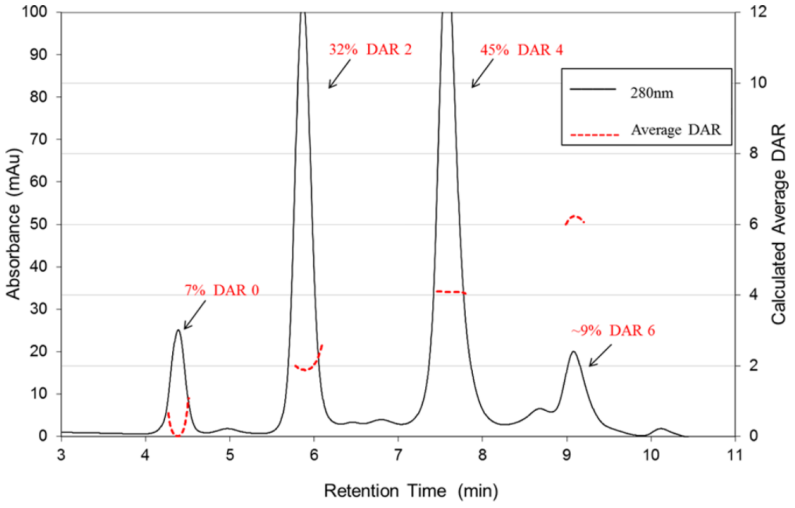


Figure 1.21: HIC chromatogram (280 nm) of an unstressed ADC 1 DAR 3.5 t0 sample with relative percentage areas and average calculated DAR of each major species.[159]

- **SEC:** separates molecules by size (or hydrodynamic volume) under native conditions, enabling detection of aggregates, fragments, or other size variants without altering the ADC's structure.[160]

- **RPLC:** In RPLC, analytes are separated based on hydrophobic interactions with a nonpolar stationary phase under denaturing conditions, typically using gradients of organic solvents such as acetonitrile with an acidic modifier (e.g., formic acid).

This method enables efficient separation of ADC subunits, released drug-linker species, and peptides generated from enzymatic digestion. However, RPLC alone isn't sufficient for DAR determination. In fact, all chromatographic techniques are coupled to MS, allowing precise identification of drug attachment sites, linker structures, and post-translational modifications.[161,162]

1.3.8 Future perspectives

Over the past two decades, advances in bioconjugation chemistry have transformed ADCs. Looking ahead, the design and optimization of ADCs may be redefined by the combination of ML and sophisticated bioconjugation chemistry. While structural control and chemical precision have been brought about by innovations like site-specific conjugation, glycoengineering, and bioorthogonal click reactions, ML can enhance these developments by forecasting pharmacokinetic and safety profiles from extensive experimental datasets, optimizing LP combinations, and predicting conjugation outcomes. Predictive models that direct the judicious selection of conjugation sites, linker chemistries, and payload ratios prior to synthesis could be trained using high-throughput experimental conjugation data in a tandem approach, significantly speeding up development cycles and lowering resource requirements. In addition, the computer-aided design of this type of pharmaceuticals could expand ADCs technology beyond oncology, helping the development of more precise treatments in the field of autoimmune, infectious, and metabolic diseases.

1.3.9 Aim of this research project

Due to their high complexity, ADCs development and manufacturing can be very challenging, and also very expensive and time-consuming. As previously described, the choice of both linker and payload is crucial for determining ADCs characteristics such as PK, stability, potency, etc... As well as a proper conjugation chemistry that allows the correct

linkage of the LP system to the desired antibody. Looking at the huge increase in ML applications in drug discovery and related research and development fields in the last decade, this project aims to develop a new predictive method with the use of an ML model to speed up bioconjugation processes and the development of new ADCs by predicting the conjugation of specific LP systems to a mAb. In particular, this project will exploit the potential of ML algorithms in predicting the DAR in the synthesis of ADCs, evaluating the performance of the tested networks, and identifying the most important features for the model in making a prediction on the bioconjugation outcome. To evaluate the robustness and generalizability of the proposed predictive model, an experimental validation was performed using a holdout dataset of bioconjugation reactions. The resulting insights provide valuable guidance for optimizing future ADC development and highlight the potential of data-driven approaches to streamline and rationalize complex conjugation workflows.

Chapter 2

Results and discussion

This chapter presents and discusses the results obtained from the application of ML models to in-house bioconjugation datasets. The performance of the tested algorithms was evaluated using several statistical metrics, and the influence of the selected features on the model predictions was analyzed. This section aims to assess the predictive capability of the proposed approach and to interpret the results in the context of ADC synthesis and conjugation efficiency.

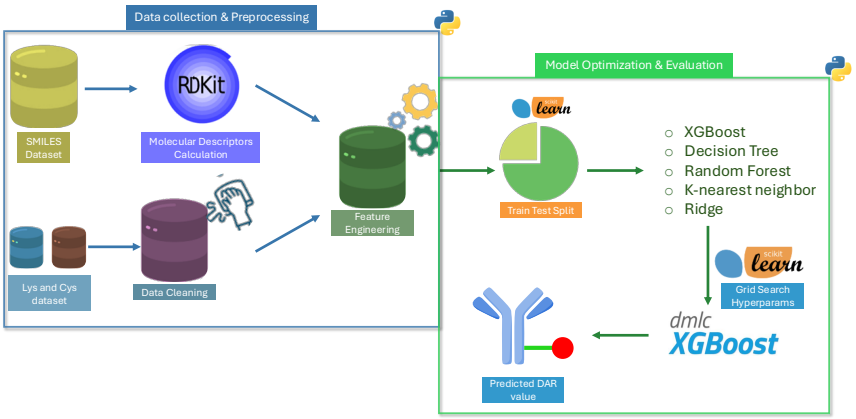


Figure 2.1: Workflow of this study.

2.1 Data acquisition and preprocessing

In the scientific world, the quality of the input data used in machine learning-based systems is a hot topic. Undoubtedly, the quality of the training data has a direct and significant impact on the fineness of the outputs produced by these prediction techniques.[163] To enhance the quality of our starting datasets, we opted to utilize the experimental data collected in our laboratories over the past decade of developing novel ADCs for diverse applications. This decision was made not only to mitigate the influence of biased outcomes stemming from misannotations, missing values, or other inconsistencies but also to facilitate the implementation of the library with novel experiments, thereby validating and refining the model. Two conventional bioconjugation approaches were investigated in this study: bioconjugation with lysine residues and bioconjugation with cysteine residues. First of all, two distinct datasets were created, collecting conjugation data from the experiments performed in our laboratories. The Lys dataset has 126 entries, of which 81 entries do not exhibit conjugation and 45 entries result in the conjugation of LP systems with mAbs. Of the 58 entries in the Cys dataset, 47 result in the conjugation of LP systems with mAbs, while 11 do not. In light of the small size of the two datasets, we adopted ML techniques tailored for small data sets, as discussed in prior research.[164] Seven different mAbs that are all members of the IgG family are included in the dataset. Commercially accessible mAbs include Bevacizumab (BVC), cetuximab (CTX), trastuzumab (TRX), and CSF-1R (BioLegend®). Azienda Ospedaliero-Universitaria Senese supplied BVC, CTX and TRX. The remaining three mAbs (J08, C-0302B17, and 4E1REC) are not marketed. Overall, our dataset includes chimeric, human, and murine isoforms; five of them are IgG1, and one is IgG2.

LP systems used in the bioconjugation experiments include both cleavable and uncleavable linkers, including those found in approved ADCs, new linkers developed in our laboratories, and others reported in the literature.[86] A vast array of structurally distinct substances

make up LP systems, from cytotoxic natural products to small molecules with a variety of actions, such as anticancer, antiviral (unpublished results), and antibacterial characteristics (unpublished results).[153,154,165–169]

A complete set of real-valued descriptors of chemical properties, including a variety of structural, topological, and physicochemical characteristics of the molecules, was created from the Simplified Molecular Input Line Entry System (SMILES) representations of the chemical compounds.[170] To calculate molecular descriptors, the RDKit cheminformatics toolkit (version 2023.09.1) was utilized.[171] The SMILES strings of molecules can be used to automatically and simply compute several molecular descriptors thanks to RDKit. To generate the input of the machine learning model, the clean data from the bioconjugation data set and the acquired molecular descriptors were combined (see section 7.1 for Python code).

2.1.1 Label encoding

A basic data preprocessing method for transforming categorical data into a numerical format appropriate for machine learning models is label encoding. Encoding is an essential step when working with features like colors, towns, or product kinds because many algorithms are unable to interpret non-numeric variables.[172] Categorical data is broadly divided into two types:

- **Nominal Data**: categories without inherent order (e.g., colors: red, blue, green).
- **Ordinal Data**: categories with a natural order (e.g., satisfaction levels: low, medium, high).

Label encoding works best for ordinal data, where the assigned numbers reflect the order. However, applying it to nominal data can unintentionally suggest an order (e.g., Red = 0, Blue = 1, Green = 2), which may mislead algorithms like linear regression. Thus,

the choice of encoding must align with the data type and the algorithm used.[173]
In this study, the LabelEncoder method was used to encode categorical variables to facilitate compatibility with ML models.[174] Each category was converted using this procedure into an integer between 0 and n-1, where n is the number of different classes for a given value. To ensure that all features contributed equally during model training and to avoid features with greater ranges controlling the learning process, numerical columns were scaled to the range $[0, 1]$ in addition to categorical encoding (see section 7.1 for Python code).

In the beginning, our goal was to predict whether or not a LP system of interest could be conjugated to a mAb. For this reason, the first ML model implementation regards the comparison of six classifiers. Working with a classifier, the outcome of a given prediction would be "1" for a positive conjugation, and "0" for a negative one. This means that the classifier would only be able to predict whether or not a compound could be successfully linked to an antibody without giving a precise quantification of the drug loading. The six classifiers evaluated for the task are:

- **Logistic regression classifier** is a widely used supervised learning method for binary and multiclass classification tasks. It models the conditional probability of a class label using a logistic (sigmoid) function applied to a linear combination of the input features. Despite its conceptual simplicity, logistic regression provides interpretable model coefficients and performs well when the classes are approximately linearly separable. In this work, a logistic regression classifier was implemented using the *LogisticRegression* class from the scikit-learn library in Python.[174]

- **Decision tree classifier** is a non-parametric model that partitions the feature space into hierarchical regions by recursively selecting features and split thresholds that maximize information gain or minimize impurity. This approach produces interpretable decision rules and can naturally handle both numerical and categorical data.

The *DecisionTreeClassifier* from scikit-learn was used in this work.[174]

- **SVM** is a supervised learning algorithm that seeks to find the optimal hyperplane that maximally separates data points of different classes in a high-dimensional feature space. By using kernel functions, SVMs can efficiently perform non-linear classification by implicitly mapping inputs into higher dimensions. In this study, the *SVC* class from the scikit-learn library was employed to train SVM classifiers.[174]

- **k-NN** algorithm is a non-parametric classification method that assigns a class to a sample based on the majority label among its k closest training instances, as measured by a distance metric such as Euclidean distance. It is simple to implement and effective for problems where class boundaries are irregular or non-linear. The *KNeighborsClassifier* implementation from scikit-learn was used in this work to evaluate the effect of neighborhood size and distance weighting on classification performance.[174]

- **linear discriminant analysis (LDA)** is a classical statistical method that projects data onto a lower-dimensional space while maximizing class separability. It assumes normally distributed classes with equal covariance matrices and computes linear decision boundaries accordingly. LDA often serves both as a dimensionality reduction technique and a linear classifier. In this work, classification was performed using the *LinearDiscriminantAnalysis* class from scikit-learn, which provides a stable and efficient implementation.[174]

- **Extreme Gradient Boosting (XGBoost)** is an ensemble learning method based on gradient-boosted DTs. It constructs a strong predictive model by iteratively adding weak learners that correct the residual errors of previous models, while incorporating regularization to prevent overfitting. XGBoost is known for its scalability, computational efficiency, and superior performance on structured data. In this study,

the XGBClassifier from the XGBoost library was employed for model training and evaluation.[175]

Next, we took a step forward, and we performed a regression analysis by comparing the performance of five regressors:

- **XGBoost regressor** is an ensemble method based on gradient boosting that builds a series of DTs sequentially, where each tree corrects the errors of the previous ones.[176] It is highly efficient and often achieves state-of-the-art performance due to its regularization and parallel computation capabilities.

- **RF regressor** is an ensemble method that constructs multiple DTs using bootstrap samples and random subsets of features, and averages their predictions to improve accuracy and robustness.[39] This approach effectively reduces overfitting and handles non-linear relationships among variables.

- **Ridge regression (Ridge)** is a linear model that incorporates L2 regularization to penalize large coefficient values, thereby reducing model variance and mitigating overfitting. It is particularly useful when multicollinearity exists among predictors.

- **k-NN regressor** is a non-parametric method that estimates the output for a given sample based on the average of its k nearest neighbors in the feature space.[177] While conceptually simple, its performance depends strongly on the choice of k and the scaling of input features.

- **Decision tree regressor** recursively partitions the data space into regions defined by feature thresholds, selecting splits that minimize prediction error.[178] The resulting tree structure is highly interpretable, though it may require pruning or ensemble methods to prevent overfitting.

These models were chosen because they are known to perform well with limited data, capturing key patterns while maintaining predictive reliability. The hyperparameters of the models were selected through a grid search, testing every possible configuration to achieve the parameter set that guarantees the best results. Details on hyperparameter tuning and metric score comparisons are provided in section 2.3.

2.2 Feature selection

Feature selection is a fundamental step in machine learning, particularly for high-dimensional datasets, as it reduces the number of input variables to those most informative for the prediction task. Proper feature selection improves model interpretability, lowers computational cost, and can enhance generalization by mitigating overfitting.[17] Techniques such as variance thresholding, recursive feature elimination, and regularization-based selection allow the systematic removal of redundant or uninformative features. This study employed variance-based feature selection to remove low-variance molecular descriptors, ensuring that the models focused on chemically relevant features, thereby simplifying model complexity while preserving predictive power.[179]

`VarianceThreshold` is a feature selector from `sklearn.feature_selection` that removes all features whose variance does not meet the threshold. For binary features (0/1), $variance = p * (1 - p)$ where p is the fraction of ones. So, $0.8 * 0.2$ removes features that are "mostly constant", e.g., a feature that is 0 in 80% of samples and 1 in 20%. Essentially, it filters out low-variance features, which are unlikely to be informative. See section 7.1 for Python code.

2.2.1 Feature scaling

A critical preprocessing step in machine learning is feature scaling, particularly when a dataset includes variables with disparate magnitudes or units. Algorithms that rely on distance metrics (k-NN, SVM) or use gradient-based optimization (linear/logistic regression, neural networks) may favor features with larger numerical values when predictors have widely different scales (e.g., one feature in the range of hundreds, another in decimals).[180] Scaling, whether through min-max normalization, standardization (zero mean, unit variance), or other techniques, guarantees that each feature contributes proportionately, enhancing the learned coefficients' interpretability, numerical stability, and rate of convergence. When it comes to molecular or chemical data, where descriptors may have widely disparate units (mass, length, electronegativity, etc.), failing to scale might result in performance deterioration or deceptive feature importance.[181] In this study, feature scaling was performed using the `StandardScaler` implementation from the `sklearn.preprocessing` module, which standardizes features to zero mean and unit variance. An alternative approach, `MinMaxScaler`, was evaluated during preliminary testing but resulted in inferior predictive performance. See section 7.1 for Python code.

2.3 Training phase

A crucial phase in the machine learning pipeline is the training phase, where models acquire knowledge from data by adjusting internal parameters to align inputs and outputs effectively. In this study, a training set and a test set were crafted using the preprocessed bioconjugation data sets. The model underwent training using the training set, which comprised 80% of the data set, to learn the hidden features. The remaining 20% of the data set constituted the test set, which was employed to evaluate the model's perfor-

mance after training. To ensure robustness, twenty runs of each network were conducted, each utilizing a distinct data set split for training and testing obtained with the method `train_test_split` from `sklearn.model_selection`. See section 7.1 for Python code.

2.3.1 Model evaluation: Classifiers

To evaluate the performance of the classifiers, we computed 4 metrics from `sklearn.metrics`:

- `accuracy_score` is a metric that measures how often a machine learning model correctly predicts the outcome.
- `recall_score` is a metric that measures how often a machine learning model correctly identifies positive instances (true positives) from all the actual positive samples in the dataset.
- `f1_score` is the harmonic mean of precision and recall, which combines both metrics into a single value that balances their importance.
- `precision_score` is a metric that measures how often a machine learning model correctly predicts the positive class.

Model	**Accuracy**	**Recall**	**F1**	**Precision**
Logistic Regression	0.800	0.777	0.734	0.711
k-NN	0.765	0.687	0.669	0.675
DT classifier	0.810	0.747	0.728	0.732
SVM	0.798	0.869	0.741	0.663
LDA	0.785	0.754	0.708	0.683
XGBoost	0.806	0.772	0.741	0.743

Table 2.1: Evaluation metrics for each classifier algorithm on the Lys bioconjugation dataset.

Model	**Accuracy**	**Recall**	**F1**	**Precision**
Logistic Regression	0.738	0.766	0.807	0.878
k-NN	0.858	1.000	0.914	0.849
DT classifier	0.871	0.911	0.917	0.927
SVM	0.662	0.632	0.749	0.940
LDA	0.825	0.906	0.893	0.889
XGBoost	0.896	0.963	0.935	0.913

Table 2.2: Evaluation metrics for each classifier algorithm on the Cys bioconjugation dataset.

Overall, all the tested networks seem to work better on the Cys bioconjugation dataset,

where in some cases the evaluation metrics have scores > 0.9. Among the evaluated classifiers, tree-based models, particularly XGBoost and DT, demonstrated superior predictive performance, achieving the highest accuracy, recall, and F1-scores. Apparently, these algorithms have effectively captured non-linear and interaction effects among features, which linear models such as Logistic Regression, LDA, and SVM were unable to model adequately. While Logistic Regression and LDA provided reasonable baseline performance, their linear assumptions probably limited their discriminative capability. Conversely, XGBoost's ensemble learning and regularization mechanisms yielded a robust balance between bias and variance, outperforming single-tree models and confirming the advantage of gradient boosting in modeling non-linear classification problems.

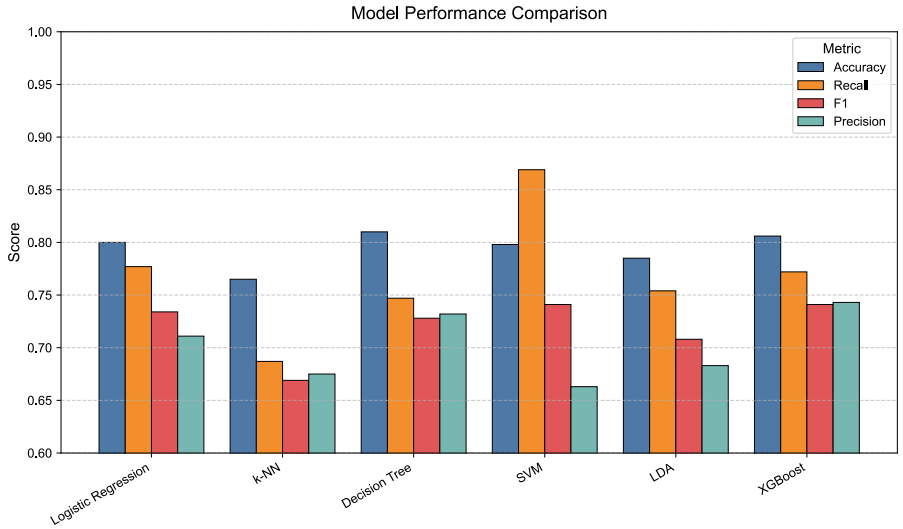


Figure 2.2: Score metrics regarding lysine dataset.

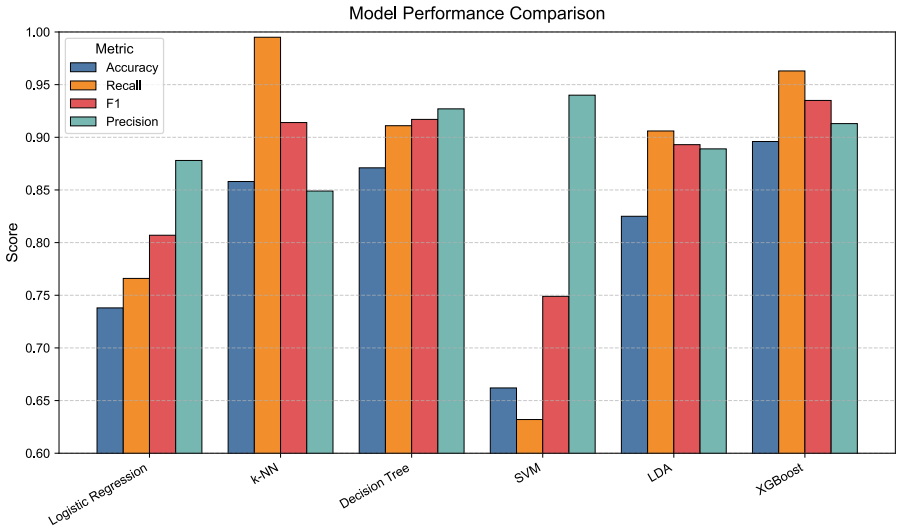


Figure 2.3: Score metrics regarding cysteine dataset.

2.3.2 Model evaluation: Regressors

Next, to assess and report the performance of the five regressors, we evaluate the coefficient of determination (R^2), mean squared error (MSE), mean absolute error (MAE), and standard deviation (SD) over the runs.

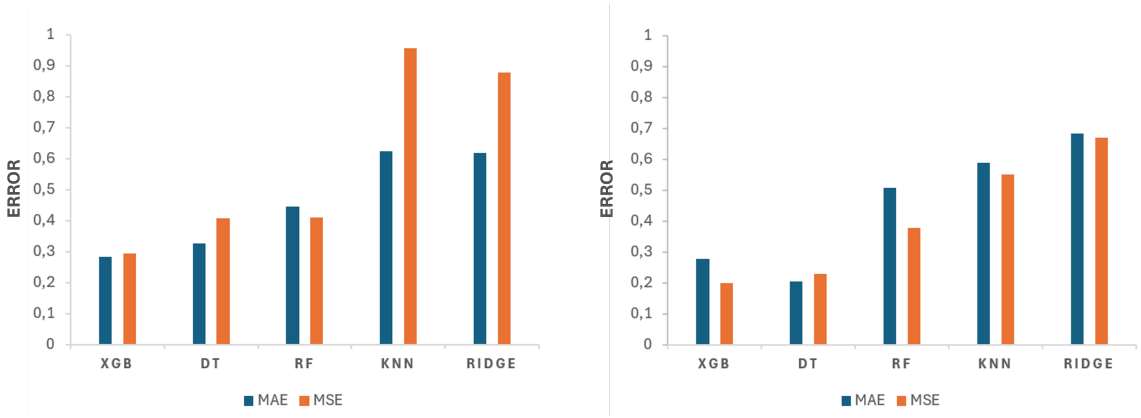
- **R^2** quantifies the proportion of variance in the dependent variable that is predictable from the independent variables in a regression model. In machine learning, it provides an interpretable measure of how well a model generalizes to unseen data, where R^2=1 indicates perfect predictive performance, and R^2=0 corresponds to performance equivalent to predicting the mean of the target variable.

- **MSE** is the average of the squared differences between predicted and true values. It penalizes large deviations more severely than MAE, making it particularly sensitive to outliers. MSE is widely used as a loss function during model training, particularly in regression and neural network optimization.

- **MAE** represents the average magnitude of prediction errors without considering their direction. It is defined as the mean of the absolute differences between predicted and observed values. Because MAE maintains the same units as the target variable, it provides a straightforward interpretation of prediction accuracy.

- **SD** quantifies the dispersion or variability of a dataset around its mean. In the context of machine learning, SD is often used to express the variability of performance metrics (e.g., accuracy, MSE, or R^2) across cross-validation folds or repeated experiments.

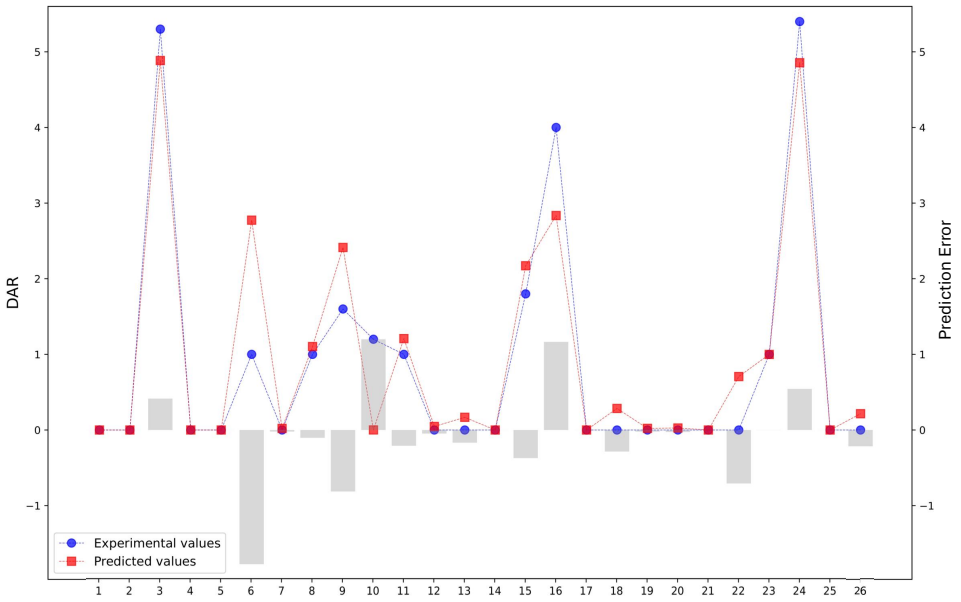
A comparison of the MAE and MSE over the runs is shown in Figure 2.4.



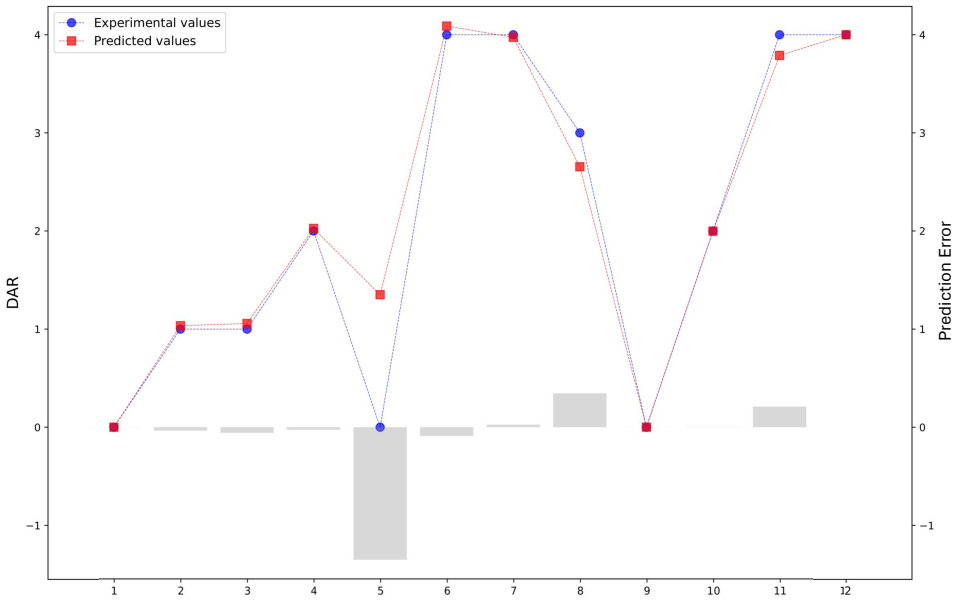
(a) (b)

Figure 2.4: Error metrics comparison for trained regression models for (a) Lys bioconjugations and (b) Cys bioconjugations.[182]

Based on the results obtained, the XGBoost model was identified as the most reliable approach for predicting the DAR (see Table S6). Although the DT model exhibited slightly superior performance for cysteine bioconjugations, the XGB model was ultimately selected due to its stronger generalization capability across both lysine and cysteine datasets. A more detailed explanation about the model optimization is reported in subsection 2.3.3. The predictive performance of the XGB model for DAR estimation is illustrated in Figure 2.5a, which presents the predicted versus experimental DAR values for a test set of 26 ADCs generated via Lys conjugation, including error bars for each prediction. A corresponding analysis for Cys-based bioconjugates, comprising a test set of 12 ADCs, is provided in Figure 2.5b. Bioconjugations are reported on the X-axis, while DAR values are on the Y-axis. Each related experimental and predicted DAR value is shown as a dashed line with markers (predicted values are shown in red and experimental values in blue). The prediction error for each bioconjugation is shown by error bars in the background, which show the difference between the experimental and projected measurements. According to the standard, the error bars' direction extends downward if the experimental value is less than the expected value and upward if it exceeds the predicted one.



(a)



(b)

Figure 2.5: Comparison between experimental and predicted DAR values on 12 Cys bioconjugations (b) and 26 Lys bioconjugations (a).[182]

An additional experimental validation was performed on both cysteine and lysine bioconjugations, taking into account a holdout test set; results are reported in 2.4 The

graphical representation of performance gives an exhaustive overview of the prediction ability of the XGBoost model. In both cases, the model accuracy in predicting DAR is really good. It is interesting to note that the model also achieved accurate predictions for LP systems that are incompatible with mAb bioconjugation, resulting in a DAR value of zero. This capability is particularly valuable for the rational design of new LP systems, as it enables the early identification of unsuitable LP candidates and thus helps prevent the unnecessary use of resources and time on compounds likely to fail during the final conjugation step.

Next, the individual contributions of the ten most influential features in our predictive model were investigated via the SHapley Additive exPlanations (SHAP) technique. SHAP assigns to each input feature a quantitative score that reflects its influence on the model's output for each observation. In our setting, the resulting SHAP values allow us to quantify how each attribute affects the model's predictions of DAR, with larger absolute values indicating features of higher importance for the model's decision process. SHAP is grounded in the concept of Shapley value from cooperative game theory, which attributes the "payout" of a game to each player by averaging their marginal contributions across all possible coalitions. In the machine-learning paradigm, the "players" are the features and the "game" corresponds to the model's prediction.[183] SHAP has been successfully applied in the interpretation of complex regression and classification models in cheminformatics.[184,185]

Figure 2.6 displays the SHAP summary plot, in which the features are ordered along the y-axis according to their overall importance, with higher positions indicating greater influence on the model's predictions. The x-axis represents the SHAP values, which quantify both the magnitude and the direction of each feature's effect on the predicted DAR. Each point corresponds to a single sample, and its color encodes the feature value, ranging from low (blue) to high (red). Therefore, this visualization illustrates how variations in each feature influence the prediction outcomes across the dataset. Both numerical and cate-

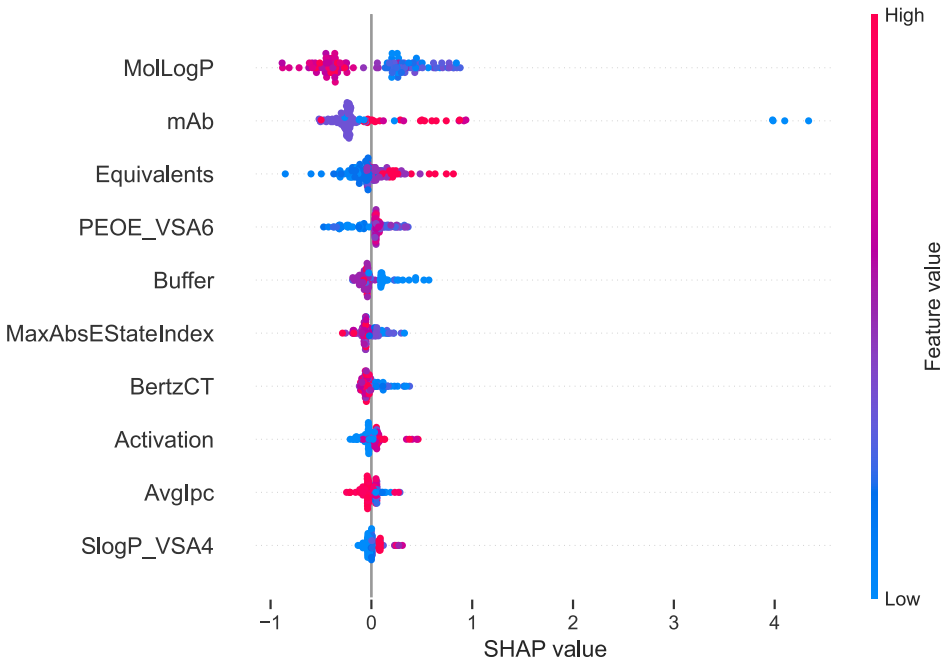
gorical variables are represented using the same color gradient; however, for categorical features, higher encoded values do not imply a quantitative relationship but rather denote different categories. The specific color schemes adopted for the categorical variables are detailed in section 7.5.

The SHAP analyses for both Lys and Cys bioconjugations reveal several overlapping features, suggesting common factors influencing the two types of conjugation processes. These shared features predominantly correspond to experimental parameters, which define critical aspects of the reaction environment necessary for efficient bioconjugation. In the case of Lys conjugation, the feature with the highest importance score is associated with the lipophilicity of the molecules. Specifically, the partition coefficient (MolLogP) estimates molecular lipophilicity by comparing solubility in nonpolar solvents (e.g., lipids or oils) versus water.[186] This descriptor, which calculates the logP value, plays a pivotal role in drug discovery and medicinal chemistry, as it influences molecular interactions, solubility, and aggregation behavior.[187–190] High lipophilicity can therefore promote the formation of undesired aggregates during or after the bioconjugation process, affecting the overall conjugation efficiency and product stability. While several descriptors are shared between Lys and Cys-based bioconjugations, certain features emerge as uniquely influential within each system. In contrast to the Lys model, where MolLogP plays a dominant role, the most important descriptor for Cys bioconjugation corresponds to a molecular operating environment (MOE-type) parameter that combines partial atomic charges and surface area contributions (PEOEVSA). This descriptor is derived from the Partial Equalization of Orbital Electronegativities (PEOE) method[191], which estimates atomic charges in σ-bonded and unconjugated π systems based solely on atomic connectivity. Such descriptors are particularly valuable for capturing the electronic distribution across a molecule and for characterizing how these electronic properties influence molecular interactions within the reaction environment. In the context of ADCs, these electronic effects can have a decisive impact on the reactivity and efficiency of linker chemistry, ultimately affecting conjugation

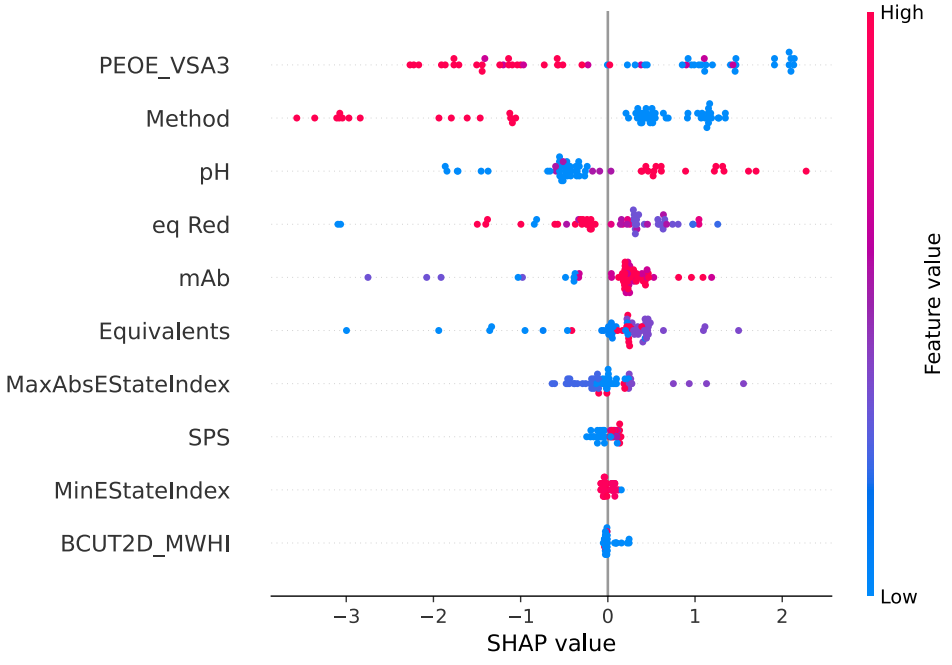
performance and the overall stability of the resulting bioconjugates[76,192].

A closer inspection of the SHAP results reveals that the PEOEVSA descriptors exert opposite effects on the model's predictions for Lys and Cys bioconjugations. Specifically, higher PEOEVSA values tend to enhance the predicted conjugation efficiency in Lys-based systems, whereas they negatively influence the prediction outcomes for Cys-based systems. This inverse relationship suggests that the local electronic environment surrounding the reactive sites contributes differently to the conjugation efficiency depending on the residue involved. Moreover, the SHAP analysis highlights that certain mAb subclasses may be inherently more favorable for bioconjugation. In particular, humanized IgG_1 antibodies, such as TRX, exhibit positive SHAP contributions, whereas IgG_2 subclasses, exemplified by CSF-1R, show negative SHAP values, likely due to their comparatively lower structural flexibility[193]. This observation aligns with current trends in ADC development, where the majority of clinically approved ADCs are based on the IgG_1 isotype, reinforcing the notion that IgG_1 antibodies possess more suitable structural and physicochemical characteristics for effective bioconjugation[194]. In addition to the influence of the electronic environment and antibody subclass, the SHAP analysis also underscores the importance of molecular complexity in determining bioconjugation efficiency[195]. Two descriptors, BertzCT and the Spatial Score (SPS), quantify the topological complexity of LP systems. In particular, the SPS provides an empirical measure of spatial complexity, integrating the fraction of sp^3-hybridized carbon atoms and the proportion of stereogenic centers within a molecule[196]. The analysis indicates that more complex molecules, especially those containing multiple stereogenic centers, exhibit positive SHAP contributions, implying a beneficial effect on bioconjugation performance. Conversely, planar structures with a predominance of sp^2-hybridized carbons may lack the optimal topological characteristics required for effective conjugation. Furthermore, both datasets reveal that van der Waals surface area (VSA) descriptors provide critical information for the predictive model.[197] These descriptors capture molecular properties such as partial charge distribution and

lipophilicity across the solvent-accessible surface, offering insight into how molecular topology and surface chemistry influence bioconjugation outcomes. In cheminformatics, VSA-based parameters are widely used to describe molecular interactions within biological environments, and in this context, they represent valuable tools for guiding the rational design of LP systems. By considering VSA, SPS, and BertzCT descriptors collectively, it becomes possible to optimize structural and topological features conducive to successful and efficient bioconjugation.



(a)



(b)

Figure 2.6: SHAP summary plots showing the 10 key features for predicting bioconjugation: (a) Lys and (b) Cys bioconjugation.[182]

2.3.3 Hyperparameter tuning

To optimize the predictive performance of the machine learning model developed for DAR prediction, a systematic hyperparameter tuning procedure was carried out using grid search. The tuning was performed independently for both Lys and Cys bioconjugation datasets to account for their distinct feature distributions. A grid search strategy combined with five-fold cross-validation (k=5) was employed to explore different combinations of hyperparameters and identify the configuration that yielded the best predictive performance. Model selection was based on multiple evaluation metrics, including the R^2 (R^2), MAE, MSE, and SD of predictions. The specific hyperparameter ranges considered for each model are summarized in Table 2.3.

Model	**Hyperparameters**
XGB	Learning rate: {0.01, 0.1, 0.3}
	Number of estimators: {100, 200, 500}
	Maximum depth: {3, 5, 7, 10}
	Subsample ratio: {1.0, 0.8, 0.6}
	Column sample by tree: {1.0, 0.8, 0.6}
	Regularization parameters: {1.0, 0.1, 10}
DT	Maximum depth: {None, 5, 10, 20}
	Minimum samples required for a split: {2, 5, 10}
RF	Number of estimators: { 100, 200, 500}
	Maximum depth: {None, 10, 20, 30}
	Minimum samples required for a split: {2, 5, 10}
	Minimum samples required per leaf: {1, 2, 5}
k-NN	Number of neighbors: {3, 5, 7, 10}
	Distance metric: {'minkowski', 'euclidean', 'manhattan'}
Ridge	Alpha: {1.0, 0.1, 10}

Table 2.3: Hyperparameter ranges considered for each tested mode.

After performing a grid search, the best hyperparameters that resulted in the highest predictive accuracy were identified. The final selected hyperparameters for Lys and Cys bioconjugation are reported in Table 2.4.

Model	Lys hyperparameters	Cys hyperparameters
XGB	Learning rate = 0.3 Estimators = 500, Max depth = 5 Subsample = 0.6 Colsample by tree = 1.0 Reg = 0.1	Learning rate = 0.01 Estimators = 500, Max depth = 10 Subsample = 0.8 Colsample by tree = 1.0 Reg = 0.1
DT	Max depth = None Min samples split = 10	Max depth = None Min samples split = 2
RF	Estimators = 500, Max depth = 10 Min samples split = 2 Min samples leaf = 1	Estimators = 200 Max depth = None Min samples split = 2 Min samples leaf = 1
k-NN	Neighbors = 10 Distance = 'minkowski'	Neighbors = 10 Distance = 'minkowski'
Ridge	Alpha = 1.0	Alpha = 1.0

Table 2.4: Optimal hyperparameters selected through grid search for Lys and Cys bioconjugation models.

2.4 Experimental validation

This section presents the results obtained on a holdout test set comprising 13 bioconjugation entries that include cysteine and lysine conjugates. The corresponding results are presented in Table 2.5. Entries B226 and B229 correspond to lysine-based conjugates; entries C58–C121, instead, represent cysteine-based conjugations, generated through se-

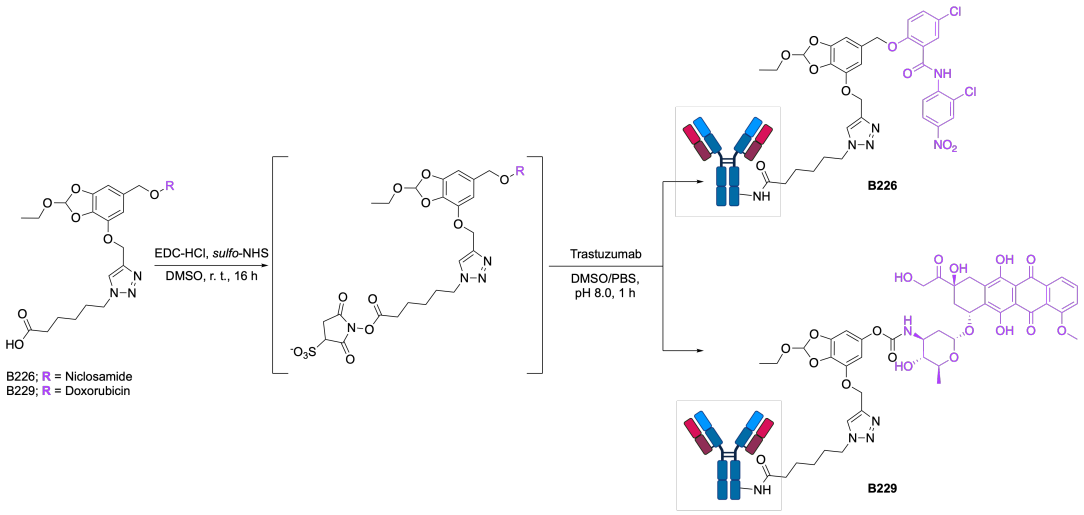
lective modification of interchain disulfide bonds. Figures shown in subsection 2.4.2 report MALDI-TOF MS spectra obtained from Lys and Cys bioconjugates and native mAb (see chapter 4 for MALDI sample preparation). The DAR value was calculated from the MALDI spectra to assess the accuracy of the prediction generated by the ML model, as shown in chapter 4.

Entry	**MW LPs**	**pred. DAR**	**exp. DAR**
B226	716.53	0.0	0.3
B229	976.94	1.0	1.2
C58	350.40	7.8	8.0
C96	469.54	7.9	8.0
C108	469.54	2.0	2.0
C110	469.54	1.9	2.0
C111	937.09	1.9	4.0
C112	469.54	8.0	8.0
C113	883.80	7.5	6.0
C114	469.54	2.0	2.0
C117	883.80	7.6	8.0
C120	883.80	7.5	6.0
C121	883.80	2.0	2.0

Table 2.5: Comparison of predicted and experimental DAR values on holdout test set.

B226 and B229 are lysine-based bioconjugates that employ the same cleavable linker but differ in their payloads: B226 incorporates the anthelmintic agent niclosamide, whereas B229 carries doxorubicin. In both cases, the carboxylic acid of the LPs was first activated *in situ* with N-hydroxysulfosuccinimide (sulfo-NHS) and 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride (EDC-HCl) as the coupling reagents at room temperature for 16 h. The activated LPs were then added to the mAb solution, and the mixture was gently stirred for 1 h to allow conjugation (Scheme 1). Subsequently,

2 molar equivalents of a glycine solution (100 mM in water) were added to quench the unreacted LP system. The resulting mixture was then purified onto a 5 kDa cutoff Sephadex desalting column. Samples were then analyzed through MALDI-TOF MS following the procedure shown in chapter 4. Table 2.6 shows the equivalents of the reagents used for the synthesis of B266 and B229.



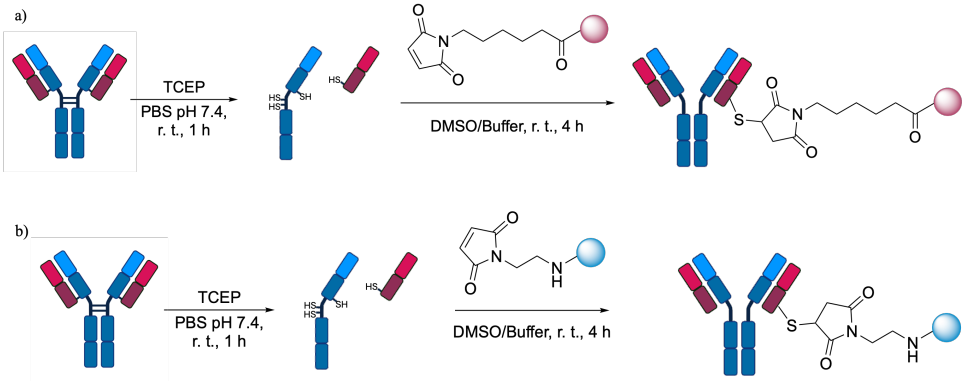
Scheme 1: Synthesis of lysine bioconjugates B226 and B229.

Entry	mAb	mAb conc. (mg/mL)	eq. EDC-HCl	eq. sulfo-NHS	eq. LP
B226	TRX	5	30	30	15
B229	TRX	5	30	30	15

Table 2.6: Equivalents of reagents and LPs used in the synthesis of B226 and B229.

On the other hand, eleven cysteine-based bioconjugates were synthesized under different reaction conditions. The mAb was initially reduced with TCEP at room temperature for 1 h to generate free thiol groups, after which the LP system was introduced and the reaction mixture was gently stirred for 4 h. The final solution was then purified onto a 5 kDa cutoff Sephadex desalting column. Samples were then analyzed through MALDI-TOF MS following the procedure shown in chapter 4. To increase the diversity of the holdout

test set, two of the LPs used contained a MC-based linker, while the remaining two were functionalized with an N-(2-aminoethyl)maleimide linker. Table 2.7 shows the reaction conditions used for the synthesis of cysteine bioconjugates present in the holdout test set.



Scheme 2: General synthesis of cysteine bioconjugates: a) synthesis of ADCs with LP systems bearing a MS and b) synthesis of ADCs with LPs bearing a shorter N-(2-aminoethyl)maleimide.

Entry	**mAb**	**mAb conc. (mg/mL)**	**eq. TCEP**	**eq. LP**	**Buffer**
C58	TRX	10	8	5	PBS 8.5
C96	TRX	5	10	10	PBS 7.4
C108	CTX	5	10	10	PBS 7.4
C110	TRX	5	10	10	PBS 7.4
C111	TRX	2.5	10	10	PBS 7.4
C112	BVC	5	10	10	PBS 7.4
C113	TRX	10	10	20	PBS 7.4
C114	TRX	10	10	10	PBS 7.4
C117	TRX	10	6	20	PBS 7.4
C120	TRX	5	10	10	PBS 7.4
C121	TRX	5	10	20	PBS 7.4

Table 2.7: Bioconjugation conditions for the synthesis of cysteine bioconjugates present in the holdout test set.

Detailed experimental procedures for the synthesis and purification of these bioconju-

gates are provided in chapter 4.

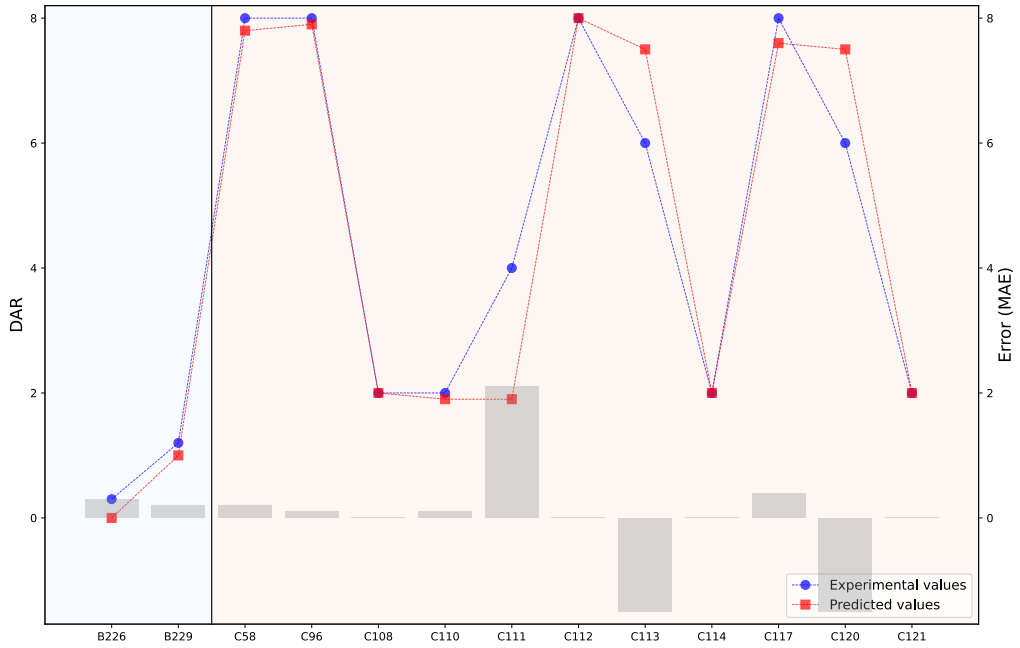


Figure 2.7: Predicted versus experimental DAR comparison on the experimental validation test.

Looking at Figure 2.7, in the X-axis, 2 Lys bioconjugations (on the left) and 11 Cys bioconjugations (on the right) are reported. The corresponding predicted and experimental DAR values are displayed as dashed lines with markers for each (red for predicted values and blue for experimental values). Error bars in gray, plotted in the background, indicate the absolute prediction error for each bioconjugation, reflecting the deviation between predicted and experimental measurements. Overall, the results obtained from the prediction on the holdout test set align with the previous ones, highlighting the robustness of the method.

2.4.1 Case study

This section presents a case study focusing on LPs that are either currently employed in approved ADCs or are of potential research interest. Specifically, two commercially available LPs and six under investigation LPs were analyzed. The approved LPs correspond to those used in ADCETRIS® and ZYNLONTA®, namely *mc-vc-PAB-MMAE* and *Tesirine*, respectively. Table 2.8 reports the model predictions for the conjugation of these two reference LPs, serving as validation cases against known, clinically established ADC constructs. Subsequently, Table 4.1 summarizes the model predictions for six additional LP candidates selected for exploratory evaluation.

	mc-vc-PAB-MMAE	**Tesirine**	**mc-vc-PAB-MMAE**
PEOE_VSA3	9.5891	9.6944	9.5891
Method	Dialyzed	Dialyzed	Dialyzed
pH	8	8	8
eq TCEP	2.75	2.75	6
mAb	BRX	LCX	BRX
eq	10	10	10
SPS	17.4043	17.0000	17.4043
BertzCT	2875.14	3623.73	2875.14
MW	1316.63	1496.65	1316.63
MolLogP	5.077	5.298	5.077
Pred. DAR	2.2	2.2	7.5
Exp. DAR	4.04	2.1	8.0

Table 2.8: DAR prediction on commercially available LP systems currently approved.

Table 4.1 summarizes the conjugation conditions employed in the study, including the conjugation method, buffer pH, the molar equivalents of reducing agent, mAb, and LP. The table also reports several molecular descriptors relevant to the model interpretation.

Among these, *PEOEVSA3*, identified through SHAP analysis as the most influential feature, stands out as a key predictor. Additional descriptors such as *SPS* and *BertzCT* quantify molecular complexity, with the former reflecting the number of stereogenic centers and the latter indicating the number of heteroatoms. The lipophilicity parameter *MolLogP* is also included; interestingly, for cysteine-based conjugations, this descriptor does not represent a limiting factor in the reaction outcome. Notably, *MolLogP* appeared as a relevant feature in the SHAP analysis for lysine-based conjugations but not for cysteine-based ones. Both LPs analyzed exhibit the same predicted DAR of 2.2, indicating a high degree of chemical similarity. Their nearly identical values of *PEOEVSA3*, *SPS*, molecular weight (MW), and *MolLogP* further support this conclusion. The model prediction for *Tesirine* aligns closely with the experimental DAR (predicted 2.2 vs experimental 2.1), whereas for *vcMMAE* the deviation is larger (predicted 2.2 vs experimental 4.0). This discrepancy can be attributed to the duration of the disulfide bond reduction step, which was not included as a variable in the dataset. In all experiments within the cysteine dataset, reduction was standardized to one hour. However, Nayak and Richter[198] reported that achieving a DAR of 4 for *vcMMAE* requires approximately four hours of reduction using 2.75 equivalents of the reducing agent. Their work also demonstrated that the reduction time constitutes the rate-limiting step in the bioconjugation process and that employing an excess of reducing agent relative to the four interchain disulfide bridges can yield DAR values up to 8.

Our model is consistent with these observations. In fact, when the equivalent amount of TCEP is increased in the predictive simulation to exceed the number of disulfide bonds, the model returns a predicted DAR of 7.5, remarkably close to the experimental value reported by Nayak and Richter. This agreement underscores the robustness and interpretability of the proposed model in reproducing experimentally observed conjugation behaviors.

	1.1	**1.2**	**1.3**	**2.1**	**2.2**	**2.3**
PEOE_VSA3	9.370	9.370	14.164	14.164	9.370	9.370
Method	Dialyzed	Dialyzed	Dialyzed	Dialyzed	Dialyzed	Dialyzed
pH	8	8	8	8	8	8
eq TCEP	2.75	2.75	2.75	2.75	2.75	2.75
mAb	CTX	CTX	CTX	CTX	CTX	CTX
eq	10	10	10	10	10	10
SPS	13.894	14.315	17.557	17.900	11.639	12.032
BertzCT	2261.938	2295.019	2643.450	2677.272	2149.231	2182.054
MW	812.42	813.43	976.65	977.66	858.52	859.53
MolLogP	5.088	5.1307	8.200	8.242	3.929	3.971
Pred. DAR	2.2	2.3	1.9	1.9	2.1	2.1

Table 2.9: DAR prediction on LPs of interest.

From the results obtained, it is important to note that LPs exhibiting *PEOEVSA3* values comparable to those of the commercial ADC linkers display the same predicted DAR. In contrast, molecules 2.1 and 2.2, characterized by higher *PEOEVSA3* values, show a lower predicted DAR. This observation is fully consistent with the SHAP analysis reported in subsection 2.3.2, which indicates that increasing *PEOEVSA3* exerts a negative influence on the DAR outcome. Consequently, LPs 1.1 and 1.2 emerge as the most promising candidates for conjugation, as their physicochemical profiles closely resemble those of the commercial reference LPs.

2.4.2 MALDI spectra

In this section, MALDI spectra of some ADCs present in the validation test set are reported; the acquisition procedure and DAR calculation formula are reported in chapter 4. MALDI spectra obtained from Cys bioconjugates represent the reduced chains of the mAb (light chain and heavy chain). For a correct calculation of the DAR on the MALDI spectra of Cys bioconjugates, it is necessary to consider that 2 light chains and 2 heavy chains are formed from each reduced mAb.

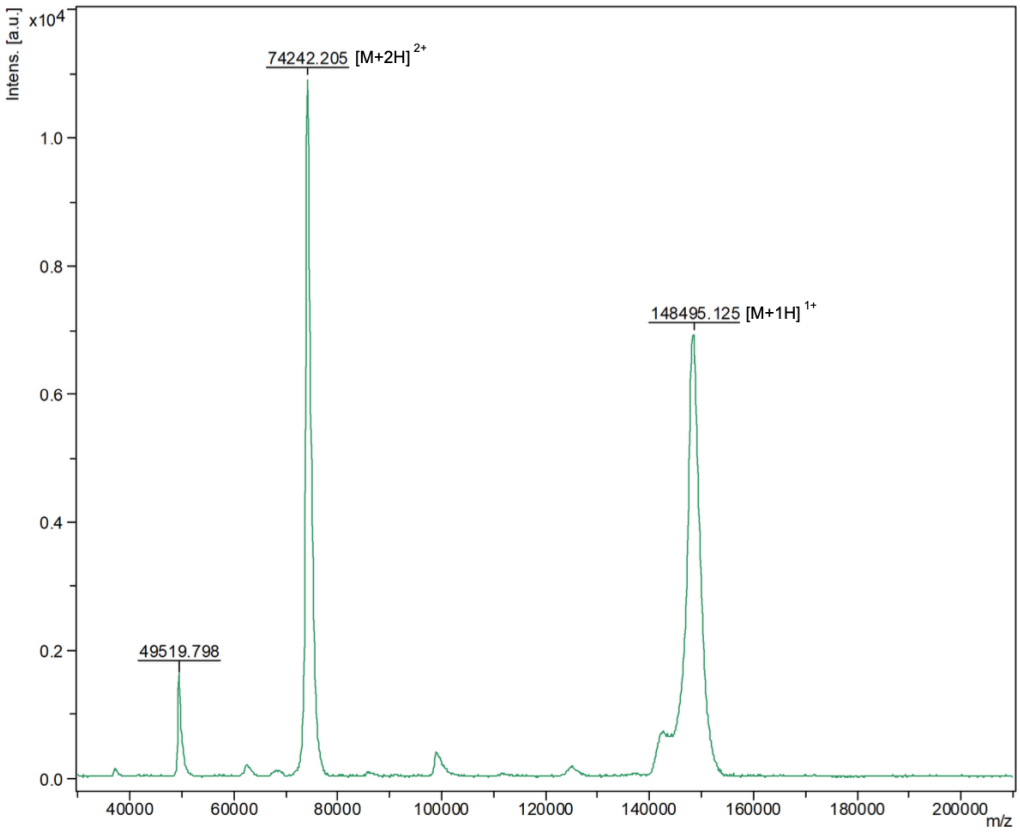


Figure 2.8: MALDI-TOF MS spectrum of intact TRX. The intact mass was determined by smoothing the peak of the doubly and mono-charged species using a Savitzky-Golay filter with 50 points, followed by automatic peak picking of the maximum of the curve.

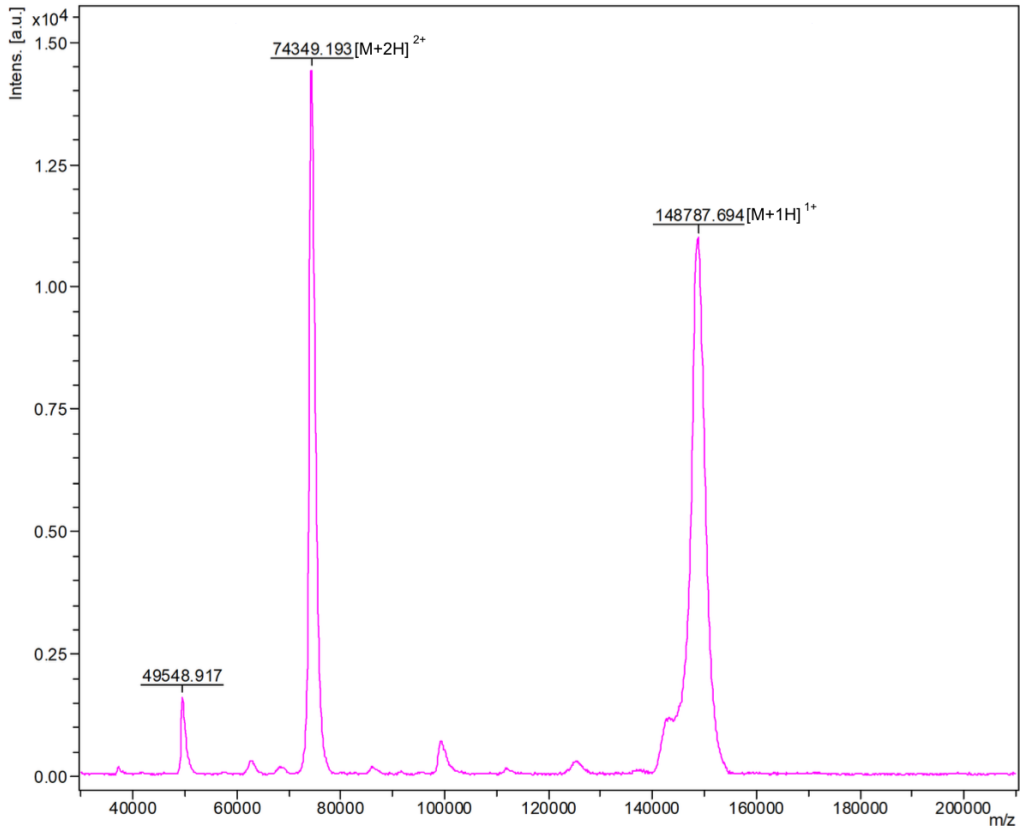


Figure 2.9: MALDI-TOF MS spectrum of lysine bioconjugate B226. The mass peak was determined as described in Figure 2.8. The DAR was calculated on the m/z value of the doubly charged species.

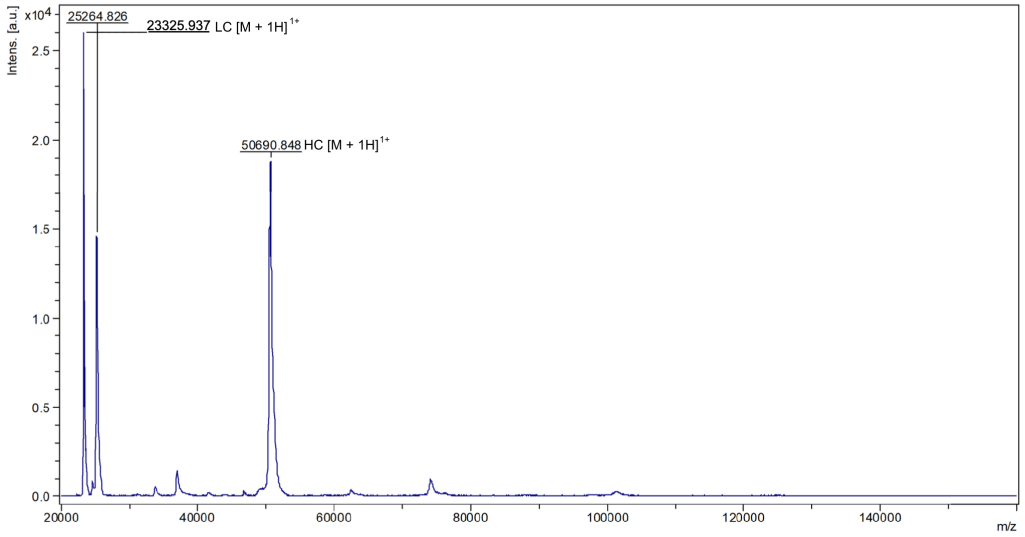


Figure 2.11: MALDI-TOF MS spectrum of reduced TRX. The mass peak was determined as described in Figure 2.8.

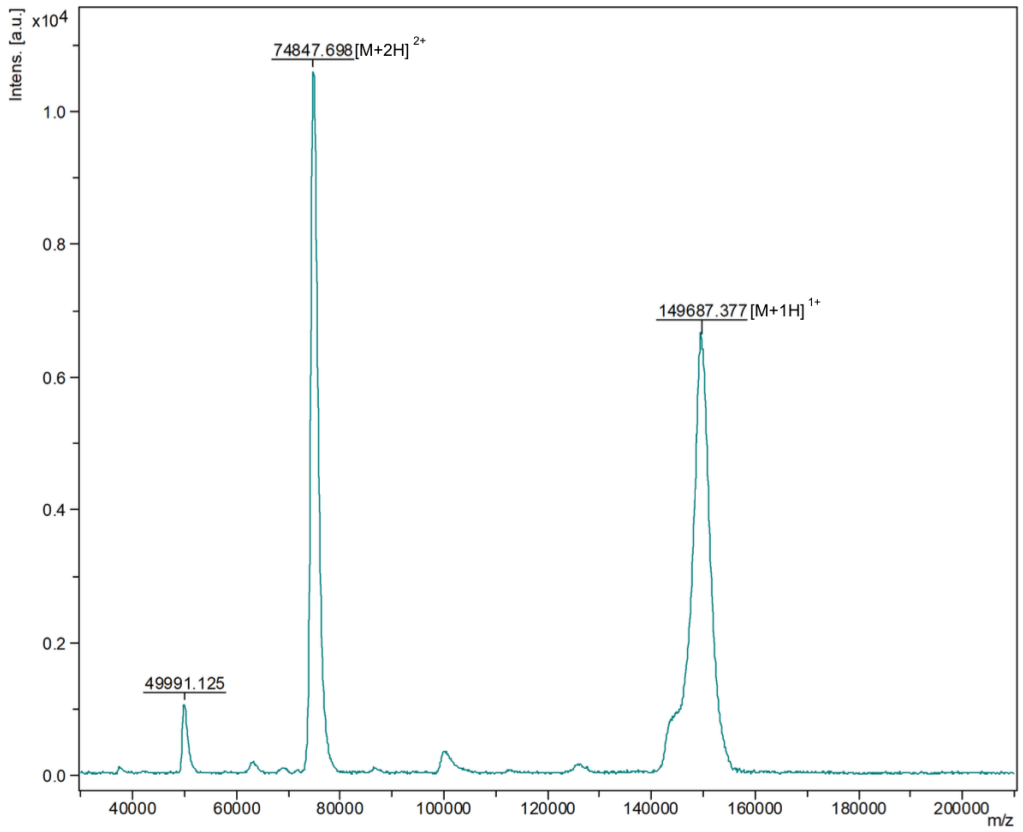


Figure 2.10: MALDI-TOF MS spectrum of lysine bioconjugate B229. The mass peak was determined as described in Figure 2.8. The DAR was calculated on the m/z value of the doubly charged species.

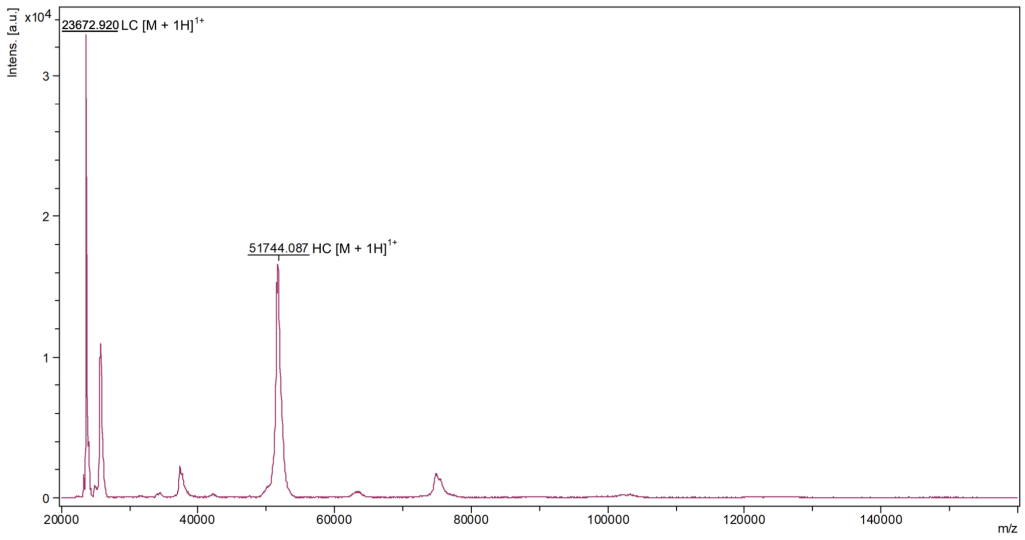


Figure 2.12: MALDI-TOF MS spectrum of C58. The mass peak was determined as described in Figure 2.8. The DAR was calculated on the m/z value of the mono-charged species of both Light and Heavy chain peaks (LC, HC).

Chapter 3

Conclusions

In this work, the implementation of an ML-based strategy is proposed to optimize bioconjugation processes that are fundamental to the development of ADCs. Among the algorithms evaluated, the XGBoost model exhibited outstanding predictive performance for bioconjugation reactions involving both lysine and cysteine-based conjugation sites across different antibodies, underscoring its potential as a powerful tool to enhance ADC design and synthesis. The main objective of this thesis was to build a reliable approach able to correctly predict the DAR in the synthesis of ADCs. The XGBoost algorithm stood out as both a classifier and a regressor to fulfill this task, underlining the great versatility of this network. By identifying key molecular descriptors, such as MolLogP, SPS, and PEOE_VSA, as the most influential features governing bioconjugation efficiency, the model effectively prioritized the physicochemical parameters that dictate conjugation outcomes. Thanks to the SHAP evaluation, we can understand how varying reaction conditions can impact the outcome of the bioconjugation reaction, but also which descriptors can be considered in order to optimize the LP system structurally and topologically. In addition, the interpretation given by the model is consistent with the literature, as discussed in subsection 2.4.1. Overall, this study establishes a robust computational framework that not only accelerates the development of ADCs but also addresses key challenges related

to conjugation selectivity, stability, and process optimization. The integration of machine learning approaches into bioconjugation design represents a significant advance toward more efficient, precise, and sustainable strategies for next-generation ADC production. Looking ahead, the insights generated in this work could enable the development of personalized bioconjugation protocols, where reaction parameters are tailored to the specific structural and chemical characteristics of a given LP system. Future efforts will focus on refining the predictive models and extending their applicability to a broader spectrum of conjugation chemistries, thereby expanding their potential well beyond the ADC domain.

Chapter 4

Experimental section

All reagents and solvents were used as received from commercial suppliers without further purification unless otherwise specified. This section describes the experimental procedures employed for the synthesis of ADCs via bioconjugation reactions, including the MALDI-TOF MS protocol and the DAR calculation formula. All sample concentrations were calculated using an Implen NanoPhotometer® N60 using the 260/280 ratio for quantitation and purity check. MALDI-TOF MS samples were prepared on an Ohaus® dry block heater with an operating temperature set to 37 °C. Unless otherwise stated, all experiments reported in this chapter were performed by the candidate.

Condition	Bioconjugation with Lys	Bioconjugation with Cys
pH	6.0 to 8.0	7.4 to 8.5
mAb	TRX, CTX, 4E1REC J08, CSF-1R, C-0302B17	TRX, CTX, CSF-1R, BVC
Buffer	MES, PBS, EPPS, BBS, TRIS, Water	PBS, BBS
Eq. of LP system	5 to 80	2 to 40
Activation type	NHS, sNHS, HOBTU, mAb-azide, PFP, DMT	maleimide moiety, alkyne moiety
Reducing agents		DTT, TCEP
Eq. of reducing agent		1.1 to 12.0
Method		Onepot - dyalized

Table 4.1: Value ranges of experimental conditions for lysine and cysteine bioconjugation.

4.0.1 Conjugation to Lys residues

General procedures here reported refer to protocols previously developed inside the research group[86,199].

4.0.1.1 General procedure

A solution of the purified mAb (100μL, 10 mg/mL) was buffer exchanged into a conjugation buffer. 5 to 80 molar equivalents (as described in Table 1) of the previously activated LP system (as defined in Table 1) were added, and the solution was gently mixed for 1 h. Subsequently, 2 molar equivalents of a glycine solution (100 mM in water) were added to quench the unreacted LP system, and the solution was mixed for 0.2 h. The final solution was then purified onto PD SpinTrap-G25 (5000 MWCO, Cytiva®) as suggested by the manufacturer. The final product was stored at 4 °C in phosphate buffered saline (PBS) (pH 7.4) until needed.

4.0.1.2 In-situ procedure

The LP system was first activated as sulfo-NHS derivative using 2 molar equivalents of EDC-HCl and sulfo-NHS (Merck®) at room temperature for 16 h[200]. Following this step, the synthetic protocol for the conventional procedure was followed as described above.

4.0.1.3 Pre-functionalization procedure

100 μL (10 mg/mL) of the conjugation buffer solution of the purified mAb were added with 5 to 80 molar equivalents of an NHS-PEG4-N3 solution (10 mM in dimethylsulfoxide (DMSO)), and the mixture was gently mixed for 1 h. Subsequently, the solution was purified onto PD SpinTrap-G25 (5000 MWCO, Cytiva®) as suggested by the manufacturer. After purification, 5 to 80 molar equivalents of the LP system of interest (10 mM in DMSO) were added and the solution was incubated at 4 °C for 16 h. The final product was then purified as previously described and stored at 4 °C in PBS (pH 7.4) until needed.

4.0.2 Conjugation to Cys residues

4.0.2.1 Two steps protocol

1.1 to 12.0 molar equivalents (as described in Table 1) of TCEP·HCl (1 mM, H_2O) or dithiothreitol (DTT) (1 mM, H_2O) were added to a solution of mAb (100μL, 10 mg/mL)in PBS (pH 7.4), and the solution was incubated at room temperature for 1h. After this time, the solution was purified on PD-SpinTrap (5000 MWCO, Cytiva®), and the buffer was swapped into the conjugation buffer. Then, 2 to 40 molar equivalents (as described in Table 1) of a solution of the LP system (10 mM, DMSO) was added to the reduced mAb at room temperature, and the solution was gently mixed for 4 h at room temperature. Excess reagents were removed by centrifugation on PD-SpinTrap (5000 MWCO, Cytiva®) as previously described, and the final product was stored at 4 °C in PBS (pH 7.4) until needed.

4.0.2.2 Onepot protocol

2 to 40 molar equivalents (as described in Table 1) of a DMSO solution of the LP system (10 mM) were added to 100 μL of mAb (10 mg/mL) in conjugation buffer, and the solution was incubated at room temperature for 15 minutes. After that, 1.1 to 12.0 molar equivalents of a TCEP·HCl solution (1 mM, H_2O) were added and the mixture was gently mixed at room temperature for 4 h. Excess reagents were removed by centrifugation on PD-SpinTrap (5000 MWCO, Cytiva®) as previously described, and the final product was stored at 4 °C in PBS (pH 7.4) until needed.

4.0.3 MALDI-TOF analysis of bioconjugates to estimate DAR

According to the manufacturer's protocol, bioconjugate samples were desalted using PD SpinTrap-G25 (5000 MWCO, Cytiva®). For the preparation of the thin layer, a saturated solution of sinapinic acid in ethanol was deposited onto the target plate and left to dry. The sample solution and the matrix solution (sinapinic acid 20 mg/mL, acetonitrile/water 0.1% TFA (70/30)) were mixed in a 1:1 ratio. Subsequently, 1μL of the mixed solution was spotted onto the target plate and left to dry. For every bioconjugate, two sample spots were prepared from the same solution. The spots on the target were chosen so that all samples were near the reference antibody. MALDI-TOF measurements were performed on a Bruker UltrafleX TOF/TOF in linear mode, voltage polarity POS, and 10000 laser shots were accumulated to obtain a spectrum. The intact mass of each Lys bioconjugate was calculated on the doubly charged peak after smoothing with a Savitzky-Golay filter with 50 points for the entire peak. The reference antibodies (TRX, CTX, BVC, 4E1REC, J08, C-0302B17, and CSF-1R) followed the same procedure. For Cys bioconjugates, the intact mass of the light chain and the heavy chain was calculated on the mono-charged peak after smoothing with a Savitzky-Golay filter with 50 points for the entire peak.

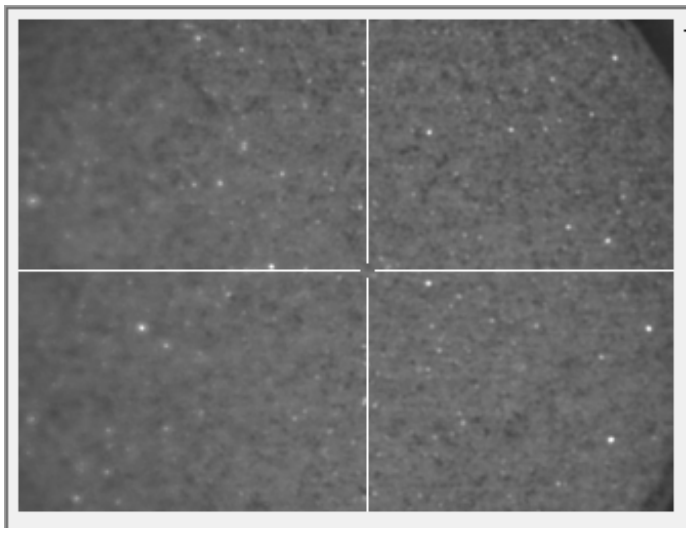


Figure 4.1: Example of a MALDI-TOF sample spotted on a Bruker MTP 384 ground steel target plate.

DAR was calculated using the equation reported below:

$$DAR = \frac{MW_{bioconjugate} - MW_{mAb}}{MW_{drug}}$$

where $MW_{bioconjugate}$ denotes the molecular weight (MW) of the ADCs, MW_{mAb} is the MW of the mAb, while MW_{drug} represents the MW of the LP systems.

Part II

Development of a pseudo-atom approach to optimize reactive docking of covalent inhibitors

Chapter 5

Introduction

Research on pharmacological small molecules is mostly focused on finding and improving reversible medications. Covalent medications, which create a link with their target protein, were not well studied until recently because of potential toxicity issues. From direct tissue injury to protein haptenization, the production of chemically reactive drug metabolites from reversible medications has been considered a risk concern in drug development.[201] However, there are numerous instances of extremely successful and safe covalent pharmaceuticals, despite the pharmaceutical industry's historical reluctance to covalent drug research.[202]

5.1 Targeted covalent inhibitors

In recent years, it has become clear that the advantages of covalent and non-covalent drug binding can be combined by designing ligands with finely tuned electrophilicity and high structural complementarity to their targets. Such compounds are known as targeted covalent inhibitors (TCIs).[203] TCIs have particular functional groups that are intended to react with a matching location in the target, usually the side chain of an amino acid. The side chain could be a non-catalytic site near the binding pocket or a component of the target

enzyme's catalytic machinery.[204] The development of covalent inhibitors offers the potential for enhanced efficacy compared to their non-covalent counterparts, as the formation of a covalent bond provides substantially greater interaction strength than typical reversible binding forces. Because the target protein remains inhibited until it is degraded and subsequently resynthesized, covalent inhibitors generally exhibit a prolonged duration of action. This extended activity arises from the typically irreversible nature of the covalent bond formed between the inhibitor and the protein.[205] Notable examples include epidermal growth factor receptor (EGFR) and KRAS, where achieving high selectivity using traditional reversible ligands is particularly challenging.[206,207] In the TCI sector, covalent engagement of non-catalytic cysteines has become the most popular targeting strategy. Acrylamides and related α,β-unsaturated amides are by far the most prevalent warhead class in this regard. There have been innumerable studies employing α,β-unsaturated amides to target cysteines since the FDA approved the first acrylamide-based covalent kinase inhibitors, ibrutinib (2, targeting Bruton's tyrosine kinase (BTK)) and afatinib (1, targeting EGFR). Several factors contributed to the remarkable success of α,β-unsaturated amides in the field of cysteine-targeted covalent drugs and chemical probes, including the moderate and tunable reactivity of these functional groups, the generally high chemoselectivity for cysteine, the relative stability of the products of the β-thioether reaction, and the chemical accessibility via simple amide coupling.[208,209] Alongside the interest in TCIs, there has been a renewed focus on computational techniques that can simulate these inhibitors, despite the particular modeling difficulties they present.[210] Nowadays, there are plenty of methods that tend to focus on the evaluation of the endpoint of the reaction, like tethered or biased approaches that model the ligand in the bound state.[211–214] Although these techniques are appropriate for analyzing well-defined groups of compounds into well-characterized binding sites, they are not the best for predicting binding sites across nucleophilic residues in proteins, much less proteomes.[215]

5.2 Reactive docking

The reactive docking method is formalized in the work presented by Bianco et al.[216] This approach was specifically developed to enable both the prediction of reactive binding sites and the identification of optimal reactive ligands. The method employs a modified version of the AutoDock4 force field,[217], in which the parameters are adjusted to favor near-attack conformations (NAC).[218] The NAC concept, originally described by Hur and co-workers, refers to ground-state conformations of the enzyme-substrate complex that closely resemble the geometry of the transition state, particularly in terms of key interatomic distances and angles. Because the NAC represents a critical thermodynamic checkpoint along the reaction pathway and involves minimal geometric rearrangements of either the ligand or the target, this species is particularly suitable for modeling the ligand-target interaction. Accordingly, in this framework, the ligand is docked in its unmodified form, with the reactive warhead intact before covalent bond formation. As described in the reactive docking method,[216] the AutoDock standard forcefield[219] is modified by adding a 13-7 Lennard-Jones potential,[220] chosen to give a slightly narrower potential with respect to the standard van der Waals, between the ligand reactive atom i^* and the receptor reactive atom j^* to define incipient bond formation.

$$\Delta_{react} = \frac{E_{i*j*}}{r_{i*j*}^{13}} - \frac{F_{i*j*}}{r_{i*j*}^{7}} \tag{5.1}$$

Where

$$E_{i*j*} = 7\left(\frac{\epsilon r_{eq}^{13}}{13-7}\right) \tag{5.2}$$

$$F_{i*j*} = 13\left(\frac{\epsilon r_{eq}^{7}}{13-7}\right) \tag{5.3}$$

r_{eq} represents the equilibrium distance (Å) and ϵ represents the equilibrium energy (kcal/mol). The reactive potential is softened in accordance with the standard AutoDock

force field description.[221] The introduction of this "reactive interaction" causes atoms within two bonds of the reacting pair to adopt distances shorter than their standard van der Waals equilibrium separations Figure 5.1. To avoid unrealistic steric compression, the equilibrium distances for pseudo 1-3 and 1-4 interactions surrounding the reactive atoms are scaled by empirical weighting factors $W_{1,3}$ and $W_{1,4}$, respectively.

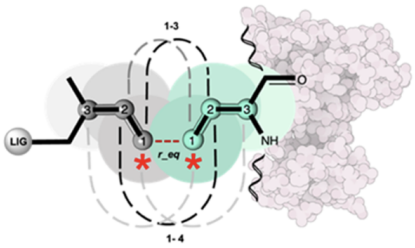


Figure 5.1: An illustration of the NAC-like state and the reactive docking model that displays the pseudo 1-3 and 1-4 interactions together with their van der Waals radii.[216]

During the docking process, ligands are allowed to sample a wide range of binding conformations, including orientations that do not support covalent bond formation. The target residue involved in the reaction is modeled as flexible, enabling the exploration of binding modes that may or may not lead to a productive covalent interaction.[217] The distance between the reactive atoms of the ligand and residue is measured by analyzing the result with the best docking score after dockings are finished, which establishes whether the reaction is covalent or not.

5.3 Aim of the project

During my six-month visiting research period at The Scripps Research Institute, joining the lab of Prof. Stefano Forli, I worked on a project focused on the improvement of the existing reactive docking method discussed above. The primary objective of this project was to refine and extend the existing reactive docking methodology by introducing Psuedo-Atoms

(PAs) on the ligand warhead. In this approach, PAs are employed to encode the geometry and spatial orientation required for covalent bond formation, thereby providing an implicit description of the reaction trajectory in order to predict the optimal NAC. To implement this strategy, a series of Python scripts was developed to automatically compute the required reference coordinates and place the PAs onto the ligand structures in a consistent and reproducible manner. Following the integration of these modifications into the workflow, comprehensive docking simulations were performed to assess the performance of the enhanced method on the reactive docking dataset (see Supporting Information). This evaluation enabled the systematic analysis of the accuracy and applicability of the modified reactive docking protocol.

Chapter 6

Results and discussion

This section describes the computational procedures employed to investigate ligand-protein interactions and reactive binding modes. The protocols detailed below outline the preparation of structural input files, parameter assignment, docking setup, and post-processing steps used to analyze and compare the resulting complexes. Two main warheads were studied in this work, acrylamides and chloroacetamides.

Warhead	Training set	Test set
Acrylamides	20	104
Chloroacetamides	20	6

Table 6.1: Dataset studied for PAs method.

6.1 Ligands preparation

The restored ligands' SMILES were used to compute 3D coordinates using Scrubber (`https://github.com/forlilab/molscrub`), modeling protonation states at pH 7.4, and generating 3D coordinates using RDKit's ETKDGv3 and UFF minimization. PAs calculations were performed using class_PseudoAtom.py (see section 7.4) starting from

the 3D coordinates generated with Scrubber and with SMARTS patterns to define the anchoring atom on which to place PAs. Partial charges, torsions, and standard and reactive atom type parameters were assigned according to the AutoDock protocol using Meeko (`https://github.com/forlilab/Meeko`) with SMARTS patterns to define the warhead atoms and assign the reactive docking force field parameters (Equation 5.1).[222]

6.2 Receptors preparation

Target protein structures were obtained from the Protein Data Bank, and hydrogen atoms were added using Reduce2 (`https://github.com/cctbx/cctbx_project/tree/master/mmtbx/reduce`). Protonation and tautomeric states of histidine residues were assigned by Reduce2 based on local hydrogen-bonding patterns and steric considerations. Explicit pKa calculations were not carried out; histidines were treated as neutral unless indicated otherwise by their structural environment. Ligands were processed with Meeko to assign partial charges, torsional degrees of freedom, and both standard and reactive atom type parameters in accordance with the AutoDock protocol.[222] For each reactive site, a cubic docking box with a side length of 30 Å (corresponding to 80 grid points in the AutoGrid parameter file[217]) was centered on the Cα atom of the target residue.

6.3 Reactive docking

All cysteine residues present in the receptors' structure were considered for docking calculation (except for those involved in disulfide bridges). Each ligand was docked against the individual target residues selected for evaluation. During docking, ligands were modeled in their unmodified form, with reactive warheads preserved, using a standard docking approach (untethered), while the side chain of the target residue was treated as flexible.[217] Docking simulations were performed with AutoDock-GPU,[221] generating 50

poses per ligand using the default Lamarckian Genetic Algorithm (LGA) parameters.[222] In the standard reactive docking protocol, a ligand-residue pair was considered reactive if the distance between their respective reactive atoms in the lowest-energy pose was ≤ 2.0 Å. In the reactive docking protocol incorporating PAs, two distances are evaluated: the distance between the reactive atom on the flexible residue and the PAs on the ligand warhead, and the distance between the actual atoms of the reactive pair. A ligand-residue pair is considered reactive only if both distances satisfy the threshold in the lowest-energy pose. Residues were subsequently ranked according to their predicted reactivity, based on the most favorable binding energy of ligands predicted to react with them. A detailed, step-by-step description of the reactive docking protocol, including setup and execution, is available at `https://github.com/forlilab/Meeko`.

Docking simulations were performed on the Garibaldi High Performance Computing (HPC) cluster, available at The Scripps Research Institute, La Jolla, CA, USA, consisting of:

- 62 Dell Poweredge R420 servers with two intel quad core E5-2450 processors, 48GB RDIMM memory;
- 8 Dell Poweredge R720 servers with two intel quad core E5-2650 processors, 126GB RDIMM memory;
- 64 Dell Poweredge M610 blades with two 2.40 GHz Intel quad core E5530 XEON-EMT processors, 48GB of ECC DDR3 memory;
- 96 Dell Poweredge M610 blades with two 2.27 GHz Intel quad core E5520 XEON-EMT processors, 48GB of ECC DDR3 memory;
- 96 Dell Poweredge M600 blades with two 2.66 GHz Intel quad core E5430 XEON-EMT processors, 32GB of ECC DDR2 memory.

6.4 Parameters selection

As done for the standard reactive docking protocol,[216] the training set was used to calibrate the PAs reactive docking parameters. The r_{eq} value for the pair 1S4-C1, corresponding to the sulfur atom of the cysteine side chain and the electrophilic carbon of the ligand warhead, was set to 1.8 Å to approximate a typical covalent C-S bond length, and the equilibrium distance for the pair 1S4-XX, that in this particular protocol is the actual "reactive" pair, was set to 0.25 Å. The ϵ value was obtained after an exhaustive search (from 2.0 to 10 kcal/mol^{-1}). During the calibration, each parameter set was evaluated against the full training set by docking every ligand to its corresponding protein target while considering all cysteine residues as potential reactive sites.

6.5 Calibration and success rate

In contrast to the standard reactive docking protocol, in this study, the scaling factors applied to the $W_{1,3}$ and $W_{1,4}$ distances were intentionally reduced. This adjustment was introduced to preserve, at least partially, the steric constraints imposed by the native van der Waals radii of the atoms involved. The rationale behind this choice is that maintaining a realistic steric environment can enhance the ability of the model to capture the directional characteristics of covalent bond formation in the PAs reactive docking approach. In the conventional protocol, by contrast, the $W_{1,3}$ and $W_{1,4}$ distances are typically scaled to prevent steric clashes that may arise from the shortened interatomic distances around the reactive pair. According to the results analysis, accuracy really begins to decline when ϵ values fall below 2.0 kcal/mol.

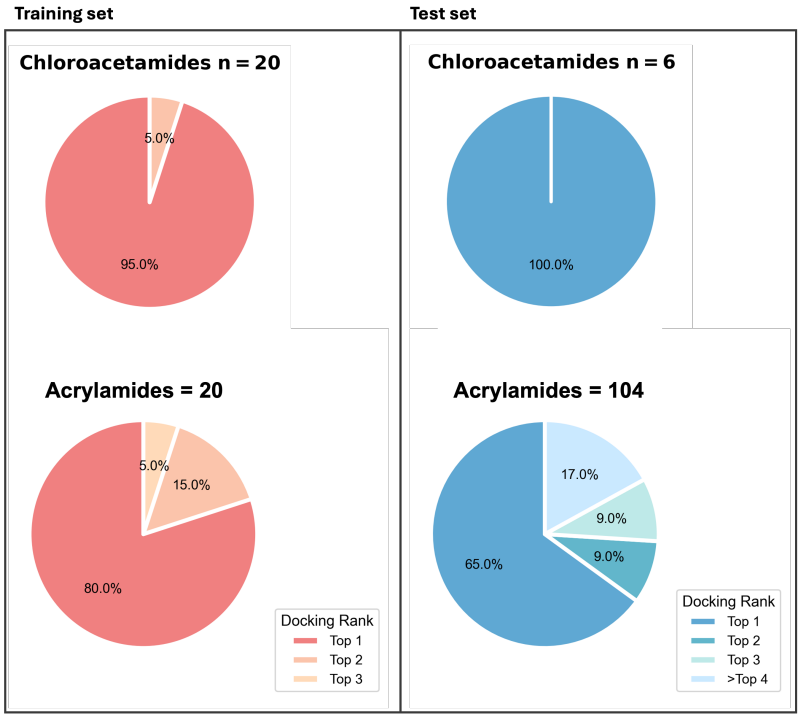
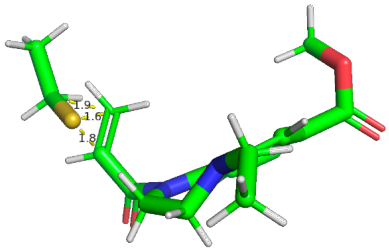
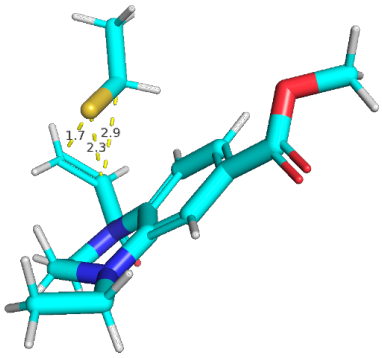


Figure 6.1: Training and test set docking performance in reactive cysteine predictions.

Preliminary results are highly encouraging. For chloroacetamide ligands, the method achieved a success rate of 95% on the training set and 100% on the test set. In the case of acrylamides, the corresponding success rates were 80% and 65%, respectively. Overall, the PAs reactive docking approach reached an accuracy of 87.5% in correctly identifying the labeled cysteine as the top-ranked residue within the training set, and 83.6% within the top three predictions in the test set. It is important to note that these success rates were obtained by simultaneously evaluating two geometric constraints, namely 1S4-C1 and 1S4-XX distances. Consequently, the top-ranked docking poses not only correspond to reactive configurations but also exhibit the appropriate geometry for covalent bond formation.



(a)



(b)

Figure 6.2: Comparison between a docked pose obtained with standard reactive docking 6.2a and a docked pose obtained with reactive docking including PAs 6.2b.

6.6 Virtual screening

In addition, a virtual screening study was conducted. Based on experimental findings published by Resnik et al.,[223] a virtual screening set was constructed using a library of 993 reactive fragments functionalized with moderate electrophiles, such as acrylamides (n = 241, 24%) and chloroacetamides (n = 752, 76%) that target cysteines. Due to time constraints, only 1 out of 10 cysteine-containing receptors was studied across chloroacetamide warheads. The entire fragment library was docked against the target Nudix Hydrolase 7 (NUDT7)[224] in a blind docking fashion using the set of parameters derived from the calibration step. This process took into account all solvent-accessible cysteines for each target and predicted the most likely ligands to react as well as the most likely residues to

be modified. The method was able to predict the correct cysteine, highlighting its reactive residue prediction capacity. To evaluate the virtual screening results, the top 0.5%, 1.0%, and 10% docking poses were considered.

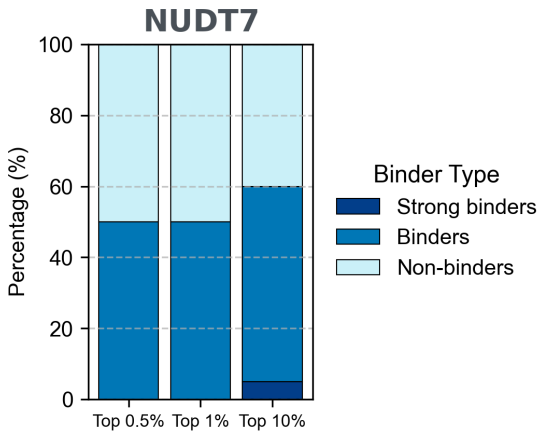


Figure 6.3: Virtual screening success rates in binders recovery (0.5, 1.0, and 10% fractions).

The PAs reactive docking protocol achieved a true-positive rate of 50% within both the top 0.5% and top 1.0% ranked poses, and an overall rate of 60% within the top 10% (comprising 5% strong binders and 55% binders). In comparison, the standard reactive docking protocol applied to the NUDT7 dataset yielded 100% true positives in the top 0.5% (50% of which were strong binders), 70% in the top 1%, and 40% in the top 10%. Despite the inherent difficulty of evaluating virtual screening performance with fragment-based libraries, as opposed to drug-like compounds[225], these results indicate that the PAs protocol can successfully predict both reactive residues and corresponding ligands with an enhanced geometric approach for the reactive pair, supporting its potential utility for future large-scale virtual screening applications.

Chapter 7

Conclusions

In this preliminary study, a potential enhancement to the existing reactive docking method was investigated through the introduction of pseudo-atoms, designed to improve the geometric accuracy of ligand-residue reactive poses. The approach was evaluated on two electrophilic warheads, chloroacetamides and acrylamides, by redocking each ligand into its corresponding protein target to assess the protocol's ability to correctly identify the reactive cysteine with an appropriate covalent geometry. A subsequent virtual screening analysis on the pyrophosphatase NUDT7 was performed to further evaluate the method's predictive performance in identifying both the reactive residue and its cognate ligand. Overall, the results obtained with the new pseudo-atom protocol are encouraging, particularly for the chloroacetamide series. Nevertheless, further optimization is ongoing to improve performance across both warheads, including a more exhaustive search of parameter combinations for ϵ, $W_{1,3}$, and $W_{1,4}$, as well as exploring the introduction of additional pseudo-atoms on the reactive residue side chain to better capture reaction directionality.

Appendix A

The following Python scripts were developed using *Jupyter Notebook* to implement and evaluate the machine learning models described in Chapter 2. Each section corresponds to a specific stage of the workflow, including data preprocessing, model training, and hyperparameter optimization. All scripts were executed using Python 3.10 (Anaconda distribution) and rely on the libraries *rdkit*, *scikit-learn*, *xgboost*, *numpy*, *pandas*, *meeko*, *molscrub*, and *matplotlib*.

7.1 Data Preprocessing Scripts

This section includes the Python scripts used to clean, normalize, and prepare the dataset before model training. The preprocessing steps include handling of missing values, feature scaling, and dataset partitioning into training and test sets.

```
from rdkit import Chem
from rdkit.Chem import MolFromSmiles
from rdkit.Chem import Descriptors
from rdkit.Chem.Descriptors import CalcMolDescriptors
import pandas as pd

# Load cysteine bioconjugation dataset
df_cys_bioconjugations = pd.read_excel("cys_dataset.xlsx")

```

```
# Load SMILES data
df_smiles = pd.read_excel("SMILES_Cys.xlsx")

# Convert SMILES strings to RDKit Mol objects
mols = [Chem.MolFromSmiles(s) for s in df_smiles["Smiles"]]

# Calculate molecular descriptors
descriptors = []
for mol in mols:
    desc = CalcMolDescriptors(mol)
    descriptors.append(desc)

df_desc = pd.DataFrame(descriptors)
df_desc['ID'] = df_smiles['ID']
# Combine descriptors with bioconjugation data
df_final = pd.merge(df_cys_bioconjugations, df_desc, on='ID')
```

Listing 7.1: Python script for molecular descriptor calculation on cysteine bioconjugation dataset.

```
from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()
cols_to_encode = ["mAb", "Reductant", "Buffer", "pH", "Method"]

for col in cols_to_encode:
    df_final[col] = le.fit_transform(df_final[col])
```

Listing 7.2: Python script to encode categorical variable using the *LabelEncoder* method from scikit-learn.

```
from sklearn.feature_selection import VarianceThreshold

# Split predictors and target
X = df_final.drop(["DAR", "Outcome"], axis=1)
```

```
y = df_final["Outcome"]

# Set variance threshold and feature selection
sel = VarianceThreshold(threshold=(.8 * (1 - .8)))
X2 = sel.fit_transform(X)

# Retrieve names of features that were kept
concol = X.columns[sel.get_support()].tolist()

# Rebuild a DataFrame with selected features
df_X = pd.DataFrame(X2, columns=concol)
```

Listing 7.3: Python script to apply a feature selection using the *VarianceThreshold* method from scikit-learn.

```
from sklearn.preprocessing import MinMaxScaler, StandardScaler


std_scaler = StandardScaler() # StandardScaler achieved the best results
    overall
norm_scaler = MinMaxScaler()

X_scaled = std_scaler.fit_transform(df_X)
X_norm = norm_scaler.fit_transform(df_X)
```

Listing 7.4: Python script to apply a feature scaling using the *StandardScaler* and *MinMaxScaler* methods from scikit-learn.

7.2 Model Training and Evaluation Scripts

The following scripts were used to train and evaluate multiple machine learning classifiers, including Logistic Regression, k-NN, Decision Tree, SVM, LDA, and XGBoost. Model performance was assessed using accuracy, recall, precision, and F1-score averaged across

repeated runs.

```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, recall_score, f1_score,
    precision_score
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
import xgboost as xgb

# Define models
models = {
    "Logistic Regression": LogisticRegression(class_weight='balanced'),
    "k-NN": KNeighborsClassifier(),
    "Decision Tree": DecisionTreeClassifier(),
    "SVM": SVC(class_weight='balanced'),
    "LDA": LinearDiscriminantAnalysis(),
    "XGBoost": xgb.XGBClassifier()
}

# Define the number of runs and storage
n_runs = 20
results = []

# Evaluation loop
for name, model in models.items():
    scores, recalls, f1s, precisions = [], [], [], []

    for _ in range(n_runs):
        X_train, X_test, y_train, y_test = train_test_split(
            X_scaled, y, shuffle=True, random_state=None, test_size=0.20)
```

```
        model.fit(X_train, y_train)
        y_pred = model.predict(X_test)
        if name in ["Logistic Regression", "SVM"]:
            r = recall_score(y_test, y_pred, zero_division=0)
            f = f1_score(y_test, y_pred, zero_division=0)
            p = precision_score(y_test, y_pred, zero_division=0)
        else:
            r = recall_score(y_test, y_pred)
            f = f1_score(y_test, y_pred)
            p = precision_score(y_test, y_pred)

        # Store metrics
        scores.append(accuracy_score(y_test, y_pred))
        recalls.append(r)
        f1s.append(f)
        precisions.append(p)

    # Append averaged results
    results.append({
        "Model": name,
        "Accuracy": np.mean(scores),
        "Recall": np.mean(recalls),
        "F1": np.mean(f1s),
        "Precision": np.mean(precisions)
    })

# Results DataFrame
df_results = pd.DataFrame(results)
```

Listing 7.5: Python script to train and evaluate classifiers' performance across 20 runs.

```
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
```

```
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error
from sklearn.linear_model import Ridge
from sklearn.neighbors import KNeighborsRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
import xgboost as xgb
import pandas as pd
import numpy as np

scaler = MinMaxScaler()
X_norm = scaler.fit_transform(X)

# Define models
models = {
    "XGBoost Regressor": xgb.XGBRegressor(),
    "Ridge Regressor": Ridge(),
    "Random Forest Regressor": RandomForestRegressor(),
    "Decision Tree Regressor": DecisionTreeRegressor(),
    "k-NN Regressor": KNeighborsRegressor()
}

# Define the number of runs and storage
n_runs = 20
results = []

# Evaluation loop
for name, model in models.items():
    r2_scores, maes, mses, stds = [], [], [], []

    for _ in range(n_runs):
        X_train, X_test, y_train, y_test = train_test_split(
            X_norm, y, shuffle=True, random_state=None, test_size=0.20
```

```
        )

        # Train model
        model.fit(X_train, y_train)
        y_pred = model.predict(X_test)

        # Store metrics
        r2_scores.append(r2_score(y_test, y_pred))
        maes.append(mean_absolute_error(y_test, y_pred))
        mses.append(mean_squared_error(y_test, y_pred))
        stds.append(np.std(y_pred))

    # Append averaged results
    results.append({
        "Model": name,
        "R^2": np.mean(r2_scores),
        "MAE": np.mean(maes),
        "MSE": np.mean(mses),
        "Std. Dev. (pred)": np.mean(stds)
    })

# Results DataFrame
df_results = pd.DataFrame(results)
```

Listing 7.6: Python script to train and evaluate regressors' performance across 20 runs.

7.3 Hyperparameter Optimization Scripts

The script below was used to perform the grid search to select the optimal hyperparameters for the best-performing model XGBoost regressor.

```
from xgboost import XGBRegressor
```

```
from sklearn.model_selection import GridSearchCV

xgb = XGBRegressor()

parameters = {"colsample_bytree": [0.6, 0.8, 1],
              "max_depth": range(3, 25),
              "min_child_weight": range(5, 10, 1),
              "n_estimators": range(40, 300, 20),
              "learning_rate": [0.1, 0.01, 0.05]}
grid_search = GridSearchCV(estimator=xgb,
                           param_grid=parameters,
                           scoring='r2',
                           n_jobs=4,
                           cv=5,
                           verbose=True)
grid_search.fit(X_norm, y)
```

Listing 7.7: Python script for hyperparameter optimization of the model *XGBRegressor*.

7.4 Pseudo-atom coordinates calculation

The Python script in this section reports an example to calculate PAs coordinates.

```
import import_ipynb
from class_PseudoAtom import PseudoAtoms
from rdkit import Chem

pseudo = PseudoAtoms(aromatics=False, acrylamide_addition=True,
    acrylamide_lenght=1.7)

# Load molecules as sdf
suppl_mol = Chem.SDMolSupplier("ligand_scrubbed.sdf", removeHs=False)
```

```
9
10 suppl_list = []
11 names = []
12 pseudo_coord = {}
13
14 # Get each molecule and ID
15 for mol in suppl_mol:
16     if mol is None:
17         continue
18     suppl_list.append(mol)
19     names.append(mol.GetProp("_Name"))
20
21 # Define acrylamide warhead with SMARTS and convert it to a RDKit mol format
22 smarts_acrylamide = "[N]C(C=C)=O"
23 acrylamide_pattern = Chem.MolFromSmarts(smarts_acrylamide)
24
25 # Check the acrylamide group in each ligand molecule and calculate pseudo-
       atoms coordinates if found
26 for m, n in zip(suppl_list, names):
27     conf = m.GetConformer()
28     acrylamide_matches = m.GetSubstructMatches(acrylamide_pattern)
29     if acrylamide_matches:
30         for match in acrylamide_matches:
31             pos_C1 = conf.GetAtomPosition(match[1])
32             pos_C2 = conf.GetAtomPosition(match[2])
33             pos_C3 = conf.GetAtomPosition(match[3])
34             new_pos1, new_pos2 = pseudo._acrylamide_michael_pos(
35                 pos_C1, pos_C2, pos_C3)
36             pseudo_coord[n] = new_pos1, new_pos2
```

Listing 7.8: Python script to calculate pseudo-atoms coordinates using class_PseudoAtom.

7.5 Categorical features encoding

The figures below represent the encoded categorical features involved in the SHAP analysis.

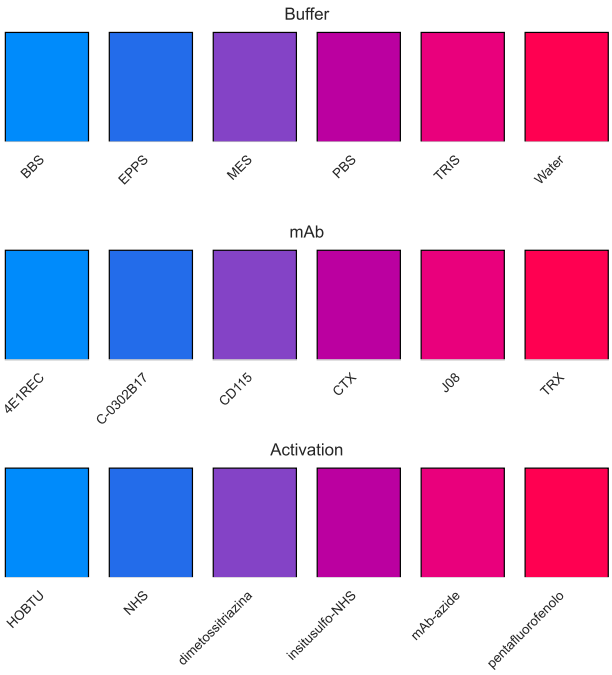


Fig. S7: SHAP Color Mapping for Lysine categorical features.

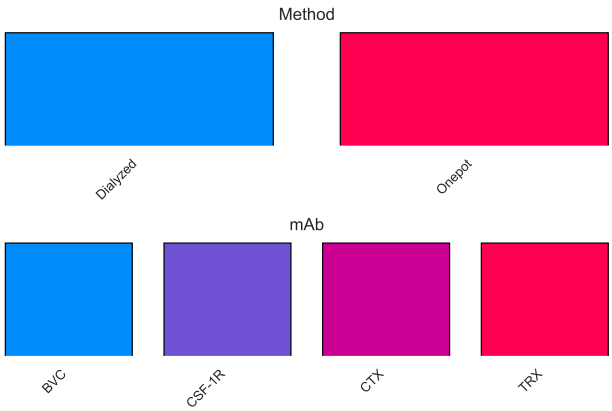


Fig. S8: SHAP Color Mapping for Cysteine categorical features.

Bibliography

(1) Ocana, A.; Pandiella, A.; Privat, C.; Bravo, I.; Luengo-Oroz, M.; Amir, E.; Gyorffy, B. *Biomark. Res.* **2025**, *13*, 45.

(2) Ferreira F. J. N.; Carneiro, A. S. *ACS Omega* **2025**, *10*, 23889–23903.

(3) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477.

(4) Blanco-González, A.; Cabezón, A.; Seco-González, A.; Conde-Torres, D.; Antelo-Riveiro, P.; Piñeiro, Á.; Garcia-Fandino, R. *Pharmaceuticals* **2023**, *16*.

(5) McCulloch, W. S.; Pitts, W. *The bulletin of mathematical biophysics* **1943**, *5*, 115–133.

(6) Morris, R. *Wiley* **1999**, *50*, 437.

(7) Wilson, C. J. *Brain Res Bull* **1999**, *50*, 335.

(8) Turing, A. M. *MIND* **1950**, *59*, 433–460.

(9) Goodfellow, I.; Bengio, Y.; Courville, A., *Deep Learning*, `http://www.deeplearningbook.org`; MIT Press: 2016.

(10) Khanna, R.; Awad, M., *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*; Apress: 2015.

(11) Cunningham, P.; Cord, M.; Delany, S. J. In *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*, Cord, M., Cunningham, P., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008, pp 21–49.

(12) Ghahramani, Z. In *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*, Bousquet, O., von Luxburg, U., Rätsch, G., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2004, pp 72–112.

(13) Zhu, X. *Semi-Supervised Learning Literature Survey*; tech. rep.; University of Wisconsin-Madison Department of Computer Sciences, 2005.

(14) Sutton Richard S.; Barto, A. G., *Reinforcement Learning: An Introduction*; MIT Press: 1998.

(15) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. *J. Chem. Inf. Model.* **2018**, *58*, 287–298.

(16) Cao, D. S.; Xiao, N.; Xu, Q. S.; Chen, A. F.; Liang, Y. Z. *J. Chemometrics* **2017**, *31*, e2930.

(17) Guyon, I.; Elisseeff, A. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.

(18) Bergstra, J.; Bengio, Y. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.

(19) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. *Front. Environ. Sci.* **2016**, *3*, 80.

(20) Ganji, Z.; Hakkak, M. A.; Zare, H. *Neurol. Res.* **2022**, *44*, 1142–1149.

(21) Mehrpour, O. e. a. *BMC Med. Inform. Decis. Mak.* **2023**, *23*, 137.

(22) Sterkenburg T. F.; Grünwald, P. D. *Synthese* **2021**, *199*, 9979–10015.

(23) Nazir, R.; Bucaioni, A.; Pelliccione, P. *Journal of Systems and Software* **2024**, *207*, 111860.

(24) Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V. I. *BMC Medical Informatics and Decision Making* **2020**, *20*, DOI: 10.1186/s12911-020-01332-6.

(25) Semnani, P.; Gubernatis, J. E.; Lookman, T.; Bartel, C. J. *The Journal of Physical Chemistry C* **2024**, *128*, 17431–17443.

(26) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A., et al. *Nature Biotechnology* **2019**, *37*, 1038–1040.

(27) Mater, A. C.; Coote, M. L. *Journal of Chemical Information and Modeling* **2019**, *59*, 2545–2559.

(28) Müller, A. T.; Hiss, J. A.; Schneider, G. *Journal of Chemical Information and Modeling* **2018**, *58*, PMID: 29355319, 472–479.

(29) Grisoni, F.; Moret, M.; Lingwood, R.; Schneider, G. *Journal of Chemical Information and Modeling* **2020**, *60*, PMID: 31904964, 1175–1183.

(30) Wu, X.; Kumar, V.; Quinlan, J. R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G. J.; Ng, A.; Liu, B.; Yu, P. S.; Zhou, Z.-H.; Steinbach, M.; Hand, D. J.; Steinberg, D. *Knowledge and Information Systems* **2008**, *14*, 1–37.

(31) Neugebauer, A.; Hartmann, R. W.; Klein, C. D. *Journal of Medicinal Chemistry* **2007**, *50*, PMID: 17705363, 4665–4668.

(32) Cortes, C.; Vapnik, V. *Machine Learning* **1995**, *20*, 273–297.

(33) Schneider, H. W.; Raiol, T.; Brigido, M. M.; Walter, M. E. M. T.; Stadler, P. F. *BMC Genomics* **2017**, *18*, 804.

(34) Whitley, D. *Statistics and Computing* **1994**, *4*, 65–85.

(35) Meiler, J.; Will, M. *Journal of the American Chemical Society* **2002**, *124*, PMID: 11866596, 1868–1870.

(36) Polikar, R. In *Ensemble Machine Learning: Methods and Applications*, Zhang, C., Ma, Y., Eds.; Springer New York: New York, NY, 2012, pp 1–34.

(37) Breiman, L. *Machine Learning* **1996**, *24*, 123–140.

(38) Freund, Y.; Schapire, R. *icml* **1996**, *96*, 148–156.

(39) Breiman, L. *Machine Learning* **2001**, *45*, 5–32.

(40) Bernal, A.; Crammer, K.; Hatzigeorgiou, A.; Pereira, F. *PLoS Comput Biol* **2007**, *3*, e54.

(41) Xue, C.; Zhou, M. *Biology* **2025**, *14*, DOI: 10.3390/biology14050554.

(42) Ritchie, M. D.; White, B. C.; Parker, J. S.; Hahn, L. W.; Moore, J. H. *BMC Bioinformatics* **2003**, *4*, 28.

(43) Rathee, M.; Kumar, T. *International Journal of Applied Evolutionary Computation* **2014**, *5*, 84–108.

(44) Jumper, J. et al. *Nature* **2021**, *596*, 583–589.

(45) Varadi, M. et al. *Nucleic Acids Research* **2022**, *50*, D439–D444.

(46) Qureshi, R.; Irfan, M.; Gondal, T. M.; Khan, S.; Wu, J.; Hadi, M. U.; Heymach, J.; Le, X.; Yan, H.; Alam, T. *Heliyon* **2023**, *9*, e17575.

(47) Ehrlich, P. *Br Med J* **1913**, *2*, 353–359.

(48) Matsuda, Y.; Chang, J. R.; Mendelsohn, B. A. *Chembiochem* **2025**, e2500305.

(49) Chen, B.; Zheng, X.; Wu, J.; Chen, G.; Yu, J.; Xu, Y.; Wu, W. K. K.; Tse, G. M. K.; To, K. F.; Kang, W. *Molecular Cancer* **2025**, *24*, 279.

(50) Ricart, A. D. *Clinical Cancer Research* **2011**, *17*, 6417–6427.

(51) Senter, P. D.; Sievers, E. L. *Nature Biotechnology* **2012**, *30*, 631–637.

(52) Lambert, J. M.; Chari, R. V. J. *Journal of Medicinal Chemistry* **2014**, *57*, PMID: 24967516, 6949–6964.

(53) Lamb, Y. N. *Drugs* **2017**, *77*, 1603–1610.

(54) Dhillon, S. *Drugs* **2018**, *78*, 1763–1767.

(55) Dean, A. Q.; Luo, S.; Twomey, J. D.; Zhang, B. *MAbs* **2021**, *13*, 1951427.

(56) Monk, B. J.; Enomoto, T.; Kast, W. M.; McCormack, M.; Tan, D. S. P.; Wu, X.; González-Martín, A. *Cancer Treat Rev* **2022**, *106*, 102385.

(57) Maecker, H.; Jonnalagadda, V.; Bhakta, S.; Jammalamadaka, V.; Junutula, J. R. *MAbs* **2023**, *15*, 2229101.

(58) Blair, H. A. *Drugs* **2025**, *85*, 965–975.

(59) Blair, H. A. *Drugs* **2025**, *85*, 1171–1176.

(60) Phuna, Z. X.; Kumar, P. A.; Haroun, E.; Dutta, D.; Lim, S. H. *Life Sciences* **2024**, *347*, 122676.

(61) Dumontet, C.; Reichert, J. M.; Senter, P. D.; Lambert, J. M.; Beck, A. *Nature Reviews Drug Discovery* **2023**, *22*, 641–661.

(62) Baah, S.; Laws, M.; Rahman, K. M. *Molecules* **2021**, *26*.

(63) Sheyi, R.; de la Torre, B. G.; Albericio, F. *Pharmaceutics* **2022**, *14*, DOI: 10.3390/pharmaceutics14020396.

(64) Balamkundu, S.; Liu, C.-F. *Biomedicines* **2023**, *11*.

(65) Fu, Z.; Li, S.; Han, S.; Shi, C.; Zhang, Y. *Signal Transduction and Targeted Therapy* **2022**, *7*, 93.

(66) He, J.; Zeng, X.; Wang, C.; Wang, E.; Li, Y. *MedComm (2020)* **2024**, *5*, e671.

(67) Tai, Y.-T. et al. *Blood* **2014**, *123*, 3128–3138.

(68) Radocha, J.; van de Donk, N. W. C. J.; Weisel, K. *Cancers (Basel)* **2021**, *13*.

(69) Oostra, D. R.; Macrae, E. R. *Breast Cancer (Dove Med Press)* **2014**, *6*, 103–113.

(70) Köhler, G.; Milstein, C. *Nature* **1975**, *256*, 495–497.

(71) Mach, J. P.; Carrel, S.; Forni, M.; Ritschard, J.; Donath, A.; Alberto, P. *N Engl J Med* **1980**, *303*, 5–10.

(72) Corraliza-Gorjón, I.; Somovilla-Crespo, B.; Santamaria, S.; Garcia-Sanz, J. A.; Kremer, L. *Front Immunol* **2017**, *8*, 1804.

(73) Argollo, M.; Kotze, P. G.; Kakkadasam, P.; D'Haens, G. *Nature Reviews Gastroenterology & Hepatology* **2020**, *17*, 702–710.

(74) Troisi, M.; Marini, E.; Abbiento, V.; Stazzoni, S.; Andreano, E.; Rappuoli, R. *Frontiers in Microbiology* **2022**, *Volume 13 - 2022*, DOI: 10.3389/fmicb.2022.1080059.

(75) Tsuchikama, K.; An, Z. *Protein Cell* **2016**, *9*, 33–46.

(76) Su, Z.; Xiao, D.; Xie, F.; Liu, L.; Wang, Y.; Fan, S.; Zhou, X.; Li, S. *Acta Pharm Sin B* **2021**, *11*, 3889–3907.

(77) King, H. D.; Dubowchik, G. M.; Mastalerz, H.; Willner, D.; Hofstead, S. J.; Firestone, R. A.; Lasch, S. J.; Trail, P. A. *Journal of Medicinal Chemistry* **2002**, *45*, PMID: 12213074, 4336–4343.

(78) Finbloom, D. S.; Abeles, D.; Rifai, A.; Plotz, P. H. *J Immunol* **1980**, *125*, 1060–1065.

(79) Zhao, R. Y.; Wilhelm, S. D.; Audette, C.; Jones, G.; Leece, B. A.; Lazar, A. C.; Goldmacher, V. S.; Singh, R.; Kovtun, Y.; Widdison, W. C.; Lambert, J. M.; Chari, R. V. J. *Journal of Medicinal Chemistry* **2011**, *54*, PMID: 21517041, 3606–3623.

(80) Lyon, R. P.; Bovee, T. D.; Doronina, S. O.; Burke, P. J.; Hunter, J. H.; Neff-LaFord, H. D.; Jonas, M.; Anderson, M. E.; Setter, J. R.; Senter, P. D. *Nature Biotechnology* **2015**, *33*, 733–735.

(81) Kern, J. C. et al. *Journal of the American Chemical Society* **2016**, *138*, PMID: 26745435, 1430–1445.

(82) Teicher, B. A.; Morris, J. *Curr Cancer Drug Targets* **2022**, *22*, 463–529.

(83) Khongorzul, P.; Ling, C. J.; Khan, F. U.; Ihsan, A. U.; Zhang, J. *Mol Cancer Res* **2019**, *18*, 3–19.

(84) Senter, P. D. *Curr Opin Chem Biol* **2009**, *13*, 235–244.

(85) DiJoseph, J. F.; Dougher, M. M.; Kalyandrug, L. B.; Armellino, D. C.; Boghaert, E. R.; Hamann, P. R.; Moran, J. K.; Damle, N. K. *Clin Cancer Res* **2006**, *12*, 242–249.

(86) Migliorini, F.; Cini, E.; Dreassi, E.; Finetti, F.; Ievoli, G.; Macrì, G.; Petricci, E.; Rango, E.; Trabalzini, L.; Taddei, M. *Chem. Commun.* **2022**, *58*, 10532–10535.

(87) Bargh, J. D.; Isidro-Llobet, A.; Parker, J. S.; Spring, D. R. *Chem. Soc. Rev.* **2019**, *48*, 4361–4374.

(88) Gondi, C. S.; Rao, J. S. *Expert Opin Ther Targets* **2013**, *17*, 281–291.

(89) Dubowchik, G. M.; Firestone, R. A.; Padilla, L.; Willner, D.; Hofstead, S. J.; Mosure, K.; Knipe, J. O.; Lasch, S. J.; Trail, P. A. *Bioconjugate Chemistry* **2002**, *13*, PMID: 12121142, 855–869.

(90) Jeffrey, S. C.; Andreyka, J. B.; Bernhardt, S. X.; Kissler, K. M.; Kline, T.; Lenox, J. S.; Moser, R. F.; Nguyen, M. T.; Okeley, N. M.; Stone, I. J.; Zhang, X.; Senter, P. D. *Bioconjugate Chemistry* **2006**, *17*, PMID: 16704224, 831–840.

(91) Jeffrey, S. C.; De Brabander, J.; Miyamoto, J.; Senter, P. D. *ACS Medicinal Chemistry Letters* **2010**, *1*, 277–280.

(92) Kern, J. C.; Dooney, D.; Zhang, R.; Liang, L.; Brandish, P. E.; Cheng, M.; Feng, G.; Beck, A.; Bresson, D.; Firdos, J.; Gately, D.; Knudsen, N.; Manibusan, A.; Sun, Y.; Garbaccio, R. M. *Bioconjugate Chemistry* **2016**, *27*, PMID: 27469406, 2081–2088.

(93) Doronina, S. O.; Mendelsohn, B. A.; Bovee, T. D.; Cerveny, C. G.; Alley, S. C.; Meyer, D. L.; Oflazoglu, E.; Toki, B. E.; Sanderson, R. J.; Zabinski, R. F.; Wahl, A. F.; Senter, P. D. *Bioconjugate Chemistry* **2006**, *17*, PMID: 16417259, 114–124.

(94) Younes, A.; Bartlett, N. L.; Leonard, J. P.; Kennedy, D. A.; Lynch, C. M.; Sievers, E. L.; Forero-Torres, A. *N Engl J Med* **2010**, *363*, 1812–1821.

(95) Peters, C.; Brown, S. *Biosci Rep* **2015**, *35*.

(96) Bouchard, H.; Viskov, C.; Garcia-Echeverria, C. *Bioorg Med Chem Lett* **2014**, *24*, 5357–5363.

(97) Beck, A.; Goetsch, L.; Dumontet, C.; Corvaïa, N. *Nature Reviews Drug Discovery* **2017**, *16*, 315–337.

(98) Akaiwa, M.; Dugal-Tessier, J.; Mendelsohn, B. A. *Chem Pharm Bull (Tokyo)* **2020**, *68*, 201–211.

(99) Boger, D. L.; Johnson, D. S. *Proc Natl Acad Sci U S A* **1995**, *92*, 3642–3649.

(100) Gébleux, R.; Casi, G. *Pharmacol Ther* **2016**, *167*, 48–59.

(101) Yang, F.; Teves, S. S.; Kemp, C. J.; Henikoff, S. *Biochim Biophys Acta* **2013**, *1845*, 84–89.

(102) Leimgruber, W.; Stefanović, V.; Schenker, F.; Karr, A.; Berger, J. *J Am Chem Soc* **1965**, *87*, 5791–5793.

(103) Rahman, K. M.; Thompson, A. S.; James, C. H.; Narayanaswamy, M.; Thurston, D. E. *Journal of the American Chemical Society* **2009**, *131*, PMID: 19725510, 13756–13766.

(104) Rahman, K. M.; Corcoran, D. B.; Bui, T. T. T.; Jackson, P. J. M.; Thurston, D. E. *PLoS One* **2014**, *9*, e105021.

(105) Thomas, A.; Pommier, Y. *Clin Cancer Res* **2019**, *25*, 6581–6589.

(106) Pommier, Y.; Kiselev, E.; Marchand, C. *Bioorg Med Chem Lett* **2015**, *25*, 3961–3965.

(107) Yu, S.; Lim, A.; Tremblay, M. S. In *Innovations for Next-Generation Antibody-Drug Conjugates*, Damelin, M., Ed.; Springer International Publishing: Cham, 2018, pp 321–347.

(108) Pal, L. B.; Bule, P.; Khan, W.; Chella, N. *Pharmaceutics* **2023**, *15*.

(109) Buttgereit, F.; Aelion, J.; Rojkovich, B.; Zubrzycka-Sienkiewicz, A.; Chen, S.; Yang, Y.; Arikan, D.; D'Cunha, R.; Pang, Y.; Kupper, H.; Radstake, T.; Amital, H. *Arthritis Rheumatol* **2023**, *75*, 879–889.

(110) Lee, H.; Bhang, S. H.; Lee, J. H.; Kim, H.; Hahn, S. K. *Bioconjugate Chemistry* **2017**, *28*, PMID: 28107624, 1084–1092.

(111) Lehar, S. M. et al. *Nature* **2015**, *527*, 323–328.

(112) Schade, A. E.; Schieven, G. L.; Townsend, R.; Jankowska, A. M.; Susulic, V.; Zhang, R.; Szpurka, H.; Maciejewski, J. P. *Blood* **2007**, *111*, 1366–1377.

(113) Lee, K. C.; Ouwehand, I.; Giannini, A. L.; Thomas, N. S.; Dibb, N. J.; Bijlmakers, M. J. *Leukemia* **2010**, *24*, 896–900.

(114) Wang, R. E.; Liu, T.; Wang, Y.; Cao, Y.; Du, J.; Luo, X.; Deshmukh, V.; Kim, C. H.; Lawson, B. R.; Tremblay, M. S.; Young, T. S.; Kazane, S. A.; Wang, F.; Schultz, P. G. *Journal of the American Chemical Society* **2015**, *137*, PMID: 25699419, 3229–3232.

(115) Tang, S.-C.; Wynn, C.; Le, T.; McCandless, M.; Zhang, Y.; Patel, R.; Maihle, N.; Hillegass, W. *Cancer Metastasis Rev* **2024**, *44*, 18.

(116) Colombo, R.; Tarantino, P.; Rich, J. R.; LoRusso, P. M.; de Vries, E. G. E. *Cancer Discov* **2024**, *14*, 2089–2108.

(117) Thermo Fisher Scientific, *Bioconjugation and Crosslinking Technical Handbook*; Thermo Fisher Scientific: Waltham, MA, 2021.

(118) Matsuda, Y.; Mendelsohn, B. A. *Expert Opin Biol Ther* **2020**, *21*, 963–975.

(119) Chari, R. V. J. *Accounts of Chemical Research* **2008**, *41*, PMID: 17705444, 98–107.

(120) Lazar, A. C.; Wang, L.; Blättler, W. A.; Amphlett, G.; Lambert, J. M.; Zhang, W. *Rapid Commun Mass Spectrom* **2005**, *19*, 1806–1814.

(121) Hamann, P. R.; Hinman, L. M.; Beyer, C. F.; Lindh, D.; Upeslacis, J.; Flowers, D. A.; Bernstein, I. *Bioconjugate Chemistry* **2002**, *13*, PMID: 11792177, 40–46.

(122) Doronina, S. O.; Toki, B. E.; Torgov, M. Y.; Mendelsohn, B. A.; Cerveny, C. G.; Chace, D. F.; DeBlanc, R. L.; Gearing, R. P.; Bovee, T. D.; Siegall, C. B.; Francisco, J. A.; Wahl, A. F.; Meyer, D. L.; Senter, P. D. *Nature Biotechnology* **2003**, *21*, 778–784.

(123) Matsuda, Y.; Mendelsohn, B. A. *Chem Pharm Bull (Tokyo)* **2021**, *69*, 976–983.

(124) Fontaine, S. D.; Reid, R.; Robinson, L.; Ashley, G. W.; Santi, D. V. *Bioconjugate Chemistry* **2015**, *26*, PMID: 25494821, 145–152.

(125) Knight, P. *Biochemical Journal* **1979**, *179*, 191–197.

(126) Shen, B.-Q. et al. *Nature Biotechnology* **2012**, *30*, 184–189.

(127) Alley, S. C.; Benjamin, D. R.; Jeffrey, S. C.; Okeley, N. M.; Meyer, D. L.; Sanderson, R. J.; Senter, P. D. *Bioconjugate Chemistry* **2008**, *19*, PMID: 18314937, 759–765.

(128) Strop, P. et al. *Chemistry & Biology* **2013**, *20*, 161–167.

(129) Lyon, R. P.; Setter, J. R.; Bovee, T. D.; Doronina, S. O.; Hunter, J. H.; Anderson, M. E.; Balasubramanian, C. L.; Duniho, S. M.; Leiske, C. I.; Li, F.; Senter, P. D. *Nature Biotechnology* **2014**, *32*, 1059–1062.

(130) Behrens, C. R. et al. *Molecular Pharmaceutics* **2015**, *12*, PMID: 26393951, 3986–3998.

(131) Robinson, E.; Nunes, J. P. M.; Vassileva, V.; Maruani, A.; Nogueira, J. C. F.; Smith, M. E. B.; Pedley, R. B.; Caddick, S.; Baker, J. R.; Chudasama, V. *RSC Adv.* **2017**, *7*, 9073–9077.

(132) Badescu, G. et al. *Bioconjugate Chemistry* **2014**, *25*, PMID: 24791606, 1124–1136.

(133) Gubens, M.; Sen, S.; Salkeni, M.; Vandross, A.; Gandhi, N.; Edenfield, W.; Aggarwal, R.; Parsons, K.; Chen, I.; Powderly, J. *Journal of Thoracic Oncology* **2024**, *19*, Abstracts from the 2024 World Conference on Lung Cancer, S243–S244.

(134) Walsh, S. J.; Bargh, J. D.; Dannheim, F. M.; Hanby, A. R.; Seki, H.; Counsell, A. J.; Ou, X.; Fowler, E.; Ashman, N.; Takada, Y.; Isidro-Llobet, A.; Parker, J. S.; Carroll, J. S.; Spring, D. R. *Chem. Soc. Rev.* **2021**, *50*, 1305–1353.

(135) Junutula, J. R. et al. *Nature Biotechnology* **2008**, *26*, 925–932.

(136) Adhikari, P.; Zacharias, N.; Ohri, R.; Sadowsky, J. In *Antibody-Drug Conjugates: Methods and Protocols*, Tumey, L. N., Ed.; Springer US: New York, NY, 2020, pp 51–69.

(137) Kolb, H. C.; Finn, M. G.; Sharpless, K. B. *Angewandte Chemie International Edition* **2001**, *40*, 2004–2021.

(138) Agarwal, P.; Bertozzi, C. R. *Bioconjugate Chemistry* **2015**, *26*, PMID: 25494884, 176–192.

(139) Van Geel, R.; Wijdeven, M. A.; Heesbeen, R.; Verkade, J. M. M.; Wasiel, A. A.; van Berkel, S. S.; van Delft, F. L. *Bioconjugate Chemistry* **2015**, *26*, PMID: 26061183, 2233–2242.

(140) De Bever, L.; Popal, S.; van Schaik, J.; Rubahamya, B.; van Delft, F. L.; Thurber, G. M.; van Berkel, S. S. *Bioconjugate Chemistry* **2023**, *34*, PMID: 36857521, 538–548.

(141) Rabuka, D.; Rush, J. S.; deHart, G. W.; Wu, P.; Bertozzi, C. R. *Nature Protocols* **2012**, *7*, 1052–1067.

(142) Appel, M. J.; Bertozzi, C. R. *ACS Chemical Biology* **2015**, *10*, PMID: 25514000, 72–84.

(143) Yamada, K.; Shikida, N.; Shimbo, K.; Ito, Y.; Khedri, Z.; Matsuda, Y.; Mendelsohn, B. A. *Angew. Chem. Int. Ed.* **2019**, *58*, 5592–5597.

(144) Fujii, T. et al. *Bioconjugate Chemistry* **2023**, *34*, PMID: 36894324, 728–738.

(145) Matsuda, Y.; Shikida, N.; Hatada, N.; Yamada, K.; Seki, T.; Nakahara, Y.; Endo, Y.; Shimbo, K.; Takahashi, K.; Nakayama, A.; Mendelsohn, B. A.; Fujii, T.; Okuzumi, T.; Hirasawa, S. *Organic Letters* **2024**, *26*, PMID: 38639400, 5597–5601.

(146) Chen, Y. In *Antibody-Drug Conjugates*, Ducry, L., Ed.; Humana Press: Totowa, NJ, 2013, pp 267–273.

(147) Hamblett, K. J.; Senter, P. D.; Chace, D. F.; Sun, M. M. C.; Lenox, J.; Cerveny, C. G.; Kissler, K. M.; Bernhardt, S. X.; Kopcha, A. K.; Zabinski, R. F.; Meyer, D. L.; Francisco, J. A. *Clin Cancer Res* **2004**, *10*, 7063–7070.

(148) Quiles, S.; Raisch, K. P.; Sanford, L. L.; Bonner, J. A.; Safavy, A. *Journal of Medicinal Chemistry* **2010**, *53*, PMID: 19958000, 586–594.

(149) Safavy, A.; Bonner, J. A.; Waksal, H. W.; Buchsbaum, D. J.; Gillespie, G. Y.; Arani, R.; Chen, D.-T.; Carpenter, M.; Raisch, K. P. *Bioconjugate Chemistry* **2003**, *14*, PMID: 12643740, 302–310.

(150) Valliere-Douglass, J. F.; McFee, W. A.; Salas-Solano, O. *Analytical Chemistry* **2012**, *84*, PMID: 22384990, 2843–2849.

(151) Signor, L.; Boeri Erba, E. *JoVE* **2013**, e50635.

(152) Tscheuschner, G.; Schwaar, T.; Weller, M. G. *Antibodies (Basel)* **2020**, *9*.

(153) Cini, E.; Faltoni, V.; Petricci, E.; Taddei, M.; Salvini, L.; Giannini, G.; Vesci, L.; Milazzo, F. M.; Anastasi, A. M.; Battistuzzi, G.; De Santis, R. *Chem. Sci.* **2018**, *9*, 6490–6496.

(154) Cianferotti, C.; Faltoni, V.; Cini, E.; Ermini, E.; Migliorini, F.; Petricci, E.; Taddei, M.; Salvini, L.; Battistuzzi, G.; Milazzo, F. M.; Anastasi, A. M.; Chiapparino, C.; De Santis, R.; Giannini, G. *Chem. Commun.* **2021**, *57*, 867–870.

(155) Huang, R. Y.-C.; Chen, G. *Drug Discovery Today* **2016**, *21*, 850–855.

(156) Fiala, J.; Schuster, D.; Heck, A. J. R. *Journal of the American Society for Mass Spectrometry* **2025**, *36*, PMID: 40408263, 1395–1403.

(157) Kempen, T.; Cadang, L.; Fan, Y.; Zhang, K.; Chen, T.; Wei, B. *MAbs* **2024**, *17*, 2446304.

(158) Rodriguez-Aller, M.; Guillarme, D.; Beck, A.; Fekete, S. *Journal of Pharmaceutical and Biomedical Analysis* **2016**, *118*, 393–403.

(159) Beckley, N. S.; Lazzareschi, K. P.; Chih, H.-W.; Sharma, V. K.; Flores, H. L. *Bioconjugate Chemistry* **2013**, *24*, PMID: 24070051, 1674–1683.

(160) VanAernum, Z. L.; Busch, F.; Jones, B. J.; Jia, M.; Chen, Z.; Boyken, S. E.; Sahasrabuddhe, A.; Baker, D.; Wysocki, V. H. *Nature Protocols* **2020**, *15*, 1132–1157.

(161) Le, L. N.; Moore, J. M. R.; Ouyang, J.; Chen, X.; Nguyen, M. D. H.; Galush, W. J. *Analytical Chemistry* **2012**, *84*, PMID: 22913809, 7479–7486.

(162) Chen, T.-H.; Yang, Y.; Zhang, Z.; Fu, C.; Zhang, Q.; Williams, J. D.; Wirth, M. J. *Analytical Chemistry* **2019**, *91*, PMID: 30661356, 2805–2812.

(163) Wang, H. et al. *Nature* **2023**, *620*, 47–60.

(164) Dou, B.; Zhu, Z.; Merkurjev, E.; Ke, L.; Chen, L.; Jiang, J.; Zhu, Y.; Liu, J.; Zhang, B.; Wei, G.-W. *Chemical Reviews* **2023**, *123*, PMID: 37384816, 8736–8780.

(165) Pietrobono, S.; Santini, R.; Gagliardi, S.; Dapporto, F.; Colecchia, D.; Chiariello, M.; Leone, C.; Valoti, M.; Manetti, F.; Petricci, E.; Taddei, M.; Stecca, B. *Cell Death & Disease* **2018**, *9*, 142.

(166) Petroni, M. et al. *Cell Death & Disease* **2018**, *9*, 895.

(167) Vesci, L.; Milazzo, F. M.; Stasi, M. A.; Pace, S.; Manera, F.; Tallarico, C.; Cini, E.; Petricci, E.; Manetti, F.; De Santis, R.; Giannini, G. *European Journal of Medicinal Chemistry* **2018**, *157*, 368–379.

(168) Maresca, L. et al. *Pharmacological Research* **2023**, *195*, 106858.

(169) Manetti, F.; Maresca, L.; Crivaro, E.; Pepe, S.; Cini, E.; Singh, S.; Governa, P.; Maramai, S.; Giannini, G.; Stecca, B.; Petricci, E. *ACS Medicinal Chemistry Letters* **2022**, *13*, 1329–1336.

(170) Weininger, D. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.

(171) Landrum, G. **2016**.

(172) Bolikulov, F.; Nasimov, R.; Rashidov, A.; Akhmedov, F.; Cho, Y.-I. *Mathematics* **2024**, *12*, DOI: 10.3390/math12162553.

(173) Zhu, W.; Qiu, R.; Fu, Y. Comparative Study on the Performance of Categorical Variable Encoders in Classification and Regression Tasks, 2024.

(174) Pedregosa, F. et al. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(175) Chen, T.; Guestrin, C. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, ACM: New York, NY, USA, 2016, pp 785–794.

(176) Friedman, J. H. *Ann. Stat.* **2001**, *29*, 1189–1232.

(177) Altman, N. S. *Am. Stat.* **1992**, *46*, 175–185.

(178) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J., *Classification and Regression Trees*; Wadsworth International Group: Belmont, CA, 1984.

(179) Ali, M.; Aittokallio, T. *Biophys Rev* **2018**, *11*, 31–39.

(180) Pinheiro, J. M. H.; de Oliveira, S. V. B.; Silva, T. H. S.; Saraiva, P. A. R.; de Souza, E. F.; Godoy, R. V.; Ambrosio, L. A.; Becker, M. The Impact of Feature Scaling In Machine Learning: Effects on Regression and Classification Tasks, 2025.

(181) De Amorim, L. B.; Cavalcanti, G. D.; Cruz, R. M. *Applied Soft Computing* **2023**, *133*, 109924.

(182) Angiolini, L.; Manetti, F.; Spiga, O.; Tafi, A.; Visibelli, A.; Petricci, E. *Journal of Chemical Information and Modeling* **2025**, *65*, PMID: 40214655, 5847–5855.

(183) Lundberg, S. M.; Lee, S.-I. In *Advances in Neural Information Processing Systems*, ed. by Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., Curran Associates, Inc.: 2017; Vol. 30.

(184) Rodríguez-Pérez, R.; Bajorath, J. *Journal of Computer-Aided Molecular Design* **2020**, *34*, 1013–1026.

(185) Veríssimo, R. F.; Matias, P. H. F.; Barbosa, M. R.; Neto, F. O. S.; Neto, B. A. D.; de Oliveira, H. C. B. *Journal of Chemical Information and Modeling* **2025**, *65*, PMID: 40300554, 7874–7886.

(186) Leo, A.; Hansch, C.; Elkins, D. *Chemical Reviews* **1971**, *71*, 525–616.

(187) Buecheler, J. W.; Winzer, M.; Weber, C.; Gieseler, H. *J Pharm Sci* **2019**, *109*, 161–168.

(188) Lundahl, M. L. E.; Fogli, S.; Colavita, P. E.; Scanlan, E. M. *RSC Chem. Biol.* **2021**, *2*, 1004–1020.

(189) Lucas, A. T.; Price, L. S. L.; Schorzman, A. N.; Storrie, M.; Piscitelli, J. A.; Razo, J.; Zamboni, W. C. *Antibodies (Basel)* **2018**, *7*.

(190) Siew, A. *Pharmaceutical Technology APIs, Excipients, & Manufacturing* **2018**.

(191) Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219–3228.

(192) Yang, X.; Pan, Z.; Choudhury, M. R.; Yuan, Z.; Anifowose, A.; Yu, B.; Wang, W.; Wang, B. *Med. Res. Rev.* **2020**, *40*, 2682–2713.

(193) Liu, H.; May, K. *mAbs* **2012**, *4*, PMID: 22327427, 17–23.

(194) Dean, A. Q.; Luo, S.; Twomey, J. D.; Zhang, B. *MAbs* **2021**, *13*, 1951427.

(195) Bertz, S. H. *Journal of the American Chemical Society* **1981**, *103*, 3599–3601.

(196) Krzyzanowski, A.; Pahl, A.; Grigalunas, M.; Waldmann, H. *Journal of Medicinal Chemistry* **2023**, *66*, PMID: 37651653, 12739–12750.

(197) Yang, L.; Zhao, Y.; Fu, X.; Zhang, W.; Xu, W. *Journal of the American Society for Mass Spectrometry* **2025**, *36*, PMID: 40168520, 991–998.

(198) Nayak, S.; Richter, S. M. *Organic Process Research & Development* **2023**, *27*, 2091–2099.

(199) Milazzo, F. M.; Vesci, L.; Anastasi, A. M.; Chiapparino, C.; Rosi, A.; Giannini, G.; Taddei, M.; Cini, E.; Faltoni, V.; Petricci, E.; Battistuzzi, G.; Salvini, L.; Carollo, V.; De Santis, R. *Frontiers in Oncology* **2020**, *9*, DOI: 10.3389/fonc.2019.01534.

(200) Anderson, G. W.; Zimmerman, J. E.; Callahan, F. M. *Journal of the American Chemical Society* **1964**, *86*, 1839–1842.

(201) Singh, J. *Journal of Medicinal Chemistry* **2022**, *65*, PMID: 35439421, 5886–5901.

(202) Park, B. K. et al. *Nature Reviews Drug Discovery* **2011**, *10*, 292–306.

(203) Singh, J.; Petter, R. C.; Baillie, T. A.; Whitty, A. *Nature Reviews Drug Discovery* **2011**, *10*, 307–317.

(204) Potashman, M. H.; Duggan, M. E. *Journal of Medicinal Chemistry* **2009**, *52*, PMID: 19203292, 1231–1246.

(205) Lonsdale, R.; Ward, R. A. *Chem. Soc. Rev.* **2018**, *47*, 3816–3830.

(206) Lu, X.; Smaill, J. B.; Patterson, A. V.; Ding, K. *Journal of Medicinal Chemistry* **2022**, *65*, PMID: 34962782, 58–83.

(207) Li, H.-y.; Qi, W.-l.; Wang, Y.-x.; Meng, L.-h. *Genes & Diseases* **2023**, *10*, 403–414.

(208) Fry, D. W. et al. *Proceedings of the National Academy of Sciences* **1998**, *95*, 12022–12027.

(209) Jackson, P. A.; Widen, J. C.; Harki, D. A.; Brummond, K. M. *Journal of Medicinal Chemistry* **2017**, *60*, PMID: 27996267, 839–885.

(210) Bianco, G.; Goodsell, D. S.; Forli, S. *Trends Pharmacol Sci* **2020**, *41*, 1038–1049.

(211) Bianco, G.; Forli, S.; Goodsell, D. S.; Olson, A. J. *Protein Sci* **2015**, *25*, 295–301.

(212) Schröder, J.; Klinger, A.; Oellien, F.; Marhöfer, R. J.; Duszenko, M.; Selzer, P. M. *Journal of Medicinal Chemistry* **2013**, *56*, PMID: 23350811, 1478–1490.

(213) Ouyang, X.; Zhou, S.; Su, C. T. T.; Ge, Z.; Li, R.; Kwoh, C. K. *J Comput Chem* **2012**, *34*, 326–336.

(214) Ai, Y.; Yu, L.; Tan, X.; Chai, X.; Liu, S. *Journal of Chemical Information and Modeling* **2016**, *56*, PMID: 27411028, 1563–1575.

(215) London, N.; Miller, R. M.; Krishnan, S.; Uchida, K.; Irwin, J. J.; Eidam, O.; Gibold, L.; Cimermančič, P.; Bonnet, R.; Shoichet, B. K.; Taunton, J. *Nat Chem Biol* **2014**, *10*, 1066–1072.

(216) Bianco, G.; Holcomb, M.; Santos-Martins, D.; Tillack, A.; Hansel-Harris, A.; Forli, S. *Journal of Chemical Information and Modeling* **2023**, *63*, PMID: 37639635, 5631–5640.

(217) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. *J Comput Chem* **2009**, *30*, 2785–2791.

(218) Hur, S.; Bruice, T. C. *Proc Natl Acad Sci U S A* **2003**, *100*, 12015–12020.

(219) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639–1662.

(220) Backus, K. M.; Correia, B. E.; Lum, K. M.; Forli, S.; Horning, B. D.; González-Páez, G. E.; Chatterjee, S.; Lanning, B. R.; Teijaro, J. R.; Olson, A. J.; Wolan, D. W.; Cravatt, B. F. *Nature* **2016**, *534*, 570–574.

(221) Santos-Martins, D.; Solis-Vasquez, L.; Tillack, A. F.; Sanner, M. F.; Koch, A.; Forli, S. *J Chem Theory Comput* **2021**, *17*, 1060–1073.

(222) Forli, S.; Huey, R.; Pique, M. E.; Sanner, M. F.; Goodsell, D. S.; Olson, A. J. *Nature Protocols* **2016**, *11*, 905–919.

(223) Resnick, E. et al. *Journal of the American Chemical Society* **2019**, *141*, PMID: 31060360, 8951–8968.

(224) Reilly, S.-J.; Tillander, V.; Ofman, R.; Alexson, S. E. H.; Hunt, M. C. *J Biochem* **2008**, *144*, 655–663.

(225) Hubbard, R. E.; Chen, I.; Davis, B. *Curr Opin Drug Discov Devel* **2007**, *10*, 289–297.

Acknowledgements

Da dove cominciare... gli ultimi quattro anni sono stati ricchi di emozioni, avventure e persone. Mi porto dietro un'infinità di esperienze e ricordi che mi hanno reso oggi una persona molto diversa da quella che ha cominciato questo percorso. Prima di tutto vorrei ringraziare la Prof.ssa Petricci, a lei va un grandissimo abbraccio ed un sincero GRAZIE per avermi dato modo di intraprendere questo percorso e di poter arrivare a questo risultato. Il viaggio non è stato semplice, molti intoppi e difficoltà, ma alla fine ci siamo tolti anche qualche soddisfazione!

Ringrazio anche tutti i professori che hanno preso parte a questo viaggio, la Prof.ssa Cini e il Prof. Taddei, sempre pronti a dare consigli ma anche ad aprire spunti di riflessione sul nostro lavoro. Vorrei poi ringraziare tutte le persone che hanno partecipato a questo viaggio, chi per un internato di tesi o una borsa di studio, altri dottorati o borse post doc e ricercatori (già chiedo scusa a chi andrò a dimenticare perché sicuramente qualcuno me lo perdo...): Elena, Sofia, Giorgia, Filippo, Giulia, Simone, Giusy, Samuele, Davide, Maria, Irene, Leonardo, Federica, Antonino e Davide (sempre a coppia vanno scritti), Elisa, Alessia, Viviana, Chiara, Evelyn, Roberto, Marta... oddio non mi ricordo più chi altro c'era! Scusatemi e grazie mille perché ognuno di voi ha contribuito, anche se in piccola parte, nel viaggio che mi ha portato fino a qui!

Demetra e Giovanni, pensavate mi fossi dimenticato di voi e invece no! Due righe ve le meritate (se le meriterebbero tutti ma poi c'è da pagare la copisteria...). Abbiamo iniziato questo dottorato insieme con tanta emozione, passo dopo passo abbiamo superato tutti

gli scogli che abbiamo trovato, dai primi poster alle presentazioni ai congressi, bright night varie, fino ad allontanarci, solo per un pò, per andare a fare ricerca altrove. Oggi siamo giunti alla fine di questo percorso che però sarà per tutti l'inizio di uno nuovo, ci ritroveremo a bere uno spritz da qualche parte insieme per tenerci aggiornati!

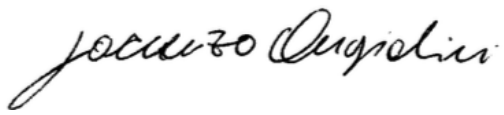
Vorrei poi ringraziare tutte le persone con cui ho avuto l'occasione di collaborare: il Prof. Tafi, il Prof. Manetti, la Prof.ssa Spiga e Anna. Te Anna, mi hai letteralmente aperto un mondo la prima volta che mi hai fatto scrivere due righe con Python e da questo mondo penso proprio che non riuscirò mai ad uscire, quindi grazie di cuore.

I would like to thank all the people I've met during my stay in San Diego. Starting with Stefano, you managed to help me understand and do things I never even imagined. Under your guidance, I immediately felt comfortable, maybe that's your superpower... I don't know, but I'm sure many people would confirm it. Thank you for giving me the opportunity to have this experience in the legendary States! I will carry a wonderful memory of it with me. Alessandra and Pietro came to pick Vale and me up at the airport, after we had quite a stressful moment with the customs officer while entering the country. We shared some fantastic moments together. I don't know what I wouldn't give to go back and watch a Padres game with you again (though I think I might have been the only one who really liked it...). A special thanks also goes to all the people at the Forli Lab, highly skilled individuals who managed to teach me a profession that was completely new to me: Peter, Allison, Manu, Matthew, Renhao, Diogo, Yuting, Niccolò, Ishan, Joani, Quentin, Kam, Althea, Chris, and Boyuan.

Jenn, Kai, and Stella, you became our new family, and we truly miss you so much! You managed to make us call home a place that was completely new and far from what we were used to; we couldn't have asked for anything better. I just hope I managed to cancel all the periodical Amazon purchases; I already know that a package of laundry detergent ended up being delivered to you...

Un enorme grazie va poi ai miei genitori, mi appoggiate sempre in tutto e ci siete per

qualsiasi cosa. Riuscite a rendermi fiero di quello che faccio, anche quando vi racconto il mio lavoro, e voi non ci capite niente, siete sempre soddisfatti e questo mi fa super piacere. Infine, Valentina, amore mio, non smetterò mai di ringraziarti, sei la prima persona che mi supporta e anche che mi sopporta. Sei sempre pronta ad ascoltarmi quando ho un problema e ti impegni da morire per aiutarmi a risolverlo. L'ultimo anno è stato forse il più bello che abbiamo passato insieme, abbiamo visto e fatto cose incredibili che se tornassi indietro le rifarei tutte quante! Anche dall'altra parte del mondo eri con me, pronta all'avventura! Non vedo l'ora di iniziare il prossimo capitolo, ovunque sia.



Lorenzo Angiolini
06.03.2026 18:43:16 GMT+01:00