Thèse n° 10 392

Development of tools and methods for protein identification: From single molecules to in vivo applications.

Présentée le 28 juin 2024

Faculté des sciences de base
Laboratoire de chimie biophysique des macromolécules
Programme doctoral en chimie et génie chimique

pour l'obtention du grade de Docteur ès Sciences

par

Salome Riccarda PÜNTENER

Acceptée sur proposition du jury

Prof. Ph. Schwaller, président du jury
Prof. B. Fierz, Prof. P. Rivera Fuentes, directeurs de thèse
Prof. C. Chan, rapporteuse
Prof. S. Schmid, rapporteuse
Prof. B. Correia, rapporteur

École
polytechnique
fédérale
de Lausanne

2024

Acknowledgments

I would like to take this opportunity to express my gratitude to the people and institutions that have supported me throughout my doctoral journey. Their contributions, guidance, and encouragement have been instrumental in shaping my personal and academic growth during this period.

First, I am grateful to my PhD supervisor, Professor Pablo Rivera-Fuentes, for allowing me to do my doctorate in his group. Thank you for your guidance, feedback, and motivational support when the way was not clear to me. Your enthusiasm and excitement for the science we do in the lab are truly inspiring and the experiences I was able to gather during that time are invaluable.

I extend my thanks to Professor Beat Fierz for taking over the role as my official supervisor and to the members of the jury Professors Chan Cao, Sonja Schmid, Bruno Correia, and Philippe Schwaller for their time and effort in co-examining my thesis.

Many thanks to the staff of the facilities that keep all the equipment running and perform crucial analyses we heavily rely on. Thanks to the BIOP, the MS and NMR service, the workshop, and the shop teams at EPFL. Thanks to Emanuela for helping to get us started and keep things running in Lausanne. Thanks to the MS and NMR service, the Cytometry facility, S3IT, and the staff of FGCZ at UZH. Many thanks to Dr. Roman Briskine for your work and support.

I am grateful for all the students I had the privilege to supervise during this time. Lionel, Peter, Gianluca, and Julian, thank you for trusting me and dedicating time to the projects, the mentoring experiences helped me grow as a teacher and as a person.

I would like to thank the current and past LOCBP-gang for the great work environment and all the fun times inside and outside of the laboratory and I hope I will be able to enjoy your company a little while longer.

Thank you, Elias, for being the best mentor one can imagine. Your infinite positivity is an inspiration and a source of energy for the community around you and has motivated me time and time again. I enjoyed the time in the lab (day and night), the coffee breaks, floating on the Limmat and the late nights out. I learned a lot from you scientifically and on a personal level. I hope we keep in touch.

Thank you, Alina, for setting up the laboratory in the first place and instilling the spirit of a clean and organized laboratory that still can be found today (at least some days). I enjoyed our time spent in- and outside of the laboratory with some (caffeinated) drinks and your refreshing directness.

Thank you, Jade, for being a great colleague and even more so for being a great friend. You shaped our laboratory incredibly and always brought the people together. Thanks for keeping the spirits up even when times were hard. I know that I can always rely on you. I have the fondest memories of our time working together in the lab, un-, packing, jogging, baking, dining, going out, and more than I can list here. I am looking forward to future adventures together!

Thank you, Zach, for showing me around the laboratory in the very beginning and for your knowledgeable support. Thanks for bringing climbing, bouldering, and crazy plant/fungi experiments to the laboratory.

Thank you, Adam, for always making the time spent in the chemistry laboratory fun, no matter the late hours. Thanks for the weird and intense discussions, the infinite guessing games, business ideas, and the interesting music choices. I will never drink a cherry Coke again without thinking of you.

Krzysztof, thanks for being a great colleague and working with me on this project, embracing my morning moods and turning them around. Thank you for all the kitchen-parties, movie nights, karaoke, and food blogging moments, I will remember them fondly.

Thank you, Gina, for bringing a relaxed and always optimistic spirit to the lab. Thank you for starting the yeast work and for your patience and support whenever I was blocked with it.

Thank you, Carla, for being the "Lab Mom" and a dear friend. I deeply appreciate your social awareness, your care for all members of the lab, your attention to detail, and that you pick up the slack when everyone else was too lazy. I am grateful for your encouragement, monitoring of work hours, and the activities outside of the lab. I learned a lot from you.

Annabell, thank you for being a kind and supportive colleague. Thanks for being part of the yeast team "SPAM" and working together on implementation and troubleshooting. Thanks for the morning coffees we shared and your delicious cake creations.

Sarah, thanks for being a great colleague with an endless supply of smiles and a positive attitude. I enjoyed the intense bouldering sessions and supportive discussions on planning the future.

Thank you, Set, for your smiles, your weekend travel stories, and the great Thai diners!

Thank you, Hen, for joining the biking-, bouldering-, and crazy-plant-project crew, and other projects even if they included Lac Leman, early mornings, and freezing temperatures. I hope you find many more projects and people to share them with.

Thank you, Doro, for joining the chemistry lab safety team. There could have been no one better, with your attention to detail and spotting opportunities for improvement. Thank you for organizing lab parties and events, the carnival craziness, the intricate drinks, and the interesting chemistry discussions.

Thank you, Juan, for bringing a positive and relaxed vibe to the lab, for your awesome dance moves, and for being a loyal “Friends” follower. Thanks for your Mexican treats and the lunch and late-evening discussions, scientific and otherwise, and I hope there will be more such conversations in the future.

An additional thanks to Carla, Doro, Hen, Jacqueline, and Andreas for proofreading this thesis from the half-finished-sentence- to the almost-final-version.

My sincerest acknowledgements go to my family and friends with a special thanks to Jacqueline. I can always count on you personally and for an outside perspective. Thanks to my family for their constant support and their encouragement to do whatever would make me happy.

And last but not least thanks to Andreas for being my stronghold and for your endless support in any way possible, you are the best!

Disclosure: For writing this thesis a large language model (ChatGPT3.5) was used for text optimization. However, no content or conceptual aspects were generated. The german translation of the abstract is based on the translation with DeepL.com.

Zusammenfassung

Als grundlegende funktionelle Einheit der Zelle, beeinflussen Proteine deren Zustand fundamental. Zellen unterscheiden sich untereinander stark unterscheiden, was die Analyse des Proteoms auf der zellulären Ebene unerlässlich macht, um ihr Verhalten und ihre Entwicklung zu verstehen.

Während die Analyse des Genoms und des Transkriptoms einzelner Zelle in kommerziellem Maßstab möglich ist, ist dies für die Proteomik nicht der Fall. Die Proteomanalyse ist aufgrund der enormen chemischen Komplexität, der großen Anzahl von Proteinen und Spielraum der Proteinkonzentrationen herausfordernd. Außerdem können Proteine, im Gegensatz zur DNA, nicht direkt amplifiziert werden, so dass hochempfindliche Methoden notwendig sind. In den letzten Jahren wurden innovative Einzelmolekülansätze vorgeschlagen, die jedoch immer noch mit vielen technischen Schwierigkeiten konfrontiert sind, insbesondere hinsichtlich der Komplexität der gemessenen Proben und des Durchsatzes.

In dieser Arbeit werden das Konzept und die erste Umsetzung einer neuen Herangehensweise, «Blinkognition», für die gezielte Charakterisierung einzelner Moleküle, insbesondere von Peptiden und Proteinen vorgestellt mittels vorgestellt spontan blinkende Fluorophore. Die neue Methode nutzt die Sensitivität des Fluoreszenzblinken für die chemische Umgebung, um Moleküle in der Nähe des Fluorophors zu erkennen. Es wurde ein datenbasierter Ansatz des maschinellen Lernens verwendet. Ein Modell wurde mittels Fluoreszenzintensitätssignale einzelner Moleküle von synthetisch reinen Proben trainiert. Das resultierende Modell konnte verschiedene Peptid-Farbstoff-Konjugate von Peptiden mit unterschiedlichen Sequenzen, Phosphorylierungs- und Epimerisierungsmustern erkennen. Des Weiteren nutzten wir die Unsicherheit der Modellvoraussage, um Signale geringer Qualität herauszufiltern. Unsere jüngsten Arbeiten haben gezeigt, dass die Methode auf lipidumschlossene, Fluorophor-Protein-Konjugate und sogar auf Konjugate desselben Proteins erweitert werden kann. In Kombination mit einer angepassten Markierungsstrategie hat die Methode Potenzial für die Erstellung von Proteinprofilen und bietet viele Möglichkeiten für die weitere Entwicklung auch über Proteine hinaus.

Weiterführen beschreibt diese Arbeit beschreibt diese Arbeit die erste Erforschung eines kleinen Peptid-Tags, der einen fluorogenen Farbstoff binden kann, um Proteine in Zellen zu visualisieren. Eine frühere Arbeit beschrieb ein Boronsäuren modifiziertes Fluorophor, das ein Peptide, welches zwei Serinpaare enthält, mit hoher Affinität bindet. Diese Probe weist jedoch ein hohes Hintergrundsignal in Zellen auf, das möglicherweise auf die natürliche Präsenz der Zielaminosäuresequenz zurückzuführen ist. Wir haben vorgeschlagen die Hintergrundfluoreszenz zu reduzieren durch die Kombination eines fluorogenen Rhodamin-Farbstoff mit Boronsäuren für die Peptidbindung. Wir implementierten ein Hefe-Display von Peptiden mit zwei Serinpaaren, die durch zwei oder acht randomisierte Aminosäuren getrennt sind. Das Screening nach einem Zielpeptid, das die neuen Farbstoffe bindet und fluoreszent macht, blieb erfolglos, was zum Teil an der geringen Löslichkeit des Farbstoffs lag. Unser Ansatz hat aber weiterhin Potential mit verbesserten Fluorophoren sowie erweiterten Bibliotheken. Eine kleine Markierung in Kombination mit einem fluorogenen Molekül könnte auf Zieleproteine angewendet werden, die keine große Modifikation wie fluoreszierende Proteine erlauben.

Schlüsselwörter: Blinkognition, spontan blinkende Fluorophore, Einzelmolekül-Spektroskopie, Protein und Peptid Erkennung, Maschinelles Lernen, Bisboronsäurn, Hefezellen Oberflächenpräsentation, Fluorogenizität

Abstract

As the fundamental machinery orchestrating cellular functions, proteins influence the state of every cell profoundly. As cells exhibit significant variations from one to another, analyzing the proteome on a single-cell level is imperative to unravel their behavior and development. Beyond proteomic composition, the cellular and temporal context adds another layer of complexity to biological processes.

Whereas the analysis of the genome and transcriptome of a single cell can be done readily on a commercial scale, this is far from reality for proteomics. Proteome analysis is challenging due to the vast chemical complexity, diverse number of proteins, and their concentration range. Moreover, unlike DNA, protein analysis lacks a method for amplification and therefore requires highly sensitive methods. In recent years innovative single-molecule approaches have been proposed but they still face many technical difficulties, especially sample complexity and throughput.

This thesis introduces the concept and development of a mechanistically novel system, termed blinkognition. It employs spontaneously blinking fluorophores for the targeted characterization and analysis of single molecules, particularly peptides and proteins. The fluorescence blinking is sensitive to the chemical environment. We proposed that the signal modulation due to the presence of a molecule of interest could be leveraged for peptide identification. A data-driven approach was employed, where a machine learning model was trained on fluorescence intensity time traces from synthetically pure samples measured in single-molecule total internal reflection fluorescence microscopy. The resulting model could discern different peptide-dye conjugates of peptides with different sequences, phosphorylation, and epimerization patterns. Furthermore, we proposed to leverage model uncertainty to filter low-quality signals. Our latest work showed that the method can be extended to encapsulated, labeled proteins and even different sites on the same protein. Combined with an adapted labeling strategy, the method shows potential for protein profiling with many opportunities for further development even beyond proteins.

Beyond the single-molecule level and toward measurements of proteins in cells, this thesis describes the initial exploration for a small peptide tag that can bind a fluorogenic dye. Previous work has reported high-affinity binding of a peptide that contains two serine pairs by a fluorophore modified with two boronic acid moieties. The reported system features high background in cells, potentially due to natively present targets. We proposed combining fluorogenic rhodamine dye with boronic acids for peptide binding to lower this background fluorescence. We implemented yeast display of peptides with two serine pairs separated by two or eight randomized amino acids.

Screening these yeast libraries for a peptide candidate that can bind and turn on the bisboronic probes remained unsuccessful in part hampered by low dye solubility. Nevertheless, with improved dyes and extended libraries this approach holds the potential to provide a small peptide for protein labeling. A small tag in combination with a fluorogenic small molecule could be applied to targets that do not tolerate large protein modifications such as fluorescent and self-labeling proteins while offering the photophysical advantages of a small molecule.

Keywords: blinkognition, spontaneously blinking fluorophores, single-molecule total internal reflection fluorescence imaging, peptide and protein identification, machine learning, bisboronic probes, yeast surface display, fluorogenicity

Table of contents

Abbreviations and symbols

°	degree
µ	micro
δ	chemical shift
σ	standard deviation
a.u.	arbitrary units
15:0 PC	1,2-dipentadecanoyl-sn-glycero-3-phosphocholine
16:0 biotinyl PE	1,2-dihexadecanoyl-sn-glycero-3-phosphoethanolamine-*N*-(biotinyl) (sodium salt)
16:0 NBD PE	1,2-dipalmitoyl-sn-glycero-3-phosphoethanolamine-*N*-(7-nitro-2-1,3-benzoxadiazol-4-yl) (ammonium salt)
2-PCA	2-pyridinyl carboxaldehyde
a.u.	aribtrary units
AEAPTMS	*N*-(2-aminoethyl)-3-aminopropyl-trimethoxysilane
AFM	atomic force microscopy
AIBN	2,2'-azobis(2-methylpropionitrile)
aPEG	azide-polyethylene glycol
BBA	bisboronic acid
BG	*O*^{6}-benzylguanine
BPBS	0.5% bovine serum albumin in PBS
BPC	base peak chromatogram
BuLi	butyllithium
Bzl	benzyl group
C	Celsius
calcd.	calculated
CNN	convolutional neural network
CT	certainty threshold
CuAAC	copper(I)-catalyzed azide-alkyne cycloaddition
CV	cross-validation
CWP	cell wall protein
Cy3	Cyanine 3
d	doublet
dba	dibenzylideneacetone
DC	display control
DHFR	dihydrofolate reductase
DIPEA	*N*,*N*-diisopropylethylamine
DL	deep learning
DLS	dynamic light scattering
DMF	*N*,*N*'-dimethylformamide
DMPC	1,2-dimyristoyl-sn-glycero-3-phosphocholine
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
dppf	1,1’-ferrocenediyl-bis(diphenylphosphine)
DPPC	1,2-dihexadecanoyl-sn-glycero-3-phosphocholine
DPPE	1,2-dipalmitoyl-sn-glycero-3-phosphoethanolamine
DST	disuccinimidyl tartrate
EBX2	ethynylbenziodoxolone
EDC HCl	*N*-ethyl-*N′*-(3-dimethylaminopropyl) carbodiimide hydrochloride

EDTA	ethylenediaminetetraacetic acid
EIC	extracted ion chromatogram
EMCCD	electron multiplying charged coupled device
EPFL	Swiss Federal Institute of Technology in Lausanne
equiv.	equivalents
ESI	electrospray ionization
EtOAc	ethyl acetate
FACS	fluorescence-activated cell sorting
FAST	fluorescence-activating and absorption shifting tag
FGCZ	Functional Genomic Center Zürich
FlAsH	fluorescein arsenic helix-binder
Fmoc	9-fluorenylmethoxycarbonyl
FP	fluorescent protein
FRET	Förster resonance energy transfer
FRET X	FRET via DNA eXchange
FSC	forward scattering
g	gram(s)
GPU	graphical processing unit
GRU	gated recurrent unit
Grx	glutaredoxin
h	hour(s)
HA	human influenza hemagglutinin
hAGT	human O^6- alkylguanine-DNA-alkyltransferase
HATU	*O*-(7-azabenzotriazol-1-yl)-*N*,*N*,*N*",*N*"-tetramethyluronium hexafluorophosphate
HBTU	*O*-(1*H*-Benzotriazole-1-yl)-*N*,*N*,*N′*,*N′*-tetramethyluroniumhexafluorophosphate
HCTU	*O*-(1*H*-6-chlorobenzotriazole-1-yl)-1,1,3,3-tetramethyluronium
HeLa cells	Henrietta Lacks human cervical cancer cells
HMSiR	hydroxymethyl silicon rhodamine
HOBt	1-hydroxybenzotriazole
HPLC	high-performance liquid chromatography
HRMS	high-resolution MS
HTHTL	HaloTag7 labeled with HMSiR-Halo 34
HTIA	HaloTag7 labeled with HMSiR-IA 22
HTS	high-throughput screen(ing)
Hz	Hertz
IA	iodoacetamide
IUPAC	International Union of Pure and Applied Chemistry
K_D	dissociation constant
KNN	k-Nearest Neighbors
QTOF	quadrupole time-of-flight
L	liter
LB	lysogeny broth
LC-MS	liquid chromatography-MS
LC-MS/MS	liquid chromatography-tandem MS
LSTM	long-short term memory
m	milli; meter; multiplet
M	molar
m/z	mass over charge ratio

MCD	Monte Carlo dropout
McoTI	*Momocordia cochinchinensis* trypsin inhibitor
min	minute(s)
ML	machine learning
MLE	maximum likelihood estimation
mPEG	methoxy-polyethylene glycol
MLP	multilayer perceptron
MPLC	medium pressure liquid chromatography
mRNA	messenger RNA
MS	mass spectrometry
MTBE	methyl tert-butyl ether
MW	microwave
n	nano
NAA	N-terminal amino acid
NBS	*N*-bromosuccinimide
NCV	nested cross-validation
NGS	next generation sequencing
NHS	*N*-hydroxysuccinimide
NMR	nuclear magnetic resonance
OOD	out-of-domain
PB	sodium phosphate buffer
PBS	sodium phosphate buffer saline
PCA	principal component analysis
PCR	polymerase chain reaction
PC	principal component
PeT	photoinduced electron transfer
PEG	polyethylene glycol
PITC	phenyl isothiocyanate
PLL-*g*-PEG	poly-L-lysine grafted polyethylene glycol
POI	protein of interest
ppm	parts per million
PPOI	peptide of interest
PTM	post-translational modification
q	quartet
ReLU	rectified linear unit
RhoBo	rhodamine bisboronic acid
RiPPs	ribosomally synthesized and post-translationally modified peptides
RNA	ribonucleic acid
RNN	recurrent neural networks
rpm	rounds per min
rSAM	*S*-adenosyl-L-methionine radical superfamily
s	second(s); singlet
SAMMI	spectrally accurate modeling of multiply charged ions
sCGrx1p	C30S mutant of yeast glutaredoxin 1 (single-cysteine glutaredoxin)
SD medium	synthetic dextrose medium
SEC	size exclusion chromatography
SG medium	synthetic galactose medium
SiO_2	silica gel

SiRhoBo	silicon rhodamine bisboronic acid
SMLM	single-molecule localization microscopy
SMPC	1-octadecanoyl-2-tetradecanoyl-sn-glycero-3-phosphocholine
SNP	single nucleotide polymorphism
SPPS	solid-phase peptide synthesis
SSC	side scattering
STP	sulfotetrafluorophenyl
SVM	support vector machine
t	triplet
TBAF	tetrabutylammonium fluoride
TFA	trifluoroacetic acid
THF	tetrahydrofuran
THPTA	tris((1-hydroxy-propyl-1*H*-1,2,3-triazol-4-yl)methyl)amine
TIC	total ion chromatogram
TIPS	triisopropylsilyl
TIRF	total internal reflection fluorescence
TLC	thin layer chromatography
TSTU	*N*,*N*,*N'*,*N'*-Tetramethyl-*O*-(*N*-succinimidyl)uronium tetrafluoroborate
T_t	transition temperature
UV	ultraviolet
UZH	University of Zürich
W	watt
XantPhos	4,5-bis(diphenylphosphino)-9,9-dimethylxanthene
YPD	yeast peptone dextrose
YSD	yeast surface display

Chapter 1 Introduction

This thesis is dedicated to the development of new tools for identifying proteins. In the first part of this work, we tested the validity of the new proposed method, blinkognition, and started the exploration of its scope. In the second part, we implemented a yeast display-based library screen for a genetically encodable small peptide tag and the synthesis of potential small-molecule binding partners.

1.1 Introduction to part one: blinkognition

Diversity is one of the fundamental principles of biology as it allows for a system to be more resilient and flexible to adapt to environmental change over time.[1] Nevertheless, classical measurements in biology are based on analyzing cell populations in bulk under the assumption that all cells of the population are the same and behave similarly. This approach neglects the fact that the average does not necessarily describe any specific cell within the considered population. For example, if the analyzed population consists of multiple extreme subpopulations, the average does not describe the cellular composition and behavior.[1] However, cellular behavior and interactions of single cells or small subpopulations can be crucial as can be observed in developmental biology.[2] In the early stages of development, a few single cells give rise to the complex system of an organism. These few cells are imperatively heterogeneous, and every cell is decisive simply because no bulk has developed yet. It is therefore essential to study cell-to-cell differences to better understand cellular interactions and communication, as they form the basis of any systemic effect both in a healthy or a pathological context.

1.1.1 The central dogma of molecular biology and proteome complexity

In the central dogma of molecular biology put forward by Francis Crick in 1958 and his commentary in 1970, the hypothesis of the directionality of the information flow between deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins is described (Figure 1). It is emphasized that the information cannot be transferred back to any form of nucleic acid or other proteins once it has reached the level of proteins. This hypothesis still holds based on the current knowledge of molecular biology.[3]

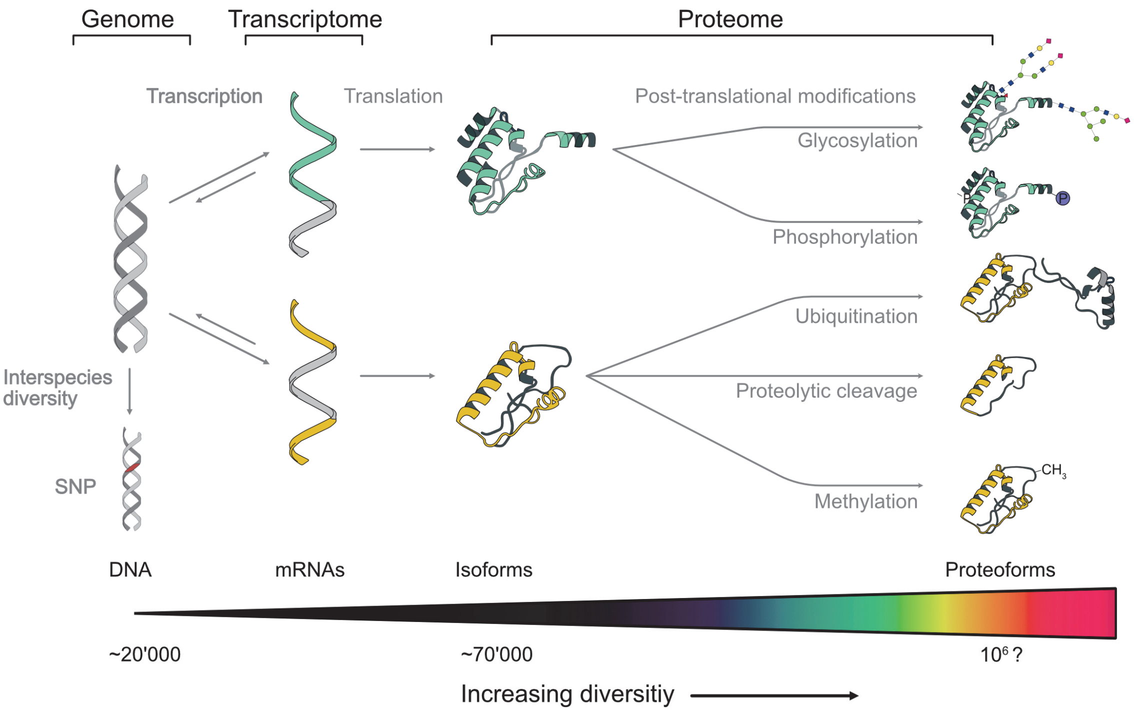


Figure 1. The central dogma of molecular biology and the various processes leading to the vast proteome complexity.[4]

Based on the central dogma the determination of the proteome might seem trivial: Once it is known what is written in the DNA, we can infer the proteins that are present in a cell. However, there is an increasing diversity of versions of proteins generated from one gene (proteoforms) along the information trajectory.[5] From one protein-encoding gene, multiple versions with differing sequences (isoforms), can be generated by alternative splicing, errors in transcription, or translation. In addition to isoforms, further variation is generated after translation by a whole array of post-translational modifications (PTMs) that can alter its location, function, or activity (Figure 1). PTMs range from chemical modification with small groups such as phosphoryl, acetyl, or methyl, to large, complex modification with glycans or ubiquitin. Furthermore, this process is not limited to a single proteoform with a single set of PTMs per protein but a proteoform can carry multiple combinations. Currently, the number of different proteoforms present in a single cell is unknown but can be estimated from the number of protein-encoding genes in the human genome (Figure 1).[4] Based on extrapolations of transcriptomics and current proteomics research, only half of the genome (~10’000) is expected to be expressed in a given cell.[4] Considering a minimally required copy number and the small absolute amount of protein within a cell, the proteoforms are limited well below the theoretically possible number ($\sim 6 \cdot 10^6$).[4]

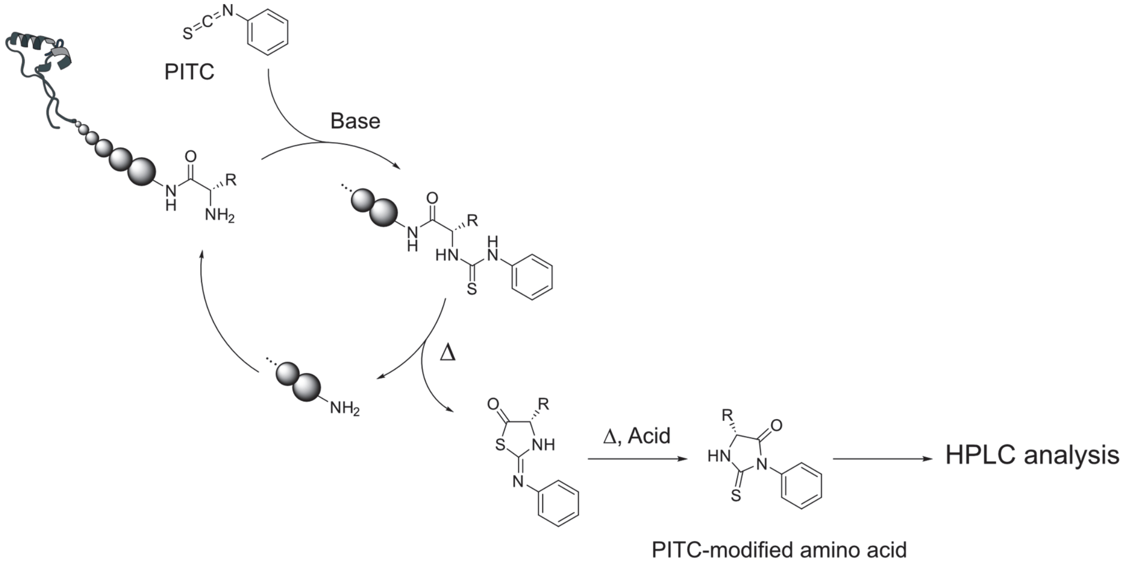
1.1.2 Fundamental challenges in proteome analysis

The analysis of genomic sequences has become extremely fast and affordable since the initial sequencing of RNA and subsequently DNA in the 1970s.[6] The analysis of the transcriptome informs on the cell state and the gene expression but the proteome diverges from the transcriptome due to variable translation efficiencies, different protein stabilities, protein degradation, and modification of translated proteins.[4] Therefore it is paramount to analyze the proteome directly to obtain a full picture of what constitutes a certain phenotype. Furthermore, the proteome varies from cell to cell requiring the analysis of each cell separately.[7] Each cell provides little analyzable protein mass containing a large number of proteins present in abundances that can vary over six orders of magnitude or more (e.g., in serum, 10-10^{11} copies).[8] Furthermore, proteins cannot be amplified nor the protein information be transferred into DNA.[9] Therefore, any method would need to be highly sensitive and massively parallelizable to detect rare proteins, present in as few as 10-100 copies among the very abundant proteins.[10–12] Additional difficulties arise because of the physicochemical diversity of proteins due to the large number of building blocks (20 amino acids), different sizes, shapes, charges, solubilities, affinities, and interprotein interactions making experimental handling difficult.[8]

1.1.3 State-of-the-art proteome analysis and limitations

1.1.3.1 Origins and approaches

One of the first approaches toward protein sequence analysis was Edman degradation, published by Edman Pehr in 1950 and has been developed to be commercialized as peptide sequencers.[13] In this process, phenyl isothiocyanate (PITC) reacts with the necessarily free N-terminus of a peptide under basic conditions. Decreasing the pH leads to cyclization and cleavage of the PITC-modified peptide N-terminal amino acid (Scheme 1). The phenylthiohydantoin amino acid derivative can then be determined via its retention time in a high-performance liquid chromatography (HPLC) run. Even though the reaction has been adapted to commercial automated systems, the process is slow with ~50 min per residue, requires pure protein, and is limited to ~30 residues.[14] Edman degradation gives full sequence information but it is inherently slow and requires large amounts of sample for reliable amino acid derivative detection.

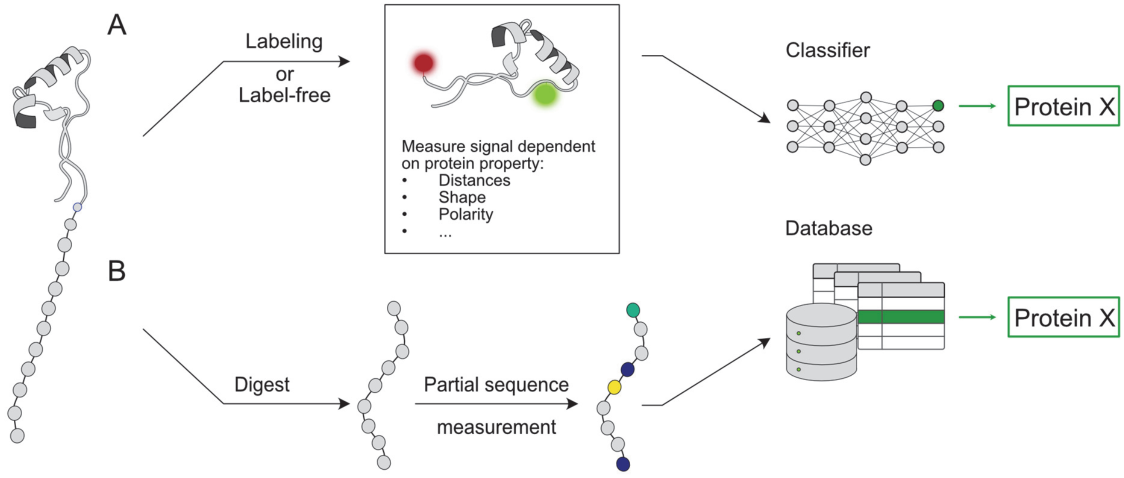


Scheme 1. The Edman degradation process.[13]

1.1.3.2 Fingerprinting and targeted analysis

As an alternative to a full sequencing approach, methods have been proposed to lower the complexity by subsampling proteins. These are generally termed fingerprinting approaches.[15] These methods take advantage of the limited sequence coverage observed in organisms (Scheme 2, B). Based on this premise only partial knowledge of a protein sequence is required to find its identity in a database with absolute or high certainty. Multiple simulation studies have shown that only the relative position of a subset of the proteogenic amino acids is needed for protein identification.[15–18] This approach lowers the number of distinctive readouts required by a method. However, it relies on the availability of reference databases and fails when proteins only differ in segments of unsampled regions. Such a scenario likely occurs in the presence of single nucleotide polymorphisms (SNPs) and homologous proteins or PTMs.

In a targeted analysis, a selected set of proteins is analyzed in a classification approach, also called protein profiling (Scheme 2, A). However, this setting usually requires preselection of the protein of interest (POI) by fractionation, pull-down, or immunoprecipitation.[14]



Scheme 2. Approaches of targeted or fingerprinting approaches. A, Targeted approaches where properties of a protein are directly measured either via a label or directly from the protein. These methods require previous knowledge or properties that can be modeled. B, Fingerprinting approaches obtain partial sequence knowledge and the relative distribution, which allows the prediction of the POI with high certainty from databases. Usually, proteolytic digestion is required before amino acid determination.

1.1.3.3 Protein profiling using antibodies

Antibodies are immunoglobulin proteins that are secreted by plasma cells to support the clearance of an entity recognized as foreign by innate immune cells. Each antibody recognizes a specific region, epitope, on an antigen that it will bind.[19] Antibodies raised against a specific POI can be used to check the presence of a protein by labeling the antibody with a fluorophore or another label that is independent of spectral separability to increase multiplexing, such as DNA barcodes.[20,21] Theoretical detection limits of antibodies have been estimated to be 100 to 1000 POI copies thus they are highly sensitive.[14] However, the use of antibodies can be hampered by low specificity resulting in off-target labeling and false positives. Furthermore, the use of antibodies is limited by their availability for desired targets, which is intensified if an antibody is required for a specific site or modification, like a PTM.[14] Nevertheless, antibodies have become an indispensable, powerful tool for target analysis although one should be aware of potential issues.

1.1.3.4 Mass spectrometry

To date, the proteomics methods of choice are based on mass spectrometry (MS), usually in the form of liquid chromatography-tandem MS (LC-MS/MS). The coupling of MS to LC allows for the fractionation of the proteome before the MS measurement.[8] The standard procedure of label-free MS for the analysis of a proteomics sample is a “bottom-up” (shotgun) approach in which the proteins in the sample are digested by a protease before MS analysis.[22] After digestion, the sample is separated by chromatography in flow with electrospray ionization (ESI), where the ions required for MS analysis are generated. The ions are detected by their mass-to-charge ratio (m/z). After the m/z measurement of the full peptide ions, they are fragmented into smaller peptide ions. The assembly of fragment ions reveals the full or partial sequence of the peptides obtained from the enzyme digest.[23] Bottom-up MS resembles protein fingerprinting, as it generally does not deliver full sequencing information but infers protein presence from characteristic tryptic peptides predicted through database searches. The assignment is complicated for isobaric amino acids (i.e., iso- and leucine) and the small size of the non-overlapping tryptic peptides.[8] Hence this approach is problematic for proteins with sequence similarities such as homologs, isoforms generated by alternative splicing, SNPs, or with the correlations of multiple PTMs.[22] Beyond these fundamental aspects, LC-MS/MS is limited by its sensitivity, which currently lies at 5000-50000 required copies in a mammalian proteome.[8,12,24] Although theoretically not restricted by the MS detectors, the bottleneck is the ionization efficiency of the sample.[8] Ion economics (e.g., a third of the generated ions is measured), as well as the low signal-to-noise ratio (especially in complex samples), complicate data analysis and reduce assignment rates to ~15-50% of the measured fragments, contributing to lower sensitivity.[8,14,25,26] As a result MS. has a limited dynamic range of 10^4-10^6.[27,12,10,8] While the common “bottom-up” approach alleviates some of the challenges of the physicochemical variability amongst proteins and helps increase the proteome coverage, it aggravates the problem of the dynamic range of protein concentrations.[8] There have been many advancements toward increasing the throughput by streamlining sample preparation, various multiplexing approaches, and improved data analysis.[7,28–30] Currently the throughput limits lie at ~8000 quantified proteins per hour in a multiplexing approach with ~1000 proteins per human cell.[26,31] As opposed to bottom-up, top-down MS measures the full protein without a digestion step. Hence this approach circumvents the problem of assigning peptides to a protein and in principle allows the analysis of isoforms, proteoforms, and even protein complexes.[32,33] However, it is experimentally complex and not widely accessible. Furthermore, it does not offer a mode for parallelization and is currently low throughput.[34]

Ultimately, the choice between top-down and bottom-up proteomics depends on the specific goals of the experiment, the nature of the biological samples, the available instrumentation, and expertise. Nevertheless, MS practically still faces practical challenges when analyzing the full depth of a proteome of a small sample, such as a single cell.[23,35,14,24]

1.1.4 Single-molecule protein analysis and emerging tools

Sparked by the successes in single-molecule DNA and RNA sequencing, new innovative single-molecule approaches have been proposed for the analysis of the proteome. Single-molecule methods hold great potential due to the ultimate sensitivity and potential for massive parallelization.[36] These methods are in principle capable of detecting every species eventually if they can screen enough or all molecules in a sample. The later conditions will eventually depend on the speed and scalability that can be reached by any of the proposed methods.[8] At the time of writing, all the mentioned methods are in the proof-of-principle or early development stage, and the applicability to complex mixtures with throughput and parallelization remains to be demonstrated. We will discuss promising methods organizing them into four main groups, nanopore-based, DNA- and fluorescence-assisted methods.

1.1.4.1 Nanopore-and nanochannel-based approaches

Nanochannel technologies aim at identifying molecules via a signal elicited when they are confined in a channel. We distinguish here between nanopore methods that use changes in the ion flow through the pore or nanogaps that measure electric current between electrodes via electron tunneling (i.e., the quantum tunneling effect) across the gap. In both cases, an electric current change is induced by the molecule in the constriction.[37,38]

Quantum tunneling

The applicability of the quantum tunneling effect for analyte detection was initially demonstrated for DNA bases. The electrical currents induced by the bases were distinctive and independent from the adjacent bases.[39] These efforts were subsequently translated to attempts for peptide sequencing using similar setups (Figure 2). To make tunneling signals more specific and lower background noise, the electrodes were modified with recognition molecules (Figure 2, B). When a single molecule, e.g., an amino acid, is tightly bound to the recognition molecules, this creates a short path between the two electrodes allowing for tunneling currents across the electrode gap. Tunneling currents depend exponentially on the position, resulting in a pattern of current spikes upon binding due to thermal vibrations of the analyte (Figure 2, C).[40] These spike clusters in each binding event depend on the analyte properties and the binding with the recognition molecules, therefore they are a molecule-specific signature.[40] It has been shown that current signals obtained from solutions containing different amino acid building blocks or peptides differing in a few amino acids or phosphorylation can be distinguished using a machine learning (ML) approach based on features obtained from the automatically identified signal clusters (Figure 2).[40,41]

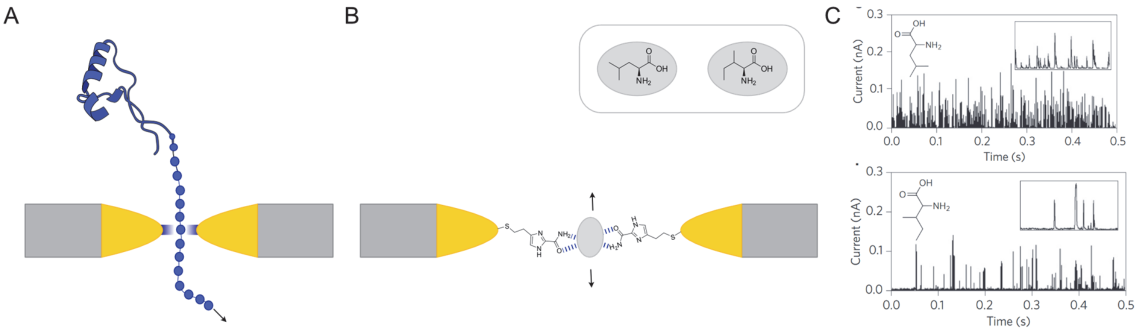


Figure 2. Working principle of quantum tunneling approach. A, Ideally the protein will be moving through the pore and, as each amino acid interacts and allows for a tunneling current, the amino acids are read. B, Iso- and leucine were distinguishable with a 4(5)-(2-mercaptoethyl)-1*H*-imidazole-2-carboxamide modified nanogap. C, Example currents measured from the analytes in B. Data adapted from Y. Zhao, et al., *Nat. Nanotechnol.* 2014, *9*, 466–473.[40]

Nanopore sequencing

The idea of nanopores for sequencing sparked in 1996 when Kasianowicz et al. first observed the translocation of DNA and RNA across a nanopore when an electric field is applied.[42] Since this first proposal of nanopores as a potential solution to DNA sequencing, they have been developed extensively and have been commercialized mainly by the company Oxford Nanopore.[43,44] In the measurements, the stream of ions moving through the pore, due to an applied electric field, will be obstructed by analytes passing across the pore. The decrease in the flow depends on the analyte's volume, shape, and intrinsic electric field, which results in a lower ion current in the measurement.[43,44] To eventually sequence a protein, it needs to be unfolded and move through the pore. As each amino acid reaches the sensitive (narrowest) area in the pore it will be recognized by a specific change in current (Figure 3). The success of nanopore approaches for protein sequencing will depend on two important aspects: first, the accuracy with which each single amino acid can be detected, and second the availability of a translocation mechanism that allows reliable matching of the measured signal to an amino acid.

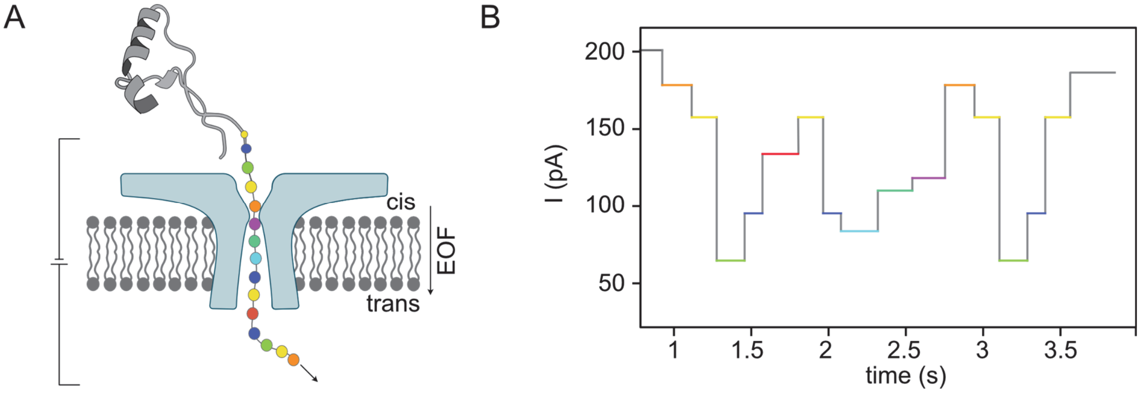


Figure 3. Idealized nanopore sequencing experiment. A, Setup of a nanopore sequencing experiment. A nanopore analyzes each amino acid while unfolding the protein through a pore. B, A model result of a nanopore reading each amino acid is resolved with different current levels and reproducible dwell time.

Accurate and independent distinction between amino acids in a nanopore is intrinsically difficult due to the small variety of amino acid volumes (0.06-0.23 nm^3) and the short distance between residues (~0.38 nm).[45] The specific current assignment could potentially be improved using pores with multiple sensing regions (applied for DNA).[44,46] Residue-wise sequencing is challenged by the lack of a controlled means to unfold a protein and translocate the amino acid chain through a nanopore. The translocation should ideally be unidirectional and steady, to allow relative sequence assignment with a speed resulting in sufficiently long dwell times per amino acid for identification.[45] However, in comparison to DNA, the capture and translocation of proteins are complicated by their vast range of physicochemical properties. The inhomogeneous charge distribution does not allow for uniform translocation by an electrophoretic force.[44] Alternative approaches, such as the use of an electroosmotic flow or an enzyme-driven translocation, have been demonstrated but the translocation is too fast (< 0.1 ms per amino acid vs ~2 ms per base pair in DNA sequencing) and irregular for amino acid identification.[47–53]

Nevertheless, a vast body of work, that has been summarized elsewhere, has shown that nanopores can distinguish peptides differing in an amino acid or a modification in a targeted setting with known analytes and mechanisms to translocate peptides.[37,45] However, to the best of our knowledge, no published work has achieved amino acid by amino acid de novo sequencing as it has been reported for DNA.[44,45,54] In the future, the field could profit substantially from current bioinformatics advances in both data analysis and experimental design. For example, computational protein design could potentially be used to design pores that feature additional sensing mechanisms in addition to the dipole and amino acid volume.[55,56] Alternatively, a recent report proposed using a peptidase conjugated to a pore that allows the analysis of the released amino acids, circumventing the issue of analyzing an intact protein.[57–59]

Full protein recognition with a nanopore

Alternatively, targeted protein profiling applications using nanopores have been proposed. Multiple groups have worked toward measuring fully folded proteins in a nanopore. The resulting signal depends on the size, shape, dynamics, and charge distribution of the POI. Additionally, such analysis could be applied to distinguish proteins of the same isoform but with different conformations or PTMs. To obtain as much information as possible, the protein needs to be retained within the nanopore for extended times. Approaches to prolonging the confinement time have used large biological and solid nanopores combined with lipid-tethered proteins, plasmonics, or an electro-osmotic trap (NEOtrap).[60,61] In the NEOtrap the protein is confined by an electroosmotic flow toward a pore that is blocked for protein but not for ion passage by a DNA origami.[60,61] This principle allows for extensive observation (over hours) of proteins in the pore and how they affect the current by their presence.[60,61] Hence it could potentially be used to fingerprint proteins or their modification by current characteristics using ML. Further, this approach could be extended to observe protein dynamics, interactions, or even reactions such as polymerization.[62] Key advantages of this method are its label-free nature and the long periods it can observe a target.

1.1.4.2 Optical approaches

Beyond MS and nanopores, there is a great interest in developing proteomics analysis tools based on fluorescence. Fluorescence measurements in a controlled setting allow for single-molecule sensitivity which is applied in DNA sequencing methods such as Illumina systems.[63–65] To date, however, to the best of our knowledge, all single-molecule protein analysis methods that rely on fluorescence are not capable of de novo sequencing due to a restricted number of fluorophores that can be spectrally resolved as well as lacking chemistry to target each amino acid selectively. Therefore, the discussed methods propose protein fingerprinting or profiling (see 1.1.3.1). The high sensitivity and spatial control are achieved by immobilizing the construct of interest on a functionalized glass surface. Surface treatment allows the addition of an attachment handle and gives the surface anti-fouling properties to reduce non-specific binding. For imaging, total internal reflection fluorescence (TIRF) microscopy is used. This microscopy technique allows for minimal sample irradiation, close to the surface by creating an evanescent field (~100 nm thick) through total reflection of the excitation laser. This limited irradiation substantially increases the signal-to-noise ratio for signals close to the surface, making TIRF microscopy a highly sensitive method.

Fluorosequencing

One of the earliest methods for single-molecule protein fingerprinting using fluorescence was fluorosequencing, demonstrated in 2018.[16,36] This method uses a combination of Edman degradation (see 1.1.3.1) and fluorescence microscopy for protein fingerprinting, which makes it highly parallelizable.[36] In this technology, small peptides are labeled with a fluorophore of a designated wavelength per amino acid (e.g., cysteine, lysine) and attached to a glass surface (Figure 4, top row). The peptides are then degraded using classical Edman conditions (PITC, strong acid, heat) and after each cycle, the fluorescence is monitored. Depending on the number of labels, and hence amino acids, present a higher intensity level will be observed (Figure 4, middle row). If the level remains the same after each cycle, it can be concluded that the position was occupied by an amino acid that does not carry the fluorophore. Conversely, when a stepwise decrease in fluorescence is observed the position can be attributed to the amino acid of the corresponding channel (Figure 4, lower row).[36] The method is based on well-established and highly optimized Edman degradation (yields of 91-97%).[36] Nevertheless, the error rate of unsuccessful cleavage and the resulting dephasing of amino acid annotation limits the method to peptides with less than 30 amino acids.[14] The harsh conditions required for peptide cleavage pose an additional hurdle. The authors have shown that it is indeed compatible with an imaging system and surface chemistry. However, the degradation conditions can damage the fluorophores and are incompatible with certain labeling strategies. Dye decoupling or destruction would lead to additional wrong assignments, but it has been proposed that these cases could be partially filtered out later on during analysis.[15] Similar problems can occur due to photobleaching of the dyes as they need to be imaged after each cleavage cycle. In addition to these technical issues, the currently accessible labeling strategies remain a limitation. PTMs add another layer of complexity as they can only be detected if they can be labeled directly or derivatized into a fluorophore, which has been only demonstrated for phosphorylation.[36,66] Lastly, since fluorosequencing requires a digestion step, multiple fragments of each protein are generated that will have to be sequenced. This step increases the number of reads required to detect all proteins in a cell. An analogous effect is observed in RNA and DNA sequencing due to short read lengths, but the dynamic range of RNA (10^3) is significantly smaller than for proteins (10^6).[8,66]

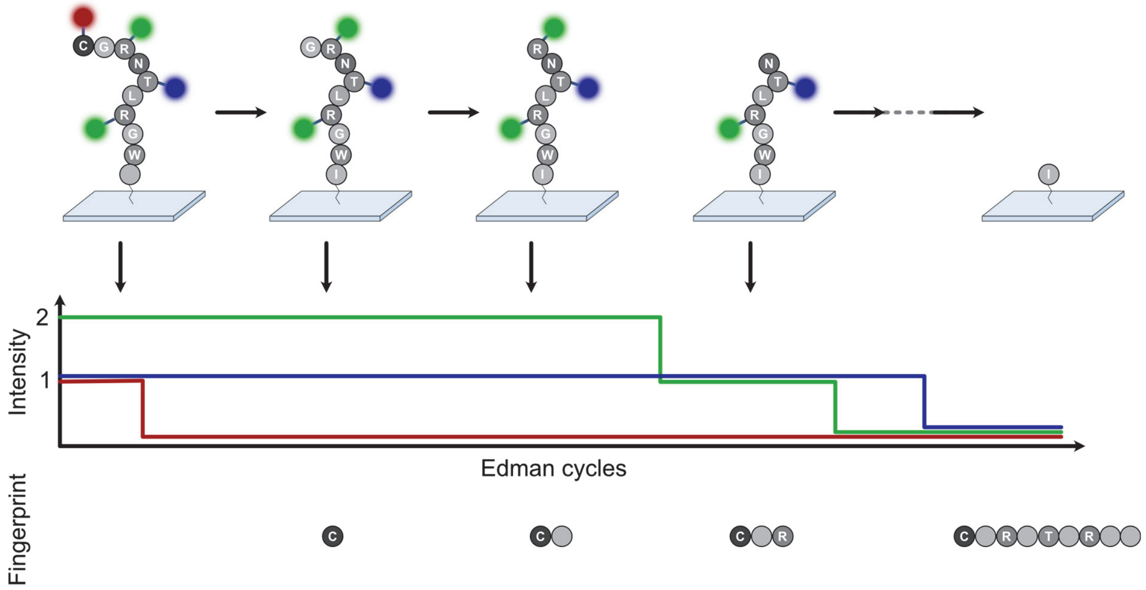


Figure 4. Working principle of fluorosequencing. Consecutive degradation and imaging reveal the fingerprint of a labeled peptide. The figure was adapted from J. Swaminathan, A. A. Boulgakov, E. M. Marcotte, *PLoS Comput. Biol.* 2015, *11*, e1004080.[16]

Förster Resonance Energy Transfer via DNA eXchange (FRET X)

In FRET X certain amino acid types (i.e., cysteine and lysine) are labeled on the native protein with their own set of DNA docking strands that can be transiently bound by an imager strand labeled with a FRET donor (Figure 5, A). At the same time, the N- or C-terminus is modified with a docking strand targeted by a FRET acceptor labeled imager strand (Figure 5, A). The docking strand also allows for binding to a functionalized glass surface to immobilize the construct.[18,67,68] Then images of the construct in the presence of each type of imager corresponding to the applied docking strands are acquired. The recorded fluorescence intensities in the acceptor channel during excitation (with the donor laser) can be used to calculate the FRET efficiency, yielding histograms of all FRET efficiencies (Figure 5, B and C). As a result, a histogram containing all highly resolved, relative distances of the labeled amino acid to the fixed point (N- or C-terminus) is obtained. This histogram is the FRET fingerprint of the respective protein (Figure 5, D) and can be compared to a reference database of computationally predicted fingerprints.

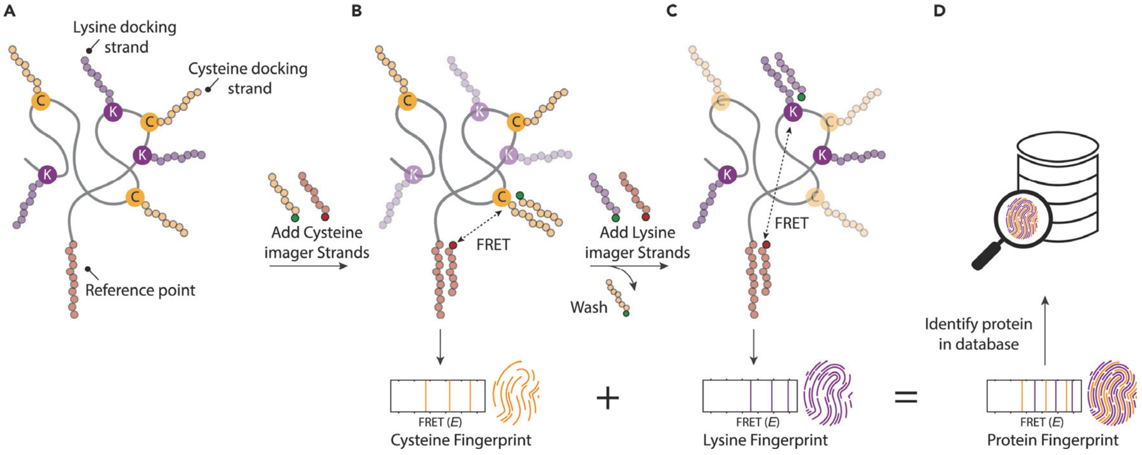


Figure 5. The working principle of FRET X. A, A POI is modified with the docker strand for the amino acid types and the reference point respectively. B and C, The docker strands are imaged in the FRET channel, separately but in the presence of the fix point imager strand. D, The final fingerprint is compared to FRET fingerprints in a database of experimental or computationally predicted FRET fingerprints. This figure was adapted from C. V. de Lannoy, M. Filius, R. van Wee, C. Joo, D. de Ridder, *iScience* 2021, *24*, 103239.[18]

Protein sequence binders

The founders of the company Nautilus Biotechnology have proposed an alternative fingerprinting method using fluorescently labeled binders of amino acid trimers in a protein identification approach called: Protein Identification by Short-epitope Mapping.[69] The trimer-specific binders are added sequentially to proteins that have been denatured and attached to defined regions on a surface.[70] The binding stochastics are evaluated with an ML model to determine the likelihood of the presence of the amino acid sequence. After many cycles of binder addition and washing, the presence of certain amino acid trimers narrows down the number of possible proteins that contain these trimers. Based on simulations it is estimated that 150–200 different binders would be required to cover <90% of the whole proteome not including PTMs.[69] In combination with the high-density functionalization of surfaces that can host 10^{10} proteins (fully occupied), it is predicted to allow for the detection of >95% of a single copy of each protein in a HeLa cell. This coverage drops significantly when serum samples with notoriously large dynamic ranges are simulated.[69] Indeed, the method has the advantage of sampling full proteins but does not currently propose a solution for PTMs and to the best of our knowledge, no experimental evidence of the method has been published.[69]

Protein end-binders

The most recently commercialized (by QuantumSi) fluorescent-based single-molecule approach toward full proteome uses *N*-terminal amino acid (NAA) binders and proteolytic degradation instead of chemical degradation (Figure 6).[71] NAA binders are derived from ATP-dependent Clp protease adaptor proteins that naturally function as *N*-recognins in the bacterial degradation machinery.[72] The technology immobilizes peptides on the surface of a waveguide and probes the NAA binding by measuring the fluorescence lifetime rather than the fluorescent intensity. The limited number of covered amino acids by available NAA binders (three) and eventually limited spectral resolution is extended to more amino acids by including NAA-characteristic binding patterns via the mean pulse duration over many binding events. However, based on the number of *N*-terminal binders available and the amino acid identities they reveal, the method to date remains a fingerprinting approach. The cleavage and hence sensing of the next amino acid is mediated in a stochastic manner by aminopeptidases that are permanently present in the solution with average sampling times on the order of 30–90 min but can be substantially longer.[71] Therefore, the method is currently slow and takes multiple hours even for short peptides, although cleavage rates could in the future be optimized by protein evolution. A similar approach for NAA-binders evolution or application of aptamers was proposed for PTMs, which have currently not been included in the analysis. The technology is promising but requires future development to cover the analysis of samples of unknown protein and higher complexity. So far, no de novo sequencing has been shown nor disclosed how amino acids are unequivocally assigned considering the exhibited variance of pulse durations depending on the neighboring sequence.

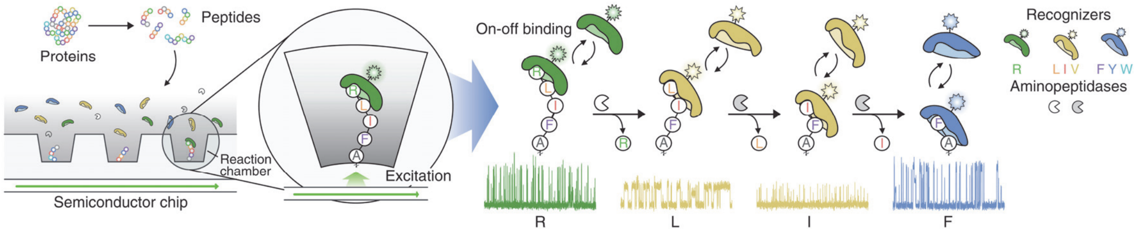


Figure 6. Working principle of real-time dynamic protein sequencing. Proteins are proteolytically digested and tethered in a reaction chamber. Signal is acquired in the presence of all NAA binders and aminopeptidases. The NAA is cleaved stochastically revealing a new NAA to be probed. The temporal sequence is determined considering the binder identity and binding kinetics. The Figure was taken from B. D. Reed et al, *Science* **2022**, *378*, 186–192. Reprinted with permission from AAAS.

1.1.5 The concept underlying blinkognition

We discussed challenges posed by analyzing the proteome and approaches that have been put forward by the field. However, to date, no solution has been developed that fulfills all requirements and there is likely a combination of methods needed to explore different aspects of the proteome. Therefore, new ideas and approaches are critical to access new opportunities. In this work, we will present the exploration of a novel approach for capturing information about peptides and proteins based on the fluorescent signal of spontaneously blinking fluorophores. The information contained in the blinking signals could be used to reidentify a POI from a mixture of proteins labeled with the same fluorophore in a fingerprinting approach. It benefits from the single-molecule detection limit of fluorescence microscopy without the need for sequential protein degradation. We termed the method "blinkognition", a portmanteau from blinking and recognition.

The intermittency behavior of fluorophores (or blinking) describes the alternation of fluorophores from a fluorescent state to a dark and back under continuous irradiation. Blinking can be caused by multiple phenomena and quenching mechanisms, which are usually not fully understood. Among the explanations that have been proposed are triplet-state quenching, or changes in orientation of the fluorophore in the molecule. However, the potential mechanisms at play will depend on the system that is imaged (fluorophore and environment).[73–78] Blinking has been considered a nuisance for continuous single-molecule fluorescence imaging as it can interfere with the interpretation of the phenomena under investigation.[77,79] Conversely, some studies have used the blinking behavior of the well-known dyes Cyanine 3 (Cy3) and Atto655 to probe biological systems, more specifically distinct conformations of nucleic acids.[80–82] In the first case, the blinking of Cy3 results from the cis-trans isomerization of a double bond in the conjugation pathway. This mechanism is strongly dependent on steric effects.[78,80] Therefore, Cy3 bound to DNA shows different blinking kinetics depending on the DNA conformation. In the second case, the blinking of Atto655 was modulated by placing it on a DNA strand that would allow for conformation-dependent accessibility to a specific redox buffer required for the blinking to occur.[81] Beyond these examples, the advent of super-resolution microscopy techniques, more specifically in single-molecule localization microscopy (SMLM) approaches, blinking became a desired property.[83] In the initial protocols for SMLM, most dyes would have to be bleached to achieve sparse labeling or photoactivatable dyes would be used. However, these require cycles of photoactivation, imaging with intense irradiation leading to phototoxicity in live-cell imaging.[84] In 2014, Uno et al. introduced spontaneously blinking fluorophores, i.e., hydroxymethyl silicon rhodamine (HMSiR), as an alternative for SMLM (Figure 7).[83]

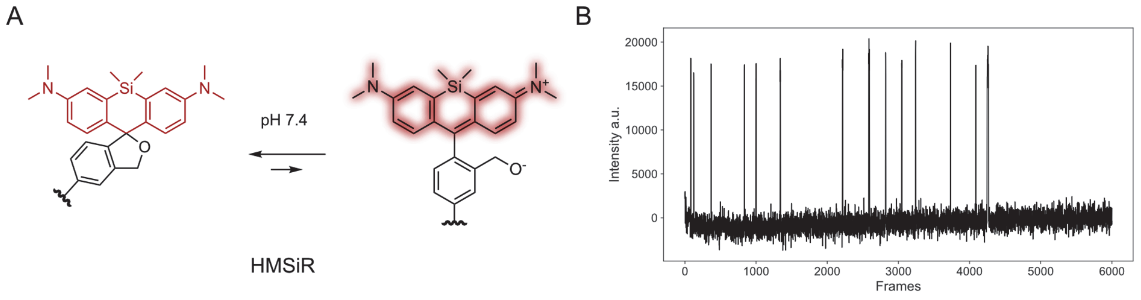


Figure 7. Spontaneously blinking fluorophore. A, The chemical structure, and equilibrium of HMSiR.[83] B, Example trace obtained from immobilized HMSiR on a glass surface.

In this work, we show that spontaneously blinking dyes can be used for a more general application. The fluorescence intermittency of spontaneously blinking fluorophores is the result of a ground-state isomerization, i.e., a thermal equilibrium, between a fluorescent and a non-fluorescent isomer with a low barrier of interconversion. If the two isomers in the equilibrium are chemically dissimilar, i.e., if they differ in charge, polarity hydrogen-bonding ability, π-surface, protonation, etc., we can expect the relative stability of the isomers and their barrier of interconversion to depend on the environment (Figure 8, A). More specifically, the relative energy levels and transition trajectory will vary depending on the availability of interactions in the environment of the fluorophore.[85–88] As a consequence, the fluorescent signal over time of a spontaneously blinking dye represents fluorophore-environmental interactions such as electrostatic and hydrophobic interactions, hydrogen bonding, etc.[85–89]

We hypothesized that in any dye conjugate these interactions would be dominated by the conjugation partner. For instance, if we consider fluorophore-peptide or -protein conjugates, the interactions with amino acid side chain functionalities and the backbone of the given construct will impact the thermal equilibrium (Figure 8, B). Such constructs are dynamic hence they will be moving with and adapting to each other, resulting in a construct-dependent blinking pattern (Figure 8, C). Considering the complexity of the information, the numerous processes, the interactions, and the time scale involved in generating such a fluorescence time trace make it currently impossible to predict or model such blinking patterns.

Nevertheless, we propose that it could be possible to harness the information in a data-driven approach using ML. In such a methodology, a large amount of data for training a model is required. The ground-truth data can be obtained by measuring the fluorescence signal from a purified synthetic sample. As in fluorosequencing and other previously discussed methods, the ideal experimental method for low background signal is TIRF microscopy.[90] By fixing the pure sample on a glass coverslip, the labeled training data can be recorded and used to recognize a specific peptide in a future mixture (Figure 8, D). The viability and initial proof-of-principle are explored in this work using pure samples of peptides and proteins. However, further exploration and improvements will be necessary to determine achievable sensitivity and robustness for the practical application of this approach to more complex samples.

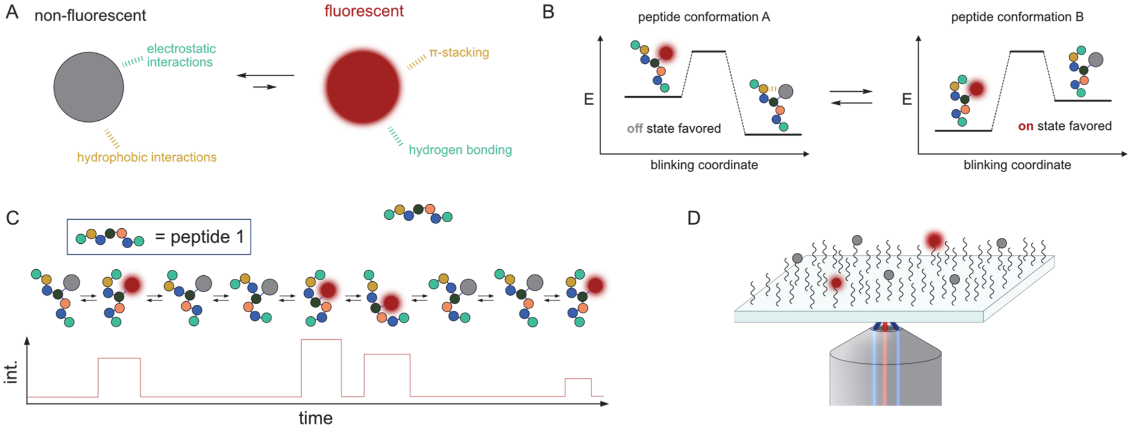


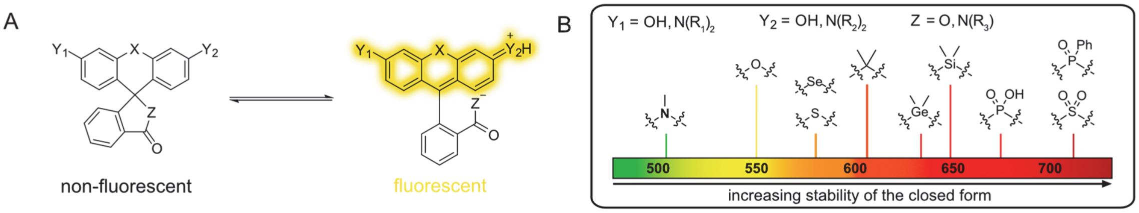
Figure 8. The hypothesis underlying blinkognition. A, Spontaneously blinking fluorophores are in a thermal equilibrium between a fluorescent and non-fluorescent form and the equilibrium will be influenced by de- or stabilizing interactions of the two forms. B, The environment of the two forms can be a larger molecule (i.e. a peptide) via a covalent link. The interactions with the conjugation partner will de- or stabilize the two forms. C, The interaction within the conjugate is dynamic and will change with conformational changes of the peptide over time. D, The single molecules can be observed by surface fixation and TIRF microscopy. Adapted from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* 2023, *145*, 1441–1447.[91]

1.2 Introduction to part two: small-peptide tags for protein imaging

Proteins execute most cellular functions, and thus greatly impact the state of the cell and an organism. To fully understand the biological role of a protein it needs to be examined in its cellular context. Fluorescence microscopy offers a minimally invasive way of localizing and tracking proteins in a cell spatially and temporally. A crucial aspect of imaging proteins is targeting fluorescence to a POI. Most used fluorophores for cell imaging are small molecule dyes or fluorescent proteins (FP). In the following, we will introduce currently available tools and strategies.

1.2.1 Xanthene dyes and fluorogenicity

Rhodamine derivatives are a prominent class of small molecular fluorophores that are composed of a xanthene core fused to an isobenzofuran moiety (Scheme 3, left, X, Z = O). An important characteristic of these dyes is that they are present in an equilibrium between a spirocyclic form (with the isobenzofuran moiety) that is not fluorescent ("closed") and a zwitterionic, fluorescent ("open") form. A wide range of such dyes has been developed by tuning the donor/acceptor moieties to shift the excitation and emission wavelength and improve brightness and photostability by restricting rotation or mitigating fluorophore damage.[92–95] The excitation and emission wavelengths can be further modulated by changing the bridging heteroatom in the conjugated ring system (Scheme 3, atom X).[96–99] Beyond tuning the photophysical properties of rhodamines the modulation of the intramolecular equilibrium opens up unique opportunities to adapt dyes for specific purposes such as improving cell permeability, molecular sensing, fluorogenicity, or spontaneous blinking.



Scheme 3. Fluorescence modulation in xanthene dyes. A, Open (fluorescent) and closed (non-fluorescent) equilibrium of xanthene dyes. B, The photophysical changes can be tuned by exchanging the heteroatom X. Visualization in B courtesy Annabell Martin.[97]

Fluorogenicity describes the property of fluorophores to be initially non-fluorescent but become fluorescent when bound to their cellular target. Fluorogenic rhodamine derivatives have been developed for binding of cellular structure and self-labeling proteins (see 1.2.2.3) by shifting the equilibrium towards a more closed form.[100–104] The tendency of rhodamine dyes to be in the open or closed form can be adjusted by modifying the xanthene core or the intramolecular nucleophile.

A (decreased) increased electrophilicity of the core or a (weaker) stronger nucleophile will lead to a more (open) closed rhodamine derivative.[105,97,106] The electrophilicity of the xanthene core can be modulated by changing the electron-withdrawing/-donating properties of donor/acceptor substituents or by functionalizing the core directly with electron-withdrawing substituents.[107,102,103] Exchanging the oxygen in the xanthene system also affects its electrophilicity and not only the excitation and emission wavelength. In silicon rhodamines, the oxygen is replaced with a *gem*-dimethyl silicon moiety resulting in a higher electrophilicity by stabilizing the unoccupied molecular orbital of the core via a Si-C σ*- π* interaction (Scheme 15, atom X).[83,108,109] Therefore, silicon rhodamines are more closed than their rhodamine analogs. A drawback of altering the xanthene core is that such modifications inevitably impact the photophysical characteristics of the dye, and can render it susceptible to intermolecular reactions.[105] The tuning of the nucleophile, instead, avoids these issues and has thus been a valuable alternative, which can be combined with core modifications to fine-tune the equilibrium.[103,104]

1.2.2 Available tools for protein identification in a cellular context

1.2.2.1 Direct labeling of specific proteins

Some proteins can be directly targeted if a protein-specific ligand is available, which can be conjugated with a fluorophore. Prominent examples are Paclitaxel derivatives that bind to β-tubulin and can be used to image microtubules, or Jasplakinolide that binds actin filaments. However, some of these are active molecules that modulate the function of their target. In addition, these are very selective ligands and not generally available.[101,110,111]

1.2.2.2 Immunolabeling

In immunolabelling, proteins are targeted using an antibody. Antibodies can be visualized by fluorescence either by directly labeling the primary antibody (the one targeting the POI) or using a second labeled antibody that recognizes the first one. Multiple secondary antibodies can bind to each primary antibody, which leads to signal amplification. However, antibodies are very large molecules (~150 kDa) and can only be used for extracellular or endocytosed proteins. For intracellular targets the cells need to be fixed and permeabilized, which precludes live-cell imaging and can introduce artifacts.[112,19,113]

1.2.2.3 Genetically encoded tools

Approaches based on genetic engineering represent an alternative to directly targeting an epitope on a POI for labeling. In this case, the POI is modified by fusing it to a tag that can be followed by fluorescence. In the last 20 years, many strategies have been developed to genetically label a POI specifically and selectively. The most common approach is the use of FPs.

FPs

FPs were originally discovered and isolated from a jellyfish (*Aequorea victoria*).[112] This 238-amino acid protein forms a β-barrel structure that provides the environment for a part of the protein sequence to form an internal chromophore via a sequence of spontaneous cyclization and oxidation to adopt a conformation that fluoresces. The elucidation of the mechanism has allowed researchers to engineer them for improved photophysical properties such as brightness, photostability, and excitation wavelength.[114,115] FPs are attached genetically to the sequence of the POI via a linker and are therefore absolutely specific. Although FPs have been extensively improved, they suffer multiple drawbacks, especially in comparison to small-molecule fluorophores. FPs exhibit low brightness and photostability, as well as broad absorption and emission spectra, which complicates multicolor imaging.[116] Furthermore, the conditions required for fluorophore formation preclude acidic, hypoxic environments and very fast processes.[112] Moreover, FPs are large (~25 kDa), which can disturb the folding (i.e., induce aggregation), localization, function, and interactions of proteins, especially in the case of small targets.[117]

Self-labeling proteins

Self-labeling proteins are exogenous or engineered endogenous proteins that bind a small-molecule ligand that can be conjugated to a fluorophore. Hence, they combine the absolute specificity of FPs through their genetic link and the advantageous photophysical properties of small molecules. Prominent examples of self-labeling include Halo-, SNAP-, CLIP-, and TMP-tag (Figure 9, A-D).[118–121] HaloTag is derived from a bacterial dehalogenase enzyme and reacts with chloroalkane-modified molecules (Figure 9, A). SNAP- and CLIP-tag have been evolved from the DNA repair enzyme human *O*^{6}-alkylguanine-DNA-alkyltransferase (hAGT) that transfers alkyl groups from alkylated DNA bases to itself.[122] SNAP recognizes *O*^{6}-benzylguanine (BG) derivatives whereas CLIP recognizes cytosine derivatives (Figure 9, B).[120] TMP tag is based on dihydrofolate reductase from *E. coli* (eDHFR) that binds trimethoprim with a very high affinity (~1 nM) unlike the endogenous mammalian DHFR.[121,123] In principle any fluorophore can be attached to the respective self-labeling protein ligand. However, the use of fluorogenic rhodamine dyes is advantageous, due to their good cell permeability and minimal fluorescence background. These benefits have motivated the development of fluorogenic rhodamine dyes, such as the Janelia Fluors JF526, JF552, and JF646, by fine-tuning the electronic properties of the xanthene core.[102] Alternatively, fluorogenicity has been achieved by incorporating electron-deficient amides as closing moiety, yielding cell-permeable fluorogenic rhodamine dyes (MaP dyes) of a variety of colors.[104,124] Another tag that does not depend on a ligand fluorophore conjugate is the fluorescence-activating and absorption shifting tag (FAST). FAST binds a fluorogen that becomes fluorescent due to its stabilization analogous to the chromophore in the green FP.[125,126] These tags have many advantages over FPs but they are still relatively large with sizes ranging from ~14 kDa (FAST) to ~33 kDa (HaloTag). Hence they remain problematic for sensitive or small targets.[127]

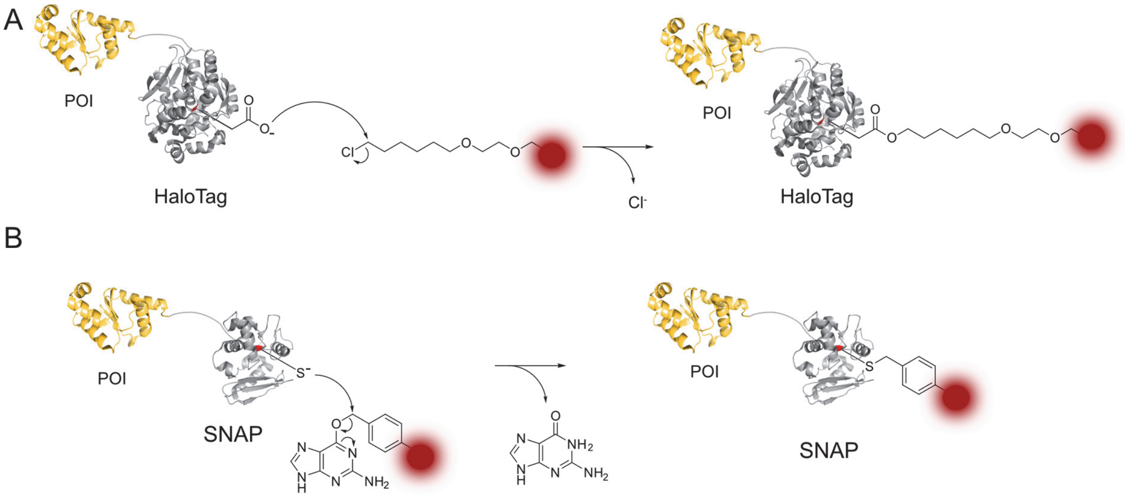


Figure 9. Examples of self-labeling proteins. A, B, The mechanisms of HaloTag and SNAP-tag self-labeling.[118,119,128–130]

Small peptide-modifying proteins

To reduce the tag size, peptide-modifying enzymes such as ligases and transferases have been adapted for protein labeling.[127] In these technologies, a small peptide tag, which gets modified by an enzyme, has to be added to the sequence of the POI.[127,131,132] In most cases, the enzyme attaches a moiety similar to the natural substrate but carries a handle that can be used in a bio-orthogonal reaction. A few enzymes are more promiscuous enabling direct conjugation of a fluorophore. The first example was biotin ligase (Figure 10), an enzyme from *E. coli* that naturally transfers a biotin moiety to the lysine residue in the peptide -GLNDIFEAQ**K**IEWHE- (AP-tag). The biotin can be subsequently labeled with a modified streptavidin or an alkyne- or azide-modified biotin can be transferred by the more promiscuous biotin ligase of *Saccharomyces cerevisiae* or *Pyrococcus horikoshii* and then combined with click chemistry.[127] Many alternative enzymes and tags have been introduced that are more or less flexible in terms of the position of the tag in the POI and the modification of the transferred moiety (Figure 10).[127] Although these labels are small (5-22 amino acids), they face other difficulties. Depending on the organism of interest the enzyme needs to be expressed in addition to the modified POI, the reaction times of these enzymatic conversions are usually long or require high reagent concentrations. These limitations are further amplified if a second bio-orthogonal reaction is required.[127]

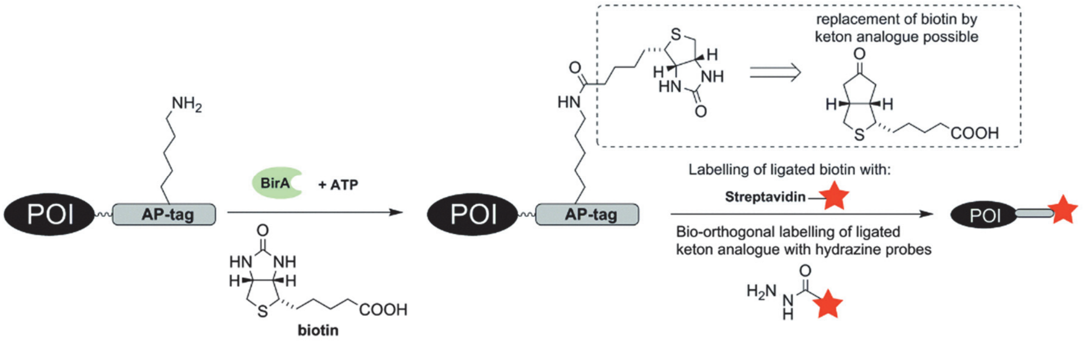

replacement of biotin by keton analogue possible
NH2
BirA
+ ATP
POI
AP-tag
POI
AP-tag
biotin
COOH
Labelling of ligated biotin with:
Streptavidin
Bio-orthogonal labelling of ligated keton analogue with hydrazine probes
POI

Figure 10. Biotin ligase BirA from *E. coli* as an example of a small peptide-modifying enzyme. Figure adapted from J. Lotze, U. Reinhardt, O. Seitz, A. G. Beck-Sickinger, *Mol. BioSyst.* 2016, *12*, 1731–1745.[127]

Small protein tags

Another type of small peptide tag was developed by Tsien and co-workers in 1998 and does not require the action of a protein. This tag consists of 6 amino acids with a tetracysteine motif -CCxxCC-. The first example was the biarsenical fluorophore fluorescein-arsenical-helix-binders (FlAsH) binding to the tetracysteine motif -CCPGCC- (FlAsHTag, Figure 11, A). For labeling in live cells, the protected FlAsH-ethanedithiol (FlAsH-EDT$_2$, Figure 11, A) is required for cell permeability and to prevent toxicity induced by the arsenic atoms in the molecule. FlAsH-EDT$_2$ is non-fluorescent but shows a large fluorescence increase upon FlAsHTag binding. The fluorescence quenching is not yet fully understood and multiple mechanisms such as internal conversion and photoinduced electron transfer (PeT) have been proposed.[133] Additional biarsenical probes with improved photophysical properties, robustness, and different colors such as ReAsH, CHoXAsH, CrAsH (Figure 11, A), F$_2$FlAsH have been reported.[134–137] Furthermore, a Cy3-based probe AsCy$_3$ with an increased interatomic distance (14.5 Å) between the two As(III) and larger target peptide tag -CCKAEAACC- has been reported for orthogonal labeling.[138] Biarsenical dyes have several limitations in addition to their toxicity: substantial background through binding of other thiols, misfolding of the tag-containing proteins due to unintended disulfide bridges, and a limited range of available colors.[127,134,139,140]

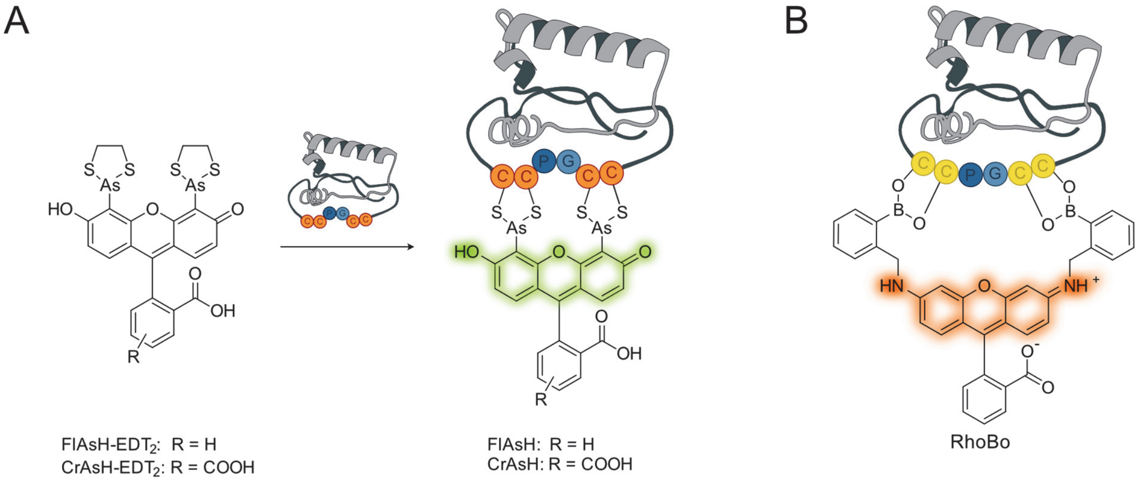


Figure 11. Structure and schematic binding of FlAsH-EDT$_2$ and CrAsH-EDT$_2$ (A) and RhoBo (B).

In 2009, the Schepartz group proposed the use of an analogous, non-toxic tetraserine tag that would be targeted by a rhodamine-derived bisboronic acid (RhoBo, Figure 11, B).[141] RhoBo was originally introduced as a monosaccharide sensor that showed a fluorescence increase upon glucose binding when sugars are used in high concentrations (mM). However, the authors confirmed their hypothesis that the binding of a tetraserine motif -SSPGSS- containing peptide would result in a more stable complex (K_d ~450 nM) compared to monosaccharide binding with significant fluorescence turn-on (Figure 11, B).[141,142] Even though the probe is promising *in vitro*, the presence of the tetraserine motif in more than 100 human proteins results in low selectivity and a high background in cell imaging, making it currently inapplicable for biological live-cell imaging experiments.[127,141] The fluorescence increase is based on the *ortho*-aminomethylboronic acid moiety contained in RhoBo. This motif has been employed in many sugar sensors, yet the working mechanism of fluorescence quenching has been a topic of longstanding debate.[143] The current model supported by most data evokes the “loose-bolt effect”, a quenching phenomenon through internal conversion of the fluorescence energy into O-H bond vibrations of the boronic acid.[143] However, the mechanism in RhoBo specifically has not been studied but in similar fluorophores, a PeT mechanism has been proposed.[133,144]

Unnatural amino acids

The smallest available labels are unnatural amino acids carrying a bio-orthogonal reaction handle. Conceptionally, the incorporation is achieved by reassigning the least abundant stop codon (amber codon) to the new unnatural amino acid. This requires the evolution of an orthogonal tRNA synthetase and tRNA (that recognizes the amber codon) pair. This tRNA synthetase can load the new amino acid onto the tRNA, which is used by the cellular machinery to incorporate the unnatural amino acid during protein synthesis. Although successfully applied, the method is experimentally laborious as an orthogonal tRNA synthetase, the tRNA, and the POI need to be co-expressed, and the unnatural amino acid needs to be provided. In addition, background might occur due to the incorporation of the new amino acid into naturally occurring amber codons.

1.3 Outline

In this work, we develop new tools to identify and localize proteins. The first part of the thesis is dedicated to the implementation and exploration of a new method, blinkognition, that can distinguish peptides and proteins in isolated pure samples on the single-molecule level. In the second part, we describe our work towards a genetically encodable peptide tag that can be bound by a fluorogenic fluorophore using a yeast-display screening approach.

Chapter 2 describes the design, sample preparation, and acquisition of our first model set up to test the basic hypothesis of blinkognition that was presented in section 1.1.5. Following the experimental data collection, the data extraction and preparation are presented. Initial analysis indicated that a different approach and model is required.

In Chapter 3 we further explore the data analysis. ML models are tested and show that blinkognition can work for small peptides with differing sequences and smaller chemical modifications. Uncertainty estimation is introduced to improve classification accuracy by removing traces that exhibit low certainty.

Chapter 4 details the further development of blinkognition toward the analysis of native proteins. A new experimental setup for protein immobilization is introduced and tested. We prepared a reactive version of the spontaneously blinking dye to label proteins, and labeling reactions were tested. Finally, we test the previously developed model of the protein data.

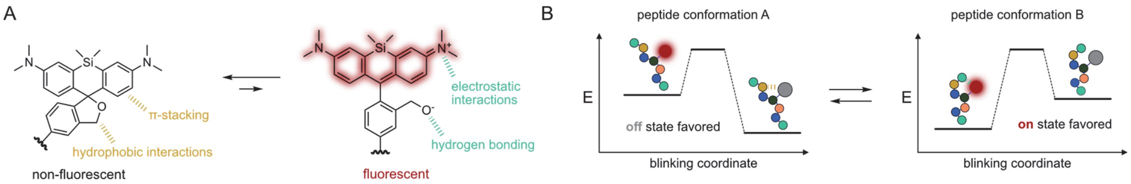
In Chapter 5 we move beyond single-molecule level analysis of isolated peptides and proteins and embark on the search for a new tetraserine-based peptide tag. We combined the concepts of fluorogenicity and the boronic acid-based serine binding of RhoBo toward a new binder with low background fluorescence. We employ yeast display to present libraries of randomized tetraserine peptides to screen for a peptide that can bind and turn on our newly synthesized boronic acid probes.

Chapter 2 Measuring single-molecule fluorescence traces and their analysis.

Parts of this chapter were published in the *Journal of the American Chemical Society*.[91] We thank the imaging center BIOP at the Swiss Federal Institute of Technology in Lausanne (EPFL) and Dr. Luc Patiny for their initial help and discussion on the analysis. We further thank Dr. Thomas Moehl from the laboratory of Professor David Tilley for conducting the AFM measurements and Dr. Eduard Ebberink from the laboratory of Professor Beat Fierz for the introduction and support with the peptide synthesizer. Additionally, we thank Lionel Rumpf for preparing the compounds P15-P17 and Dr. Adam Eördögh for preparing compound 2.

2.1 Goal of the project

In Chapter 1, we discussed the concept of our new tool "blinkognition" for peptide and protein identification. We propose to harness the information on the chemical environment contained in fluorescence intermittency patterns of spontaneously blinking fluorophores (HMSiR) to identify peptides. The interactions and processes contributing to the blinking behavior of a spontaneously blinking dye are highly complex and dynamic over the lifetime of the fluorophore. Therefore, it is not possible to predict or model all the factors ab initio. Instead, we aim at a data-driven ML approach.



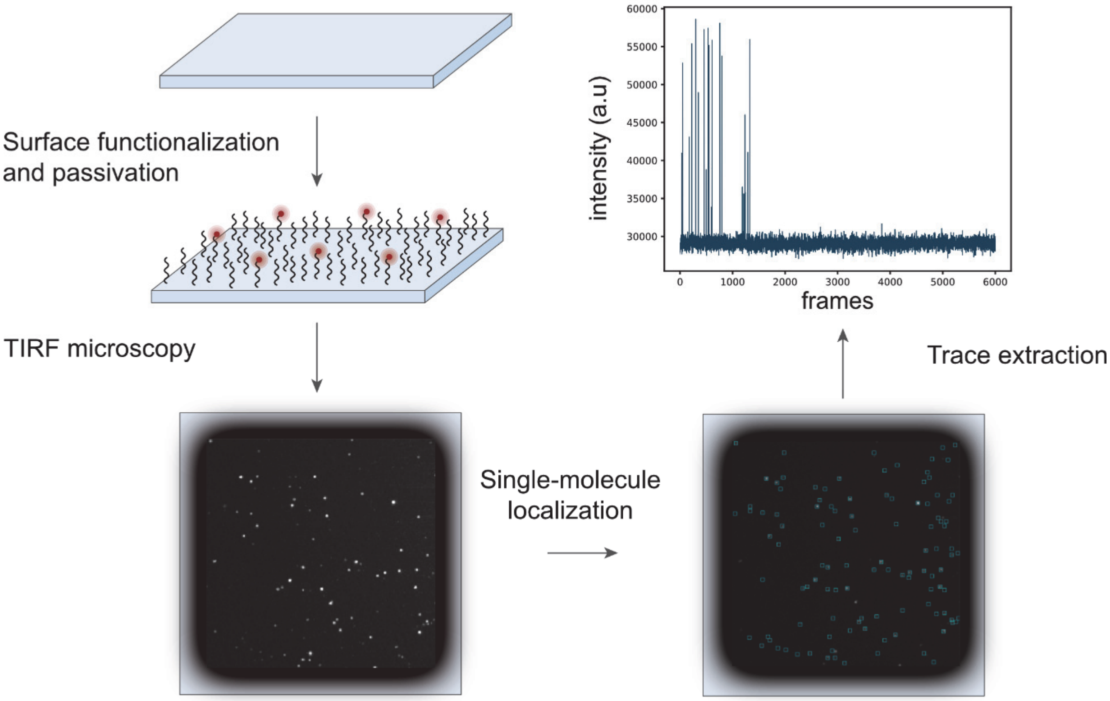
Scheme 4. Overview of the principle employed for blinkognition. A, Spontaneous equilibrium of HMSiR, that is influenced by a multitude of interactions depending on the environment. B, The influence of the microenvironment on the energy level of a dye-peptide conjugate. Adapted from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* 2023, *145*, 1441–1447.[91] 🅭🅯

ML algorithms are used to find hidden patterns in large, complex datasets and build a predictive model for the problem at hand. They have been extensively applied to the analysis of signals obtained from nanopores, which fundamentally have a similar data structure to our fluorescence measurement.[37,145] As demonstrated by this example, ML methods can be used for approaching any problem that meets the mathematical criteria without the need for a deep understanding of the specific mechanisms and causal relationships at play. The context of the data is not a mandatory prerequisite for these algorithms. Consequently, ML methods have been successfully applied in various disciplines, from natural sciences such as chemistry and biology to social sciences and finance, but there are also prominent examples in daily life applications, for instance, voice and face recognition.[146–149]

In this chapter, the development of a workflow to test our hypothesis is presented, including the design and synthesis of peptides labeled with our dye. Further, we would use features obtained from the fluorescence signal over time to attempt classification using various ML algorithms. The workflow can be divided into three steps that each require optimization: data generation, data extraction (Scheme 5), and data analysis. Next, we give an overview of each step in the workflow.

Data generation entails the synthesis of the dye-peptide constructs, sample preparation for imaging, and TIRF microscopy. To obtain the fluorescence of all single molecules over time, we immobilized them on a glass surface. Fixing many molecules on a glass slide allows us to measure the fluorescence continuously in a parallelized manner (Scheme 5).

Data extraction consists of an automated pipeline for determining the position of the molecules in the recorded single-molecule movies. After localization, the fluorescence traces are extracted by integrating the fluorophore area frame per frame (Scheme 5).



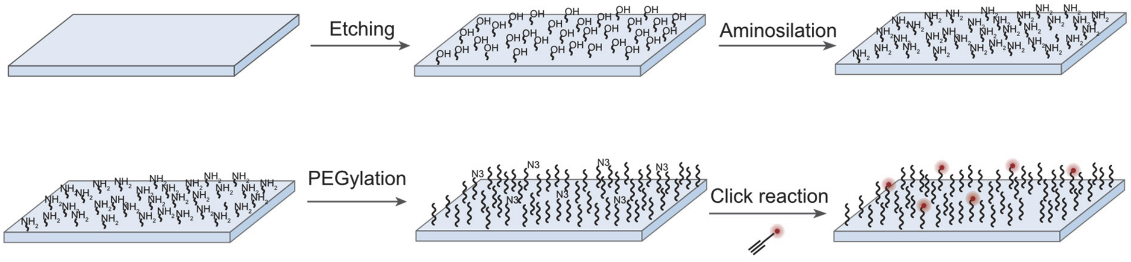
Scheme 5. Schematic overview of the steps of the workflow to obtain large amounts of single-molecule data for analysis.

The initial step for data analysis is preprocessing the extracted time traces. Preprocessing consists of filtering out traces resulting from noise or non-specifically bound molecules and scaling to correct for image inhomogeneities between different acquisitions. The normalized filtered traces are used to compute trace-specific features like the number of peaks, the time of the first peak, etc.

2.2 Surface preparation/functionalization

Spatial restriction on the surface allows us to image the fluorophore molecule during its whole fluorescence lifetime without having to account for signal movement. To covalently attach the molecules to the glass surface we adapted a surface functionalization protocol originally developed and published by the group of T. J. Ha.[150] The general principle of the protocol is based on installing long polyethylene glycol (PEG) chains capped with a terminal methoxy group (mPEG), to prevent unspecific binding ("passivate the surface"). For single molecules to be distinguishable during imaging sparsely functionalized surfaces are required. Such low-density labeling of the glass surface can be achieved with low amounts of PEG carrying a modification handle to mPEG ratios.

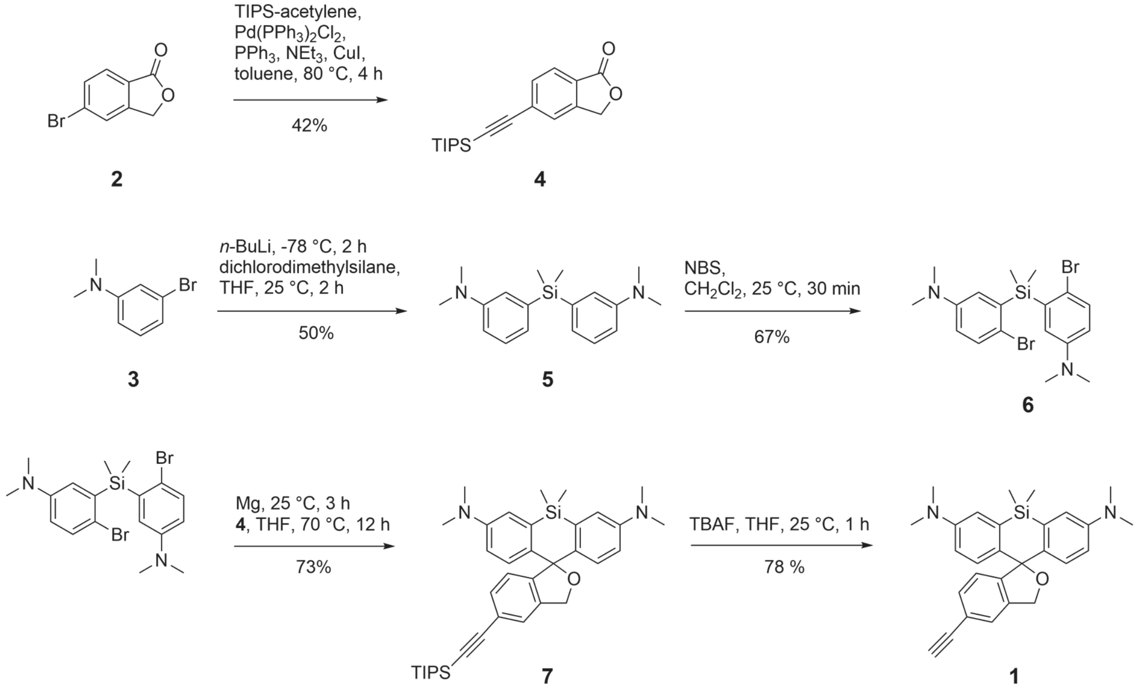
For our application, we used an azide as a reactive handle (aPEG), as this functionality allows us to attach the peptides via a copper(I)-catalyzed azide-alkyne cycloaddition (CuAAC).[151] By employing CuAAC instead of previously reported biotin-streptavidin binding, we circumvent the close presence of the protein streptavidin, which could exert a stronger influence on the intermittently blinking dye compared to the molecule it is linked to and is to be sampled. If streptavidin becomes the determining factor in the environment, it could potentially diminish the specificity of the approach. To obtain single-molecule signals of specifically bound molecules, we needed to implement the mentioned surface functionalization in our laboratory. In the following sections, we described the exploration of the surface functionalization protocol (Scheme 6) and the conditions in the crucial steps that we have tested to lower noise signals on the glass surface. To test the CuAAC reaction and successful surface passivation we employed a mock reagent alkyne-HMSiR 1.



Scheme 6. Overview of the different steps in the surface functionalization procedure. The steps include cleaning and etching of the surface followed by installation of amine groups via aminosilation. The free amines are reacted with a bifunctional amine-reactive PEG reagent. The sparse aPEG can be modified in a CuAAC reaction.

2.2.1 Synthesis of alkyne-HMSiR **1**

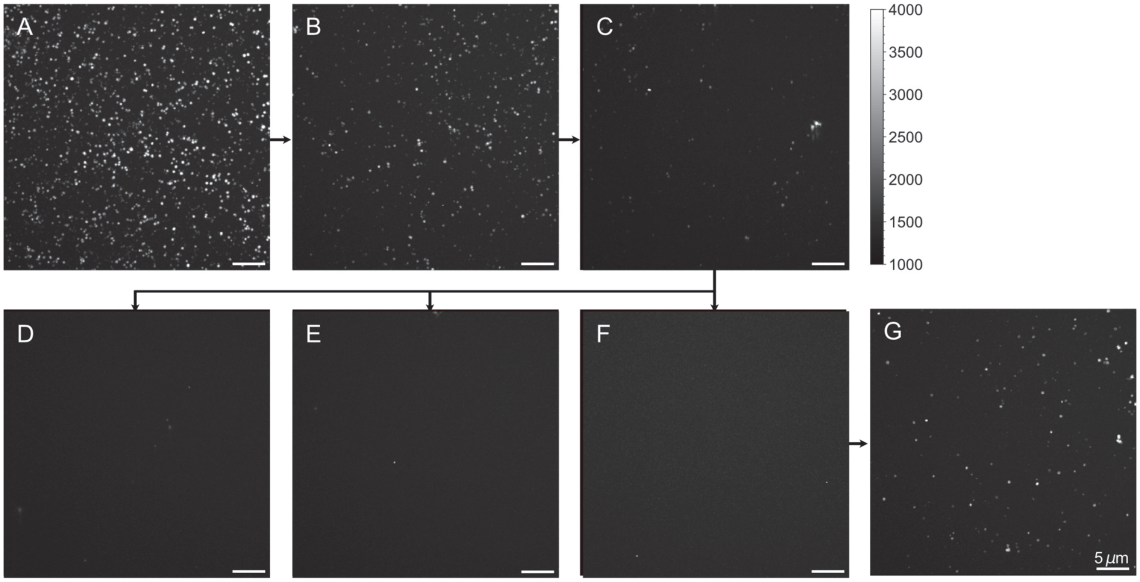
For the optimization of the surface functionalization protocol, we synthesized an alkyne-bearing spontaneously blinking dye HMSiR compound 1 (Scheme 7) from commercially available 5-bromophthalide and 3-bromo-*N*,*N*-dimethylaniline 2. The alkyne functionality was introduced via a Sonogashira reaction of 5-bromophthalide and triisopropylsilyl (TIPS)-acetylene yielding compound 3. The xanthene rings were constructed by lithiation of 2 with *tert*-butyllithium (*t*-BuLi), which then reacted with dichlorodimethylsilane to form the silicon-bridged dimer 4. Compound 4 was brominated using *N*-bromosuccinimide (NBS). The resulting brominated product 5 was used to form the Grignard reagent, to which compound 3 was added giving TIPS-protected derivative 6 in good yields. Deprotection with tetrabutylammonium fluoride (TBAF) gave the compound 1 (Scheme 7). With this compound in hand, we could test the functionalization procedure and especially the new CuAAC step.



Scheme 7. Synthesis of alkyne-HMSiR 1. THF = tetrahydrofuran, dba = dibenzylideneacetone

2.2.2 Optimization of the cleaning and etching method

The first step after the removal of the organic contamination through washing and sonication steps (Scheme 8, A-C) is the etching of the glass surface (Scheme 6). This step is crucial to remove any residual organics that could lead to background signals and create free hydroxyl groups for conjugation in the following polymerization step, i.e., the aminosilation. In literature, the most used methods are extended etching with KOH solutions,[152] (additionally) with Piranha solutions,[153] air/oxygen plasma,[154] and ultraviolet (UV)/ozone.[155,156] To determine which method resulted in the cleanest surface, we prepared a batch of glass coverslips that were treated including the last solvent sonication step (Scheme 8, C). At each step, one slide was removed to check the progress of the surface cleaning. The remaining slides were etched using the following methods: air-plasma etching, UV/ozone cleaning, sonication in 1 M KOH, and sonication in 1 M KOH with additional Piranha treatment (Scheme 8).



Scheme 8. Images of glass slides at every step of the cleaning procedure. A, Commercial glass, out-of-the-box. B, Result after sonication in 10% Alconox^{TM} solution for 20 min. C, Result after sonication in acetone for 20 min. D, Result after plasma cleaning using air plasma for 5 min. E, Result after UV/ozone cleaning for 10 min with 5 min resting time. F, Result after 40 min of sonication in 1 M KOH. G, Result after KOH treatment with an additional wash in freshly prepared Piranha solution for 20 min. All images were obtained as the maximum projection of 6000 frames from slides of the same batch measured using TIRF microscopy using a 647 nm excitation laser. Scale bar = 5 μm. The color bar represents the fluorescence intensity (a.u.). Adapted from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* 2023, *145*, 1441–1447.[91] 🅭🅯

The results looked promising across all methods, except for the additional Piranha treatment, which appeared to introduce more background signals. Considering the substantial generation of basic waste in the KOH etching process and the accessibility of UV/ozone, we selected the latter cleaning method moving forward. To further ensure that we were generating an even, homogenous surface without creating crevasses due to excess cleaning, we conducted an atomic force microscopy (AFM) experiment.[157] Similar to the microscopy experiment, we imaged a representative area on a glass slide by AFM at different steps of the procedure (Figure 12).

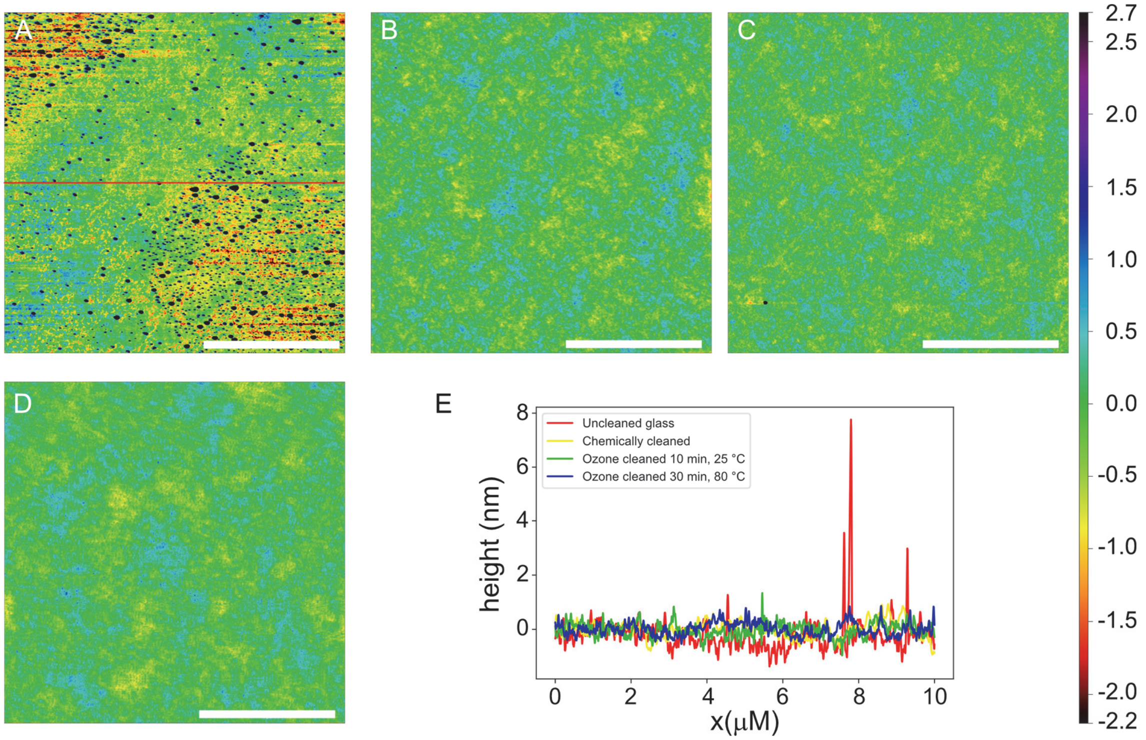


Figure 12. AFM acquisition of 10 x 10 µm areas of glass coverslips at different steps during the cleaning procedure. A, Uncleaned glass coverslip taken from the box and wiped with a KimWipe tissue. B, Glass coverslip after the wet cleaning steps, sonication in detergent (10% Alconox^{TM}), MiliQ and acetone. C, Glass coverslip after the wet cleaning steps followed by ozone cleaning at 25 °C for 10 min followed by 5 min rest in the chamber. D, Glass coverslip after the wet cleaning steps followed by ozone cleaning at 80 °C for 30 min followed by 5 min rest in the chamber. E, Extracted profiles along the horizontal midline of all images (exemplified as the red line in A). The color bar represents the distance in the z-axis (nm). Scale bar = 4 µm.

The results showed that the cleaning removes organic contamination and creates a homogenously flat surface with variability in height < 2 nm even after extensive UV/ozone treatment. Based on these combined results, we decided to use UV/ozone treatment as the etching step for our microscopy measurements.

2.2.3 Optimization of the aminosilation and PEGylation

2.2.3.1 Aminosilation

The next step after etching the surface is the installation of the free amino groups on the surface. Reacting the newly formed hydroxyl groups on the surface with the amine-functionalized silyl ether *N*-(2-aminoethyl)-3-aminopropyl-trimethoxysilane (AEAPTMS) (Figure 13, A), which at the same time polymerizes, should ideally lead to a polymeric layer of amino silane covalently bound to the glass surface. The silyl ether group can react with the surface, hydrolyze, or polymerize with the hydrolyzed product 7 forming siloxane bonds, in any order and available reaction partner (Figure 13, C).[154,158] Furthermore, the polymerization reaction is catalyzed by the water present in the reaction and the intermolecular interaction of the secondary amine in AEAPTMS (Figure 13, B).[154,158]

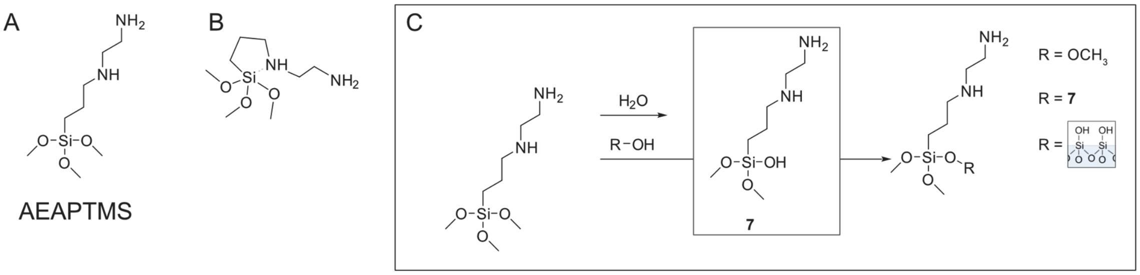


Figure 13. Reactions involved in the aminosilation step. A, The structure of the bifunctional reagent AEAPTMS. It contains an amine group, which is necessary for later steps, and methyl silyl ether groups. B, The intramolecular interaction catalyzing the hydrolysis or polymerization of AEAPTMS.[154] C, AEAPTMS can either undergo hydrolysis with a water molecule, polymerization with another AEAPTMS molecule (or its hydrolysis product), or anchoring to the surface through a reaction with a surface silanol.

The speed of polymerization in comparison to the reaction with the surface silanols will determine the structure of the formed layer. Hence the quality of the surface functionalization will depend on the water content of the solvent (environment humidity) and the quality of the AEAPTMS reagent. To investigate whether a different solvent might result in a better surface quality, we compared the proposed aminosilation conditions 5% AEAPTMS in methanol + 5% acetic acid with another published condition using 2% AEAPTMS in acetone (HPLC grade) (Figure 14).[150,154,158,159] The two conditions looked qualitatively similar. However, the aminosilation using acetone seemed less crowded with more homogeneous signals. This result was not fully conclusive, yet we decided to use acetone with 2% AEAPTMS in future experiments, taking into consideration the number of reagents. The latter condition uses fewer components, which could potentially result in less variability between different aminosilation results if the reagent quality remains consistent.

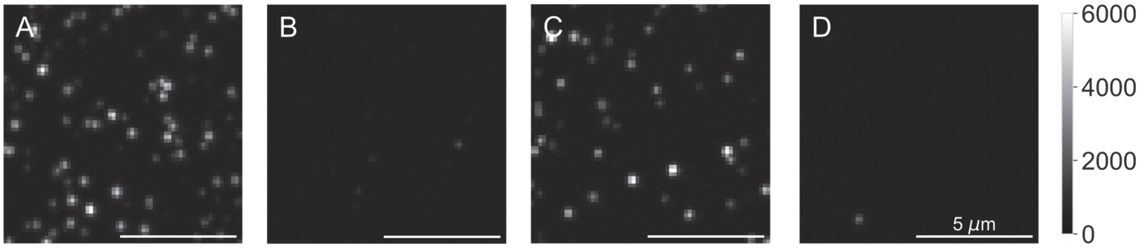


Figure 14. Comparison of two different aminosilation conditions in the standard slide functionalization using alkyne-HMSiR 1 in the click reaction. A, B, methanol + 5% acetic acid, freshly prepared, was used as a solvent with 5% AEAPTMS. C, D, acetone was used as a solvent with 2% AEAPTMS. A, C, are the positive control cover slips that were modified with aPEG and mPEG chains. B, D, are the negative control coverslips that were modified with mPEG only. The color bar represents the fluorescence intensity (a.u.).

2.2.3.2 PEGylation and CuAAC

In our initial test experiments, we observed considerable fluorescent contamination on control coverslips that are missing the azide group and cannot undergo a click reaction with any fluorophore. Therefore, we needed to optimize the protocol to prevent such unspecific adsorption of the fluorescent impurities onto the surface.

We prepared two negative controls for the click reaction: first, the slides were functionalized with a mixture of $aPEG_{5000}$-*N*-hydroxysuccinimide (NHS) diluted in $mPEG_{2000}$-NHS, but in the click buffer, no copper was added (Figure 15, E, and J). Second, the slide was functionalized only with the $mPEG_{2000}$-NHS but the complete click buffer with alkyne 1, copper, and ascorbate was added (Figure 15, C and I). The comparison of these two negative control slides revealed that copper promotes significant unspecific adsorption. In fact, with the azide moiety but no copper, the fluorescent background is lower (Figure 15, E and J). Therefore, we introduced a washing step with 10 mM ethylenediaminetetraacetic acid (EDTA) at pH 8.5 after the click reaction (Figure 15) to remove the copper from the surface (Figure 15, B, G, D, and I).[160] In addition, we also imaged the same conditions using $mPEG_{5000}$-NHS to test if the expected change in PEG densities on the surfaces (shorter PEG chains have been reported to pack more densely) would show a clear effect on the unspecific adsorption (Figure 15, F to J).[161] The additional washing step noticeably decreased the fluorescent background in all negative controls independent of the mPEG-NHS chain length employed in the preparation (Figure 15 C, H).

HMSiR	+	+	+	+	-
Cu	+	+	+	+	-
N3	+	+	-	-	+
EDTA	-	+	-	+	-

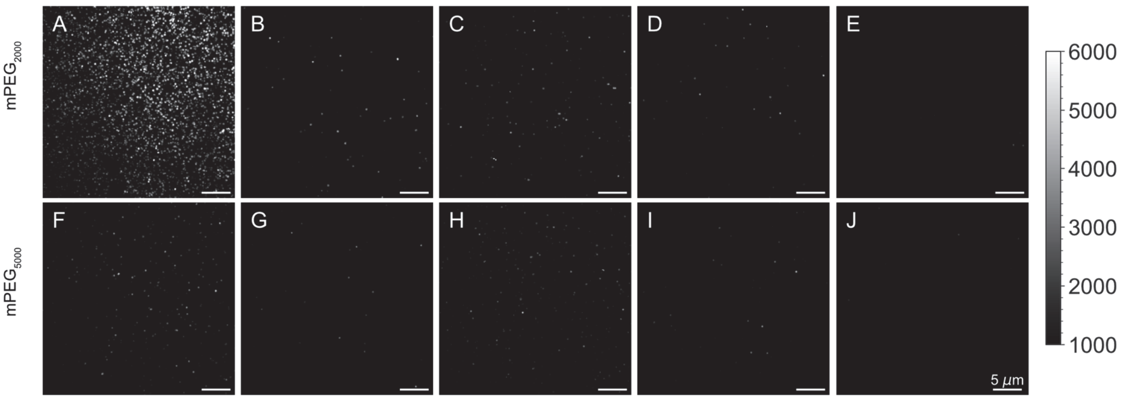


Figure 15. Maximum projection of movies of 6000 frames length imaged under TIRF conditions in surface protocol optimization experiments. In all the experiments, except for the negative controls E and J, copper was used in the click buffer. A-E, Short mPEG$_{2000}$-NHS was used for PEGylation and dilution of aPEG$_{5000}$-NHS. F-J, Long mPEG$_{5000}$-NHS was used for PEGylation and dilution of aPEG$_{5000}$-NHS. A, F, Positive controls: Slides with a- and mPEG-NHS. B, G, slides from A and F after washing with 10 mM EDTA solution for at least 10 min. C, H, Negative controls: Only mPEG-NHS but with copper added in the click buffer. D, R, slides from C, and H after washing with 10 mM EDTA solution for at least 10 min. E and J, Negative controls: Slides with a- and mPEG-NHS but no copper and no alkyne-HMSiR 1 added in the click buffer. Scale bar = 5 µm. The color bar represents the fluorescence intensity (a.u.).

The PEGylation quality seemed to be better in the case of long mPEG$_{5000}$-NHS, resulting in a lower density of aPEG$_{5000}$-NHS and generally lower signal in the negative controls. To further improve the functionalization, we tested the protocol with an additional PEGylation step using a very short mPEG$_4$-NHS reagent after the first round of PEGylation (Figure 16).[162] We again imaged positive controls with the aPEG$_{5000}$-NHS for both mPEG$_{2000}$-NHS (Figure 16, A) and mPEG$_{5000}$-NHS (Figure 16, E) and the negative controls in the absence of azide (Figure 16, C and G). We also imaged the slides after washing with 10 mM EDTA (Figure 16, C, F, D and H). This additional step resulted in a modest decrease in background signal on the negative control slides compared to the addition of the EDTA washing step (Figure 16). Nonetheless, the best conditions seemed to be obtained by adding both new steps when comparing the two negative controls with copper (Figure 16, D and I) and (Figure 16, D and H).

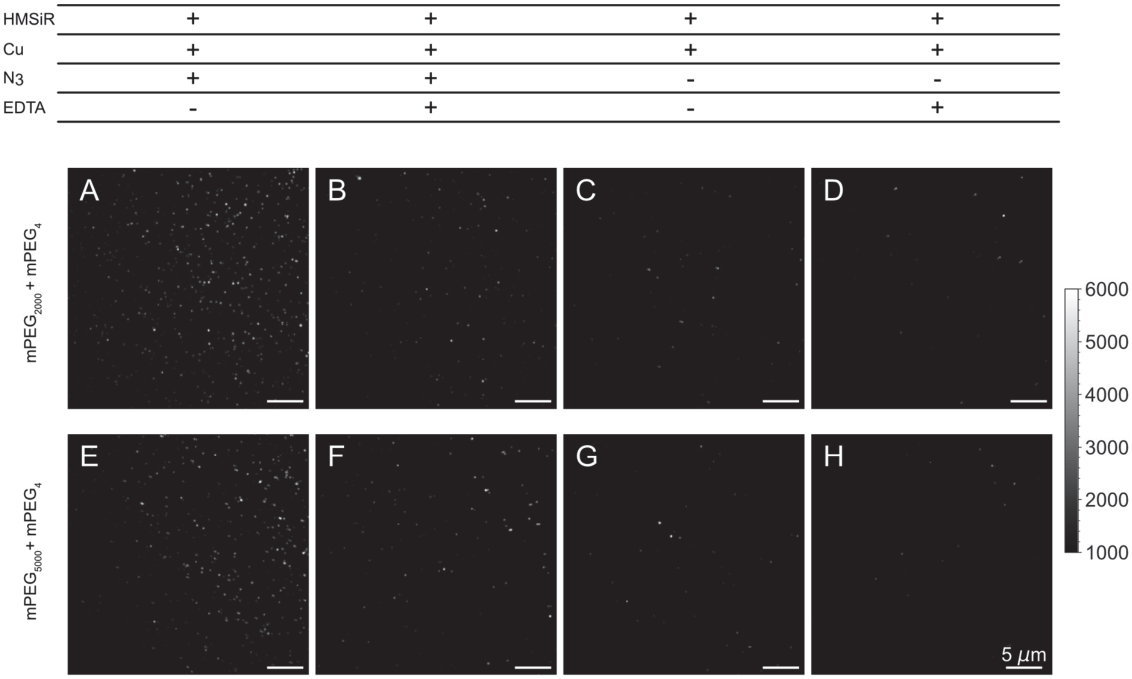


Figure 16. Maximum projection of movies of 6000 frames length imaged under TIRF conditions in surface protocol optimization experiments with an additional PEGylation step. A-D, $mPEG_{2000}$-NHS and $aPEG_{5000}$-NHS were used for PEGylation. E-H, $mPEG_{5000}$-NHS and $aPEG_{5000}$-NHS were used for PEGylation. A, E, Positive controls: slides with a- and mPEG-NHS. B, F, slides from A and E after washing with 10 mM EDTA solution for at least 10 min. C, G, Negative controls: Only mPEG-NHS. D, H, slides from C and G after washing with 10 mM EDTA solution for at least 10 min. Scale bar = 5 µm. The color bar represents the fluorescence intensity (a.u.).

These results encouraged us to test further passivating reagents and washes to enhance surface quality. Yet, the overall observation suggested that incorporating more supplementary steps introduced additional sources of contamination, resulting in comparable or even inferior outcomes. Attempts to increase the PEGylation density by adding a second PEGylation step with $mPEG_{2000}$-NHS before the final $mPEG_4$-NHS PEGylation led to more background signals (Appendix: Figure 66). Similarly, incorporating a passivation step with disuccinimidyl tartrate (DST) after the second passivation ($mPEG_4$-NHS PEGylation) showed no improvement (Appendix: Figure 67).

The effective decrease in background, likely resulting from removing surface copper ions with EDTA, inspired experiments with a copper ligand in the click reaction. A ligand might help prevent the initial adsorption of copper ions onto the surface and facilitate their removal during the subsequent washes. Experiments using the water-soluble ligand tris((1-hydroxy-propyl-1*H*-1,2,3-triazol-4-yl)methyl)amine (THPTA) in the CuAAC showed promising results, with good signal in the positive control and low signal in the negative control with surface missing aPEGs (Figure 17, A and D respectively). Trying to optimize the CuAAC step, we experimented with longer reaction times resulting in a notable increase in unspecific signals on the negative control slide without an increase on the coverslips containing aPEGs (Figure 17, B and E). Similar results of unspecific binding were observed with higher concentrations of compound 1 (Figure 17, C and F).

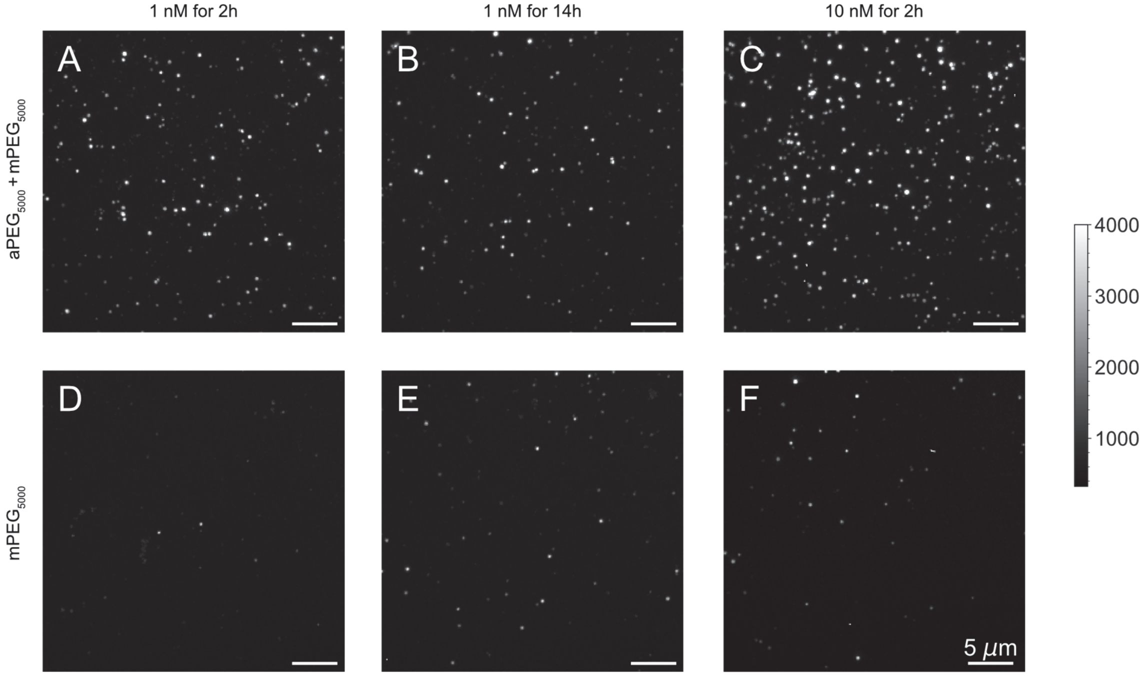


Figure 17. Maximum projection of movies of 3000 frames length imaged under TIRF conditions in surface protocol optimization experiments with an additional PEGylation step, $mPEG_{5000}$-NHS and $aPEG_{5000}$-NHS were used for PEGylation. CuAAC reaction mixture containing 10 mM THPTA, 2 mM $CuSO_4$ was used for all coverslips. A-C, $mPEG_{2000}$-NHS, and $aPEG_{5000}$-NHS. D-F, $mPEG_{5000}$-NHS was used for PEGylation. A, D, The CuAAC reaction was conducted using 1 nM alkyne-HMSiR 1 for 2 h. B, E, The CuAAC reaction was conducted using 1 nM alkyne-HMSiR 1 for 14 h C, F, The CuAAC reaction was conducted using 10 nM alkyne-HMSiR 1 for 2 h. Scale bar = 5 µm. The color bar represents the fluorescence intensity (a.u.).

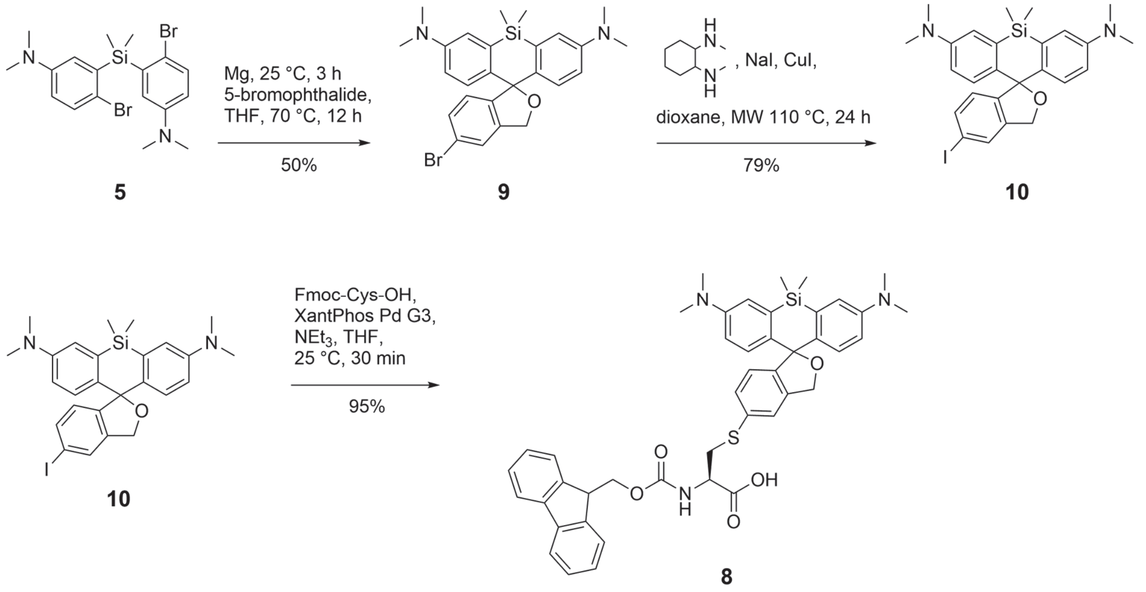
Based on these experiments, we decided to include the ligand THPTA in the standard conditions. These newly defined conditions were used to prepare the slide for acquiring the movie datasets. The data produced from these acquisitions are then used to test our hypothesis, trying to classify the imaged molecules. Given our prior experience that surface functionalization will unavoidably vary between batches, data for any training dataset must be collected across multiple days and experiments.

2.3 Design and synthesis of the test peptide sets

With a suitable surface modification protocol, we focused on the synthesis of peptides modified with the spontaneously blinking fluorophore HMSiR.

2.3.1 Synthetic design

We set out to synthesize a small test set of pentameric peptides that contained a cysteine modified with the blinking dye HMSiR. To increase possible interactions of the fluorophore with all the amino acids in the peptide, we chose to initially place the residue carrying the HMSiR moiety at the central position in the peptide. Furthermore, we aimed for a conjugation strategy that does not introduce a large spatial distance of the dye to the test peptide. To guarantee full control over the attachment of the dye HMSiR moiety in our proof-of-concept experiments, we conjugated the dye to the thiol group of a 9-fluorenylmethoxycarbonyl (Fmoc)-protected cysteine by palladium-catalyzed arylation (Scheme 9). The preparation of such a prelabeled building block would allow for compound **8** to be used like any other Fmoc-amino acid building block in solid-phase peptide synthesis (SPPS) following a standard Fmoc protocol.[163] The synthesis of the SPPS building block **8** started from compound **5**. Compound **5** was used to form the Grignard reagent with good conversion but after adding **2** only modest yields of **9** were obtained.[164] The resulting bromo-HMSiR **9** was transformed into the corresponding iodinated compound **10** in a Finkelstein reaction. The arylation of Fmoc-cysteine-OH with **10** proceeded to give the desired product **8** in good yields.



Scheme 9. Synthesis of the cysteine building block **8** modified with HMSiR for SPPS. MW = microwave, XantPhos = 4,5-bis(diphenylphosphino)-9,9-dimethylxanthene. Adapted from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* 2023, *145*, 1441–1447.[91]

With this modified amino acid 9 in hand, any small peptide can be synthesized with full control over the site cysteine labeling. Therefore, we needed to define the first peptides to synthesize as a proof-of-concept of our proposed method. The peptides for the proof-of-concept experiments should be small and easy to handle. Moreover, they should allow insight into the sensitivity of blinkognition toward different chemical influences that might naturally occur for peptides.

2.3.2 Peptide sets: charged peptides

For the first set of peptides (C11, C12, C13, C14) we opted for very small peptides composed of the two amino acids aspartic acid and leucine in addition to the modified cysteine (Figure 18). These two amino acids were chosen because they share the same bond skeleton and therefore are sterically similar. Their polarity, on the other hand, is different because aspartate contains a negatively charged carboxylic acid functional group. The exchange of leucine for an aspartate corresponds to a shift toward a more polar microenvironment and should lead to a modulation of the fluorescence blinking. In addition, pentynoic acid was coupled to the last amino acid on the N-terminus (Figure 18). This modification allowed us to immobilize the peptide on the glass surface.

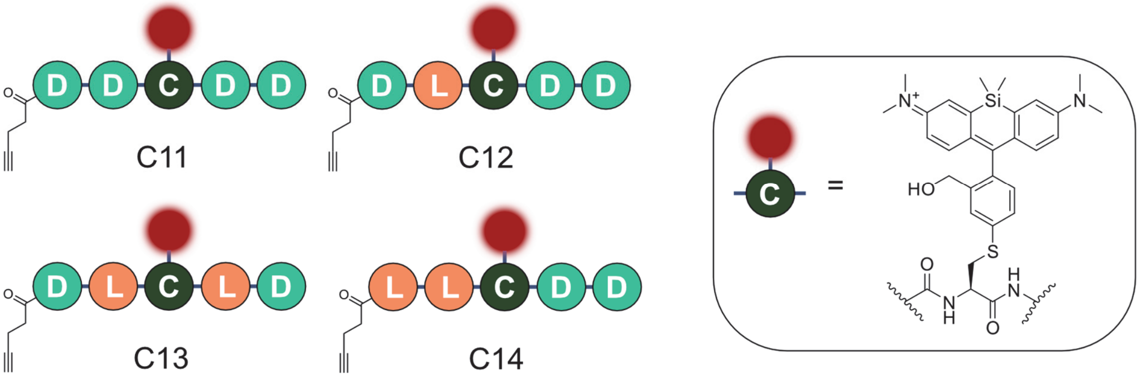


Figure 18. Set of small, charged peptides C11-C14 containing the HMSiR-modified cysteines and peptides of different polarities and polarity distributions.

These unnatural peptides are primarily interesting from a chemical aspect, rather than a biological standpoint, as they will show that changes in charge and polarity can be sensed using blinkognition.

Apart from differentiation between different peptide sequences, the method could potentially also be used to detect subtle changes in an amino acid side chain. In biology, such chemical changes occur in the form of PTMs on proteins and peptides. More than 200 PTMs have been reported, which can vary from a small structural rearrangement over an addition of a chemical group for instance methylation, acetylation, or glycosylation, to a proteolytic cleavage of a protein into its mature form.[165] PTMs add another layer of complexity to the proteome based on the

20 proteinogenic amino acids encoded in the DNA of an organism. Irregularities in PTMs are implicated in many pathological conditions like cancer and muscular dystrophies.[166,167] The analysis of PTMs is conventionally done with time-consuming biochemical methods like radioactive isotope labeling, and antibody analysis i.e., by western blotting, peptide or protein arrays, or assays based on PTM-specific enzymes.[168] The current state of the art for analysis of PTMs is MS. Nonetheless, MS analysis of PTMs is still challenging due to the low relative and absolute abundance of modified protein. In addition, handling and preparation of a sample for MS is complicated by increased hydrophilicity when modified with certain PTMs.[166] Moreover, many PTMs are unstable under MS conditions or prevent efficient ionization or detection.[166] Our proposed method has a single-molecule detection limit and provides conditions in which PTMs are stable. Therefore, PTM substrates are an interesting target to validate our method and further explore the sensitivity of our approach.

2.3.3 Peptide sets: phosphorylated peptides

The most frequent and best-studied PTM is the phosphorylation of serine, threonine, and tyrosine, although other phosphorylated residues (histidine, lysine, arginine, cysteine, aspartic acid, and glutamic acid) have been reported.[165,169] The introduction of a phosphate group by a kinase is a reversible modification. It often functions as a switch, which can turn on and off certain biological functions. A class of such switch molecules crucial for signal transduction are the small monomeric GTPases of the Ras superfamily. Rap1b is a member of this family and is involved in the regulation of cell adhesion and cell growth. This protein can be phosphorylated by cAMP-dependent protein kinase A.[170] Altschuler et al. determined the exact position of phosphorylation by introducing site-specific mutations to Rap1b.[170] Proteolytic digestion of Rap1b with the protease trypsin would result in cleavage of the amide bond between residues K178 and S179 resulting in the C-terminal fragment SSCQLL. The cleavage product contains an ideally situated cysteine to attach the blinking fluorophore needed for our method. With the model peptides containing a fluorophore-modified cysteine SSC(HMSiR)QLL **P15**, SpSC(HMSiR)QLL **P16**, and pSSC(HMSiR)QLL **P17** (Figure 19), we can test whether our method will be able to distinguish the phosphorylation pattern.

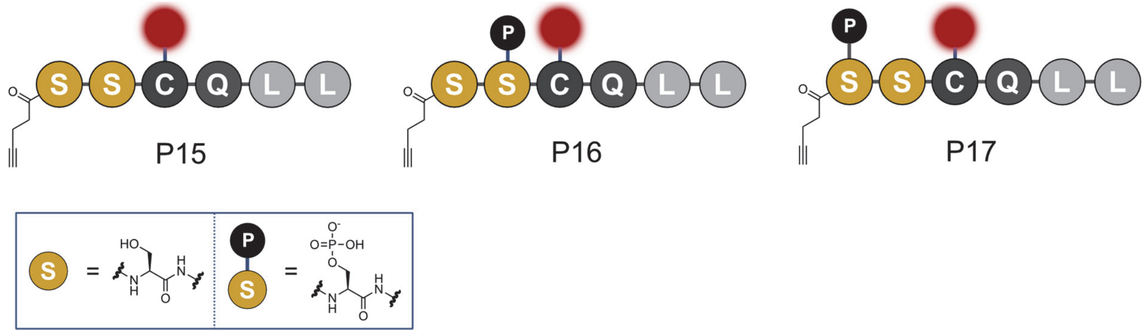
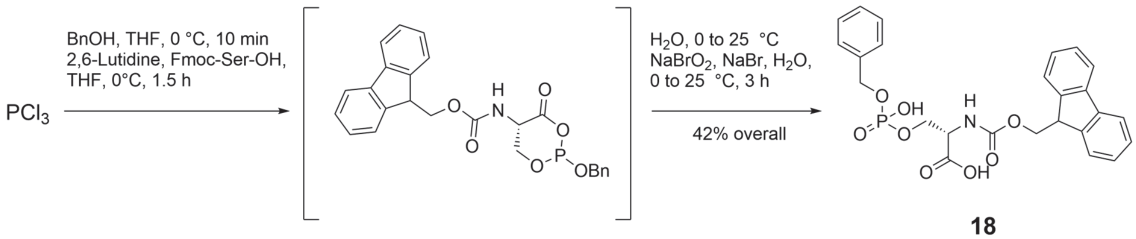


Figure 19. The model peptides envisioned for the study of phosphorylation.

To synthesize the new selection of peptides via SPPS using the same building block as for the charged small peptides, we needed a serine building block carrying a phosphate group Fmoc-Ser(PO(OBzl)OH)-OH 18 that could be incorporated during SPPS (Scheme 10).



18

Scheme 10. Synthesis of the building block of the phosphorylated serine for SPPS.

2.3.4 Peptide sets: epimerized peptides

A second PTM we were interested in studying is epimerization. In an epimer, the stereocenter of a single amino acid is inverted. In nature, amino acids with an inverted stereocenter at the α-carbon center occur in ribosomally synthesized and post-translationally modified peptides (RiPPs) and non-ribosomally produced peptides. RiPPs are a class of natural products consisting of heavily modified peptides. Their biosynthesis starts from a precursor peptide containing an *N*-terminal leader sequence for recognition by the modifying enzymes, a core peptide resulting in the modified peptide, and sometimes a C-terminal follower peptide, which can also be required for recognition.[171] After proteolytic cleavage, the mature core peptide is released and becomes active. One type of modification occurring in RiPPs is epimerization. The inverted stereochemistry is usually achieved by transforming the natural L-form of an amino acid into the D-form through racemases generating a D- and L- mixture that is resolved by downstream modifying enzymes. Alternatively, they are generated by a dehydrogenases and hydrogenases reaction sequence or a radical process performed by a single enzyme of the *S*-adenosyl-L-methionine radical superfamily (rSAM).[172–174] As a model system, we chose the peptide corresponding to the RiPP core peptide OspA (Figure 5) from *Oscillatoria sp*. PCC 6506. This model peptide has been studied previously in vitro due to its solubility properties and low product complexity. The original OspA core peptide **E19** is only epimerized at two positions by the epimerase OspD (Figure 5, Ile-4 and Val-13) as opposed to other RiPP substrates of rSAM epimerases like Polytheonamide A. Polytheonamide A is a 48 amino acids long RiPP with every third position inverted.[172] The epimerase OspD first epimerizes Ile-4, yielding an intermediate core peptide **E20** before the fully epimerized OspA core **E21** is formed.

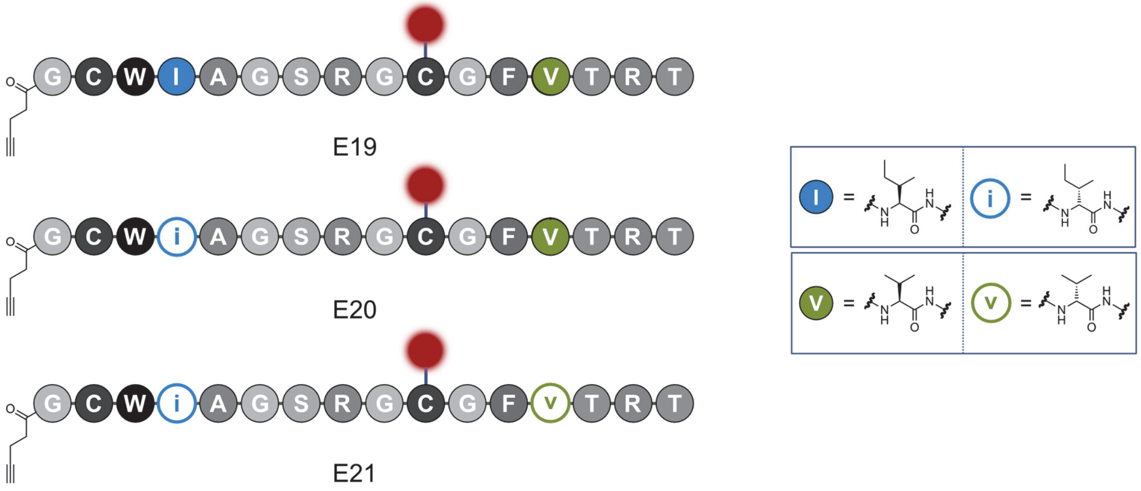

G C W I A G S R G C G F V T R T
E19
G C W i A G S R G C G F V T R T
E20
G C W i A G S R G C G F v T R T
E21
I =
i =
V =
v =

Figure 5. The RiPP core peptide OspA. The positions epimerized by OspD are shown in lowercase and are not filled.

After modification of the OspA core with HMSiR and a pentynoic acid at the N-terminus, we can use these model peptides to test whether our method can distinguish subtle chemical changes like epimerization on a single-molecule level with our proposed method. The synthesis of these peptides was achieved by manual SPPS for peptides **C11-C14** and **P15-P17**, using a standard Fmoc-protocol in good yields. Peptides **E19-E21** were synthesized on a peptide synthesizer except for the coupling of the modified C, which was done manually.

2.4 Data acquisition

With the alkyne-modified peptides in hand, we immobilized them via CuAAC, followed by TIRF microscopy. The single-molecule movies were obtained from at least five independent measurements for each molecule, performed on five different days, and using independent glass slide preparations for each day. The peptides of each set were measured on the same days. The imaging conditions were kept constant in all acquisitions: 641 nm, 110 mW with 30 ms exposure time on an area of 314x314 pixels) and were acquired on the same microscope setup.

2.5 Data extraction and preprocessing

Since the movies would not be used directly in the ML analysis, we set up a trace extraction and preprocessing pipeline. In initial attempts, we considered using a mask based on the maximum projection of the movies, however, this approach had multiple issues. We would miss molecules that were dim in comparison to other molecules and record noise bursts of the camera or impurities in the sample. Furthermore, it would be difficult to distinguish two molecules that are very closely overlapping. These issues can be circumvented by using the same approach as in the analysis of stochastic optical reconstruction microscopy data, i.e., by localizing the molecules in each frame individually (Figure 20). For the framewise localization, we used the Picasso package with the settings listed in Table 10. The package fits a Gaussian to each frame and localizes the center with a maximum likelihood approach.[175] The obtained framewise x- and y-coordinates were assigned to the same molecule location if their center was within a radius of 2 pixels. The average coordinates of all linked localizations were calculated, which were used to define a bounding box with a side length of 5 pixels.[175] Localizations with boxes that overlap more than 2 pixels were discarded as their traces cannot be separated and the signal of the neighboring particle would be visible in each fluorescence intensity trace (Figure 20, red boxes shown in B).

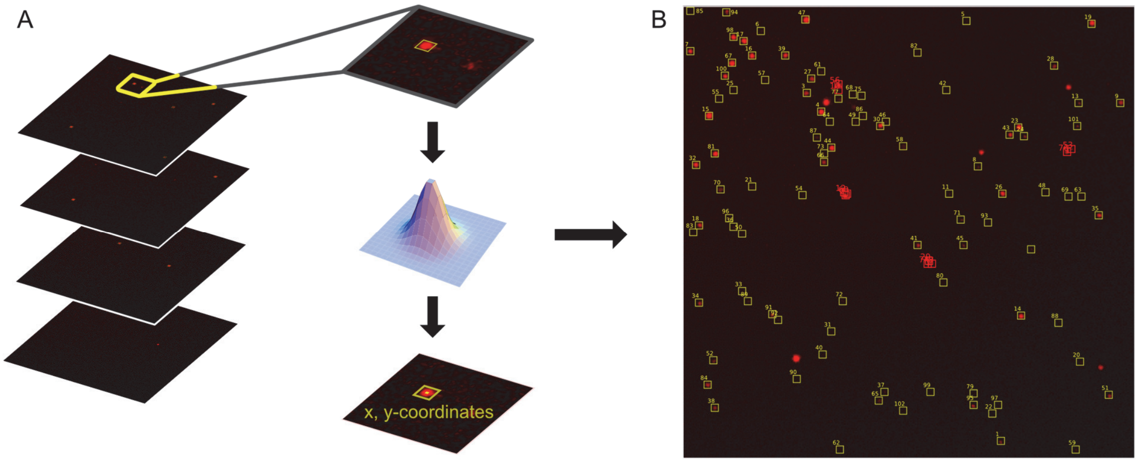


Figure 20. Localization of single molecules. A, Single molecules were localized using the localization module and linked using the postprocessing module contained in the Picasso package.[175] B, Resulting overlay of a maximum projection with the found bounding boxes. Boxes that overlapped were discarded (yellow = accepted, red = rejected). Adapted from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* 2023, *145*, 1441–1447.[91] 🅭🅯

The pixel intensities within the box were summed to give the particle intensity for each frame, with all frames in the movie resulting in a single-molecule trace per averaged x-, y-coordinate (Figure 21, A). To remove the background signal, which might vary between different experiments, position on the coverslip, or the area in the field of view, the average signal of the last 500 frames was calculated. The movies were acquired longer than the blinking of the molecules, therefore the molecules are bleached during the last 500 frames, and we can assume they should contain mostly background noise. The average of this background was subtracted from the whole signal trace (Figure 21, B). To normalize the trace across different experiments, the signal was scaled relative to the standard deviation of the signal (z-scoring) (Figure 21, C).

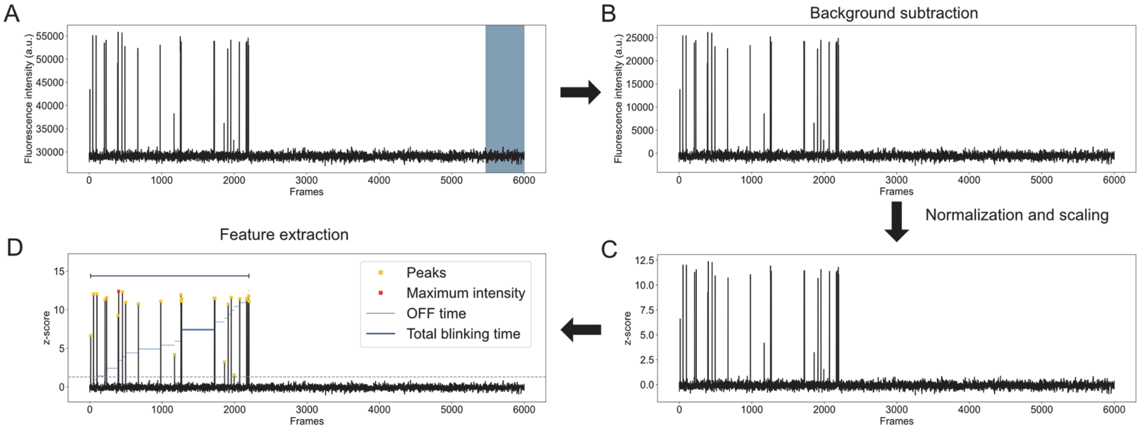


Figure 21. Signal processing after obtaining the trace from the single-molecule localization. A, Raw trace obtained from summing the intensity in the movie within a box of 5x5 pixels around the localization. The area of the last 500 frames is shaded in dark blue, which is used to determine the average background. B, The trace obtained after subtraction of the mean calculated from the last 500 frames (shaded in A). C, The trace obtained after z-score normalization of the background subtracted trace (B). D, Exemplary presentation of the trace properties, that were measured to calculate the features for classification. Adapted from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* 2023, *145*, 1441–1447.[91] 🅭🅯

We need to extract descriptive features from the traces to evaluate whether the fluorescence time traces that we had extracted from the single-molecule movies are characteristic of the corresponding peptides. With numerical features derived from the traces, we can investigate differences between the signals of different molecules.

We used the Python SciPy signal module to annotate each trace with features.[176] Peaks are algorithmically detected by choosing a minimum height of a peak as well as a minimum width of a peak (Figure 21, D yellow crosses show detected peaks). We set the minimum width to one frame and the minimum peak height to eight times the standard deviation of the corresponding trace. Hence, a peak needs to deviate strongly from the average baseline compared to any noise. Based on these detected peaks we filtered out traces that likely originate from noise, such as single bursts, or traces that are extremely short and would contain little information. The conditions and thresholds that we used for filtering the traces can be found in Table 11. For the remaining traces, we extracted numerical features derived from the detected peaks, like the number of peaks, the first and last frame a peak occurred, and the total blinking time, etc. (Table 12). We also calculated the maximum, minimum, mean, and standard deviation of the times between peaks, peak height, peak width, and peak area. Furthermore, we used another package, librosa, which provides a function to estimate beats per minute in audio. We hoped that this feature could capture some time-related information.[177] To get a first idea of how distinctive these features would be we plotted some cross-correlation plots of the extracted features for peptide **C11-C14** (Figure 22).

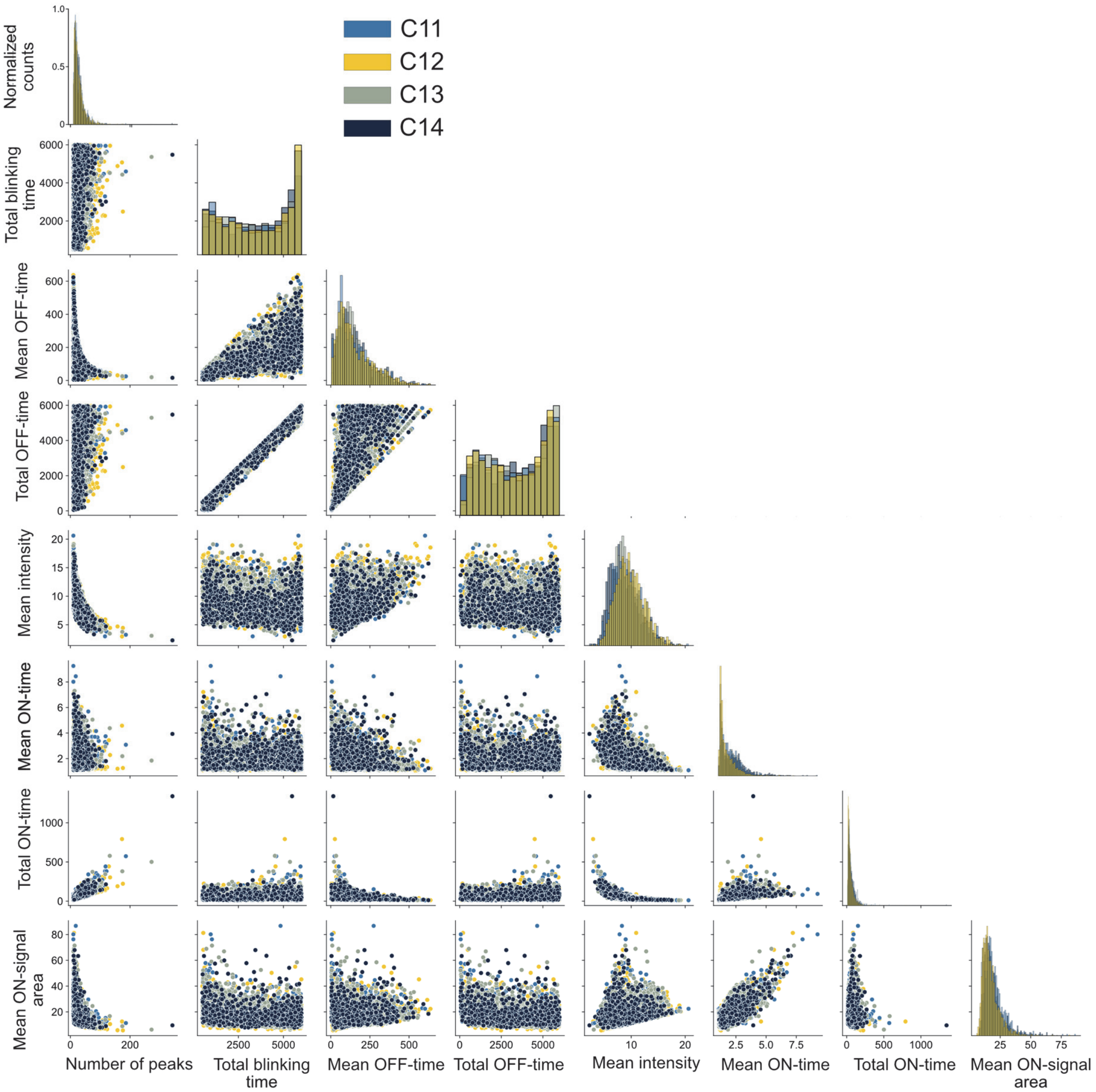


Figure 22. Initial visualization of the extracted features. Examples of cross-correlation plots and histograms (diagonal). Adapted from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* **2023**, *145*, 1441–1447.[91] 🅭🅯

This first explorative analysis did not reveal any peptide-specific clusters, nor the features by themselves (Figure 22, histograms on the diagonal) nor combined with a second feature (Figure 22, off-axis).

2.6 Data analysis of extracted features

Based on the histograms and correlation plots we were not able to distinguish any of the peptide classes. However, we hypothesized that an ML-based approach could be a viable solution for the analysis of fluorescent blinking patterns. ML algorithms are used to find hidden patterns in large, complex datasets and build a predictive model for the problem at hand (a classification or a regression). Model fitting, i.e., building a predictive model, is what makes ML extremely valuable. In this step, the algorithm tries to optimize a function f(x) = y using input data x to predict an output value y.[178] It improves the function by learning from experience obtained from the training data. In other words, the algorithm that solved the given problem does not need to be programmed specifically by the user but is the result of the ML algorithm itself based on the training data. The output value y can be a continuous numeric value, in which case the problem is a regression, or it can be a categorical label in the case of a classification problem.[178] The problem in the present project falls into the category of classification, where x corresponds to the recorded time traces and the output value y should be the corresponding peptide label. For the algorithm to learn, it requires a set of training data from which it can gain experience and optimize its accuracy. In the case of supervised algorithms, the training data are provided as a labeled set, i.e., tuples of (x_1, y_1), …, (x_n, y_n), in which case the algorithm tries to improve its accuracy of the predicted y values. However, there are also unsupervised algorithms that only require (x_1, …, x_n) values without labels. These algorithms then try to find patterns and classes by themselves, a prominent example is the principal component analysis (PCA).[178,179]

2.6.1 Unsupervised ML - PCA

In our initial data exploration, we only consider a linear combination of two features, however, the extracted data is of higher dimensionality. One approach to visualize differences in high-dimensional data can be the use of dimensionality reduction methods like PCA. The idea is to condense the relevant information of a multidimensional dataset into fewer dimensions and hence make them more visually accessible. PCA linearly transforms the data into a new coordinate system, also called feature space, that best represents the variation in the multidimensional data. Consequently, we would hope to see clusters of traces corresponding to our molecules in this reduced feature space (Figure 23).

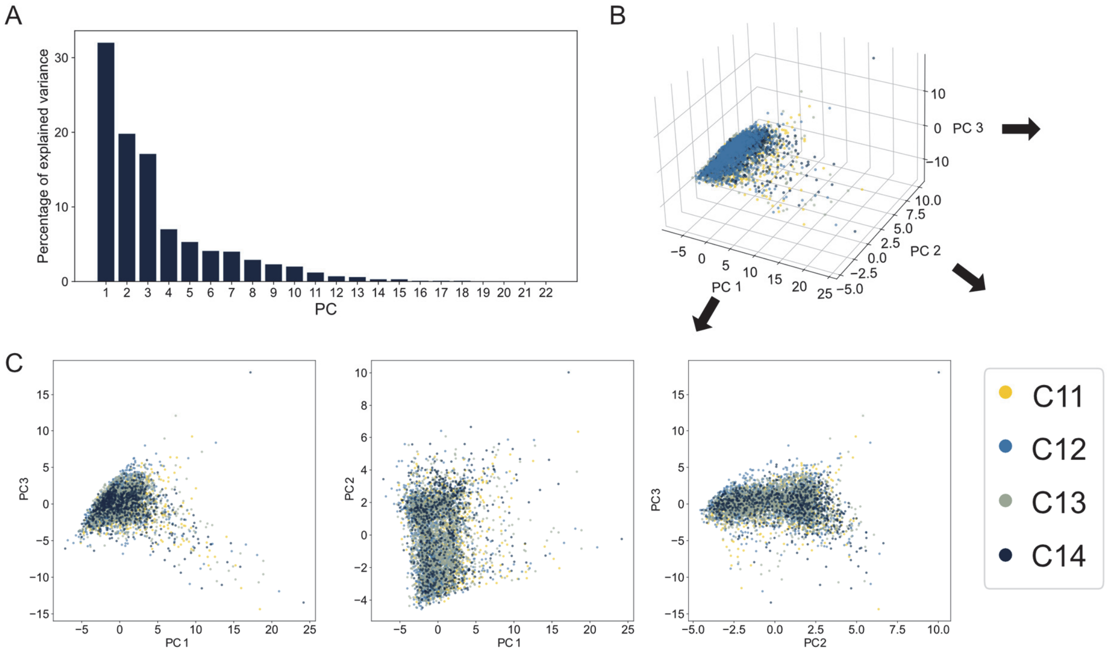


Figure 23. The outcome of the analysis using the PCA approach for peptides C11-C14. A, Scree plot resulting from the PCA shows that most variability is explained by the first 3 PCs, however, the explained variance is very low. B, Results from a PCA using 3 principal components (PCs) as a 3D view. The arrow indicates the corresponding projections of the 3 PCs (C). C, Projections of the 3D PCA (B). Reprinted in modified form from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* **2023**, *145*, 1441–1447.[91]

The result of our PCA was inconclusive without revealing any clustering or shifting populations in the new feature space (Figure 23, B and C), with a low explained variance by the principal components (Figure 23, A). One of the weaknesses of PCA is that it is limited to linear combinations of the input variables, which might not be the case for the data at hand.[180] There are nonlinear approaches to dimensionality reduction, like Sammon mapping and t-distributed stochastic neighbor embedding to name a few.[180] However, we decided not to investigate dimensionality reduction methods but instead explore supervised ML algorithms for classification.

2.6.2 Supervised ML

The features we extracted for the PCA can be directly used to train classical and ensemble supervised ML models since we know the label for all data. However, some classifiers are sensitive to feature scales. Therefore, it is important to normalize the trace features. We used the Scikit-learn inbuilt scaling function 'Robust Scaler'.[181] The scaled features were split into a training and a test set (20%). We trained a list of classifiers available in the Scikit-learn package to test their performance on our data.[181] We tested classic ML classifiers ('k-Nearest Neighbors (KNN)', 'Support Vector Machine (SVM)', 'Decision Tree'), ensemble classifiers ('Random Forest', 'AdaBoost'), a small neural network ('Multilayer Perceptron' (MLP)), and a stack classifier based on all of them (Figure 24, A).

In the following, we present the performance of an algorithm in the form of a confusion matrix. It is a presentation of all predictions made by the classifier in a matrix m_{rc}. The row r corresponds to the true label and the column c to the predicted label. Hence, the sum of a row corresponds to all data points of a class and the sum of a column to the number of times the class was predicted, normalized respectively.[181] A position (r,c) describes the case that class c was predicted for an instance of class r, hence the diagonal corresponds to all the correct cases. To ensure robustness, the fluorescence intensity traces are split into a train/test set (85:15). The test set is kept aside while a model is trained using the training/validation set in a five-fold cross-validation (CV) approach. Each model resulting from CV training is evaluated on the test set.

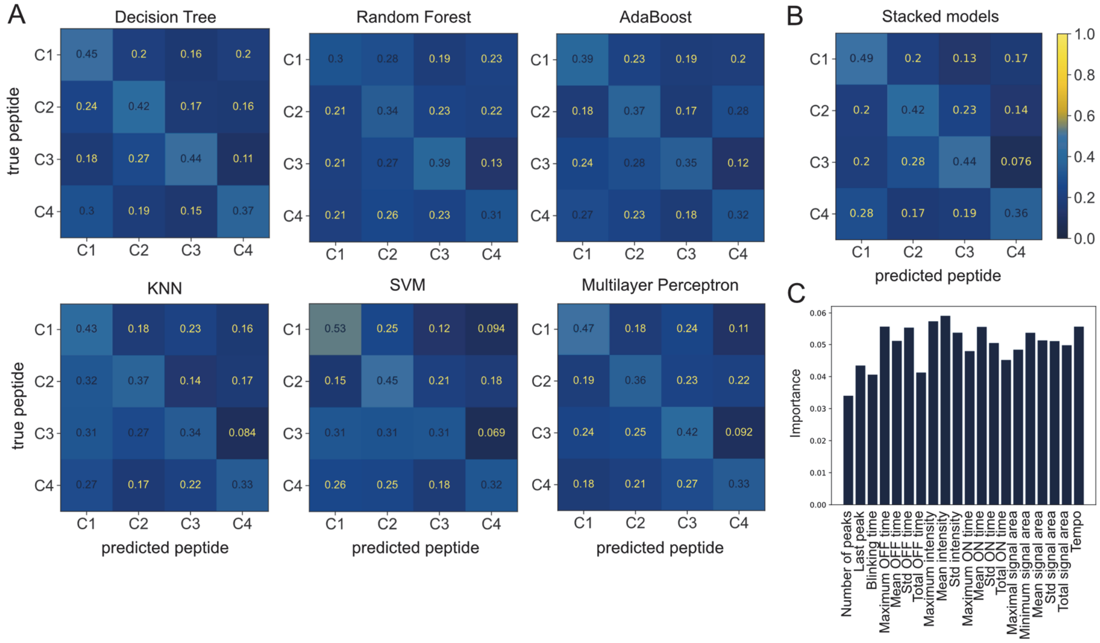


Figure 24. The mean classification results of a CV approach with five folds using the corresponding model on the extracted feature data of peptides **C11-C14** from 479 peptide traces in the test set. A, Confusion matrices obtained from predictions on the test set using the corresponding model. B, Feature importance obtained for the random forest model, most features have rather similar importance. C, Confusion matrix obtained from the model stack built from the models in A. Reprinted in modified form from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* **2023**, *145*, 1441–1447.[91] 🅭🅯

The models trained on features extracted from traces could only identify molecules with moderate accuracies ranging from 35% to 45% (Figure 24). While these accuracies are not sufficiently high for practical use, they do illustrate that interpretable ML models can extract certain sequence-related information from blinking patterns. This implies that blinkognition holds promise as a strategy for single-molecule peptide fingerprinting.

2.7 Conclusion and outlook

In this chapter, we implemented a surface functionalization protocol and optimized the steps to lower some of the unspecific background signals. Furthermore, we designed three peptide sets as well as a synthesis strategy using a prelabeled cysteine building block that can be applied in the standard Fmoc protocol. We immobilized the peptides on the passivated glass surfaces through a click reaction and imaged them using TIRF microscopy in multiple sessions for each peptide set. We then implemented a data processing pipeline using an SMLM Python package for localization and custom code for trace extraction and preprocessing.

However, visualization methods, e.g., correlation plots, and PCA analysis did not reveal obvious clustering for the different peptides C11-C14. Therefore, we moved to supervised classical ML models using the same features. The tested algorithms showed very modest accuracies that were 10-20% higher than a random guess. This result demonstrated that classical ML approaches could detect some differences in the traces modified by the microenvironment determined by the peptide and hence proved that the basic hypothesis has validity. Nevertheless, the results obtained so far do not suffice for practical application but encouraged us to investigate further analysis approaches. The primary difficulty when using classical ML algorithms is the design of appropriate features, as they need to capture all the relevant information contained in the data. If the crucial information is not preserved in these numerical values, any model is doomed to fail.

An option to automate the feature engineering process is the application of certain artificial neural networks. In convolutional neural networks (CNNs), the convolutional layer will learn to extract the relevant features during model training.[182] Furthermore, it would be interesting to implement and train a CNN because they profit from data that are ordered, these approaches are presented in Chapter 3.[179,183]

Chapter 3 A deep look at blinkognition

Parts of this chapter were published in the *Journal of the American Chemical Society*.[91] We thank Dr. Roman Briskine from the informatic service S3IT at the University of Zürich (UZH) for technical assistance with the implementation of Monte Carlo dropout.

3.1 Project goal

In Chapter 2 we optimized the data acquisition and implemented a data processing workflow for our proposed approach of peptide and protein analysis method blinkognition. The fluorescence intensity time traces of spontaneously blinking molecules are highly complex data. They reflect the combined effects of multiple factors on the blinking behavior of the fluorophore, such as polarity, pH, hydrogen bonding, steric and hydrophobic interactions, etc. To describe these blinking patterns quantitatively, it would be necessary to model all these processes over a long time (mins), which is currently not possible. In Chapter 2 we have explored classical and ensemble ML algorithms, like SVM, random forest classifiers, etc., and a small neural net (multilayer perceptron), using features that we have extracted from the measured single-molecule traces. However, based on the modest results we obtained from these approaches we posited that the manually engineered features did not capture the complexity of the phenomenon producing the traces. Advances in the field of ML, or more precisely in the subfield of deep learning (DL) offer an alternative to time-consuming and domain-specific feature engineering. Here, data can be directly utilized as input for the algorithm, superseding the need for separate feature extraction, as this extraction process is concurrently optimized during model training.[179,184]

In the past two decades, DL has enabled breakthroughs in important challenges in pattern recognition across many fields, to name a few examples: computer vision, natural language processing, and the prediction of protein structures based on their sequence.[185–195] Many different designs have been developed and have supported these advances, to offer a glimpse: CNNs, recurrent neural networks (RNNs), transformers, autoencoders, and generative adversarial networks.[185] DL models are organized in layers, with the first layer being the input layer, the last one constituting the output layer, and the layers in between called "hidden" layers. The more hidden layers a model has the deeper it is. In such a model, the scaled input data is transformed from one layer to the next. Each transformation creates a more abstract representation of the input data and can be non-linear. The last layer maps this abstract representation of the input to the required output format. In the case of a classification task, the output would be a class vector with a value corresponding to each class, whereas a regression problem would result in a numeric prediction. The result is then compared to the expected outcome and an error ("loss") is calculated in the form of a mathematical distance. To decrease this error and hence optimize the model, its internal transformation parameters ("weights") are adjusted. The direction and amount of change are defined by the gradient vector for each parameter. This gradient indicates how the error changes upon modification of the weights, the adjustment is chosen along the steepest gradient toward a minimum. The most used gradient function is a procedure called stochastic gradient descent. The resulting model is then evaluated on new data that was excluded from the original data, the test set. This additional evaluation allows for checking whether the model can make reasonable predictions on data it has not encountered during training.[179] Therefore, the user does not necessarily require problem-specific knowledge, and the dauntingly difficult task of manually designing features that fulfill these requirements can be circumvented by automated data-based training.[179] Nevertheless, understanding data requirements and choosing an appropriate model type is crucial.[178,179]

In this chapter, we set out to test and evaluate the performance of different DL architectures on the data set that was acquired in Chapter 2.

3.2 Data augmentation

DL models have showcased remarkable accuracies and performance across various applications, offering numerous advantages such as applicability to diverse problems and circumventing the need for feature engineering. However, these successes and benefits come with a specific requirement: data. Generally, if more data are available for model training, the model's performance will improve and overfitting can be alleviated.[179] Overfitting describes the problem of DL models in that they learn the training data by heart and therefore will perform extremely well on the training data but will not generalize well to unseen data.[196] Such behavior can be recognized when comparing the performance of the training set and the test set. When the model starts to overfit, the testing error (or loss) starts to increase while the training error keeps decreasing (Figure 25).

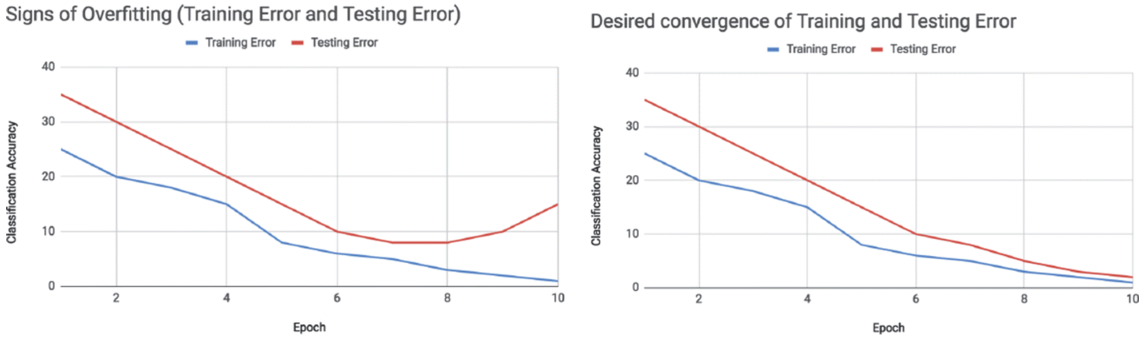


Figure 25. Example learning curves. In the case of overfitting the training error diverges from the testing error. Taken from C. Shorten, T. M. Khoshgoftaar, *Journal of Big Data* 2019, *6*, 60.[196] 🅭 🅯

The need for large amounts of data can pose a challenge in certain scenarios where either there is insufficient data or it exists in formats that are not immediately usable, lacks annotations, or is in varying formats that demand extensive data curation.[197] Consequently, a prevalent strategy involves leveraging available data by making slight modifications that preserve the content (and hence the label) but present it differently to the ML model. For instance, for an image classifier CNN, identical images could be introduced with alterations like rotations or the addition of artificial noise.[196] Such data augmentation increases model performance by expanding the amount of training data and also helps to alleviate the problem of overfitting.

For blinkognition, once the test sample has been synthesized, the amount of data depends solely on the time that is available to acquire, in principle, as many movies as desired. To maximize the data obtained from the movies we acquired in Chapter 2, we inverted the time axis of traces and kept the label. We reasoned this to be valid as the system is observed in the ground state under constant conditions, therefore processes at play should not be unidirectional. However, the fluorophores are very likely to bleach during the time of the acquisition it would not make chemical sense for a spontaneously blinking fluorophore to be in a non-fluorescent state, i.e., bleached in the beginning and then to suddenly start blinking. Therefore, we decided to only invert the time axis on the part of the sequence that shows blinking activity. This allows us to duplicate the number of available traces while maintaining the physically expected shape of a fluorescence intensity trace.

3.3 Model optimization

After augmenting our data by inverting the trace area that shows active blinking, we sought to build a neural network architecture that could process the experimental data. A widely and successfully applied architecture for array-structured data is CNN.

3.3.1 CNN model

CNNs have been developed and applied to many applications since the 1990s. However, significant uptake only happened in 2012 with the publication of AlexNet for image classification in combination with the advent of more powerful graphical processing units (GPUs). These GPUs enabled highly parallelized computation, facilitating training on extensive datasets.[198,199]

Standard CNNs are built mainly from an input, multiple stacks of convolutional, pooling, and activation layers, followed by fully connected (dense) layers and an output layer (Figure 26).[179] As implied by the name, the crucial layer in these networks is based on the mathematical operation of convolution. The layer applies a set of learnable parameters called filters (or kernels), which are smaller than the input data, in a sliding manner to the input. The calculated dot product of the kernel results in the new “pixel” (Figure 26, A) in the output (activation map). Hence the output will be smaller than the input data. The extent of size reduction depends on the kernel size and the offset applied during the sliding (stride). Therefore, the convolutional layers help to down-sample the input data while extracting relevant local information (depending on the kernel size). This size reduction can be extended when using a pooling layer that aggregates multiple neighboring pixels into one with a chosen summary function, like a local maximum or an average. In contrast to the linearity of a convolution operation, the activation layer introduces non-linearity by mapping the direct output value to a value between 0 and 1 according to a non-linear function, in most cases. Commonly used non-linearities are the rectified linear unit (ReLU; f(x) = max(0,x)), the logistic (sigmoid) function, and the hyperbolic tangent (tanh). After the last convolutional layer block, the resulting representations are passed through dense fully connected layers, where the abstract representation is mapped to the classification output.[198,200]

CNNs are networks designed to process data in the form of arrays. Originally it was inspired by the mammalian visual cortex and hence was first applied to 2D data like the pixels in an image. In a CNN, as opposed to a fully connected network, nodes are only connected to a small region of the previous layer and the input data.[201] Although originally developed in the context of 2D image data, analogous processing layers have been developed for 1D array data. In this implementation, the matrix calculations are replaced with array operations.[198] 1D CNN layers have been successfully applied to complex signal data, prominently in the analysis of electrocardiogram analysis.[202,203]

3.3.1.1 Applying convolution to blinking data

Based on the structure of the fluorescence intensity, inherently a time series, we decided to experiment with a CNN model architecture and test a similar model based on 1D convolutional layers (see the used architecture Figure 26 and Appendix Table 16).

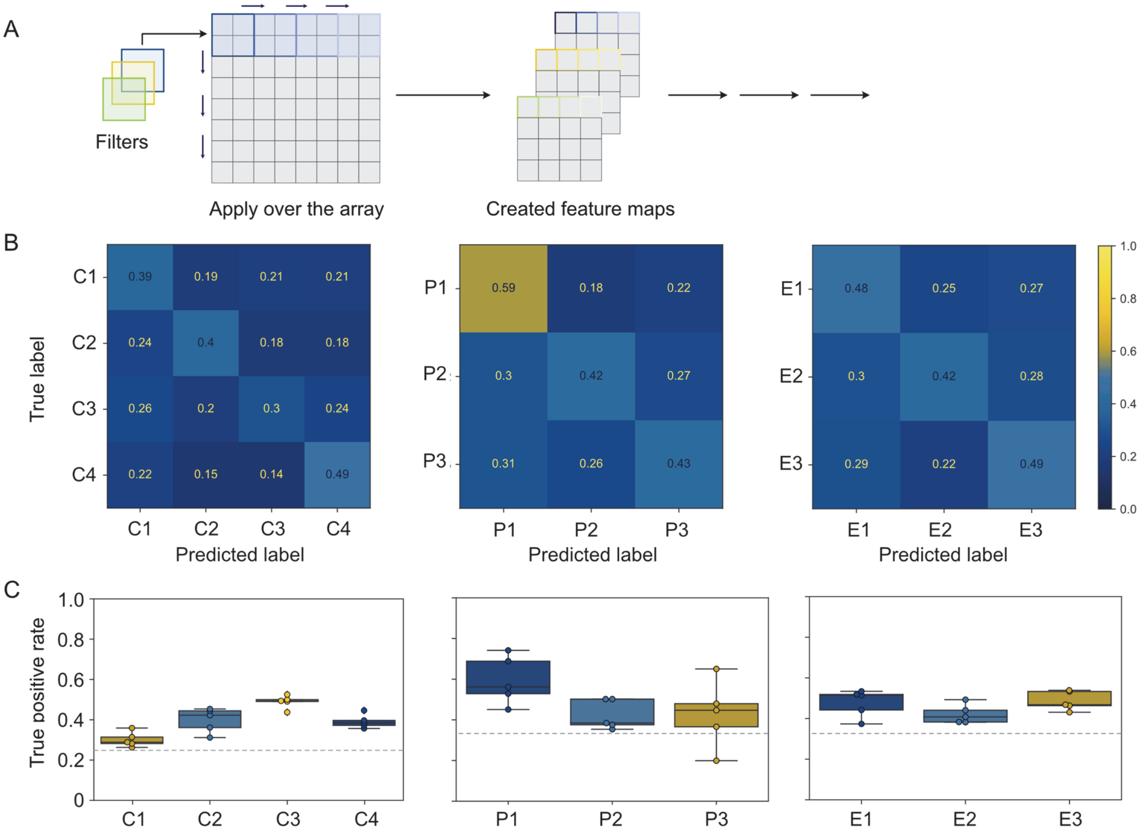


Figure 26. Five-fold CV of a CNN model using the full z-scaled and min-max scaled traces as input. A, Visualization of convolution, filters are applied over the whole array resulting in a new feature map per filter. Hence, original the dimension decreases while the lateral dimension increases. B, Evaluation of the model (Appendix: Table 16) on all datasets. C, Plots of the true positive values obtained for all the five models trained in the five random splits. The middle line represents the median of all the values, while the whiskers represent the quartiles, without outliers as determined by the Seaborn implementation.[204] The grey dashed line indicates a rate of 0.25 (left) and 0.33 (middle, right).

The results we obtained from our modified CNN architecture (Table 16) did not outperform the best results obtained in Chapter 2 using the manually extracted features in combination with an ensemble model (Stacked model) or even the MLP. Nevertheless, the CNN model achieved an accuracy ranging between 40% and 45%, showing a mean overall accuracy of the fivefold validation set, comparable to the conventional approach and significantly higher than random guessing (Figure 26, B, C). We concluded that the CNN architecture can capture some sequence specificity from the single-molecule traces but might not be the optimal model for this task as it does not take a directional temporal component into account. However, models like RNNs have been designed specifically to incorporate this temporal aspect and may be more appropriate for this purpose.[205]

3.3.2 Recurrent model

RNN has been developed to work with sequential data such as text in the context of natural language processing.[200] In RNN, all operations are calculated for each timestep with the same parameters. The key component is the hidden state of the RNN, which mimics a memory across the sequence. At every timestep, the hidden state from the previous step is updated with the new information from the current step. Although RNNs can "remember" previous information in the sequence they face difficulties in training. Training an RNN with long sequences can result in a vanishing or exploding gradient, preventing convergence.[206] To solve this problem adapted architectures like long-short-term memory (LSTM) and gated-recurrent units (GRUs) have been designed.[198]

The basic idea remains the same, in that each times step is processed sequentially. However, in LSTMs the so-called cell state takes over the function of memory and is separated from the hidden state. Furthermore, there are three gates (operations on the hidden state-, cell-state, and input), the input, forget, and output gate. The forget gate modulates the information obtained from the previous step, while the input gate determines the relevance of the current input so both gates together determine the new cell state. Lastly, the output gate determines the new hidden state based on the cell state. All gates have their learnable parameters that are optimized during training. These mechanisms allow an LSTM model to learn long-term relations if they are relevant to the task while mitigating the problem of vanishing gradients. GRUs are a simplified version of LSTM and use only two gates, the update, and the reset gate, which decide which information to add and to discard respectively. In general, GRUs are less complex and faster to train but come with a reduced memory range.[201,207,208]

3.3.2.1 Implementing a recurrent model

Although LSTMs and GRUs were conceptualized for long-range memory and perform better on long sequences, using the 6000 frames of fluorescence traces would be too long for efficient training.[201] Therefore, we decided to keep our 1D-CNN model and only insert the GRU layers after the convolutional layers. The convolutional layers are used to reduce the length of the traces while extracting or emphasizing features that then can be learned by the inserted GRU layers. Initial experimentation indicated improved performance. Therefore, we decided to go ahead and optimize our new 1D-CNN-GRU model.

The addition of recurrent layers to the CNN layers improved the model performance from the previously observed ~40% to ~65% (Figure 26 and Figure 27). We then explored whether the model is sensitive to changes in the hyperparameters. We experimented with a low-level hyperparameter search using an available Keras implementation for hyperparameter tuning called Keras-Tuner in combination with TensorBoard for visualization and early stopping to prevent overfitting and lower run times (Figure 27).[209,210] We could observe improvements when using adapted parameters, with most models reaching accuracies of 60-70% (Figure 27).

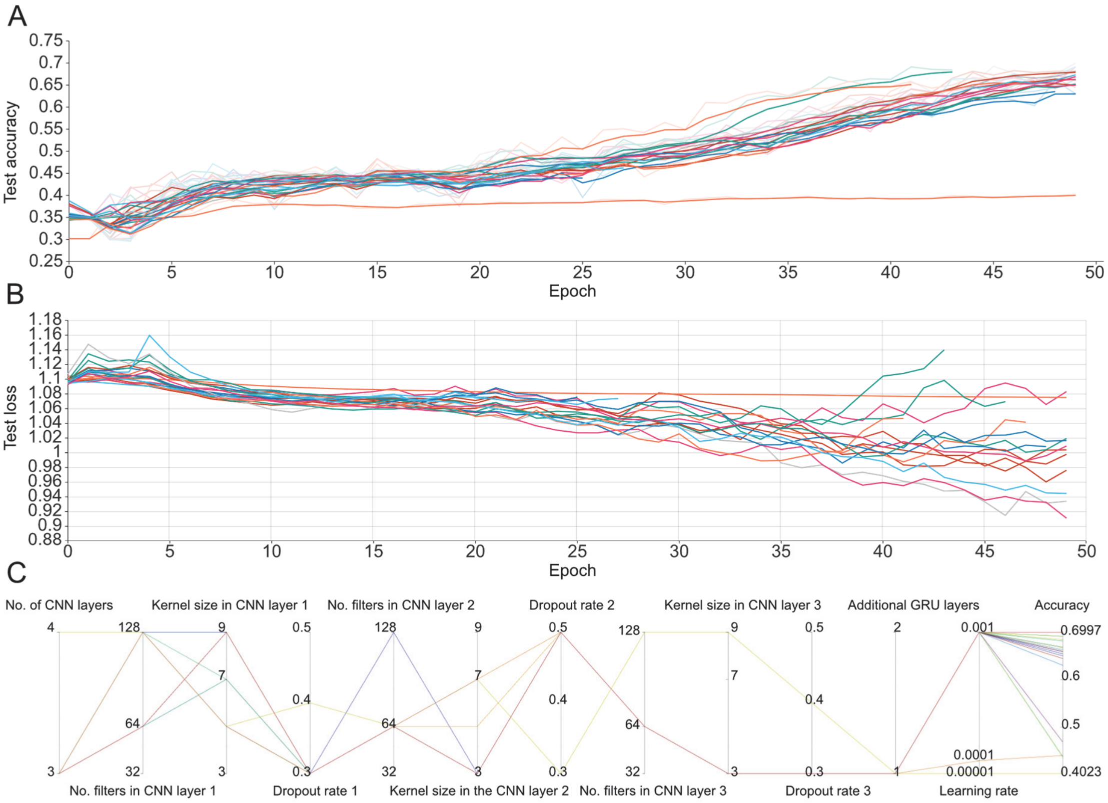


Figure 27. Hyperparameter optimization results in the model (red) used for further analysis. The optimization was run for 30 trials using the BayesianOptimization Tuner on the peptide **P15-P17**.[209] A and B are obtained from TensorBoard. C was replotted for better visibility using Plotly.[211] Colors in A and B do not match C. A, Overall accuracy in the test set over the training epochs. B, Loss on the test set over the training epochs. C, Hyperparameter combinations chosen by the BayesianOptimization Tuner and the resulting accuracy on the test set. The best model over 30 runs on the peptides achieved 70% accuracy in the run and led to the model architecture described in Table 17.

3.3.2.2 Validation

To test whether the model architecture would perform reliably and estimate its generalizability we performed nested CV (NCV). In CV a test set is set aside and from the remaining data, a fivefold CV with another split into a training validation set is performed and the model is finally evaluated on the test set (Figure 28). In NCV this is repeated five times with new validation splits to ensure that the initial validation data split does not influence the outcome. The results are summarized in Appendix Table 14 and Figure 28, B. The model performs reliably on all data sets with accuracies between 60 and 70%, except for the epimer set where we can observe outliers in the evaluation results, from models that did not converge most likely due to exploding gradients during training (Figure 28, B).

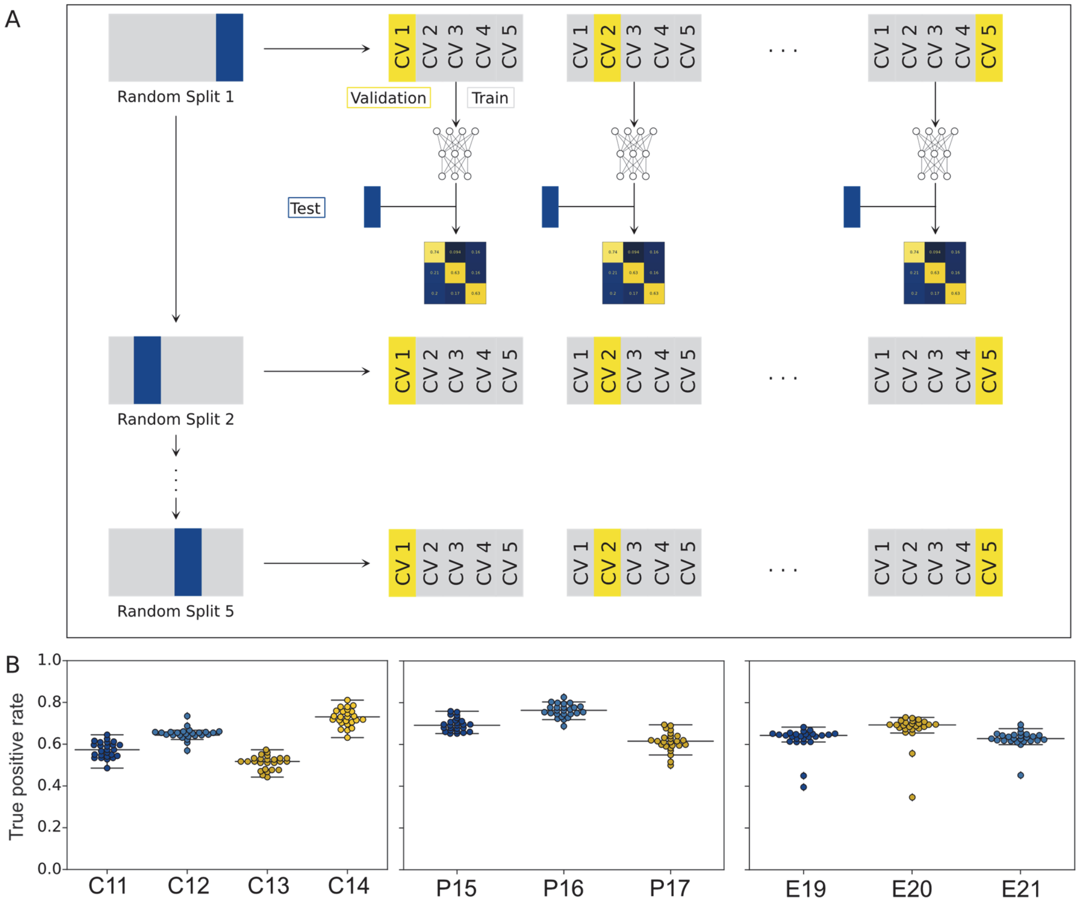


Figure 28. NCV. A, Evaluation of the model (Table 17) on a dataset. B, Plots of the true positive values obtained for all the 25 models trained in the five random splits with five-fold CV. Values are displayed for the deterministic 1D-CNN-GRU model. The middle line represents the median of all the values, while the whiskers represent the quartiles, without outliers as determined by the Seaborn implementation.[204] Reprinted in modified form from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* 2023, *145*, 1441–1447.[91]

The results of this model are very promising and clearly show that there is information contained in the fluorescence blinking traces that can be leveraged to distinguish the immediate environment of a single spontaneously blinking fluorophore.

We speculate that reaching a higher accuracy might be feasible with an improved dataset possessing less noise or fewer background signal traces. To verify this assumption, dedicating more time to refining sample preparation would be necessary. However, achieving this within our current procedures is challenging and would demand extensive testing and perhaps a meticulously controlled environment, such as in a clean room.

Given our existing dataset, we propose an alternative strategy to further eliminate noise or uninformative traces. Currently, our evaluation method is deterministic, i.e., each test signal trace provided to the model generates a definite prediction. In an optimal scenario, we would want to measure the certainty level of each prediction. Such information would enable a user to exclude traces from evaluation if their predictions do not meet the specified standards.

3.4 Model uncertainty

Model uncertainty is becoming more and more important in many applications of ML models. Especially when the outcome has direct consequences in real life, such as in evaluating human health or steering autonomous vehicles. With a quantified model confidence, uncertain cases can be handled differently, or be removed altogether from a test set. The experimentalist should be able to adapt the model output to the problem at hand in terms of certainty of the proposed class at the cost of coverage of the recorded traces.[212]

However, DL models do not have a built-in way of determining model uncertainty. The output of the Softmax layer, which is generally used as the last activation function and maps the output from the last hidden layer to the class vector, cannot directly be interpreted as a probability.[212] For this purpose, we implemented a probabilistic adaptation of our model that outputs a certainty measure of the corresponding prediction. Multiple approaches have been proposed in the ML field to extend the deterministic models to probabilistic models, with the most applied being ensemble methods like Bayesian models or Monte Carlo dropout (MCD).[212–215]

3.4.1 MCD

MCD is a straightforward technique for estimating prediction uncertainty for models featuring dropout layers. Dropout layers are applied during training as a measure to lower overfitting and then turned off during evaluation. When dropout is applied, activation outputs of random nodes are set to zero, preventing the model from relying on a few nodes.

In the MCD approach, dropout is applied during prediction as well. A distribution of predicted values from slightly varied models can be obtained by performing multiple rounds of predictions. Consequently, the uncertainty can be estimated from the obtained distribution.

We implemented the approach into our deterministic model, while slightly adapting the dropout values from the original model (1D-CNN-GRU-MCD, Table 19). The output corresponds to a distribution of Softmax outputs per class in the classification problem. To express the model's uncertainty for each test trace as a singular value, which could subsequently serve as a filtering criterion, we utilized a statistical distance metric known as the Wasserstein distance. It is a non-parametric distance measure, also referred to as the earth-movers distance. The distance is equal to the minimum cost of reshaping one distribution into another, thus factoring in both the distance and shapes of the distributions (Figure 29, C-D). The intuition of the Wasserstein distance can be described as a mass at location x, y (*π(x,y)*) that is moved to a new location (*dπ(x,y)*) requiring a certain moving cost (*c(x,y)*) (Figure 29, Eq. 1). However, the mass can be moved along many paths. Hence, to obtain the path with the lowest costs (which correspond to the distance |x-y|), the minimal solution in the ensemble needs to be found (Figure 29, Eq. 2). While the Wasserstein distance can be calculated between all classes, the relevant one in our case is between the class with the highest average Softmax output and the next closest (Figure 29, D). This value can be used as a filtering criterion comparing the calculated value to a chosen certainty threshold (CT), to decide whether to keep traces or discard them due to potentially unreliable predictions. The larger the distance, i.e., the more different the prediction distributions are, the more certain the model prediction is expected to be.

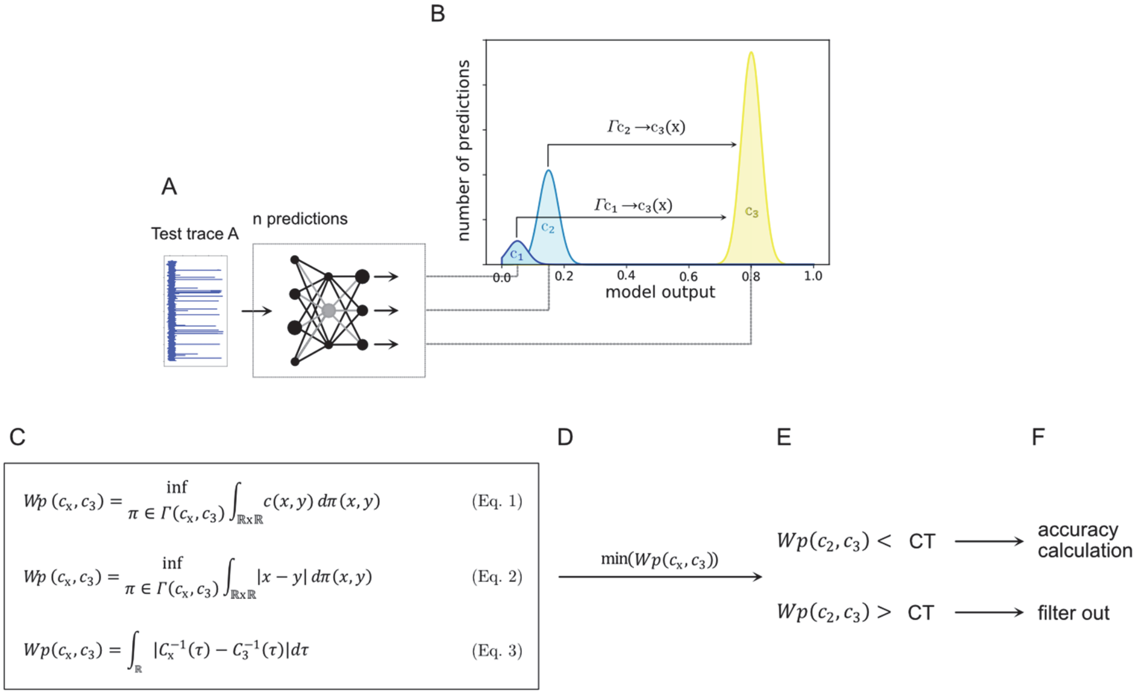

A
Test trace A
n predictions
B
number of predictions
model output
0.0
0.2
0.4
0.6
0.8
1.0
$\Gamma c_2 \rightarrow c_3(x)$
$\Gamma c_1 \rightarrow c_3(x)$
c_1
c_2
c_3
C
$Wp\,(c_x, c_3) = \inf_{\pi \in \Gamma(c_x, c_3)} \int_{\mathbb{R}\times\mathbb{R}} c(x,y)\,d\pi(x,y)$ (Eq. 1)
$Wp\,(c_x, c_3) = \inf_{\pi \in \Gamma(c_x, c_3)} \int_{\mathbb{R}\times\mathbb{R}} |x-y|\,d\pi(x,y)$ (Eq. 2)
$Wp(c_x, c_3) = \int_{\mathbb{R}} |C_x^{-1}(\tau) - C_3^{-1}(\tau)|d\tau$ (Eq. 3)
D
$\min(Wp(c_x, c_3))$
E
$Wp(c_2, c_3) <$ CT
$Wp(c_2, c_3) >$ CT
F
accuracy calculation
filter out

Figure 29. Principle of MCD and definition of CT. A, To get statistics for each test trace in the test set, 100 predictions are calculated using the model 1D-CNN-GRU-MCD with active dropout. B, The Wasserstein distance between the class with the highest probability and the other classes is calculated. Γ represents the mapping of one distribution into the other. C, The formal description of the Wasserstein distance. The W_p distance can be expressed in a closed form through the cumulative distribution function, which is implemented in the SciPy Python library (Eq. 3).[176] D, The minimal distance of all distances to the class with the highest model output is saved as the quantified uncertainty. E, The minimal distance is compared to the CT, which is a number chosen by the experimentalist (0 ≤ CT ≤ 1). F, If the distance is larger than the CT, the trace is included for the accuracy calculation if not it will be discarded. Reprinted in modified form from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* 2023, *145*, 1441–1447.[91] 🅭🅯

We tested the new metric and used it to filter out potentially uncertain traces. We found that traces with a smaller Wasserstein distance (less certain predictions) were more likely to be wrongly predicted. By filtering low-certainty traces, the overall accuracy increased (Figure 30, A-C). When comparing the different peptide sets, we can see that in the case of E19-E21 more traces are lost at the same CT (convex curve), potentially indicating that this data set is noisier than C11-C14 and P15-P17.

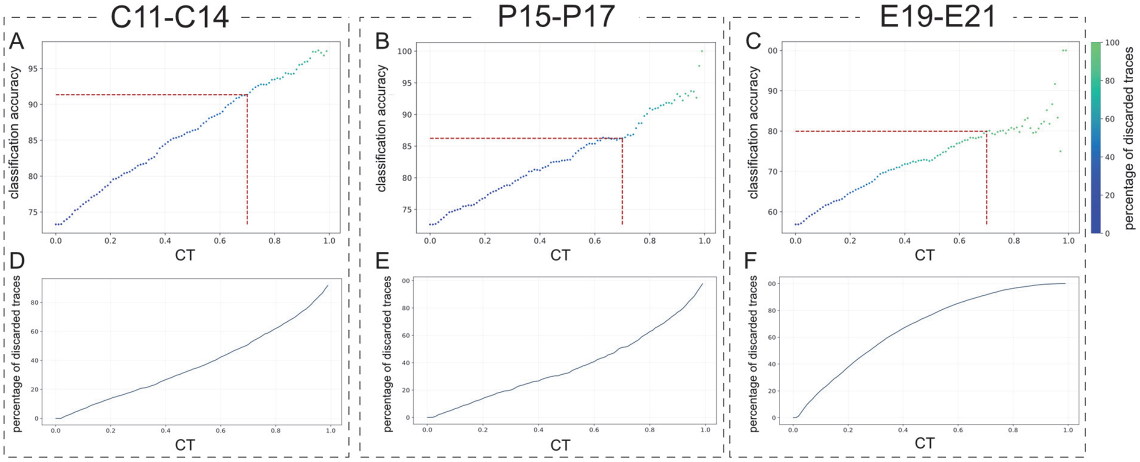


Figure 30. Certainty filtering, accuracy, and discarded traces. A-C, Relationship between the classification accuracy and the CT used for filtering with the percentage of lost traces during filtering being color-coded. D-F, Relationship between the classification accuracy and the CT used for filtering. The color bar represents the percentage of discarded traces. Reprinted in modified form from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* 2023, *145*, 1441–1447.[91] 🅭🅯

We conducted a control experiment to confirm that a high CT is associated with a higher accuracy and the same could not be achieved by randomly removing traces (Figure 31). Traces were eliminated independent of their CT and no increasing accuracy was observed. Toward very high CTs, the overall accuracy tended to become erratic due to the small number of traces left in the corresponding test set.

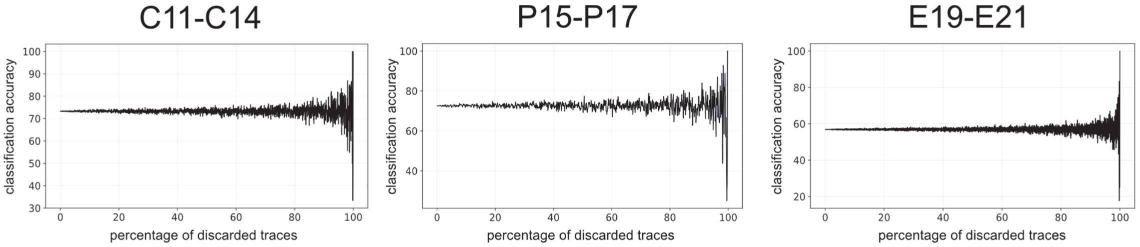


Figure 31. Classification accuracy after discarding an increasing percentage of random traces from the test set. This experiment shows erratic behavior when the number of traces becomes very low, but no general increase in overall accuracy is observed. Reprinted from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* 2023, *145*, 1441–1447.[91] 🅭🅯

Additionally, we checked the composition of the test set with more stringent filtering. Since there is no obvious reason why one class of peptide should contain more experimental noise than another class acquired under the same conditions, we would expect the composition to remain approximately the same (Figure 32). At very high CTs (>0.8) the number of traces approaches or reaches zero resulting in an apparent large shift of the relative number of traces in the different classes, which corresponds however to a small absolute difference.

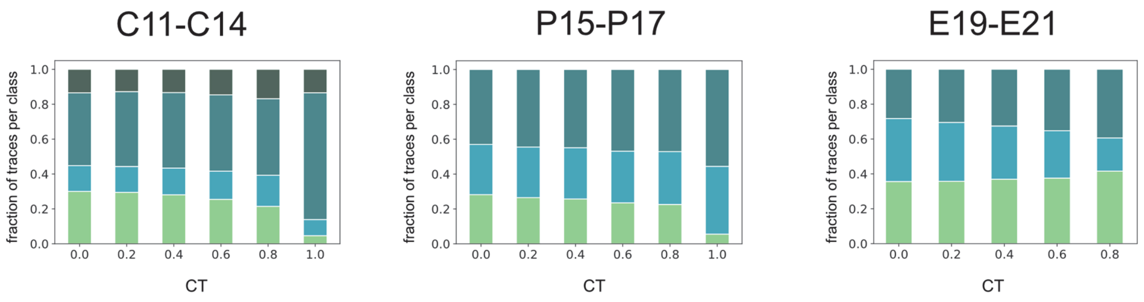


Figure 32. Composition of the test set after discarding traces based on their CT value. The bars in the plots represent the percentage of traces of a class in the traces used for calculating the accuracy after filtering with a specific CT value for all peptide sets. Reprinted from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* 2023, *145*, 1441–1447.[91] 🅭🅯

Based on these results, we were confident that MCD in combination with the Wasserstein distance as a certainty metric can be used to improve the classification results, at the expense of coverage. This approach and a CT of 0.7, which leads to a loss of approximately 50% of traces for **C11-C14** and **P15-P17** results in very high accuracies, especially for the small peptides (**C11-C14**) set with 90.2% (Figure 33 and Table 18).

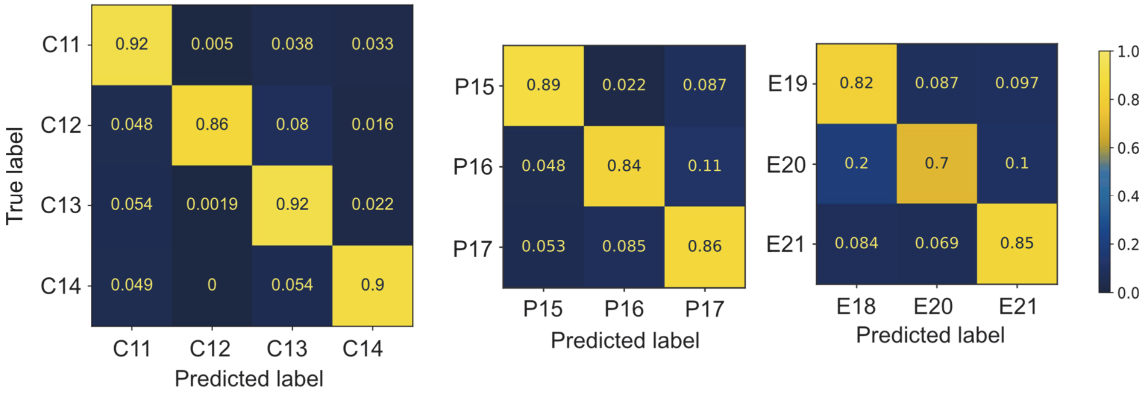


Figure 33. Classification results using the 1D-CNN-GRU-MCD model with a CT of 0.7 for all peptide sets. Reprinted in adapted form from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* 2023, *145*, 1441–1447.[91] 🅭🅯

To further solidify our results and ensure that the model (1D-CNN-GRU-MCD) does not pick up information inadvertently introduced during the experiments we conducted an additional control classification experiment.

3.4.1.1 Control experiment

We designed a control experiment to exclude that the model learns any background information or there is incorrect information flow during training and classification. Starting from the raw data of the peptide set C11-C14, we attempted to classify traces extracted from the background noise in the movies. To get background signal traces, only the maximum projection of the movies (Figure 34, A first and second tile) was segmented using the thresholding algorithm Yen (Figure 34, A middle tile).[216] Then, random x-, and y-positions were chosen such that a 5x5 pixels area could be placed in the background segment. The intensity over time within this box was used as the background traces (Figure 34, A). The label for the trace was assigned corresponding to the experimentally applied peptide. The extracted time traces (15 per movie) were postprocessed in the same way as the signal traces without the filtering step. Attempts at training the same model architecture under the same conditions resulted in low accuracies (Figure 34, B). When filtering with high CT, the accuracy increased, however, it shows a similar behavior to the random dropping experiment (Figure 31). Hence, the increase might be the result of the low trace numbers. Nevertheless, it could be valuable to inspect the remaining traces for visible abnormalities or features. This control experiment indicates that the model learns from the blinking behavior of the fluorophore in the respective samples, and it is not an experimental bias allowing the model to achieve higher accuracies. However, we do not yet know what characteristics allow the model to do so.

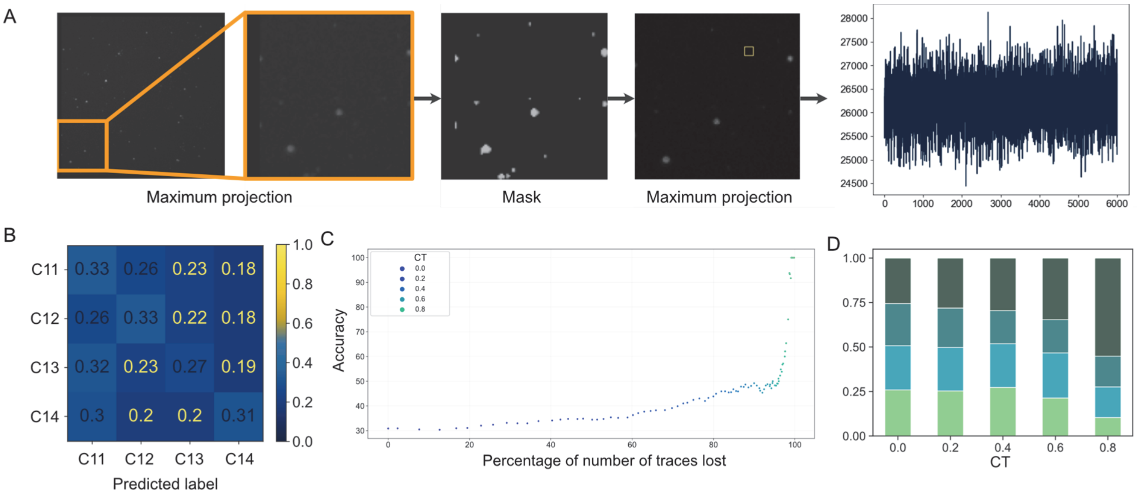


Figure 34. Control experiment to exclude learning of any eventual background information for the initial peptide set **C11-C14**. A, Extraction of background signal. B, Confusion matrix of the classification obtained for the control experiment (1218 test traces). C, Accuracies obtained when filtering out a given number of traces based on their uncertainty. The color code depicts the corresponding CT. D, The change in the composition of the test set with changing CT. Each color in the bar represents a different "peptide". E, Reprinted in adapted form from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* 2023, *145*, 1441–1447.[91]

3.5 Conclusion and outlook

Throughout this chapter, we have worked toward a DL model architecture that could be applied in our tool "blinkognition". Such a model had to be able to extract relevant information from the fluorescence time traces of spontaneously blinking fluorophores conjugated to peptides, aiming to differentiate between these conjugates. While we had gathered the raw data in Chapter 2, our attempts to differentiate between them were unsuccessful. We demonstrated that an architecture based on CNN and GRU layers is capable of classifying traces across all measured peptide sets significantly more accurately than one would anticipate with a random guessing approach. Our control experiment provided additional evidence supporting that the model effectively learns the information inherent in the spontaneous blinking of the fluorophore. We further integrated the MCD technique to estimate uncertainty in the predictions made by our model. For quantification of the obtained prediction value distribution, we applied a statistical distance, the Wasserstein distance. Based on this distance quantification we were able to filter out uncertain predictions and improve the classification accuracies.

These results demonstrated that blinkognition is a viable tool for the identification of small peptides at the single-molecule level with high sensitivity in a controlled setting. Not only were we able to distinguish peptides of different sequences and charges (C11-C14) but also small chemical modifications like PTMs and their position. Even in an as subtle change as in the case of epimerization overall classification accuracies from 57-79% could be obtained, depending on the acceptable loss of traces.

As it stands, blinkognition can be used for targeted problems with a clear hypothesis. It is, however, not able to sequence or identify unknown peptides. To adapt blinkognition to a new problem, the corresponding pure standards need to be prepared, measured, and the acquired data used for model training. A notable potential advantage of blinkognition is that, unlike other single-molecule approaches for peptide and protein analysis, no sequencing based on a degradations step is required. Consequently, the performance of blinkognition is theoretically independent of the sequence length of the targets, provided it can interact with the spontaneously blinking fluorophore. However, so far, our demonstrations have solely involved the classification of synthetically pure peptide samples equipped with a preinstalled immobilization handle, specifically an alkyne modification for click chemistry. A natural next step would be to demonstrate blinkognition performance on larger molecules, e.g., proteins. This demonstration would give insight into the scope of the method and its applicability to larger molecules. Furthermore, this application would require the development of new strategies to label the fully formed protein and subsequently immobilize it on the surface. This work will be presented in Chapter 4. Yet, we envision that blinkognition is not limited to the realm of peptides and proteins but could extend its applicability to other macromolecules with limited analytical methods, such as oligosaccharides. Besides exploring the scope of the approach, numerous avenues for further development of experimental and analytical aspects could be explored to fully exploit the potential of the technology and improve its performance.

On the experimental side, noise reduction and the availability of high-quality training data are crucial. Consequently, improvements in the experimental preparation such as immobilization strategies and anti-fouling properties of the surface could benefit the method. Another valuable approach could be the addition of a second blinking fluorophore of a different color attached to the same peptide or protein. This additional signal would facilitate the filtering of noise signals because colocalization of signals in the two channels would be less likely for noise. Additionally, the signal would further increase the available information about the peptide of interest (PPOI). A second fluorophore would report on the environment at a different position of the same peptide or protein and might do so with a differing sensitivity. Moreover, a second dye could be chosen such that interaction via FRET is photophysically possible, which would allow for information on spatial distances between the two residues to be implicitly included in the fluorescence traces.

With an increasing number of employed fluorophores, developing new, robust blinking dyes with longer lifetimes and potentially new blinking mechanisms could support the development of blinkognition.

Following HMSiR, other blinking fluorophores of different colors have been developed, e.g., green versions such as HEtetTFER and HM-JF526 as well as a yellow carborhodamine HMCR550.[87,217,218] Beyond the expansion of the color palette other nucleophiles such as spirolactams and sulfonamides as well as a different class of dyes (polymethines) have been introduced.[88,104,219–224] These alternative fluorophores might interact differently with biomolecules and display variable sensitivity. Therefore, chemically diverse spontaneously blinking dyes could be systematically tested for improved classification performance and an increased sensitivity range, which could broaden the scope of blinkognition. The dynamic range of new probes could be evaluated on a standardized protein classification problem or initially on bulk experiments. In bulk, the sensitivity range for environment parameters (e.g., polarity or viscosity) could potentially be approximated by the slope of the fluorescence intensity curve when solvent properties such as pH or the dielectric constant are varied.[88,220] A flat slope would indicate that on the single-molecule level, the number of molecules that are florescent at the same time changes slowly with the given parameter. Hence the blinking behavior, e.g., average ON- or OFF time should correlate with the bulk fluorescence and differences can be observed over a larger range of environmental changes.

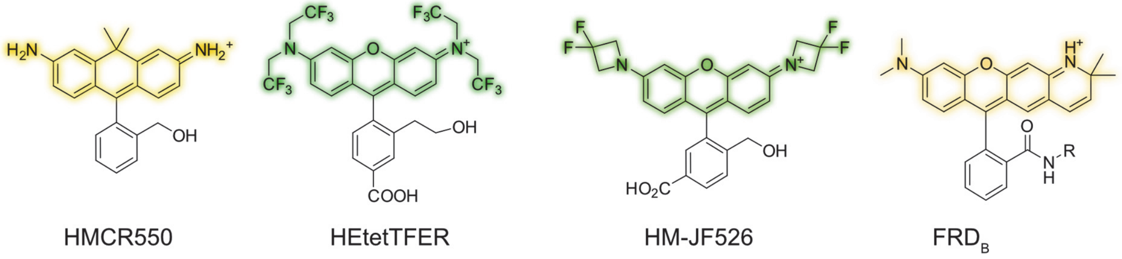


Figure 35. Examples of spontaneously blinking fluorophores of different colors and an alternative nucleophile.[87,102,217,220]

Beyond improving the experimental procedure and fluorophore properties, the data analysis could be explored, potentially also integrating new developments in the ML field. The model architecture used in our studies resulted from exploratory investigations followed by a low-level hyperparameter search. Although the model did not seem to be sensitive toward changes in hyperparameters (the best models of the explored hyperparameters resulted in overall accuracies of ~70%), it cannot be excluded that more extensive testing or a different model architecture would result in a better performance. Therefore, exploring other architectures for time series classifications could provide an architecture that is better at learning relevant features from our data.[225,226] Transformers deliver state-of-the-art results in natural language processing, computer vision, speech processing, and they are increasingly investigated for time series analysis tasks. Applying transformers offers a future avenue to advance blinkognition but so far has not been tested due to their impracticality for long-time sequences and large data requirements for training and generalizations. However, these issues are being actively addressed in the field, and a modified transformer architecture, transfer learning, or alternative preprocessing of our data offers opportunities for future exploration.[227] Another approach to improve preprocessing and analysis is to gain a better understanding of what the algorithm learns. Initial insights could be obtained by comparing traces (visually or through extracted features) that have been assigned labels with high and low uncertainty. This comparison may reveal discernible trends and guide future analysis. Beyond such a comparison, post-hoc interpretability tools could be explored and applied to our results.[228]

So far, we have tested three and four class problems, however, it will be crucial to explore the upper limits of classes (peptides and proteins) that can be distinguished and how strongly they depend on the species that are compared. Furthermore, it will be crucial to assess blinkognition's performance in handling more complex samples containing unknown species. This problem setting corresponds to out-of-domain (OOD) detection, i.e., the model needs to recognize OOD data while still producing a good classification of the known classes. Currently, we are relying on MCD for uncertainty estimation of our model and applying it to filter out traces that have a low prediction certainty. These traces might be either noise or traces missing class-distinctive features but no OOD data. Unknown but labeled molecules would generate OOD data that have single-molecule characteristics with similar features. Hence, this task will be extremely challenging. However, OOD data detection is an active research topic in ML and many alternative approaches to MCD have been proposed, which could offer opportunities in the future.[229–231] To test ODD detection and explore the space of different blinking behaviors that can be observed, it would be beneficial to record an OOD dataset. Such a dataset could be obtained in a synthetic approach, e.g., split-and-mix SPPS using the previous building block and surface immobilization protocol.[232] Alternatively, a biological sample could be obtained from a trypsin digest of the cellular proteome sample (i.e., a cell lysate). This approach would require the experimental optimization of labeling the obtained peptides with HMSiR and an alkyne for surface immobilization but could be based on a procedure published in the context of fluorosequencing.[233] A large amount of un- and labeled data could open opportunities for applying semi(self)-supervised or a contrastive learning approach to detect OOD by improving the quality of the extracted features.[234,235]

Chapter 4 Toward the analysis of proteins

The results reported in this chapter are from collaborative work with Dr. Krzysztof Bielec during his postdoctoral time in the laboratory. He conducted some experiments and prepared some figures presented here. We thank Dr. Zacharias Thiel for the preparation of Map555-Halo and compound 33, Henriette Lämmermann for purified HaloTag7 protein and compound 26, Sarah Emmert for the HaloTag7 plasmid, Annabell Martin for the purified SNAP_f protein, and Carla Miró Vinyals for the sCGrx1p protein. Gianluca Quargnali contributed to the labeling strategy and synthesis protocol of compound 22 during his M.Sc. Thesis. Furthermore, we would like to thank Besim Fazliji from the laboratory of Professor Roland K.O. Sigel at UZH for the initial encapsulation protocol.

4.1 Project goal

In Chapter 3 we provide evidence for the fundamental idea of blinkognition, that the fluorescence output of spontaneously blinking fluorophores is sensitive to its chemical environment. We were able to use fluorescence time traces of peptide-fluorophore conjugates to distinguish small peptides using a DL model that was trained on data from a pure synthetic standard. Furthermore, we demonstrated the sensitivity of blinkognition beyond changes in the peptide sequence. PTMs such as phosphorylation and epimerization could also be distinguished in our test molecules. To date, we have applied our analysis to synthetic peptides that are relatively small (up to 16 amino acids) and have been modified during synthesis to carry an immobilization handle. Since blinkognition is formally independent of the sequence length it should be extendable to proteins. For the method to be applicable in settings where samples of naturally occurring peptides or proteins should be analyzed, further developments are necessary. The two main aspects that require optimization are the bioconjugation of the fluorophore to a protein and the immobilization on the surface for imaging.

There are important considerations for the fluorophore conjugation and the immobilization. Ensuring proximity between the fluorophore and the target protein throughout the fluorophore's lifespan provides the basic interactions required for blinkognition. Therefore, the fluorophore is ideally covalently conjugated to the POI. To obtain pure training data the labeling needs to be specific and consistent. Based on our current understanding, the location and quantity of dye binding to the protein should not critically impact the performance of blinkognition. We hypothesize that multiple dyes could be beneficial as they will report on different parts of a protein in a combined blinking signal (as they are not spatially separable by TIRF microscopy). However, it is imperative that the labeling remains consistent across both training and testing samples, hence the proteins should be labeled to the highest degree possible.

For immobilization, covalent attachment as has been used for the peptides is one possibility but would require a dual-reactive handle. This handle would need to attach to a specific site on the protein and possess a compatible moiety on the surface. Ideally, the immobilization process should be universally applicable to all proteins. Immobilization merely needs to allow for the molecule to be imaged during the whole lifetime of the fluorophore but does not need to be directly via the protein. Therefore, an alternative approach could be the encapsulation of the protein in lipid particles of a controlled size. Such an approach has been successfully used in single-molecule FRET studies of messenger RNA (mRNA) constructs.[236]

In this chapter, we aim to tackle these two main challenges and attempt the classification of proteins using blinkognition. We first designed and synthesized a spontaneously blinking probe for labeling recombinantly expressed proteins. With the substrates in hand, labeling experiments were conducted to optimize conditions. Encapsulation was adopted in our laboratory and adapted to our use case by Dr. Krzysztof Bielec. We attempted to classify the generated labeled proteins using a model based on the previously employed ML architecture.

4.2 Protein labeling

4.2.1 Labeling strategy

Proteins are commonly labeled with fluorescent molecules through a reactive amino acid side chain. Common targets are the primary amine of lysine, the *N*-terminus, or the thiol group of cysteines due to their nucleophilic reactivity.[237] When determining the amino acid and group to target in general, three main aspects should be considered: the abundance of natural amino acids, which determines the maximum coverage of proteins (Figure 36, A), the achievable amino acid selectivity, and the efficiency of labeling. In our case, targeting a specific amino acid of a protein should be less important compared to obtaining complete labeling of the respective residue. Zanon et al. (2021) recently assessed the selectivity and efficiency of established electrophilic protein modification agents on the SH1000 *Staphylococcus aureus* test proteome using MS (Figure 36, B). Their analysis revealed widespread and specific labeling of cysteine and lysine residues. However, lysine-targeting reagents sometimes encountered challenges by also reacting at the *N*-terminus. Their analysis did not quantify the specific residue labeling percentage over the total residues present, considering both the relative abundance (Figure 27, A, where cysteine is 3.8 times less common) and the total labeled residues, we opted to target cysteine in our preliminary study for protein blinkognition, following the approach in Chapters 2 and 3. Nonetheless, lysine residues remain an appealing target for potentially introducing a secondary color channel in the future.

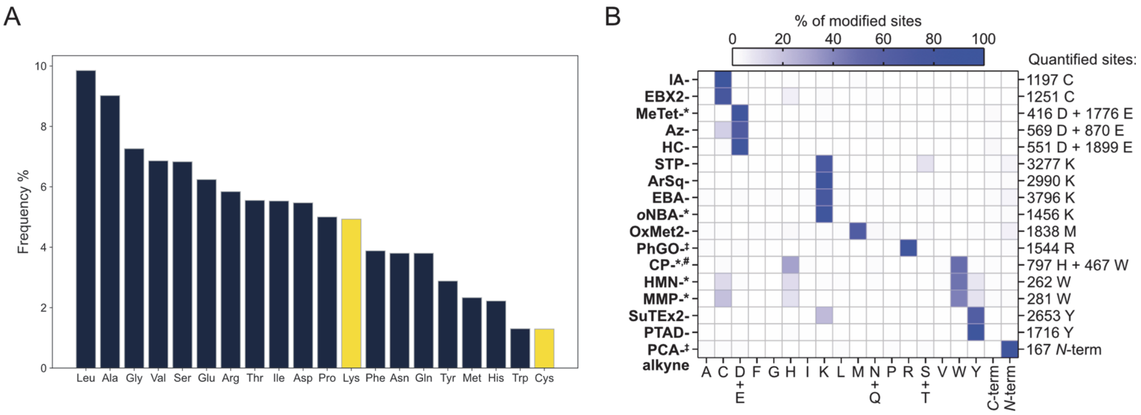


Figure 36. Relative amino acid abundance and reactivity with common reagents. A, Amino acid abundances data replotted from UNIPROTKB/TREMBL protein database release 2023_05 statistics (Release 2023_05 of 08-Nov-2023 of UniProtKB/TrEMBL contains 251,131,639 sequence entries, comprising 88,223,298,202 amino acids), lysine and cysteine residues are colored in yellow.[238] B, Reactivity studies of 17 electrophiles toward the proteome of SH100. The plot was taken from P. R. A. Zanon et al., *chemRxiv* 2021, DOI 10.26434/chemrxiv-2021-w7rss-v2.[239]

Cysteines can be efficiently targeted by reagents containing iodoacetamide (IA) or ethynylbenziodoxolone (EBX2) moieties. The cysteine coverage of EBX reagents is reportedly higher than for IA. However, EBX2 can form the covalent bond to cysteine through multiple reaction pathways leading to variability in the final covalent connection.[239,240] Such a change in connectivity could induce changes in the environment, detectable by blinkognition. Hence this behavior would effectively increase the signal variability per protein class in the classification, thereby making the task more challenging. Consequently, we introduced the reactive IA modification to the spontaneously blinking dye HMSiR at position 5 in the lower ring (Figure 37) resulting in compound IA-HMSiR **22**.

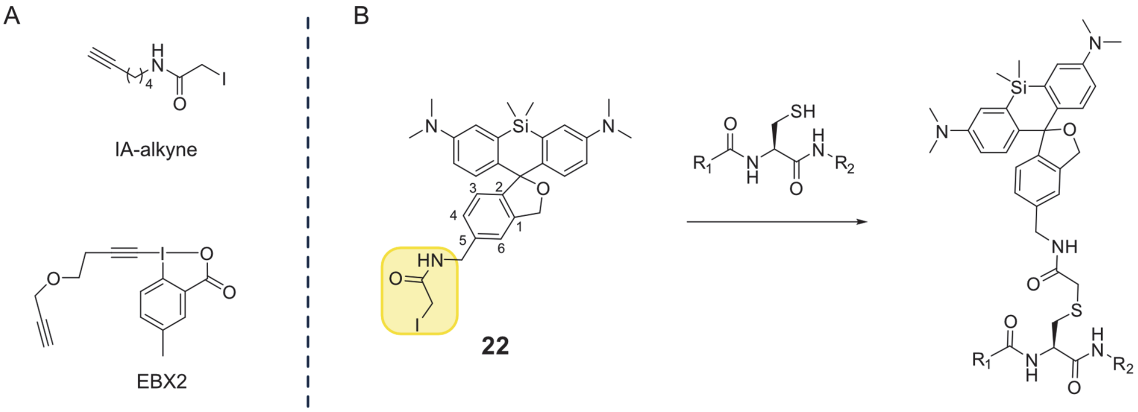
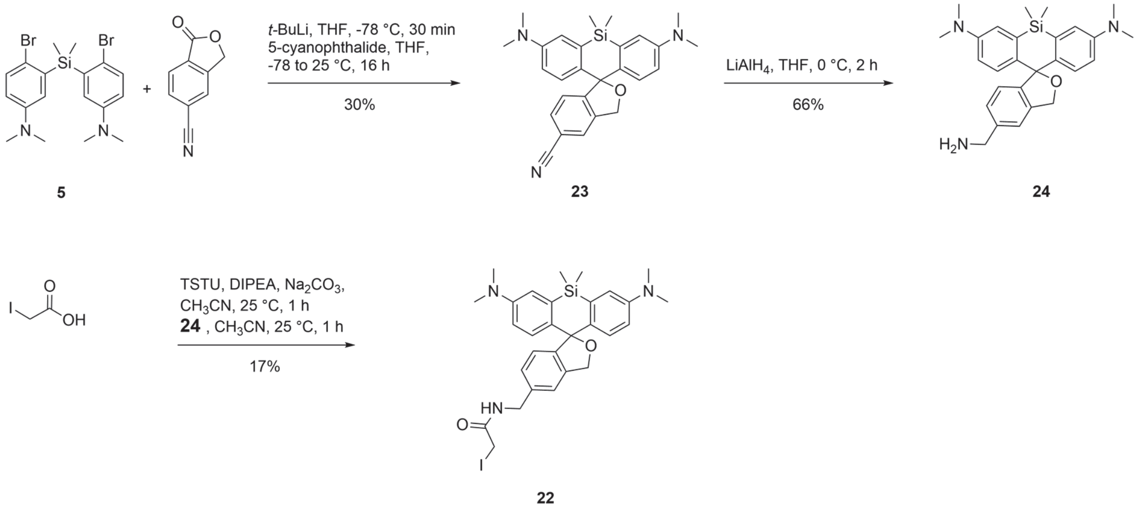


Figure 37. Reagents for cysteine-specific modification. A, Two potential reagents for cysteine modification evaluated in P. R. A. Zanon et al., *chemRxiv* 2021, DOI 10.26434/chemrxiv-2021-w7rss-v2.[239] B, The final IA (marked in yellow) modified HMSiR (HMSiR-IA) 22 and the proposed reactivity with a cysteine residue in a protein sequence.

4.2.2 Synthesis of the reactive probe **22**

HMSiR-IA 22 was synthesized from compound 5 (Scheme 11). 5 was transformed into its lithium salt, 5-cyanophthalide was added for nucleophilic attack onto the lactone, followed by the loss of a molecule of water which yielded the HMSiR with a nitrile group in the 5-position 23 in mediocre yield. Reduction of the nitrile group gave the primary amine 24, without the need for reoxidation of the xanthene core, in good yield. The amide coupling proceeded via the formation of the activated NHS-ester of iodoacetic acid under basic conditions and the addition of amine 24.

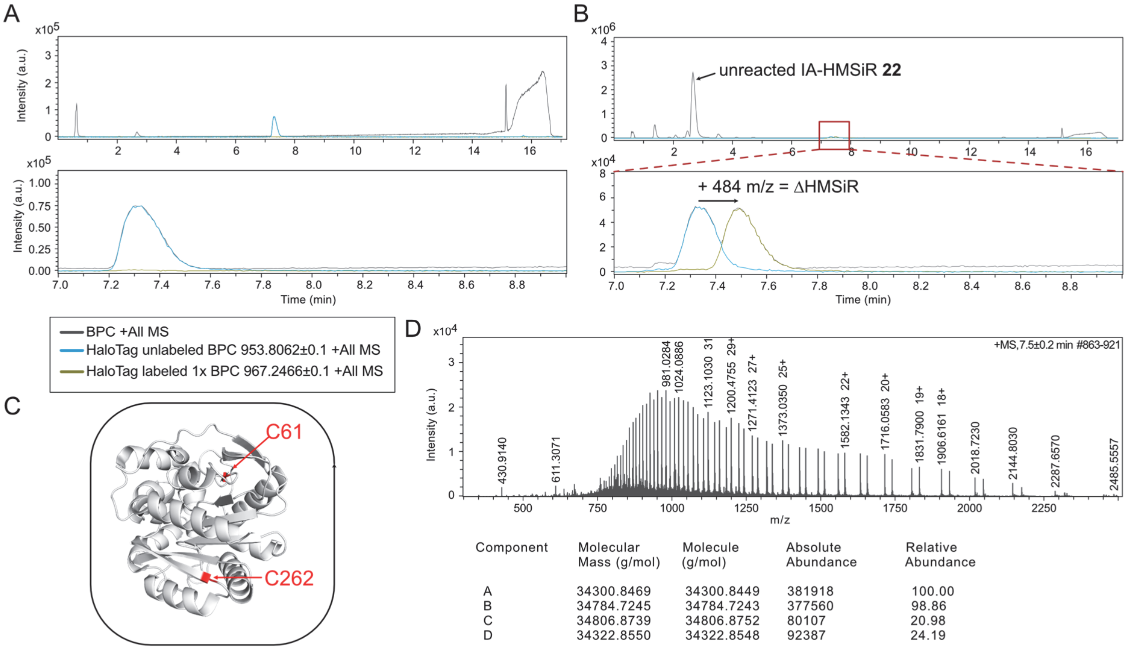

t-BuLi, THF, -78 °C, 30 min
5-cyanophthalide, THF,
-78 to 25 °C, 16 h
30%
LiAlH$_4$, THF, 0 °C, 2 h
66%
TSTU, DIPEA, Na$_2$CO$_3$,
CH$_3$CN, 25 °C, 1 h
24 , CH$_3$CN, 25 °C, 1 h
17%
5
23
24
22

Scheme 11. Synthesis of a cysteine targeting HMSiR-IA 22 over three steps. TSTU = *N*,*N*,*N′*,*N′*-tetramethyl-*O*-(*N*-succinimidyl)uronium tetrafluoroborate, DIPEA = diisoproylethylamine

4.2.3 Labeling experiments with HaloTag7

With the compound HMSiR-IA **22** in hand, we conducted labeling experiments to optimize the reaction conditions. As a first sample protein, we used the self-labeling protein HaloTag, more specifically the variant HaloTag7. HaloTag7 offers multiple advantages for initial experiments. Firstly, this protein possesses two cysteines that play no role in enzymatic activity nor are they involved in disulfide bonds (Figure 38, C).[241] Moreover, its relatively compact size of ~33 kDa, allows for efficient handling and manipulation. HaloTag is readily accessible and has been successfully expressed in our laboratory, ensuring reliability. Furthermore, HaloTag has been extensively mutated with success. The availability of various mutants associated with HaloTag provides diverse opportunities that could be leveraged for further classification investigations. Lastly, as we have discussed in Chapter 1 HaloTag can be labeled with another fluorophore that has been modified with a chloroalkane linker. This reactivity could allow for control experiments and the analysis of the accuracy of blinkognition.

To label the protein, we added our prepared reagent HMSiR-IA **22** in excess (5 equiv.) to a 15 µM protein solution in sodium phosphate buffer (PB) (100 mM pH 7.5) with shaking for 2 h at 20 °C. We monitored the reaction progress with native protein LC-MS directly from the reaction mixture (Figure 38) and quantified the labeling through deconvolution of the combined peaks (Figure 38, C and D) under the assumption that the dye would not affect the behavior of the protein in the MS analysis.



Component	Molecular Mass (g/mol)	Molecule (g/mol)	Absolute Abundance	Relative Abundance
A	34300.8469	34300.8449	381918	100.00
B	34784.7245	34784.7243	377560	98.86
C	34806.8739	34806.8752	80107	20.98
D	34322.8550	34322.8548	92387	24.19

Figure 38. HaloTag7 labeling with HMSiR-IA **22**. A and B, The grey trace corresponds to the base peak chromatogram (BPC), and the light blue trace is the extracted ion chromatogram (EIC) for the ion (z=36+) 953.8062±0.1 m/z and the green trace for the ion (z=36+) 967.2466±0.1 m/z corresponding to the un- and labeled HaloTag7, respectively. A, Complete LC-MS run (top) of unlabeled HaloTag7 in PB and close-up (bottom) of the same run. B, Complete LC-MS run (top) of a labeling reaction at 100 mM pH 7.5 after 2 h at 20 °C and close-up (bottom) of the same run. The unreacted dye is marked by an arrow. The peak shift and the mass shift are indicated with a black arrow. C, Crystal structure of HaloTag7 with cysteines marked in red.[130,242] D, The extracted mass spectrum over the time range 7.4±0.2 min. Deconvolution report from the Compass HyStar (Bruker Daltonics) analysis software, showing unlabeled HaloTag7 with mass A: ~34300 Da and the labeled HaloTag77 B: ~34784 Da. The view was generated using PyMOL.[130]

The initial labeling conditions were based on the procedures for similar commercial probes. We further compared varying amounts of excess dye compared to the protein (1, 5, 10, 15, and 20 equivalents). The labeling was conducted at 15 µM protein concentration in sodium phosphate buffer saline (PBS) at pH 7.4. After adding HMSiR-IA **22** from a 1 mM or 10 mM stock solution in dimethyl sulfoxide (DMSO), the sample was shaken for 2 h, at 20 °C and 200 rpm. In all experiments, one of the cysteines per protein (Figure 37, B and D) was labeled to ~60% exclusively and no double-labeled protein could be observed. After 2 h of reaction time, HaloTag7 labeling did not increase even when adding more than 5 equivalents or after extended labeling times at 8 °C (Figure 39, A and C). We hence conducted all following experiments with 5 equivalents of compound **1** from a 10 mM stock in DMSO, unless mentioned otherwise.

Considering the partial labeling of only one cysteine, we hypothesized that the buffer concentration and hence its capacity might be too low and the resulting pH unfavorable for further conjugation. Therefore, we tested three different pHs 7.5, 8.0, and 8.5 PB each at two concentrations (Figure 39, B) at 20 °C for 2 h. Counter to our expectation, with increased pH and buffer concentration, similar labeling extents of a single cysteine were observed. At pH 8.5 labeling was decreased, which could be due to the increased instability of the protein at such high pH levels. In a last optimization experiment, we examined if labeling could be increased when using a higher overall concentration of the reaction mixture (25 µM, Figure 39, D) and an increased reaction temperature of 37 °C for 2 h. Indeed, we observed a substantial increase of labeling at 37 °C whereas a higher reaction concentration seemed ineffective.

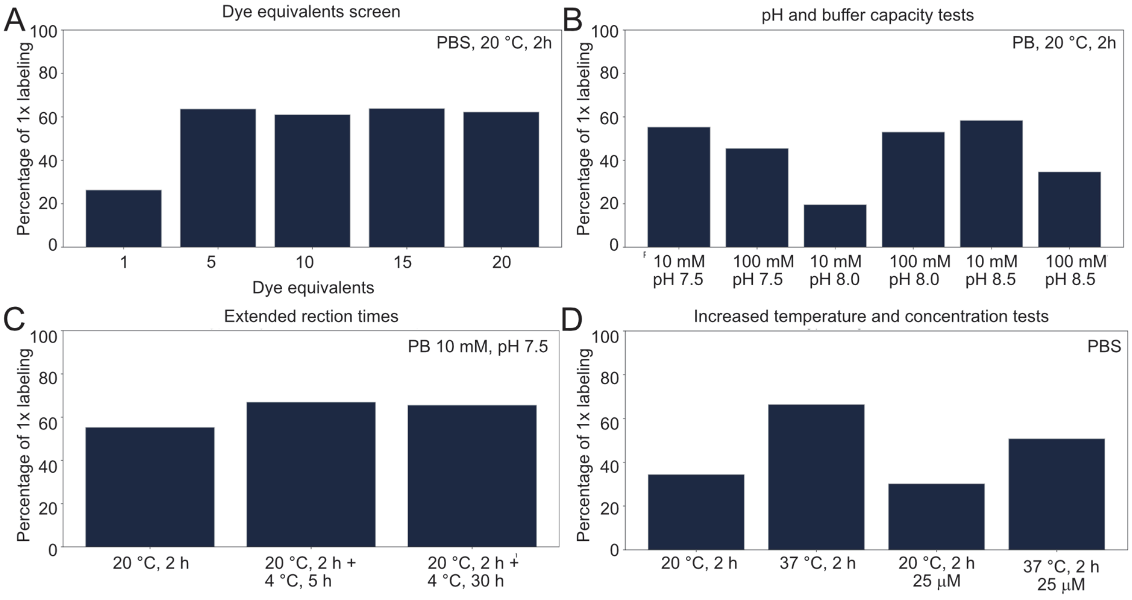


Figure 39. Percentages of singly labeled HaloTag7 under various conditions. A, Labeling extents at different HMSiR-IA **22** concentrations. B, Labeling extents in buffers of different concentrations and pH. C, Labeling extents at extended reaction times. D, Labeling extents of reactions at higher concentrations and increased reaction temperature. Results are from one reaction per condition on the same day from the same batch of protein and HMSiR-IA **22**.

With the current labeling, we could only observe the conjugation of one cysteine per molecule. Based on the LC-MS experiments presented above we cannot exclude that there is a mixture of labeled cysteines in the reaction. We therefore prepared a larger batch for potential imaging experiments and LC-MS/MS analysis to determine the site of labeling. HaloTag7 was labeled with 10 equivalents of HMSiR-IA **22** in PBS at pH 7.9 for 4 h. Excess dye was removed by ultrafiltration resulting in 68% of singly labeled HaloTag7 (Figure 40). A sample was sent for MS/MS analysis at the Functional Genomics Center Zürich (FGCZ). The analysis showed that the main labeling site was indeed C61 (Appendix, Table 20 and Figure 69). The HaloTag7 labeled with HMSiR-IA **22** is referred to as HTIA in the following work. However, as expected from our LC-MS data, the labeling was not complete (Figure 40, B). In a further test experiment, HaloTag7 was labeled with HMSiR-IA **22** and in a separate reaction with IA. After this reaction, we could only observe the mass of HaloTag7 labeled with a single IA without unlabeled or doubly labeled HaloTag7 (Appendix, Scheme 23, A). This result supports that C262 might be inaccessible or unreactive in general and HMSiR-IA **22** works as expected although the additional hydrophobic bulk added by HMSiR lowered the reactivity of the reagent. For improvement of the reagent in general, the linker between the IA moiety and HMSiR could be extended. However, such an approach might decrease the sensitivity of blinkognition, as the dye is presented with a more diverse environment due to the extended radius that it can sense on the protein surface and solvent. Such adaptations remain to be explored. LC-MS data of the final labeling reactions were analyzed using mass calibration in the software MassWorks (Cerno Bioscience) and spectrally accurate modeling of multiply charged ions (SAMMI, Cerno Bioscience) (Figure 40, C).[243–245] The analysis software calibrates a measured spectrum using a calibration filter, yielding a spectrum with a calibrated m/z and known peak-shape function. This instrument and condition-dependent filter is determined from a measured and the theoretical spectrum of known reference ions. The sample components are subsequently determined by multiple charge deconvolution and a multiple linear regression identifying and relatively quantifying the components.[243–245]

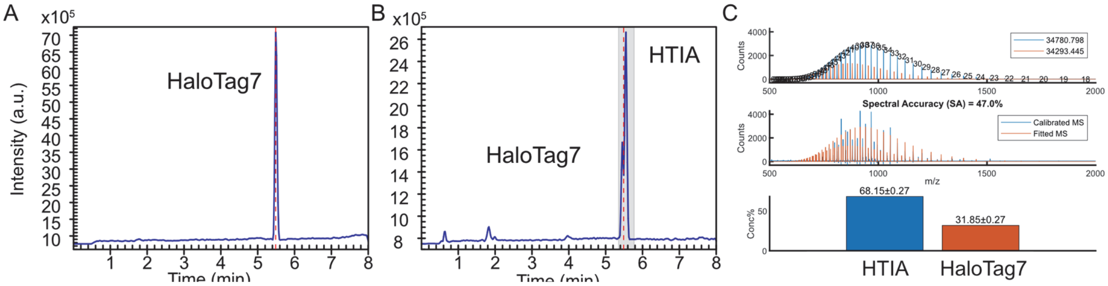


Figure 40. LC-MS analysis of the final labeling of HaloTag7. A, + total ion chromatogram (TIC) trace of the unreacted HaloTag7 from recombinant expression. The red line marks the retention time of the unlabeled protein. B, +TIC after the reaction and purification, with the red line marking the retention time of the unlabeled protein from A. C, Deconvolution results from SAMMI (Cerno Bioscience) of the area marked in grey in B. Top: the theoretically calculated component spectra. Middle: overlay of calibrated mass spectrum and the linear combination of the theoretical predictions of the components. Bottom: composition of the m/z signal from the predicted masses.[243]

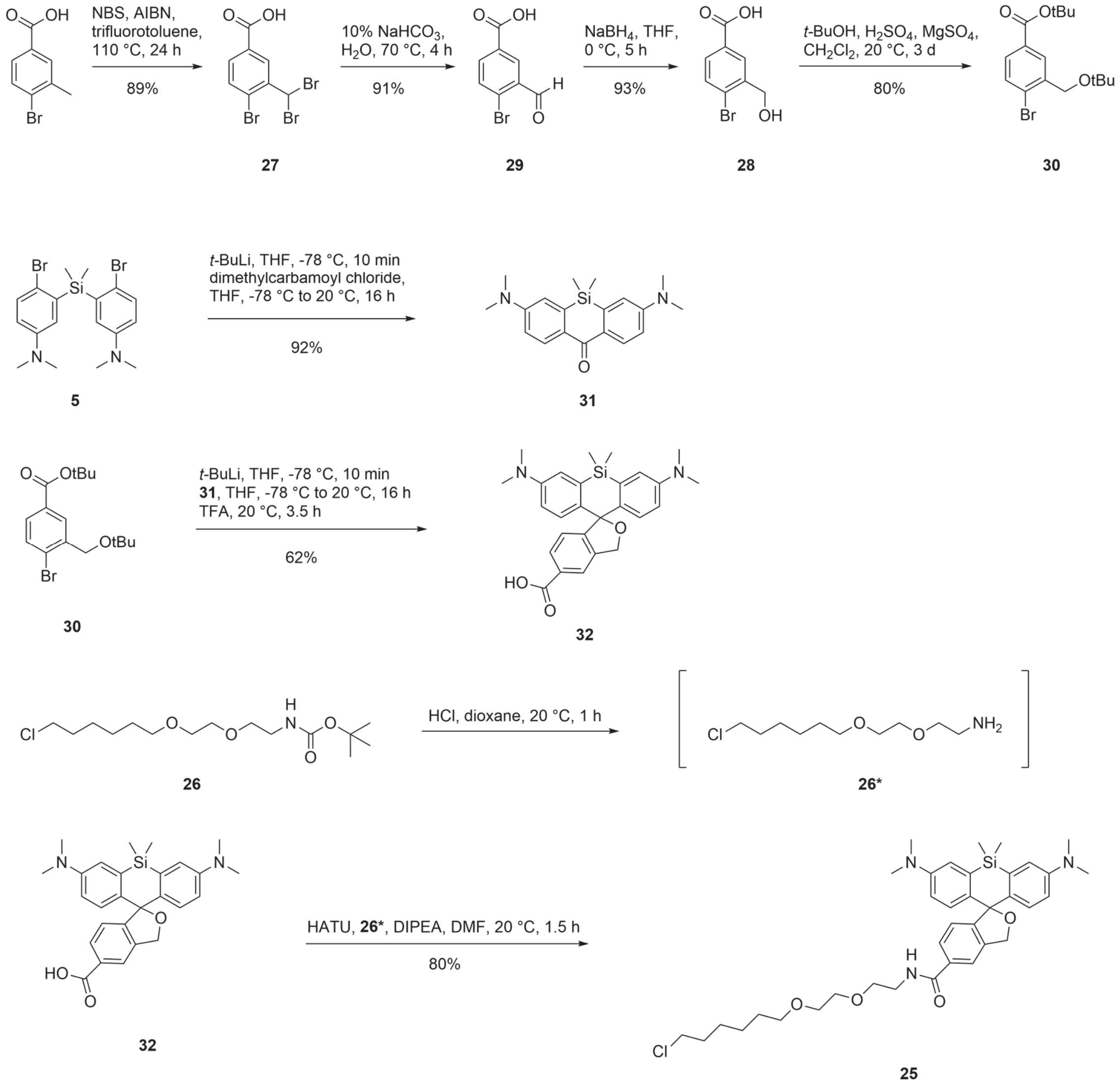
4.2.4 Preparation of proteins for single-label classification

Equipped with a singly labeled HaloTag7 (Figure 40), we proceeded to assemble a protein set for the preliminary examination of blinkognition's capability to discern between proteins labeled with a single spontaneously blinking dye. Incompletely labeled protein samples do not pose a problem in these proof of principle experiments, since they are not fluorescent.

*4.2.4.1 Synthesis of HMSiR-Halo **25** and HaloTag7 labeling via its intrinsic function*

Having labeled HaloTag7 at C61, we can explore the intriguing option of labeling it through its intrinsic reactivity. In the two samples, HTIA and HaloTag7 labeled with HMSiR-Halo 25 (HTHTL), the HMSiR resides in different areas of the protein. If blinkognition can differentiate between these samples, it supports the idea that labeling a protein with two spontaneously blinking fluorophores enhances the information in the fluorescence time trace. Thus, using multiple labels may improve the accuracy and sensitivity of blinkognition.

We synthesized HMSiR-Halo 25 from the commercially available 4-bromo-3-methylbenzoic acid, intermediate 5 and 26 (courtesy of Henriette Lämmermann) (Scheme 12). The synthesis of the fluorophore core is based on published procedures.[83] The radical bromination of 4-bromo-3-methylbenzoic acid yielded the highly brominated intermediate 27. Compound 28 was obtained through hydrolysis of the *gem*-dibromomethylarene 27 followed by reduction of the aldehyde 29. Protection of the alcohol and acid functionality with *tert*-butyl groups under acidic conditions resulted in the lower ring intermediate 30. The di-lithium salt of compound 5 was reacted with dimethylcarbamoyl chloride to give the ketone of the top ring 31. Compound 30 was treated with *t*-BuLi and a solution of the ketone was added to the lithium salt of 30. After an acidic work-up the intermediate was directly deprotected with trifluoroacetic acid (TFA) furnishing the free carboxylic acid of HMSiR 32 with 62% yield. The protected HaloTag ligand 26 was deprotected under acidic conditions and directly used in the amide coupling with 32 to furnish the final probe HMSiR-Halo 25.



Scheme 12. Synthesis of HMSiR-Halo 25 starting from 5, 26, and commercial product 4-bromo-3-methylbenzoic acid. AIBN = 2,2'-azobis(2-methylpropionitrile), DMF = *N*,*N*'-dimethylformamide, HATU = *O*-(7-azabenzotriazol-1-yl)-*N*,*N*,*N*'',*N*''-tetramethyluronium hexafluorophosphate.

After the successful synthesis of HMSiR-Halo 25 we employed it to label HaloTag7 (Scheme 12). After 1 h at 37 °C with 1.5 equivalents of HMSiR-Halo 25 in PBS at pH 7.2 the protein was fully labeled (Figure 41, B).

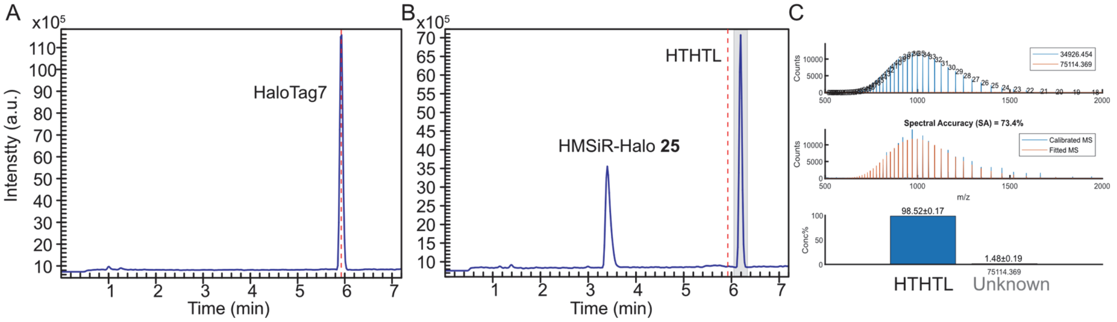


Figure 41. LC-MS analysis of the final labeling of HaloTag7 with HMSiR-Halo 25. A, +TIC trace of the unlabeled HaloTag7 from recombinant expression. The red line marks the retention time of the unlabeled protein. B, +TIC after the reaction before purification, with the red line marking the retention time of the unlabeled protein from A. The remaining unbound 25 is marked. C, Deconvolution results from SAMMI (Cerno Bioscience) of the area marked in grey in B. Top: the theoretically calculated component spectra. Middle: overlay of calibrated mass spectrum and the linear combination of the theoretical predictions of the components. Bottom: composition of the m/z signal from the predicted masses.[243]

With HTHTL we have a first test pair to explore if blinkognition can distinguish two different sites on the same protein. Another challenging task would be to differentiate two proteins with similar reactivity. We assume that the reactive pockets of two proteins that catalyze an analogous chemical reaction will provide comparable chemical environments. Hence the spontaneously blinking fluorophore would experience similar interactions and influences, which would likely result in similar blinking behaviors.

4.2.4.2 SNAP$_f$-tag

Another self-labeling protein that we have mentioned in Chapter 1 is SNAP-tag. SNAP-tag has been evolved by directed evolution from the human DNA repair protein hAGT.[119] hAGT is a suicide enzyme that reverts DNA damage by transferring alkyl groups from *O*^{6}-alkylated guanine bases in DNA to its reactive cysteine residue.[119] Based on this activity, it has been evolved to recognize BG derivatives, for intracellular stability after self-alkylation, and later for improved kinetics leading to a faster version (SNAP$_f$-tag).[119,246] Comparing the labeled SNAP$_f$-tag with HTHTL will allow us to see if proteins of similar reactivity labeled in the reactive site still offer enough variable interactions to render the fluorescence time traces distinguishable. Furthermore, SNAP$_f$-tag offers the fallback of using an HMSiR modified with BG for selective labeling.[83] However, we first attempted to label the reactive cysteine of purified SNAP$_f$-tag (Figure 42, A) selectively over two other cysteines present in SNAP$_f$-tag (Table 20) using HMSiR-IA 22, which are involved in binding a structural zinc(II)-ion.[128] By LC-MS we could only observe singly labeled SNAP$_f$-tag after 2 h at

37 °C using 5 equivalents in PBS at pH 7.4 (Figure 42, B and C). A sample of the labeled protein was submitted for analysis at FGCZ, confirming exclusive labeling at the reactive cysteine (C151, Appendix, Table 20 and Figure 70).

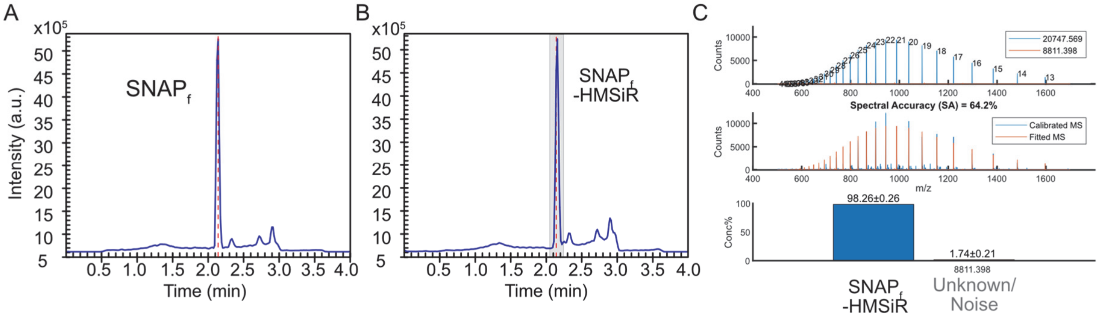


Figure 42. LC-MS analysis of labeling SNAP$_f$-tag with HMSiR-IA **22**. A, +TIC trace of the unlabeled SNAP$_f$-tag from recombinant expression. The red line marks the retention time of the unlabeled protein. B, +TIC after the reaction and purification, with the red line marking the retention time of the unlabeled protein from A. C, Deconvolution results from SAMMI (Cerno Bioscience) of the area marked in grey in B. Top: the theoretically calculated component spectra. Middle: overlay of calibrated mass spectrum and the linear combination of the theoretical predictions of the components. Bottom: composition of the m/z signal from the predicted masses.[243]

4.2.4.3 Single-cysteine glutaredoxin

To enrich our protein set for the initial proof of principle experiment, we sought a third small protein with distinctly different functionality and chemistry. Moreover, the selected protein should be easily expressible in *E. coli* and ideally feature a single cysteine to avoid off-target labeling. Following these criteria, we identified a suitable candidate from the group of glutaredoxins (Grxs).

Grxs are small proteins in the range of 9-15 kDa that are involved in protecting proteins from oxidative damage or restoring their function as part of regulatory processes.[247] Redox homeostasis is highly important for cell functioning and is actively maintained by enzymes. Under oxidative stress, or as a regulatory mechanism, reduced glutathione, a redox buffer in the cell, can form disulfide bonds with cysteines in proteins either to protect them from irreversible damage or modulate protein function. Grxs remove these GSS-protein disulfide bonds and restore the protein function via thiol-disulfide exchange reactions. Most Grxs contain a CXXC motif, however, only one of the two cysteines is known to be involved in these reactions. The C30S mutant from yeast Grx 1 (sCGrx1p), which has the non-participating cysteine exchanged for serine, is still active and the remaining cysteine exhibits a very low pK_a.[247] This mutant has been used to study its mechanism and has been applied in the determination of the redox state in the endoplasmic reticulum (ER) and for the Golgi apparatus (unpublished work from our lab).[247] Considering the small size, the highly reactive cysteine, the promising efficient labeling, and the availability in our lab we decided to include sCGrx1p in our first classification test.[129,247,248]

We labeled recombinantly expressed sCGrx1p (Figure 43, A) with HMSiR-IA **22**. We tested if we could lower the amount of excess dye, as this would facilitate the removal of excess and hydrolyzed **22** after the reaction. Using the standard conditions for 1 h at 37 °C with only 1 equivalent of **22** little labeling was observed. Therefore, we labeled the final batch with 4 equivalents of **22** for 2 h at 37 °C. Excess dye was removed by ultrafiltration yielding a batch of protein with ~91% labeled and 9% unlabeled sCGrx1p (Figure 43, B and C). The labeled protein was submitted for analysis at FGCZ, confirming exclusive labeling at the cysteine residue (C40, Appendix, Table 20 and Figure 71).

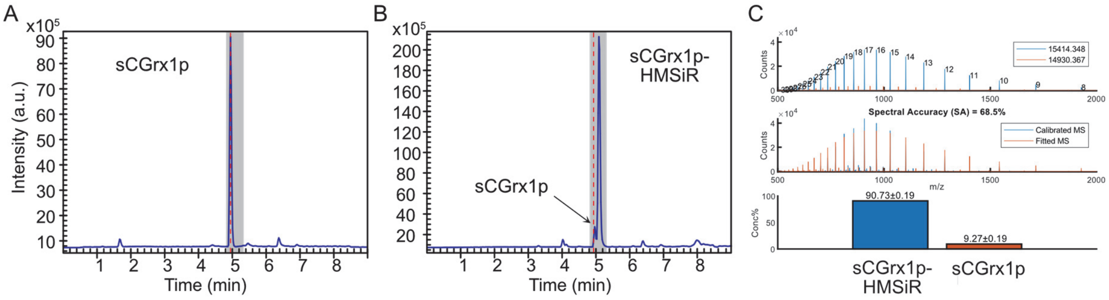


Figure 43. LC-MS analysis of sCGrx1p labeling with HMSiR-IA **22**. A, +TIC trace of the unlabeled sCGrx1p from recombinant expression. The red line marks the retention time of the unlabeled protein. B, +TIC after the reaction and purification, with the red line marking the retention time of the unlabeled protein from A. C, Deconvolution results from SAMMI (Cerno Bioscience) of the area marked in grey in B. Top: the theoretically calculated component spectra. Middle: overlay of calibrated mass spectrum and the linear combination of the theoretical predictions of the components. Bottom: composition of the m/z signal from the predicted masses.[243]

4.3 Encapsulation and surface immobilization

The second challenge, in addition to modifying the protein with spontaneously blinking dyes, involves finding a solution to confine the proteins to the surface, enabling their imaging. The two approaches we considered were encapsulation into liposomes and direct attachment to a residue of the protein. For direct binding, the *N*-terminus would be an attractive option, since all proteins have this reactive moiety, and its chemical properties differ from the amine group of the lysine residues. These properties have extensively been sought to be exploited for specific modification.[249–253] However, the reactions developed for *N*-terminal modification do not reach completion and are not generalizable in that they prefer or tolerate only certain amino acids at the *N*-terminus.[251] Encapsulation offers an alternative approach but also poses certain biases depending on the protein and the lipids employed.[254] However, this approach might benefit from future developments in medical applications of vesicles, or more specifically, lipid nanoparticles used in mRNA vaccines.[255] Liposome vesicles can be conveniently tethered to a surface through the inclusion of lipids with modified head groups, such as a biotin moiety, without a need for further modifying the POI. When paired with a biotinylated surface, the liposomes can be immobilized using streptavidin or its deglycosylated variant, NeutrAvidin. Beyond serving as immobilization handles, lipids can be functionalized with fluorophores, enabling experiments with a known ground truth where different species can be combined with distinct dye-lipid conjugates. With such an experiment it would be possible to empirically verify the prediction made by a ML model. Therefore, we sought to attempt the first classification of proteins with blinkognition via the encapsulation approach.

4.3.1 Workflow of encapsulation

The encapsulation in phospholipid liposomes we tested has been extensively used by other groups for ribozyme encapsulation for single-molecule FRET experiments.[236] The main lipid component is a phosphatidylcholine (initially: 1,2-dimyristoyl-sn-glycero-3-phosphocholine (DMPC)) that is overall uncharged (Figure 44, A). The length of the fatty acid chain and the saturation determines the transition temperature, i.e., the temperature required for the lipid layer to become unstable and change the physical state.[256]

The lipids that were used to form the final liposomes were solubilized in chloroform and the resulting lipid solutions were mixed in the appropriate ratio. After mixing, the solutions of the different lipid components were dried again to form the so-called lipid cake (Figure 44, A and B). In addition to DMPC, we used a biotin-modified lipid, 1,2-dihexadecanoyl-sn-glycero-3-phosphoethanolamine-*N*-(biotinyl) (16:0 biotinyl PE), and a lipid with a fluorescent dye, 1,2-dipalmitoyl-sn-glycero-3-phosphoethanolamine-*N*-(7-nitro-2-1,3-benzoxadiazol-4-yl) (16:0 NBD PE) (Figure 44, A). An aqueous solution containing the substrate, i.e., the POI, was added to the lipid cake and shaken extensively at a temperature of 4 °C above the transition temperature. During this time the vesicles can form and reshape while encapsulating the POI in solution. However, this process generates a large variety of liposome sizes (Figure 45, A). To homogenize the sizes of the vesicles they were extruded multiple times through a filter with a pore size of 100 nm, resulting in a small size distribution around 100 nm (Figure 45, B). The mixture contained empty and occupied vesicles of 100 nm diameter and some freely diffusing protein in the solution.

To remove the free protein, a size-exclusion chromatography (SEC) step was included, where the free protein has a longer retention time, and the clean vesicles could be collected (Figure 44, C and D).

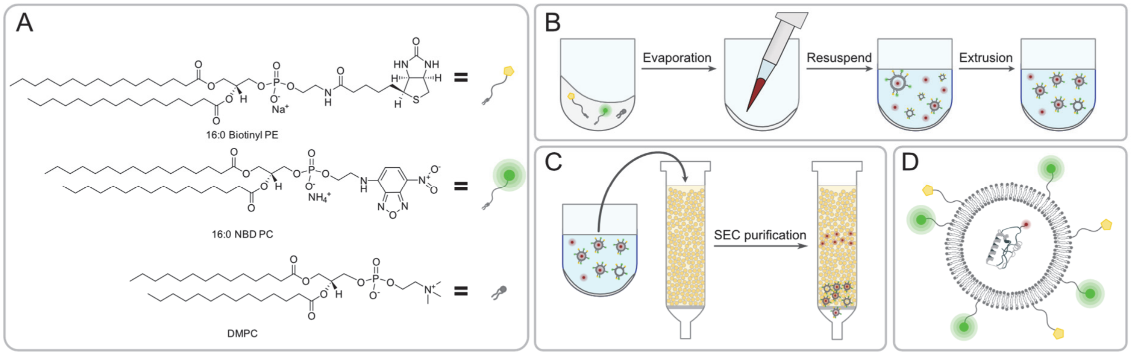


Figure 44. Components and processes involved in the preparation of lipid vesicles. A, Chemical structures of the lipids for vesicle formation. B, Steps involved in the preparation of the lipid vesicles. C, The crude homogenized liposomes are purified via SEC column. D, Simplified schematic structure of a completely assembled lipid vesicle.

4.3.2 Evaluation of SEC columns (collaboration with Dr. Krzysztof Bielec)

We evaluated the resolution of gravity size exclusion chromatography columns. Specifically, we compared the performance of an 8 cm height equilibrated bed of a column with that of a height of 21 cm. Fractions of both columns were collected for unloaded vesicles made from DMPC, a 20 µM solution HaloTag7 labeled with the fluorogenic dye Map555, and a solution of the small molecule dye of fluorescein (100 µM).[124] We determined the fraction in which each sample would elute to simulate the separation of a mixture without cross-talk between the constituents. We collected fractions from the column at 1 mL intervals and analyzed them using dynamic light scattering (DLS) to determine the presence and size distribution by scattering (Figure 45). The labeled protein and small molecules (fluorescein) fractions were analyzed by fluorescence measurements. The samples were measured on different experimental days. After each run, the column was purged with at least two volumes of equilibration buffer (10 mM PB, pH 7.4), then exchanged for 20% EtOH and stored at 4 °C overnight. The experiment showed that the 8 cm SEC column in gravity mode would not suffice to separate vesicles from proteins that escaped encapsulation. In contrast, we could see good separation in the 21 cm column. The vesicles could be detected by thin-layer chromatography (TLC) stained with permanganate solution. This method allows for easy and fast confirmation that the vesicles were indeed in the expected fraction.

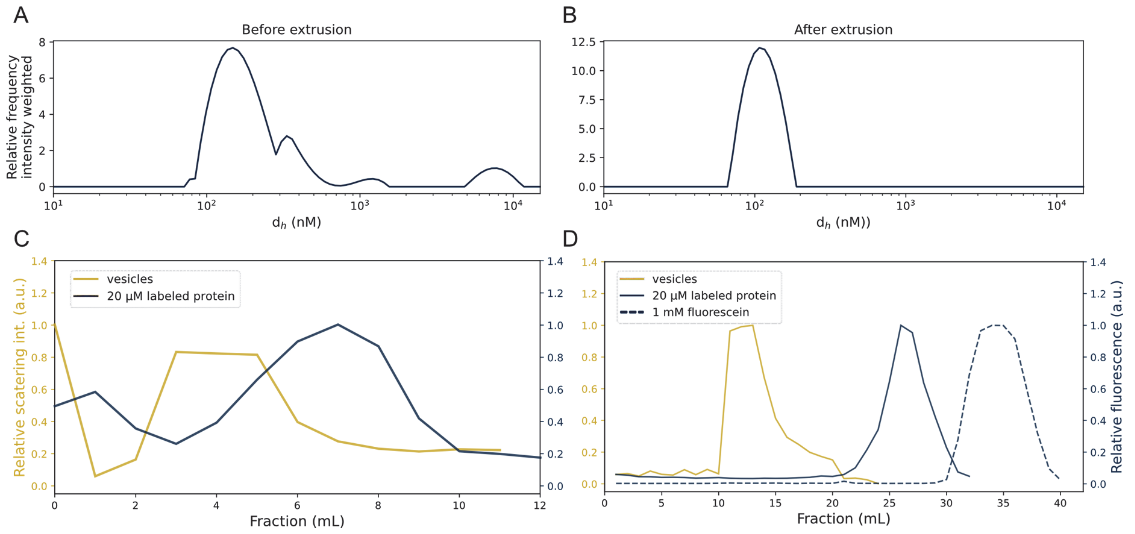


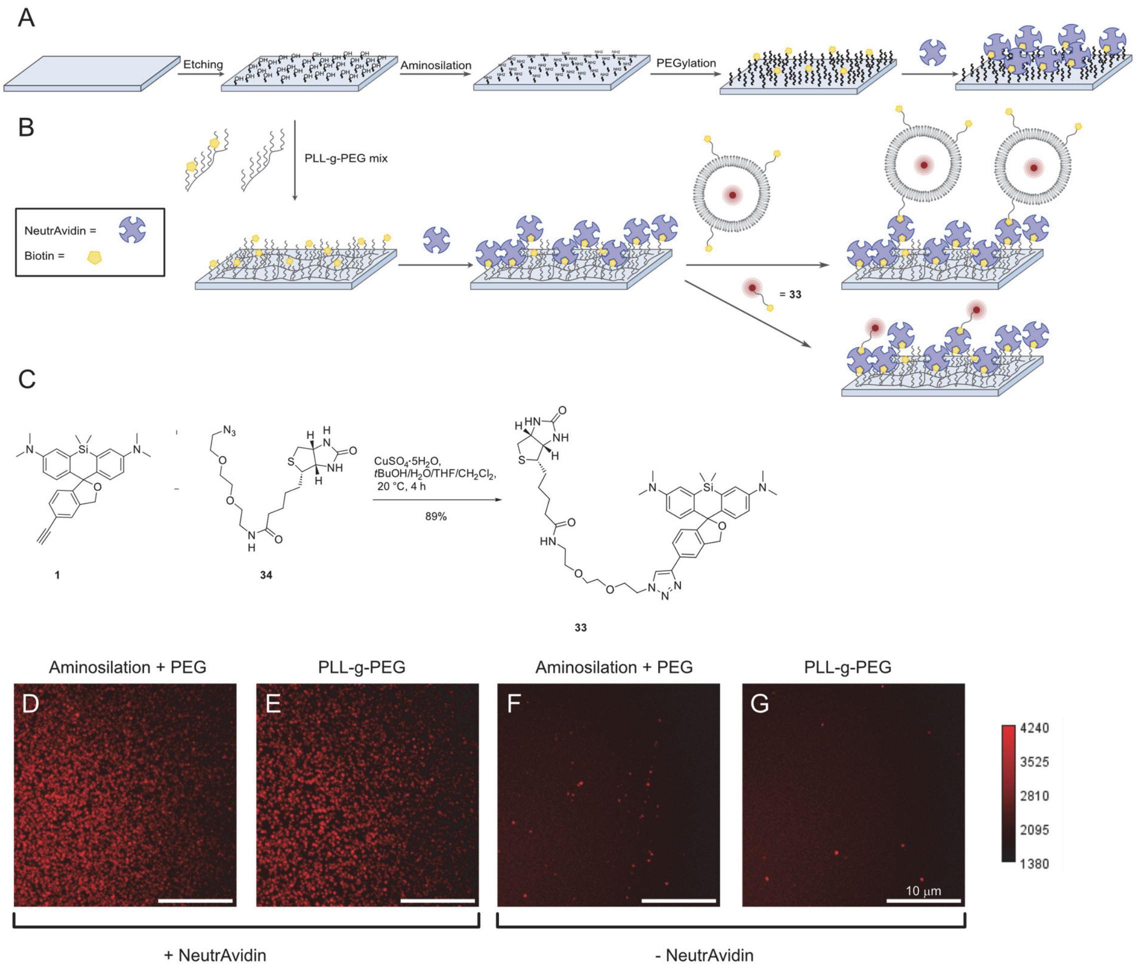
Figure 45. DLS experiments to evaluate vesicle creation and purification. All vesicles were made using DMPC only. A, Vesicles after resuspension and extensive shaking. B, Vesicles from A after extrusion. C and D, Retention time of a fluorescently labeled protein, 20 µM HaloTag7 labeled with Map555-Halo, in comparison to the retention time measured for the vesicles. For the labeled protein the fluorescence at the emission maximum and for the vesicles the scattering was measured respectively. C, A short (8 cm) SEC column was used. D, A long (21 cm) SEC column was used.

Based on these results all samples were purified by a 21 cm SEC column and the fractions starting from 9 to 14 were collected and screened for the presence of the vesicle by TLC, prior to imaging experiments.

4.3.3 Changes to the surface modification

In Chapter 2 we adapted an established method for the functionalization of glass surfaces.[257,258] In our hands, this method yielded unreliable results regarding the surface fouling properties with noise signals of unknown origin. Even after many optimization experiments (Chapter 2), it was unclear which step caused these signals or rather how we could prevent them. In the original procedure, aminosilation is a crucial step to install surface amine functionality to attached (functionalized) PEG chains (Scheme 13, A). This step has been extensively researched and many have tried to optimize the procedure.[257,259–261] However, the formation of a monolayer is intrinsically difficult to control due to multiple potential reactions that could occur other than the attack of the surface hydroxyl groups on the siloxane.[257,260,261] The reaction is highly sensitive to the amount of water in the reaction, which is difficult to control in a standard open laboratory setting.[257,262] Considering the tediousness of the procedure and the large number of steps, we sought alternative ways of functionalizing a glass surface. Another surface modification principle that has been reported for single-molecule TIRF microscopy is the electrostatic adsorption of a mixture of graft poly-L-lysine PEG co-polymer (PLL-*g*-PEG) mixed with biotinylated PLL-*g*-PEG in low amounts (Scheme 13, B).[263–265] Following the adsorption of PLL-*g*-PEG onto the treated glass surface, the biotins are bound to NeutrAvidin. Any molecule of interest connected to biotin can be attached to the free NeutrAvidin such as HMSiR-biotin **33**. A change to the PLL-*g*-PEG protocol would lower the number of steps after cleaning the glass surface and remove time-consuming and sensitive steps from the current surface functionalization protocol.

In order to conduct some initial tests, we prepared HMSiR-biotin **33** from HMSiR-alkyne **1** and azide-modified biotin **34** via a CuAAC (Scheme 13, C). In a test experiment, we compared the two surface functionalization protocols, while exchanging the previously employed heterobifunctional NHS-PEG$_{5000}$-azide with NHS-PEG$_{5000}$-biotin (Scheme 13, A). The glass slides (round, ø 25 mm) were cleaned with detergent, and solvent, and etched by UV/ozone treatment. Then a solution of biotinylated PLL-*g*-PEG (150 µL, 0.1 mg mL^{-1}) was applied for 30 min before rinsing it off with MiliQ and 10 mM PB at pH 7.4. The positive controls were treated with a solution of NeutrAvidin (0.1 mg mL^{-1}, ThermoFisher Scientific) for 30 min before rinsing with MiliQ and 10 mM PB at pH 7.4. Following NeutrAvidin binding, all samples were incubated for 30 min with a 1 nM solution of HMSiR-Biotin in 10 mM PB at pH 7.4. In the negative control samples, the NeutrAvidin binding step was left out, hence no conjugation is possible and all observed signals must stem from noise or non-specifically adsorbed molecules of HMSiR-biotin **33** (Scheme 13, F and G). For each sample, movies of the spontaneously blinking dye on the surface were acquired. The experiment successfully demonstrated that with both protocols, functionalized surfaces with immobile spots can be observed (Scheme 13, D and E). Concerning the anti-fouling properties, the PLL-*g*-PEG surface displayed similar or fewer signals in the negative control (Scheme 13, F and G). These promising results and the considerably fewer steps and shorter preparation time led us to adapt this surface modification agent for future experiments.



Scheme 13. Adaptation of the surface functionalization protocol. A, Protocol applied in the previous project, including an aminosilation and PEGylation step. B, Protocol using the PLL-*g*-PEG reagent. C, Synthesis of the HMSiR-biotin **33**, from **1** and **34**. D-G, Maximum projections of the movies of HMSiR-biotin **33** on differently prepared surfaces (230x230 pixels, 6000 frames, 638 nm, 30 ms, 90%). D and F use surface functionalization via A (aminosilation and PEGylation) whereas E and G employ B (PLL-*g*-PEG). D and E, Positive controls include the NeutrAvidin binding step. F and G, Negative controls without the addition of NeutrAvidin. The color bar represents the fluorescence intensity (a.u.). Scale bar = 10 µm.

4.3.4 Lipid composition (collaboration with Dr. Krzysztof Bielec)

To test whether we were able to encapsulate proteins we used 100 µL of 1 µM sCGrx1p labeled with HMSiR-IA **22** to resuspend a lipid cake containing DMPC, the fluorescently labeled lipid 16:0 NBD PE and 16:0 biotin PC for surface binding. After extrusion, the solution was directly applied to a biotinylated PLL-*g*-PEG cover slip. After removing the excess vesicle solution, we observed the fluorescence of the NBD dye to visualize the vesicles and the fluorescence of HMSiR to check the labeled protein (Figure 46, A, 488 nm and 640 nm, left). However, we could not observe labeled protein contained in vesicles (Figure 46, left). We hypothesize that the reason might be the low transition temperature of DMPC (T_t = 24 °C) which might allow proteins to be released from the vesicles. Furthermore, a temperature close to the transition temperature of the lipid allows for further flexibility and for the vesicles to fuse and reorganize, which might in turn explain the bigger vesicles that could be observed.[256] In follow-up experiments, we exchanged DMPC for a phospholipid with a higher melting temperature 1,2-dihexadecanoyl-sn-glycero-3-phosphocholine (DPPC, T_t = 41 °C). We resuspended the corresponding lipid cake with a solution of sCGrx1p labeled with HMSiR-IA **22** (100 µL, 3 µM) in 10 mM PB at pH 7.4 and purified the sample by SEC column after extrusion. Introducing this change, we were able to observe signals overlapping in the channel of the lipid and the labeled protein (Figure 46, A, right).

To confirm that we indeed have encapsulated labeled proteins in the vesicles, we extracted fluorescent time traces in the red channel (HMSiR-IA **22**) of vesicles loaded with HTHTL (100 µL, 1 µM) for visualization. However, the signals did not show classical single-molecule characteristics but rather clustered signals with many different intensity levels or random spikes (Figure 46, B, top). We reasoned that the temperature required during the preparation of vesicles with DPPC (45 °C) was too high and would be incompatible with most proteins.[266] Consequently, we explored two other commercially available lipids 1-octadecanoyl-2-tetradecanoyl-sn-glycero-3-phosphocholine (SMPC) and 1-octadecanoyl-2-tetradecanoyl-sn-glycero-3-phosphocholine (15:0 PC) with a transition temperature of 30 °C and 35 °C respectively. Both lipids furnished stable liposomes (Appendix, Figure 72) with fewer such artifacts (Figure 46, B, bottom and middle). Given the temperature sensitivity of proteins and the aim to minimize damage during encapsulation, we opted to employ SMPC to acquire our dataset.

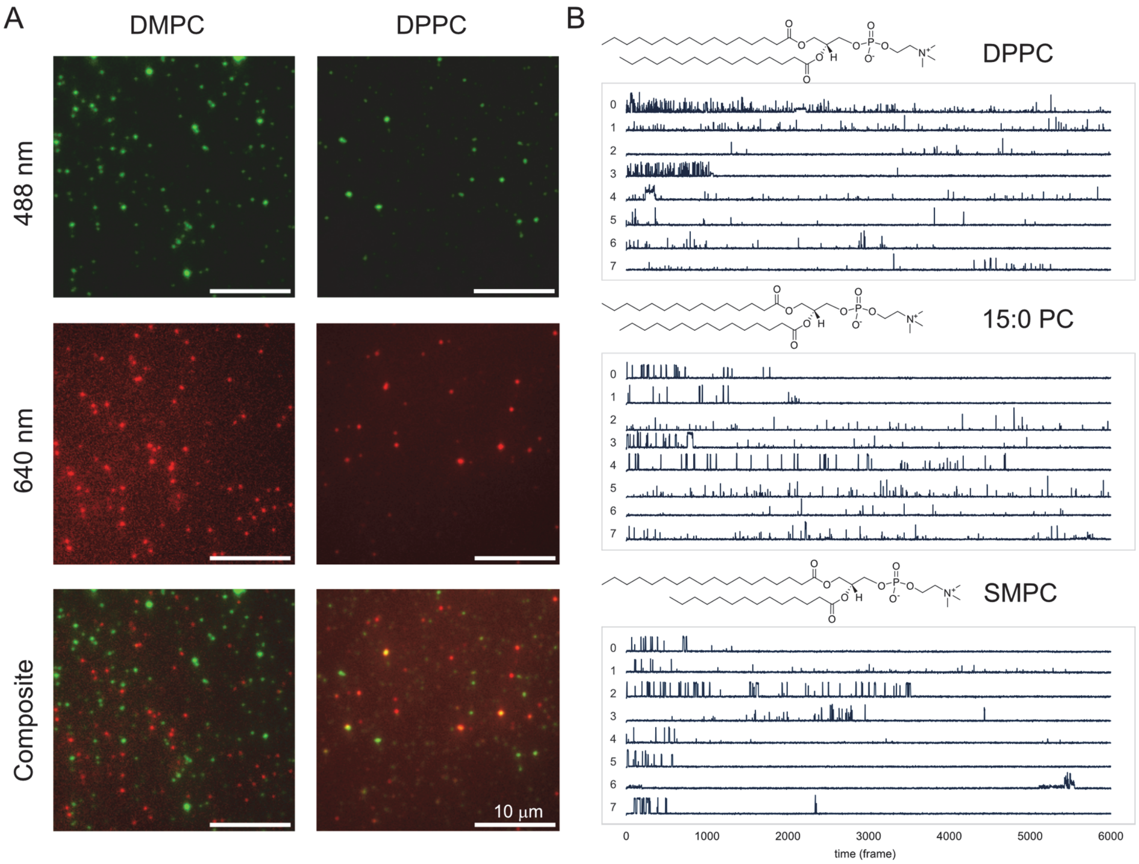


Figure 46. Experimental results for the determination of the ideal lipid for liposome formation. A, Images of vesicles made from the corresponding lipid, 16:0 biotinyl PE and 16:0 NBD. NBD dye indicating the vesicle was imaged with a 488 laser (30 ms, 2 mW for DMPC 60 ms, 2 mW for DPPC) and HMSiR on the labeled protein cargo with a 638 nm laser (60 ms, 90 mW for DMPC and 30 ms, 90 mW for DPPC). For DPPC a 230x230 pixels box from a 512x512 pixels image is shown. B, Extracted sample traces of signals in the red channel of liposomes loaded with HTHTL. The liposomes are composed from the lipids DPPC (T_t = 41 °C), 15:0 PC (T_t = 35 °C), and SMPC (T_t = 30°C). Scale bar = 10 µm.

4.4 Data set acquisition and classification of proteins

With the labeled protein samples and the lipid formation for surface functionalization, we were ready to obtain initial insights on key questions for blinkognition on proteins (Figure 47). When comparing the two protein samples HTHTL and HTIA we can observe whether blinkognition has the sensitivity to distinguish fluorescence traces that have been obtained from the same protein but that carry the fluorophore at different sites and were exposed to different environments given by the protein (Figure 47, A). The comparison of HTHTL and SNAP$_f$, both labeled at the protein active site, will provide insight into whether an environment that provides a similar reactivity is still differentiable using our approach (Figure 47, B). Lastly, when adding sCGrx1 to the mix, we increase the diversity by including another protein class which allows us to observe if this additional complexity will decrease the predictability (Figure 47, C).

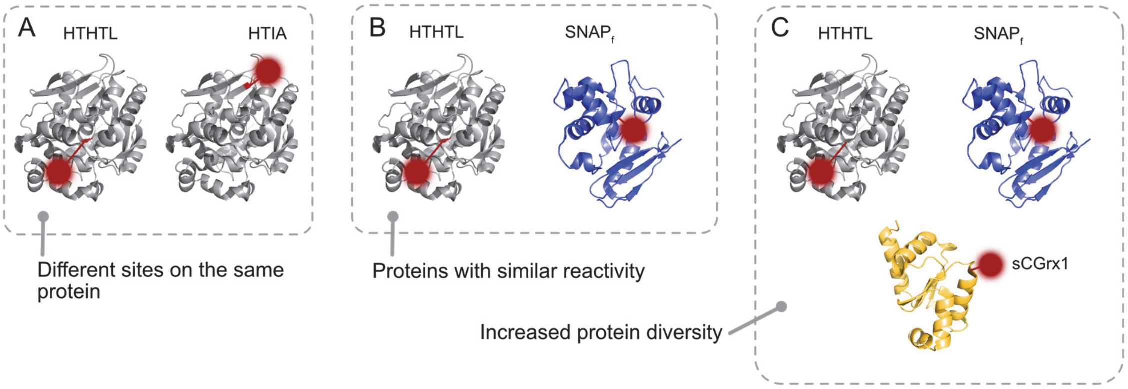

A
HTHTL
HTIA
Different sites on the same protein
B
HTHTL
SNAPf
Proteins with similar reactivity
C
HTHTL
SNAPf
sCGrx1
Increased protein diversity

Figure 47. Scenarios to be tested using protein blinkognition. The views were generated from published structures using PyMOL.[128–130]

4.4.1 Sample preparation and acquisition

We prepared vesicles for all protein samples HTIA, HTHTL, HMSiR-IA **22** labeled sCGrx1, and SNAP$_f$ on at least three different days, to decrease a potential bias through experimental noise. Parallelization was limited by the extrusion and the available SEC columns. Therefore, two proteins were always run in parallel and in two rounds of SEC purification per day. The order of sample preparation was alternated to avoid any potential systematic experimental biases. The lipid cakes prepared from SMPC, 16:0 biotinyl PE, and 16:0 NBD were prepared ahead of the experiments and stored at –20 °C. The lipid cakes were resuspended with the respective protein solution (100 µL, 2 µM) and prepared at 34 °C. After extrusion, the sample was stored protected from light at 20 °C until the second sample was ready and two SEC columns were run separately. The relevant fraction was determined by TLC and permanganate stain. The collected fraction was diluted 100-fold before applying it to the functionalized cover slip. Coverslips were prepared on the same day with a mixture of 1:50 biotin PLL-*g*-PEG and unfunctionalized PLL-*g*-PEG at 0.1 mg mL^{-1}. After 15 min of incubation, the density of the vesicles was observed in the green channel (vesicles, 488 nm) the solution was removed and the slide was washed five times without completely removing the solution to avoid drying the vesicles. The samples were imaged after preparation for 10 frames in the vesicle channels (488 nm, 30 ms, 6 mW) and 6000 frames in the channel of the labeled protein (638 nm, 30 ms, 90 mW).

4.4.2 Analysis of the protein datasets

4.4.2.1 HTHTL vs HTIA

The movies were analyzed with the sample pipeline and parameters used in Chapter 2.[91] In brief, single molecules were localized in the 638 nm channel using the program Picasso (Chapter 2, Appendix Table 10) followed by custom scripts to extract the time traces.[175] The extracted time traces were filtered using the same criteria as described previously (Chapter 2, Appendix: Table 11) and augmented by mirroring the peak-containing area of each trace.[91] The current implementation does not include a comparison of vesicle signals and protein signals. In the future, only protein signals that are associated with a signal in the vesicle channel will be included in data sets. This approach could help to remove noise signals and ensure homogeneous conditions for the proteins. We employed the same model architecture as in Chapter 3 (Appendix: Table 19), a 1D-CNN-GRU-MCD (Figure 48, A and Figure 73). To mitigate problems due to the long input sequence, which can result in disappearing gradients, we tested classification performance with shortened traces on HTHTL and HTIA (Figure 48). We observed increasing accuracies when training models with traces from 2000 to 4000 frames. However, when using more than 4000 frames the accuracy does not increase (Figure 48, A). In addition, we can observe that the number of traces that contain their last peak (a signal with an intensity >4 σ above the mean) after a certain number of frames decreases to a minimum of around 4000 to 5000 frames before it increases again (Figure 48, C and D). However, we hypothesize that this observation is due to sporadic signals toward the end of traces caused by contamination or noise and hence should not contribute to a compound-specific signal. In conclusion, we decided to train the following models on traces shortened to 4000 frames to capture relevant signals while avoiding noise contribution or experimental bias.

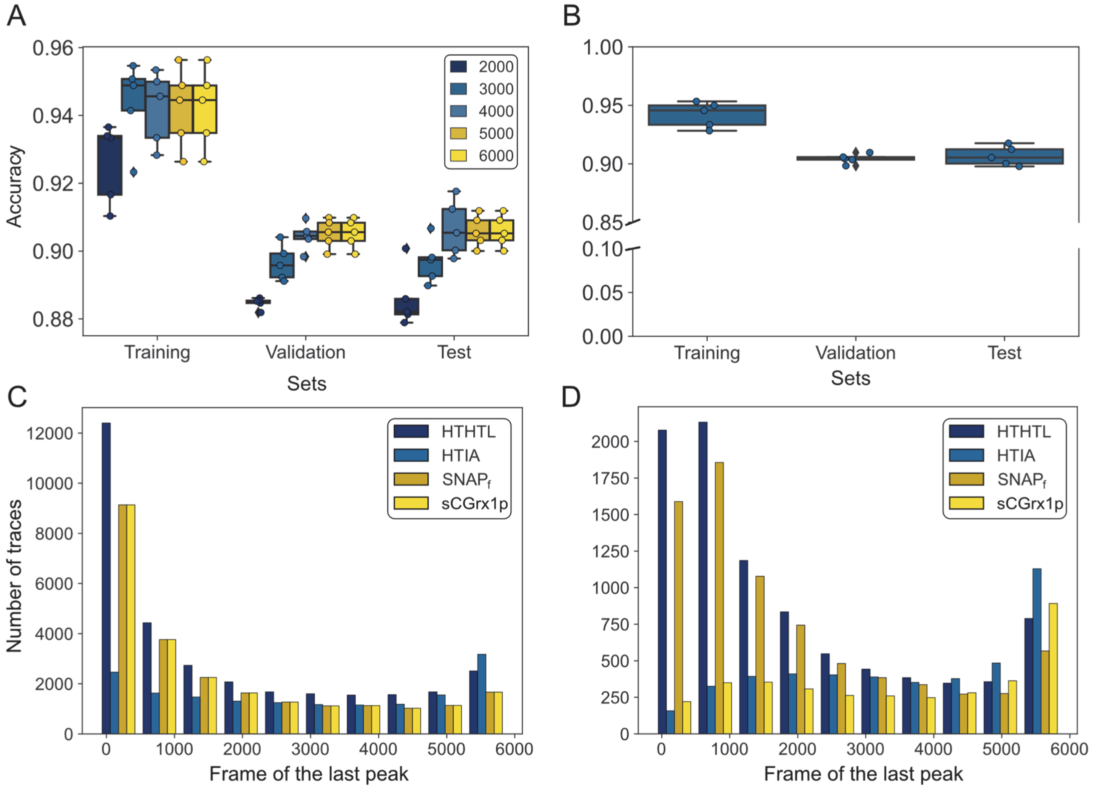


Figure 48. Influence of trace length on the classification of HTHTL vs HTIA. A, The traces were cut to the respective length prior to training and evaluated using five-fold CV. B, The accuracies obtained in five-fold CV results of HTHTL bs HTIA with traces 4000 frames long. C, The distribution of the frame number of the last peak of all extracted traces without filtering. D, The distribution of the frame number of the last peak of all extracted traces post filtering (Appendix: Table 11). A peak threshold is defined as at least four standard deviations above the mean of a trace. For A and B, The middle line represents the median of all the values and the whiskers represent the quartiles without outliers as determined in the Seaborn implementation.[204]

The evaluation of the comparison of HTHTL and HTIA using 4000 frames looked very promising with an overall accuracy of ~89.7% before filtering, which could be improved to ~96.3% with a CT of 0.7 (Figure 49, A and B). However, in this case, the CT could be increased without excessive loss of traces (Figure 49, C). This decreased loss indicates that the two labeling variants are distinguishable with high confidence, indicating that indeed different sites on the same protein provide different environments resulting in distinct blinking patterns. However, HaloTag poses a special case as it has been designed to bind rhodamines. Therefore, the favored open form of rhodamines might be part of the reason for the observed relatively large population of traces that exhibit early bleaching (Figure 48, D).[267] Further evidence and data exploration is required to support such an argument. Nevertheless, these results support our hypothesis that labeling a protein with multiple dyes provides more information about the protein and hence could support the protein identification. This effect could further be supported by applying two spontaneously blinking dyes of different colors, as the information is linked to a site, in addition to easier differentiation from noise.

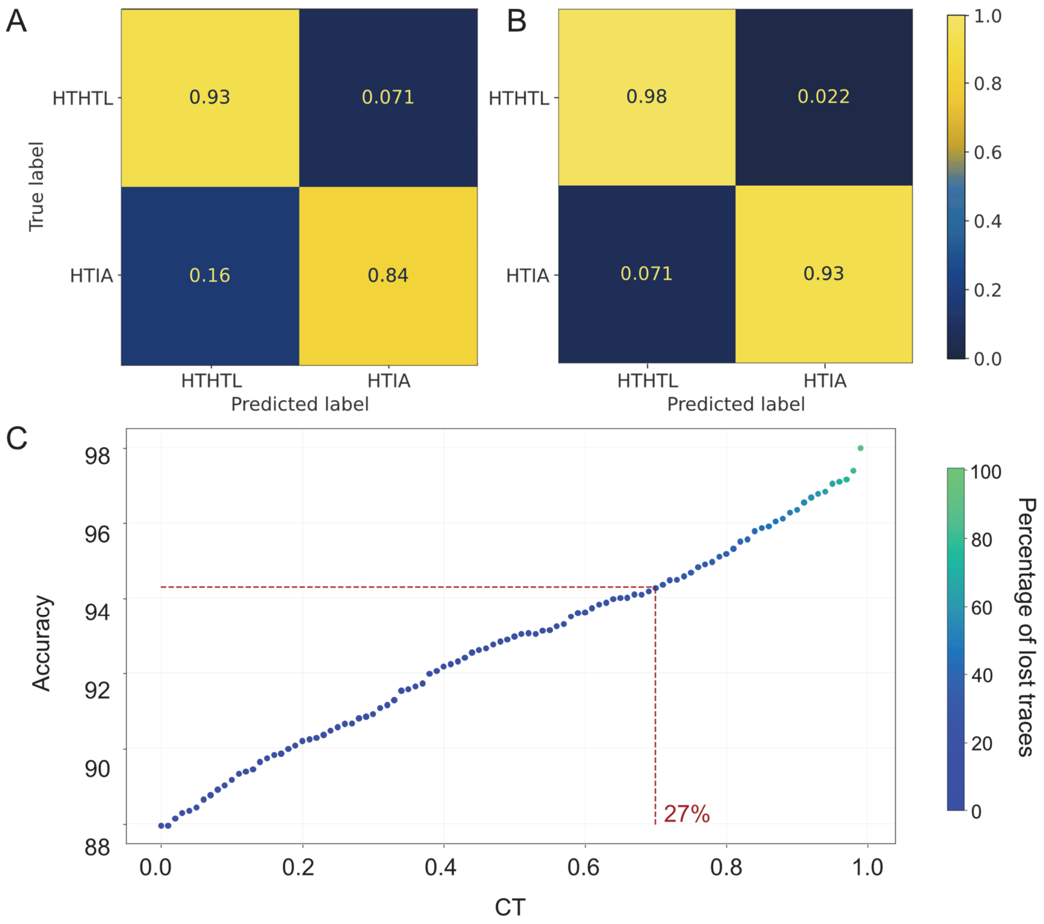


Figure 49. The results of the model with a 1D-CNN-GRU-MCD architecture for the two-class analysis of HTHTL and HTIA. A, Classification result before any filtering. B, Accuracies obtained after filtering with a CT of 0.7 (see C). C, The relation of a more stringent CT vs the obtained accuracy and the lost traces are indicated by the color code, the red number shows the percentage of lost traces at CT 0.7. The label is determined by evaluating 100 predictions.

4.4.2.2 HTHTL vs SNAP$_f$

The classification results of the two self-labeling proteins that were modified in their active pocket, present itself as a more difficult problem with lower accuracies of ~69.7% before and ~96.5% after filtering with a CT of 0.7 at a loss of 70.2% of all traces in the test set (Figure 50). Although unintuitive at first glance, the two self-labeling proteins have been engineered to react with rhodamine dyes, which could explain a similar chemical environment in their active site. For a future experiment to support this idea, labeling of SNAP$_f$ at another residue could be attempted, by first blocking the highly reactive cysteine with a BG-derivative through its enzymatic activity or a reactive moiety such as IA, followed by targeting one of the other cysteines on the protein. If comparing the two differently labeled SNAP$_f$ proteins would also result in higher accuracy and a lower trace loss upon filtering than the SNAP$_f$-HTHTL classification problem, it would support our argument.

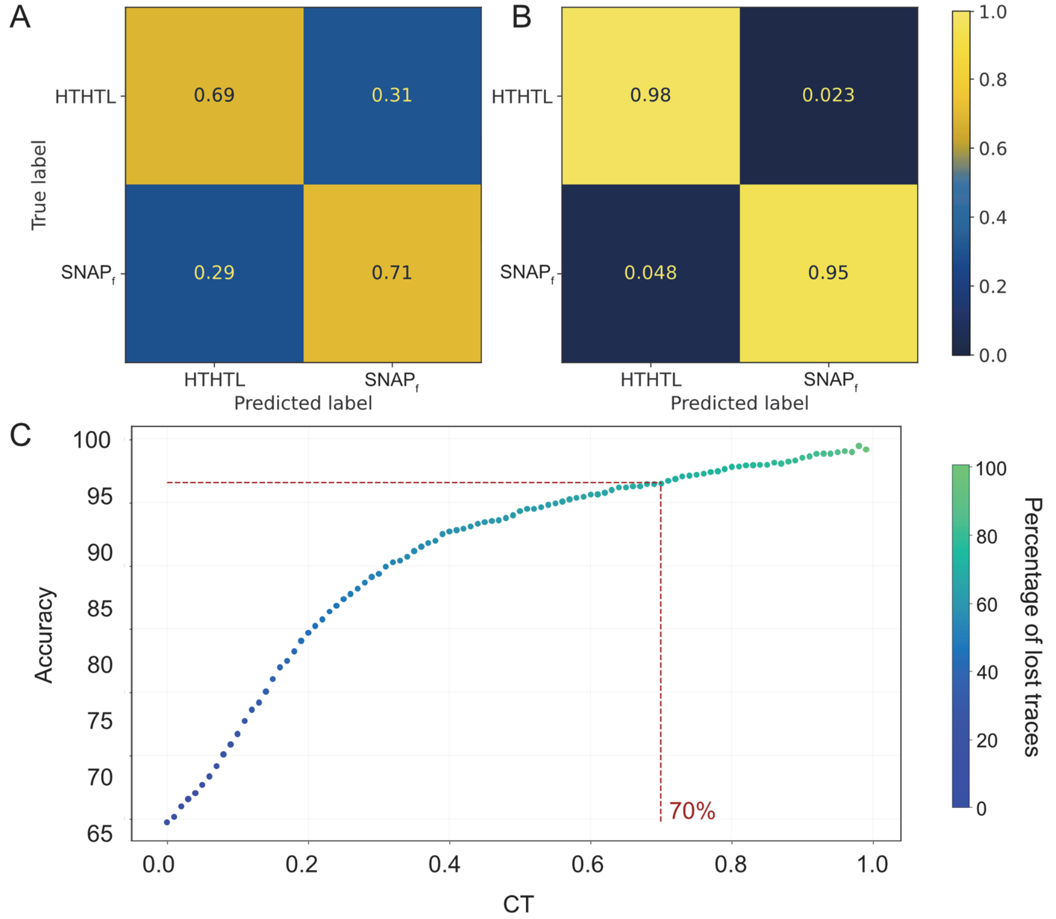


Figure 50. The results of the model with a 1D-CNN-GRU-MCD architecture for the two-class analysis of HTHTL and SNAP$_f$. A, Classification result before any filtering. B, Accuracies obtained after filtering with a CT of 0.7 (see C). C, The relation of a more stringent CT vs the obtained accuracy and the lost traces are indicated by the color code, the red number shows the percentage of lost traces at CT 0.7. The label is determined by evaluating 100 predictions.

4.4.2.3 HTHTL, SNAP$_f$, and sCGrx1p

Finally, we compared three proteins by adding labeled sCGRx1p to the more difficult case: HTHTL and SNAP$_f$. The classification accuracies were comparable to the previous problem with ~65.4% that could be improved to ~93.0% with a CT of 0.7 and a loss of ~72.6% of all traces in the test set (Figure 51). Therefore, the addition of a third protein seemed to have decreased accuracies only slightly and when comparing the CT vs accuracy curve it is less upward convex than the two-protein case (Figure 51, C and Figure 50, C). These observations indicate that blinkognition can be used for more proteins than just two or three. However, the performance in the current configuration will depend on a case-by-case basis, as we could observe in the previous two cases. Yet, it remains to be explored on a much larger dataset, in terms of diversity, to which extent blinkognition can be multiplexed. Our initial results indicate that the developed method and the model architecture can be extended to targeted protein analysis on clean samples.

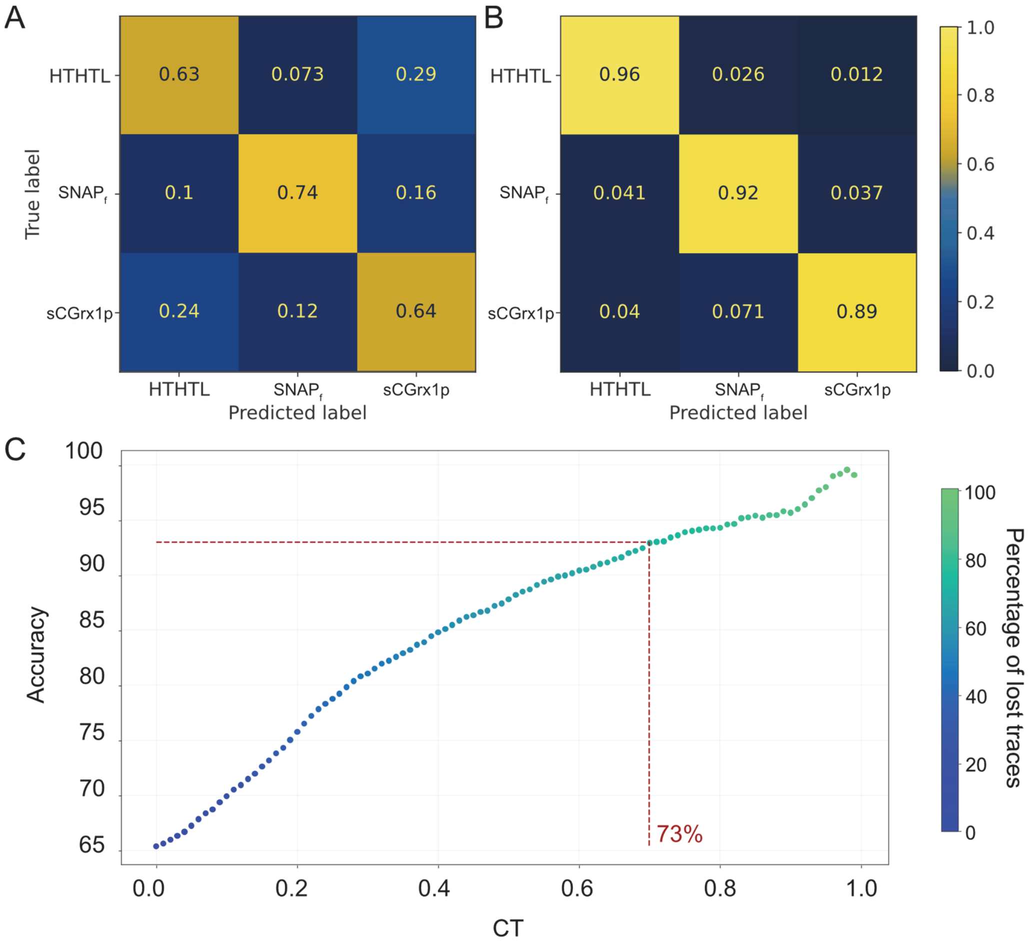


Figure 51. The results of the model with a 1D-CNN-GRU-MCD architecture for the two-class analysis of HTHT, SNAP$_f$, sCGrx1. A, Classification result before any filtering. B, Accuracies obtained after filtering with a CT of 0.7 (see C). C, The relation of a more stringent CT vs the

obtained accuracy and the lost traces are indicated by the color code, the red number shows the percentage of lost traces at CT 0.7. The label is determined by evaluating 100 predictions.

4.5 Conclusions and outlook

In this chapter, we have presented the work on the progress toward analyzing proteins with blinkognition. To this end, we synthesized an IA-modified version of HMSiR with reactivity toward cysteines. With this dye, we conducted labeling experiments and prepared three labeled proteins: HTIA, SNAP$_f$, and sCGrx1. Further, we prepared HTHTL using the enzyme's self-labeling action with HMSiR-Halo 25. In addition, we introduced protein immobilization to the glass surface via encapsulation into liposomes. The lipids carry modifications such as a biotin handle and fluorescent dyes by adding the respectively modified lipid to the liposome preparation. After acquiring datasets for all proteins via the vesicle protocol, we conducted a preliminary classification test using the same model architecture applied previously to peptide classification (Chapter 3).

The results presented herein indicated that blinkognition can be applied to proteins. The classifier was able to differentiate between labeling at different sites of the same protein (HaloTag7) readily. However, when two different proteins with similar reactivity labeled at their active side were compared, the accuracy was reduced. These observations support blinkognition's concept that blinking relies on the microenvironment and available interactions, which are likely to be similar for proteins that promote similar reactivity. When we added sCGrx1 to the comparison of the two self-labeling proteins the accuracy remained in a similar range. Hence, based on the current observations, adding more proteins does not make the classification more difficult by default but it should depend on the "similarity of the environment". However, further investigations and potentially systematic testing of proteins could establish a correlation between protein property descriptors and predictability with blinkognition.[268] A potential series of experiments could be set up by choosing a protein with a single cysteine (for ease of labeling with HMSiR 22). In this protein, amino acids in spatial proximity to the labeled cysteine could be mutated systematically and hence modulate the environment in a controlled manner. For example, the classification accuracy could be compared when an amino acid changes the polarity e.g. a valine-to-isoleucine mutation vs. a valine-to-aspartic acid mutation. Additionally, mutations could be tested at defined distances from the labeling site. In such a setting the experimental accuracies could be compared to the calculated surface properties of the mutants. Such a causal relationship could help predict the scope of the application of blinkognition.

The classification results of HTHTL and HTIA imply that our initial proposal of attaching multiple dyes to the same protein could be beneficial when comparing proteins. Multiple labels would allow the integration of information on multiple sites on the protein. Labeling multiple sites can be achieved with the same fluorophore at multiple occurrences of the reactive amino acids or by adding a fluorophore with a different color and residue reactivity. Using multiple fluorophores can provide information about the interaction of the dyes with their microenvironment, the presence of exposed reactive amino acids, and potentially the distances and dynamic changes through (homo-) FRET, if appropriate dyes are used. The signal of each color channel used will contain the integrated information of all these processes and should therefore be more specific to the protein and improve the classification.

However, it is impossible to predict the expected improvement theoretically as it will depend on the similarity of the proteins or rather the two sites and the resulting combined signal. In addition, important practical considerations limit this approach. Every additional fluorophore labeling site adds possible labeling variants due to incomplete labeling. Hence the additional information is counteracted by increasing the variability of the sample. Furthermore, protein solubility is reduced by adding many hydrophobic fluorophores causing sample loss and changes in the sample composition. Therefore, we hypothesize that only a limited number of fluorophores will prove beneficial. Experimental studies are necessary to determine the most effective implementation of blinkognition for specific and general use cases. A first model system for the influence of incomplete labeling could be constructed by mixing HTIA and HTIA additionally labeled with HMSiR-Halo **25**. In this case, HTIA would mimic partially labeled data.

To optimize the multi-fluorophore approach, new dyes are needed to label alternative residues and improve the labeling observed using HMSiR **22**. So far we could not detect a double-labeled HaloTag7 variant in our MS analysis, even though the second cysteine appears similarly accessible for the solvent based on the crystal structure (see 4.2.3, Figure 38, C) as the currently labeled cysteine. A potential cause could be protein precipitation upon labeling with HMSiR at this site, which could be reduced by increasing the solubility (in aqueous media) of the current reagent **22**.[269] The reactivity of **22** could potentially be improved by elongating the linker between the benzyl alcohol and the IA moiety, e.g. via PEG units. Multicolor labeling of a single protein requires another residue-specific reaction. After cysteine, the next commonly targeted amino acid is lysine with the most promising reactive groups being sulfotetrafluorophenyl (STP) ester or an NHS ester, with a preference for STP due to increased stability.[239,270]

The ongoing work has preliminarily shown promising results for protein blinkognition and offers many follow-up investigations and further developments. An important parameter to add to the analysis in the future is the prerequisite of any localization to co-localize with a vesicle signal. This filtering could substantially contribute to removing noise from our datasets as the chances that a noise signal occurs in the same location are low. Furthermore, we can be more certain that we only compare signals from similar experimental conditions.

The herein-implemented vesicle approach opens the door to validate the performance of blinkognition with experimental ground truth. Toward this aim, two proteins will be encapsulated in lipid vesicles harboring lipids labeled with fluorophores of a specific color each. Training sets could then be acquired for the proteins separately in the respective colored vesicle. After successful model training the trained model can also be used to evaluate traces obtained from experiments with mixed vesicle samples as the label for evaluation can be obtained from the color of the encapsulating vesicle. Such experiments will be invaluable in confirming the usefulness of blinkognition. We have conducted a preliminary experiment to ensure that proteins cannot be exchanged between vesicles (Figure 52). In this setup, unlabeled HaloTag7 was encapsulated in vesicles with a blue labeled lipid, 1,2-dipalmitoyl-sn-glyero-3-phosphoethanolamine (DPPE)-Atto425 (Figure 52, A in blue), to mimic relevant experimental conditions where both vesicles are partially filled. HTHTL was contained in green vesicles DPPE-Atto520 (Figure 52, A in cyan). The two vesicle solutions were mixed before addition to the coverslip under standard conditions and ten frames were acquired in the vesicle channels and 8000 frames were acquired in the red channel (638 nm) of HMSiR. When overlaying the maximum projections of the three channels we can observe very few red signals that are either not overlapping with any vesicle or with the green vesicles as expected. We hypothesize that the free red signal corresponds to noise, which is supported by manually extracted time traces (average intensity in a 5x5 pixel box) (Figure 52, B in magenta). In contrast, the signals that are associated with the green vesicle correspond to a single-molecule-like signal (Figure 52, B in green). However, the number of red signals, i.e., the labeled protein, was exceptionally low, and we need more signals to confirm these results. Nevertheless, to date, we have not observed red signals associated with blue vesicles nor vesicles that show both colors which would indicate lipid exchange or vesicle fusion.

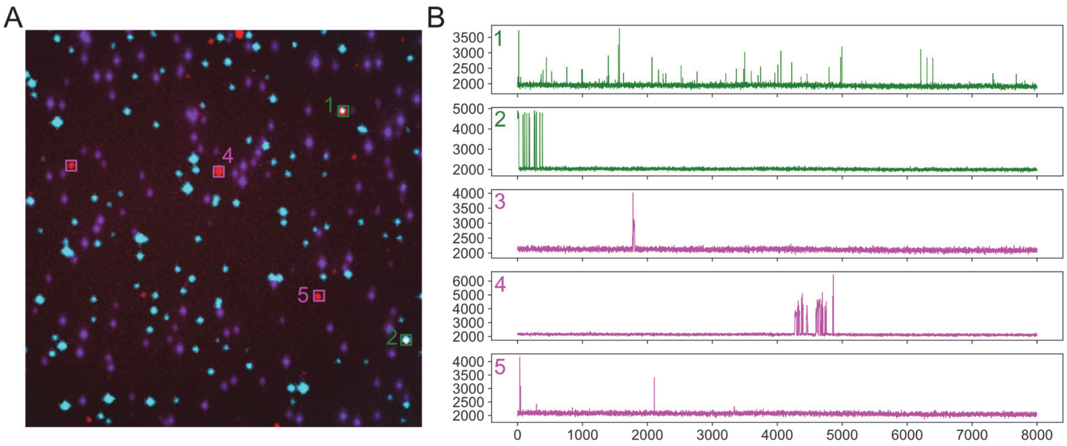


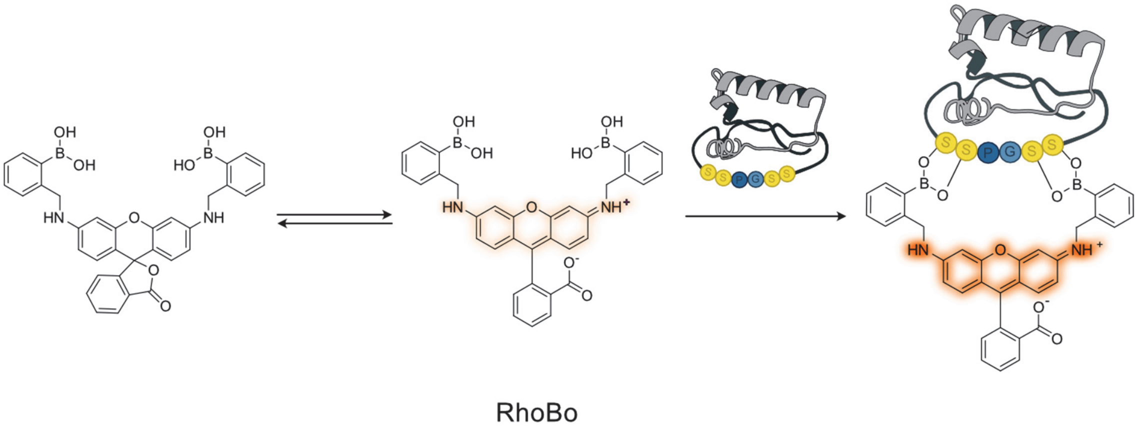
Figure 52. Preliminary multicolor experiment. A, Overlay of the maximum projections of the vesicles containing unlabeled HaloTag7 (blue, 10 frames, 405 nm, 15 mW, 30 ms), of vesicles loaded with HTHTL (cyan, 10 frames, 488 nm, 15 mW, 30 ms), and the red channel for HMSiR fluorescence (red, 8000 frames, 638 nm, 90 mW, 30 ms).

Chapter 5 Small tetraserine peptide tags for protein labeling with bisboronic acid probes

The results reported in this chapter are from collaborative work with Annabell Martin. Effort and credit for the preparation of the yeast display libraries are shared equally. We thank Dr. Gina Grammbitter for the initial implementation of the yeast display protocol in our laboratory. We also thank Dr. Elias A. Halabi Rosillo, a former member of our lab, for the preparation of compound 43.

5.1 Project goal

In our preceding projects, our focus was centered on the identification of proteins at the single-molecule level. The efforts in the previous chapters aimed at detecting protein species and their modifications with the highest sensitivity in a quantifiable manner. However, it is vitally important to consider proteins in their cellular environment to understand their function and their role in the cellular processes. In Chapter 1, we discussed available genetically encoded fluorescent tags such as FPs, self-labeling proteins, and small peptide tags that are targeted by enzymes or fluorescent dyes. An important factor that limits the usage of common tags like FPs and self-labeling proteins is their size. The addition of a large second protein to a POI can lead to misfolding or mislocalization.[139] In this project, we sought to develop fluorogenic bisboronic acid (BBA) probes with the same binding mode as RhoBo (Scheme 14), aiming to lower the background fluorescence in cell experiments. However, for the fluorophore to be of use, it requires a partner peptide tag that can render the molecule fluorescent upon binding. We employed a high throughput (HTS) yeast surface display (YSD) screen to search for a peptide capable of opening the fluorophore.[271,272] The resulting compound-peptide pair would be suitable for imaging of small proteins or protein-protein interactions, which is difficult to achieve with larger tags.

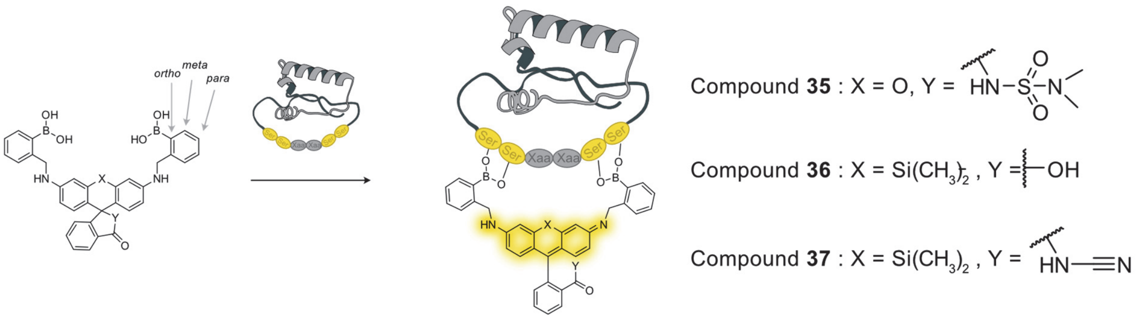


Scheme 14. Working principle of RhoBo, a small-molecule peptide tag binder.

Annabell Martin, a collaborator from our lab, was working closely with me on this project with the final goal of developing analogous BBA carbocyanines based on her recently published strategy to make fluorogenic carbocyanines.[224] We aimed for these BBA probes to be turned on with different peptide tags and thus would be orthogonal tags to the rhodamine-based derivates.

5.2 Probe design

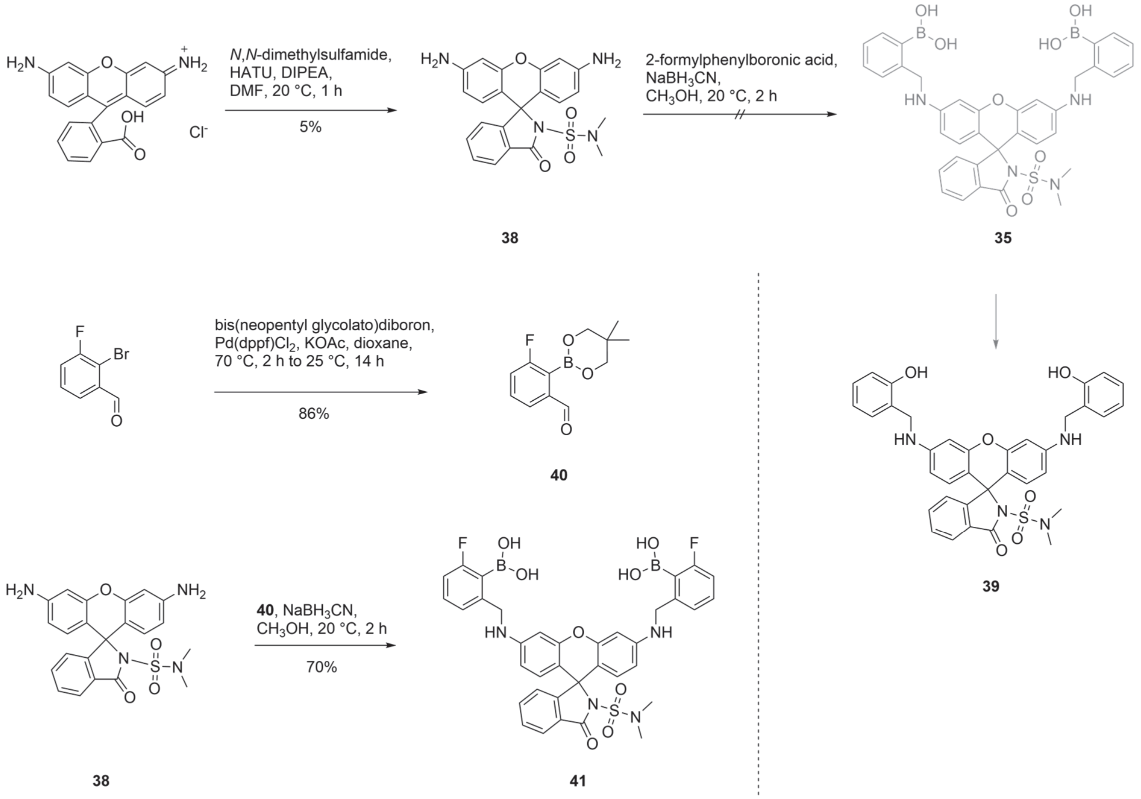
The major problem of RhoBo is the strong background fluorescence in cells due to the wide availability of the target tetraserines in the cell. Upon binding, the quenching mechanism of the free boronic acid is lost and RhoBo shows an increase in fluorescence. However, in the rhodamine structure, we do not have to solely rely on the quenching mechanism of the boronic acid. To lower the background fluorescence of a new BBA-rhodamine derivative, we aimed to shift the intramolecular open-closed equilibrium toward the closed form in comparison to RhoBo. The equilibrium can be tuned by increasing the nucleophilicity of the moiety attacking the sp2-hybridized central C atom.[124] The Johnsson group introduced electron-deficient amides to establish probes that turn fluorescent upon binding to a self-labeling protein.[124] We hypothesized that it should be possible to combine the fluorogenic compound MaP555 with the analogous boronic acid modification of RhoBo to create a fluorogenic *N*,*N*-dimethylsulfamide RhoBo probe 35 (Scheme 15, compound 35) upon tetraserine binding.[124] Silicon rhodamines have also been shown to be fluorogenic and are red-shifted compared to rhodamines, which is advantageous for live-cell imaging due to a lower autofluorescence and lower cell toxicity of the excitation light.[100,101] We could analogously use the silicon rhodamine scaffold and combine it with boronic acid moieties to create an orthogonal tetraserine BBA pair that could be used in multicolor experiments. Therefore, we decided to explore silicon rhodamine BBA (SiRhoBo) 36 (Scheme 15, compound 36). However, it is important to note that we cannot predict if the boronic acid quenching effect applies to silicon rhodamines, as the quenching mechanism has not been elucidated and it is unclear whether this is affected by the changes in the energy levels of the silicon rhodamine. Consequently, we additionally tested silicon rhodamines that carry electron-deficient amide nucleophiles such as compound 37 (Scheme 15), analogous to compound 35.[104] Furthermore, we wanted to investigate different positions of the boronic acids on the phenyl rings (Scheme 15). This change increases the interatomic distance of the boronic acid residues. As a result, we expected a larger tetraserine peptide as a binding partner, increasing the chance for orthogonal tetraserine-dye pairs.


ortho
meta
para
OH
B
OH
HO
HN
NH
X
Y
O
Ser
Xaa
N
Compound **35** : X = O, Y = HN–S(=O)(=O)–N
Compound **36** : X = Si(CH$_3$)$_2$, Y = OH
Compound **37** : X = Si(CH$_3$)$_2$, Y = HN—≡N

Scheme 15. Introducing a Si-atom (compounds 36 and 37) or electron-deficient amides (compounds 35 and 37) to create fluorogenic RhoBo alternatives upon tetraserine binding.[103,104]

5.2.1 Synthesis toward probe **35**

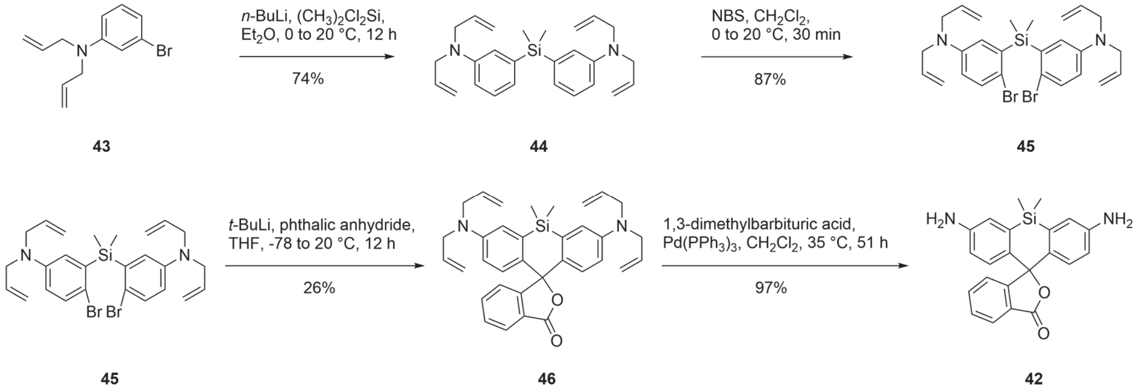
The synthesis of the target compound 35 started from the commercial rhodamine 110-chloride (Scheme 16) by installing first the *N*,*N*-dimethylsulfamide. The amide coupling of the carboxylic acid using HATU for its activation yielded the desired compound 38 in low yield.[273] After reductive amination with 2-formylphenylboronic acid, the targeted compound 35 could not be isolated. Instead, the mass corresponding to the side product 39 resulting from oxidation of the boronic acid to the alcohol, was observed as the main component. We hypothesized that we could lower the tendency of the boronic acid to oxidize by decreasing the electron density of the phenyl ring by the addition of fluorine in the *ortho*-position to the boronic acid. The fluorinated 2-formylphenylboronic acid analog 40 was obtained by Suzuki coupling of 2-bromophenylboronic acid and bis(neopentylglycolato)diboron. The intermediate 42 was directly used for reductive amination without deprotection. The corresponding fluorinated compound 41 could be isolated in good yield.



Scheme 16. Synthesis attempts of *N*,*N*-dimethylsulfamide RhoBo 35 and the final probe, the fluorinated *N*,*N*-dimethylsulfamide RhoBo 41. dppf = 1,1’-ferrocenediyl-bis(diphenylphosphine)

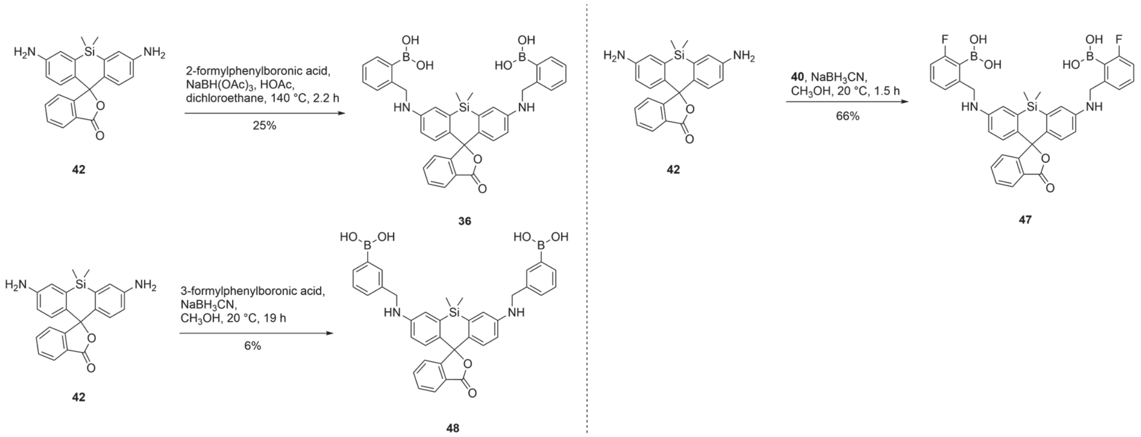
5.2.2 Synthesis of SiRhoBo and derivatives

The synthesis of SiRhoBo 36 and derivatives with shifted boronic acid positions required the synthesis of an alkylated silicon rhodamine that could be further modified using varying aldehydes via reductive amination. The common intermediate 42 was prepared in four steps starting from compound 43 (Scheme 17). Lithium-halogen exchange and nucleophilic substitution of dimethyldichlorosilane furnished the intermediate 44. Bromination of 44 with NBS resulted in compound 45. Lithium-halogen exchange with *t*-BuLi and addition to phthalic anhydride, followed by loss of water resulted in allyl-protected silicon rhodamine 46. The general intermediate 42 was obtained by palladium-catalyzed deprotection of the allyl-protected amines.


Br
N
43
n-BuLi, $(CH_3)_2Cl_2Si$,
Et_2O, 0 to 20 °C, 12 h
74%
N
Si
N
44
NBS, CH_2Cl_2,
0 to 20 °C, 30 min
87%
N
Si
N
Br Br
45
N
Si
N
Br Br
45
t-BuLi, phthalic anhydride,
THF, -78 to 20 °C, 12 h
26%
N
Si
N
O
O
46
1,3-dimethylbarbituric acid,
$Pd(PPh_3)_3$, CH_2Cl_2, 35 °C, 51 h
97%
H_2N
Si
NH_2
O
O
42

Scheme 17. Synthetic pathway to general SiRhoBo intermediate 42.

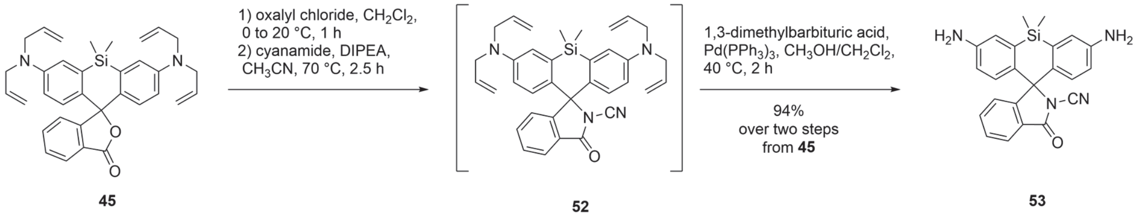
The silicon rhodamine intermediate 42 provided free amine groups for reductive amination with various formylphenyl boronic acids (Scheme 18). As opposed to the *N*,*N*-dimethylsulfamide RhoBo 35, both the non-fluorinated SiRhoBo 36 and fluorinated SiRhoBo 47 could be isolated. Compound 36 was prepared using reported reaction conditions using catalytic amounts of acid, the mild reducing agent sodium triacetoxyborohydride and 2-formylphenyl boronic acid with MW heating.[144] A modified procedure using sodium cyanoborohydride in anhydrous methanol was used for all other preparations of final BBA probes, including the fluorinated analog 47 with compound 40 as the aldehyde. The *meta*-SiRhoBo 48 was obtained via the same reaction conditions using 3-formylphenyl boronic acid.



Scheme 18. Synthesis of SiRhoBo 36 as well as the fluorinated SiRhoBo 47 and *meta*-SiRhoBo 48 starting from the common intermediate 42.

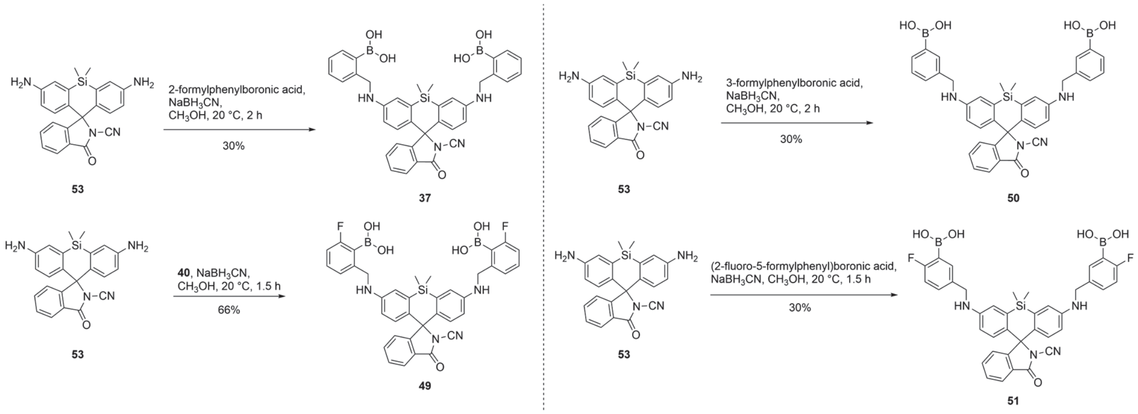
5.2.3 Synthesis of SiRhoBo-cyanamide and derivatives

The synthesis of the cyanamide derivatives of SiRhoBo 37, 49, 50, and 51 continued from compound 46 (Scheme 19). The cyanamide was introduced via the formation of the acyl chloride followed by nucleophilic attack of the cyanamide under basic conditions, resulting in the allyl-protected form 52 of the second common intermediate 53 (Scheme 19).


45
1) oxalyl chloride, CH$_2$Cl$_2$,
0 to 20 °C, 1 h
2) cyanamide, DIPEA,
CH$_3$CN, 70 °C, 2.5 h
52
1,3-dimethylbarbituric acid,
Pd(PPh$_3$)$_3$, CH$_3$OH/CH$_2$Cl$_2$,
40 °C, 2 h
94%
over two steps
from 45
53

Scheme 19. Synthesis of the cyanamide core intermediate 53.

Analogous to the preparation of the SiRhoBo derivatives, the cyanamide SiRhoBos were obtained through reductive amination of the free amine SiRhoBo 53 and the corresponding formylphenyl boronic acids (Scheme 20). The conditions were kept constant with slightly varying reaction times. SiRhoBo cyanamide 37 was obtained using 2-formylphenylbroronic acid, *meta*-SiRhoBo cyanamide 50 using 3-formylphenylbroronic, fluorinated SiRhoBo cyanamide 49 using compound 40, and fluorinated *meta*-SiRhoBo cyanamide 51 with commercial 2-fluoro-5-formylphenylboronic acid.


53
2-formylphenylboronic acid,
NaBH3CN,
CH3OH, 20 °C, 2 h
30%
37
53
3-formylphenylboronic acid,
NaBH3CN,
CH3OH, 20 °C, 2 h
30%
50
53
40, NaBH3CN,
CH3OH, 20 °C, 1.5 h
66%
49
53
(2-fluoro-5-formylphenyl)boronic acid,
NaBH3CN, CH3OH, 20 °C, 1.5 h
30%
51

Scheme 20. Preparation of cyanamide RhoBo derivatives 37, 49, 50, and 51.

5.2.4 Spectroscopic evaluation of the probes

The spectra of all newly prepared BBA probes were measured to evaluate their basal fluorescence in PBS and methanol. The boronic acid quenching effect should be observable in water, but it should be decreased in methanol due to the spontaneous formation of methyl boronic esters (Figure 53, first row A and first row B). The equilibrium of the open and closed forms of xanthene dyes is dependent on the pH of the medium. At lower pH, the nucleophile on the lower ring should be preferentially protonated and hence be less nucleophilic, leading to a larger population of the fluorescent form. The absorbance and fluorescence spectra of all BBA probes were measured at 5 µM in PBS and methanol with and without the addition of 0.1% TFA (Figure 53 and Appendix Figure 76, A and B, top and bottom row respectively). For all BBA silicon rhodamine cyanamide compounds 36, 49, 50, and 51, no absorption nor emission could be observed in any medium supporting their strong tendency to persist in the closed state even upon the addition of TFA (Appendix Figure 76). The SiRhoBo variants 33, 47, and 48 as well as *N*,*N*-dimethylsulfamide 41 exhibited low fluorescence in PBS and unexpectedly even lower fluorescence in methanol. This observation might be attributed to the dyes preferentially existing in the closed state in solvents with a lower dielectric constant and not to the potential boronic acid quenching.[124] A clear increase in absorbance and fluorescence could be observed upon the addition of 0.1% TFA confirming acid-induced ring-opening. As expected, the dyes exhibited very low fluorescence, especially the silicon rhodamines modified with an electron-deficient amide moiety. However, these experiments were carried out in solution without any of the potential interaction partners that are present in a cell. Therefore, it was crucial to test the background fluorescence in cells.

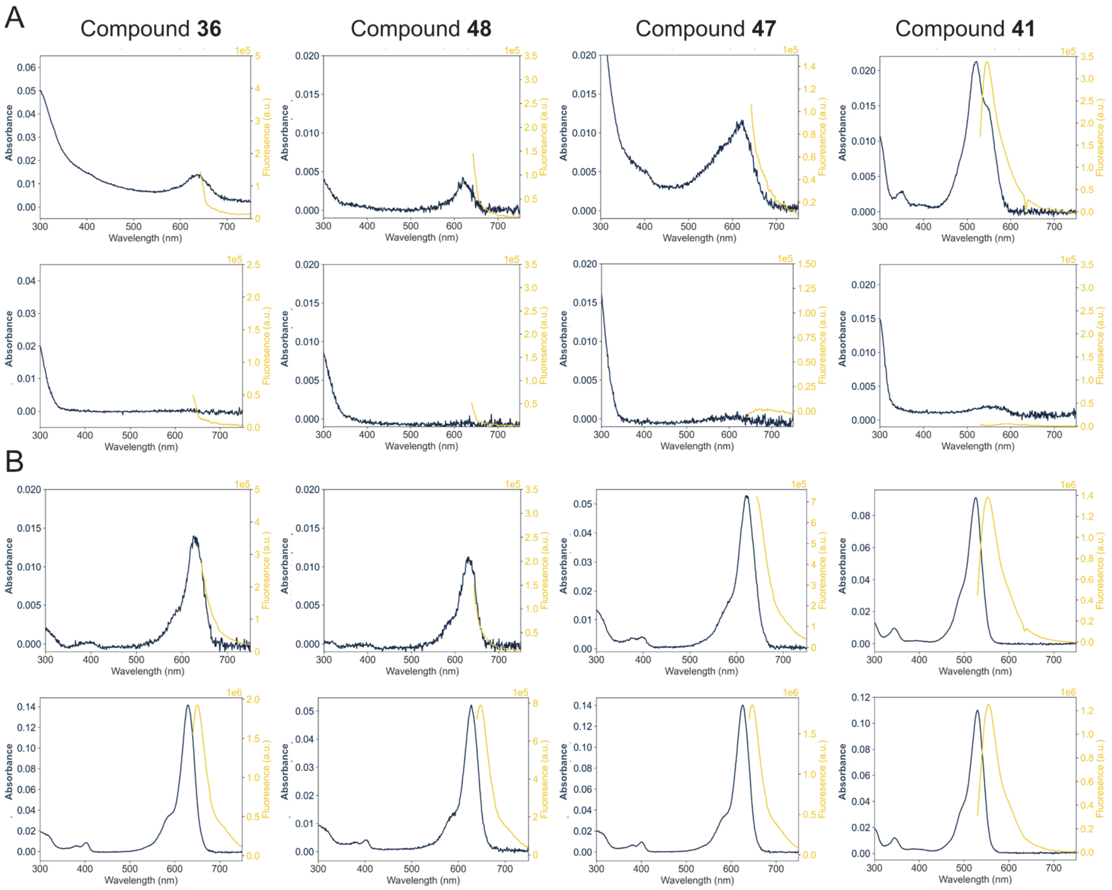


Figure 53. Absorbance and fluorescence spectra of 5 μM compounds **36**, **48**, **47**, and **41**. A, Spectra were measured in PBS (top row) and methanol (bottom row). B, Spectra were measured in PBS (top row) and methanol (bottom row) with 0.1% TFA in each. Compounds **36**, **47**, and **48** were excited with 633 nm light and compound **41** with 519 nm light.

5.2.5 Background evaluation in live cells

With the new BBA compounds, we conducted imaging experiments in HeLa cells to qualitatively evaluate if we could indeed observe low background fluorescence in the presence of native tetraserines and other cellular components like sugars. HeLa cells were incubated with 100 nM, 500 nM, 1 μM, and 2 μM dye concentrations. The cells were imaged at different time points after incubation (10 min, 1 h, 3 h, and 15 h) to assess the time scale of potential binding of the compound. All silicon rhodamine probes were imaged using a 638 nm laser and *N*,*N*-dimethylsulfamide rhodamine **41** was imaged using a 515 nm laser. The images of the different conditions compared at the same image contrast can give an idea of the relative background over time and concentration (Appendix Figure 78, Figure 79). For comparison, the average fluorescence intensity of ten randomly chosen cells was extracted after 3 h (Figure 54, E-G). In general, the fluorescence background increased clearly with increasing concentration after labeling over 3 h (Figure 54 and Appendix: Figure 78, Figure 79). The SiRhoBo derivatives 36, 47, and 48 without the electron-deficient amide exhibited an increased background fluorescence in comparison to the more closed cyanamides (Figure 54, B). The results did not show a clear difference comparing the *ortho*-and *meta*-placement of the boronic acid group. The average fluorescence seemed to be comparable for the *ortho*- vs *meta*-boronic acid SiRhoBo cyanamides (49 and 51 or 50 and 37) independent of the fluorination state (Figure 54, G). However, SiRhoBo 36 appeared to result in a higher background compared to *meta*-SiRhoBo 48. In all cases, the fluorinated compounds gave lower fluorescence (Figure 54, F, G). It is important to note that we could not determine the cause of lower background fluorescence in this experiment. Therefore, we do not know if the lower fluorescence of the fluorinated compounds was due to a shifted equilibrium of the ring closing, less background binding, or a lower cell permeability of the compound. The spectra previously measured (Figure 53 and Appendix Figure 76) suggested that the equilibrium between the open and the closed forms at cellular pH is not the determining factor. Concerning the fluorinated derivatives, we theorized that the electron-withdrawing properties of the fluorine would decrease the pK_a of the boronic acid, which in turn would result in a larger amount of negatively charged probes. Since negatively charged probes are generally less membrane-permeable, the lower pK_a could have been the reason for the decreased background.[269] Further, the structures that the probes seemed to localized resembled the endoplasmic reticulum (Figure 54, D). However, we did not conduct a formal colocalization experiment with a probe known to be targeted to the endoplasmic reticulum to test this hypothesis. Furthermore, we could not observe staining of the cell membrane or the glycocalyx, similar to the original RhoBo (Figure 54, A).[141] If sugars were effective turn-on binders for our probes, we could expect staining of the glycocalyx, which is composed of heavily glycosylated proteins and glycans attached to lipids and is known to be present in HeLa cells.[274] In conclusion, we have prepared BBA compounds with low background fluorescence, which at least partially is due to a shifted intramolecular equilibrium toward the closed form even in a cellular environment (Figure 54, Appendix Figure 78 and Figure 79).

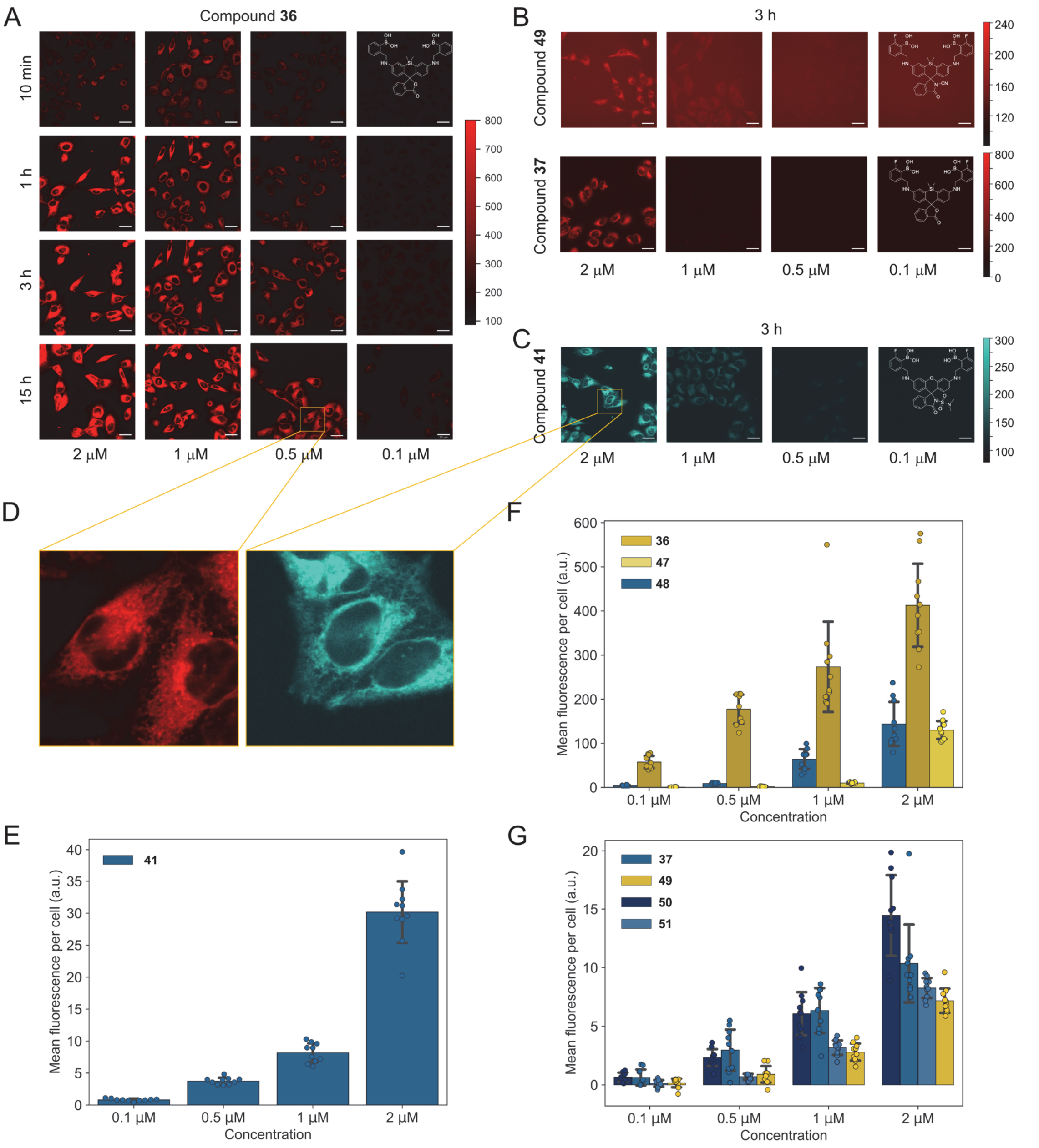


Figure 54. HeLa cell imaging experiments for the estimation of the expected background fluorescence with all silicon rhodamine-based compounds imaged using a 638 nm laser (100 ms, 30 mW). A, Images for all conditions conducted for each probe, with SiRhoBo **36** as an example. B, Images of different concentrations after 3 h incubation time of compound **47** and the analogous cyanamide compound **49**. C, Images of different concentrations after 3 h incubation time of compound **41** (515 nm, 100 ms, 30 mW). D, Enlarged area of the two respective areas in the images in A and C. E, Average fluorescence intensity of compound **41** after 3 h of incubation. F, Average fluorescence intensity of all SiRhoBo derivatives without cyanamides. G, Average

fluorescence intensity of all SiRhoBo derivatives with cyanamides. A-C, Scale bars 20 µm. D-G, n = 10. The color bar represents the fluorescence intensity (a.u.).

5.3 Surface display screening approach

To maximize the opportunities to identify a tag that would be able to bind and turn on our fluorophore, we opted for an HTS approach. Cell surface display technology has been successfully used to select peptides for improved function in multiple organisms: phage, bacteria, yeast, and mammalian cells.[132,275–278] In surface display technologies, a PPOI is fused to a surface-anchored protein. Hence the PPOI is exposed on the surface of an organism, allowing it to be probed easily (Figure 55). From the different platforms proposed in the literature, we opted to use YSD (Figure 55, A).[132,275–278] The main reason for this choice is the relatively large size of a yeast cell in comparison to bacteria or phages. The larger cell size makes yeast readily compatible with fluorescence-activated cell sorting (FACS). In addition, the larger cell surface results in an increased amount of protein, and consequently an increased detection sensitivity.[279,280] Although mammalian cells are even larger, their handling and construction of larger libraries is more challenging.[281] In yeast display presentation is achieved through a fusion of the PPOI to a cell wall protein (CWP). The CWPs themselves are fixed at the C-terminus either via a glycosylphosphatidylinositol anchor or they contain internal repeats (PIR family) that are covalently attached to β-1,3-glucan, a component of the yeast cell wall.[282] The α-agglutinin system is the most used and consists of two subunits, the anchorage subunit Aga1p and the adhesion subunit Aga2p.[282–284] Aga1p consists of 725 residues and anchors the assembly to the yeast cell wall via β-glycan covalent linkage. Aga2p contains 69 residues and is linked to Aga1p via two sulfide bonds.[285] Consequently, both termini of Aga2p are free to be fused to a PPOI.

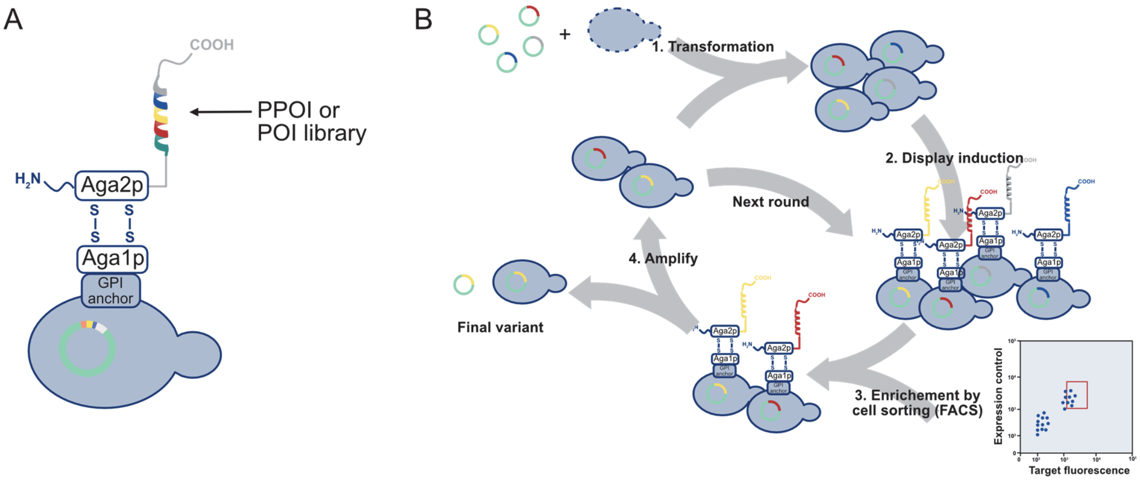


Figure 55. Summary of YSD using the α-agglutinin system. A, General composition and structure of a target peptide or protein fused to the C-terminus of the α-agglutinin system displayed on a yeast cell. B, Principal workflow of YSD for PPOI selection using a randomized plasmid library and selection via FACS.

5.3.1 Library design

The currently used tags for FlAsH, ReAsH, and RhoBo have been derived from initial rational design, followed by selected substitutions, extensions, and adaptation to new probes.[138,286,287] However, to the best of our knowledge, there has not been an exhaustive screening approach to improve the current tags by varying the amino acid sequence separating the dicysteines or diserines. Based on the successful binding of RhoBo to the peptide motif -SSPGSS-, we sought to screen all peptides of the pattern -SSXXSS- (Figure 56, A, 2aa-Library) for fluorophore turn-on. Although many hexapeptides with a tetraserine are present in the proteome, we suspect that even when bound, only a small number might trigger fluorescence turn-on of our dyes.[141]

The initial rational approach toward tag design envisioned placing the two cysteine pairs on the same face of an α-helix at a distance matching the interatomic distance between the arsenic atoms on the corresponding biarsenical fluorophore.[133] However, follow-up studies indicate that slightly longer peptides have the potential for improved binding properties. T. Wang et al. screened the proteome of *Shewanella oneidensis* MR-1 and *E. coli* K-12 for alternative FlasH tags with improved properties and obtained longer tags.[139] The Schepartz group investigated the binding of AsCy3 by systematically extending the peptide tag one amino acid at a time by repeating the original tag. They found that the binding increased significantly after adding six amino acid residues to the original tag.[288] Although the interatomic distance of RhoBo is similar to FlAsH (~6 Å) and hence smaller than AsCy3 (~14.5 Å), the relative distance between the boronic acid is more flexible, which could allow for more variable binding modes than FlAsH and probes **48**, **50** and **51** have an increased interatomic distance.[134,138] Extending the tetraserine tag also offers the advantage of an increasing number of variants and hence more opportunities to find a peptide that can open a fluorogenic dye. In addition to the increase in possible variants, the coverage of tetraserines by the human proteome decreases and the chance that a hit peptide tag is not present in the proteome increases (Figure 56, A). With these prospects, we decided to screen an additional library with eight amino acids flanked by two serine pairs (Figure 56, B, 8aa-Library). For both libraries -SSXXSS- (2aa-Library) and -SSXXXXXXXXSS- (8aa-Library), X corresponds to a random amino acid. However, we excluded leucine based on the reasoning that it is chemically very similar to isoleucine and hence would not add chemical diversity to the library. In consequence, we increased the density of chemical diversity.[289] Further, we avoided cysteine to prevent disulfide formation. As a result, the 2aa-library offers 361 and the 8aa-library $\sim 1.1 \cdot 10^{10}$ possible peptides.

We utilized the α-agglutinin YSD system and fused our randomized target peptide libraries (purchased from EllaBiotech) to the C-terminus of the subunit Aga2p. Furthermore, we introduced the human influenza hemagglutinin (HA)-tag in tandem with the target peptides separated by a $(G_4S)_3$-repeat linker (Figure 56, C).[284] As a control and background check, we utilized a display control (DC) that expressed the same plasmid backbone with the Aga2p and the HA tag. Hence, the HA peptide was expressed with the display protein and used to control for proper display and background fluorescence (potential dye binding) without the presence of a TS-containing peptide.

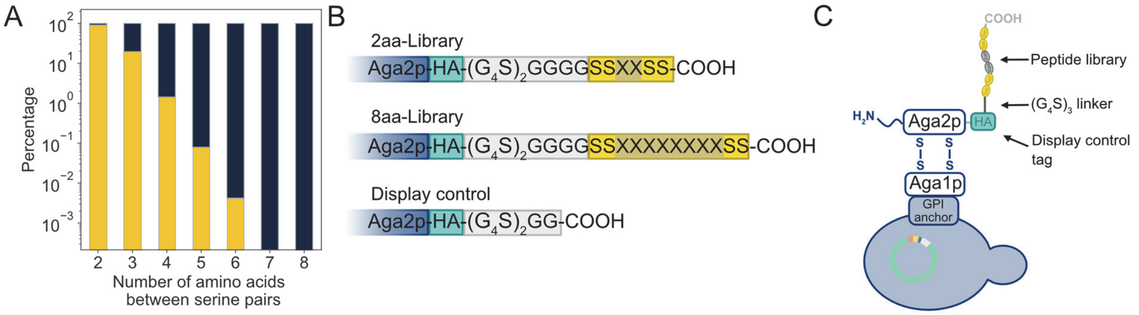
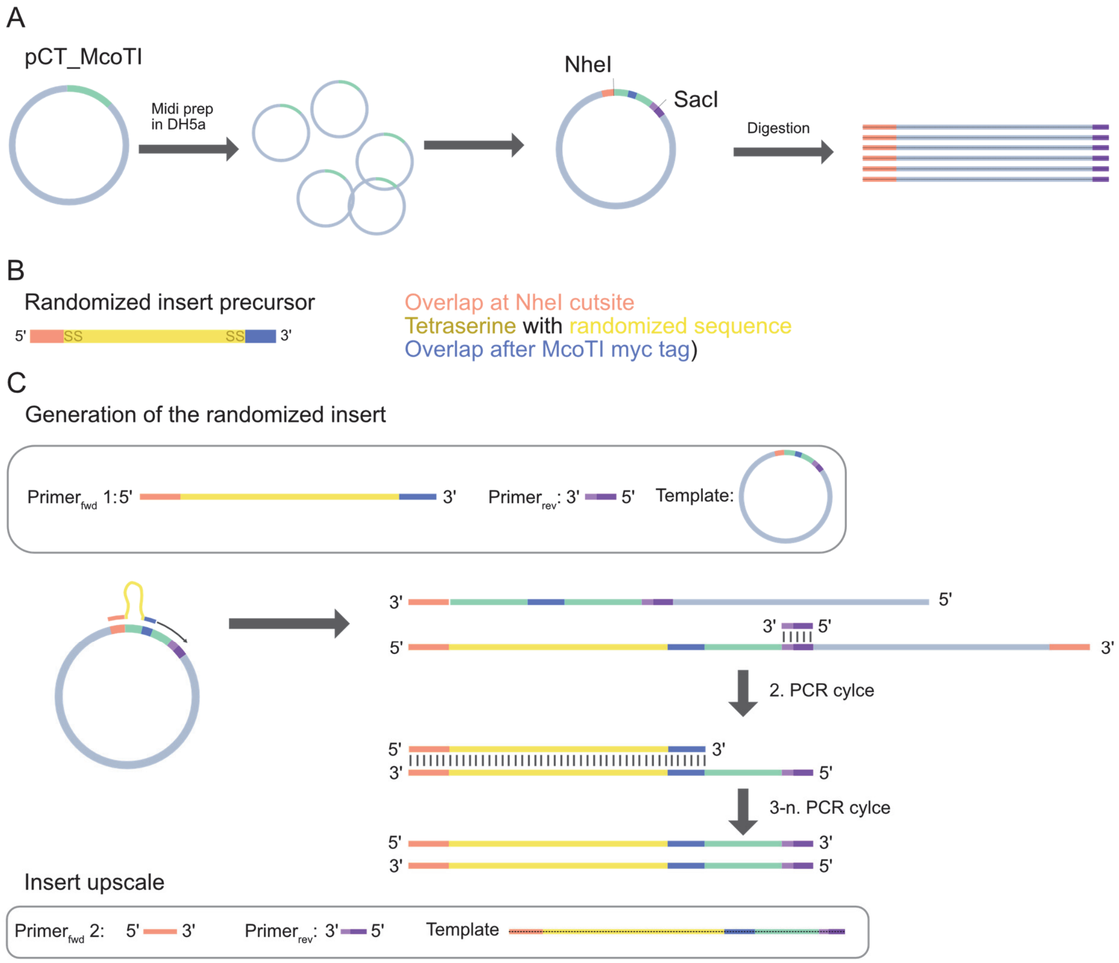


Figure 56. Library design. A, Percentage of TSs that are present in the human proteome (yellow, based on proteins present in the UniProtKB/TrEMBL and UniProtKB/Swiss-Prot, January 2024) compared to all possible variants.[290] B, Explicit peptide sequences are shown after the HA-tag (turquoise) and the C-terminus (-COOH) for the two envisioned libraries and the DC. C, Yeast with α-agglutinin system fused with HA, a $(G_4S)_3$-linker, and the randomized tetraserine library.

5.3.2 Cloning strategy

To generate a construct that contains Aga2p, our desired peptide library, and a surface presentation control, we exploited the efficient homologous recombination of yeast to assemble the plasmid with the display protein-tetraserine fusion. Naturally homologous recombination is a DNA repair mechanism that restores double-strand breaks using the homology of the sister chromosome to generate overlaps.[291,292] This mechanism can be used to combine linear DNA fragments with overlapping flanking regions (20 to 40 base pairs) in the yeast cell and has been optimized for YSD library generation (Figure 55, A).[289,293]

We prepared the two linearized fragments with overlapping regions for transformation into yeast. The first (backbone) fragment was obtained from the plasmid pCT_McoTI (Appendix Figure 75), an *E. coli/S. cerevisiae* shuttle vector, allowing for replication in both organisms. The plasmid encodes Aga2p fused to the HA-tag, the $(G_4S)_3$-linker, the trypsin inhibitor McoTI (*Momocordia cochinchinensis* trypsin inhibitor), and the myc-tag in tandem under a galactose-inducible promoter. We amplified the plasmid in *E. coli* and digested it using the restriction enzymes NheI and SacI. This resulted in a linear fragment from which the sequence encoding McoTI and the myc-tag was removed (Scheme 21, A). The second (insert) fragment was generated using an oligonucleotide library (EllaBiotech) that encoded a randomized tetraserine peptide and was flanked with sequences overlapping on the plasmid pCT_McoTI (Scheme 21, B). This oligonucleotide was used as the forward primer in a polymerase chain reaction (PCR) with pCT_McoTI as the template and a reverse primer partially overlapping on the SacI cut site (Scheme 21, C purple). This step yielded the insert fragment with the required overlapping regions on the backbone fragment (34 and 36 base pairs) for homologous recombination (Scheme 21, C and Figure 57, A). The insert was amplified in a second PCR to obtain sufficient material for transformation. The two fragments were transformed into yeast cells by electroporation.



Scheme 21. Cloning strategy for the preparation of the linearized vector backbone and library insert needed for yeast transformation. A, The backbone plasmid was amplified and digested to create the linearized backbone. B, Design and synthesis of the randomized library primer (EllaBiotech). B and C, The library primer has two overlap regions on pCT_McoTI flanking the previously displayed protein. Employing this primer in the PCR with Primer$_{rev}$ results in the linearized insert fragment, where the randomized library replaces the previous protein.

5.3.3 Transformation and next-generation sequencing

We carried out the transformation of yeast cells (EBY100) via electroporation to generate the libraries with the corresponding insert.[289] Following transformation cells were grown in synthetic dextrose (SD) dropout medium. This medium does not contain tryptophane, which is essential for yeast cells of the strain EBY100. Therefore, only cells that had taken up both fragments and successfully recombined them could produce tryptophane using the gene supplied in the backbone and survive. The cells were then frozen in aliquots and freshly thawed before analysis and selection. Homologous recombination was checked by picking ten colonies per transformed library (2aa and 8aa) for colony PCR (Figure 57, B). The sequencing results showed that homologous recombination was successful and different amino acids were introduced between the serine pairs. The diversity of the obtained 8aa-library was estimated as the number of viable transforms obtained from electroporation via dilution series and determined to be approximately 7 · 10^{6}. Considering the vast possible diversity of an 8aa-library, we checked the quality and the amino acid distribution of the generated yeast library by next-generation sequencing (NGS). The adapter sequences were added via PCR during the amplification of the gene sequence encoding the tetraserine peptide (Primers, Appendix Table 24). Further preparation of the sample and amplicon sequencing was conducted by FGCZ at the UZH. Sequencing was conducted with Illumina NextSeq 2000 technology with >30 Mio paired-end reads of 150 bases per sample. Processing of reads and data analysis was performed by Dr. Natalia Zajac (FGCZ). The sequencing data confirmed the even distribution of all expected amino acids between the flanking serine pairs (Figure 57, C).

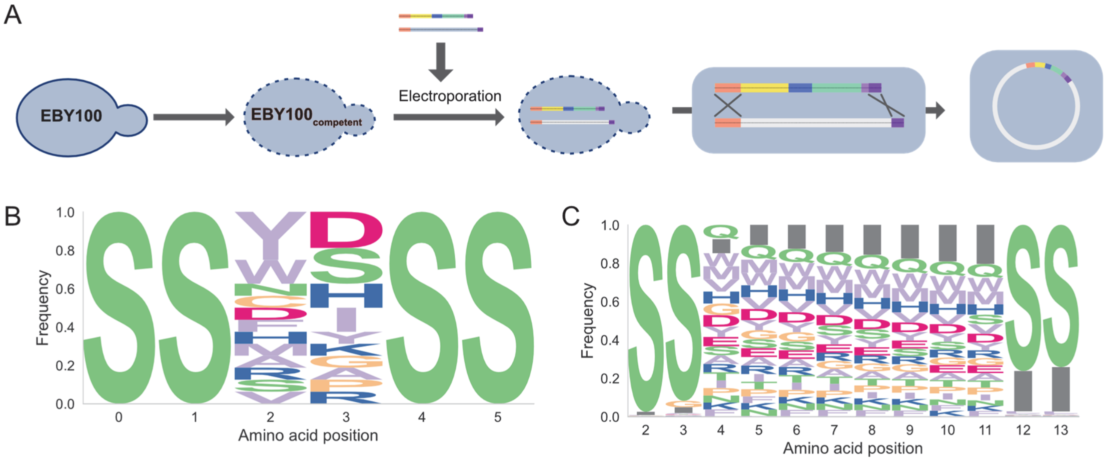
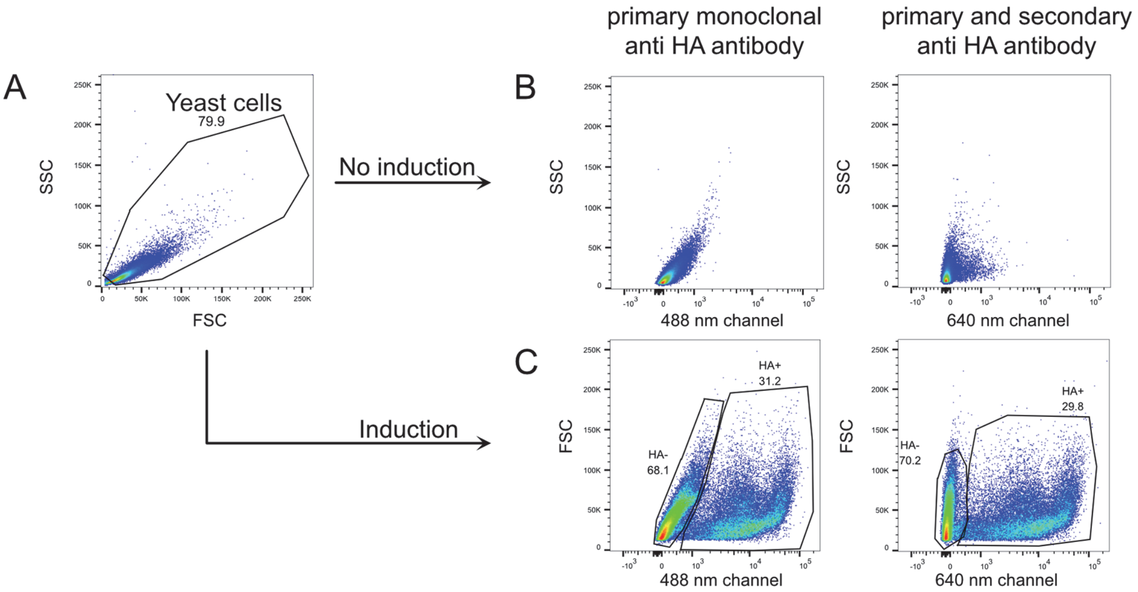


Figure 57. A, Transformation on freshly prepared competent EBY100 cells via electroporation. B, Amino acid distribution in tetraserine peptides sequenced by colony PCR of the 2aa library, n = 16. C, Amino acid distribution determined by NGS, the dark grey bar represents a stop codon, n = 25,856,781.

5.3.4 Induction test

After growing the cells, they were diluted in synthetic galactose (SG) medium, in which dextrose is substituted by galactose. The expression of the display construct is controlled by a galactose inducible promotor. Hence only cells grown in SG medium should display Aga2p, the HA-ag, and any fused PPOI (Scheme 22). We tested YSD by inducing the cells in SG medium, followed by flow cytometry analysis. We labeled the HA tag on the surface of the cells expressing the construct with an anti-HA antibody (Scheme 22, B). We found approximately 30% of the measured cells to display the HA tag upon induction, independent of whether a directly labeled primary anti-HA antibody was used or a primary antibody in combination with a labeled secondary antibody (Scheme 22, B).



Scheme 22. Monitoring of cell surface display of induced yeast cells by antibody staining. A, Yeast cells were gated in a side-scatter (SSC) vs. forward-scatter (FSC) plot to exclude very small particles such as debris. B, Not induced cells that were stained with the monoclonal antibody (left) and the secondary staining protocol from (right). C, Induced yeast cells that were stained with the monoclonal antibody (left) and the secondary staining protocol (right). $1 \cdot 10^5$ cells were measured per sample.

5.4 Enrichment of peptides by FACS

5.4.1 FACS of the 2aa-library

The 2aa-library was chosen in analogy to the strong binding of RhoBo to -SSPGSS-. In RhoBo, the boronic acids are placed in *ortho*-position, hence we decided to test peptide enrichment with all our BBA probes with the same relative position of the boronic acid (compounds 36, 37, 47, 49, and 41. In an initial enrichment experiment, we co-stained the 2aa-library with a fluorescent anti-HA antibody, to ensure display and the compounds (Appendix Figure 80 to Figure 84).[132,285,294] However, we could not observe enrichment of fluorescent cells (Appendix Figure 82 to Figure 84). Based on these results and taking into consideration the substantial size of the antibody, we hypothesized that antibody labeling of the HA tag close to the potential peptide tag might interfere with BBA probe binding. Hence, we decided to sort cells based only on the fluorescence intensity in the channel relevant to the respective probe. To determine the gates for selecting the cells we measured induced DC cells directly preceding the sorting of the induced yeast library. DC cells should exhibit the same off-target binding to the display machinery and other displayed features (HA tag, $(G_4S)_3$-linker). Therefore, a higher fluorescence for cells presenting a peptide that can bind and turn on our probes in addition to any potential background fluorescence would be expected. For each sorting round, $2 \cdot 10^7$ cells were incubated for 1 h with a 1 µM solution of the respective probe. Directly before sorting, the dye was removed, and the cells were diluted and strained. Cells stained with 36, 37, 47, and 49 were measured in the 640 channel (633 nm, 780/60) vs FSC, and compound 41 was measured in the 561 channel (561 nm, 582/15) vs FSC (Figure 58). The gates were set based on the DC sample while collecting 0.1-1% of the screened cells (Figure 58). Sorting was terminated after detecting $1 \cdot 10^7$ events. The sorted cells were collected in SD medium and grown for 3 days before continuing with the next round of sorting and storing surplus cells in aliquots at –80 °C. After the first round of sorting, no population at higher fluorescence intensity could be observed, but the samples looked comparable to the measurement before any sorting (Figure 58, B and D). In the second round (Figure 58, D), we decided to sort the samples into multiple populations for the compounds for which a broad distribution of sizes and fluorescence of labeled cells was observed (Figure 58, C and D, compound 36, 37, and 47). This would allow us to see whether the prominently stained population (Figure 58, D, P5) that was already observed in DC cells and the first FACS attempt corresponds to dead cells or cell debris. If these cells are dead or apoptotic, the population should not grow after sorting.

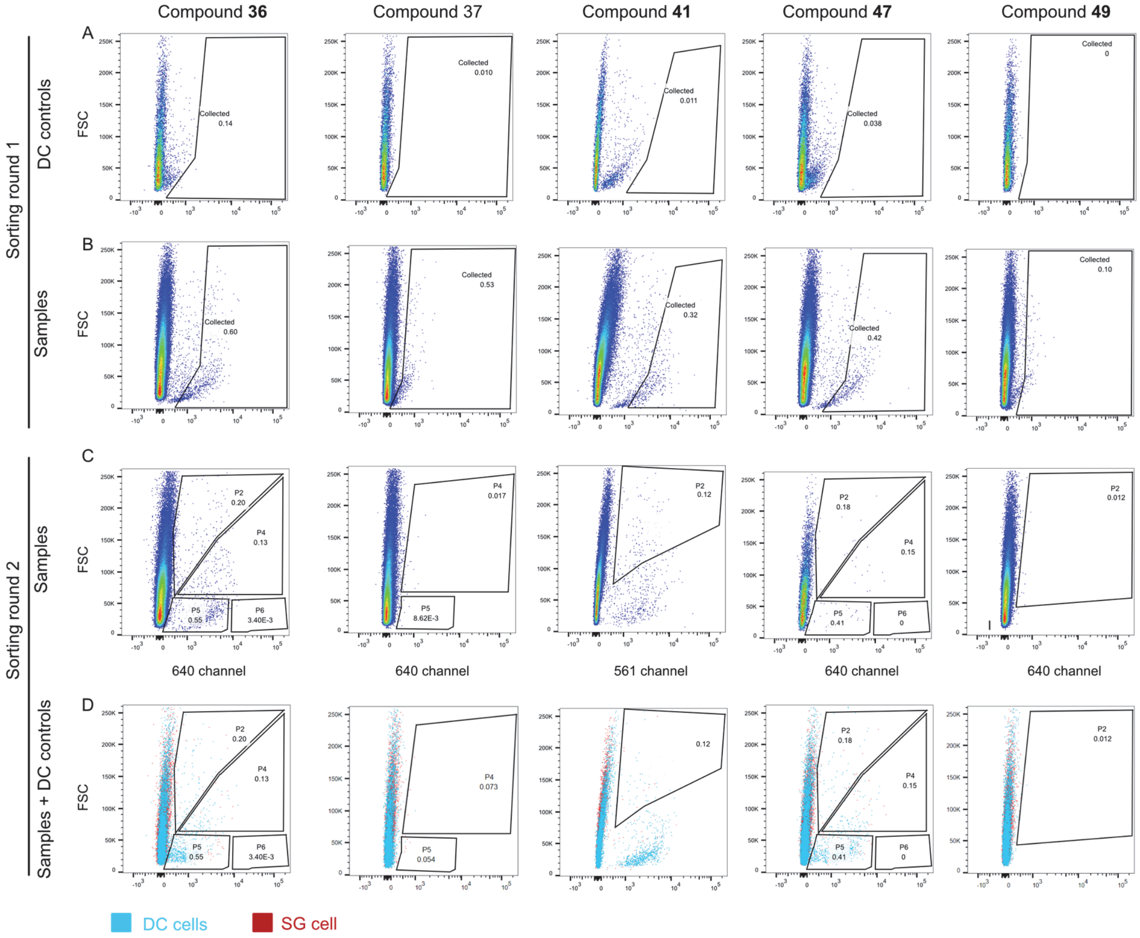


Figure 58. FACS for enrichment of small peptides from the 2aa-library using our compounds 36, 37, 41, 47, and 49 over two rounds. A, Control samples stained with one compound before sorting. B, Induced library samples stained with one compound before sorting. C, Induced library samples stained with one compound after the first round of sorting. The black polygons correspond to the chosen gates for the respective round. $1 \cdot 10^5$ cells are shown in each plot. D, Control samples (blue) stained with one compound after the first round of sorting overlapped with the stained, induced library samples (red).

After sorting into the different populations, these were grown separately and induced, before analyzing them by flow cytometry (BBAs 36, 37, 47, and 49: 640 nm, 660/20, BBA 41: 488 nm, 585/42) (Figure 59). None of the populations died and all grew to similar extents, hence populations P5 and P6 are not dead cells. Furthermore, P6 seemed to have a slightly shifted population of smaller, more intense cells compared to the DC control for both compounds 36 and 47 (Figure 59, zoom-in). Therefore, populations P5 and P6 of compounds 36 and 47 appear to be more fluorescent.

However, the results of these two rounds were not promising considering the low number of variants (361) in the 2aa-library. In the case of a low number of variants in comparison to the number of screened or analyzed cells ($1 \cdot 10^7$ cells or $1 \cdot 10^5$ cells, respectively), we would expect to observe each variant many times, which should result in a well-visible population. Nevertheless, based on flow cytometry analysis, we cannot evaluate the location of labeling, which could give insight into whether the displayed peptide is labeled or whether we only observed the higher background of compounds **36** and **47** (see 5.2.5).

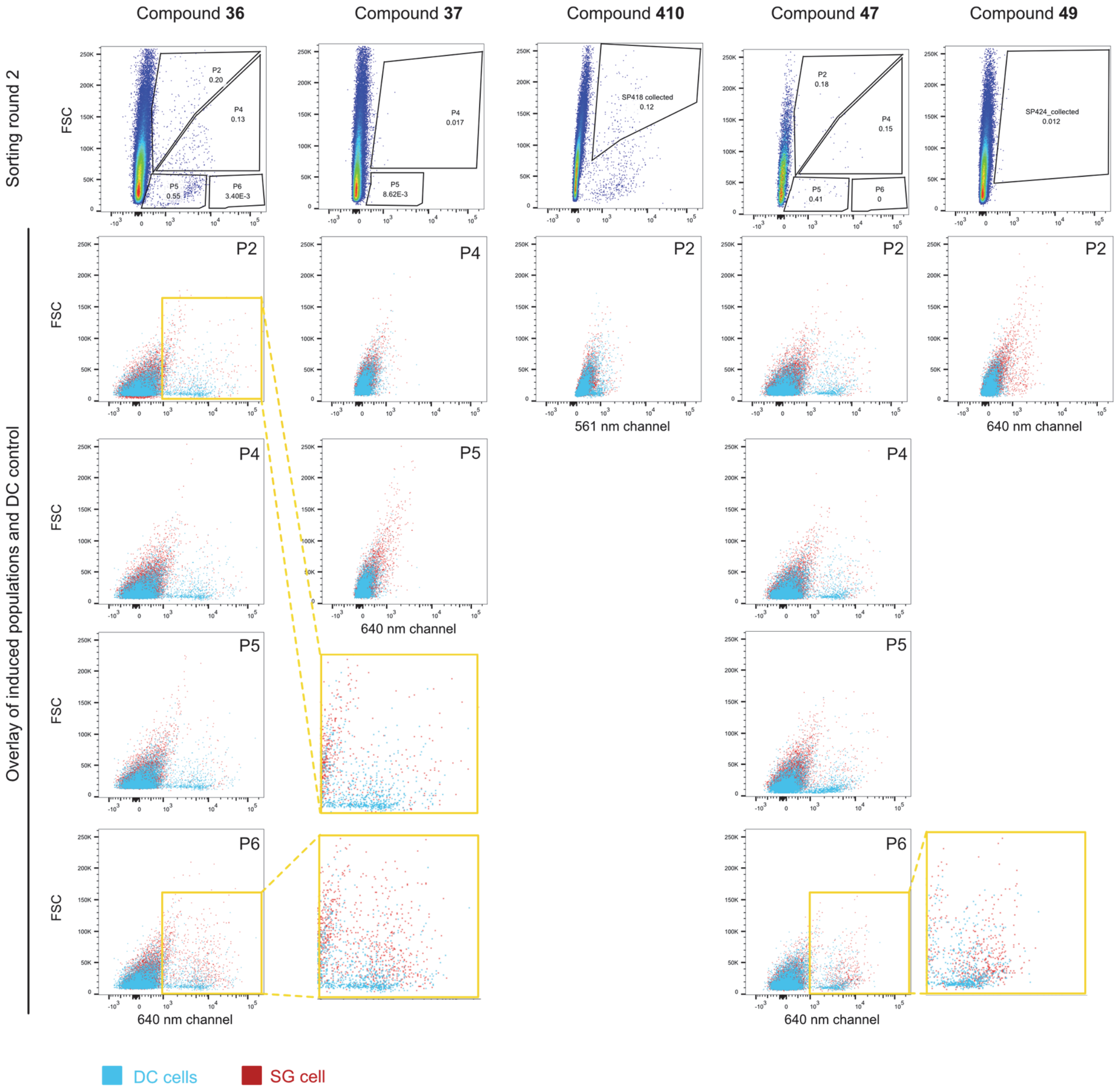


Figure 59. Flow cytometry analysis of the second round of the second attempt to enrich a peptide that can bind and turn on the fluorescence of our dyes. Top: panels from the sorting. Below: overlays of the populations after cultivation and induction from the collected cells of the respective population (red) and the DC control (blue). Flow cytometer settings used for, **36**, **37**, **47** and **49**: 640 nm, 660/20, compound **41**: 488 nm, 585/42. $1 \cdot 10^5$ cells are shown in each plot.

5.4.2 Imaging of sorted yeast populations

During enrichment, we collected different populations with a slight increase in fluorescence. In addition to the flow cytometer analysis of these populations, we imaged the yeast cells using a confocal microscope following the same labeling procedure (1 µM, 1 h) as applied for flow cytometry. Compounds **36**, **37**, **47**, and **49** were imaged in the 640 channel (638 nm, 200 ms, 60 mW) and BBA **41** in the 515 channel (515 nm, 200 ms, 60 mW). If a dye can bind an enriched peptide in any of the populations, we would expect to see predominantly staining of the yeast cell wall, where the displayed protein is located. This staining of the outside of the yeast cells was visible when staining the HA peptide in induced cells with an anti-HA antibody conjugated to AlexaFluor488 (Figure 60, B and C). However, we could not see similar staining, exclusive to the cell wall, in any of the populations collected and stained with the respective BBA probe (Examples: Figure 60, E and F). Only very few cells could be observed to be fluorescent and generally, they were stained throughout the cell body without a higher intensity on the outer ring (Figure 60, B and C). Importantly, such cells were also observed in populations grown from the sorted cells that were not stained with the dye but only with the antibody, which carries a fluorophore of a different color (Figure 60, D and zoom-in). Furthermore, these fluorescent cells exhibited a particular morphology in the brightfield image, and the same area is fluorescent in multiple channels not corresponding to the fluorophore (Figure 60, E and F, zoom-in). Based on these observations, we concluded that we were not able to find a peptide in the 2aa-library that binds to one of the BBA probes and generates a turn-on while displayed on the surface of a yeast cell.

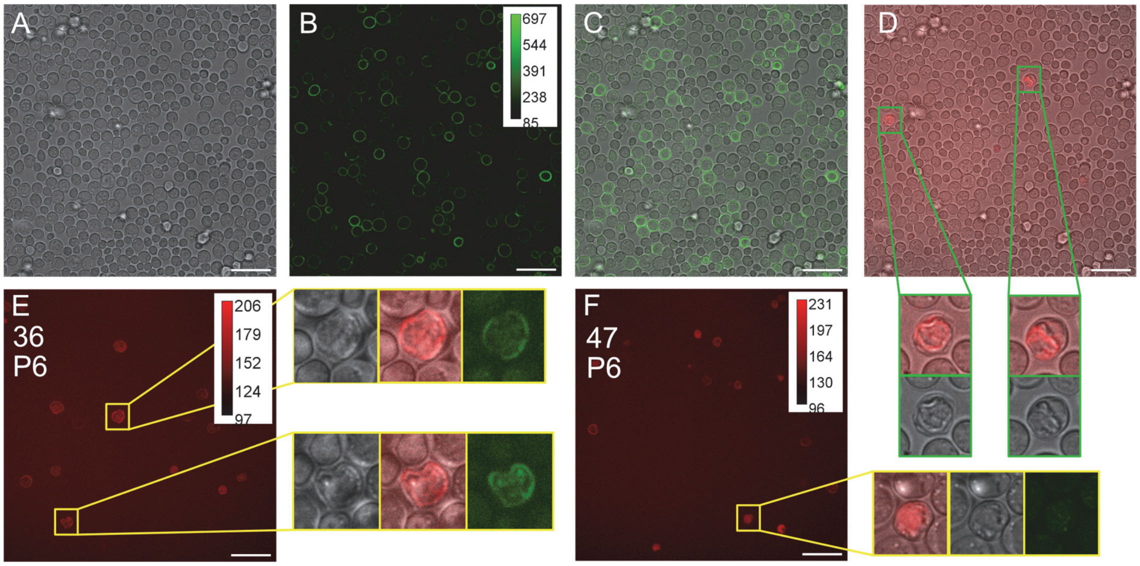


Figure 60. Confocal microscopy of induced yeast cells stained with the respective BBA fluorophore (638 nm, 200 ms, 60 mW) or the anti-HA antibody imaged in the 488 channel (488 nm, 200 ms, 60 mW). A to D, Cells from sorting using compound 47 P6 stained with the anti-HA antibody only. A, Brightfield image. B, Image in the 488 nm channel. C, Overlay of the two channels. D, Fluorescence image in the 640 channels with the cells showing autofluorescence in this channel framed in green and expanded alongside the same brightfield area. E and F, Selected images in the 640 nm channel of the population P5 and P6 labeled with the respective BBA (36 or 47). Fluorescent cells are pointed out in a yellow frame and zoomed in alongside the same cells that are imaged in the brightfield and the 488 channel of the respective cells. The color bar represents the fluorescence intensity (a.u.). Scale bar = 20 µm.

5.4.3 FACS and flow cytometry analysis of the 8aa-library

The results for the smaller 2aa-library showed that we were unable to extract a binder for the compounds 36, 37, 41, 47, or 49. However, these are very short peptides offering less variability and interactions with a potential binder that might support the turn-on of any of the dyes. Therefore, we moved to our 8aa-library, which offers a very high number of possible different peptides that might provide more specific interactions with a BBA fluorophore. To reduce the number of samples for FACS, we decided to combine similar dyes. We labeled one sample of $2 \cdot 10^7$ cells with all probes that carry the boronic acid in the *ortho*-position (36, 37, 47, 49), one with all the *meta*-compounds (49, 50, 51), and a separate sample for compound 41 due to the different wavelength. After enriching a peptide, we would then prepare a sample for each compound in the mix to resolve which BBA probe(s) were bound and turned on by the respective peptide. In addition to checking the display on a separate sample (Figure 61, A), we stained cells with ZombieUV^{TM} (BioLegend), an indicator of cell viability. This approach aimed to avoid collecting dying and autofluorescent cells that we observed in the 2aa-library confocal imaging. In each round, we collected 0.1-1% of cells with a low fluorescence in the ZombieUV^{TM} channel (ZombieUV–) and the highest observed fluorescence in the BBA probe (BBA+) channel (Figure 61, B). We ensured correct gating and collection settings during the FACS by resorting 100 cells of the collected sample. The increased number of cells in the gate in comparison to the unsorted sample confirmed the correct collection set-up (Figure 62, Backsorting). Multiple sorters had to be used in the different rounds. Therefore, samples from the first and the third rounds of selection were thawed and analyzed by flow cytometry on the same day for direct comparison. The samples were measured for all dyes separately to assess whether a peptide was enriched throughout the three rounds for a specific BBA probe (Figure 62).

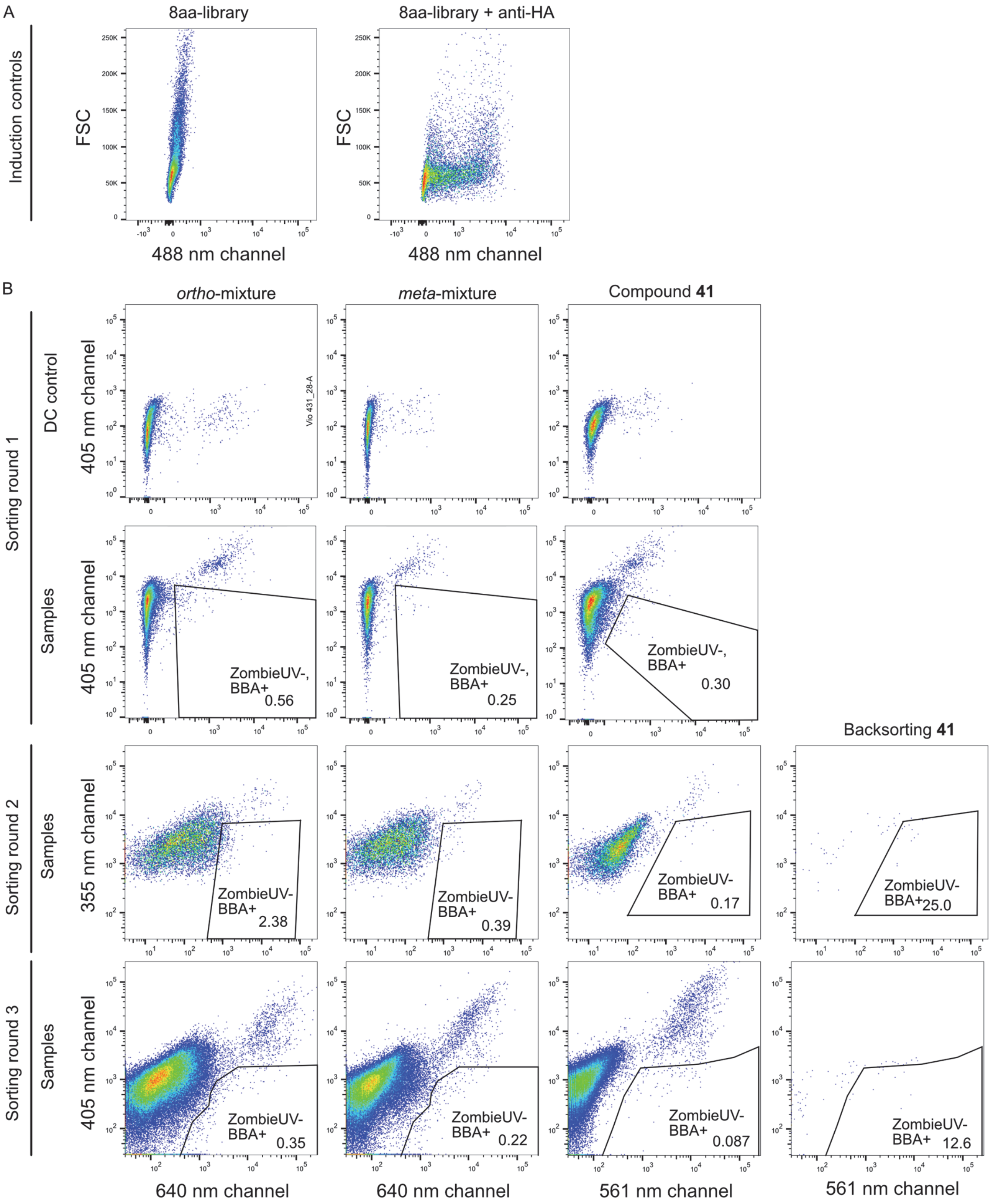


Figure 61. FACS enrichment of the 8aa-library for the two dye mixtures and compound **41** separately. The number in the gate corresponds to the percentage of cells within the gate from all measured cells. A, Induction control example measured on a separate sample from the same batch as used for sorting. B, Chosen gates ZombieUV–, BBA+ for sorting in the three rounds on two different machines. ZombieUV^{TM} was measured in the 405 (402 nm, 431/28 or 405 nm, 450/50) and whenever available the 355 channel (355 nm, 515/30). $1 \cdot 10^4$ cells are shown in each plot except for the last row where $1 \cdot 10^5$ cells are shown.

The flow cytometry analysis indicated that the enrichment was not successful (Figure 62, Q3). The number of cells with a high background fluorescence in a DC sample that had been stored for two weeks at 4 °C (Figure 62, Q3 first rows) was higher than in the samples of cells from later rounds. The percentage of cells in the desired Q3 gate decreased or remained constant from the first to the second round on a generally low level (Figure 62, middle and bottom rows). These results showed that none of our BBA probes was able to bind to any displayed peptide and turn fluorescent under the screened conditions.

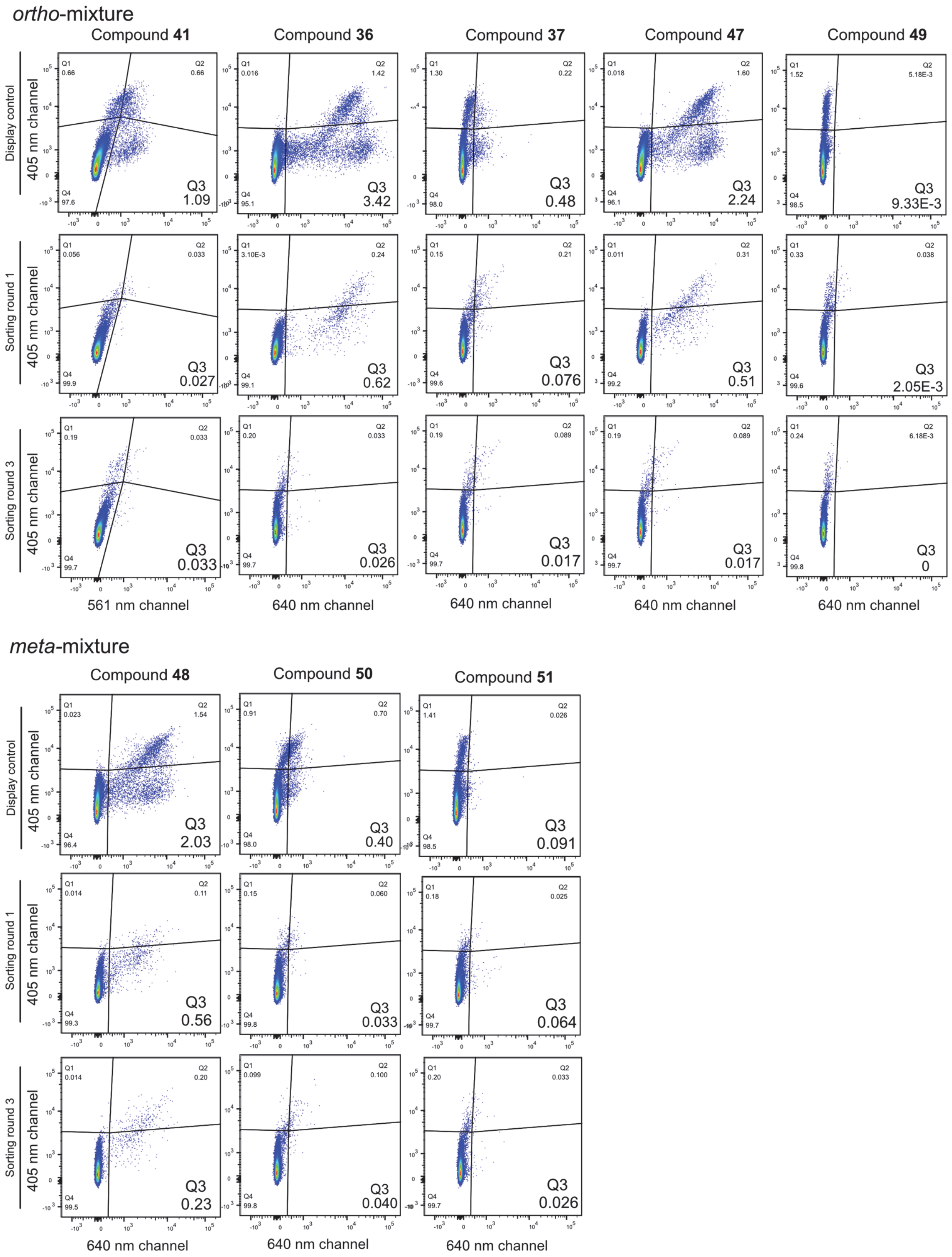


Figure 62. Flow cytometry analysis of DC cells (top row) and yeast cells collected during the first (middle row) and third rounds (bottom row) of the cells sorted of the *ortho*-BBA (top panel) and *meta*-BBA mixture (bottom panel). The number in the gate corresponds to the percentage of cells within the gate from all measured cells. $1 \cdot 10^5$ cells are shown in each plot.

5.4.4 Confocal imaging of sorted yeast cells with increased dye concentration

The results from flow cytometry lead us to question whether the chosen labeling conditions (1 µM, 1 h) were too stringent. The concentration of 1 µM was chosen based on the reported K_D of ~400 nM of RhoBo to its tetraserine tag.[141,142] However, in our, we do not necessarily require such a small K_D given the low background fluorescence. Furthermore, if a binder with a substantially higher K_D could be discovered, the analysis of the binding mechanism and individual influence of the amino acids in the sequence could allow for rational design toward an optimized peptide sequence that might not have been present in our library or a peptide of different length. To see if we could observe cells with the expected binding behavior, we used the cells from the flow cytometry analysis (5.4.3) in an imaging experiment with increased dye concentrations (10 µM) (Figure 63 and Figure 64). However, we observed that all the BBA probes based on SiRhoBo were insoluble at 10 µM concentration and formed precipitates that seemed to adhere to the yeast cells (Examples in Figure 63 and Figure 64, right). For all SiRhoBo derivatives, faint staining of the cells was observed at increased contrasts (Figure 63). However, this staining was very homogeneous over all cells in the sample, but we know that only a subset of cells will display a peptide-Aga2p-fusion at all (e.g., Scheme 22 and Figure 60). Furthermore, the same pattern was observed in the DC and uninduced cells that were kept in SD medium (Figure 64, left).

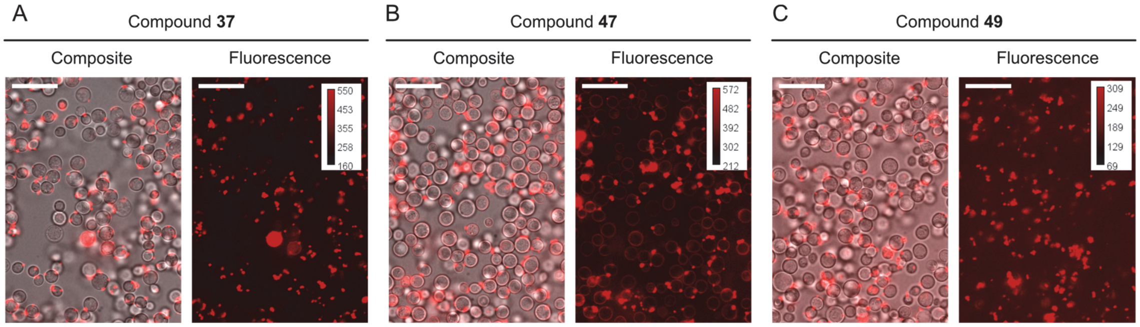


Figure 63. Confocal images of yeast cells after the third sorting round incubated with different *ortho*-BBAs including fluorinated and cyanamide derivatives at 10 µM. A, Cells stained with SiRhoBo **36**. B, Cells stained with fluorinated SiRhoBo **47**. C, Cells stained with fluorinated cyanamide SiRhoBo **49**. The color bar represents the fluorescence intensity (a.u.). Scale bar = 20 µm.

The rhodamine-based compound **41** was soluble at 10 µM. This compound showed cell wall staining more clearly than the SiRhoBo derivatives but also in the DC sample and uninduced cells. Hence, this binding must be either unspecific or specific to a cell wall component of yeast that is not unique to the displayed construct (Figure 64, left). Based on these observations, **41** is unsuitable for screening in yeast display at higher concentrations due to background staining and the SiRhoBo derivatives due to their low solubility. Therefore, more soluble SiRhoBo analogs need to be designed and synthesized to allow for further screening for binders with low affinity and hence at higher concentrations.

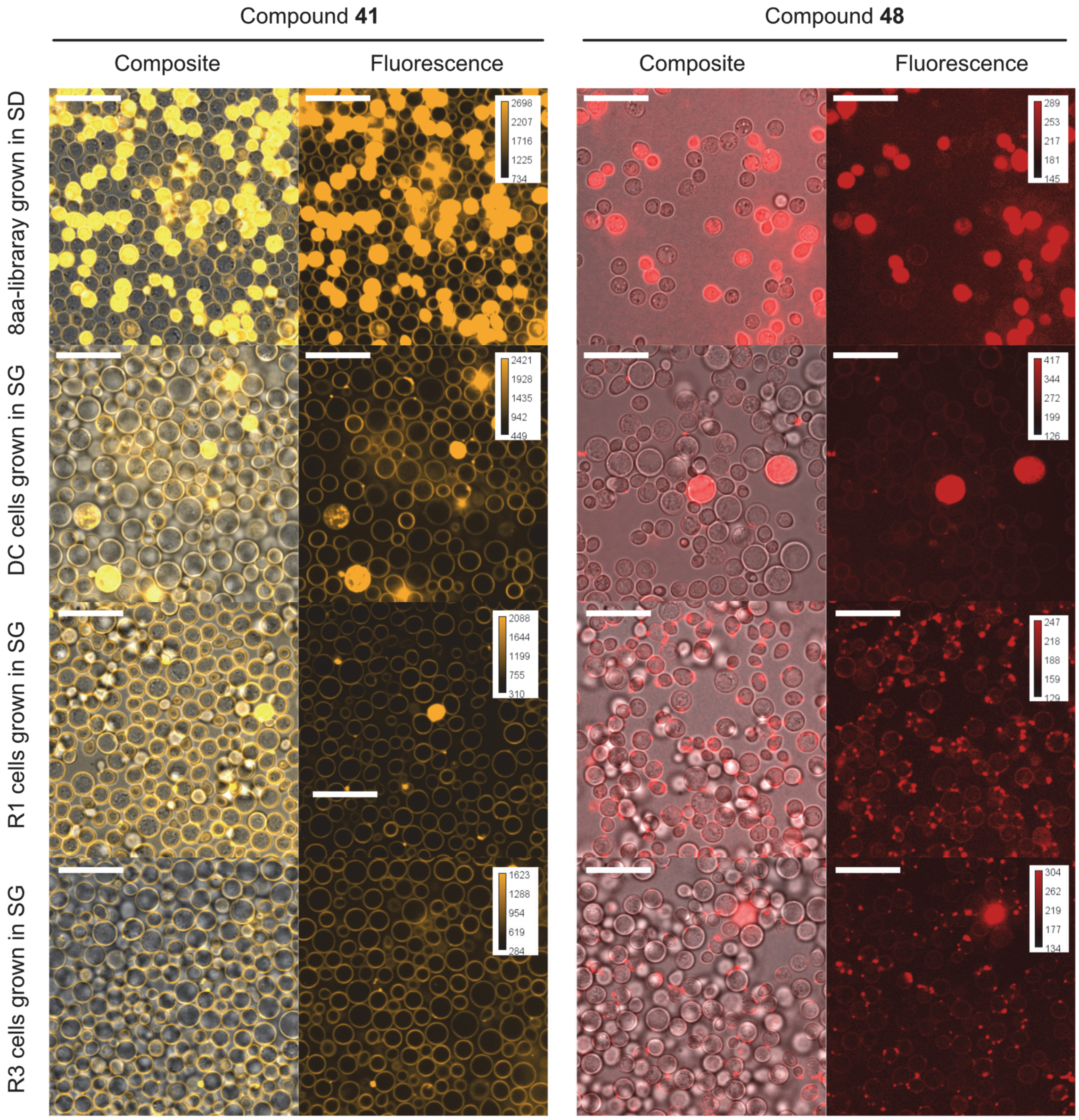


Figure 64. Confocal and composite images of the fluorescence and brightfield images of yeast cells from the original 8aa library kept in SD medium, DC cells in SG, and cells from round 1 (R1) and 3 (R3) grown in SG medium. The cells were stained with compounds 41 (left) and 48 (right) at 10 µM concentration for at least 1 h and imaged in confocal microscopy (left column: 515 nm, 400 ms, 80 mW, right column: 638 nm, 200 ms, 80 mW). The color bar represents the fluorescence intensity (a.u.). Scale bar = 20 µm

5.5 Conclusion and outlook

In this chapter, we presented the efforts toward the development of a small peptide tag containing two serine pairs and a fluorogenic BBA small molecule binding partner. The binding was predicted to be analogous to RhoBo, a compound published as a tetraserine binder in the work of the Schepartz laboratory.[141] We aimed to combine the binding of a fluorophore with two boronic acid moieties and the approach toward fluorogenic dyes proposed by the laboratory of Kai Johnsson. We hypothesized that the resulting fluorophores would be non-fluorescent in the unbound state but become fluorescent once bound by a tetraserine-containing peptide. We synthesized a set of probes that exhibited low background fluorescence in spectroscopic measurements and live-cell imaging in HeLa cells. Furthermore, we created yeast display libraries that express two serine pairs separated by a stretch of randomized amino acids. The two libraries that were implemented included either two or eight amino acids. The diversity of the libraries was confirmed by yeast colony PCR and NGS for the larger library. Screening the prepared libraries for binding and consequent fluorescent turn-on by FACS failed to select a peptide. From the data we gathered, it is unclear whether the problem lies in the binding or a lack of stabilization of the open form, as both result in the same experimental outcome. However, we hypothesized that the chosen experimental setting imposed too stringent selection conditions. The concentration applied during FACS and the dilution prior to sorting require a very strong binder. Followingly, the low solubility of the tested compounds hampers the use of higher concentrations despite the low background fluorescence of the current probes.

First, the conditions and set-up during sorting must be improved. The current protocol for FACS-based peptide enrichment could be optimized in terms of sample preparation and selection with a positive control such as RhoBo or FlAsH. We know that RhoBo binds with high affinity to the peptide -SSPGSS-.[141] This peptide should be present in the 2aa-library considering the small diversity. Alternatively, this specific peptide displayed in tandem with the Aga2p system could be prepared separately and added to the 2aa-library. Under optimized conditions, we should be able to enrich the peptide when applying RhoBo to this library. In addition, it could be valuable to reintroduce the anti-HA antibody in addition to a probe for staining dead cells. Using this combination could allow us to discern highly displaying cells, good binders, and dead cells. Testing such an approach initially on the RhoBo system could give a conclusive answer to whether antibody binding could potentially interfere with BBA probe binding and should indeed be excluded from the FACS panel for future experiments. Furthermore, we could test other aspects of the experiment, such as incubation times and the handling during the preparation of the FACS sample: washing the cells in fresh buffer or keeping them at the chosen dye concentration in the FACS tube.

Second, the design of our probes needs to be re-evaluated. Optimization of the probes should prioritize enhancing solubility. We imagine that cell permeability can be optimized afterward, as cell entry is not required during surface display-based screening. A potential first step could be the introduction of a PEG chain on the lower ring (Figure 65, 54). To estimate the solubility of proposed compounds we can consider the predicted partition coefficient (cLogP) that describes the ratio of concentrations of a compound, in immiscible solvents such as 1-octanol and water. The addition of a PEG chain with 3 units (Figure 65, A, 54) would lead to a decrease in predicted cLogP values from ~6.4–7.0 of our initial cyanamide compounds to ~5.6 and hence to a similar range as compound 41, which was observed to be more soluble. The PEG chain could be introduced via a Click reaction after replacing phthalic anhydride with its 5-TIPS-alkyne derivative 55 analogous to phthalide 3 reported herein (Figure 65, B).

We hypothesize that the solubility of our probes is significantly diminished by the two hydrophobic phenyl rings carrying the boronic acid in addition to the necessary dye core. Therefore, we propose to replace the *ortho*-aminomethylboronic acid moiety with an alternative group to introduce the boronic acids. This replacement would not interfere with the fluorogenicity of the probe, as our design is based on the nucleophilic ring closing onto the xanthene core and does not rely on the boronic acid moiety for fluorescence quenching. Replacing the phenyl moiety with an ethyl linker lowers the predicted cLogP by approximately 5 units and hence should lead to very soluble probes (Figure 65, 56, 57, and 58). Alkyl boronic esters have been reported to effectively bind diols and hence should not preclude tetraserine binding.[295,296] The ethyl linker would be the preferred option over a methylene linker due to the stability concerns of α-amino boronic acids.[297] The ethyl boronic acid moiety could be introduced via a reductive amination from intermediate 59 using a reported α-MIDA boryl aldehyde (Figure 65, B).[298,299] These probes could allow testing of peptide enrichment of the 2aa- and 8aa-yeast display library at increased concentrations and hence allow for a binder with decreased affinity.

As an alternative to modifying the dye scaffold, the peptide library could be enhanced with additional amino acids flanking the diserines to potentially create a more defined structure of the peptides, improving the binding due to better preorganization. Designing improved target peptide libraries could be supported using new protein design tools. A possibility could be to start from a binding site grafted onto a protein followed by minimization of the required scaffold to a shorter sequence.[55,300,301] Sequences obtained from modeling could be randomized at selected positions and screened using yeast display and FACS.

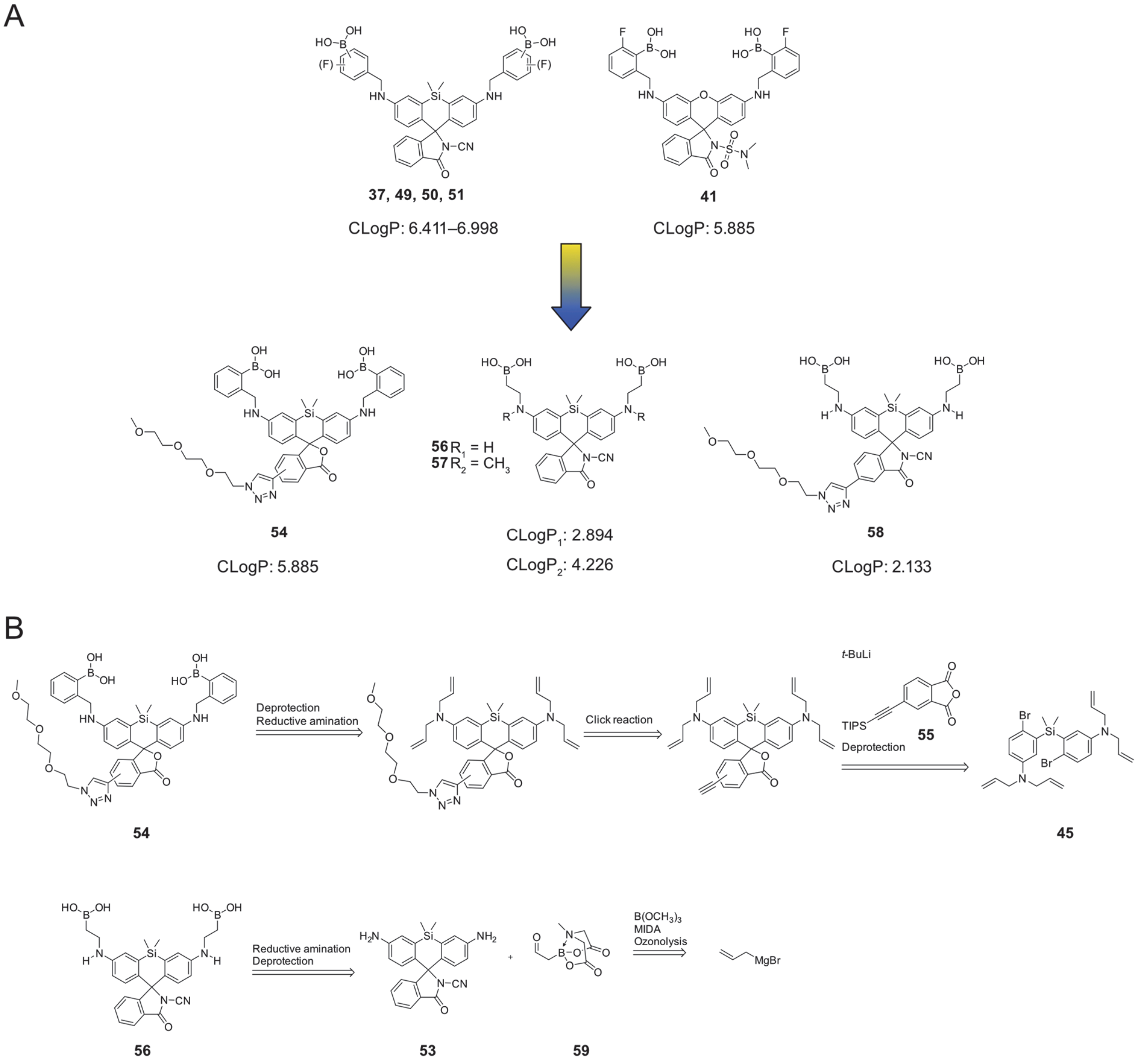


Figure 65. A, Proposed BBA probes predicted to improve solubility. cLogP values were predicted using ChemDraw 22.2.0. B, Retrosynthesis of proposed probes 54 and 56.

Chapter 6 Conclusions and perspective

6.1 Conclusion

In this thesis, we describe the efforts toward developing protein detection methods not only at the single-molecule level but also at the cellular level. We first proposed a new approach to detect peptides or proteins by employing spontaneously blinking dyes that have been developed in other contexts. We took advantage of the environment-dependent blinking of spontaneously blinking fluorophores for the distinction of peptide and protein species in a data-driven approach. The technology uses the fluorescence of these dyes after covalently linking them to the PPOI or POI. The fluorescence intensity over time is used to recognize the PPOI or POI using a model that has been trained on clean training data of the respective PPOI or POI conjugate.

To fundamentally verify our hypothesis, we prepared peptide HMSiR conjugates synthetically by SPPS. We chose three sets of short proteins (five to 16 amino acids) to explore the sensitivity of the method. All peptides in the sets carry a cysteine modified with HMSiR, they differ either in the peptide sequence and hence the polarity, the phosphorylation states, or the epimerization of certain residues. We then developed the workflow from surface preparation, immobilization, TIRF imaging, and data extraction to data analysis. Our initial classification efforts based on manually extracted features of the fluorescence time traces failed. However, applying a DL model architecture based on convolutional layers, which extract features directly in the training, and GRU layers proved successful. The results could be further improved by implementing uncertainty-based filtering. With these results supporting our hypothesis we moved to test whether the method would work on larger substrates such as proteins in pure samples. We adapted our workflow to avoid the need for direct immobilization of the proteins but instead encapsulated them in liposomes, which can be attached to the surface. The first proteins we tested were the self-labeling proteins HaloTag7 and SNAP$_f$ and an unrelated protein sCGrx1p. Using HaloTag7 we could compare the sensitivity toward different sites of the same protein. Preliminary results indicate that the method is sensitive enough to distinguish proteins and multiple locations on the same protein.

The initial proof of blinkognition and the experiments with proteins presented herein demonstrate the high sensitivity of the method but also reveal potential fundamental limitations. Blinkognition is not sensing the sequence but an integration of interactions of the spontaneously blinking fluorophore with the environment resulting in a modulated blinking pattern, which means that any varying interaction combinations resulting in a similar tendency might result in a similar fluorescence output. Multiple sites on different proteins will likely provide a similar combined influence on the blinking behavior of the fluorophore and hence would not be distinguishable and limit the number of proteins that can be analyzed. However, if we can combine signals from different sites by multiple labeling of proteins, they would only be indistinguishable if all sites of each color channel provide the same overall influences on the fluorophores. Consequently, each additional fluorophore would increase the number of proteins that can be distinguished theoretically. However, these considerations are based on theoretical assumptions in well-defined systems. In any practical application such as a complex sample, many additional challenges and limitations arise and need to be evaluated in the application of blinkognition.

A challenge inherent to all methods requiring protein modification for detection is incomplete labeling. The numerous labeling variants created, lead to reduced accuracy due to likely misclassification. In addition, protein modification can result in sample loss due to precipitation, which then alters the sample composition. These factors will have to be addressed in the future by experimentation and optimization of the number of fluorophores used and the labeling strategy to obtain optimal results. Another limitation in the currently proposed form of blinkognition is that for each protein that should be recognized a pure sample needs to be prepared and trained for, which limits the number of proteins for classification in terms of time but also computational capacity. Therefore, it will be crucial for practical application in complex samples to develop a means to handle OOD data so that the targeted proteins can be recognized among unknown signals. Alternatively, if a correlation of blinking patterns with surface properties, i.e. a set of predictable surface descriptors, can be established even unknown signals might provide information on what kind of protein or a range of potential proteins are likely to give rise to the recorded signal and might offer an alternative application of blinkognition. In conclusion, the work herein provides evidence of the validity of the fundamental hypothesis, but future exploration and developments are required to demonstrate the practical applicability of blinkognition in complex settings.

Whereas the first part of this work was focused on providing proof-of-principle and initial exploration of the scope of a highly sensitive protein fingerprinting method that could potentially be used for protein detection or profiling, the second part has been focused on further developing tools to target a POI in cells. Fluorescent labeling of a POI in live cells aims at understanding the protein function in context such as expression, localization, interactions, and temporal changes. To observe proteins of low abundance live, we require bright labeling with a low background, which can, for instance, be achieved with self-labeling proteins. These combine the good photophysical characteristics of small molecule fluorophores and the absolute specificity of genetic fusions. For some of these proteins, fluorogenic dyes have been developed to tackle the background problem.

However, a problem that persists is the large size of these N- or C-terminal tags, which can lead to issues with proper folding or localization of the POI.

We sought to improve a BBA-modified rhodamine dye RhoBo, which has been shown to bind a tetraserine peptide with high affinity but produces significant background fluorescence in cells.[141] To reduce background we synthesized potentially fluorogenic BBA analogs by introducing an electron-deficient amide into RhoBo. Furthermore, we derivatized the scaffold by introducing a *gem*-dimethyl silicon resulting in a silicon rhodamine analog with and without electron-deficient amides. Moreover, we introduced fluorine atoms to improve the stability of the boronic acids and generate a wider range of boronic acid reactivity. Lastly, we explored SiRhoBo derivatives with a wider interatomic distance between the boron atoms, which could allow for the binding of a larger tag. While synthesizing these compounds, we prepared two yeast display libraries that carry tetraserine peptides with randomized amino acids between the serine pairs. In the first library, two amino acids were randomized, in analogy to the RhoBo peptide, and the second eight amino acids were introduced, which could be beneficial for the larger spanning compounds and would additionally provide less background due to natively present tetraserines. These libraries were tested with the BBA probes in FACS-based screening. However, our efforts remained unsuccessful. We hypothesized that the applied concentration demanded a high-affinity binder, which would not be necessary for a fluorogenic probe. However, increasing the concentration proved impossible due to the insolubility of the dyes and will require design probes with improved solubility.

6.2 Perspective

In the introduction, we discussed the challenges involved in analyzing the proteome. Briefly, the proteome is diverse in chemistry and numbers. The diversity arises from the large number of building blocks (20 amino acids) and mechanisms that diversify the 20,000 protein-encoding genes during the expression in a non-linear and unpredictable manner, additionally, they can be modified post-translationally. Single-cell proteome analysis faces the hurdle of analyzing a small amount of highly diverse proteins with varying copy numbers in a large range. Current methodologies either lack sensitivity or throughput. Furthermore, they exhibit substrate biases, e.g., through stability, solubility, ionization (MS), labeling, or immobilization challenges (fluorescence-based methods).[8,233,239,251] Currently, only a few methods have the potential for protein sequencing; nanopore and MS^2 experiments most other methods work as protein fingerprinting or profiling methods.

Identifying all proteins in a cell, despite high-throughput sequencing, would remain a time-consuming and resource-intensive task, with the majority of read sequences corresponding to the most abundant protein.[8] Therefore, the author of this thesis is convinced that even with the option to sequence all proteins, it would rarely be necessary. Instead, methods to fractionate the proteome will be crucial at least to remove the most abundant proteins. Then full proteome sequencing would only be required or beneficial to validate the fractionation ensuring that no biases and unintended alterations of the proteome are introduced. In MS analysis fractionation is most used in the form of LC-MS(/MS) during bottom-up approaches, whereas the full-protein MS has developed many gel-based (1D or 2D) fractionation approaches before sample analysis. Another method is immunoprecipitation of the most abundant proteins (immunodepletion) which significantly decreases the number of proteins to be analyzed and hence lowers the dynamic range and its masking effect.[8] The MS community has put great effort into improving sample preparation and isolation to reduce sample loss and biases introduced due to steps preceding the actual measurement. It could be of great value if the single-molecule technologies combined their sensitivity with these advances to explore the analysis of more complex samples. This could lead to promising applications, especially in biomarker detection. In these applications, it is not required to measure a full proteome but to quantify one or a limited number of targets with high sensitivity.

Hence, a heavier focus on testing new and innovative methods on more complex samples, and addressing fundamental challenges and biases could have a great impact. For instance, in improving labeling chemistries and surface attachment methods. The current most used reagent for N-terminal modification, 2-pyridinyl carboxaldehyde (2-PCA) has been successfully applied for surface attachment. However, 2-PCA cannot react with *N*-terminal prolines and has known substrate preferences.[233,239,251] Alternatively to small molecules, enzymatic N-terminal modifications have been proposed. Although proteins also exhibit substrate preferences, they can be engineered to be more promiscuous or a mixture of proteins with different substrate preferences can be combined. Protein-based approaches could be promising, especially with the currently emerging and highly promising protein prediction/design models.[194,195,302] Optimizing these aspects could allow for the successful application of single-molecule fingerprinting and protein profiling methods to be applied for targeted analyte detection even in complex mixtures such as those required for biomarkers.

Whereas broad proteome analysis is in its beginning, many approaches have been conceived to target a POI in a cell. Although our approach toward a small peptide tag binding a fluorogenic dye has not delivered a desired dye peptide pair, we have obtained important insight into possible future steps that could lead to a successful tag. The yeast display library has been constructed successfully and is of high diversity, therefore, it can be used to perform screens with improved fluorophores. As we proposed in Section 5.5, the crucial next steps will be to improve the dye properties toward more soluble dyes. Concerning the tetraserine libraries we have only varied the amino acid sequence between the serines resulting in extremely small peptides. However, a slightly larger tag that includes amino acids on the C- and N-terminus of the tetraserine could benefit the formation of a defined conformation supporting dye opening, similar screens using mammalian cells as the display platform, have discovered extended target peptides containing a tetracysteine for FlAsH.[287] Further developments for a larger tetraserine tag could include a proximity-induced labeling mechanism that employs an additional reactive moiety in the fluorophore which can be attacked by an amino acid in the tag. Similar approaches have been proposed for a tag called “fluorette” or a cysteine-alanine-tetraaspartate (CA_6N_4)-tag that is bound by a zinc(II)-ion based complex with a reactive chloroacetamide moiety.[303,304] Such a cysteine-tetraserine system could be employed to support the reversible boronic acid-binding. In addition to screening approaches, specific interactions could be designed with the newly emerging protein design algorithms, especially with further developments including potential ligand interactions in a newly designed structure.[194,302,305,306] Combining such predicted structures could be further a way to create improved, biased libraries for screening and hit identification. Successful hits would result in a small peptide tag with fluorogenic dyes that can be used to image low-abundance proteins with minimal background and interference with protein folding, localization, and function.

Chapter 7 Materials and methods

7.1 General remarks

7.1.1 General synthetic methods and analyses

All reagents were purchased from commercial sources and used as received. Anhydrous solvents were procured from Acros Organics and used as received. All solvents used in the preparation of the coverslips were HPLC grade. Flash column chromatography was carried out using prepacked Buchi Reveleris SiO_2 or C_{18} cartridges on a Buchi Reveleris PREP instrument or a Biotage® Selekt System. Reverse-phase HPLC was conducted on an Agilent 1290 Infinity II System with a C_{18}-column (Phenomenex). MW reactions were performed with a Biotage Initiator. TLCs were run on SiO_2-60-F_{254} plates (Sigma-Aldrich) and visualized using UV light at 254 nm and 366 nm or permanganate stain. Nuclear magnetic resonance (NMR) spectra were acquired on a Bruker Ultrashield 400 instrument. ^1H NMR chemical shifts are reported in ppm relative to $SiMe_4$ (δ = 0) and were referenced internally with respect to residual protons in the solvent (δ = 7.26 for $CDCl_3$, δ = 1.94 for CD_3CN, δ = 3.31 for CD_3OD and δ = 3.58 for THF). Coupling constants are reported in Hz. ^{13}C NMR chemical shifts are reported in ppm relative to $SiMe_4$ (δ = 0) and were referenced internally with respect to solvent signal (δ = 77.16 for $CDCl_3$ and δ = 1.32 for CD_3CN). Preliminary peak assignments are based on calculated chemical shifts. Multiplicities were abbreviated as follows: singlet (s), doublet (d), triplet (t), quartet (q), multiplet (m). High-resolution MS (HRMS) was conducted by the staff at the ISIC Mass Spectrometry facility (at EPFL) employing a Waters Xevo G2-2 quadrupole time-of-flight (QTOF) or Agilent Technologies 6530 Accurate-Mass QTOF LC-MS or by staff at Mass Spectrometry Laboratory of the Department of Chemistry at the UZH employing a Dionex Ultimate 3000 ultra-HPLC system (ThermoFischer Scientific) connected to a QExactive MS with a heated ESI source (ThermoFisher Scientific). Low-resolution mass spectra of peptides and small molecules presented here were acquired on a Shimadzu LC-MS 2020 spectrometer with ESI and a single quadrupole detector. IUPAC names of all compounds are provided and were determined using CS ChemBioDrawUltra 22.2

7.1.2 Fluorescence microscopy

7.1.2.1 Fluorescence microscope Nikon N-STORM BIOP:

The Nikon N-STORM microscope (Nikon) is equipped with an SR Apochromat TIRF 100x 1.49 N. A. oil immersion objective lens. A piezo-electronic focus-lock system (perfect focus system) was used to prevent axial drift during data acquisition. The illumination powers of the solid-state lasers were measured at the tip of the optical fiber as 561 nm, 70 mW, and 641 nm, 300 mW. The emission was passed through filters with bandpass windows at 677–740 nm. Fluorescence was detected with an electron-multiplying charge-coupled device (EMCCD) camera Andor iXon3 897. The microscope was operated using the NIS Elements (Nikon) software.

7.1.2.2 Fluorescence microscope Nikon CSU W1 EPFL:

The Nikon N-STORM microscope (Nikon) is equipped with an SR HP Apochromat TIRF (100x, numerical aperture = 1.49) oil immersion objective lens. A piezo-electronic focus-lock system (perfect focus system) was used to prevent axial drift during data acquisition. The illumination powers of the solid-state lasers were measured at the tip of the optical fiber at 488 nm: 118 mW and 638 nm: 112 mW. The emission was passed through filters with bandpass windows at 582–618 nm and 683–783 nm. Fluorescence was detected with an EMCCD camera Andor iXon 888. The microscope was operated using the NIS Elements (Nikon) software.

7.1.2.3 Fluorescence *Nikon N-STORM W1 UZH*

The Nikon W1 spinning disk microscope is a dual-mode microscopy system for confocal and TIRF imaging. It is equipped with a sCMOS dual-camera system (Photometrix) and an EMCCD camera iXon 888 (Andor). TIRF images were collected with an SR HP Apochromat TIRF 100x 1.49 N. A. oil immersion objective lens and confocal images with a CFI Plan Apochromat Lambda D oil immersion objective (60x, numerical aperture = 1.4). A piezo-electronic focus-lock system (perfect focus system) was used to prevent axial drift during data acquisition. Brightfield imaging was performed with a white LED. Available laser lines for fluorescence imaging were 405 nm, 445 nm, 488 nm, 514 nm, 561 nm, and 638 nm. Appropriate filter cubes were configured within the light path. The microscope was operated using the NIS Elements software.

7.1.2.4 Image analysis

General image processing and analysis were conducted with Fiji – ImageJ software or the Python package scikit-image.

7.2 Methods by Chapter

7.2.1 Chapter 2

7.2.1.1 Reagents and analyses

Reagents for PEGylation were purchased from Biopharma PEG Scientific (azide-PEG$_{5000}$-NHS: #HE006023-5K) and Laysan Bio (mPEG$_{5000}$-NHS: #MPEG-SVA-5000-5g). All solvents used in the preparation of the coverslips were HPLC grade. Stock solutions of synthesized chemical probes and peptides were prepared in DMSO spectrophotometric grade > 99.9% (Acros Organics) at a concentration of 20 mM, diluted aliquots were prepared according to experimental requirements and stored at –20 °C. All stock solutions were thawed immediately before use.

7.2.1.2 *Chemical surface cleaning*

Round cover glasses (25 mm, 1.5 H, ThorLabs) were marked on one side at the bottom right with a vertical line (or in any asymmetrical way) with a glass cutter to indicate the functionalized surface at the end of the slide preparation. The slides were sonicated in 10% Alconox^{TM} for 20 min, rinsed with MiliQ water until all the Alconox^{TM} was removed, and further rinsed at least five times by filling the glass staining jar with MiliQ water. The slides were sonicated in MiliQ water for 5 min, rinsed three times with MiliQ water, rinsed once with acetone, and sonicated for 20 min in acetone. The acetone was disposed and the slides were rinsed at least two times more with acetone.

7.2.1.3 *KOH cleaning*

The chemically surface-cleaned slides were transferred into polypropylene containers and rinsed with MiliQ water. The containers were filled with 1 M KOH solution and sonicated for 40 min. The slides were rinsed three times with MiliQ and dried under a stream of nitrogen.

7.2.1.4 *KOH cleaning with Piranha step*

The chemically surface-cleaned slides were transferred into polypropylene containers and rinsed with MiliQ water. The container was filled with 1 M KOH solution and sonicated for 40 min. The slides were rinsed three times with MiliQ and transferred back into the glass container dedicated to the next step. Piranha solution, i.e., a 3:1 mixture of sulfuric acid and 30% hydrogen peroxide, was freshly prepared by adding the hydrogen peroxide solution to the sulfuric acid in a glass container. The solution was allowed to cool down for a few minutes before adding it to the completely dry coverslips. The slides were kept in the solution for 20 min, the Piranha solution was disposed of, and the slides were rinsed extensively. They were dried by a stream of nitrogen and placed into the aminosilation chamber.

7.2.1.5 *Plasma cleaning*

The chemically surface-cleaned slides were placed in a plasma cleaner and treated for 5 min (Harrick Plasma PDC-32G) and then imaged directly.

7.2.1.6 UV/Ozon

The chemically surface-cleaned coverslips were placed on a piece of aluminum foil with the marked side toward the UV lamp cleaner (Jelight Company Inc., UVO-Cleaner Model No. 256-220). They were irradiated for 10 min and left in the chamber for 5 min. The slides were directly used in further steps.

7.2.1.7 *Surface passivation*

The glass containers and Erlenmeyer dedicated to the aminosilation step were sonicated first with 1 M KOH, rinsed with MiliQ five times, sonicated for 20 min in CH_3OH, and dried under a stream of nitrogen. The slides were transferred into the aminosilation containers. The aminosilation solution was prepared in the Erlenmeyer by adding 2 mL of AEAPTMS to 98 mL of acetone and mixing with a dedicated glass rod (for the aminosilation test the original protocol: a solution of methanol with 5% of acetic acid for the preparation of the 1% of AEAPTMS was used).[150] The pre-mixed solution was added to the slides in the aminosilation containers and put in the dark for 10 min. Then the containers were sonicated for 1 min before putting them back in the dark for 10 min. The aminosilation solution was discarded and the slides were rinsed three times with acetone and three times with MiliQ before drying them with compressed nitrogen gas. The slides were placed in the PEGylation container (a pipette tip box with tips to prevent the coverslip from touching the box). For each PEGylation the solution was freshly prepared. Five pairs of coverslips required 1 mg azide-PEG_{5000}-NHS and 32 mg methoxy PEG_{2500}-NHS that were added to an Eppendorf tube and dissolved in 320 µL PEGylation buffer (1 mM $NaHCO_3$). This solution was mixed gently and centrifuged for 1 min at 10’000 rpm. 70 µL of PEGylation solution was added onto the marked side of half of the coverslips and another coverslip with the marked side down was added on top of the solution without introducing any bubbles. Water was added to the bottom of the pipette box and the slides were kept in a dark, well-leveled place for 3 h. The coverslips were taken apart, rinsed with MiliQ and CH_3OH from a squirt bottle, and dried with compressed nitrogen.

7.2.1.8 Click functionalization

Procedure A without copper ligand: The coverslips were placed back into the pipette tip boxes with the passivated surface face up. The click reaction solution containing $CuSO_4$ (1 mM), sodium ascorbate (100 mM), and the alkyne dye (1 nM) (from a 1 µM DMSO stock solution) in a tris(hydroxymethyl)aminomethane buffer (200 mM) at pH = 8.5 was prepared. On the passivated side of the slides, 70 µL of click reaction solution was added and a second cover slip with the passivated surface down was added on top. The coverslips were kept for 2 h in a dark, well-leveled place. The slide pairs were taken apart and rinsed with MiliQ, CH_3OH, acetone, ethyl acetate (EtOAc), hexane, and again MiliQ before drying them using compressed nitrogen. The slides are placed in a container with 10 mM EDTA solution for 10 min, rinsed with CH_3OH and MiliQ, and dried under a stream of nitrogen. The coverslips were either imaged directly or stored overnight in a dry and dark place at 25 °C and imaged the next day.

Procedure B with THPTA ligand: The coverslips were placed back into the pipette tip boxes with the passivated surface face up. For the measurement of three different peptides in one experiment, a master mix with the common components was prepared: 45 µL of a 100 mM THPTA (10 mM, final concentration) was combined with 9 µL of a 100 mM $CuSO_4$ solution (2 mM, final concentration). The reagents were diluted with 351 µL of a 3:1 water/glycerol mixture. From this diluted solution, 135 µL were transferred to a new Eppendorf tube. Then 0.15 µL of the corresponding 1 µM peptide-dye conjugate stock solution was added, followed by 15 µL of 1 M sodium ascorbate. The reaction mix was vortexed thoroughly before applying 70 µL of the final reaction solution to the reactive surface of a passivated glass slide. A second glass coverslip is gently placed on top, with the passivated side facing down, i.e., toward the reaction solution, without introducing any bubbles. After 2 h the slides were separated and rinsed with MiliQ water, 25 mM EDTA solution, MiliQ, acetone, EtOAc, CH_3OH, and finally again MiliQ. The functionalized coverslips were dried under a stream of nitrogen, placed on a KimWipe^{TM}-lined container, and transported to the microscope for imaging.

7.2.1.9 Fluorescence imaging

For imaging, the slides were fixed in a cell chamber (Aireka Scientific) and rinsed once with 1 mL of MiliQ water and at least twice with 1 mL of imaging buffer (10 mM PB at pH = 7.4). The coverslip was then imaged in TIRF mode in the 640 nm channel under constant conditions at 80% laser power with an exposure time of 30 ms. The conditions were kept constant for all the acquisitions. For the acquisition of the dataset used for the analysis and ML in Chapter 2 and 3, only the microscope setup described in 7.1.2.2 was used.

7.2.1.10 *Data analysis*

To perform the classification analysis, we localized the single-molecule signals and extracted the numeric intensity values of the signal over all frames followed by data normalization and filtering to remove noise signals (Table 11). All analysis and classical ML analysis were conducted on a desktop computer with an AMD Ryzen 9 3900X 12-Core, 3800 MHz Processor, a GeForce RTX 2070 Super graphics card, and 128 GB of physical memory.

7.2.1.11 *Fluorescence trace extraction*

In each frame, all the single molecules were localized using the freely available single-molecule localization package Picasso[175] with the settings listed in Table 10. The obtained framewise localizations were combined using the Picasso postprocess module. All the localizations within a two-pixel distance from the first appearance of a localization were combined as one particle and the final location of the particle was calculated from the averaged x- and y-coordination of the combined localization that were detected. No significant lateral drift was detected during the acquisition. The mean locations were exported for all the particles and further used to obtain the fluorescence intensity trace of each particle with custom-written code. A mean location was used to define a box with a side length of 5 pixels and the mean as its center. The pixel intensities within this box were summed up and recorded as the particle intensity for the respective frame, resulting in an intensity trace over all the frames. The boxes that overlapped in more than 2 pixels were not considered, as well as boxes that overlapped with the edge of the field of view. All traces were recorded and saved in reference to their peptide of origin (C11-C14, P15-P17, or E19-E21).

7.2.1.12 *Fluorescence trace preprocessing and filtering*

Traces obtained from the movies needed to be standardized to remove potential influences from the different technical replicates (different coverslips and different days). Therefore, we used the last 500 frames, which were bleached in most traces, to estimate the mean value of the background of the acquisition. This value was subtracted from each frame in the trace. Furthermore, we standardized the fluorescence intensity values by calculating the z-score for each value in the trace. The z-scored traces were used to determine the peaks in each trace using the function find_peaks, contained in the SciPy python library.[176] As a threshold for peak detection, we set a minimum intensity of eight standard deviations (8σ) from the background. These peaks were the basis for the first step of trace filtering to remove spurious traces stemming from background noise or impurities in the experiment. The applied filtering criteria are listed in Table 11.

7.2.1.13 *Feature determination, visualization, and classical ML*

The z-scored traces were used to determine the peaks in each trace using the function find_peaks, contained in the SciPy python library.[176] Peak-based features were calculated from the found peaks and their properties. The peak width corresponds to the ON-time of the fluorophore, while the time between peaks corresponds to the OFF-time of the peaks. The total blinking time corresponds to the time from the left bounds of the first peak to the right bounds of the last peak. The approximate signal area is the product of the peak height and the peak width. For these aspects of each trace and its peaks, we determine the maximum, minimum, mean, and standard deviation values.[307] Furthermore, we used the librosa package to calculate the tempo of the trace, when interpreting the signal trace as a musical onset envelope, which exhibited a low variation in all the traces.[177] The Fourier transformation of some traces was calculated for visual inspection but not used further for feature engineering. For visualization and inspection of the features, correlation plots were used with a diagonal corresponding to density-normalized distribution plots. The PCA implementation of scikit-learn was applied using the standard normalized features. For classical ML models, the features were normalized using RobustScaler in scikit-learn.[181] The normalized features were used to train a selection of classical ML models using a grid search with five-fold CV using a seed 13102022.[181]

7.2.2 Chapter 3

The experimental procedures for the sample preparation and acquisition of the single-molecule movies raw data are described in Chapter 2 (the same raw data was used for analysis). For the slide preparation this corresponds to UV/Ozone cleaning, click procedure B, and for image acquisition the fluorescence setup described in 7.1.2.2.

7.2.2.1 *Data analysis*

To perform the classification analysis, we localized the single-molecule signals and extracted the numeric intensity values of the signal over all frames followed by data normalization and filtering to remove noise signals (Table 11). For DL, the standardized and normalized traces were used directly as input for the model. All preprocessing and deterministic ML analysis was conducted on a desktop computer with an AMD Ryzen 9 3900X 12-Core, 3800 MHz Processor, a GeForce RTX 2070 Super graphics card, and 128 GB of physical memory. The DL model was trained using the GPU. The MCD models were trained on the ScienceCluster (Tesla V100-SXM2-32GB GPU) and evaluation and analysis on a ScienceCloud virtual machine instance using 8 cores of an AMD EPYV 7702 Hypervisor CPU and an nVidia Tesla T4 GPU, service infrastructure provided by S3IT (www.s3it.uzh.ch), the Service and Support for Science IT team at the UZH.

7.2.2.2 *Data augmentation for DL*

We generated traces with a mirror-image peak region. We defined the peak region between the left-bound of the first and the right-bound of the last detected peak. This partial sequence was inverted in place, to obtain the trace with a mirrored peak region.

7.2.2.3 *Deep-learning settings using a one-dimensional convolutional neural network (1D-CNN) classification using a deterministic approach (1D-CNN-GRU)*

The DL model was implemented using Python and TensorFlow 2.7 [210] with the Keras API.[308] The hyperparameters used in the model are described in Table 17. We used Adam[309] optimizer, with default settings, with a mini-batch size of 32 for the compounds peptides set C11-C14 and P17-P21 and a size of 128 for the set E19-E21. The traces obtained after preprocessing, filtering, and data augmentation were used for the DL approach. The z-scored and min-max-scaled traces were used as separate features for each trace and directly used as input vectors to the model. To prevent overfitting and stop the training at the best performance on the validation set, we used early stopping with a patience of 10 epochs. For NCV we report the mean value of all the folds in the confusion matrices and the variation of the true positive rate along with the mean overall accuracies and standard deviations of the model evaluation on a separate test set that was not used in model training or validation. We performed all the modeling experiments using a seed of 42 unless mentioned otherwise.

7.2.2.4 *MCD implementation (1D-CNN-GRU-MCD*)

The traces were preprocessed as in the deterministic approach and the same input was used for the probabilistic model 1D-CNN-GRU-MCD (Table 19). To make the model probabilistic, custom Dropout and GRU layers were used, which use MCD at inference time. As MCD was used in the GRU layer itself, the input Dropout layer was removed. For evaluation, 100 predictions were calculated for each trace in the validation/test set yielding 100 potentially different predictions depending on the dropped nodes by the dropout mechanism. The mean model output over the 100 predictions was calculated per class used to predict the label.

7.2.2.5 *Definition of the classification CT for trace retention*

The uncertainty quantification was calculated as the Wasserstein distance (using the SciPy implementation)[176] between the class with the highest mean probability over 100 predictions and the closest other class in the classification problem. Therefore, the prediction with the highest mean was determined and the Wasserstein distances for the other two or three classes in the problem were calculated, the minimal distance was determined and stored. A CT value was chosen (0 ≤ CT ≤ 1) and used to filter uncertain traces if their minimal distance was larger than the chosen CT (CT = 0.7 for all cases in this work). All traces that were not filtered out based on the CT were used to calculate the accuracy score.[181]

7.2.3 Chapter 4

7.2.3.1 *Reagents and analyses*

Lipids for encapsulation were ordered from Avanti Polar Lipids. (DMPC: 85034P, DPPC: 850355, 15:0 PC: 850350C, SMPC: 850464C, 16:0 NBD PE: 810144P 16:0 Biotinyl Cap PE: 870277P). The fluorescent lipids for the two-color vesicle experiment were obtained from ATTO-TEC (DPPE-Atto425 and DPPE-Atto520). Reagents for PEGylation were purchased from Laysan Bio. (biotin-PEG_{5000}-NHS: #Biotin-PEG-SVA-5000-1g) and PLL-g-PEGs from SuSoS (biotinyl PLL-g-PEG: #PLL(20)-g[3.5]-PEG(2)/PEG(3.4)biotin20%, PLL-g-PEG: # PLL(20)-g[3.5]-PEG(2)). DLS measurements were conducted on a Litesizer 500 (Anton Paar) using a refractive index of 1.334 and the absorption settings for proteins were applied. Stock solutions of synthesized chemical probes were prepared in DMSO spectrophotometric grade > 99.9% (Acros Organics) at a concentration of 20 mM, diluted aliquots were prepared and stored at –20 °C. All stock solutions were thawed immediately before use. For image acquisition, the fluorescence setup described in 7.1.2.3 was used.

7.2.3.2 *Protein expression*

BL21(DE3) competent cells (NewEngland Biolabs) were transformed with the respective plasmid by heat shock following the provided standard procedure of the vendor. LB medium was prepared with the appropriate antibiotic. A single colony was inoculated into the LB medium and incubated at 37 °C overnight. The desired volume of LB medium was inoculated from the starter culture and incubated at 37 °C until OD=0.4–0.8 was reached. Isopropyl β-D-1-thiogalactopyranoside was added to a final concentration of 0.5–1 mM and the culture was further incubated at 18 °C overnight unless stated otherwise. *E. coli* cells were harvested by centrifugation. For culture volumes > 500 mL, the bacterial pellet was resuspended in lysis buffer (300 mM NaCl, 50 mM PB at pH = 7.5, 10% glycerol, 50 mM tris(2-carboxyethyl)phosphine) as well as Turbonuclease (5 µL) and a protease inhibitor cocktail tablet (Roche). The cells were lysed by sonication (70% amplitude,10 s pulse-10 s pulse off for 2.5 min). The lysate was cleared by centrifugation and the protein was purified by Ni-His-affinity chromatography in batch mode on HisPur™ Ni-NTA resin (ThermoFisher Scientific). Fractions as well as induction controls were analyzed by SDS page gel electrophoresis and pure fractions were pooled and dialyzed against PBS.

7.2.3.3 Optical spectroscopic measurements for protein quantification

Absorbance was measured on a Multiskan SkyHigh (ThermoFisher Scientific) with a µDrop Duo Plate.

7.2.3.4 *Protein labeling reaction*

All proteins used were stored in PBS, if another reaction buffer is mentioned the buffer was exchanged via Zeba desalting spin columns (ThermoFisher Scientific) according to the manufacturer's protocol. Protein concentration was determined using the nanodrop plate for the device MultiSkan Sky (ThermoFisher Scientific). The protein was then prediluted to the reaction concentration in the respective buffer in Eppendorf tubes. The labeling was started with the addition of the reactive dye (compound 22). If multiple samples were measured without dye removal, the reactions were started with a delay time corresponding to the LC-MS runtime. The reactions were then shaken at 600 rpm in a thermoshaker (ThermoFisher Scientific) at the given temperature and time before the sample was transferred into a conical LC-MS vial or a conical 96-well plate for measurement. The LC-MS autosampler was cooled to 8 or 10 °C.

7.2.3.5 *Purification of labeling reactions*

Final protein concentrations of purified and labeled proteins were determined by bicinchoninic acid assay using the Pierce BCA Protein Assay Kit for labeled proteins (ThermoFisher Scientific).

7.2.3.6 *Protein LC-MS and analysis*

Initial data for HaloTag7 and its mutants were acquired on a Dionex Ultimate 3000 ultra-HPLC system (ThermoFisher Scientific) equipped with a DAD-3000 diode array detector and connected to a maXis QTOF high-resolution mass spectrometer (Bruker Daltonics, Bremen, Germany); injection of samples with an XRS autosampler (CTC); Acquity BEH C4 UPLC column (1.7 µm particle size, 2.1 × 50 mm, Waters) kept at 30 °C; with A: H_2O + 1% HCO_2H + 0.04% TFA and B: CH_3CN + 0.1% HCO_2H + 0.04% TFA, full scan MS in (+)-ESI mass ranges 300–2’500 m/z. Data were analyzed on Compass HyStar DataAnalysis 5.2 (Bruker Daltonics). sCGrx1, its mutants, and SNAP measurements were acquired on an Agilent 1290 Infinity II instrument coupled with an Agilent InfinityLab LC/MSD using the OpenLab CDS software. Data were calibrated with mass spectra in MassWorks ver. 4.0.0.0 (Cerno Biosciences) with reference LC/MS tuning mix, for ESI (Agilent) and deconvoluted using SAMMI (Cerno Bioscienes).[243]

7.2.3.7 *Proteomics sample preparation by staff members at FGCZ*

The proteins were alkylated by adding chloroacetamide to a final concentration of 15 mM. The samples were incubated for 30 min at 30°C; 700 rpm and protected from light. Enzymatic digestion was done in buffered trypsin solution at pH 8 (10 mM Tris and 2 mM $CaCl_2$). Following enzymatically digests the samples were dried.

7.2.3.8 LC-MS/MS data acquisition by staff members at FGCZ

MS analysis was conducted by LC-MS/MS (Orbitrap, ThermoFisher Scientific). The digested samples were dissolved in aqueous 3% CH_3CN with 0.1% formic acid, and the peptide concentration was estimated with the Lunatic UV/Vis absorbance spectrometer (Unchained Lab). Peptides were separated on an M-class UPLC and analyzed on an Orbitrap mass spectrometer (ThermoFisher Scientific).

7.2.3.9 *LC-MS/MS data analysis by staff members at FGCZ*

The acquired MS data were processed for identification using PEAKS Studio XPlus (Bioinformatic Solutions). The spectra were searched against the following modifications: Oxidation (methionine), Variable modifications: Carbamidomethyl (cysteine), and Variable modifications: blinking fluorophore (cysteine or lysine). The results were obtained and viewed in Scaffold 5 (Proteome Software).

7.2.3.10 *Lipid cake preparati*on

Lipid films were prepared in glass tubes by dissolving an appropriate amount of phospholipid in chloroform to reach a concentration of 2 mg mL^{-1}. To prepare one sample of vesicles, 100 µL of phospholipid solution was used per glass tube (i.e., 0.2 mg of lipid per tube). Depending on the experiment, the composition of the lipid cake may vary. Typically, 16:0 biotinyl PE was added at 1:100% w/w (i.e., 1 µL of 1 mg mL^{-1} solution to the 100 µL lipid solution of 2 mg mL^{-1}) and subsequently labeled phospholipid was added at ~1:400% w/w (i.e., 0.25 µL of 1 mg mL^{-1}). Subsequently, tubes were mixed and dried under nitrogen flow overnight to obtain a thin lipid film. Tubes were sealed with Parafilm and stored at -20°C.

7.2.3.11 *Vesicle preparation*

The lipid cake was wetted with 10 mM PB buffer, pH 7.4, or a solution of the POI at 1 µM (if not mentioned otherwise) in the same buffer (for 0.2 mg of the lipid per tube). The wetted lipid cake was shaken in the thermoshaker (ThermoFisher Scientific) for 30 min at 1000 rpm and 4 °C above the transition temperature of the given lipid (T_t + 4 °C), see Table 1. Afterward, an additional 500 µL of 10 mM PB was added to the vesicle solution. The speed of shaking was decreased to 800 rpm and the sample was incubated for an additional 30 min. To uniformize the vesicle size distribution, we performed extrusion through a membrane with a pore size of 100 nm. The extruder (Avanti Lipids, Inc.) was cleaned, assembled, and equilibrated at T_t + 4°C on the heating block. The vesicle mixture was taken up in a Hamilton glass syringe. The syringe was equilibrated for 10 min to reach T_t + 4°C, then extruded 31 times (the sample was collected in the initially empty syringe).

7.2.3.12 *Size exclusion chromatography of vesicles*

We prepared the size exclusion chromatography column (height of the bed was 21 cm, d = 1.5 cm) according to the manufacturer's instructions (Sepharose CL 6B, Cytiva). The column was equilibrated with 10 mM PB buffer, pH 7.4. Samples were loaded directly onto the column and passed through by gravity flow. Fractions of 1 mL were collected. For this particular column, vesicle, protein, and dye fractions are usually collected around 10-11 mL, 25-27 mL, and 34-36 mL, respectively. To additionally confirm the presence of vesicles in the collected fractions, we analyzed and performed the analysis with suitable techniques, e.g., spectroscopic measurements, DLS, or potassium permanganate-stained TLC.

7.2.3.13 *Coverslips cleaning*

The round cover glasses (25 mm, 1.5 H, ThorLabs) were transferred to the glass staining dish that allowed spacing between slides. The coverslips were initially cleaned with MiliQ water to remove dust and water-soluble contaminants. The water was exchanged with 10% Deconex (Borer Chemie) and the slides were sonicated for 20 min, followed by thorough rinsing with MiliQ water until all traces of surfactant were removed and sonicated in MiliQ water for 5 min. The slides were washed with acetone and sonicated in fresh acetone for 20 min. The coverslips were dried under a nitrogen flow and placed in the UV-ozone cleaner (PSDP, Novascan) on a piece of aluminum foil 2 cm away from the light source, the upward-facing side was marked with a vertical line on the bottom right. They were irradiated for 10 min at 25° C and then were rested for 30 min in the closed chamber.

7.2.3.14 *Coverslips modification and vesicle deposition*

After the UV-ozone treatment, we transferred clean slides to the light-protected box and deposited 150 µL of 0.1 mg mL^{-1} PLL-PEG-biotin and PLL-PEG solutions mixed in a ratio of 1:50 or, where mentioned, 0.1 mg mL^{-1} PLL solution (SigmaAldrich) in 10 mM PB buffer, pH 7.4 on the etched side of coverslip. The solution was incubated at room temperature for 30 min. After incubation, the sample was removed and rinsed with MiliQ water. In the case of PLL-PEG-biotin, slides were additionally incubated with 0.1 mg mL^{-1} Neutravidin solution in 10 mM PB buffer, pH 7.4 for 30 min. After washing with MiliQ water, the coverslip was fixed in a cell chamber, the collected vesicles were deposited directly on the slides and they were incubated for at least 30 min. Finally, we washed the slide with the same PB buffer by removing the solution (not to dryness) and diluting them at least five times. In multicolor experiments, the vesicles were mixed in the given ratio before the application to the modified coverslips and washed analogously.

TLC staining: To confirm in which fraction the vesicles are present after purification with the SEC column, we ran a TLC in the solvent system 65:25:4 (v/v/v) chloroform: methanol: water. The lipids were visualized by potassium permanganate staining (or UV irradiation with 356 nm to excite NBD).

7.2.3.15 *Data analysis*

To perform the classification analysis, we localized the single-molecule signals and extracted the numeric intensity values of the signal over all frames followed by data normalization and filtering to remove noise signals (Table 21). For DL, the standardized and normalized traces were used directly as input for the model. All preprocessing was conducted on a laptop computer. The 1D-CNN-GRU-MCD models (Table 19) were trained on the ScienceCluster (NVIDIA Tesla V100-SXM2-32GB or 16 GB GPU) including evaluation and analysis, service infrastructure provided by S3IT (www.s3it.uzh.ch), the Service and Support for Science IT team at the UZH. For classification, the previously described model was used (Table 19). For evaluation, 100 predictions were calculated for each trace in the validation/test set yielding 100 potentially different predictions depending on the dropped nodes by the dropout mechanism. The mean model output over the 100 predictions was calculated per class used to predict the label. The Wasserstein distance was calculated and used as described previously in section 7.2.2.

7.2.4 Chapter 5

7.2.4.1 *Reagents, analyses, general settings*

Randomized libraries were custom-made by EllaBiotech. All Flow cytometer analysis was done by FlowJo™ v.10.8.1 Software (BD Life Sciences). Yeast cells were spun down at 1500 xg or 4000 rpm. For an OD_{600} of 1, $1 \cdot 10^{7}$ yeast cells mL^{-1} were assumed. Stock solutions of chemical probes were prepared in DMSO spectrophotometric grade > 99.9% (Acros Organics) at a concentration of 20 mM, diluted aliquots were prepared and stored at –20 °C. All stock solutions were thawed immediately before use. For image acquisition the fluorescence setup described in 7.1.2.3 was used.

7.2.4.2 *Optical spectroscopic measurements*

Absorbance spectra were measured on a Multiskan SkyHigh (ThermoFisher Scientific) in black 96-well plates of 100 µL solution in the stated solvent. The spectrum was scanned from 250 or 300 nm to 800 nm in 2 nm steps. Fluorescence spectra were measured on an FS5 spectrofluorometer (Edinburgh Instruments) equipped with an SC-41 well plater reader. All measurements were conducted at ambient temperature and under red light ambient.

7.2.4.3 *HeLa cells preparation and confocal imaging*

HeLa cells were grown in Dulbecco's Modified Eagle Medium supplemented with fetal bovine serum (10%) and penicillin (100 U mL^{-1})/streptomycin (100 µg mL^{-1})/Fungizone (0.25 µg mL^{-1}) at 37 ºC in 5% CO_2 environment. For imaging, 50'000 cells were seeded per well of an 8-well Ibidi chambered cover glass 2–3 days before imaging. The cells were incubated with the respective probes in growth medium for the indicated time. Before imaging, the growth medium was removed, and the cells were washed twice with PBS and imaged in FluoroBrite^{TM} (ThermoFisher Scientific). Imaging was performed at 37 ºC in a 5% CO_2 environment using the Nikon W1 UZH microscopy setup.

7.2.4.4 *Cell background quantification*

The average fluorescence intensity per cell was measured using FIJI (ImageJ 1.54f, NIH). Fluorescence intensity was quantified by defining regions comprising the whole-cell body and recording the average integrated intensity within this region. The defined region was translated to an area in the image that did not contain any cells and the average intensity of this area was recorded corresponding to the background. This background was subtracted from the cell body measurement. For each estimation, ten cells distributed over at least three fields of view were randomly chosen.

7.2.4.5 *Backbone amplification*

The pCT_McoTi plasmid was obtained from the Kolmar laboratory at TU Darmstadt via Dr. Gina Grammbitter as a glycerol stock. Lysogeny broth (LB) medium and agar plates were prepared with carbenicillin (100 µg mL^{-1}). *E. coli* cells were streaked onto LB agar and incubated at 37 ºC overnight. A single colony was picked and inoculated in LB medium (5 mL) overnight (37 °C, 300 rpm). Plasmid DNA was purified using the DNA extraction Plasmid Midi Kit (Qiagen).

7.2.4.6 *Backbone digestion*

The digestion mixture was prepared as follows:

Table 1. Components for backbone digest.

Reagent	Quantity
Backbone DNA	150 µg
NheI (5 U µg^{-1})	37.5 µL
SacI (5 U µg^{-1})	37.5 µL
10x rCutsmart buffer	80 µL
MiliQ water	544 µL

The reaction mixture was incubated at 37 °C for at least 16 h. The success of the digest was checked by analytical agarose gel 0.8%. The digest was purified by preparative agarose gel 0.8% and gel extraction using the QIAquick Gel Extraction Kit (Qiagen).

Insert PCR

The PCR mixture was prepared as follows:

Table 2. PCR mixture components for 2aa-library insert PCR for a 20 µL reaction in black changes for the 8aa-library in brackets and blue.

Reagent	Concentration	Quantity
Forward primer 2aa-library (8aa-library) (Table 24)	10 µM	1 µL
Reverse primer (GG2_YL_rev, Table 24)	10 µM	1 µL
Template (pCT_McoTI)	30 ng µL^{-1}	1 µL
dNTPs	10 mM	0.4 µL
$MgCl_2$	50 mM	0.4 µL
DMSO	-	0.8 µL
5x Phusion HF buffer	5x	4 µL
Phusion® High-Fidelity DNA Polymerase	2 U µL^{-1}	0.1
MiliQ water		11.3 µL

Table 3. PCR conditions for library insert 2aa-library in black changes for the 8aa-library in brackets and blue.

Step	Temperature (°C)	Time (s)	
1	98	30	
2	98	8	
3	61.5 (72)	15	
4	72	7	Repeat 34x
5	72	180	
6	4	∞	

The success of the PCR was checked by analytical agarose gel 1.5%. The digest was purified by preparative agarose gel 1.5% and gel extraction using the QIAquick Gel Extraction Kit (Qiagen).

***Insert PCR upscale for the 2aa-library:** The PCR mixture was prepared as follows:*

Table 4. PCR mixture components for 2aa-library insert upscaling PCR for a 20 µL reaction.

Reagent	Concentration	Quantity
Forward primer (GG7_long, Table 24)	10 µM	1 µL
Reverse primer (GG2_YL_rev, Table 24)	10 µM	1 µL
Template (2aa-library PCR)	99.3 ng µL^{-1}	0.2 µL
dNTPs	10 mM	0.4 µL
DMSO	-	0.8 µL
5x Phire reaction buffer	5x	4 µL
Phire Hot Start II DNA-Polymerase	2 U µL^{-1}	0.1
MiliQ water		12.5 µL

Table 5. PCR conditions for library insert 2aa-library upscale.

Step	Temperature (°C)	Time (s)	
1	98	30	
2	98	8	
3	70	15	
4	72	7	Repeat 34x
5	72	180	
6	4	∞	

The success of the PCR was checked by analytical agarose gel 1.5%. The digest was purified by the QIAquick PCR Purification Kit (Qiagen).

Insert upscale PCR for the 8aa-library

The PCR mixture was prepared as follows:

Table 6. PCR mixture components for 8aa-library insert upscale PCR for a 20 µL reaction.

Reagent	Concentration	Quantity
Forward primer (GG7_long, Table 24)	10 µM	1 µL
Reverse primer (GG2_YL_rev, Table 24)	10 µM	1 µL
Template (8aa-library PCR)	54.3 ng µL^{-1}	0.37 µL
dNTPs	10 mM	0.4 µL
$MgCl_2$	50 mM	0.4 µL
5x Phusion HF buffer	5x	4 µL
Phusion® High-Fidelity DNA Polymerase	2 U µL^{-1}	0.1
MiliQ water		12.03 µL

Table 7. PCR conditions for library insert upscale 8aa-library upscale.

Step	Temperature (°C)	Time (s)	
1	98	30	
2	98	8	
3	72	15	
4	72	7	Repeat 34x
5	72	180	
6	4	∞	

The success of the PCR was checked by analytical agarose gel 1.5%. The PCR reaction was purified using the QIAquick PCR Purification Kit (Qiagen).

7.2.4.7 *Preparation of the DNA for yeast transformation*

The DNA was concentrated in a SpeedVac (Eppendorf, Concentrator plus/Vacufuge) at 60 °C for 45 min to obtain concentrations < 50 µL of DNA per 400 µL transformation batch. 12 µg of insert and 4 µg of backbone per transformation with concentrations 450-550 µg mL^{-1}.

7.2.4.8 *Transformation of yeast cells with the library DNA*

A colony of EBY100 cells (grown on a yeast peptone dextrose (YPD) plate for 3 days and then stored at 4 °C) was picked and inoculated in 100 mL YPD prewarmed to 30 °C for an overnight culture grown at 30 °C and shaken at 200 rpm in a 500 mL baffled flask. After the overnight culture reached an OD_{600} of 3, the cells were diluted in 1 L of YPD and grown for ~5 h in two 2 L baffled flasks until they reached an OD_{600} of ~1.6. The cells were pelleted into 10 x 50 mL Falcon tubes, from this point onwards, all solutions for the electroporation were cooled to 4 °C and kept on ice before immediate use. The cells were kept on ice whenever possible. The pellets were resuspended in 50 mL autoclaved water by vortexing the pellet horizontally and centrifuging to remove the water. The pellet was washed in electroporation buffer (1 M sorbitol, 1 mM $CaCl_2$) twice. The pellets were each resuspended in 20 mL of conditioning buffer (100 mM lithium acetate, 2 M dithiothreitol) and combined in a 250 mL baffled flask. The flask was incubated for 30 min at 30 °C and 180 rpm. The cells were pelleted and washed once with electroporation buffer, resuspended and each pellet was resuspended with 1 mL of cold electroporation buffer to reach 2 mL total volume per pellet (overall 10 mL). The cells were kept on ice, while the DNA was added to the precooled electroporation cuvettes (sterilized through 30 min UV irradiated and then stored in an autoclaved box). 4 µg of linearized backbone and 12 µg insert DNA were combined in a total volume of <50 µL (in our case 31.3 µL). 400 µL of cell solution was added to each electroporation cuvette and pipetted up and down. Cells were electroporated at 2.5 kV and 25 µF with time constants between 2.5 ms and 3.5 ms. The cells were removed from the cuvette with a sterilized glass pipette and transferred into a 250 mL baffled flask containing 114 mL of prewarmed 1:1 of 1 M sorbitol/YPD. The cuvette was rinsed twice with 1 mL of 1:1 of 1 M sorbitol/YPD the rinses were added to the 114 mL 1:1 of 1 M sorbitol/YPD. The negative control was added to 6 mL in a 50 mL Falcon tube as well as the two rinses. The flask and the Falcons were shaken at 180 rpm at 30 °C for 1 h. The cells were collected by centrifugation and resuspended in 10 mL of SD media. Then a dilutions series for the positive and the negative control was prepared by adding 100 µL of the cells to 900 µL SD medium for 6 serial dilutions. 100 µl of each dilution was plated onto an SD plate (for the negative control only 5 dilutions were prepared). The 10 mL of transformed cells were transferred into 1 L of SD medium that was subsequently split into two 2 L baffled flasks and incubated for two days at 30 °C and 200 rpm. The dilution plates are counted after three days and can be used for colony PCR.

7.2.4.9 *Sample preparation for NGS*

A frozen yeast cell stock of yeast cells transformed with the 8aa-library was thawed and $8 \cdot 10^7$ yeast cells were pelleted. The plasmid was extracted from the yeast cells using Yeast Plasmid Miniprep II Kit (Zymo Research), according to the manufacturer's procedure. The complete eluted plasmid was amplified and extended with the Illumina adaptor sequences in 11 PCR reactions of 50 µL, using the following conditions:

Table 8. PCR mixture components for 8aa-library NGS sample of one out of 11 reactions à 50 µl.

Reagent	Concentration	Quantity
Forward primer (NGS-GG28, Table 24)	100 µM	0.25 µL
Reverse primer (NGS-GG29, Table 24)	100 µM	0.25 µL
Template (plasmid extracted from yeast)	-	1 µL
dNTPs	10 mM	1 µL
5x Phire reaction buffer	5x	10 µL
Phire Hot Start II DNA-Polymerase		0.25
MiliQ water		37.25 µL

Table 9. PCR conditions for 8aa-library NGS sample

Step	Temperature (°C)	Time (s)	
1	98	30	
2	98	8	
3	67	15	
4	72	5	Repeat 24x
5	72	180	
6	4	∞	

The success of the PCR was checked by analytical agarose gel 1.5%. The PCR reactions were purified by the QIAquick PCR Purification Kit (Qiagen). Any further library preparation was conducted by staff at the FGCZ.

7.2.4.10 *Storage of yeast library and induction*

The transformed cells were concentrated and stored in yeast freezing solution (10% glycerol, 0.7 w/v% yeast nitrogen base). At least 10x the library diversity in number of cells was frozen per aliquot. One of these aliquots is diluted to an OD_{600} of 0.5-0.8 in SD medium and grown overnight at 30°C and 200 rpm. A number of cells exceeding the library complexity by a factor of 10 were harvested and resuspended in SG media to an initial OD_{600} = 1. The cells were incubated for 24 to 48 h at 30 °C and 200 rpm. After induction, the cells are stored at 4 °C in Falcon tubes.

7.2.4.11 *Antibody staining of cells for flow cytometry analysis or FACS*

The yeast cells were induced for 48 h in SG medium at 30 °C while shaking at 200 rpm. The whole preparation was done on ice. Per sample, $2 \cdot 10^7$ cells were collected in a 1.5 mL Eppendorf tube. The cells were washed twice with 0.5% bovine serum albumin in PBS (BPBS). In the case of *n* samples, the $n \cdot 2 \cdot 10^7$ cells were combined in the beginning and washed as one batch and then diluted in $n \cdot$ 100 µl and then separated into *n* Eppendorf tubes and the BPBS was removed again. For primary antibody staining per sample 40 µl of 1:200 antibody in BPBS was prepared and used to resuspend the cells. The cells were kept on ice for 30 min in the dark. The cells were diluted in 1 mL of BPBS and then washed once more with BPBS. In the case of secondary antibody binding the incubation time of the primary antibody was extended to 1 h before washing. Then 40 µl of 1:200 secondary antibody was added to the cells and incubated for 30 min. The secondary antibody solution was diluted in 1 mL of BPBS and washed once more with 1 mL of BPBS. Finally, the cells were resuspended in 1 mL of BPBS. The cells were strained into a FACS tube immediately before the measurement. For FACS analysis the cells were diluted into 2 mL of BPBS.

7.2.4.12 Dye staining of cells for flow cytometry analysis or FACS

The yeast cells were induced for 48 h in SG medium at 30 °C while shaking at 200 rpm. Per sample, $2 \cdot 10^7$ cells were collected in a 1.5 mL Eppendorf tube. The cells were washed twice with BPBS. In the case of *n* samples, the $n \cdot 2 \cdot 10^7$ cells were combined in the beginning and washed as one batch and then diluted in $n \cdot$ 100 µl and then separated into *n* Eppendorf tubes and the BPBS was removed again. The cells were resuspended in 40 µl of 1 µM dye solution in BPBS (freshly prepared from a 1 mM dye stock in DMSO) and incubated at 25 °C for 1 h. The sample was diluted to 2 mL for flow cytometry analysis and 0.5 mL for FACS with BPBS and potential further dilution as seen fit by the operator (no washing). The cells were strained into a FACS tube immediately before the measurement.

7.2.4.13 *Combined antibody and dye staining of cells for flow cytometry analysis or FACS*

The cells were first stained with the antibody as described in 7.2.4.11 followed by labeling with the dyes according to 7.2.4.12 with 2 BPBS washes after the removal of the antibody. The cells were strained into a FACS tube immediately before the measurement.

7.2.4.14 *Flow cytometry analysis*

The samples were measured by an operator at the UZH flow cytometry core facility on a FACSCanto II (BD Life Sciences). The channels used were 405 (405 nm, no mirror, 450/50), 488 (488 nm, 502LP, 530/30), and 640 (640 nm, 685LP, 660/20). The voltages were adjusted on the respective control cells. 10^5-10^6 cells were acquired per sample.

7.2.4.15 *FACS*

The samples were measured by an operator at the UZH flow cytometry core facility on a FACSAria III 3L (BD Life Sciences) or a FACSymphony S6 5L (BD Life Sciences), cell sorter. The used channels were 355 (355 nm, 505 LP 515/30), 402 (402 nm, 425 LP, 431/28), 488 (488 nm, 495 LP, 530/30), 561 (561 nm, no mirror, 586/15), 640 (639 nm, 635 LP, 670/30) and 405 (405 nm, no mirror, 450/50), 488 (488 nm, 502LP, 530/30), 561 (561 nm, no mirror, 582/15), 640 (633 nm, no mirror, 660/20), respectively. Unless otherwise described the 10^7 detected events were used as stopping-threshold. The cells were sorted directly into SD medium containing carbenicillin and chloramphenicol.

7.2.4.16 *Yeast cell preparation for microscopy*

The cells were prepared as mentioned in “Antibody staining of cells for flow cytometry analysis or FACS: 7.2.4.11” or “Dye staining of cells for flow cytometry analysis or FACS: 7.2.4.12” with the dyes and antibodies mentioned respectively. Sampled were added directly to an 8-well Ibidi chambered cover glass and let settle down for 15 min before imaging. Imaging was conducted with the setup described in *7.1.2.3.*

7.3 Synthetic procedures

7.3.1 Peptide synthesis

General information: Peptides **C11-C14** and **P15-P17** were prepared by manual SPPS, whereas peptides **E19-E21** were prepared on an automated TributeTM UV-IR peptide synthesizer (Gyros Protein Technologies AB) with manual coupling of the HMSiR labeled cysteine at an intermediate step. All peptides were synthesized on a 0.05 mmol scale. **P15-P17** were prepared by Lionel Rumpf.

Manual peptide synthesis: After every reaction and swelling step, the resin was washed three times with each CH_2Cl_2 and DMF (4 mL). The swelling and all the reaction steps were shaken at room temperature using a heating/cooling dry bath (ThermoFisher Scientific). Peptides were synthesized on 100 mg pre-loaded Wang-resins (Bachem, 0.5–0.8 mmol g^{-1}, 1 equiv.), which were swollen for at least 3 h in 4:1 CH_2Cl_2/ DMF. Standard, manual Fmoc-SPPS protocols were used for synthesis. Briefly, deprotection was performed with 20% piperidine in DMF for 20 min and coupling with *O*-(1*H*-Benzotriazole-1-yl)-*N*,*N*,*N*',*N*'-tetramethyluroniumhexafluorophosphate (HBTU) (4 equiv.), DIPEA (8 equiv.) and the respective Fmoc-protected amino acid (4 equiv.) in DMF for 1 h. Fmoc-Cys(HMSiR)-OH **8** was coupled with 1-hydroxbenzotriazole (HOBt) (2.5 equiv.), N-Ethyl-N′-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDC HCl) (2.5 equiv.), Fmoc-Cys(HMSiR)-OH **8** (2.5 equiv.). The solids for Fmoc-Cys(HMSiR)-OH **8** coupling were pre-mixed in 3 mL DMF: THF 1:1 for 5 min, added to the deprotected resin, and left shaking for 3 h. For the coupling phosphorylated amino acid **18** the equivalents of the amino acid **18**, coupling reagent, HBTU, and DIPEA were increased by one to five and nine equivalents respectively. The final coupling was carried out by pre-mixing pentynoic acid (10 equiv.) with EDC HCl (10 equiv.) in 2 mL of 1:1 CH_2Cl_2/ DMF after 2 min the mixture was added to the resin for 2 h. Cleavage from the resin and concurrent global deprotection was carried out with a mixture of TFA, water, and triisopropylsilane (90:5:5) for 3 h. The crude products were purified by preparative HPLC (A: H_2O + 0.1% TFA, B: CH_3CN; 1. 0-10 min, 20%-50% B; 2. 10-20 min, 50% B; 3. 20-24 min, 50-20% B). The obtained peptides were characterized by LC-MS, ESI-HRMS, and ESI-MS/MS. Pure peptides were stored as 1 mM stock solutions in DMSO or in lyophilized form.

Automated peptide synthesis: Peptides **E19-E21** were synthesized on a Tribute UV-IR (Gyros Protein Technologies) peptide synthesizer applying the manufacturer's standard cycles. The couplings were conducted with the respective Fmoc-protected amino acid (5 equiv.), O-(1H-6-chlorobenzotriazole-1-yl)-1,1,3,3-tetramethyluronium (HCTU) (4.75 equiv.) and DIPEA (7 equiv.) in DMF. The Fmoc-deprotection was done with 20% piperidine in DMF. The coupling of the Fmoc-Cys(HMSiR)-OH **8** amino acid was conducted manually with Fmoc-Cys(HMSiR)-OH **8**

(2.5 equiv.), HOBt (2.5 equiv.), EDC HCl (2.5 equiv.). The solids were premixed in 3 mL 1:1 DMF/ THF for 5 min, added to the deprotected resin, and left shaking for 3 h. The final coupling was done manually by pre-mixing pentynoic acid (10 equiv.) with EDC HCl (10 equiv.) in 2 mL of 1:1 CH_2Cl_2/ DMF after 2 min the mixture was added to the resin for 2 h. After the manual steps, the resin was washed with each CH_2Cl_2 and DMF (4 mL). The peptides were cleaved from the resin by the addition of 3 mL of Reagent K (TFA, phenol, water, thioanisole, 1,2-ethanedithiol in ratio; 82.5:5:5:5:2.5) mixing for 3 h. After concentration of the cleavage mixture cool *tert*-butyl methyl ether (MTBE) was added to the residue. The precipitate was centrifuged for 15 min at 4000 rpm, the supernatant was removed, the pellet was resuspended in MTBE, and centrifuged once more. The crude products were purified by HPLC. The obtained peptides were characterized by ESI-HRMS and ESI-MS/MS. They were stored as 1 mM stock solutions in DMSO or in lyophilized form.

7.3.2 Small molecule synthesis

Compound 2 was prepared by Dr. Adam Eördögh. The synthesis of 34 was conducted by Dr. Zacharias Thiel. Compound 43 was prepared by Dr. Elias A. Halabi Rosillo and compound 26 by Henriette Lämmermann.

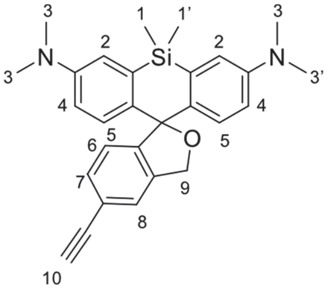
General procedure A: reductive amination of xanthene cores with aqueous workup

The diamine core (1 equiv., 5–10 mg) and the formyl phenylboronic acid (3 equiv.) were added to a flame-dried flask under a nitrogen atmosphere and dissolved in dry CH_3OH (3 mL). The reaction mixture was stirred for 20 min to 1 h at 25 °C. $NaBH_3CN$ was added slowly. The mixture was stirred at 25 °C for 1 to 4 h as judged by TLC or LC-MS measurements. The reaction was stopped by cooling the mixture to 0 °C and adding saturated $NaHCO_{3.}$ The aqueous layer was extracted with CH_2Cl_2, and the organic phase was washed with brine and dried over Na_2SO_4 before concentrating the crude under reduced pressure. The crude was directly purified by preparative HPLC (SiO_2-C_{18}; CH_3CN/H_2O 5:95 to CH_3CN/H_2O 95:5) and dried by lyophilization.

General procedure B: reductive amination of xanthene cores

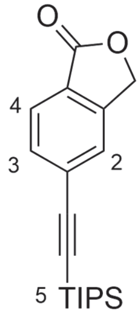
The diamine core (5–10 mg) and the formyl phenylboronic acid (2.3 equiv.) were added to a flame-dried flask under a nitrogen atmosphere and dissolved in dry CH_3OH (3 mL). The reaction mixture was stirred for 20 min to 1 h at 25 °C. $NaBH_3CN$ was added slowly. The mixture was stirred at 25 °C for 1 to 4 h as judged by TLC pr LC-MS measurements. The reaction was concentrated under reduced pressure and directly purified by preparative HPLC (SiO_2-C_{18}; CH_3CN/H_2O 5:95 to CH_3CN/H_2O 95:5) and dried by lyophilization

5'-Ethynyl-*N*^{3},*N*^{3},*N*^{7},*N*^{7},5,5-hexamethyl-3'*H*,5*H*-spiro[dibenzo[b,e]siline-10,1'-isobenzofuran]-3,7-diamine (1)



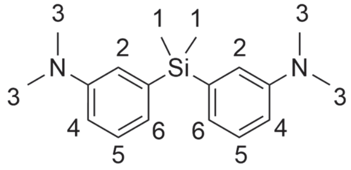
Compound **6** was dissolved in THF (1.5 mL). TBAF (0.083 mL, 0.083 mmol, 1 M) was added at 0 °C. The reaction was stirred for 2 h and then concentrated under reduced pressure and loaded onto Celite. The crude was purified by flash chromatography (SiO_2; hexane to hexane/EtOAc 25:75) and HPLC (SiO_2-C_{18}; CH_3CN + 0.1% TFA/H_2O + 0.1% 1:9 to CH_3CN + 0.1% TFA/H_2O + 0.1% 9:1) to yield 26 mg of pure compound **1** (78%) after lyophilization. ^{1}H NMR (500 MHz, $CDCl_3$) δ = 7.44 (s, 1H, H8), 7.38 (dd, *J* = 7.8, 1.4, 1H, H7), 7.01 (d, *J* = 7.9, 1H, H6), 6.97 (d, *J* = 2.9, 2H, H2), 6.95 (d, *J* = 8.8, 2H, H5), 6.62 (dd, *J* = 8.9, 2.9, 4H, H4) , 5.21 (s, 2H, H9), 3.07 (s, 1H, H10), 2.95 (s, 12H, H3), 0.61 (s, 3H, H1 or H1'), 0.54 (s, 3H, H1 or H1') ppm. ^{13}C NMR (126 MHz, $CDCl_3$) δ = 148.93, 147.35, 140.21, 137.87, 135.66, 131.60, 128.67, 125.06, 124.71, 121.05, 116.84, 113.90, 92.65, 83.86, 72.05, 40.63, 0.71, -1.10 ppm. HRMS (ESI/QTOF) $[M+H]^+$ calcd. for $[C_{28}H_{31}N_2OSi]^+$: 439.2200; found 493.2192.

5-((Triisopropylsilyl)ethynyl)isobenzofuran-1(3H)-one (3)



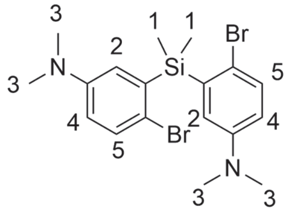
5-Bromophthalide (3 g, 14.08 mmol), $PdCl_2(PPh_3)_2$ (0.16 mg, 0.23 mmol), and CuI (0.03 mg, 0.15 mmol) were added to a flame-dried Schlenk flask and nitrogen atmosphere was established by evacuation and backfilling of the flask. The compounds were dissolved in dry toluene (18 mL) and (triisopropylsilyl)acetylene (2.57 mL, 14.083 mmol) and NEt_3 (1.56 g, 15.5 mmol) were added. The reaction mixture was heated to 80 °C and stirred for 4 h. The cooled-down mixture was filtered through Celite, the filtrate was diluted with EtOAc. The organic phase was washed with water and brine, dried over Na_2SO_4, and concentrated under reduced pressure. The crude was purified by flash column chromatography (SiO_2; hexane to hexanes/EtOAc 8:2) to give the product **3** as an off-white solid (1.36 g, 42%). ^{1}H NMR (400 MHz, $CDCl_3$) δ = 7.78 (d, *J* = 8.0, 1H, H4), 7.54 (dd, *J* = 7.9, 1.2, 1H, H3), 7.51 (t, *J* = 1.1, 1H, H2), 5.22 (s, 2H, H1), 1.14–0.99 (m, 21H, H5) ppm. ^{13}C NMR (101 MHz, $CDCl_3$) δ = 170.53, 146.57, 133.03, 129.64, 125.73, 125.59, 125.17, 105.61, 95.97, 69.41, 18.77, 11.37 ppm. HRMS (ESI/QTOF) $[M+H]^+$ calcd. for $[C_{19}H_{27}N_2O_2Si]^+$: 315.17748; found 315.1775

3,3'-(Dimethylsilanediyl)bis(*N*,*N*-dimethylaniline) (4)



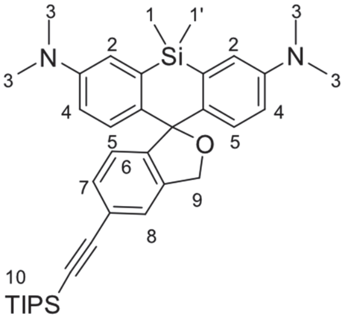
Compound 2 (5 g, 25 mmol) was dissolved in dry THF (60 mL) and the solution was cooled to –78 °C. *n*-BuLi (17.2 mL, 27.5 mmol, 1.6 M) was added dropwise and the mixture was stirred at –78 °C for 2 h. Dichlorodimethylsilane (1.52 mL, 12.5 mmol) was added dropwise and the mixture was stirred for 2 h at 25 °C. Brine (40 mL) and H_2O (10 mL) were added and the mixture was extracted three times with EtOAc. The combined organic phases were dried over $MgSO_4$ and concentrated onto Celite. The crude was purified by flash column chromatography (SiO_2; hexane to hexane/EtOAc 9:1) to give the product 4 as a light-yellow oil (3.6 g, 50%). ^{1}H NMR (400 MHz, $CDCl_3$) δ = 7.31–7.19 (m, 2H, H5), 6.97–6.90 (m, 4H, H4, H2), 6.78 (dd, *J* = 8.3, 2.8, 2H, H6), 2.94 (s, 12H, H3), 0.55 (s, 6H, H1) ppm. ^{13}C NMR (101 MHz, $CDCl_3$) δ = 154.75, 141.11, 131.49, 125.30, 120.50, 116.20, 42.84, 0.84 ppm.

3,3'-(Dimethylsilanediyl)bis(4-bromo-*N*,*N*-dimethylaniline) (5)



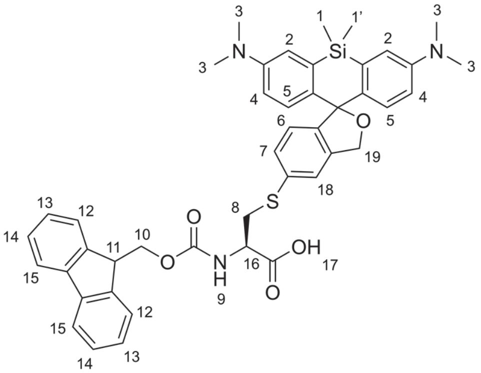
3,3'-(Dimethylsilanediyl)bis(*N*,*N*-dimethylaniline 4 (3.00 g, 0.01 mol) was dissolved in CH_3CN (75 mL) in a flame-dried flask. The solution was cooled to 0 °C and NBS (3.75 g, 0.021 mmol) was added in small portions. After complete addition, the solution was stirred for 1.5 h at 25 °C then sat. $NaHCO_3$ was added to neutralize the solution. The organic phase was washed with water. The combined aqueous phases were extracted twice with CH_2Cl_2. All CH_2Cl_2 fractions were combined, dried over $MgSO_4$, and concentrated under reduced pressure. The residue was purified by flash column chromatography. (SiO_2; hexane: EtOAc 95:2) yielding the white product 5 (2.69 g, 59%). ^{1}H NMR (400 MHz, $CDCl_3$) δ = 7.36 (d, *J* = 8.8, 2H, H5), 6.85 (d, *J* = 3.2, 2H, H2), 6.61 (dd, *J* = 8.7, 3.2, 2H, H4), 2.89 (s, 12H, H3), 0.77 (s, 6H, H1) ppm. ^{13}C NMR (101 MHz, $CDCl_3$) δ = 149.02, 138.86, 133.10, 121.92, 116.94, 115.40, 40.70, -0.78 ppm. HRMS (ESI/QTOF) $[M+H]^+$ calcd. for $[C_{18}H_{25}Br_2N_2Si]^+$: 455.0148; found 455.0146.

N[3],*N*[3],*N*[7],*N*[7],5,5-Hexamethyl-5'-((triisopropylsilyl)ethynyl)-3'*H*,5*H*-spiro[dibenzo[b,e]siline-10,1'-isobenzofuran]-3,7-diamine (6)



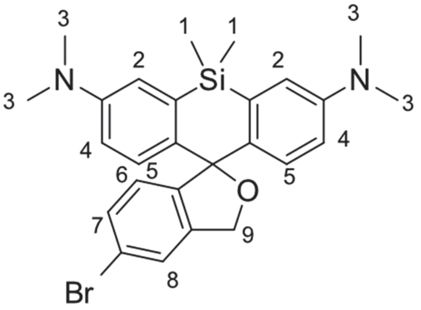
Magnesium turnings were intensely heated under a high vacuum atmosphere (109 mg, 4.49 mmol) then anhydrous THF (3 mL) was added. The suspension was heated to 60 °C then a solution of compound **5** (366 mg, 803 μmol) in anhydrous THF (10 mL) followed by 1,2-dibromoethane (20.8 μL, 241 μmol) upon which the reaction turned slowly yellow. The mixture was stirred at 25 °C for 3 h. Then the reaction mixture was heated again to 60 °C and a solution of compound **3** (184 mg, 573 μmol) in anhydrous THF (16 mL) was added, upon addition the reaction turned from deep yellow to dark brown. The reaction mixture was stirred at the same temperature overnight. Aqueous HBr (6 mL, 47%) was added and the reaction turned dark orange, some CH_2Cl_2 was added. The aqueous phase was extracted three times with CH_2Cl_2. The organic phase was dried over Na_2SO_4, concentrated under reduced pressure, loaded onto Celite, and purified by flash chromatography (SiO_2; hexane + 1% NH_3/ EtOAc + 1% NH_3 to 9:1 hexane + 1% NH_3 EtOAc + 1% NH_3). The procedure gave 200 mg of product **6** as a yellow solid (42%). ^{1}H NMR (400 MHz, $CDCl_3$) δ = 7.46 (s, 1H, H8), 7.40 (dd, *J* = 7.9, 1.4, 1H, H7), 7.08 (d, *J* = 2.9, 2H, H2), 7.01 (d, *J* = 7.6, 1H, H6), 6.98 (d, *J* = 8.9, 2H, H5), 6.70 (dd, *J* = 8.9, 2.9, 2H, H4), 5.18 (s, 2H, H9), 2.98 (s, 12H, H3), 1.14 (s, 21H, H10), 0.59 (d, *J* = 30.2, 6H, H1, H1'). ^{13}C NMR (101 MHz, $CDCl_3$) δ 148.06, 146.09, 140.21, 136.21, 131.57, 129.04, 128.86, 124.99, 124.65, 122.72, 118.00, 114.77, 107.14, 90.71, 72.02, 18.82, 11.46, 0.69, -1.32 ppm. HRMS (ESI/QTOF) $[M+H]^+$ calcd. for $[C_{37}H_{51}N_2OSi_2]^+$: 595.3534; found 595.3533.

N-(((9*H*-Fluoren-9-yl)methoxy)carbonyl)-*S*-(3,7-bis(dimethylamino)-5,5-dimethyl-3'*H*,5*H*-spiro[dibenzo[*b*,*e*]siline-10,1'-isobenzofuran]-5'-yl)-*L*-cysteine (8)



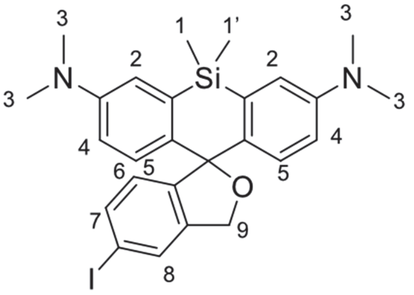
Xanthphos Pd G3 (15.8 mg, 16.7 µmol), Fmoc-L-Cys-OH (105 mg, 305 µmol) and compound **10** (150 mg, 278 µmol) were dissolved in THF (1.4 mL) in a flame-dried Schlenk flask. The mixture was stirred at 25 °C for 2–3 min, then NEt_3 (65.6 µL, 472 µmol) was added, the Schlenk tube was capped with a rubber septum, evacuated, and backfilled with argon. The solution was stirred at 25 °C for 30 min. The solvent was removed under reduced pressure. The residue was loaded onto Celite and purified by flash chromatography (SiO_2; CH_2Cl_2 to CH_2Cl_2/CH_3OH 95:5) to yield **8** as a blue/green powder (200 mg, 95%). ^{1}H NMR (400 MHz, THF) δ = 7.77 (d, *J* = 7.5, 2H, H15), 7.66 (t, *J* = 8.2, 2H, H13), 7.40 (s, 1H, H18), 7.34 (t, *J* = 7.4, 2H, H14), 7.25 (dd, *J* = 8.7, 5.4, 3H, H7, H12), 7.03 (dd, *J* = 8.7, 3.0, 3H, H5, H9), 6.96 (t, *J* = 3.1, 2H, H2), 6.86 (dd, *J* = 8.0, 4.4, 1H, H6), 6.64–6.55 (m, 2H, H4), 5.21 (s, 2H, H19), 4.49 (td, *J* = 8.0, 4.6, 1H, H16), 4.28 (dd, *J* = 7.3, 2.1, 1H, H10), 4.23 (s, 1H, H11), 3.49 (dd, *J* = 13.5, 4.8, 1H, H8), 3.25 (dd, *J* = 13.5, 7.9, 1H, H8’), 2.89 (d, *J* = 4.3, 12H, H3), 0.58 (s, 3H, H1 or H1), 0.46 (s, 3H, H1 or H1’) ppm. ^{13}C NMR (101 MHz, THF) δ = 172.37, 160.18, 159.90, 157.01, 149.88, 148.42, 145.40, 145.37, 142.39, 135.89, 129.70, 128.49, 127.93, 126.26, 126.24, 125.18, 123.93, 120.76, 118.43, 115.44, 74.24, 55.05, 54.84, 54.74, 48.38, 46.81, 42.03, 38.27, 37.32, 0.51, -0.64 ppm. HRMS (ESI/QTOF) $[M+H]^+$ calcd. for $[C_{44}H_{46}N_3O_5SSi]^+$: 756.2922; found 756.2913.

5'-Bromo-*N*3,*N*3,*N*7,*N*7,5,5-hexamethyl-3'*H*,5*H*-spiro[dibenzo[*b*,*e*]siline-10,1'-isobenzofuran]-3,7-diamine (9)



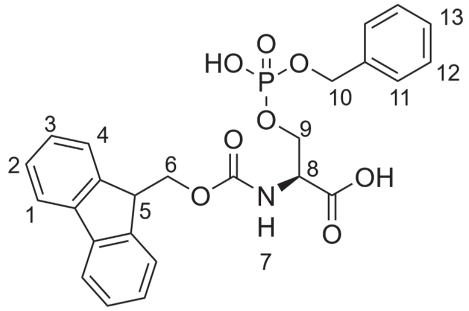
Compound **5** (1 g, 2.19 mmol) in dry THF (10 mL) was added to magnesium turnings (213 mg, 8.77 mmol) in dry THF (3 mL) at 60 °C, followed by the addition of 1,2-dibromoethane (40.6 µL, 470 µmol). The mixture was stirred at 25 °C for 3 h and the solution turned yellow slowly. The Grignard reagent was transferred into an empty flame-dried Schlenk flask. A solution of 5-bromophthalide (170 mg, 783 µmol) in dry THF (8 mL) was added at 60 °C. Upon addition, the reaction turned from deep yellow to dark brown. The mixture was stirred at the same temperature for 12 h. Aqueous HBr (3.8 mL, 47%) was added and the solution turned dark orange and CH_2Cl_2 was added. $NaHCO_3$ was added very carefully until the color moved to the CH_2Cl_2 phase. The aqueous phase was extracted three more times with CH_2Cl_2. The organic phase was dried over Na_2SO_4, concentrated under reduced pressure, and loaded onto Celite. Purification by reverse phase column chromatography (SiO_2-C_{18}; CH_3CN / H_2O + 0.1 TFA 5:95 to CH_3CN / H_2O + 0.1 TFA 95:5) gave the compound **9** as the blue trifluoroacetate salt (302 mg, 45%).^{1}H NMR (400 MHz, CD_3OD) δ = 7.90 (s, 1H, H8), 7.63 (d, *J* = 8.0, 1H, H7), 7.35 (d, *J* = 2.8, 2H, H2), 7.11–7.04 (m, 3H, H5, H6), 6.79 (dd, *J* = 9.7, 2.9, 2H, H4), 4.28 (s, 2H, H9), 3.35 (s, 12H, H3), 0.60 (d, *J* = 2.6, 6H, H1) ppm. ^{13}C NMR (101 MHz, CD_3OD) δ = 167.63, 159.17, 158.76, 155.81, 149.41, 143.39, 142.14, 137.21, 132.07, 131.11, 131.05, 128.35, 124.14, 122.30, 120.28, 117.46, 115.33, 114.63, 111.80, 66.97, 61.74, 40.94, 27.20, -1.10, -1.34 ppm. HRMS (ESI/QTOF) $[M+H]^+$ calcd. for $[C_{26}H_{30}BrN_2OSi]^+$: 493.1305; found 493.1305.

5'-Iodo-*N*3,*N*3,*N*7,*N*7,5,5-hexamethyl-3'*H*,5*H*-spiro[dibenzo[*b*,*e*]siline-10,1'-isobenzofuran]-3,7-diamine (10)



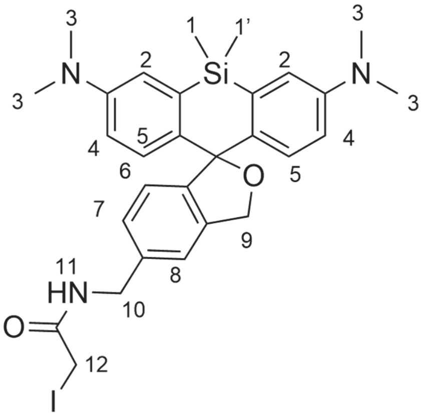
A flame-dried MW vial was charged with compound 9 (49.5 mg, 100 µmol), CuI (9.52 mg, 50 µmol), NaI (300 mg, 2.00 mmol), and briefly evacuated and backfilled with N_2. Racemic *trans*-(1R,2R)-*N*,*N*'-dimethyl-cyclohexane-1,2-diamine (1.42 mg, 10 µmol), and dioxane (0.5 mL) were added. The MW vial was sealed, and the mixture was stirred at 110 °C for 29 h. The resulting suspension was cooled to reach 25 °C, diluted with 25% aqueous NH_3, poured into water, and extracted with CH_2Cl_2. The combined organic phases were dried over Na_2SO_4 and concentrated under reduced pressure. The residue was purified by flash chromatography (SiO_2; hexane + 1% NH_3/ EtOAc + 1% NH_3 to 9:1 hexane + 1% NH_3 EtOAc + 1% NH_3) to provide product 10 as a white solid (190 mg, 77%). ^{1}H NMR (400 MHz, $CDCl_3$) δ = 7.72 (s, 1H, H8), 7.61 (d, *J*=8.0, 1H, H7), 7.27 (s, 2H, H2), 7.11 (d, *J*=8.8, 2H, H5), 6.91 (d, *J*=8.8, 2H, H4), 6.81 (d, *J*=8.0, 1H, H6), 5.16 (s, 2H, H9), 3.06 (s, 12H, H3), 0.64 (s, 3H, H1 or H1'), 0.56 (s, 3H, H1 or H1') ppm.^{13}C NMR (101 MHz, ($CDCl_3$) δ = 141.12, 136.61, 120.08, 118.10, 116.43, 114.68, 92.21, 0.38, -0.99 ppm. HRMS (ESI/QTOF) $[M+H]^+$ calcd. for $[C_{26}H_{30}IN_2OSi]^+$: 541.1167; found 541.1167.

N-(((9*H*-Fluoren-9-yl)methoxy)carbonyl)-*O*-((benzyloxy)(hydroxy)phosphoryl)-*L*-serine (18)



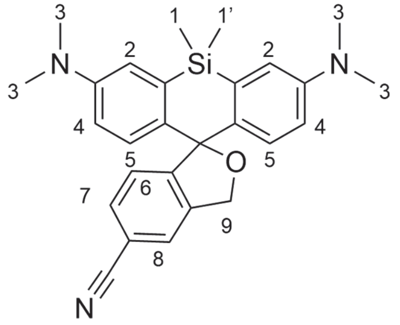
Water from Fmoc-*L*-Ser-OH·H_2O was azeotropically distilled from THF (4 mL g^{-1}) on the rotary evaporator and dried overnight. PCl_3 (379 µL, 4.34 mmol) was dissolved in THF (8.0 mL) and cooled to 0 ºC. Benzyl alcohol (518 µL, 5 mmol) was added while keeping the internal temperature below 5 ºC. The solution was stirred for 10 min at 0 to 5 ºC. The reaction was further cooled to –5 °C then 2,6-lutidine (1.17 mL, 10 mmol) was added to the flask keeping the reaction at –5 to 5 ºC, forming a thick slurry. 2,6-Lutidine (389 µL, 3.34 mmol) was added to a solution of Fmoc-Ser-OH (1.15 g, 3.34 mmol) in THF (4.0 mL) in a separate flask. This solution was added to the reaction at a rate that kept the mixture at –5 to 5 ºC. Upon reaction completion, H_2O (3.6 mL) was added to the flask, maintaining the temperature below 10 ºC followed by the addition of NaBr (789 mg, 7.67 mmol) at 0 ºC. A 20% w/w aqueous solution of $NaBrO_3$ (252 mg, 1.67 mmol) was added at 0 to 5 ºC. After the addition, the mixture was warmed to 25 °C. Upon reaction completion, an aqueous solution of $Na_2S_2O_5$ (20% in H_2O) (1.0 mL) was added to the flask in one portion. 2-CH_3-THF was added, and the layers were shaken and separated. The organic layer was washed with brine, dried with Na_2SO_4, and concentrated under reduced pressure. The crude oil was diluted with 2-CH_3-THF (7 mL g^{-1}) and was stirred at ambient temperature for 16 h. A white precipitate resulted, which was filtered off and washed with cold 2-CH_3-THF. The product **18** was obtained as a white solid (700 mg, 42%).^{1}H NMR (400 MHz, DMSO-*d*$_6$) δ = 7.89 (d, *J* = 7.6 Hz, 2H, H1), 7.84 (d, *J* = 8.3 Hz, 1H, H7), 7.73 (d, *J* = 7.5 Hz, 2H, H4), 7.45–7.27 (m, 9H, H2, H3, H11, H12, H13), 4.93 (d, *J* = 7.1 Hz, 2H, H10), 4.36–4.08 (m, 6H, H5, H6, H8, H9) ppm. ^{13}C NMR (101 MHz, DMSO-*d*$_6$) δ = 170.75, 155.98, 143.76, 143.74, 140.68, 136.77, 136.69, 128.36, 128.03, 127.64, 127.60, 127.08, 125.30, 120.10, 67.52, 67.47, 65.92, 65.28, 65.23, 54.40, 54.32, 46.55. ^{31}P NMR (162 MHz, DMSO-*d*$_6$) δ = -1.49. HRMS (ESI) $[M+H]^+$ calcd. for $[C_{25}H_{24}NO_8P]^+$: 498.1312; Found 498.1327.

N-((3,7-bis(Dimethylamino)-5,5-dimethyl-3'*H*,5*H*-spiro[dibenzo[b,e]siline-10,1'-isobenzofuran]-5'-yl)methyl)-2-iodoacetamide (22)



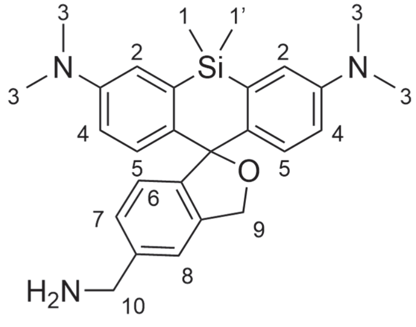
TSTU (33.9 mg, 0.113 mmol) and DIPEA (14.6 mg, 0.113 mmol) were added to a solution of iodoacetic acid (20.9 mg, 0.113 mmol) in CH_3CN (1.2 mL). After the reaction mixture was stirred at 25 °C for 1 h, a solution of compound **24** (25 mg, 0.056 mmol) in Na_2CO_3 saturated CH_3CN was added. After 1 h, 1 M HCl (1 mL) was added to the mixture, before removing the CH_3CN under reduced pressure. The residue was dissolved in $CH_2Cl_{2,}$ more water was added and extracted three times in total with CH_2Cl_2. The organic phase was washed with brine and dried over Na_2SO_4, concentrated under reduced pressure. The crude was dissolved in CH_3CN /water 50:50 with 0.1% TFA and purified by HPLC (SiO_2-C_{18}; CH_3CN + 0.1% TFA/H_2O + 0.1% 2:8 to CH_3CN + 0.1% TFA/H_2O + 0.1% 9:1) resulting in 12 mg of material **22** (17%). ^{1}H NMR (400 MHz, CD_3OH) δ = 7.67 (s, 1H, H8), 7.41 (d, *J* = 7.9, 1H, H7), 7.35 (d, *J* = 2.8, 2H, H2), 7.11 (d, *J* = 7.7, 1H, H6), 7.09 (d, *J* = 9.6, 2H, H5), 6.74 (dd, *J* = 9.7, 2.8, 2H, H4), 4.52 (s, 2H, H9), 4.31 (s, 2H, H10), 3.80 (s, 2H, H12), 3.34 (s, 12H, H3, H3'), 0.60 (d, *J* = 4.2, 6H, H1, H1') ppm.^{13}C NMR (101 MHz, CD_3OH) δ = 171.49, 169.59, 155.78, 149.42, 142.55, 141.02, 140.90, 137.34, 130.55, 128.75, 127.24, 126.93, 122.11, 115.06, 62.37, 44.31, 40.88, -1.09, -1.35, -2.18. HRMS (ESI-MS) $[M+H]^+$ calcd. for $[C_{29}H_{35}IN_3O_2Si]^{+:}$ 612.15377, found 612.15482.

3,7-bis(Dimethylamino)-5,5-dimethyl-3'*H*,5*H*-spiro[dibenzo[b,e]siline-10,1'-isobenzofuran]-5'-carbonitrile (23)



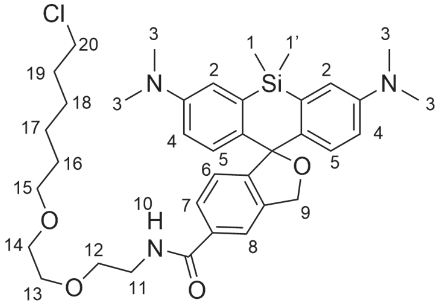
Compound **5** (450 mg, 0.986 mmol) and 5-cyanophthalide (345 mg, 2.17 mmol) were dried under reduced pressure in two flame-dried flasks and dissolved in dry THF (11.25 mL and 22.5 respectively). The solution of 3,3'-(dimethylsilanediyl)bis(4-bromo-*N,N*-dimethyl aniline) was cooled to –78 °C using an acetone/dry ice bath. *t*-BuLi solution (2.55 mL, 4.34 mmol, 1.7 M) was added, and the reaction mixture was stirred at –78 °C for 30 min. The phthalide solution was added dropwise over 60 min. The reaction mixture was allowed to warm to 20 °C and was stirred for 16 h. Saturated NH_4Cl was added to the mixture and was stirred for 30 min. The aqueous phase was extracted three times with CH_2Cl_2. The organic phase was dried over Na_2SO_4, concentrated under reduced pressure, loaded onto Celite, and purified by flash chromatography (SiO_2; hexane/EtOAc 95:5 to hexane/EtOAc 8:2). Product **6** was obtained as a white/blue powder (130 mg, 30%). ^1H NMR (400 MHz, $CDCl_3$) δ = 7.61 (s, 1H, H8), 7.51 (d, *J* = 7.9, 1H, H7), 7.11 (s, 1H, H6), 6.95 (d, *J* = 2.9, 2H, H2), 6.92 (dd, *J* = 8.9, 1.2, 2H, H5), 6.63 (dd, *J* = 8.9, 2.9, 2H, H4), 5.32 (s, 2H, H9), 2.96 (d, J=0.8, 12H, H3), 0.62 (d, *J* = 0.8, 3H, H1 or H1'), 0.54 (d, *J* = 1.2, 3H, H1 or H1') ppm. ^{13}C NMR (101 MHz, $CDCl_3$) δ 152.29, 149.01, 140.48, 137.14, 135.28, 131.72, 128.55, 125.49, 125.37, 119.16, 116.69, 113.97, 111.11, 92.85, 72.22, 40.53, 0.56, -0.76. HRMS (ESI-MS) $[M+H]^+$ calcd. for $[C_{27}H_{30}N_3OSi]^+$: 440.21527, found 440.21522.

5'-(Aminomethyl)-*N*3,*N*3,*N*7,*N*7,5,5-hexamethyl-3'H,5H-spiro[dibenzo[b,e]siline-10,1'-isobenzofuran]-3,7-diamine (24)



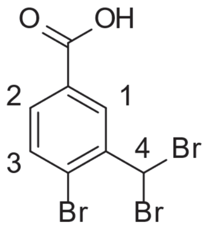
Compound **23** (130 mg, 0.3 mmol) was transferred to a flame-dried reaction flask with 2.4 mL of THF and cooled to 0 °C. The flask was flushed with nitrogen. The solution was stirred, and a $LiAlH_4$ solution (0.65 mL, 0.65 mmol, 1 M) at 0 °C was added dropwise. The color changed from green-blue to yellow to dark orange to brownish green. After 2 h the reaction mixture was left warm to 20 °C. The reaction mixture was neutralized with a few drops of saturated NH_4Cl. It was extracted three times with CH_2Cl_2. The organic layer was dried with brine and then over Na_2SO_4, concentrated under reduced pressure, and loaded onto Celite. The crude was purified by flash chromatography (SiO_2; CH_2Cl_2/CH_3OH 1:0 to CH_2Cl_2/CH_3OH 92:8), yielding compound **24** as a white powder (80 mg, 66%). ^1H NMR (400 MHz, $CDCl_3$) δ = 7.26 (s, 1H, H8, covered by $CHCl_3$ peak), 7.18 (d, *J* = 7.9, 1H, H6), 7.03 (d, *J* = 7.7, 1H, H7), 6.98 (d, *J* = 5.5, 2H, H5), 6.96 (s, 2H, H2), 6.61 (dd, *J* = 8.9, 2.9, 2H, H4), 5.21 (s, 2H, H9), 3.93 (s, 2H, H10), 2.94 (s, 12H, H3), 0.61 (s, 3H, H1 or H1'), 0.54 (s, 3H, H1or H1') ppm. ^{13}C NMR (101 MHz, $CDCl_3$): *δ* = 148.12, 144.68, 139.80, 137.76, 134.92, 127.87, 126.10, 124.11, 119.37, 116.18, 113.18, 91.75, 71.53, 45.23, 39.93, 0.00, -1.84 ppm. HRMS (ESI-QTOF) $[M+H]^+$ calcd. for $[C_{27}H_{34}N_3OSi]^+$ 444.2471, found 444.2474.

N-(2-(2-((6-Chlorohexyl)oxy)ethoxy)ethyl)-3,7-bis(dimethylamino)-5,5-dimethyl-3'*H*,5*H*-spiro[dibenzo[b,e]siline-10,1'-isobenzofuran]-5'-carboxamide (HMSiR-Halo 25)



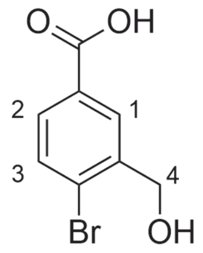
The protected ligand **26** 16.5 mg, 0.051 mmol) was dissolved in 4 M HCl in dioxane (0.35 mL) and was stirred at 20 °C for 1 h. A saturated $NaHCO_3$ solution was added to the reaction, and the mixture was extracted with CH_2Cl_2. The combined organic phases were dried over Na_2SO_4, the solvent was evaporated, and the product was directly used in the next step. In a flame-dried Schlenk flask under N_2-atmosphere, compound **32** (10 mg, 0.017 mmol) was dissolved in dry DMF (2 mL), then DIPEA(30 eq, 65.9 mg, 0.51 mmol) was added and the reaction was stirred at 20 °C for 10 min, then HATU (12.9 mg, 0.034 mmol) and the crude from the previous step was added and the reaction was stirred at 20 °C for 1.5 h. The solvent was evaporated, and the crude was stored in the freezer overnight. The crude was purified by reverse phase HPLC (SiO_2-C_{18}; CH_3CN + 0.1% TFA/H_2O + 0.1% TFA 5:95 to CH_3CN + 0.1% TFA/H_2O + 0.1% TFA 65:35). The CH_3CN was evaporated as well as parts of the H_2O, then sat. Na_2CO_3 was added until no bubbles were observed. The aqueous phase was extracted with CH_2Cl_2, the organic phases were dried over $NaSO_4$ and concentrated under reduced pressure. The oily solid (10 mg) was redissolved in CH_3CN/H_2O 1:1 and lyophilized giving 9 mg of compound **25** (80%).^{1}H NMR (400 MHz, CD_3OD) δ = 7.81 (s, 1H, H8), 7.64 (d, *J* = 8.2, 1H, H7), 7.00 (dd, *J* = 2309.9, 12.6, 2H, H2), 6.99 (d, *J* = 45.5, 2H, H5), 6.90 (d, *J* = 8.0, 1H, H6), 6.70 (dd, *J* = 8.9, 2.9, 2H, H4), 5.39 (s, 2H, H9), 3.68–3.61 (m, 4H, H19, H11), 3.60–3.54 (m, 4H, H10, H14), 3.45 (t, *J* = 6.6, 4H, H12, H13), 2.92 (s, 12H, H3), 1.73–1.58 (m, 2H, H18), 1.57–1.46 (m, 2H, H15), 1.42–1.30 (m, 4H, H16, H17), 0.59 (s, 3H, H1 or H1’), 0.49 (s, 3H, H1 or H1’) ppm. ^{13}C NMR (101 MHz, CD_3OD) δ = 170.03, 152.59, 150.53, 139.73, 139.44, 135.44, 135.03, 129.92, 127.97, 124.72, 121.76, 117.55, 115.66, 94.17, 73.83, 72.18, 71.29, 71.13, 70.47, 45.67, 41.02, 40.81, 33.68, 30.49, 27.67, 26.41, 0.22, -0.20 ppm. HRMS (ESI-QTOF) $[M+H]^+$ calcd. for $[C_{37}H_{51}ClN_3O_4Si]^+$ 664.3332, found 664.3329.

4-Bromo-3-(dibromomethyl)benzoic acid (27)



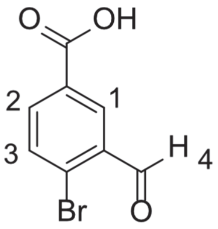
4-Bromo-3-methylbenzoic acid (10 g, 46.5 mmol), NBS (8.28 g, 46.5 mmol), and AIBN (99.3 mg, 605 μmol) were suspended in α, α, α-trifluorotoluene (149 mL) and the solution was heated to 110 °C. After it was heated more AIBN (50 mg) and 1 additional equivalent of NBS was added. AIBN (50 mg) was further added every 3 to 4 h for a total of 24 h. The mixture was cooled to 25 °C and 50 mL of heptane was added dropwise. This mixture was stirred in a salt ice bath at –10 °C for 3 h. The solid was concentrated by filtration, washed with heptane at –20 °C, and dried under reduced pressure to give compound **27** as a white solid (15.5 g, 89%). ^{1}H NMR (400 MHz, DMSO-*d*$_6$) δ = 8.47 (s, 1H, H3), 7.81 (d, *J* = 1.2 Hz, 2H, H1, H2), 7.39 (s, 1H, H4). ^{13}C NMR (101 MHz, DMSO-*d*$_6$) δ = 165.86, 140.22, 133.84, 131.88, 131.30, 131.04, 40.57. HRMS (ESI/QTOF) [M-H]$^-$ calcd. for $[C_8H_4Br_3O_2]^-$: 368.7767, found 368.7765.

4-Bromo-3-(hydroxymethyl)benzoic acid (28)



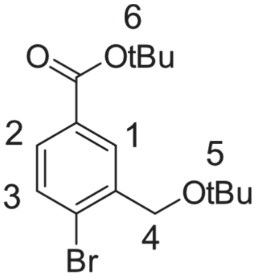
Compound **29** (1 g, 4.37 mmol) was dissolved in THF (20 mL) and then cooled to 0 °C. The mixture was stirred between 0 and 5 °C for 5 h. Then 1 M HCl was added, and the mixture was fully concentrated under reduced pressure. The residue was dissolved in EtOAc and 1 M HCl. The aqueous acidic phase was removed, and the organic phase was washed with water and brine, dried over Na_2SO_4, and concentrated under reduced pressure. Little product was obtained (clean by NMR, 70 mg) therefore the aqueous phase was further acidified by the addition of 1 M HCl. The aqueous phase was extracted twice with CH_2Cl_2, dried over Na_2SO_4, and concentrated under reduced pressure. The combined product **28** was a white solid (0.94 g, 93%). ^{1}H NMR (400 MHz, CD_3OD) δ = 8.21 (dt, *J* = 2.0, 0.9, 1H, H1), 7.81 (ddt, *J* = 8.4, 2.2, 0.7, 1H, H2), 7.66 (d, *J* = 8.3, 1H, H3), 4.69 (t, *J* = 0.8, 2H, H4) ppm. ^{13}C NMR (101 MHz, CD_3OD) δ = 169.10, 142.40, 133.68, 131.52, 130.71, 130.29, 128.00, 64.34 ppm. HRMS (ESI/QTOF) [M-H]$^-$ calcd. for $[C_8H_6BrO_3]^-$: 228.9506, found 228.9504.

4-Bromo-3-formylbenzoic acid (29)



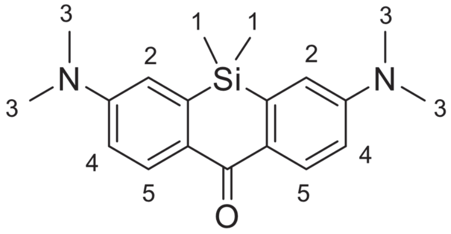
Compound 27 (4.77 g, 12.8 mmol) was added to a stirred solution of 10% aqueous Na_2CO_3 (84 mL). The solution was heated to 70 °C and stirring was continued for 5 h. The resulting precipitate was removed by filtration while the solution was still hot. The filtrate was carefully acidified with HCl until pH 1 was reached and extracted twice with EtOAc. The combined organic phase was washed with brine, dried over Na_2SO_4, filtered, and evaporated to give 29 as a white solid (9.5 g, 96%). ^1H NMR (400 MHz, $(CD_3)_2CO$) δ = 10.37 (s, 1H, H4), 8.47 (d, *J* = 2.2 Hz, 1H, H1), 8.17 (d, *J* = 8.3 Hz, 1H, H2), 7.95 (d, *J* = 8.3 Hz, 1H, H3). ^{13}C NMR (101 MHz, $(CD_3)_2CO$) δ = 190.37, 165.17, 135.62, 134.62, 133.88, 130.80, 130.59. HRMS (ESI/QTOF) $[M-H]^-$ calcd. for $[C_8H_4BrO_3]^-$: 226.9349, found 226.9356.

tert-Butyl 4-bromo-3-(*tert*-butoxymethyl)benzoate (30)



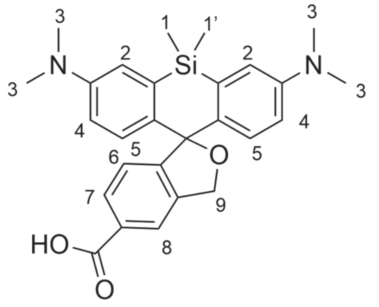
Concentrate H_2SO_4 (378 µL, 7.1 mmol) was added to a vigorously stirred suspension of anhydrous Epsom salt (3.42 g, 28.4 mmol) in CH_2Cl_2 (60 mL). The mixture was stirred for 15 min at 25 °C then compound 29 (820 mg, 3.55 mmol) and *tert*-butanol (3.37 mL) were added. The flask was closed, and the mixture was stirred for 5 d at 25 °C. After the addition of saturated $NaHCO_3$ the reaction mixture was extracted twice with EtOAc, washed with brine, dried over Na_2SO_4, filtered, and concentrated under reduced pressure. The crude was purified by flash column chromatography (SiO_2; hexane to hexanes/EtOAc 91:9) to give the product as a white oil, that crystallized when left unmoved (0.972 g, 80%).^1H NMR (400 MHz, $CDCl_3$) δ = 8.14 (dd, *J* = 2.1, 1.0, 1H, H1), 7.71 (dd, *J* = 8.3, 2.3, 1H, H2), 7.55 (d, *J* = 8.2, 1H, H3), 4.50 (s, 2H, H4), 1.59 (s, 9H, H6), 1.32 (s, 9H, H5) ppm. ^{13}C NMR (101 MHz, $CDCl_3$) δ = 165.40, 139.48, 132.34, 131.42, 130.11, 129.33, 127.26, 81.40, 74.12, 63.53, 28.30, 27.77 ppm. HRMS (ESI/QTOF) $[M+Na]^+$ calcd. for $[C_{16}H_{23}BrNaO_3]^+$: 365.0723, found 365.0721.

3,7-bis(Dimethylamino)-5,5-dimethyldibenzo[b,e]silin-10(5H)-one (31)



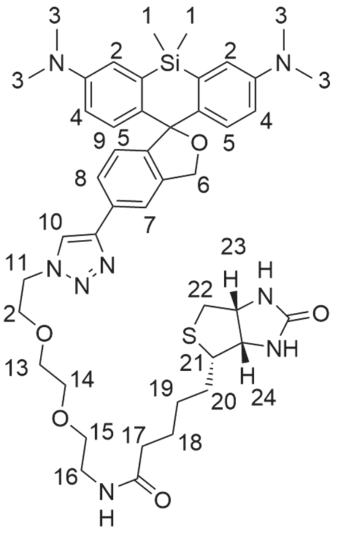
t-BuLi solution (5.15 mL, 8.76 mmol, 1.7 M) was added dropwise to a solution of compound 5 (0.999 g, 2.19 mmol) in anhydrous THF (40 mL) at –78 °C. The resulting bright yellow solution was stirred at –78 °C for 1.5 h. Neat dimethyl carbamoyl chloride (0.259 g, 2.41 mmol) was then added dropwise. The resulting mixture was stirred at –78 °C for 50 min, then allowed to warm up to 25 °C and left stirring for 16 h. The reaction was stopped with saturated NH_4Cl solution (25 mL), water was added to dissolve solids, and the mixture was extracted with EtOAc three times. The combined extracts were dried over Na_2SO_4, filtered, and the product was isolated by flash column chromatography (SiO_2; hexane to hexane/EtOAC 9:1) to give 650 mg (92%) of the ketone 31 as a bright green crystalline solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ = 8.40 (d, *J*=8.9, 2H, H2), 6.84 (dd, *J*=8.9, 2.7, 2H, H4), 6.79 (d, *J*=2.8, 2H, H5), 3.10 (s, 12H, H3), 0.47 (s, 6H, H1). ^{13}C NMR (101 MHz, $CDCl_3$) δ = 185.41, 151.59, 140.61, 131.76, 129.83, 114.38, 113.30, 40.18, -0.85. HRMS (ESI/QTOF) $[M+H]^+$ calcd. for $[C_{19}H_{25}N_2OSi]^+$: 325.1731, found 325.1763.

3,7-bis(Dimethylamino)-5,5-dimethyl-3'*H*,5*H*-spiro[dibenzo[b,e]siline-10,1'-isobenzofuran]-5'-carboxylic acid (HMSiR-carboxy 32)



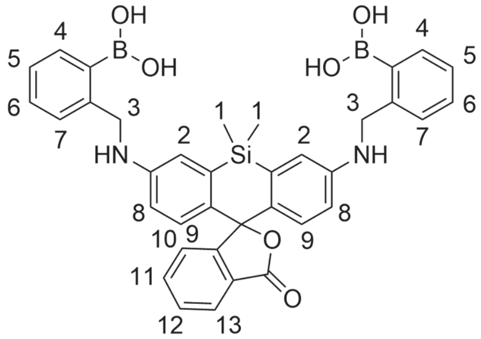
Compound **30** (291 mg, 0.847 mmol) was dissolved in THF (5 mL) under N_2 atmosphere in a pre-dried flask and cooled to –78 °C. Then, *t*-BuLi solution (0.45 mL, 0.77 mmol, 1.7 M) was added and the reaction was stirred at the same temperature for 10 min. Compound **31** (50 mg, 0.154 mmol) dissolved in anhydrous THF (2 mL) and was added. The resulting mixture was warmed to ambient temperature and stirred for 16 h. Saturated NH_4Cl was added to the reaction mixture and extracted three times with EtOAc and once with CH_2Cl_2. The combined organic phases were dried over $Na_2SO_{4,}$ and the solvent was removed under reduced pressure. The resulting intermediate was dissolved in TFA (3.5 g, 30.8 mmol) and the solution was stirred at 20 °C for 3.5 h. After removal of all volatiles under reduced pressure the crude was purified by reverse phase HPLC (SiO_2-C_{18}; CH_3CN + 0.1% TFA/H_2O + 0.1% TFA 1:9 to CH_3CN + 0.1% TFA/H_2O + 0.1% TFA 95:5) yielding 56 mg of the product **32** as a blue solid (56%). ^1H NMR (400 MHz, CD_3OD) δ = 8.41 (s, 1H, H8), 8.11 (dd, *J* = 7.8, 1.7, 1H, H7), 7.37 (s, 2H, H2), 7.28 (d, *J* = 7.9, 1H, H6), 7.04 (d, *J* = 9.6, 2H, H5), 6.78 (dd, *J* = 9.6, 2.9, 2H, H4), 4.36 (s, 2H, H9), 3.35 (s, 12H, H3), 0.61 (d, *J* = 3.7, 6H, H1) ppm. ^{13}C NMR (101 MHz, CD_3OD) δ = 169.14, 167.95, 155.83, 149.40, 142.97, 142.12, 141.52, 132.78, 130.66, 129.46, 129.22, 128.15, 122.33, 115.32, 62.02, 40.93, -1.09, -1.33 ppm. HRMS (ESI/QTOF) $[M+H]^+$ calcd. for $[C_{27}H_{31}N_2O_3Si]^+$: 459.2099; found 459.2097.

N-(2-(2-(2-(4-(3,7-bis(Dimethylamino)-5,5-dimethyl-3'H,5H-spiro[dibenzo[b,e]siline-10,1'-isobenzofuran]-5'-yl)-1*H*-1,2,3-triazol-1-yl)ethoxy)ethoxy)ethyl)-5-((3a*S*,4S,6a*R*)-2-oxohexahydro-1*H*-thieno[3,4-d]imidazol-4-yl)pentanamide (HMSiR-biotin 33)



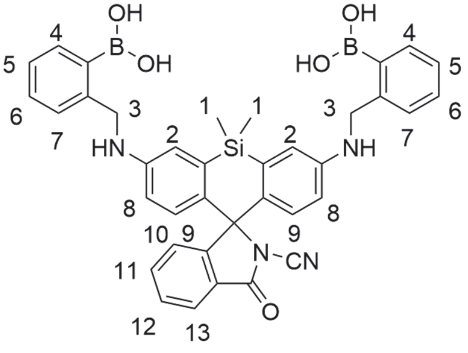
A solution of compound 34 (9.13 mg, 0.023 mmol), compound 32 (10 mg, 0.023 mmol), and $CuSO_4$ 5 H_2O (0.569 mg, 0.002 mmol) in *t*-BuOH/H_2O/THF/CH_2Cl_2 (2 mL, 1:1:1:1) at 20 °C was degassed by bubbling argon through the mixture for 5 min. A solution of sodium ascorbate (1.0 M in H_2O, 3 drops) was added. The mixture was stirred at 20 °C for 4 h. The mixture was carefully diluted with water and extracted with EtOAc three times. The combined organic phase was dried over Na_2SO_4, the solvent was evaporated and the residue was purified by reverse phase medium pressure LC (MPLC) (SiO_2-C_{18}; CH_3CN/H_2O; 5:95 to CH_3CN/H_2O 95:5) and HPLC (SiO_2-C_{18}; CH_3CN/H_2O 5:95 to CH_3CN/H_2O 95:5) giving the product 33 as a blue solid (17 mg, 98%).^{1}H NMR (400 MHz, CD_3OD) δ = 8.52 (s, 1H, H10), 8.20 (s, 1H, H7), 7.94 (d, *J* = 7.7, 1H, H8), 7.37 (d, *J* = 2.8, 2H, H2), 7.24 (d, *J* = 7.9, 1H, H9), 7.18 (d, *J* = 9.6, 2H, H5), 6.79 (dd, *J* = 9.7, 2.9, 2H, H4), 4.70 (t, *J* = 4.9, 2H, H11), 4.45 (dd, *J* = 7.9, 4.8, 1H, H23), 4.37 (s, 2H, H6), 4.25 (dd, *J* = 7.9, 4.4, 1H, H24), 3.99 (t, *J* = 4.9, 2H, H12), 3.72–3.58 (m, 4H, H15, H13), 3.54–3.44 (m, 4H, H14, H16), 3.35 (s, 12H, H3), 3.20–3.09 (m, 1H, H21), 2.87 (dd, *J* = 12.7, 5.0, 1H, H22), 2.67 (d, *J* = 12.7, 1H, H22'), 2.16 (t, *J* = 7.3, 2H, H17), 1.59 (tdt, *J* = 28.9, 16.1, 7.1, 4H, H18, H20), 1.38 (p, *J* = 7.6, 2H, H19), 0.62 (s, 3H, H1), 0.61 (s, 3H, H1) ppm. ^{13}C NMR (101 MHz, CD_3OD) δ = 176.07, 169.05, 166.05, 155.82, 149.43, 148.03, 142.46, 141.75, 138.24, 132.77, 131.11, 128.64, 125.31, 125.10, 123.66, 122.20, 115.27, 71.52, 71.23, 70.63, 70.33, 66.91, 63.33, 62.34, 61.60, 56.99, 51.60, 41.04, 40.93, 40.35, 36.71, 29.73, 29.50, 26.84, 15.44, -1.04, -1.31 ppm. HRMS (ESI-QTOF) $[M+H]^+$ calcd. for $[C_{44}H_{59}N_8O_5SSi]^+$ 839.4093, found 839.4115.

((((5,5-Dimethyl-3'-oxo-3'*H*,5*H*-spiro[dibenzo[b,e]siline-10,1'-isobenzofuran]-3,7-diyl)bis(azanediyl))bis(methylene))bis(2,1-phenylene))diboronic acid (SiRhoBo 36)



Compound **42** (10 mg, 0.027 mmol), 2-formylphenylboronic acid, and sodium triacetoxyborohydride (31.8 mg, 0.15 mmol) were dissolved in (11.2 mg, 0.075 mmol) dichloroethane (3 mL) in a MW vial and headed to 140 °C for 40 min. The MW vial was opened, acetic acid was added in one portion and the mixture was heated for 40 min at 140 °C. The vial opened and 2-formylphenylboronic acid (11.2 mg, 0.075 mmol) and sodium triacetoxyborohydride (31.8 mg, 0.15 mmol) this addition and heating were repeated twice for a total heating time of 160 min. The mixture was neutralized with saturated $NaHCO_3$, and the aqueous layer was extracted with CH_2Cl_2. After concentration under reduced pressure, the target compound was isolated by preparative thin-layer chromatography (SiO_2; CH_2Cl_2 (+NH_3 extracted from an aqueous solution)/CH_3OH 9:1) and further purified by preparative HPLC yielded 3 mg of **36** as a blue solid (12%) (SiO_2-C_{18}; CH_3CN/H_2O 5:95 to CH_3CN/H_2O 95:5).^{1}H NMR (500 MHz, CD_3OD) δ = 7.92 (d, *J* = 7.3, 1H, H13), 7.69 (td, *J* = 7.5, 1.2, 1H, H11), 7.59 (td, *J* = 7.5, 0.9, 1H, H12), 7.37–7.28 (m, 4H, H4, H5), 7.28–7.15 (m, 5H, H6, H7, H10), 6.95 (s, 2H, H2), 6.61 (d, *J* = 8.8, 2H, H9), 6.51 (dd, *J* = 8.8, 2.6, 2H, H8), 4.35 (s, 4H, H3), 0.39 (s, 3H, H1), 0.35 (s, 3H, H1) ppm. ^{13}C NMR (126 MHz, CD_3OD) δ = 172.86, 164.88, 156.16, 149.21, 145.35, 137.66, 135.37, 134.76, 132.53, 131.07, 130.24, 129.63, 128.77, 127.48, 126.45, 125.90, 124.08, 120.18, 116.61, 49.82, -0.02, -1.66 ppm. HRMS (ESI-QTOF) $[M+H]^+$ calcd. for $[C_{36}H_{35}B_2N_2O_6Si]^+$ 641.2445, found 641.2454.

((((2'-Cyano-5,5-dimethyl-3'-oxo-5*H*-spiro[dibenzo[b,e]siline-10,1'-isoindoline]-3,7-diyl)bis(azanediyl))bis(methylene))bis(2,1-phenylene))diboronic acid (37)



General procedure A with the following specifications:

Diamine: compound 53 (10 mg, 0.025 mmol)

Aldehyde: 2-formylphenylboronic acid (11.3 mg, 0.076 mmol)

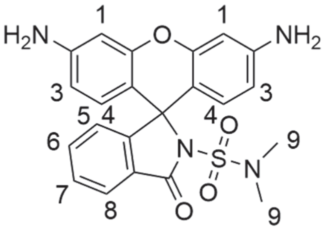
Stirring time 1: 45 min

Stirring time 2: 4 h

Yielded compound 37 as a very light blue solid (5 mg, 30%).

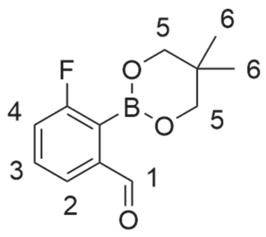
^{1}H NMR (400 MHz, CD_3OD) δ = 7.94 (dt, *J* = 7.5, 1.1, 1H, H13), 7.60 (td, *J* = 7.5, 1.4, 1H, H11), 7.54 (td, *J* = 7.5, 1.1, 1H, H12), 7.40–7.17 (m, 8H, H4, H5, H6, H7), 6.85 (d, *J* = 7.7, 1H, H10), 6.82 (d, *J* = 2.6, 2H, H2), 6.60 (dd, *J* = 8.9, 2.6, 2H, H8), 6.53 (d, *J* = 8.8, 2H, H9), 4.36 (s, 4H, H3), 0.32 (s, 3H, H1), 0.25 (s, 3H, H1) ppm. ^{13}C NMR (101 MHz, CD_3OD) δ = 169.29, 156.66, 148.84, 145.46, 136.97, 136.75, 132.60, 131.64, 130.12, 129.83, 129.65, 127.44, 126.49, 125.61, 125.09, 118.68, 118.31, 107.82, 76.01, -0.11, -0.38 ppm. HRMS (ESI-QTOF) $[M+4CH_2+H]^+$ calcd. for $[C_{37}H_{35}B_2N_4O_5Si]^+$ 665.2557, found 665.2562.

3',6'-Diamino-*N*,*N*-dimethyl-3-oxospiro[isoindoline-1,9'-xanthene]-2-sulfonamide (38)



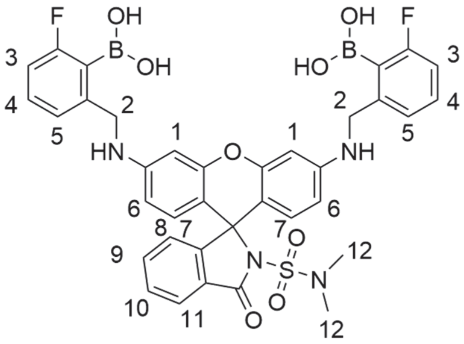
Rhodamine 110-chloride (200 mg, 0.545 mmol), *N*,*N*-dimethylsulfamide (1354 mg, 10.9 mmol), HATU (249 mg, 0.654 mmol), and DIPEA (1057 mg, 8.18 mmol) were dissolved in DMF (24 mL). The resulting mixture was stirred at 20 °C for 1 h. The solvent was removed by lyophilization. The crude was purified by flash chromatography (SiO_2; CH_2Cl_2 (+NH_3 extracted from a 25% aqueous solution)/CH_3OH 9:1) and reverse phase MPLC (SiO_2-C_{18}; CH_3CN/H_2O 5:95 to CH_3CN/H_2O 95:5) yielding 10 mg of product 38 as a red solid (5%). ^{1}H NMR (500 MHz, DMSO-d_6) δ = 8.67 (d, *J* = 7.6, 1H, H8), 8.46 (td, *J* = 7.5, 1.2, 1H, H6), 8.38 (t, *J* = 7.5, 1H, H7), 7.80 (d, *J* = 7.7, 1H, H5), 7.11 (d, *J* = 8.5, 2H, H4), 7.08 (d, *J* = 2.2, 2H, H1), 6.99 (dd, *J* = 8.5, 2.2, 2H, H3), 6.18 (s, 4H. H2), 3.42 (s, 6H, H9) ppm. ^{13}C NMR (126 MHz, DMSO-d_6) δ = 175.91, 163.13, 161.94, 159.54, 144.48, 138.43, 138.00, 136.90, 133.95, 132.67, 119.51, 115.94, 108.92, 78.00, 46.91 ppm. HRMS (ESI-QTOF) $[M+H]^+$ calcd. for $[C_{22}H_{21}N_4O_4S]^+$ 437.1278, found 437.1281.

2-(5,5-Dimethyl-1,3,2-dioxaborinan-2-yl)-3-fluorobenzaldehyde (40)



2-Bromo-3-fluorobenzaldehyde (200 mg, 0.985 mmol), bis(neopentylglycolato)diboron (245 mg, 1.08 mmol), KOAc (387 mg, 3.94 mmol) and Pd(dppf)Cl_2・CH_2Cl_2 complex (56.3 mg, 0.069 mmol) were weighted in a flame dried 10 mL Schlenkflask and the atmosphere was exchanged three times for argon. The solids were dissolved in dioxane (4 mL) and heated to 70 °C. The heating was turned off after 2 h and the reaction was allowed to cool to 25 °C over 16 h. The mixture was filtered through Celite and purified by flash column chromatography (SiO_2; hexane/EtOAc 1:0 to hexane/EtOAc 9:1) yielding compound **40** as an off-white powder (200 mg, 86%). ^{1}H NMR (400 MHz, $CDCl_3$) δ = 9.99 (d, *J* = 2.6, 1H, H1), 7.60 (d, *J* = 7.4, 0.7, 1H, H2), 7.49 (ddd, *J* = 8.2, 7.4, 5.4, 1H, H4), 7.27 (td, *J* = 7.9, 1.1, 1H, H3), 3.83 (s, 4H, H5), 1.15 (s, 6H, H6) ppm. ^{13}C NMR (101 MHz, $CDCl_3$) δ = 192.82, 192.79, 166.43, 164.00, 141.51, 141.42, 131.07, 130.99, 128.02, 127.99, 121.36, 121.11, 72.88, 32.10, 22.19 ppm.

((((2-(*N*,*N*-Dimethylsulfamoyl)-3-oxospiro[isoindoline-1,9'-xanthene]-3',6'-diyl)bis(azanediyl))bis(methylene))bis(6-fluoro-2,1-phenylene))diboronic acid (41)



General procedure B with the following specifications:

Diamine: compound **38** (5 mg, 0.011 mmol)

Aldehyde: compound **40** (8.11 mg, 0.034 mmol)

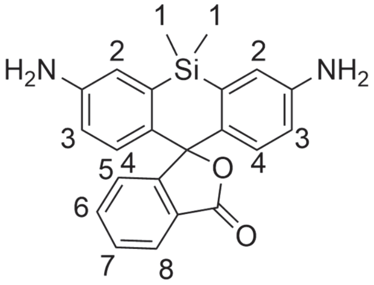
Stirring time 1: 1 h

Stirring time 2: 1 h

Yielded compound **41** as a light pink solid (6 mg, 71%).

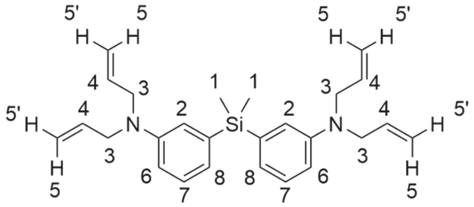
^{1}H NMR (500 MHz, CD_3OD) δ = 7.89 (d, *J* = 7.6, 1H, H11), 7.60 (t, *J* = 7.5, 1H, H9), 7.56 (t, *J* = 7.2, 1H, H10), 7.35 (ddd, *J* = 13.8, 7.7, 6.2, 2H, H3), 7.19 (d, *J* = 7.6, 2H, H5), 6.95–6.88 (m, 3H, H4, H8), 6.43 (d, *J* = 9.4, 2H, H7), 6.31–6.31 (m, 2H, H1, H6), 4.31 (s, 4H, H2), 2.66 (s, 6H, H12) ppm. ^{13}C NMR (126 MHz, CD_3OD) δ = 168.93, 167.32, 165.43, 155.35, 154.29, 151.06, 148.60, 148.53, 136.13, 131.84, 131.78, 130.19, 129.64, 129.22, 125.66, 124.46, 123.34, 123.32, 114.25, 114.05, 111.90, 110.42, 110.41, 101.31, 101.30, 70.05, 38.2 ppm. HRMS (ESI-QTOF) $[M+4CH_2+H]^+$ calcd. for $[C_{40}H_{41}B_2F_2N_4O_8S]^+$ 797.2794, found 797.2802

3,7-Diamino-5,5-dimethyl-3'H,5H-spiro[dibenzo[b,e]siline-10,1'-isobenzofuran]-3'-one (42)



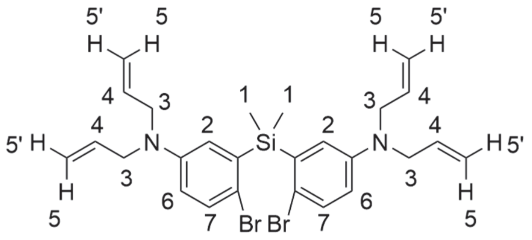
3,3'-(Dimethylsilanediyl)bis(*N*,*N*-diallyl-4-bromoaniline) 45 (108 mg, 0.193 mmol) and phthalic anhydride (62.9 mg, 0.425 mmol) were dried under reduced pressure in two flame-dried flasks and dissolved in dry THF (7 mL and 10 mL respectively). The solution of 3,3'-(diallyllsilanediyl)bis(4-bromo-*N,N*-dimethyl aniline) was cooled to –78 °C using an acetone/dry ice bath. *t*-BuLi solution (0.5 mL, 0.849 mmol, 1.7 M) was added, and the reaction mixture was stirred at –78 °C for 50 min, then warmed to -20 °C and cooled again to –78 °C. The phthalic anhydride solution was added dropwise over 90 min. The reaction mixture was then allowed to warm to 20 °C over 16 h. 0.1% TFA in water was added, during which the mixture turned deep blue and warmed to 30 °C. When it reached 25 °C again, the solution was extracted with EtOAc three times. The organic phase was dried over Na_2SO_4 and concentrated under reduced pressure. The resultant first intermediate crude was dissolved in methanol (14 mL) and sodium borohydride (81 mg, 2.14 mmol) was added. The solution was stirred at 20 °C until the intense blue color disappeared (< 1 min) and thereafter stirred further for 20 min. Water was added and the aqueous solution was extracted three times with EtOAc and three times with CH_2Cl_2. The combined organic layers were washed with water and brine, dried over $MgSO_4$, and concentrated under reduced pressure. The intermediate crude was dissolved in degassed CH_2Cl_2 (5 mL). 1,3-Dimethylbarbituric acid (307 mg, 1.96 mmol) and $Pd(PPh_3)_4$ (82.5 mg, 0.071 mmol) were added. The mixture was stirred at 35 °C for 27 h. The crude was columned by flash chromatography revealing that the deprotection was not complete. All the fractions containing a partially deprotected product were combined and dried under reduced pressure. 1,3-Dimethylbarbituric acid (307 mg, 1.96 mmol) and $Pd(PPh_3)_3$ (82.5 mg, 0.071 mmol) were added to this dry mixture under an argon atmosphere. Anhydrous CH_2Cl_2 was added (10 mL) and the reaction was heated to 30 °C for 27 h and stirred at 25 °C for another 24 h. The reaction was completed and filtered through Celite the crude was concentrated and purified using flash chromatography (SiO_2; hexane/EtOAc 1:0 to hexane/EtOAc 4:6). All the fractions containing the final product were collected and concentrated under reduced pressure. CH_2Cl_2 with extracted NH_3 was added and a white precipitate resulted, water was added, and the organic phase was washed two times with water and once with brine, to remove the baseline impurity and the barbituric acid. The organic phase was dried over Na_2SO_4, concentrated under reduced pressure, and washed with pentanes giving the compound as a yellow/brown solid. The product was further purified by reverse phase MPLC yielding compound 42 as a blue solid (30 mg, 23% over two steps). ^{1}H NMR (500 MHz, CD_3OD) δ = 8.06 (d, *J* = 7.8, 1H, H8), 7.80 (t, *J* = 7.5, 1H, H6), 7.68 (t, *J* = 7.6, 1H, H7), 7.35–7.29 (m, 3H,H2, H5), 6.91 (d, *J* = 8.8, 1H, H4), 6.80 (d, *J* = 8.3, 1H, H3), 0.64 (s, 2H, H1), 0.58 (s, 2H) ppm. ^{13}C NMR (126 MHz, CD_3OD) δ = 176.55, 135.20, 130.79, 128.63, 127.52, 124.66, 119.95, 119.12, 117.34, 116.81, 115.08, 112.81, -0.20, -1.69 ppm. HRMS (ESI-QTOF) [M+H]+ calcd. for $[C_{22}H_{21}N_2O_2Si]$+ 373.1367, found 373.1370.

3,3'-(Dimethylsilanediyl)bis(*N*,*N*-diallylaniline) (44)



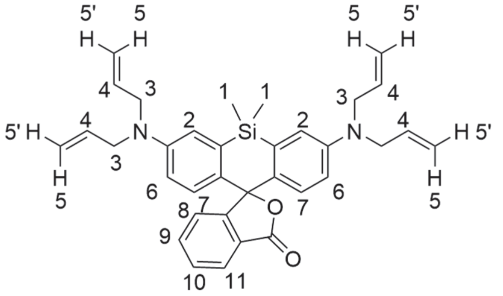
N,*N*-Diallyl-3-bromoaniline 43 (3 g, 11.9 mmol) was dissolved in diethyl ether (20 mL) in a well-dried flask flushed with argon. The solution was cooled to 0 °C, *n*-BuLi (4.0 mL, 6.4 mmol, 1.6 M) was added and the reaction mixture was stirred at 0 °C for 2 h. To this mixture dichlorodimethylsilane (0.768 g, 5.95 mmol) was added slowly (keeping the internal temperature below 5 °C, over 20 min). Then the mixture was allowed to warm up in the water ice bath for 9 h. Water was added to the reaction. The aqueous phase was extracted with diethyl ether three times and dried over Na_2SO_4. After filtration and removal of the solvent under reduced pressure, the residue was purified by flash column chromatography (SiO_2; hexane/CH_2Cl_2 1:0 to hexane/CH_2Cl_2 3:7) to give the product 44 (765 mg, 16%). ^{1}H NMR (400 MHz, $CDCl_3$) δ = 7.20 (dd, *J* = 8.3, 7.1, 2H, H7), 6.90–6.83 (m, 4H, H2, H8), 6.72 (ddd, *J* = 8.4, 2.8, 1.0, 2H, H6), 5.85 (ddt, *J* = 17.2, 10.2, 5.0, 4H, H4), 5.17 (dd, *J* = 12.8, 1.7, 4H, H5), 5.14 (dd, *J* = 5.9, 1.7, 4H, H5'), 3.91 (d, *J* = 5.0, 4H, H3), 0.51 (s, 6H, H1) ppm. ^{13}C NMR (101 MHz, $CDCl_3$) δ = 148.06, 139.01, 134.33, 128.59, 122.41, 118.28, 116.22, 113.36, 53.04, -2.20 ppm. HRMS (ESI-QTOF) $[M+H]^+$ calcd. for $[C26H_{35}N_2Si]^+$ 403.2564, found 403.2564.

3,3'-(Dimethylsilanediyl)bis(*N*,*N*-diallyl-4-bromoaniline) (45)



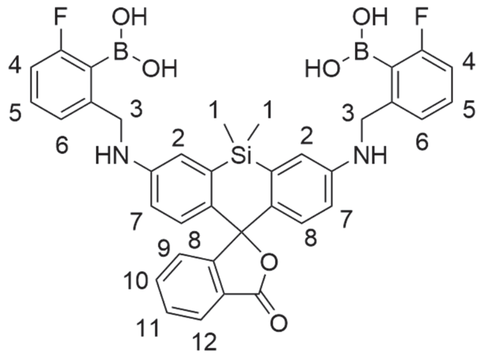
3,3'-(Dimethylsilanediyl)bis(*N*,*N*-diallylaniline) 44 (200 mg, 0.497 mmol) was dissolved in CH_2Cl_2 (15 mL) at 0 °C. NBS (195 mg, 1.09 mmol) was separately dissolved in CH_2Cl_2 (10 mL) and added dropwise to the reaction. After 30 min, the reaction was allowed to warm to 20 °C, and saturated $NaHCO_3$ solution was added. The aqueous phase was extracted with CH_2Cl_2 three times. The combined organic phases were evaporated under reduced pressure and the product was purified by flash chromatography (SiO_2; hexane/CH_2Cl_2 1:0 to hexane/CH_2Cl_2 9:1) to afford a yellowish oil that solidified after freeze drying to give compound 45 (696 mg, 87%). ^{1}H NMR (400 MHz, $CDCl_3$) δ = 7.30 (d, *J* = 8.8, 2H, H7), 6.79 (d, *J* = 3.3, 2H, H2), 6.56 (dd, *J* = 8.8, 3.3, 2H, H6), 5.78 (ddt, *J* = 17.1, 10.1, 4.9, 4H, H4), 5.15–5.05 (m, 8H, H5, H5'), 3.85 (dt, *J* = 4.9, 1.7, 8H, H3), 0.73 (s, 6H, H1) ppm.^{13}C NMR (101 MHz, $CDCl_3$) δ = 147.03, 138.77, 133.73, 133.10, 121.85, 116.42, 116.31, 115.15, 53.30, -0.72 ppm. HRMS (ESI-QTOF) $[M+H]^+$ calcd. for $[C_{26}H_{33}Br_2N_2Si]^+$ 559.0774, found 559.0777.

3,7-bis(Diallylamino)-5,5-dimethyl-3'*H*,5*H*-spiro[dibenzo[b,e]siline-10,1'-isobenzofuran]-3'-one (46)



Compound **45** (108 mg, 0.193 mmol) and phthalic anhydride (62.9 mg, 0.425 mmol) were dried under reduced pressure in two flame-dried flasks and dissolved in dry THF (2 mL and 4.25 mL respectively). The solution of compound **45** was cooled to –78 °C using an acetone/dry ice bath. A *t*-BuLi solution (0.5 mL, 0.849 mmol, 1.7 M) was added, and the reaction mixture was stirred at -78 °C for 40 min. The phthalic anhydride solution was added dropwise over 60 min by a syringe pump, during the addition it turned bright orange in color. The reaction mixture was then allowed to warm to 20 °C over 16 h. Saturated NH_4Cl was added to the reaction mixture and stirred for 30 min. The aqueous phase was extracted three times with CH_2Cl_2. The organic phase was dried over Na_2SO_4, concentrated under reduced pressure, loaded onto Celite, and purified by flash chromatography (SiO_2; hexane:EtOAc; 95:5 to 8:2) giving the product **46** light-blue powder (20 mg, 20%). ^{1}H NMR (500 MHz, CD_3OD) δ = 6.66 (d, *J* = 7.8, 1H, H11), 6.18 (t, *J* = 7.5, 1H, H9), 6.11 (t, *J* = 7.6, 1H, H10), 5.71 (s, 1H, H8), 5.69 (d, *J* = 2.8, 2H, H2), 5.42 (d, *J* = 9.5, 2H, H7), 5.16 (dt, *J* = 9.6, 2.1, 2H, H6), 4.33 (ddt, *J* = 16.0, 10.1, 4.9, 4H, H4), 3.69 (d, *J* = 10.4, 4H, H5’), 3.62 (d, *J* = 17.2, 4H, H5), 2.68 (s, 8H, H3), -1.01 (s, 3H, H1), -1.06 (s, 3H, H1) ppm. ^{13}C NMR (126 MHz, CD_3OD) δ = 168.50, 154.50, 148.75, 141.33, 133.47, 132.88, 132.72, 131.77, 131.11, 130.36, 125.48, 122.07, 120.24, 118.09, 117.96, 117.18, 115.68, 115.59, 114.92, 113.40, 54.62, -1.01, -2.06 ppm. HRMS (ESI-QTOF) $[M+H]^+$ calcd. for $[C_{34}H_{37}N_2O_2Si]^+$ 533.26188, found 533.2616.

((((5,5-Dimethyl-3'-oxo-3'*H*,5*H*-spiro[dibenzo[b,e]siline-10,1'-isobenzofuran]-3,7-diyl)bis(azanediyl))bis(methylene))bis(6-fluoro-2,1-phenylene))diboronic acid (47)



General procedure B with the following specifications:

Diamine: compound 42 (5 mg, 0.013 mmol)

Aldehyde: compound 40 (8.11 mg, 0.034 mmol)

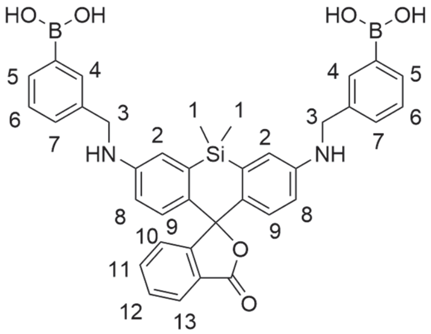
Stirring time 1: 20 min

Stirring time 2: 1 h

Yielded compound 47 as a light blue solid (6 mg, 66%).

^{1}H NMR (400 MHz, CD_3OD) δ = 7.90 (dt, *J* = 7.6, 1.0, 1H, H12), 7.68 (td, *J* = 7.5, 1.2, 1H, H10), 7.58 (td, *J* = 7.5, 0.9, 1H, H11), 7.34 (ddd, *J* = 13.9, 7.7, 6.3, 2H, H4), 7.21–7.12 (m, 3H, H6, H9), 6.95 (d, *J* = 2.6, 2H, H2), 6.89 (t, *J* = 8.2, 2H, H5), 6.63 (d, *J* = 8.7, 2H, H8), 6.53 (dd, *J* = 8.8, 2.6, 2H, H7), 4.33 (s, 4H, H3), 0.40 (s, 3H, H1), 0.36 (s, 3H, H1) ppm.^{13}C NMR (101 MHz, CD_3OD) δ = 172.85, 167.56, 165.19, 156.30, 148.78, 148.69, 148.44, 137.44, 135.51, 135.23, 131.80, 131.72, 130.27, 128.74, 127.21, 126.46, 125.45, 123.20, 120.26, 116.94, 114.21, 113.97, 93.29, -0.02, -1.57 ppm. HRMS (ESI-QTOF) $[M+4CH_2+H]^+$ calcd. for $[C_{40}H_{41}O_6N_2B_2F_2Si]^+$ 733.2883, found 733.2883.

((((5,5-Dimethyl-3'-oxo-3'*H*,5*H*-spiro[dibenzo[b,e]siline-10,1'-isobenzofuran]-3,7-diyl)bis(azanediyl))bis(methylene))bis(3,1-phenylene))diboronic acid (48)



General procedure A with the following specifications:

Diamine: compound 42 (9 mg, 0.024 mmol)

Aldehyde: 3-formylphenylboronic acid (10.9 mg, 0.072 mmol)

Stirring time 1: 45 min

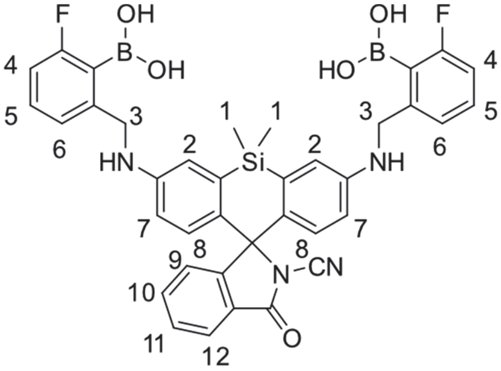
Stirring time 2: 22 h

Yielded compound 48 as a light blue solid (1 mg, 6%).

Extraction: Using MTBE as a solvent.

^{1}H NMR (500 MHz, CD_3OD) δ = 8.16 (s, 1H, H13), 7.73 (t, *J* = 7.8, 1H, H11), 7.66 (t, *J* = 7.6, 1H, H12), 7.60–7.30 (m, 8H, H4, H5, H6, H7), 7.26 (d, *J* = 7.5, 1H, H10), 7.14 (s, 2H, H2), 6.86 (s, 2H, H8), 6.60 (d, *J* = 8.9, 2H, H9), 4.54 (s, 4H, H3), 0.48 (s, 3H, H1), 0.44 (s, 3H, H1) ppm. ^{13}C NMR (101 MHz, CD_3OD) δ = 172.96, 155.92, 149.37, 142.84, 139.68, 138.17, 135.30, 133.42, 133.00, 132.90, 130.47, 130.18, 129.10, 128.66, 128.05, 126.26, 126.05, 119.42, 118.74, 114.80, 94.73, 40.42, 0.29, -1.84. ppm. HRMS (ESI-QTOF) [MH+TFA]$^+$ calcd. for $[C_{38}H_{34}B_2F_3N_2O_8Si]^+$ 753.2228, found 753.2233.

((((2'-Cyano-5,5-dimethyl-3'-oxo-5*H*-spiro[dibenzo[b,e]siline-10,1'-isoindoline]-3,7-diyl)bis(azanediyl))bis(methylene))bis(6-fluoro-2,1-phenylene))diboronic acid (49)



General procedure B with the following specifications:

Diamine: compound 53 (5 mg, 0.013 mmol)

Aldehyde: compound 40 (8.93 mg, 0.038 mmol)

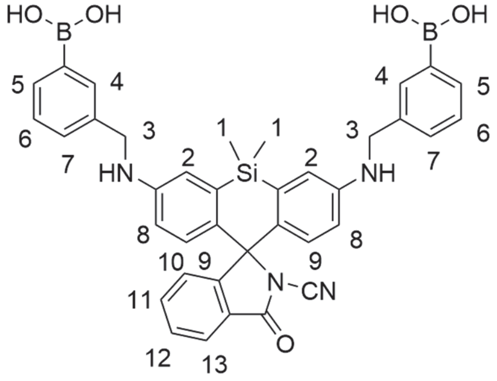
Stirring time 1: 20 min

Stirring time 2: 1.5 h

Yielded compound 49 as a light blue solid (7 mg, 66%).

^{1}H NMR (400 MHz, CD_3OD) δ = 7.95 (d, *J* = 6.9, 0.8, 1H, H12), 7.60 (td, *J* = 7.5, 1.3, 1H, H10), 7.55 (td, *J* = 7.5, 1.1, 1H, H11), 7.35 (ddd, *J* = 13.9, 7.8, 6.2, 2H, H4), 7.20 (d, *J* = 7.5, 2H, H6), 6.90 (t, *J* = 8.3, 0.9, 2H, H5), 6.86 (d, *J* = 0.9, 1H, H9), 6.83 (d, *J* = 2.6, 2H, H2), 6.62 (dd, *J* = 8.9, 2.6, 2H, H7), 6.56 (d, *J* = 8.8, 2H, H8), 4.36 (s, 4H, H3), 0.34 (s, 3H, H1), 0.26 (s, 3H, H1) ppm. ^{13}C NMR (101 MHz, CD_3OD) δ = 169.24, 167.62, 165.25, 156.59, 148.77, 148.68, 148.53, 137.01, 136.73, 132.05, 131.86, 131.77, 130.19, 129.91, 126.47, 125.68, 125.03, 123.37, 123.34, 118.81, 118.56, 114.23, 113.98, 107.77, 75.88, -0.14, -0.39 ppm. HRMS (ESI-QTOF) $[M+4CH_2+H]^+$ calcd. for $[C_{41}H_{41}B_2F_2N_4O_5Si]^+$ 757.2995, found 757.2998.

((((2'-Cyano-5,5-dimethyl-3'-oxo-5*H*-spiro[dibenzo[b,e]siline-10,1'-isoindoline]-3,7-diyl)bis(azanediyl))bis(methylene))bis(3,1-phenylene))diboronic acid (50)



General procedure A with the following specifications:

Diamine: compound 53 (10 mg, 0.025 mmol)

Aldehyde: 3-formylphenylboronic acid (11.3 mg, 0.076 mmol)

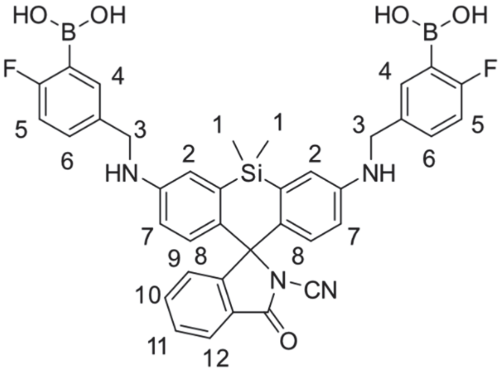
Stirring time 1: 45 min

Stirring time 2: 12 h

Yielded compound 50 as a very light blue solid (5 mg, 30%).

^{1}H NMR (500 MHz, CD_3OD) δ = 7.93 (d, *J*=7.6, 1H, H13), 7.61 (td, *J*=7.5, 1.2, 1H, H11), 7.58 (s, 2H, H4), 7.54 (dt, *J*=7.5, 0.8, 1H, H12), 7.45 (d, *J*=7.3, 1H, H10), 7.40 (d, *J*=7.7, 2H, H5), 7.31 (t, *J*=7.5, 2H, H6), 6.91 (d, *J*=7.3, 2H, H7), 6.82 (s, 2H, H2), 6.57 (dd, *J*=8.9, 2.7, 2H, H8), 6.51 (d, *J*=8.9, 2H, H9), 4.35 (s, 4H, H3), 0.37 (t, *J*=8.9, 6H, H1) ppm.^{13}C NMR (126 MHz, CD_3OD) δ = 169.49, 156.88, 149.40, 149.26, 140.09, 140.01, 137.01, 136.97, 134.16, 133.58, 133.36, 133.03, 130.45, 130.05, 129.98, 129.94, 129.73, 128.86, 128.71, 126.49, 125.46, 125.32, 117.48, 117.38, 116.84, 116.74, 108.12, 76.42, 0.12, -0.22 ppm. HRMS (ESI-QTOF) $[M+Na]^+$ calcd. for $[C_{37}H_{34}B_2N_4NaO_5Si]^+$ 687.2377, found 687.2381.

((((2'-Cyano-5,5-dimethyl-3'-oxo-5*H*-spiro[dibenzo[b,e]siline-10,1'-isoindoline]-3,7-diyl)bis(azanediyl))bis(methylene))bis(6-fluoro-3,1-phenylene))diboronic acid (51)



General procedure B with the following specifications:

Diamine: compound **53** (5 mg, 0.013 mmol)

Aldehyde: compound **40** (8.93 mg, 0.038 mmol)

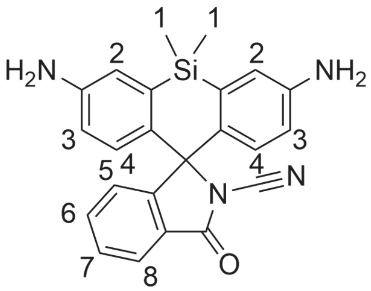
Stirring time 1: 20 min

Stirring time 2: 1.5 h

Yielded compound **51** as a light blue solid (2.8 mg, 27%).

^{1}H NMR (400 MHz, CD$_3$OD) δ = 7.94 (dt, *J* = 7.6, 1.0, 1H, H12), 7.62 (td, *J* = 7.5, 1.3, 1H, H10), 7.54 (td, *J* = 7.5, 1.0, 1H, H11), 7.43–7.35 (m, 4H, H4, H6), 6.98 (dd, *J* = 8.8, 8.8, 2H, H5), 6.91 (d, *J* = 7.8, 1H, H9), 6.81 (d, *J* = 2.6, 2H, H2), 6.57 (dd, *J* = 8.9, 2.6, 2H, H7), 6.52 (d, *J* = 8.8, 2H, H8), 4.31 (s, 4H, H3), 0.40 (s, 3H, H1), 0.39 (s, 3H, H1) ppm. ^{13}C NMR (101 MHz, CD$_3$OD) δ = 169.47, 156.85, 149.18, 137.04, 136.98, 136.56, 134.14, 131.30, 130.17, 130.01, 129.97, 126.51, 125.47, 125.32, 117.41, 116.85, 115.80, 115.55, 108.10, 76.39, 0.14, -0.20 ppm. HRMS (ESI-QTOF) [M+H]$^+$ calcd. for [C$_{37}$H$_{33}$B$_2$F$_2$N$_4$O$_5$Si]$^+$ 701.23689, found 701.2382.

3,7-Diamino-5,5-dimethyl-3'-oxo-5*H*-spiro[dibenzo[b,e]siline-10,1'-isoindoline]-2'-carbonitrile (53)



A solution of compound **45** (15 mg, 0.028 mmol) in CH_2Cl_2 (2 mL) was cooled to 0 °C. Oxalyl chloride (11.9 mg, 0.094 mmol) was added dropwise, the reaction turned deep blue, and was stirred at 20 °C for 1 h. The solvent was removed under reduced pressure and the crude under reduced pressure. In another dry flask, cyanamide (11.8 mg, 0.282 mmol) and DIPEA (36.4 mg, 0.282 mmol) were dissolved in CH_3CN (2 mL) and added to the crude followed by additional CH_3CN (4 mL). The resulting solution was stirred at 70 °C for 2.5 h. The solvent was evaporated, toluene was added and the crude was dried under reduced pressure. The crude product was used without further purification in the next step. Under an argon atmosphere, the crude cyanamide, 1,3-dimethylbarbituric acid (58.6 mg, 0.375 mmol), and $Pd(PPh_3)_3$ (11.7 mg, 0.009 mmol) were dissolved in a degassed mixture of CH_3OH/CH_2Cl_2 (5:1) (3 mL). The resulting mixture was stirred at 40 °C for 2 h. The reaction was diluted with CH_2Cl_2 and washed with saturated aqueous Na_2CO_3 solution. The aqueous phase was extracted twice with CH_2Cl_2. The combined organic phases were dried over Na_2SO_4 and the solvent was evaporated. The crude product **53** was purified by flash column chromatography (SiO_2; CH_2Cl_2/CH_3OH 1:0 to CH_2Cl_2/CH_3OH 9:1) to yield 7 mg of a light grey solid (94% over two steps). ^1H NMR (400 MHz, CD_3OD) δ = 7.96 (d, *J* = 7.6, 1H, H8), 7.65 (td, *J* = 7.6, 1.3, 1H, H6), 7.57 (td, *J* = 7.5, 1.1, 1H, H7), 7.00 (d, *J* = 2.5, 2H, H2), 6.96 (d, *J* = 7.8, 1H, H5), 6.64 (dd, *J* = 8.7, 2.6, 2H, H3), 6.55 (d, *J* = 8.7, 2H, H4), 0.54 (s, 3H, H1), 0.52 (s, 3H, H1) ppm.^{13}C NMR (101 MHz, CD_3OD) δ = 169.47, 156.87, 148.87, 137.22, 137.03, 131.23, 130.10, 130.08, 126.48, 125.55, 125.29, 119.92, 118.83, 108.03, 76.21, 33.08, 30.76, 30.61, 30.47, 30.33, 30.25, 28.11, 26.12, 23.74, 14.43, 0.15, -0.10 ppm. HRMS (ESI-QTOF) $[M+H]^+$ calcd. for $[C_{23}H_{21}N_4OSi]^+$ 397.1479, found 397.1481.

Chapter 8 Appendix

8.1 Figures and tables

8.1.1 Chapter 2

8.1.1.1 Figures

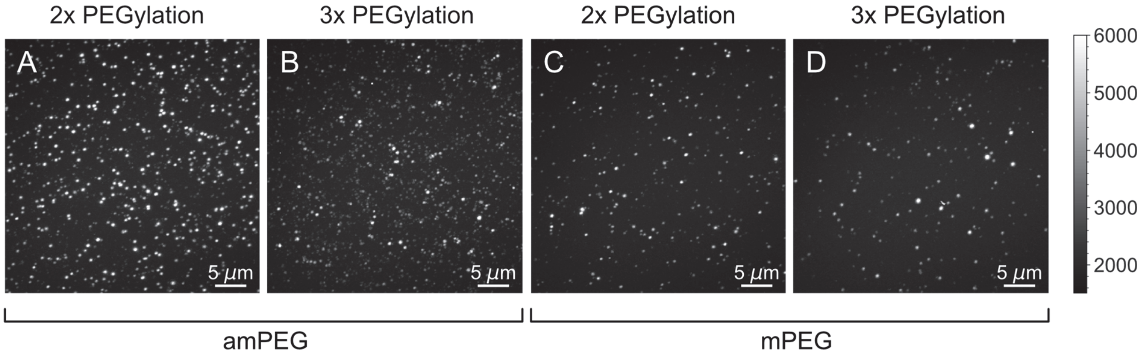


Figure 66. Maximum projections over 6000 frames of surface functionalization optimization experiments with alkyne-modified HMSiR 1. All coverslips were cleaned with the chemical cleaning procedure, before aminosilation and the first PEGylation step. A, B, Slides were PEGylated only with mPEG$_{5000}$-NHS. C, D, Slides were PEGylated with aPEG$_{5000}$-NHS and mPEG$_{5000}$-NHS. Slides A, C, Slides were directly PEGylated with mPEG$_4$-NHS and then reacted in the click reaction with alkyne-modified HMSiR 1. B, D Slides underwent an additional PEGylation step with mPEG$_{2000}$-NHS before PEGylation with mPEG$_4$-NHS. The color bar represents the fluorescence intensity (a.u.).

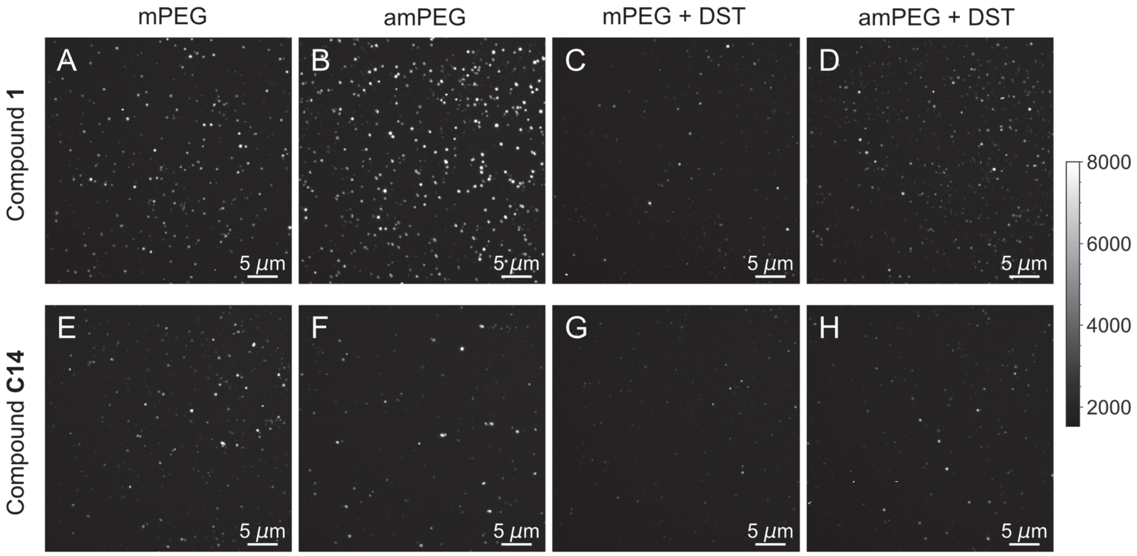


Figure 67. Maximum projections over 6000 frames of surface functionalization optimization experiments. All coverslips A-H were cleaned using a plasma cleaner for 2 min at 100 W on both sides. Slides A-D were functionalized with an alkyne-modified HMSiR **1**. E- H, Slides were reacted with peptide **C14** in the click reaction. A, E, C and G, Slides were functionalized with $aPEG_{5000}$-NHS and $mPEG_{5000}$-NHS. B, F, D and H, Slides were only PEGylated with $mPEG_{5000}$-NHS. C, D, G and H, Slides underwent an additional passivation step with DST. The color bar represents the fluorescence intensity (a.u.).

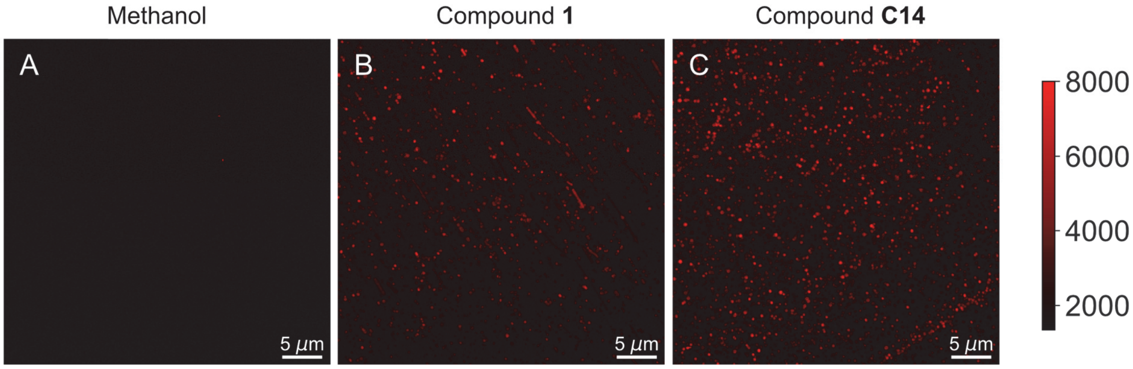


Figure 68. Spin coating experiment of coverslips cleaned using the standard protocol with a UV/Ozone cleaning step. First, a 100 µL of 1% PMMA in toluene was cast (1000 rpm, 20 s then 3000 rpm, 20 s). Then molecules were spin-cast (4000 rpm, 20 s at 4000 rpm followed by 3000 rpm, 20 s). A, Coverslip spin-coated with methanol only as a negative control. B, Coverslip spin-coated with a 1 nM solution of compound **1**. C, Coverslip spin-coated with a 1 nM solution of peptide **C14**. The color bar represents the fluorescence intensity (a.u.).

8.1.1.2 Tables

Tables 10–13 were reprinted from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* 2023, *145*, 1441–1447.[91] 🅭🅯

Table 10. Settings applied in the single-molecule localization software Picasso.[175] MLE = maximum likelihood estimation.

Parameter	Value
Box side length	5 px
Fit method	MLE
Gradient	7000
Baseline	100
Sensitivity	15.6
Gain	300
Quantum yield	0.93

Table 11. Filter parameters applied in custom code to remove noisy traces.

Parameter	Value
Threshold for peaks (in number of standard deviations)	8
Minimal peak width (in frames)	1
Minimal peak number	10
Difference in #frames between 1st and 2nd peak	1000
Last frame for a first peak to occur	1000
Minimal blinking time (earliest last peak)	500

Table 12. A list of features extracted for the correlation plots, PCA, and classical ML approaches.

Feature
Number of peaks
Last peak
Blinking time
Maximal peak height
Mean peak height
Std peak height
Total ON time (sum of peak widths)
Maximal ON time (peak width)
Mean ON time (peak width)
Std ON time (peak width)
Total OFF time
Maximal OFF time
Mean OFF time
Std OFF time
Maximal approximate peak area
Minimal approximate peak area
Mean approximate peak area
Std approximate peak area
Tempo[177]

Table 13. Parameters used for the classical ML models. If no specific numbers are mentioned the default settings were used.

Model type	Scikit-learn parameters used
Decision Tree	default
Random Forest	default
AdaBoost	default
KNN	n_neighbors=6
SVM (SVC)	cache_size=1000
MultiLayer Perceptron	alpha=0.05, learning_rate='adaptive', max_iter=1000

8.1.2 Chapter 3

8.1.2.1 Figures

8.1.2.2 Tables

Table 14–19 are reprinted from S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* **2023**, *145*, 1441–1447.[91] 🅭🅯

Table 14. Mean overall accuracies and standard deviations for the classification results on all 25 models of the five test sets of the five-fold CV (Figure 28).

Set	Mean overall accuracy	Mean loss
C11-C14	0.615450 (±0.020893)	1.306280 (±0.083175)
P15-P17	0.687844 (±0.021590)	1.108188 (±0.067480)
E19-E21	0.641567 (±0.055850)	1.055840 (±0.025304)

Table 15. Summary of the number of traces used for training and testing, as well as the required training times and evaluation time for MCD, rounded up to the minute.

Compound	Total # traces	# Traces for training	# Traces for testing	Time for training (without MCD)	Evaluation for MCD analysis
Total for charged set	11896	9517	2379	32 min (12 min)	4 min
C11	3572	2858	714		
C12	1756	1405	351		
C13	4970	3976	994		
C14	1598	1278	320		
Total for phosphorylation set	4162	3330	832	11 min (5 min)	4 min
P15	1176	941	235		
P16	1198	959	239		
P17	1788	1430	358		
Total for epimer set	31504	25204	6300	1 h 23 min (33 min)	4 min
E19	11226	8981	2245		
E20	11388	9111	2277		
E21	8890	7112	1778		

Table 16. Detailed model architecture 1D-CNN with all the parameters that were set manually. Parameters that are not mentioned were used in their default value as set in TensorFlow 2.7.0.

Layer	Layer settings	Param #
Input	input shape=(None, 6000, 2)	
Convolutional layer 1D 1	number of filters=64, kernel size=9, stride size=2, activation="relu"	1216
Batch Normalization	-	256
Dropout	dropout rate=0.3	0
Convolutional layer 1D 2	number of filters=64, kernel size=3, stride size=2, activation="relu"	12352
Batch Normalization	-	256
Dropout	dropout rate=0.5	0
Convolutional layer 1D 3	number of filters=64, kernel size=3, stride size=2, activation="relu"	12352
Flatten	-	0
Dense layer	units=number of classes, activation="Softmax"	143619 (for 3 classes)

Table 17. Detailed deterministic model architecture 1D-CNN-GRU with all the parameters that were set manually. Parameters that are not mentioned were used in their default value as set in TensorFlow 2.7.0.

Layer	Layer settings	Param #
Input	input shape=(None, 6000, 2)	
Convolutional layer 1D 1	number of filters=64, kernel size=9, stride size=2, activation="relu"	1216
Batch Normalization	-	256
Dropout	dropout rate=0.3	0
Convolutional layer 1D 2	number of filters=64, kernel size=3, stride size=2, activation="relu"	12352
Batch Normalization	-	256
Dropout	dropout rate=0.5	0
Convolutional layer 1D 3	number of filters=64, kernel size=3, stride size=2, activation="relu"	12352
Batch Normalization	-	256
Dropout	dropout rate=0.3	0
GRU layer	units=128	74496
GRU layer	units=256	296448
Dense layer	units=number of classes, activation="Softmax"	771 (for 3 classes)

Table 18. Accuracies for the classification results of the 1D-CNN-GRU-MCD models before and after filtering with a CT of 0.7 and the corresponding percentage of lost traces.

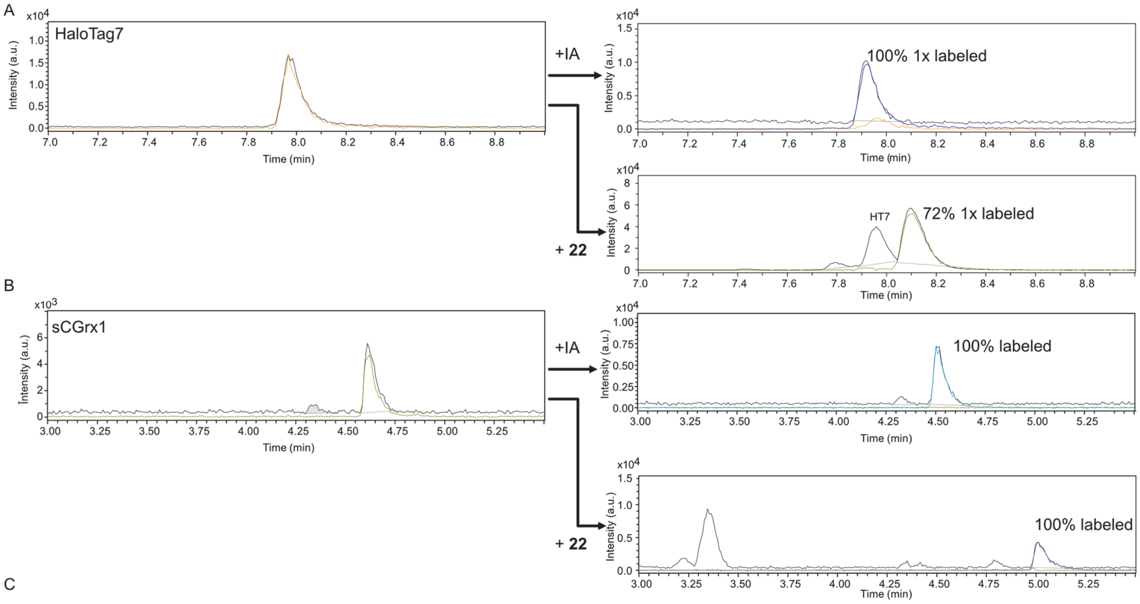
Set	Accuracy before filtering	Accuracy after filtering	Percentage of traces lost
C11-C14	0.732773	0.905629	49.2
P15-P17	0.726291	0.855234	46.1
E19-E21	0.568799	0.788026	90.2

Table 19. The model architecture including MCD in detail with all the parameters that were set manually. Parameters that are not mentioned were used in their default values as set in TensorFlow 2.7.0 and 2.10.0.

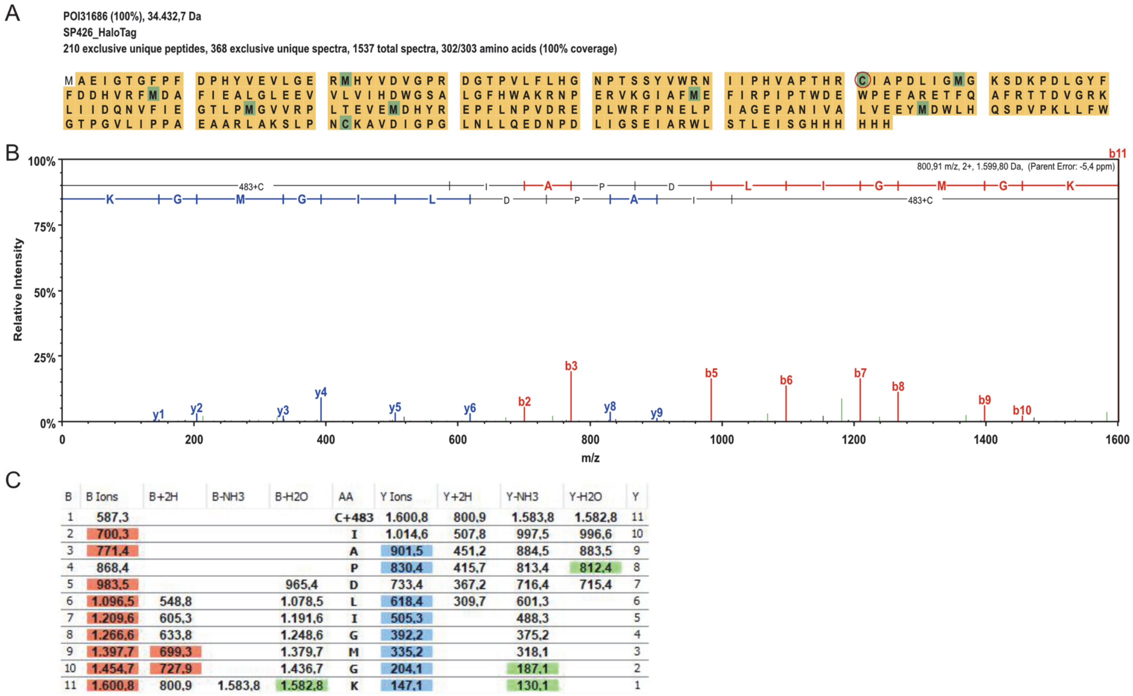
Layer	Layer settings
Input	input shape=(None, 6000, 2)
Convolutional layer 1D 1	number of filters=64, kernel size=9, stride size=2, activation="relu"
Batch Normalization	-
Dropout	dropout rate=0.1
Convolutional layer 1D 2	number of filters=64, kernel size=3, stride size=2, activation="relu"
Batch Normalization	-
Dropout	dropout rate=0.3
Convolutional layer 1D 3	number of filters=64, kernel size=3, stride size=2, activation="relu"
Batch Normalization	-
GRU layer	units=128 with Dropout=0.05
GRU layer	units=256 with Dropout=0.05
Dense layer	units=number of classes, activation="Softmax"

8.1.3 Chapter 4

8.1.3.1 Figures and Schemes

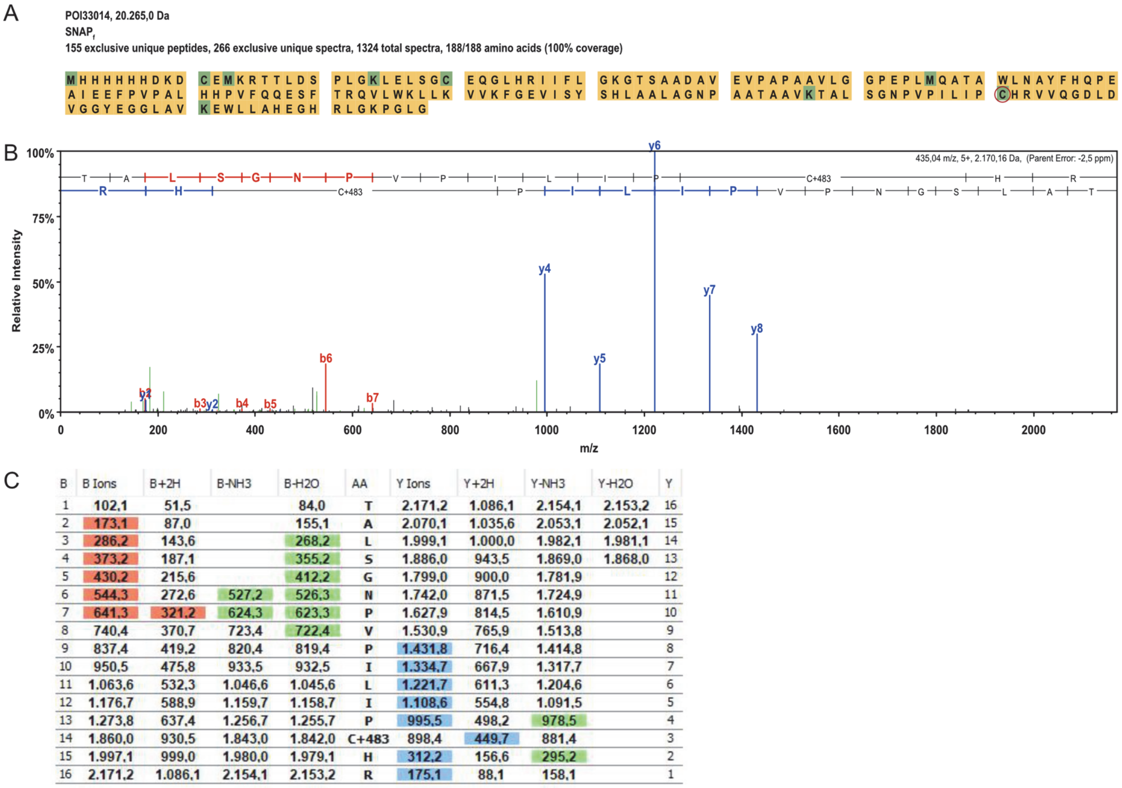


Scheme 23. Labeling reaction of HaloTag7 and sCGrx1p with IA and HMSiR-IA 22.



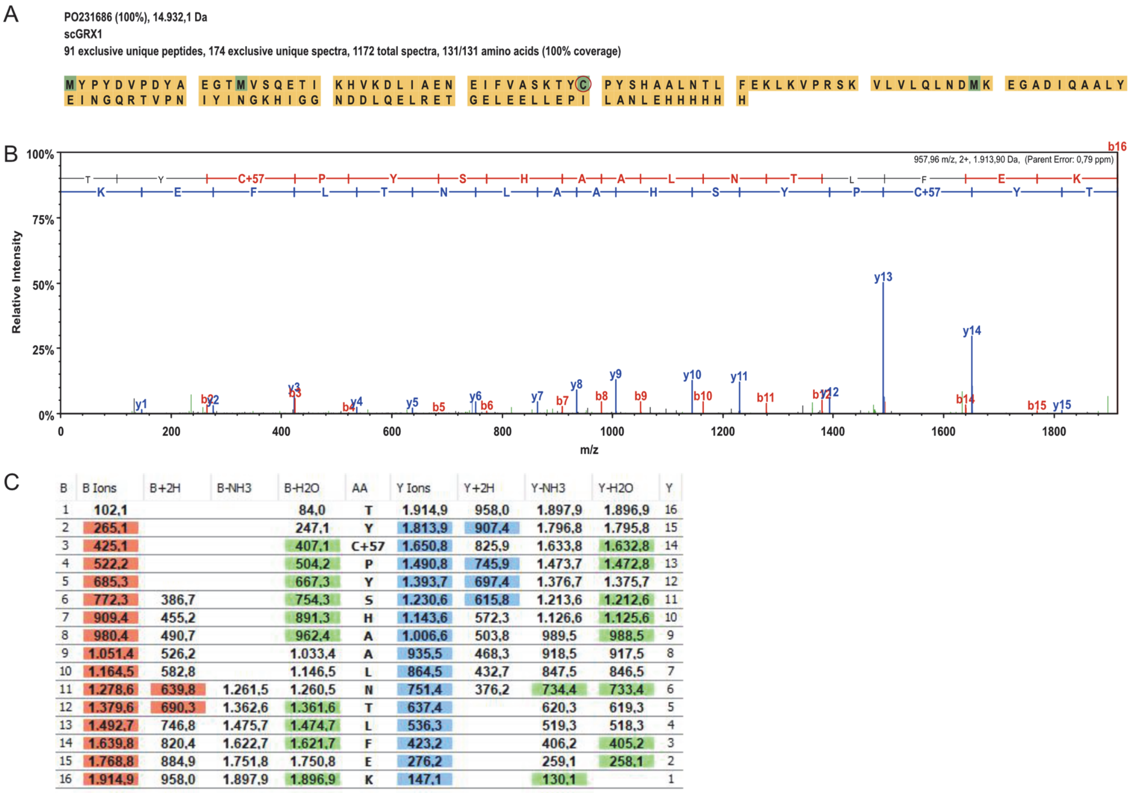
B	B Ions	B+2H	B-NH3	B-H2O	AA	Y Ions	Y+2H	Y-NH3	Y-H2O	Y
1	587,3				C+483	1.600,8	800,9	1.583,8	1.582,8	11
2	700,3				I	1.014,6	507,8	997,5	996,6	10
3	771,4				A	901,5	451,2	884,5	883,5	9
4	868,4				P	830,4	415,7	813,4	812,4	8
5	983,5			965,4	D	733,4	367,2	716,4	715,4	7
6	1.096,5	548,8		1.078,5	L	618,4	309,7	601,3		6
7	1.209,6	605,3		1.191,6	I	505,3		488,3		5
8	1.266,6	633,8		1.248,6	G	392,2		375,2		4
9	1.397,7	699,3		1.379,7	M	335,2		318,1		3
10	1.454,7	727,9		1.436,7	G	204,1		187,1		2
11	1.600,8	800,9	1.583,8	1.582,8	K	147,1		130,1		1

Figure 69. Analysis results of HTIA sample by FGCZ. A, Sequence with amino acids carrying a modification. The circled cysteine was found to be modified with the blinking fluorophore (corresponding to the + 483 mass) on three occasions whereas the second cysteine was only modified with carbamidomethyl. Methionine residues were partially oxidized. C, Fragmentation table of the spectrum shown in B, the colored numbers correspond to the m/z of the fragments found in B.



B	B Ions	B+2H	B-NH3	B-H2O	AA	Y Ions	Y+2H	Y-NH3	Y-H2O	Y
1	102,1	51,5		84,0	T	2.171,2	1.086,1	2.154,1	2.153,2	16
2	173,1	87,0		155,1	A	2.070,1	1.035,6	2.053,1	2.052,1	15
3	286,2	143,6		268,2	L	1.999,1	1.000,0	1.982,1	1.981,1	14
4	373,2	187,1		355,2	S	1.886,0	943,5	1.869,0	1.868,0	13
5	430,2	215,6		412,2	G	1.799,0	900,0	1.781,9		12
6	544,3	272,6	527,2	526,3	N	1.742,0	871,5	1.724,9		11
7	641,3	321,2	624,3	623,3	P	1.627,9	814,5	1.610,9		10
8	740,4	370,7	723,4	722,4	V	1.530,9	765,9	1.513,8		9
9	837,4	419,2	820,4	819,4	P	1.431,8	716,4	1.414,8		8
10	950,5	475,8	933,5	932,5	I	1.334,7	667,9	1.317,7		7
11	1.063,6	532,3	1.046,6	1.045,6	L	1.221,7	611,3	1.204,6		6
12	1.176,7	588,9	1.159,7	1.158,7	I	1.108,6	554,8	1.091,5		5
13	1.273,8	637,4	1.256,7	1.255,7	P	995,5	498,2	978,5		4
14	1.860,0	930,5	1.843,0	1.842,0	C+483	898,4	449,7	881,4		3
15	1.997,1	999,0	1.980,0	1.979,1	H	312,2	156,6	295,2		2
16	2.171,2	1.086,1	2.154,1	2.153,2	R	175,1	88,1	158,1		1

Figure 70. Analysis results of SNAP$_f$ sample by FGCZ. A, Sequence with amino acids carrying a modification. The circled cysteine was found to be modified with the blinking fluorophore (corresponding to the + 483 mass) in 88 spectra whereas the first cysteine in the sequence was only modified with carbamidomethyl. The second cysteine and the marked lysine were found to be modified with the fluorophore in one spectrum. Methionine residues were partially oxidized. B, Fragmentation spectrum of an example peptide containing the cysteine modified with the spontaneously blinking fluorophore. C, Fragmentation table of the spectrum shown in B, the colored numbers correspond to the m/z of the fragments found in B.



B	B Ions	B+2H	B-NH3	B-H2O	AA	Y Ions	Y+2H	Y-NH3	Y-H2O	Y
1	102,1			84,0	T	1.914,9	958,0	1.897,9	1.896,9	16
2	265,1			247,1	Y	1.813,9	907,4	1.796,8	1.795,8	15
3	425,1			407,1	C+57	1.650,8	825,9	1.633,8	1.632,8	14
4	522,2			504,2	P	1.490,8	745,9	1.473,7	1.472,8	13
5	685,3			667,3	Y	1.393,7	697,4	1.376,7	1.375,7	12
6	772,3	386,7		754,3	S	1.230,6	615,8	1.213,6	1.212,6	11
7	909,4	455,2		891,3	H	1.143,6	572,3	1.126,6	1.125,6	10
8	980,4	490,7		962,4	A	1.006,6	503,8	989,5	988,5	9
9	1.051,4	526,2		1.033,4	A	935,5	468,3	918,5	917,5	8
10	1.164,5	582,8		1.146,5	L	864,5	432,7	847,5	846,5	7
11	1.278,6	639,8	1.261,5	1.260,5	N	751,4	376,2	734,4	733,4	6
12	1.379,6	690,3	1.362,6	1.361,6	T	637,4		620,3	619,3	5
13	1.492,7	746,8	1.475,7	1.474,7	L	536,3		519,3	518,3	4
14	1.639,8	820,4	1.622,7	1.621,7	F	423,2		406,2	405,2	3
15	1.768,8	884,9	1.751,8	1.750,8	E	276,2		259,1	258,1	2
16	1.914,9	958,0	1.897,9	1.896,9	K	147,1		130,1		1

Figure 71. Analysis results of sCGrx1p sample by FGCZ. A, Sequence with amino acids carrying a modification. The circled cysteine was found to be modified with the blinking fluorophore (corresponding to the + 483 mass) in 84 spectra. Methionine residues were partially oxidized. B, Fragmentation spectrum of an example peptide containing the cysteine modified with the spontaneously blinking fluorophore. C, Fragmentation table of the spectrum shown in B, the colored numbers correspond to the m/z of the fragments found in B.

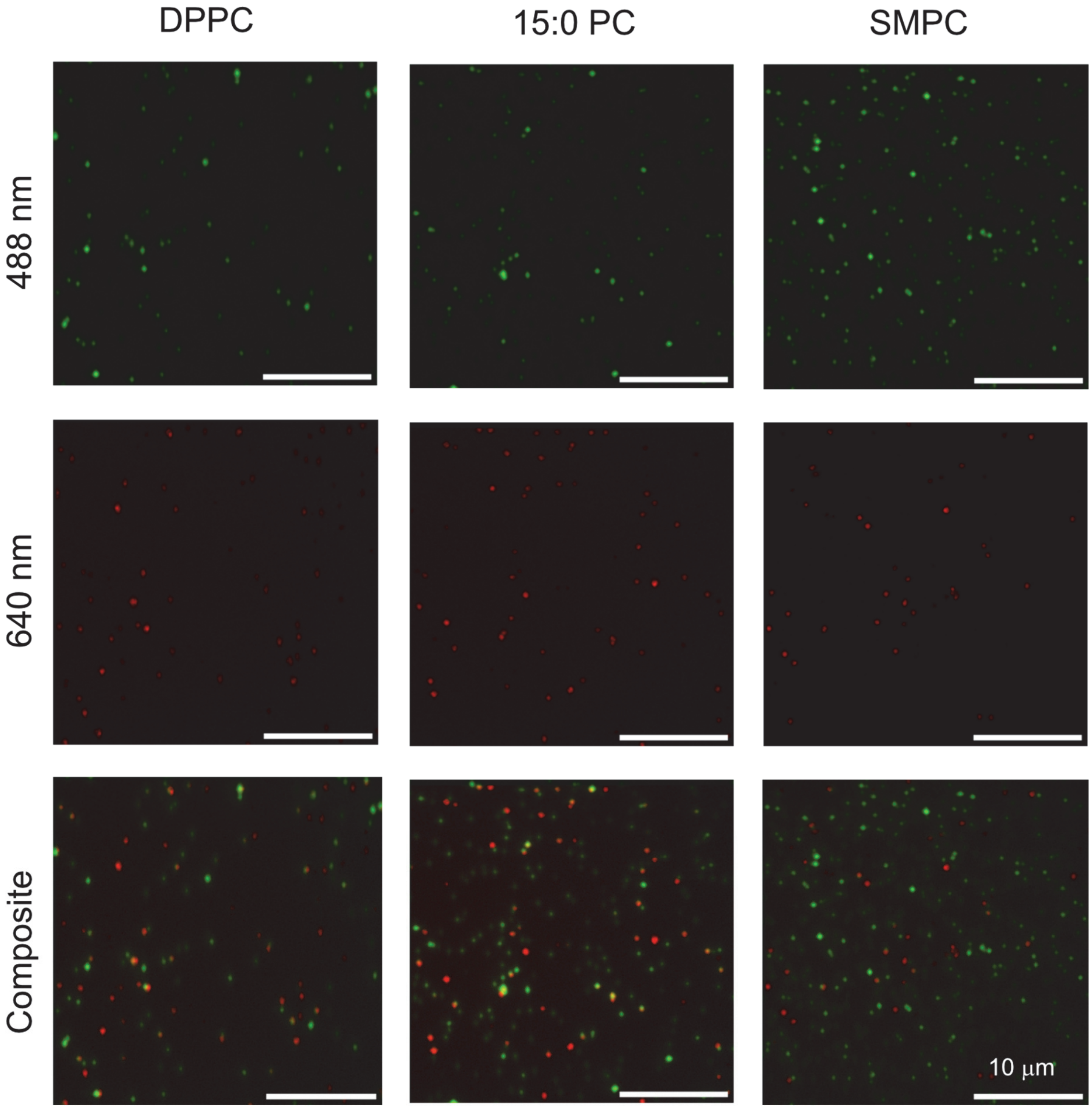


Figure 72. Maximum projections of liposomes with different base lipid DPPC, 15:0 PC, or SMPC loaded with HMSiR-Halo labeled HaloTag7. The lipids were measured via fluorescence of the NBD dye (10 frames, 488 nm, 30 ms, 2 mW) attached to a lipid and the loading via the signal of HMSiR (10000 frames, 638 nm, 30 ms, 90 mW). The images have different contrasts for the same channels.

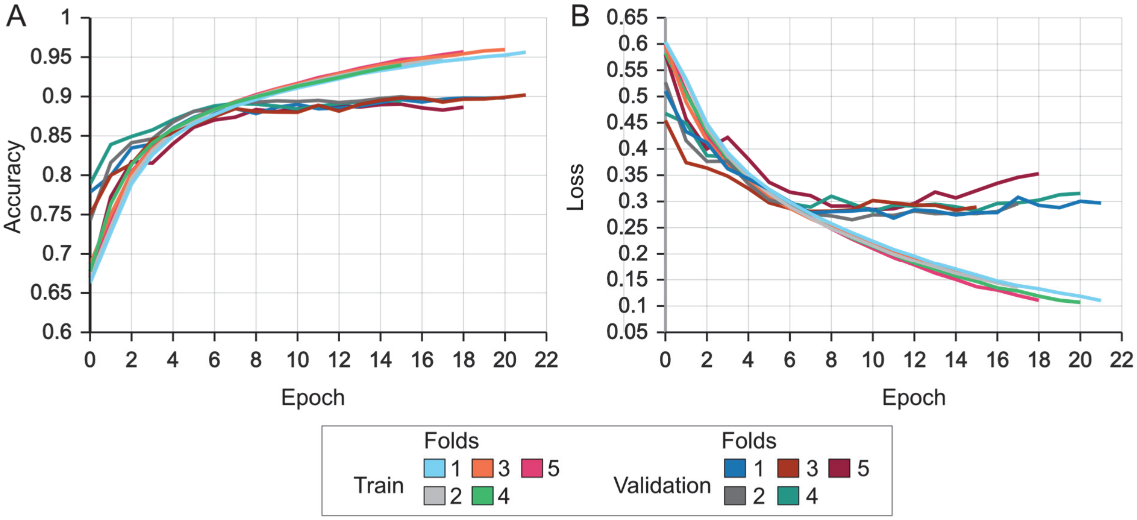


Figure 73. Accuracy and loss curves during training of the classification problem HTHTL vs. HTIA using 4000 frames. Training was stopped using early stopping with a patience of 10 epochs monitoring the validation loss. Visualization was generated with TensorBoard (smoothing = 0.6).[210]

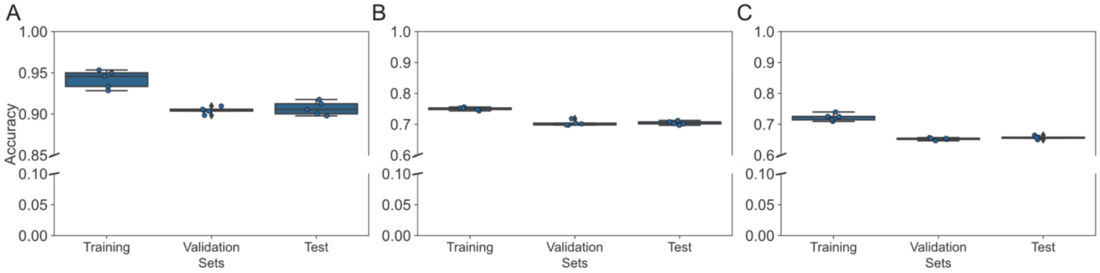


Figure 74. CV results of the protein classification using 4000 frames. A, HTHTL and HTIA. B, HTHTL and SNAP$_f$, HTHTL, SNAP$_f$, and sCGrx1p.

8.1.3.2 Tables

Table 20. Proteins used in protein labeling and classification experiments. The protein name, sequence, molecular weight, and origin of the purified protein are given. Masses were calculated using reference.[310] The labeled cysteine is marked in red.

Protein	Protein sequence	Calculated molecular weight (Da)	Origin
HaloTag7 (HaloTag7-6His)	AEIGTGFPFDPHYVEVLGERMHYVDVGPRDGTPVLFLHGNPTSSYVWRNIIPHVAPTHRCIAPDLIGMGKSDKPDLGYFFDDHVRFMDAFIEALGLEEVVLVIHDWGSALGFHWAKRNPERVKGIAFMEFIRPIPTWDEWPEFARETFQAFRTTDVGRKLIIDQNVFIEGTLPMGVVRPLTEVEMDHYREPFLNPVDREPLWRFPNELPIAGEPANIVALVEEYMDWLHQSPVPKLLFWGTPGVLIPPAEAARLAKSLPNCKAVDIGPGLNLLQEDNPDLIGSEIARWLSTLEISGHHHHHH	34301.21	Recombinant expression Initial samples: Courtesy of Henriette Lämmermann and Sarah Emmert
sCGrx1p (HA-sCGrx1p-6His)	MYPYDVPDYAEGTMVSQETIKHVKDLIAENEIFVASKTYCPYSHAALNTLFEKLKVPRSKVLVLQLNDMKEGADIQAALYEINGQRTVPNIYINGKHIGGNDDLQELRETGELEELLEPILANLEHHHHHH	14931.87	Courtesy of Carla Miró Vinyals
SNAP$_f$-tag (6His-SNAP$_f$-tag)	MHHHHHHDKDCEMKRTTLDSPLGKLELSGCEQGLHRIIFLGKGTSAADAVEVPAPAAVLGGPEPLMQATAWLNAYFHQPEAIEEFPVPALHHPVFQQESFTRQVLWKLLKVVKFGEVISYSHLAALAGNPAATAAVKTALSGNPVPILIPCHRVVQGDLDVGGYEGGLAVKEWLLAHEGHRLGKPGLG	20265.32	Courtesy of Annabell Martin

Table 21. Settings applied in the single-molecule localization software Picasso.[175]

Parameter	Value
Box side length	5 px
Fit method	MLE
Gradient	16000
Baseline	100
Sensitivity	15.6
Gain	300
Quantum yield	0.93

Table 22. Number of traces obtained for the protein classifications.

Compound	Total # traces	# Traces for training	# Traces for testing
HTHTL	18178	14542	3636
HTIA	8830	7064	1766
SNAP	15150	12120	3030
sCGrx1p	7060	5648	1412

Table 23. Accuracies for the classification results of the 1D-CNN-GRU-MCD models before and after filtering with a CT of 0.7 and the corresponding percentage of lost traces for the protein classification using 4000 frames.

Set	Accuracy before filtering	Accuracy after filtering	Percentage of traces lost
HTHTL vs HTIA	0.884857	0.915888	18.8
HTHTL vs SNAP	0.659316	0.927032	60.3
HTHTL vs SNAP vs sCGrx1p	0.646942	0.892302	65.7

8.1.4 Chapter 5

8.1.4.1 Figures and Schemes

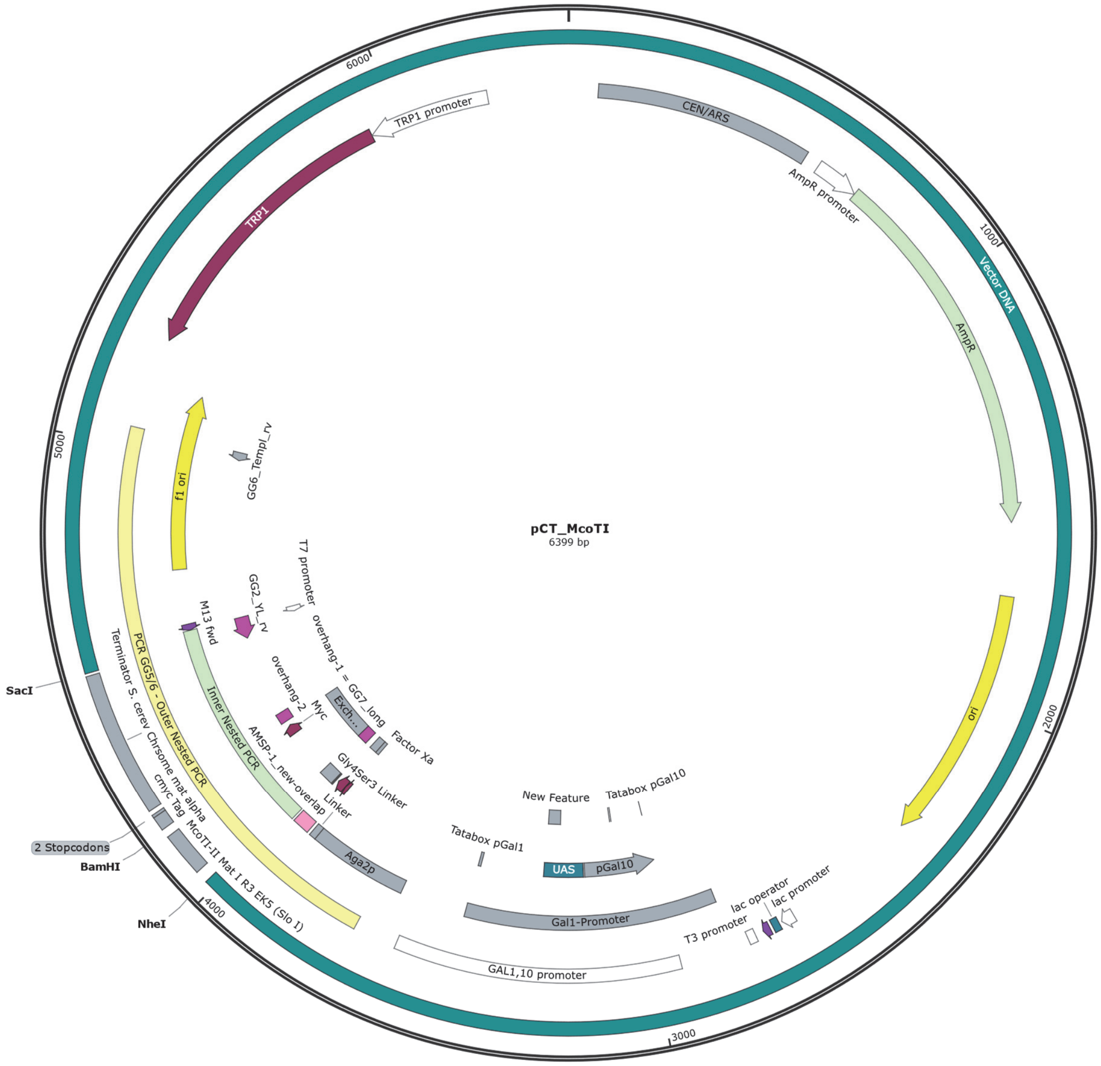


Figure 75. Plasmid map of pCT_McoTI.

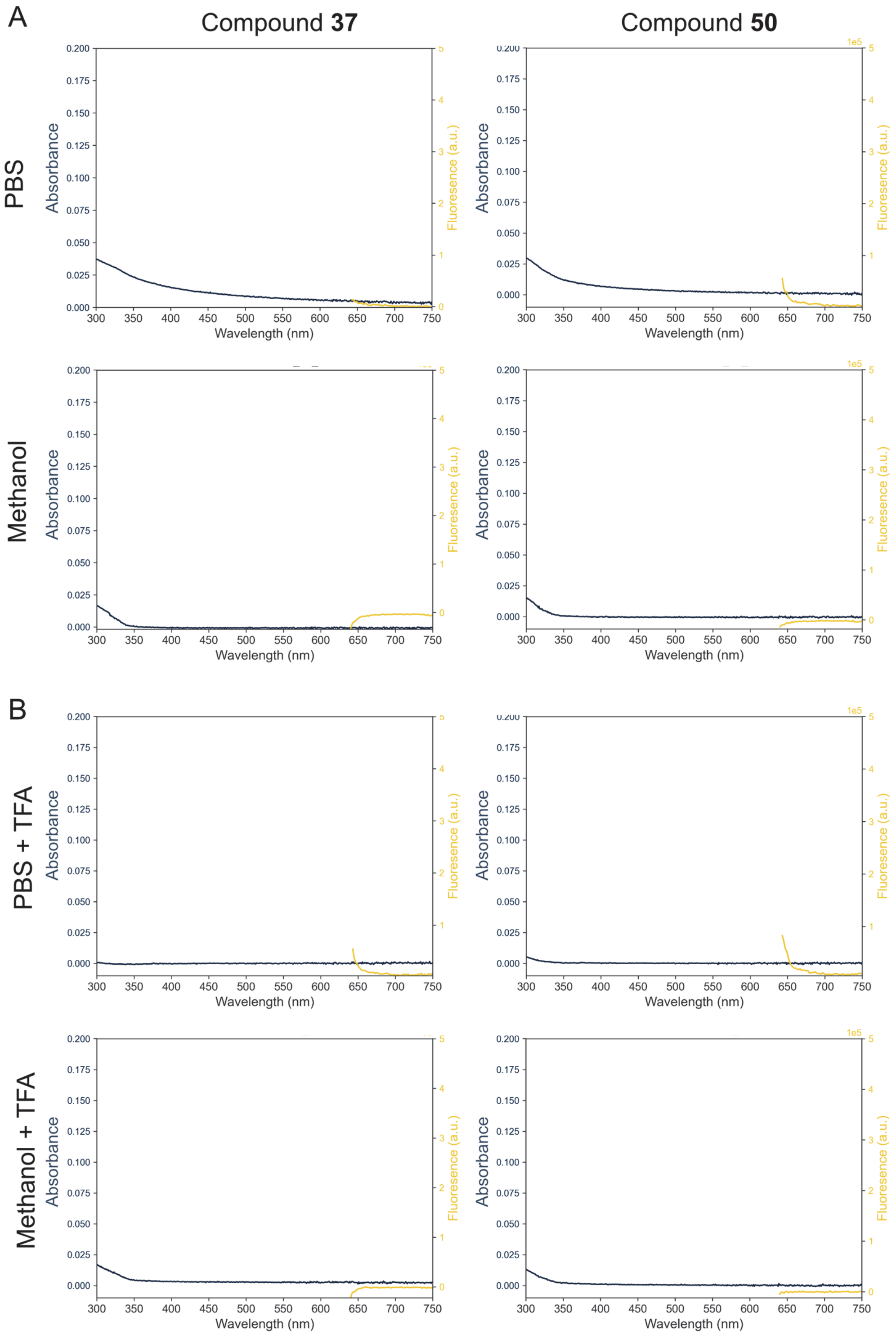


Figure 76. Absorbance and fluorescence spectra of compounds **37** and **50**. A, Spectra were measured in PBS (top row) and methanol (bottom row). B, Spectra were measured in PBS (top row) and methanol (bottom row) with 0.1% TFA in each. The fluorescence spectra were recorded while exciting the dyes with 633 nm light.

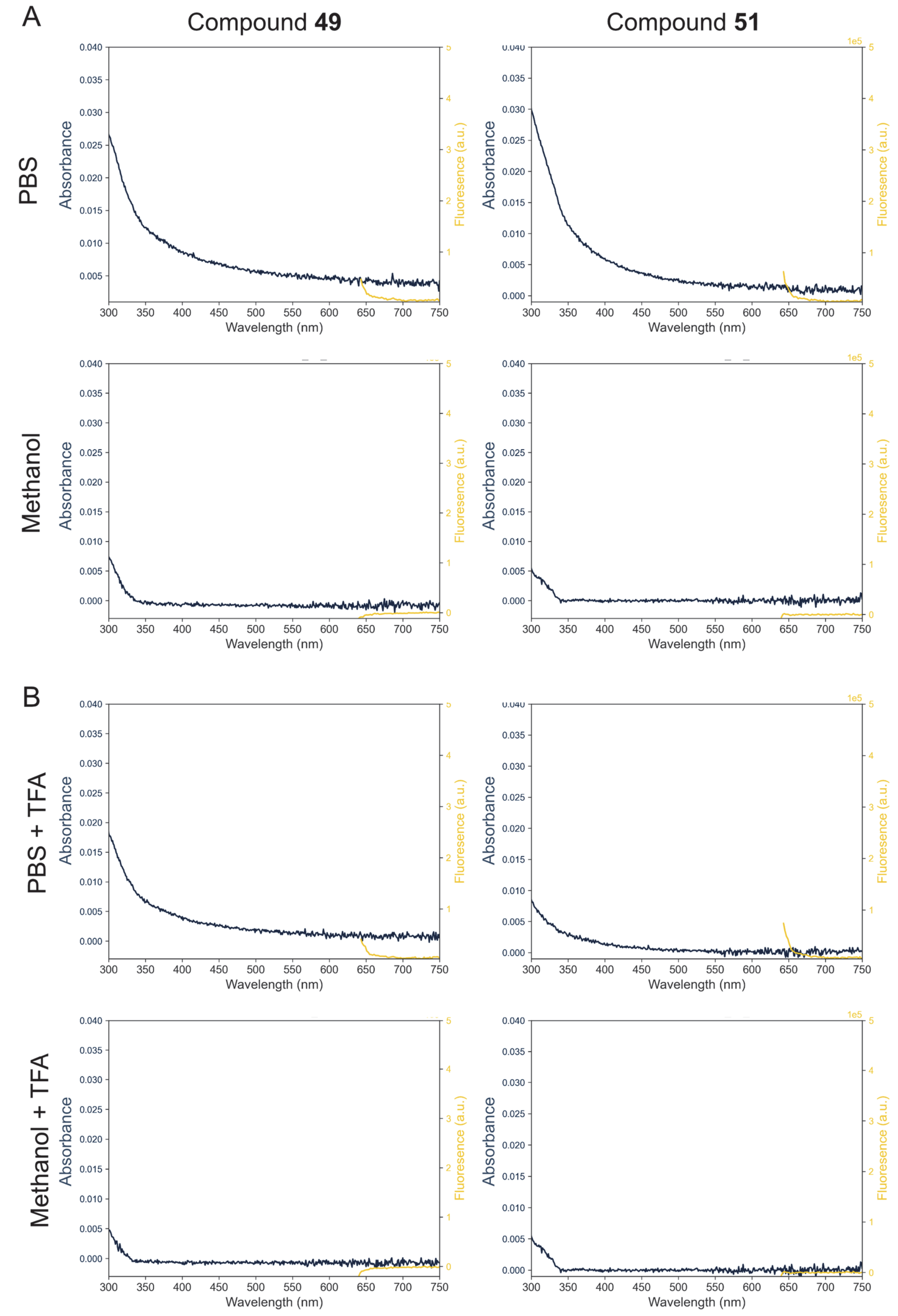


Figure 77. Absorbance and fluorescence spectra of compounds 49, and 51. A, Spectra were measured in PBS (top row) and methanol (bottom row). B, Spectra were measured in PBS (top row) and methanol (bottom row) with 0.1% TFA in each. The fluorescence spectra were recorded while exciting the dyes with 633 nm light.

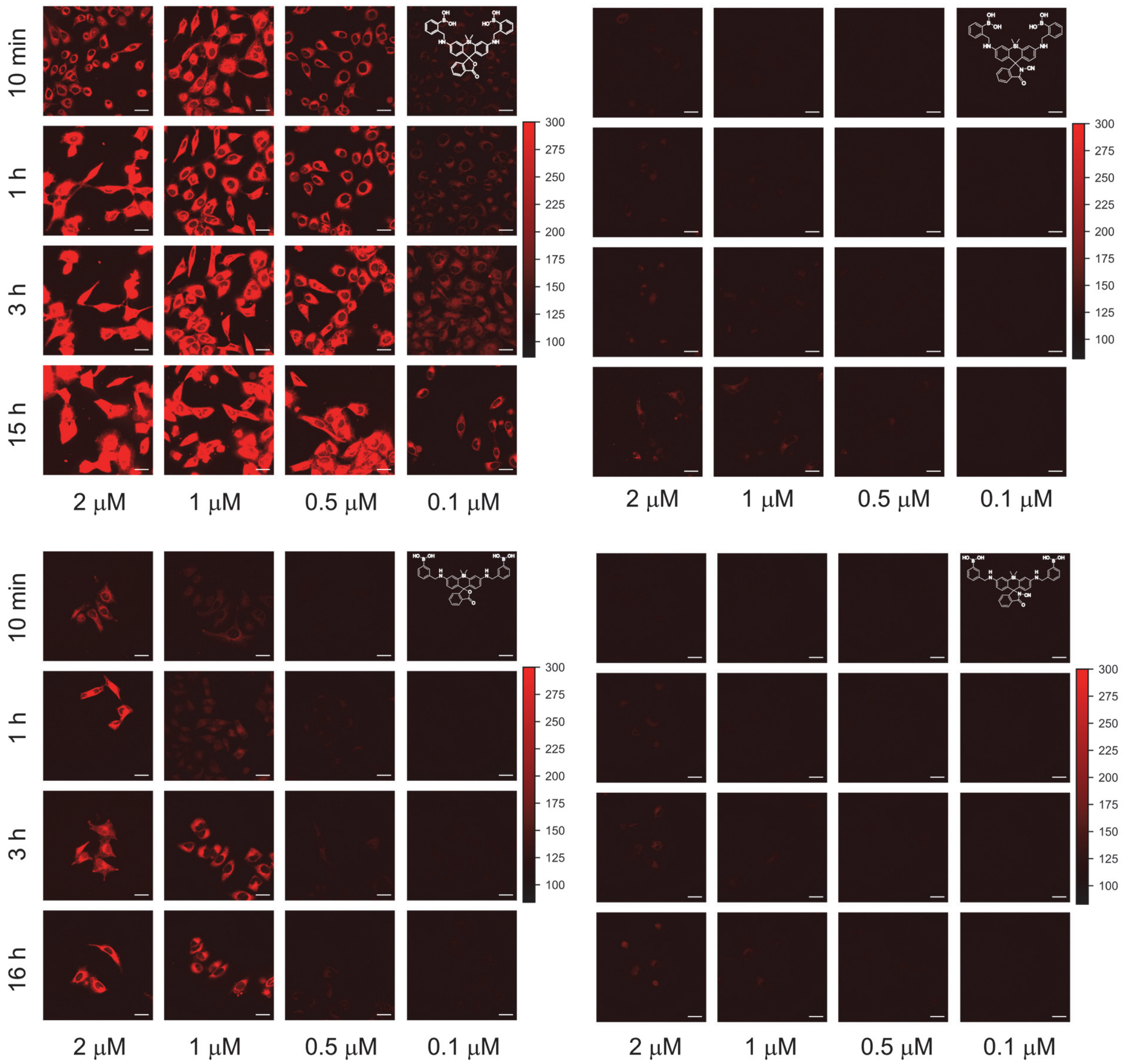


Figure 78. Imaging (638 nm, 100 ms 30 mW) of HeLa cells using compounds **36** (top left), **48** (bottom left), **37** (top right), **50** (bottom right). The cells were incubated with the corresponding concentration of compounds for the indicated amount of time, without any washing steps. Scale bar 20 μm. The color bar represents the fluorescence intensity (a.u.).

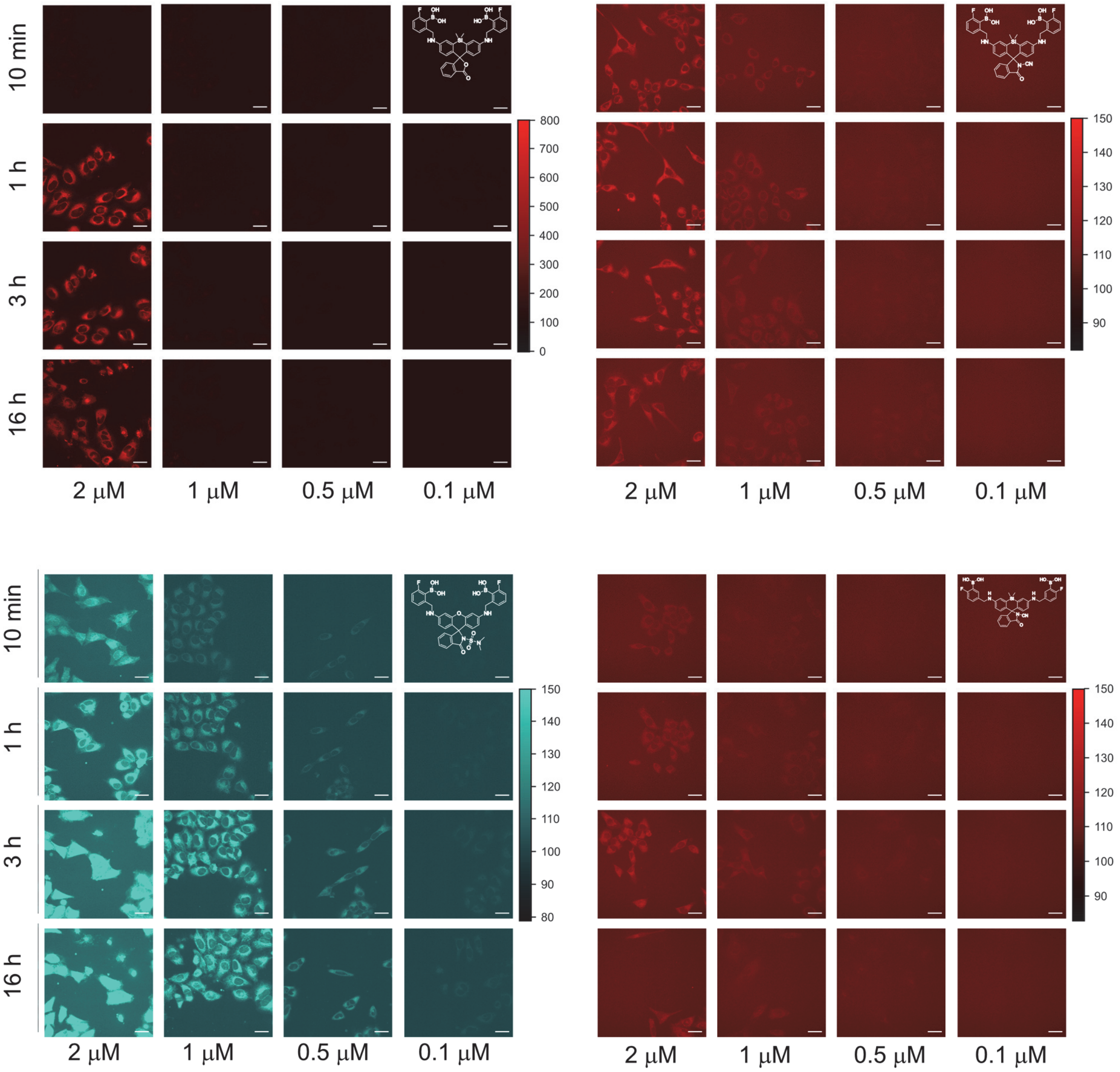


Figure 79. Imaging of HeLa cells using the fluorinated compounds 47 (top left), 41 (bottom left), 49 (top right), 51 (bottom right). Compounds 46, 49, and 50 were imaged using a 638 nm laser (100 ms, 30 mW) and compound 41 a 515 nm laser (100 ms, 30 mW). The cells were incubated with the corresponding amount of compound for the indicated amount of time. Scale bar 20 µm. The color bar represents the fluorescence intensity (a.u.).

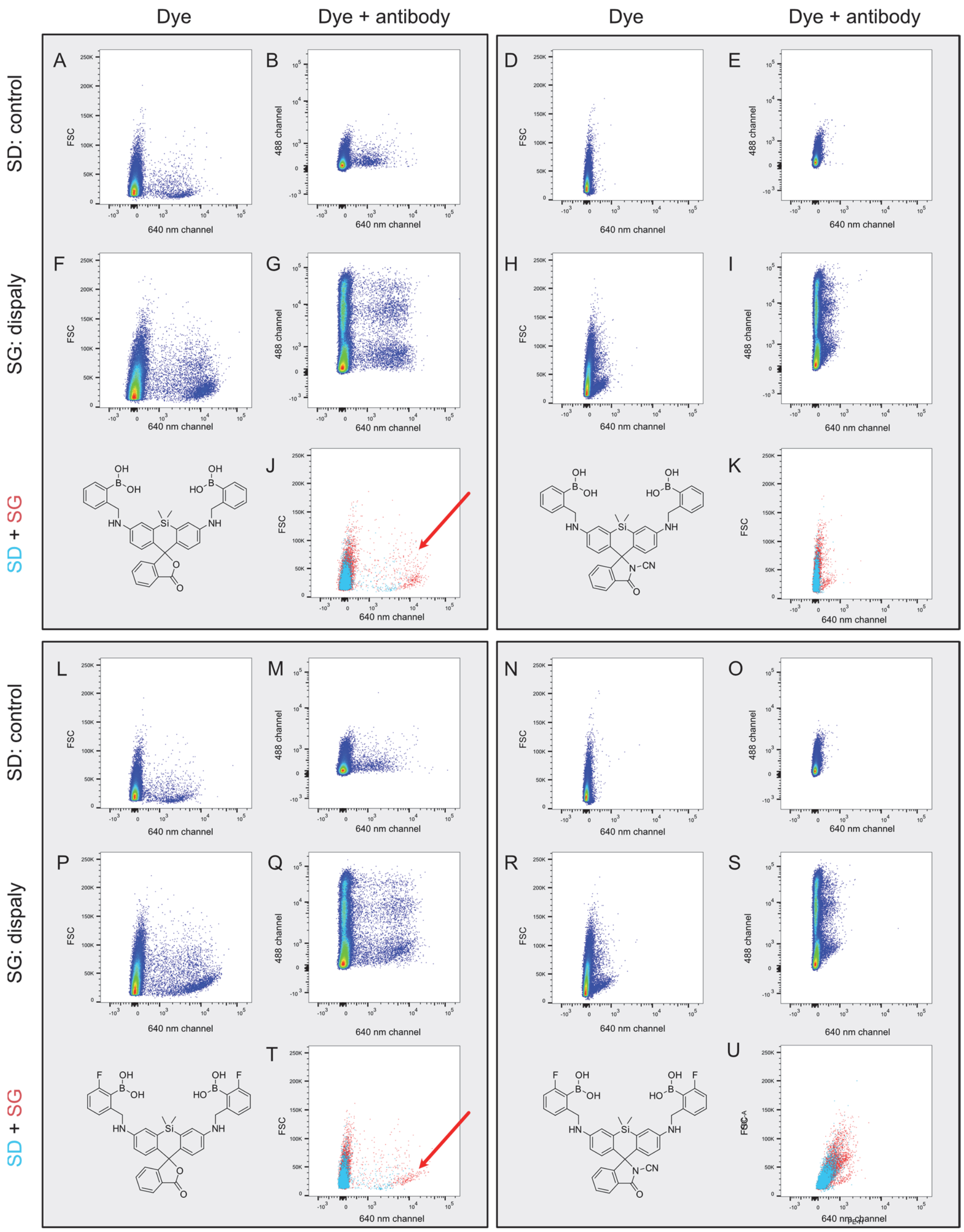


Figure 80. Panel of the flow cytometry analysis of the 2aa-library stained with the depicted dye, i.e., **36**, **37**, **47**, and **49** (1 µM) and the anti-HA antibody in the 488 nm channel. A-E, L-O, yeast cells were grown in SD medium before staining, hence not induced. F-I, P-S, yeast cells were grown in SG medium before staining, hence induced. J, K, T, U, Overlay of the induced and non-induced yeast cells after staining with the dye. The blue population is from the SD control, whereas the red is from the displaying cells. Red arrows point at potentially interesting populations with higher fluorescence intensity than the control. 10^5 cells were measured per sample.

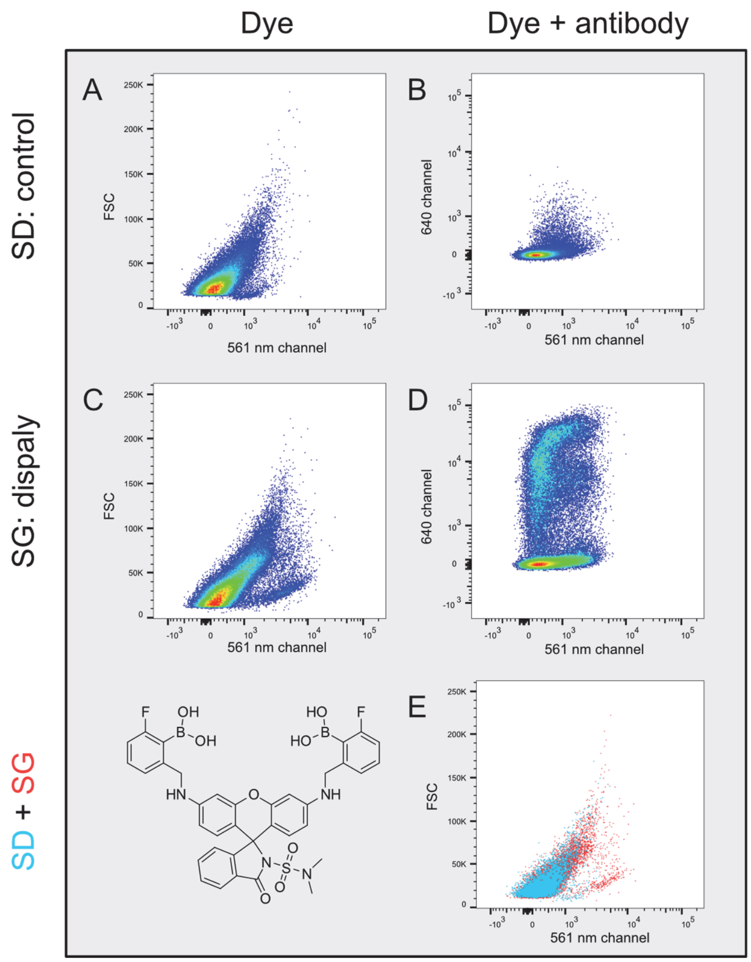


Figure 81. Panel of the flow cytometry analysis of the 2aa-library stained with compound **41** (1 mM) and the anti-HA antibody (secondary antibody in this case) in the 640 nm channel. A, B, yeast cells grown in SD medium before staining, hence not induced. C, D, yeast cells grown in SG medium before staining, hence induced. E, Overlay of the induced and non-induced yeast cells after staining with dye. The blue population is from the SD control, whereas the red is from the displaying cells. $1 \cdot 10^5$ cells were measured per sample

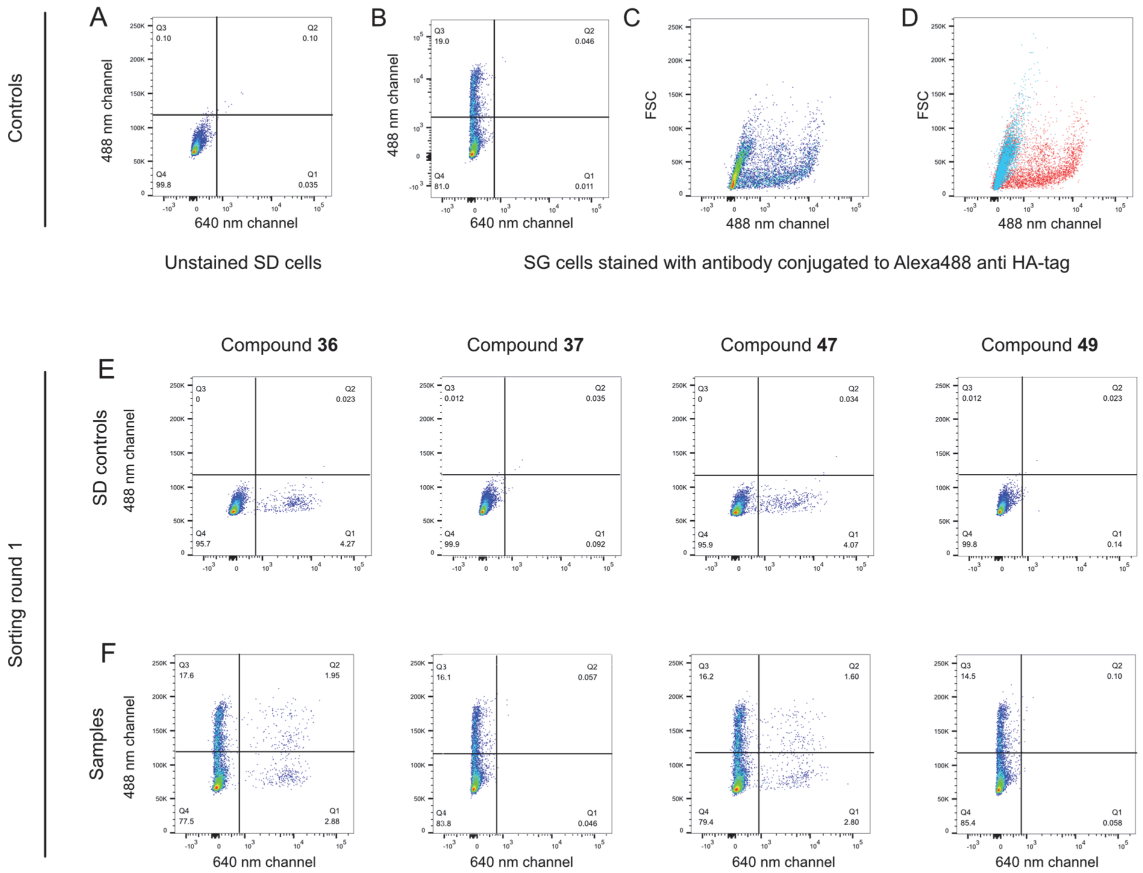


Figure 82. FACS of the 2aa-library screening for peptides binding compounds 36, 37, 47, and 49. A, unstained SD-grown cells. B and C, SG-grown cells that were stained with the anti-HA antibody in the 488 channel, confirmed yeast display. D, Overlay of induced and non-induced cells in red and blue respectively. E, Control cells that were grown in SD and stained with the corresponding dye. F, Samples grown in SG and stained with the dyes and anti-HA antibody labeled with AlexaFluor488. Cells in the Q2 quartile were collected. 5·10^5 cells were collected per sample.

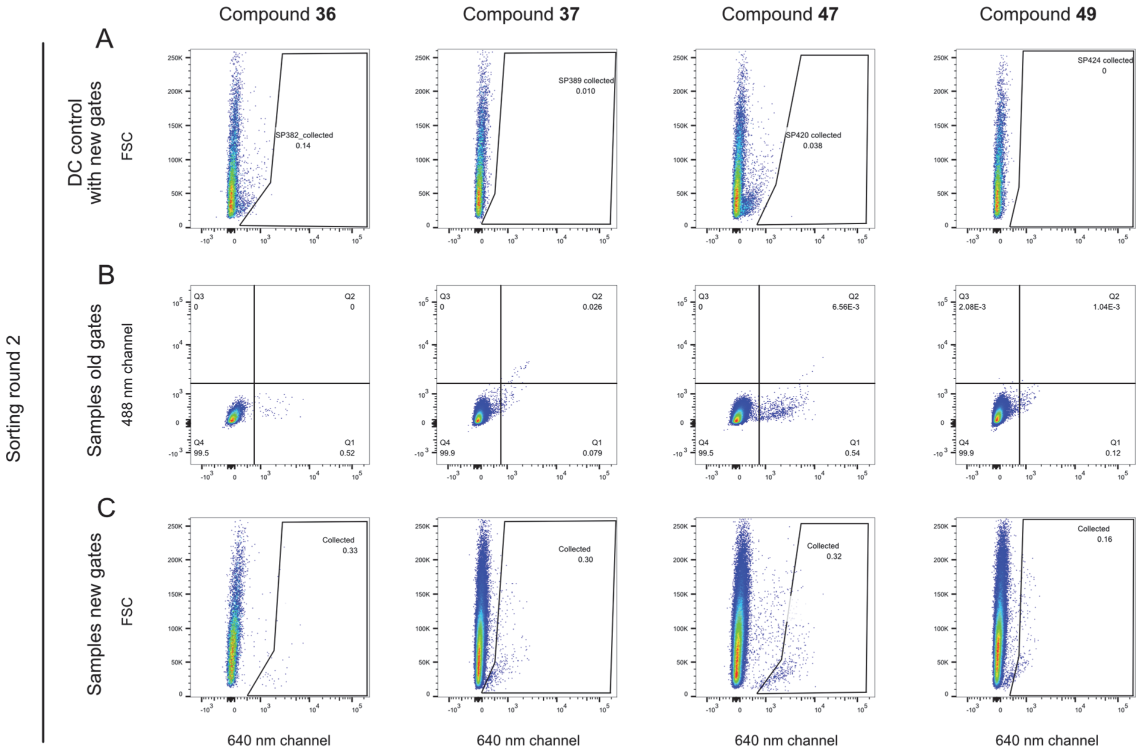


Figure 83. The second round of sorting of the 2aa-library with an adapted strategy avoiding the use of the HA antibody. A, DC cells grown in SG medium and stained with the corresponding dye. These samples are the basis for setting the gates for sorting. B, Cells collected in round one regrown and induced were labeled with the corresponding dye and are depicted with the old quartiles. B, Cells collected in round one regrown and induced were labeled with the corresponding dye and the new collections gates.

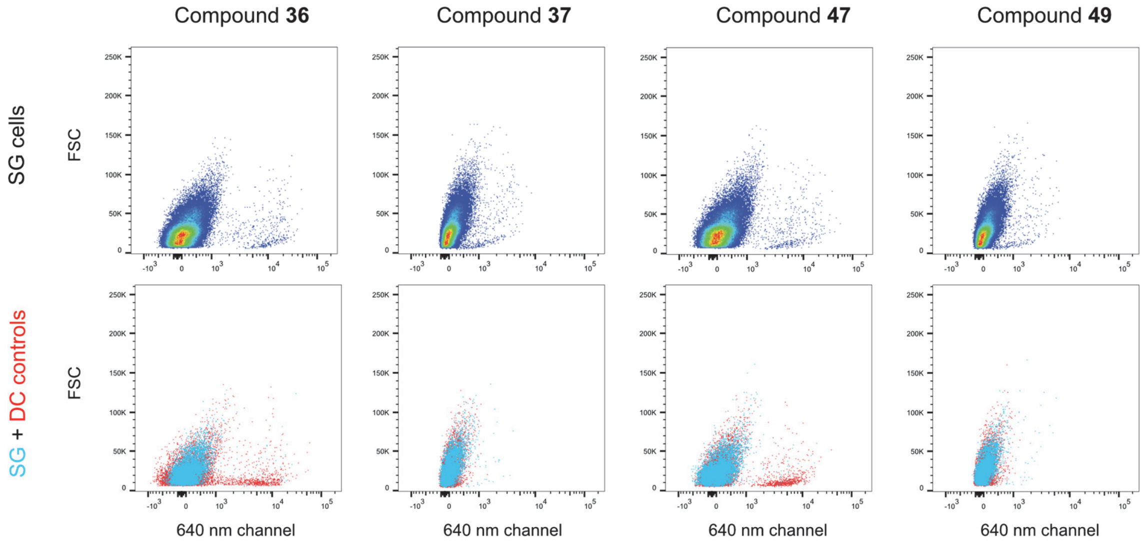


Figure 84. Flow cytometry analysis of the second round of FACS. The first frow contains the induced samples stained with the corresponding dye. In the second row the DC control cells, in red, are overlaid with the sample stained with the dye

8.1.4.2 Tables

Table 24. Primers for the generation of the 2aa and 8aa yeast display library.

Step	Primer name	Forward primer	Direction
Insert PCR	2aa-library or 8aa-library	5'-GGT TCT GGT GGT GGT GGT TCT GGT GGT GGT GGT TCT TCT X11X12X13 X21X22X23 (X31X32X33 X41X42X43 X51X52X53 X61X62X63 X71X72X73 X81X82X83) TCT TCT TGA TAA CAA CAG TGT AGA TGT AAC AAA ATC G-3'	Forward
Insert PCR and upscale	GG2_YL_rev	5'-GAA TTG TAA TAC GAC TCA CTA TAG GGC GAA TTG GAG CTC AAT TCT CTT AGG ATT CGA TTC ACA TTC-3'	Reverse
Insert upscale	GG7_long	5'-GGT TCT GGT GGT GGT GGT TCT GGT GGT GGT GGT TCT -3'	Forward
Insert upscale	GG2_YL_rev	5'-GAA TTG TAA TAC GAC TCA CTA TAG GGC GAA TTG GAG CTC AAT TCT CTT AGG ATT CGA TTC ACA TTC-3'	Reverse
Colony PCR	GG28	5'-TAG ATA CCC ATA CGA CGT TCC AGA CTA CG-3	Forward
Colony PCR	GG29	5'-CAG TGG GAA CAA AGT CGA TTT TGT TAC ATC-3	Reverse
NGS	NGS-GG28-fwd	5'-CTT TCC CTA CAC GAC GCT CTT CCG ATC TTA GAT ACC CAT ACG ACG TTC CAG ACT ACG-3'	Forward
NGS	NGS-GG29-rev	5'-GGA GTT CAG ACG TGT GCT CTT CCG ATC TCA GTG GGA ACA AAG TCG ATT TTG TTA CAT C-3'	Reverse

8.2 NMR spectra and characterization

8.2.1 Small molecule characterization

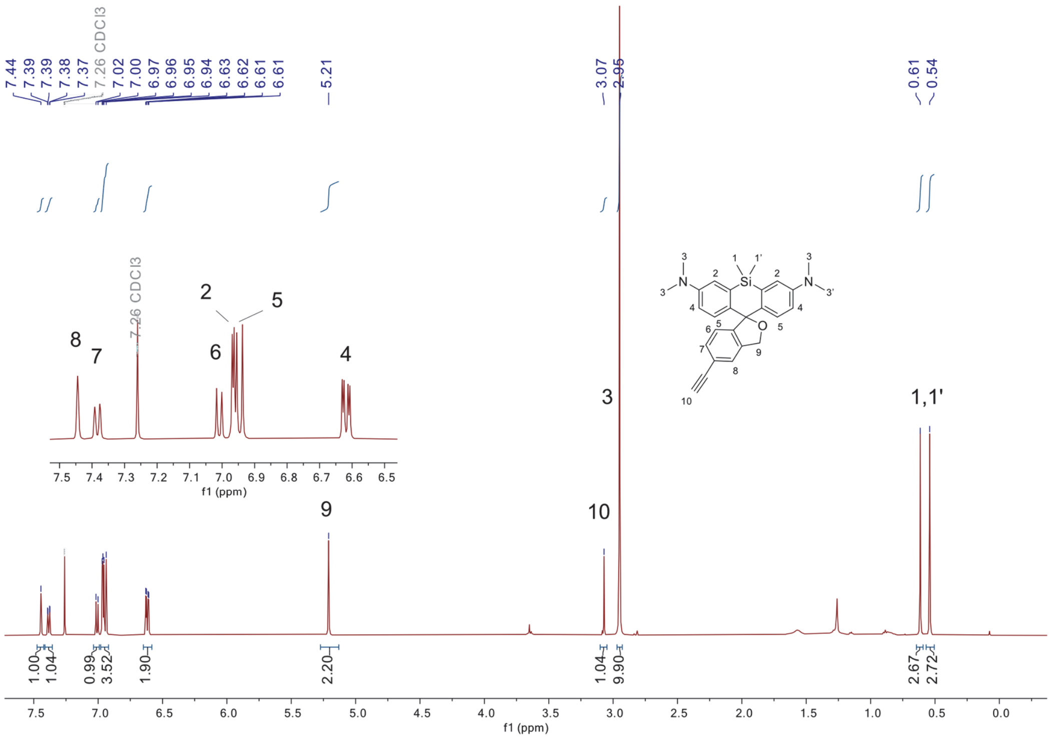


Figure 85. ^{1}H NMR (CDCl$_3$, 500 MHz) spectrum of compound 1.

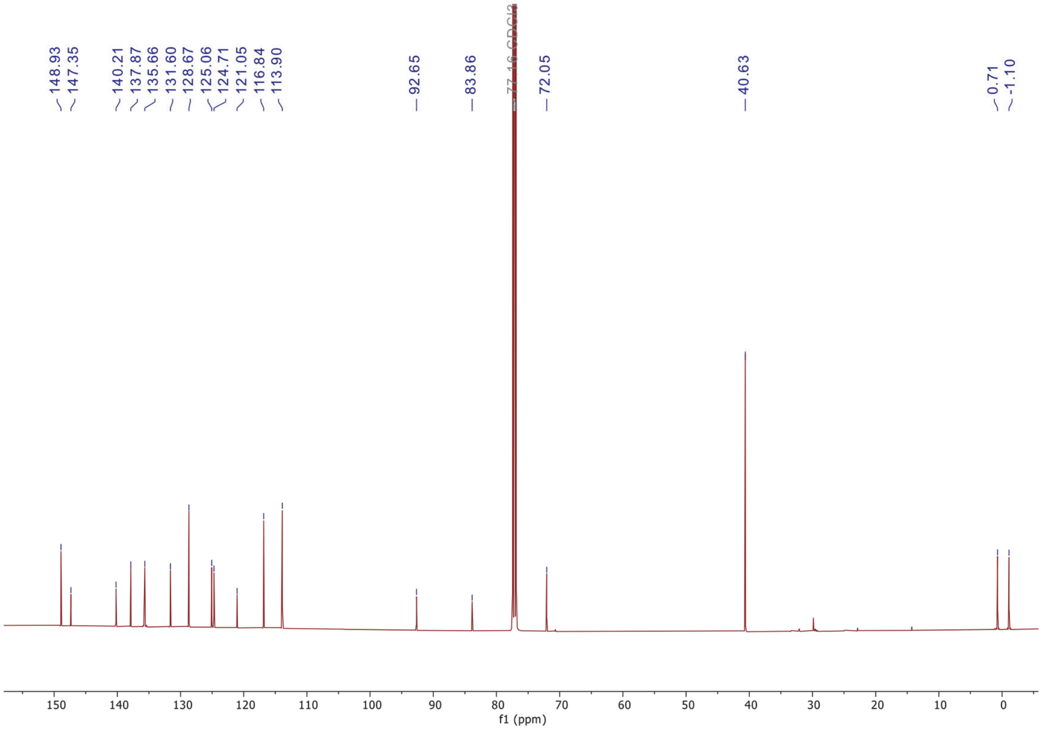


Figure 86. ^{13}C NMR (CDCl$_3$, 126 MHz) spectrum of compound 1.

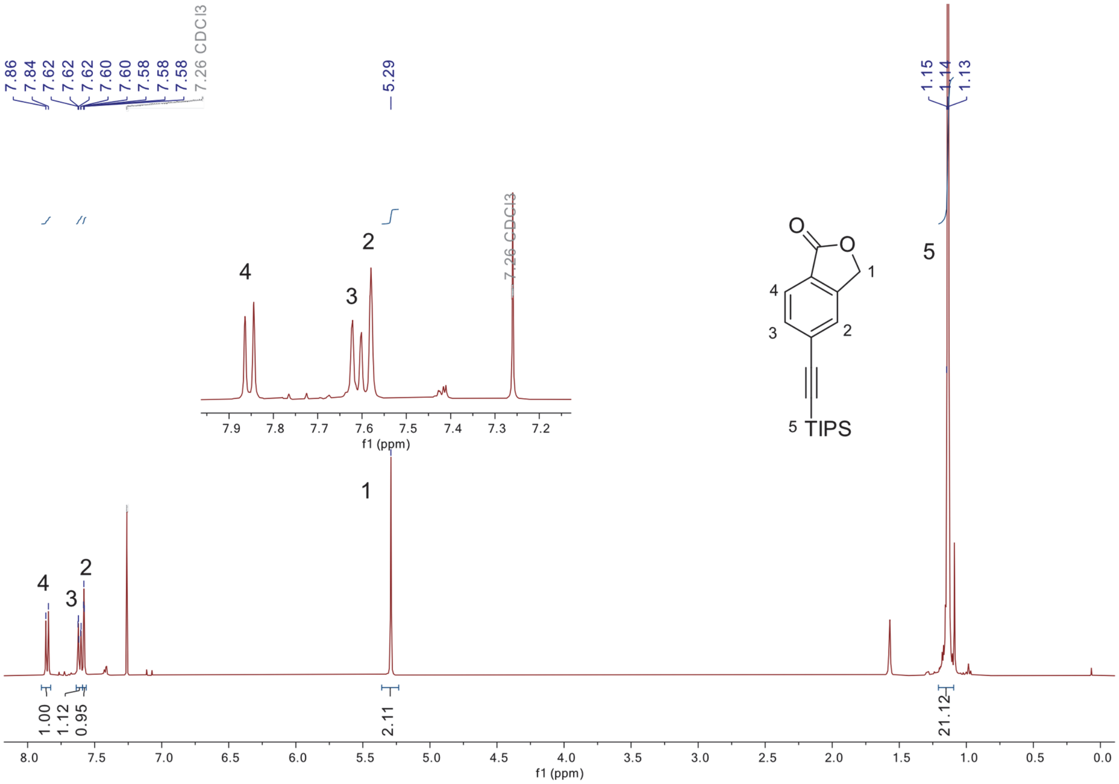


Figure 87. ^{1}H NMR ($CDCl_3$, 400 MHz) spectrum of compound **3.**

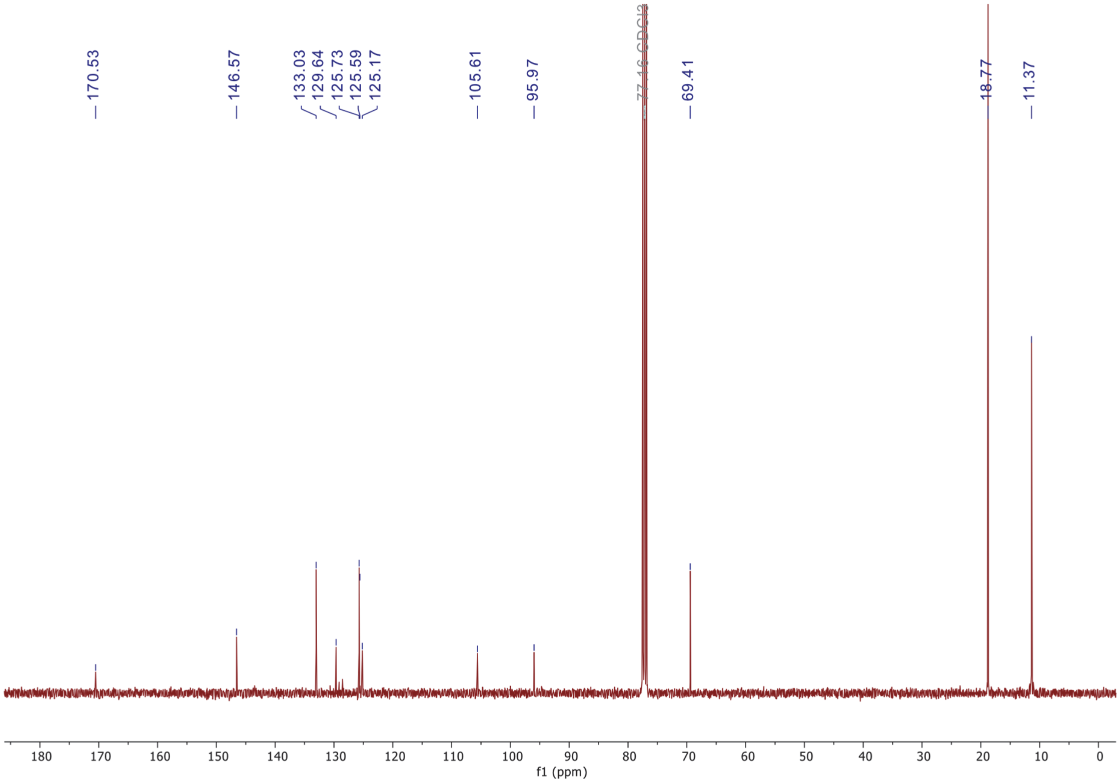


Figure 88. ^{13}C NMR ($CDCl_3$, 101 MHz) spectrum of compound **3.**

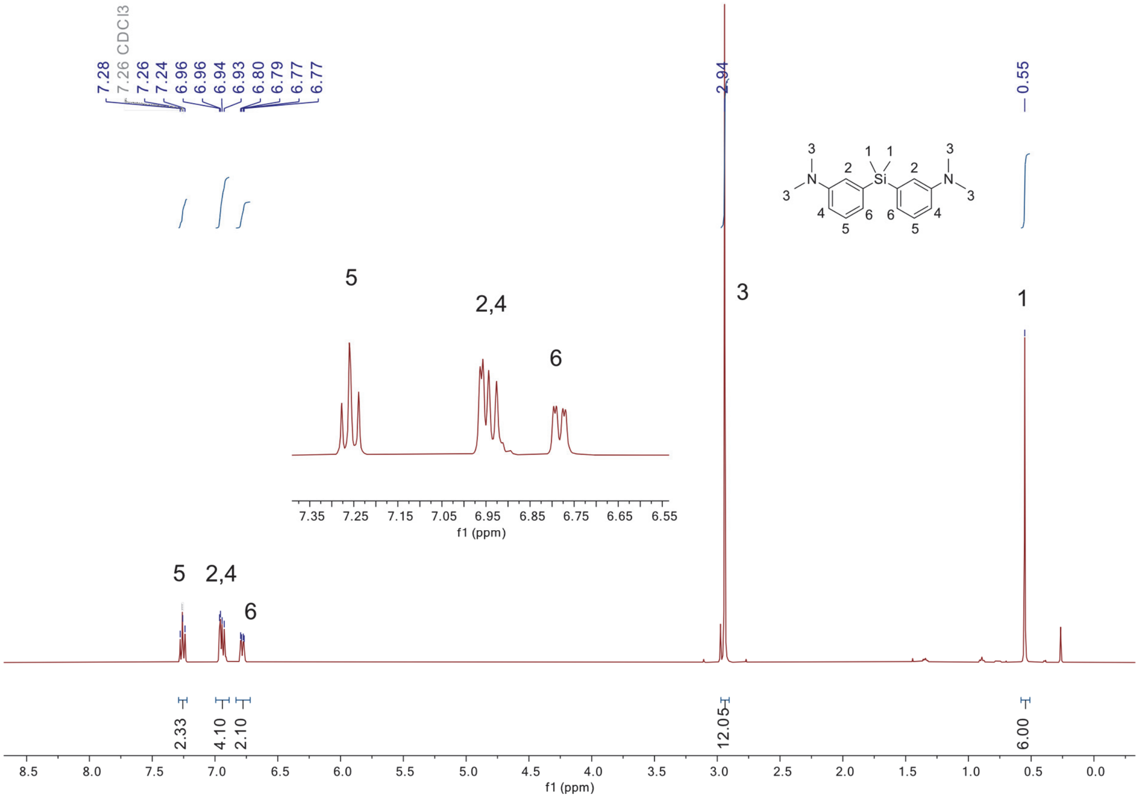


Figure 89. ^{1}H NMR ($CDCl_3$, 400 MHz) spectrum of compound **4.**

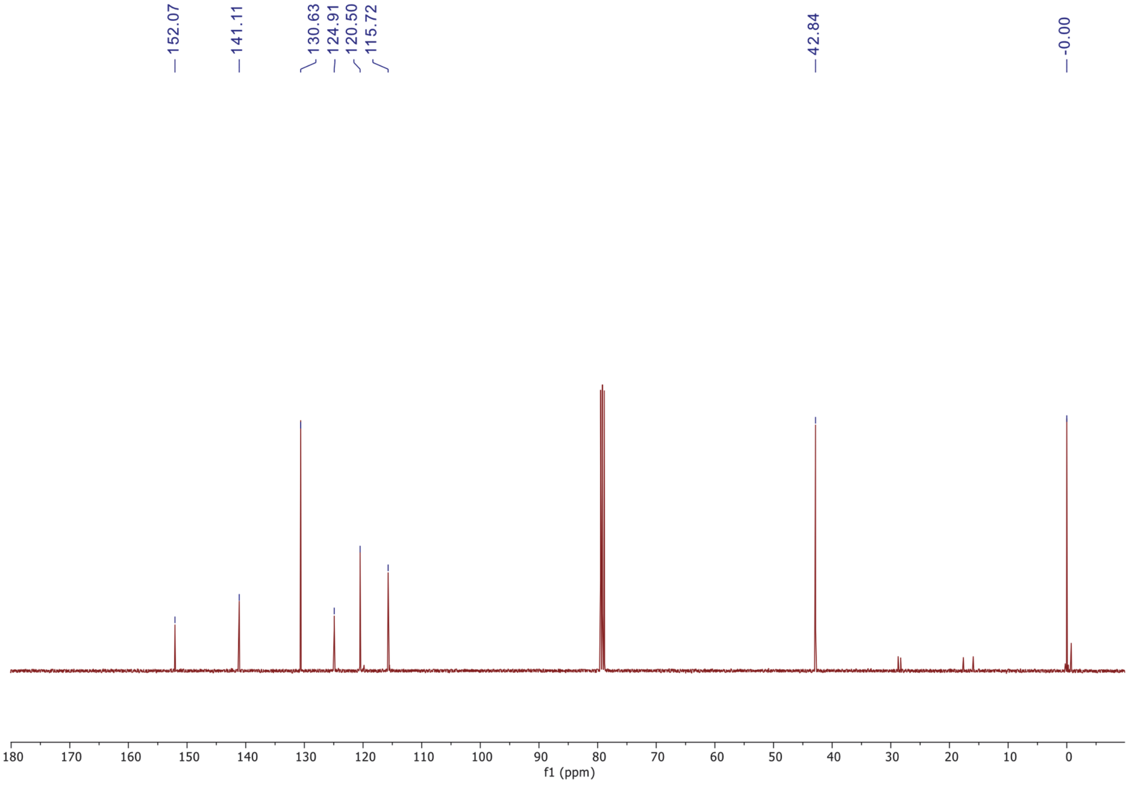


Figure 90. ^{13}C NMR ($CDCl_3$, 101 MHz) spectrum of compound **4.**

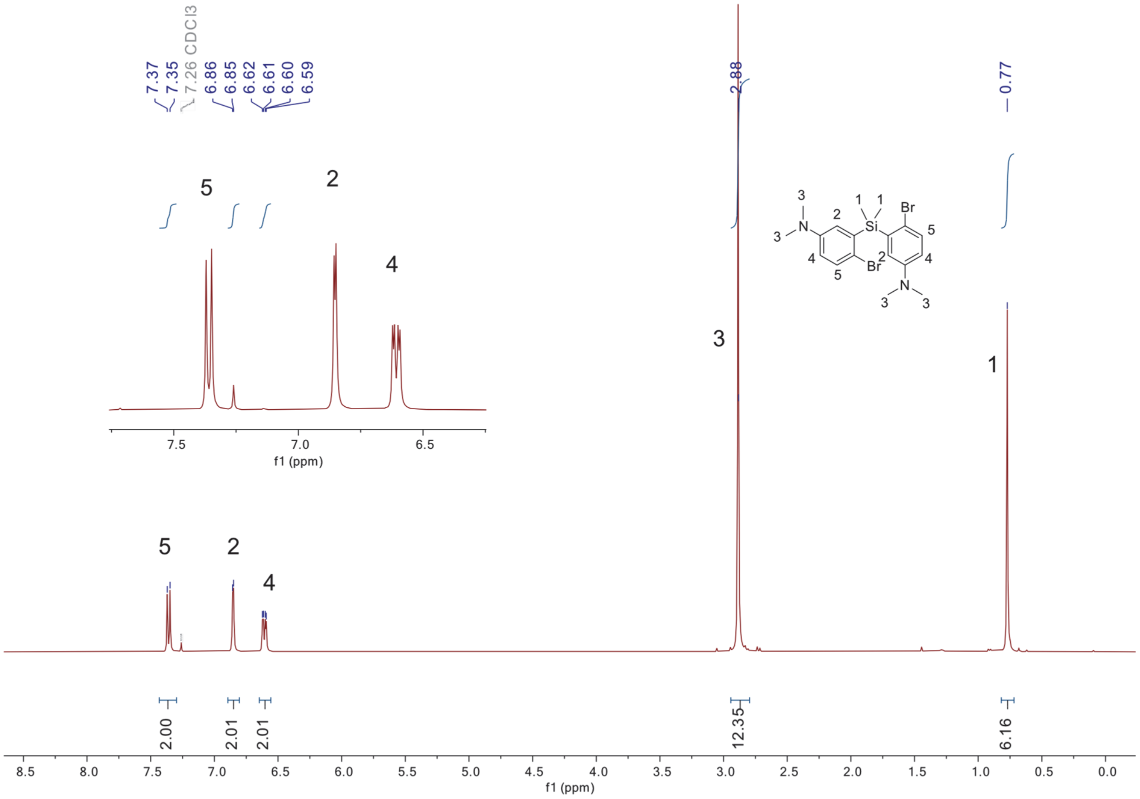


Figure 91. ^{1}H NMR ($CDCl_3$, 400 MHz) spectrum of compound **5.**

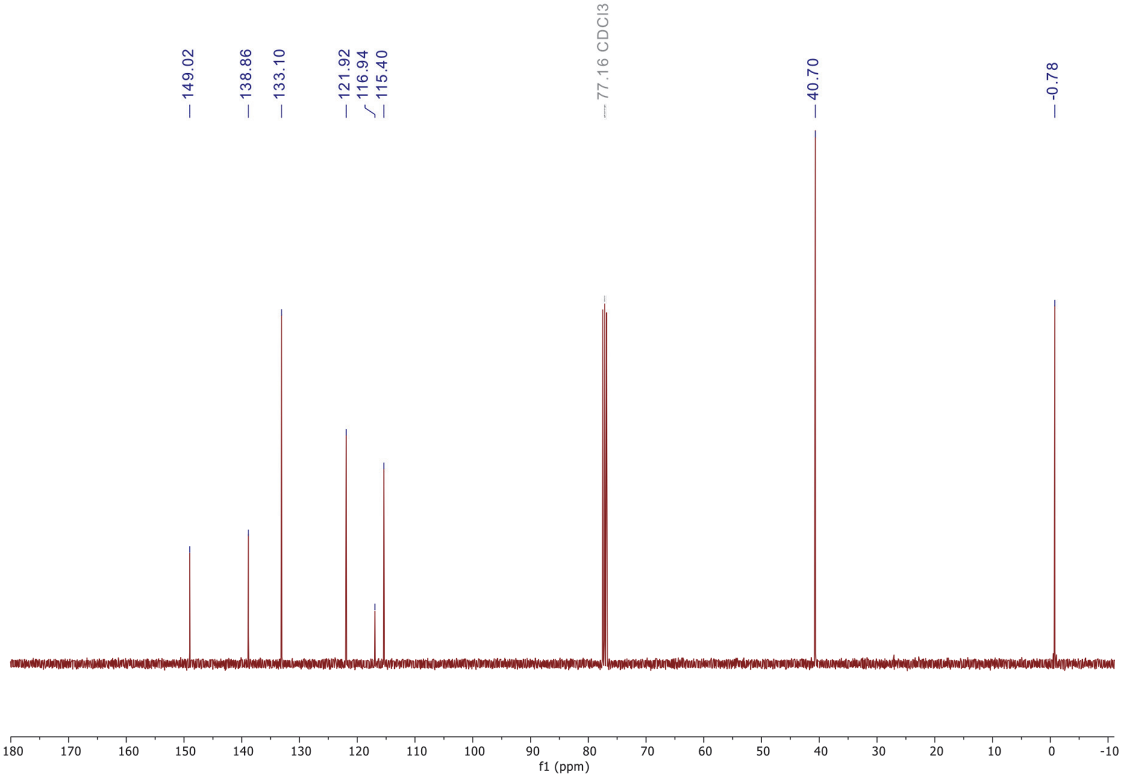


Figure 92. ^{13}C NMR ($CDCl_3$, 101 MHz) spectrum of compound **5.**

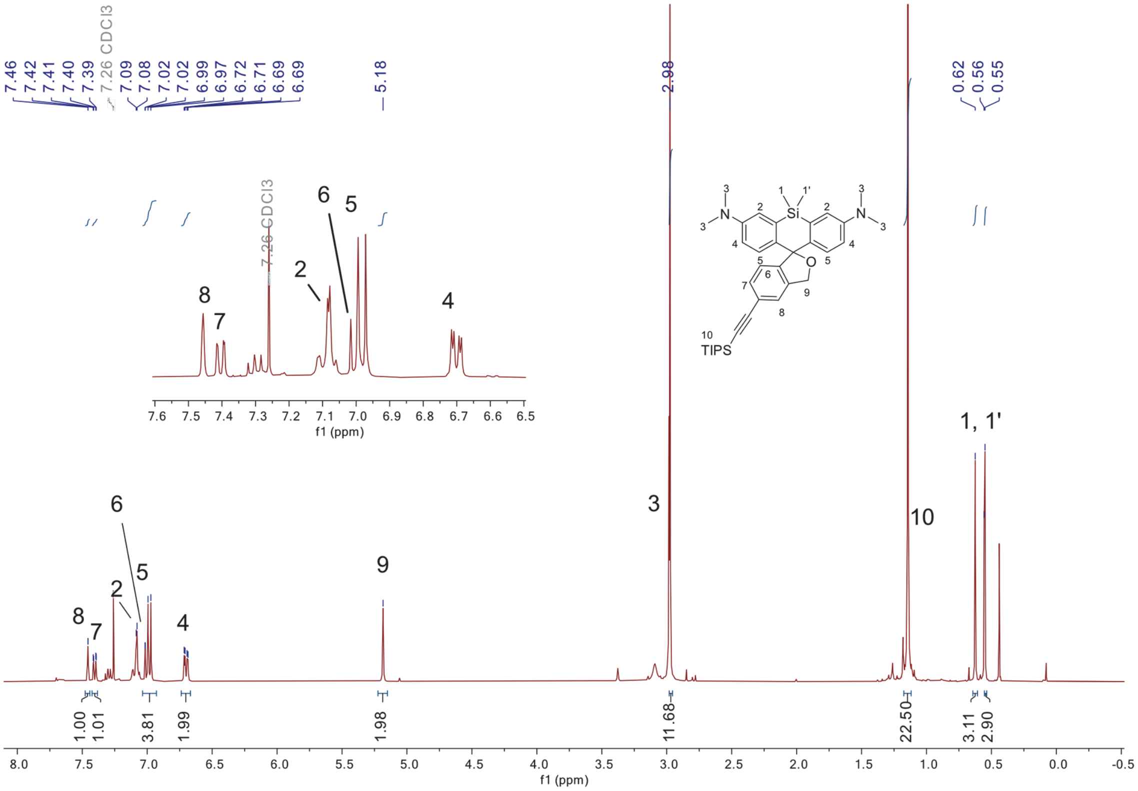


Figure 93. ^{1}H NMR (CDCl$_3$, 400 MHz) spectrum of compound **6**

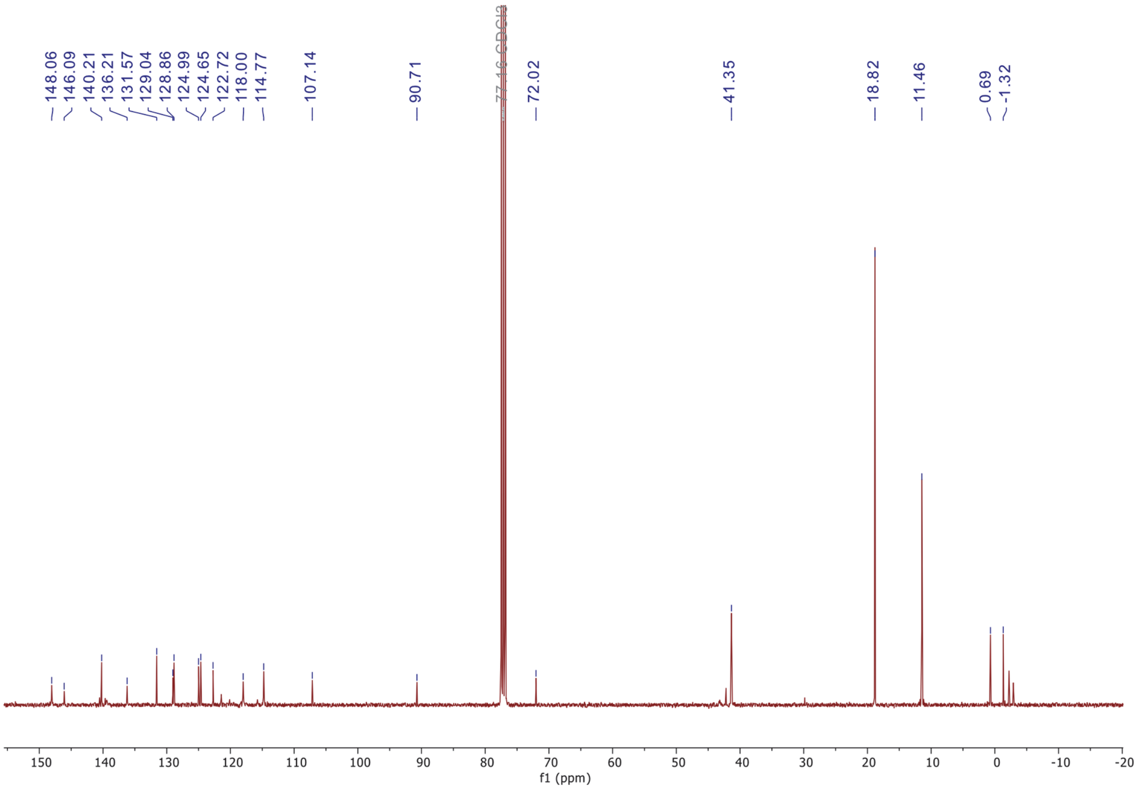


Figure 94. ^{13}C NMR (CDCl$_3$, 101 MHz) spectrum of compound **6**.

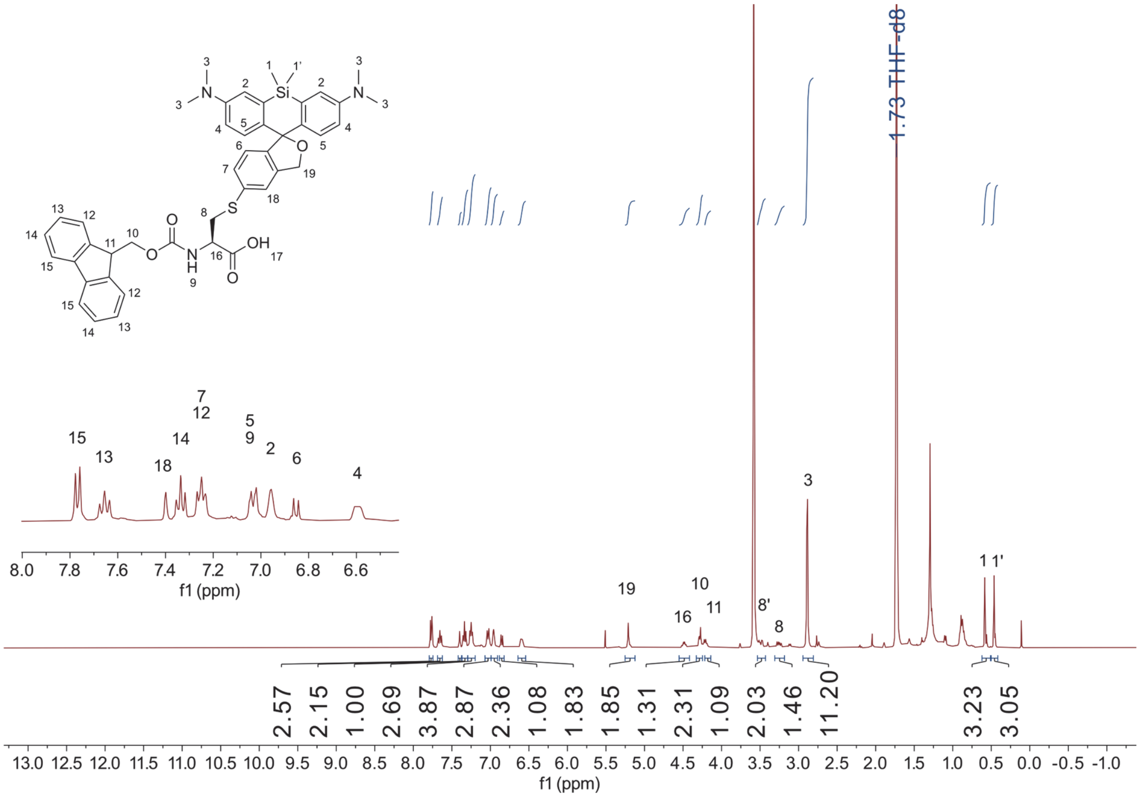


Figure 95. ^{1}H NMR (THF, 400 MHz) spectrum of compound **8.**

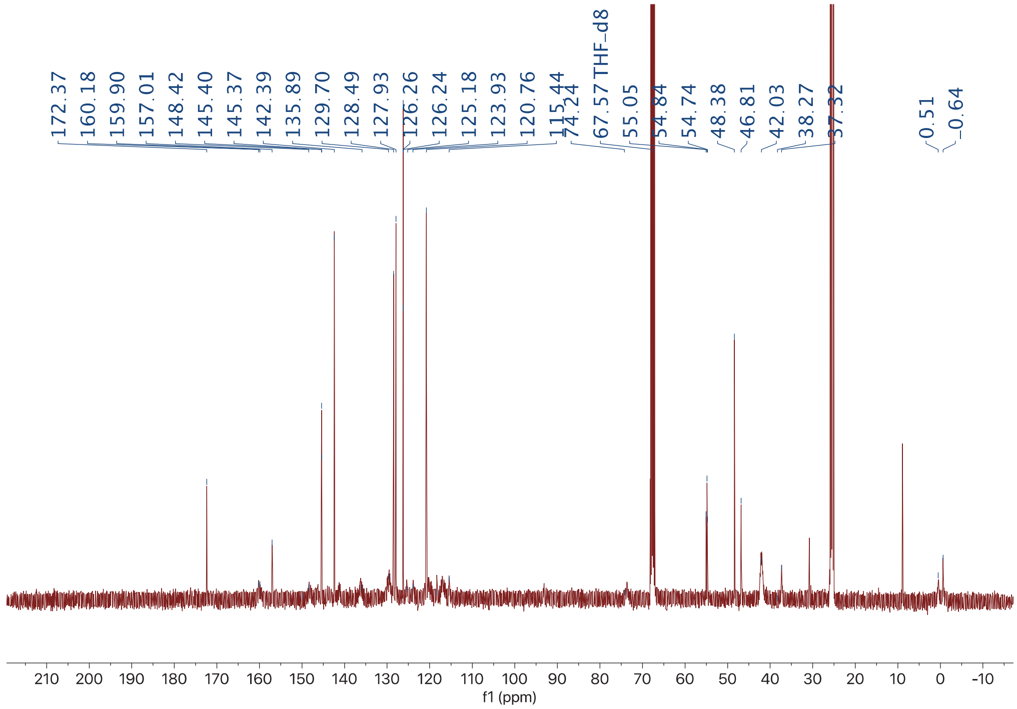


Figure 96. ^{13}C NMR (THF, 101 MHz) spectrum of compound **8.**

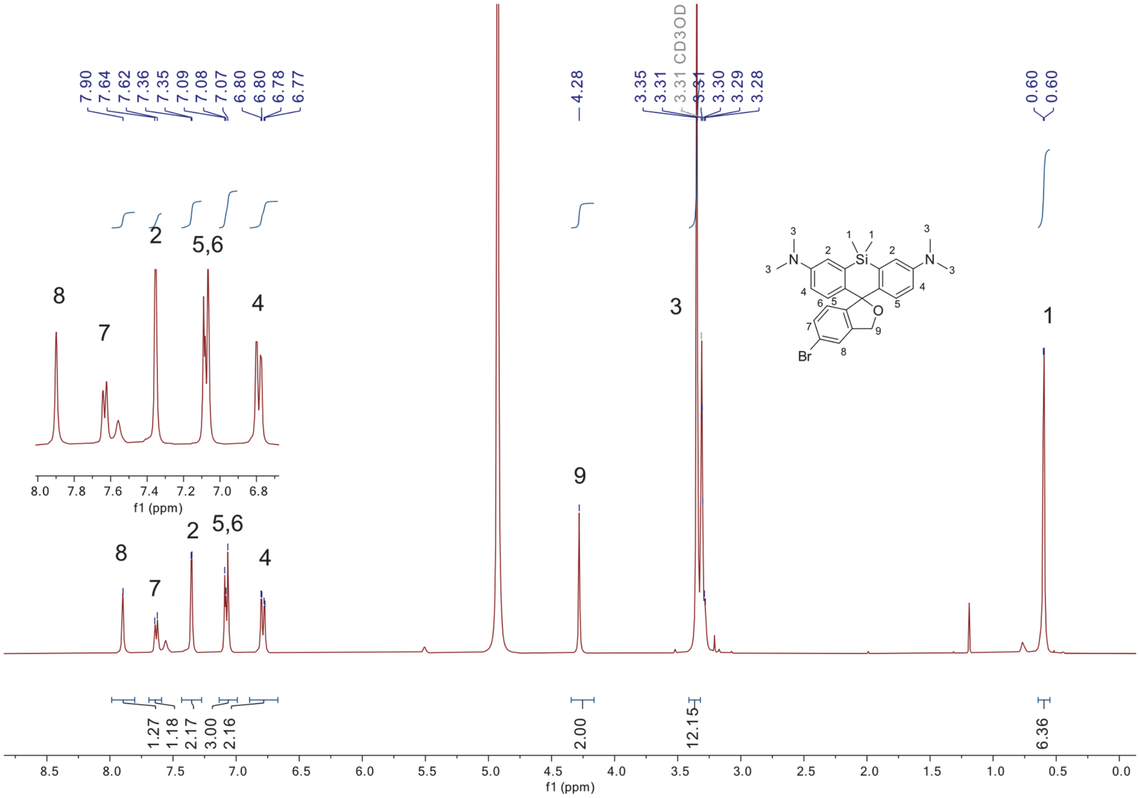


Figure 97. ^{1}H NMR ($CDCl_3$, 400 MHz) spectrum of compound **9**.

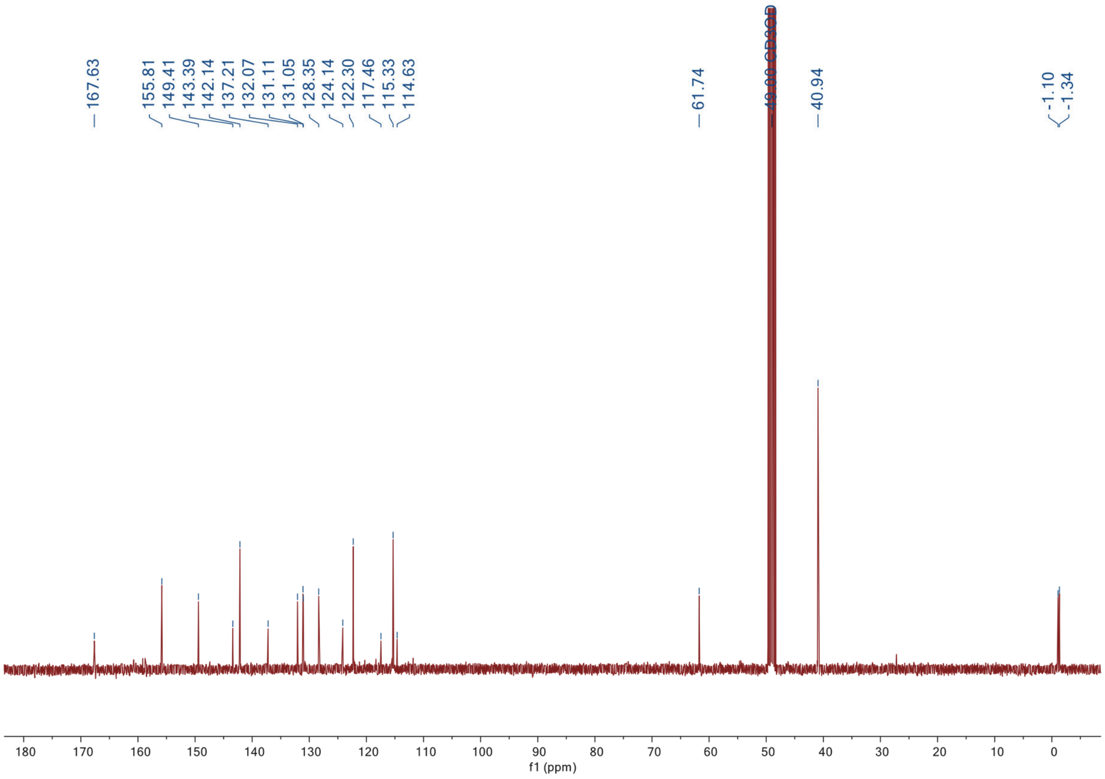


Figure 98. ^{13}C NMR ($CDCl_3$, 101 MHz) spectrum of compound **9**.

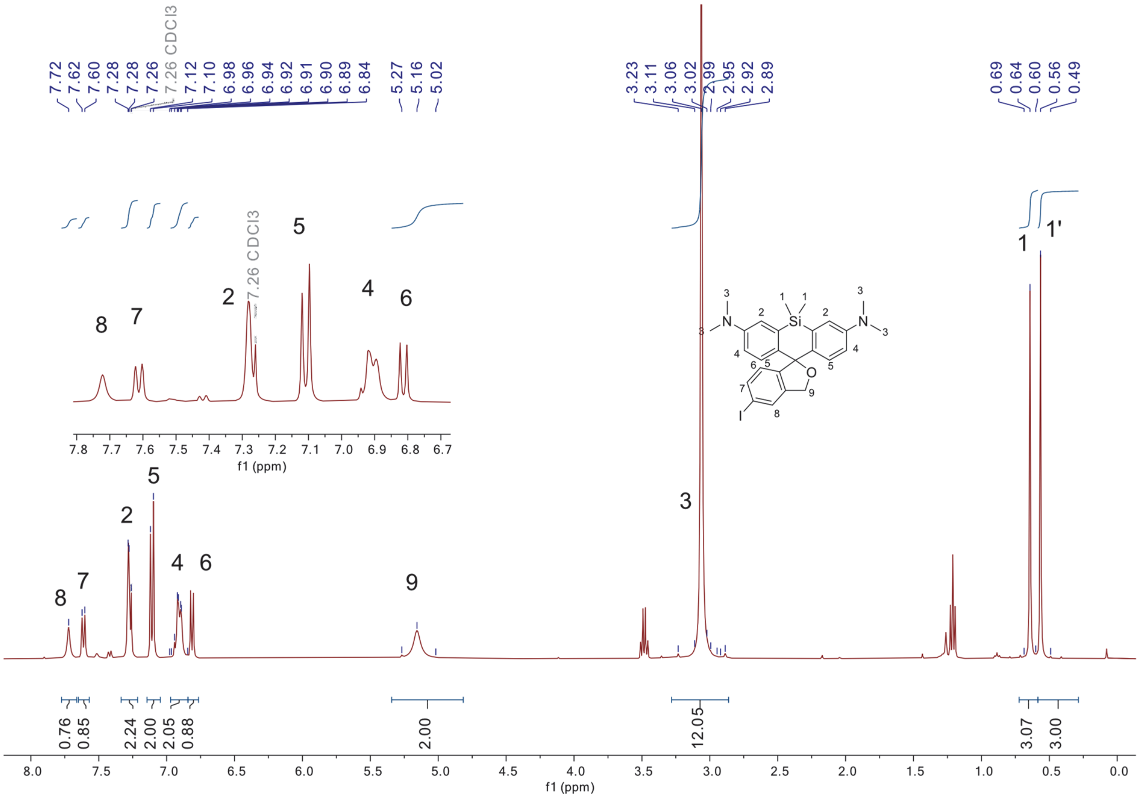


Figure 99. ^{1}H NMR ($CDCl_3$, 400 MHz) spectrum of compound **10**.

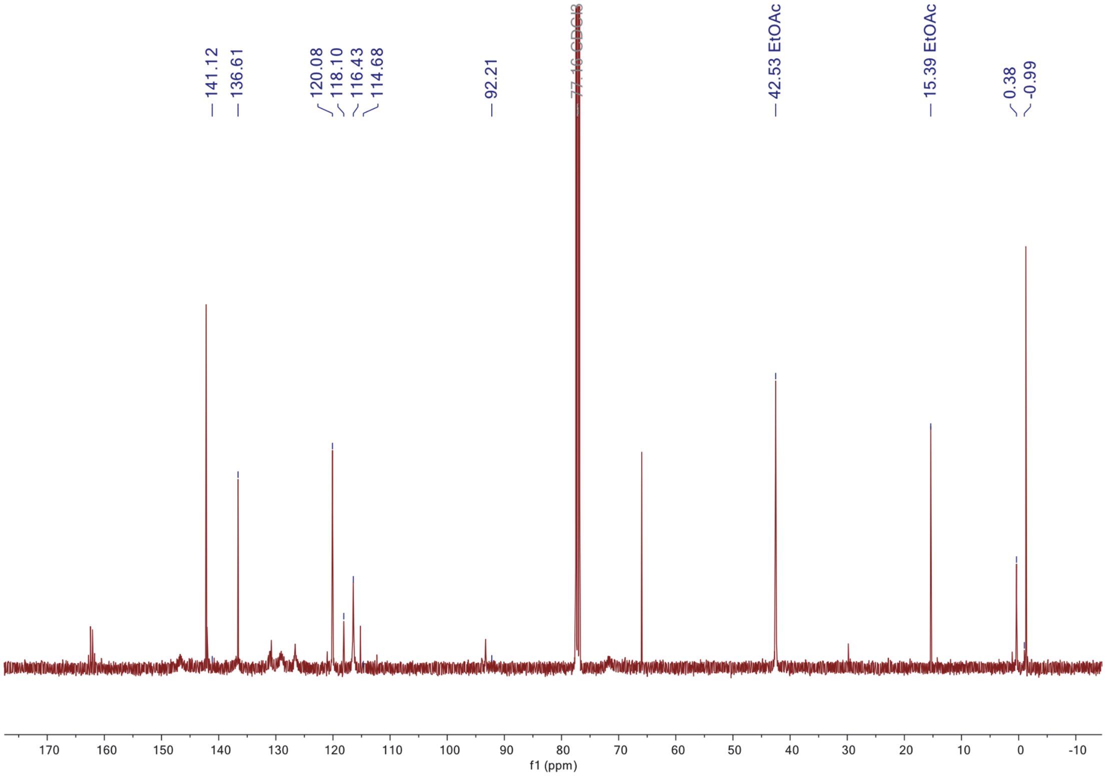


Figure 100. ^{13}C NMR ($CDCl_3$, 101 MHz) spectrum of compound **10**.

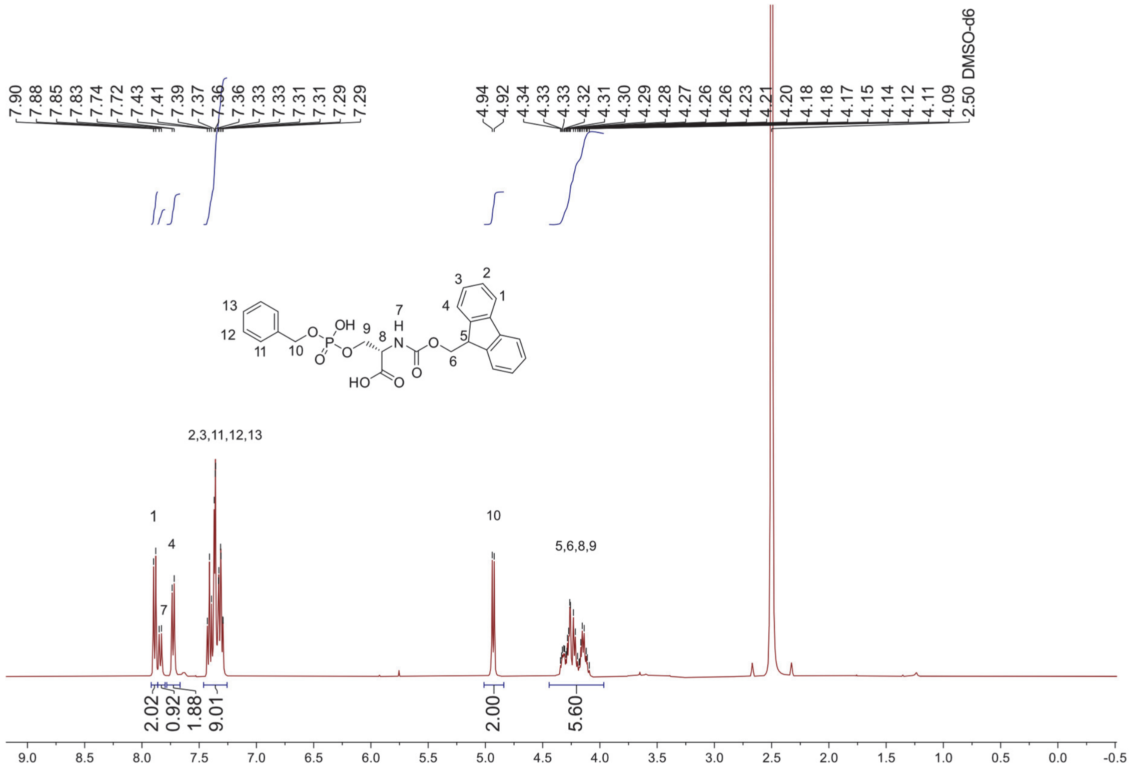


Figure 101. ^{1}H NMR (DMSO-d_6, 400 MHz) spectrum of compound **18**.

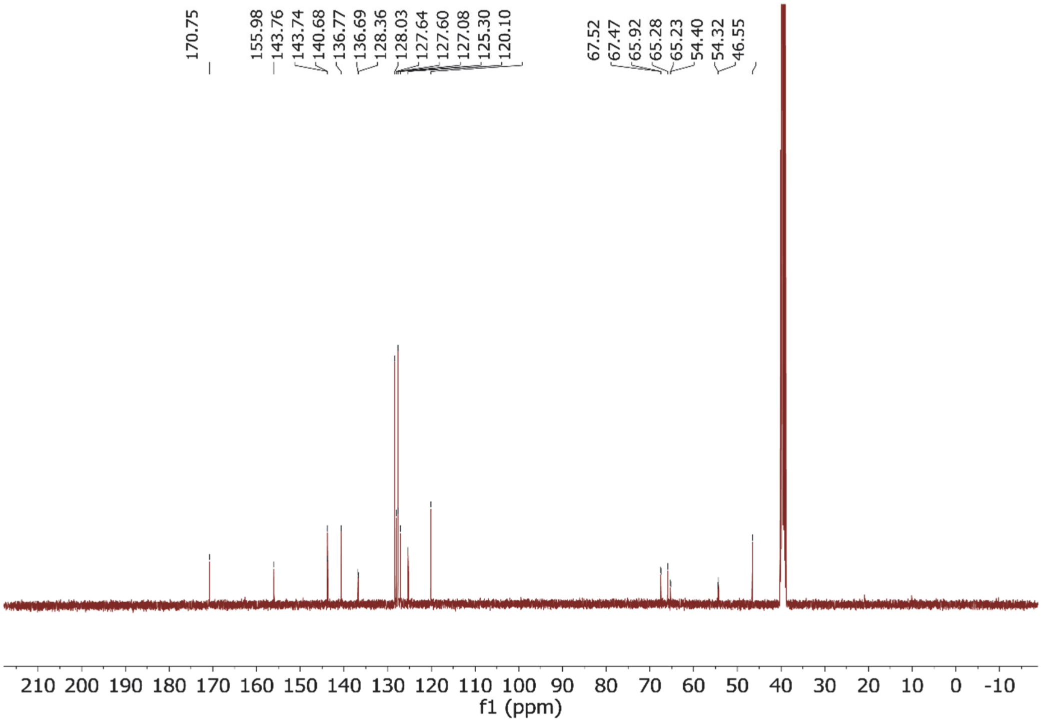


Figure 102. ^{13}C NMR (DMSO-d_6, 101 MHz) spectrum of compound **18**.

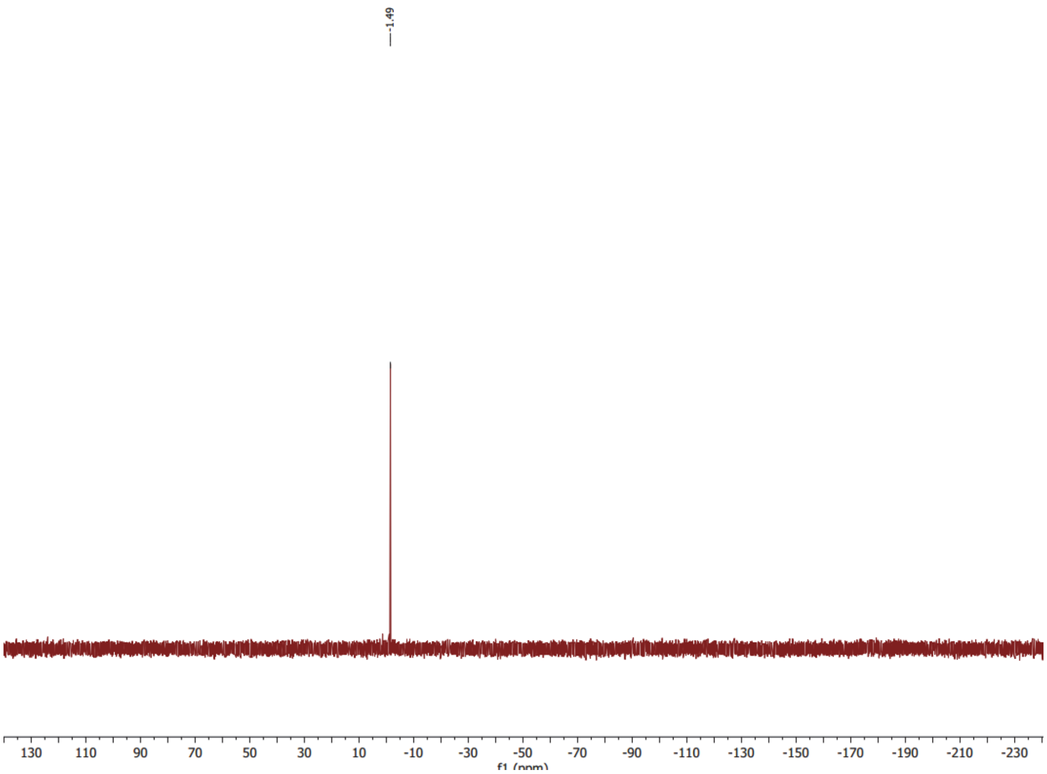


Figure 103. ^{31}P NMR (DMSO-d_6, 162 MHz) spectrum of compound **18**.

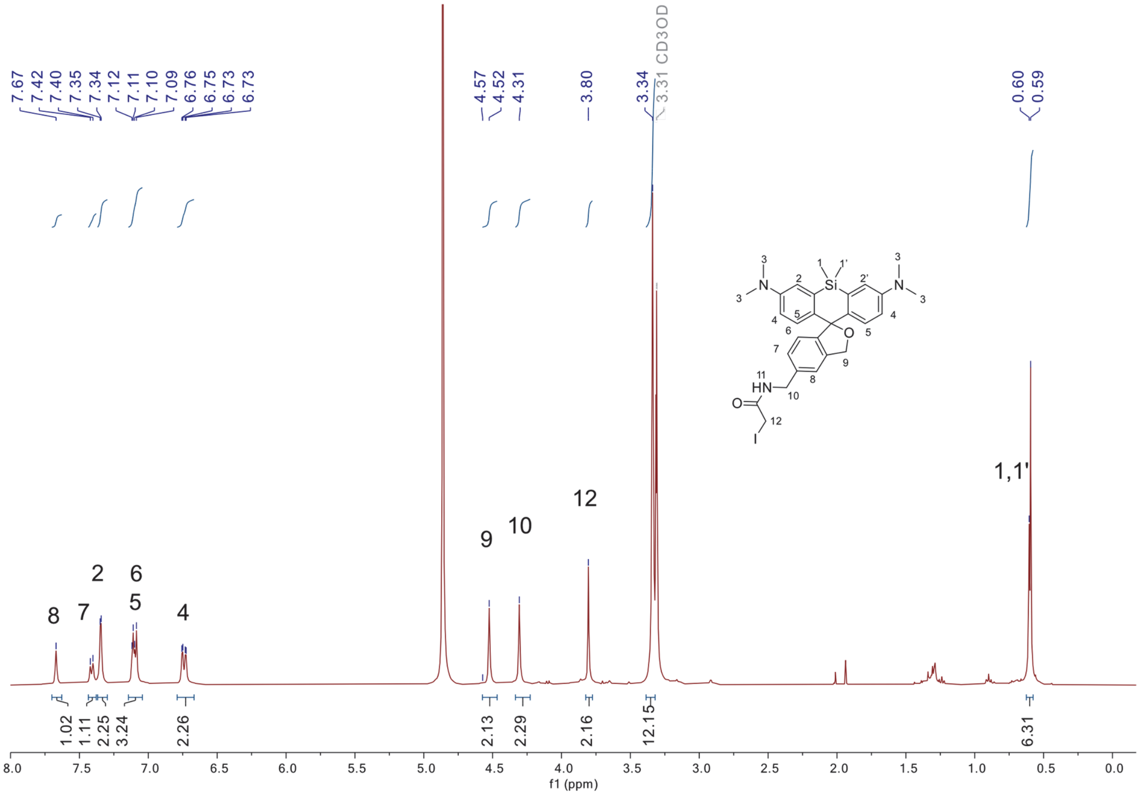


Figure 104. ^{1}H NMR (CD_3OD, 400 MHz) spectrum of compound **22**

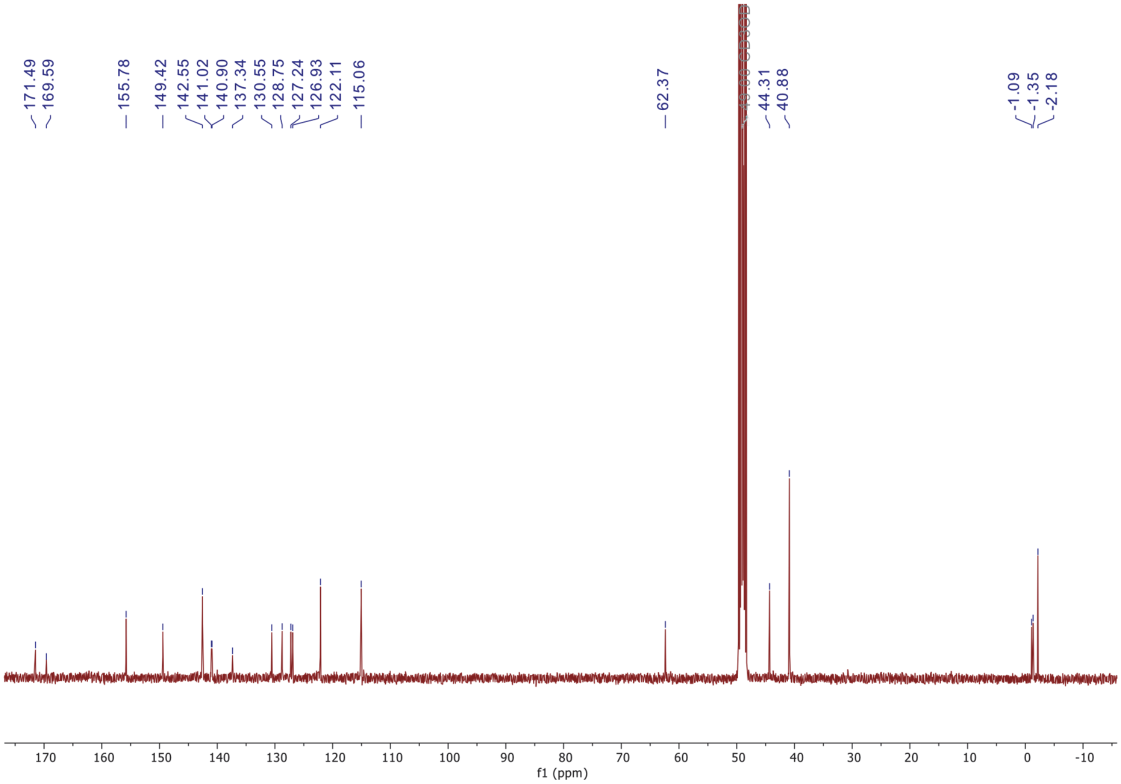


Figure 105. ^{13}C NMR (CD_3OD, 101 MHz) spectrum of compound **22**

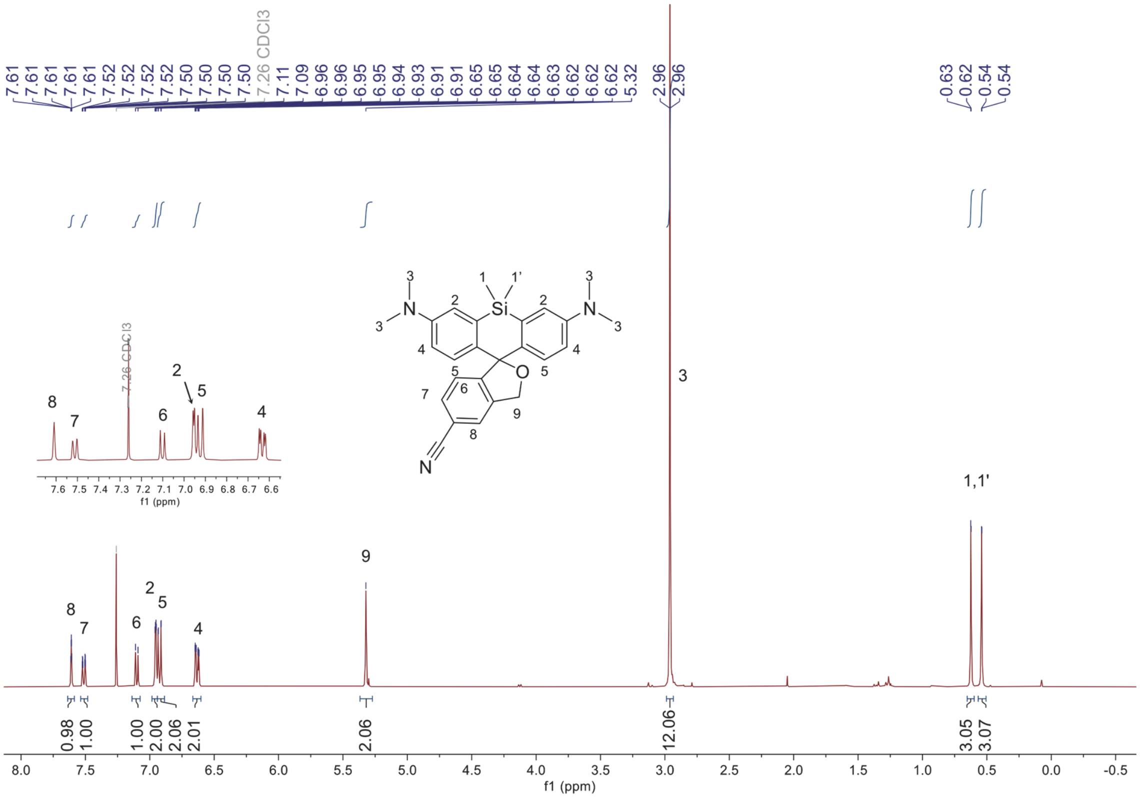


Figure 106. ^{1}H NMR (CD_3Cl, 400 MHz) spectrum of compound **23**

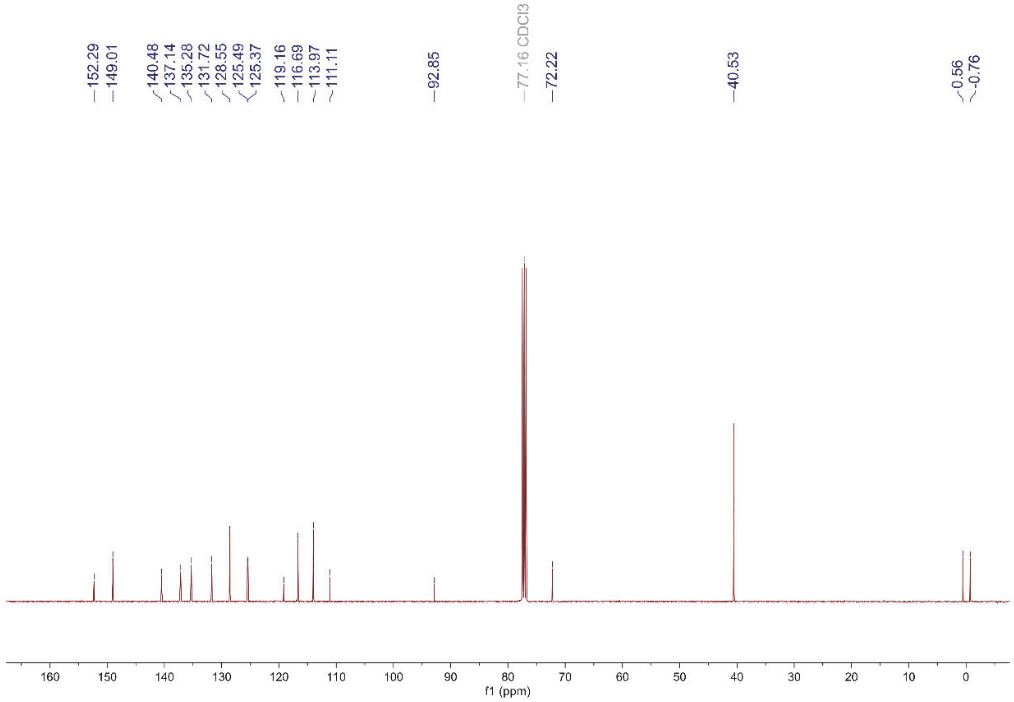


Figure 107. ^{13}C NMR (CD_3Cl, 101 MHz) spectrum of compound **23**

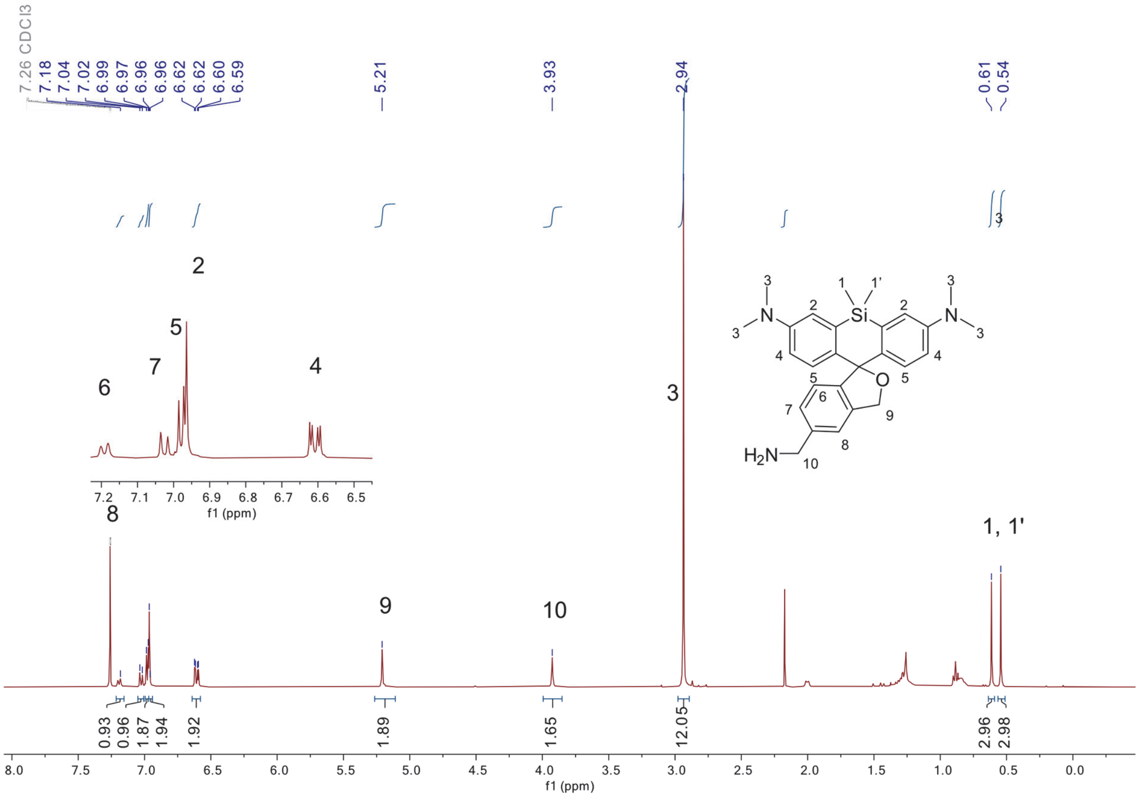


Figure 108. ^{1}H NMR (CD_3Cl, 400 MHz) spectrum of compound **24**

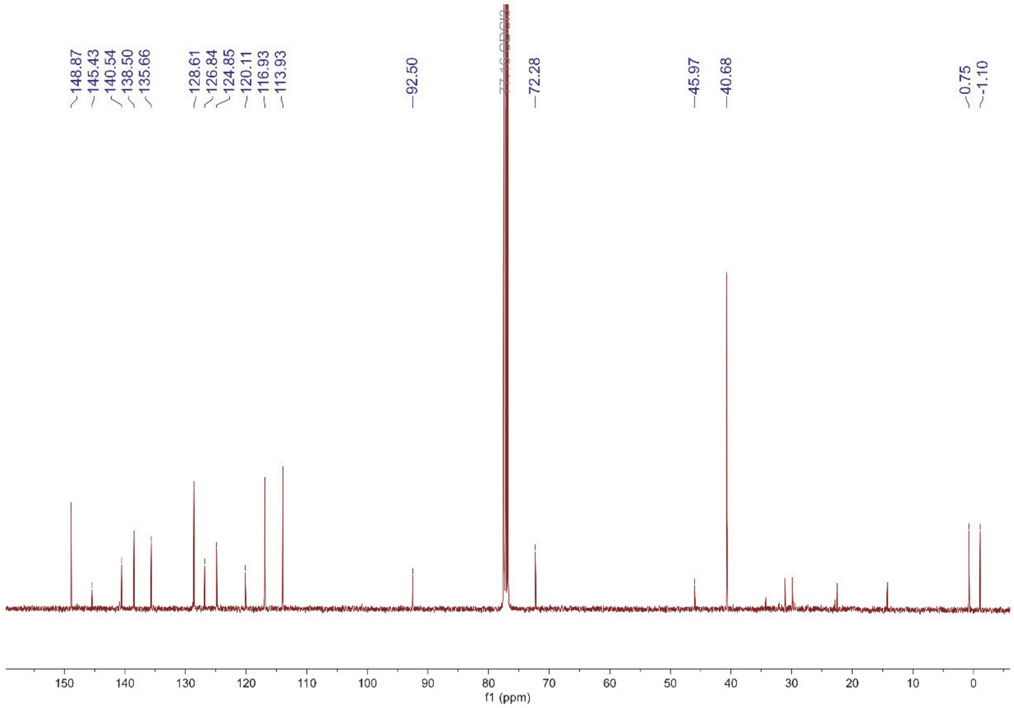


Figure 109. ^{13}C NMR (CD_3Cl, 101 MHz) spectrum of compound **24**

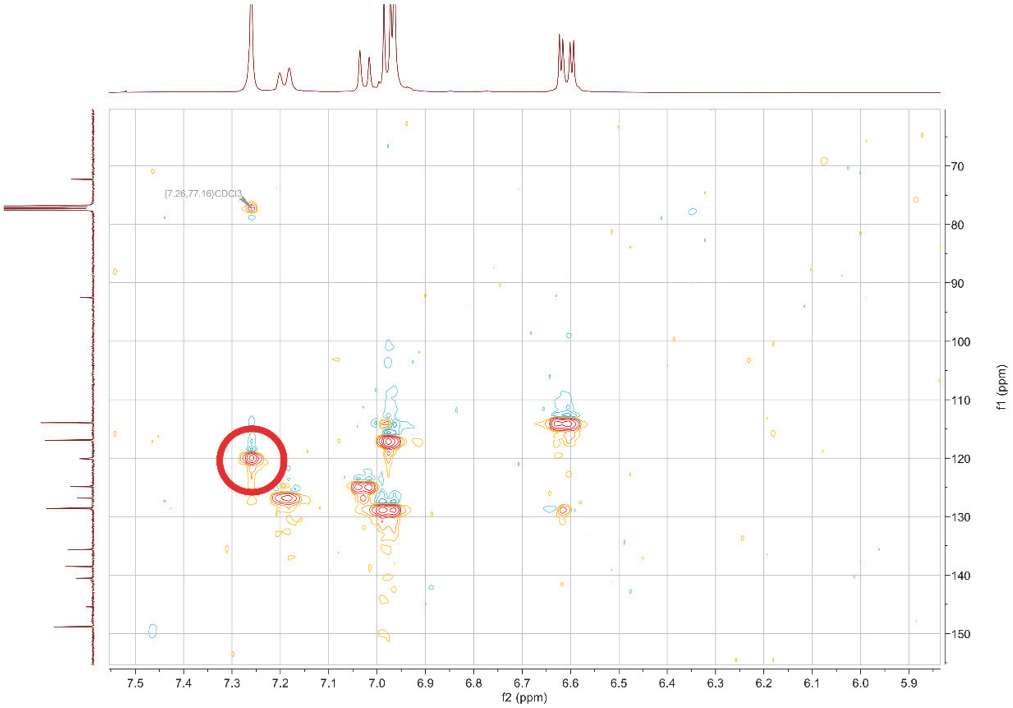


Figure 110. HSQC NMR (CD_3Cl) Blow-up of region showing the hidden peak of H8 in the HSQC spectrum of compound **24**

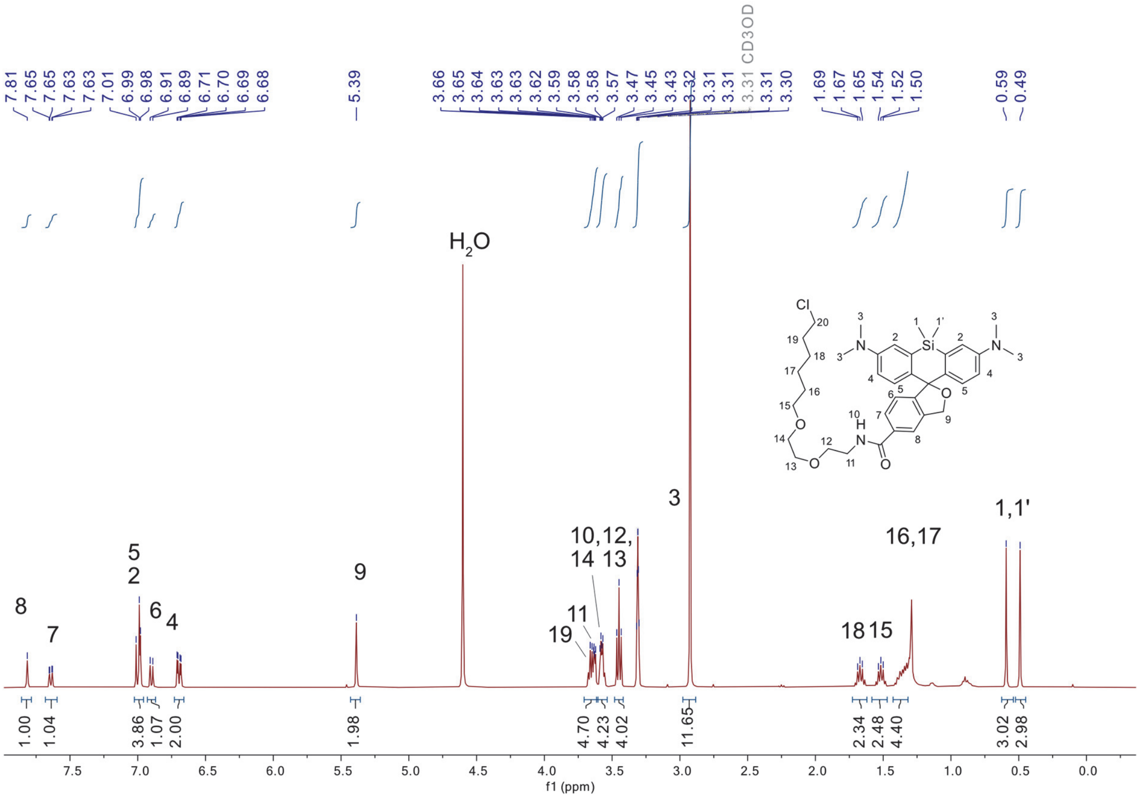


Figure 111. ^{1}H NMR (CD_3OD, 400 MHz) spectrum of compound **HMSiR-Halo 25**.

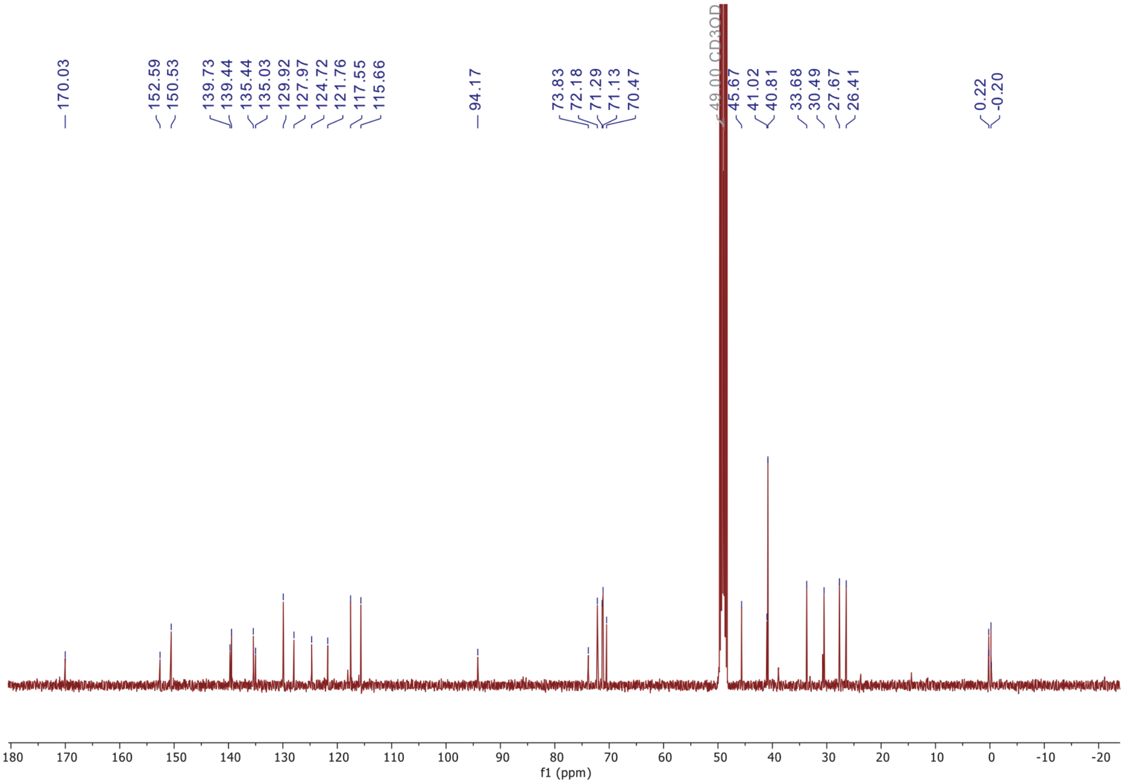


Figure 112. ^{13}C NMR (CD_3OD, 100 MHz) spectrum of compound **HMSiR-Halo 25**.

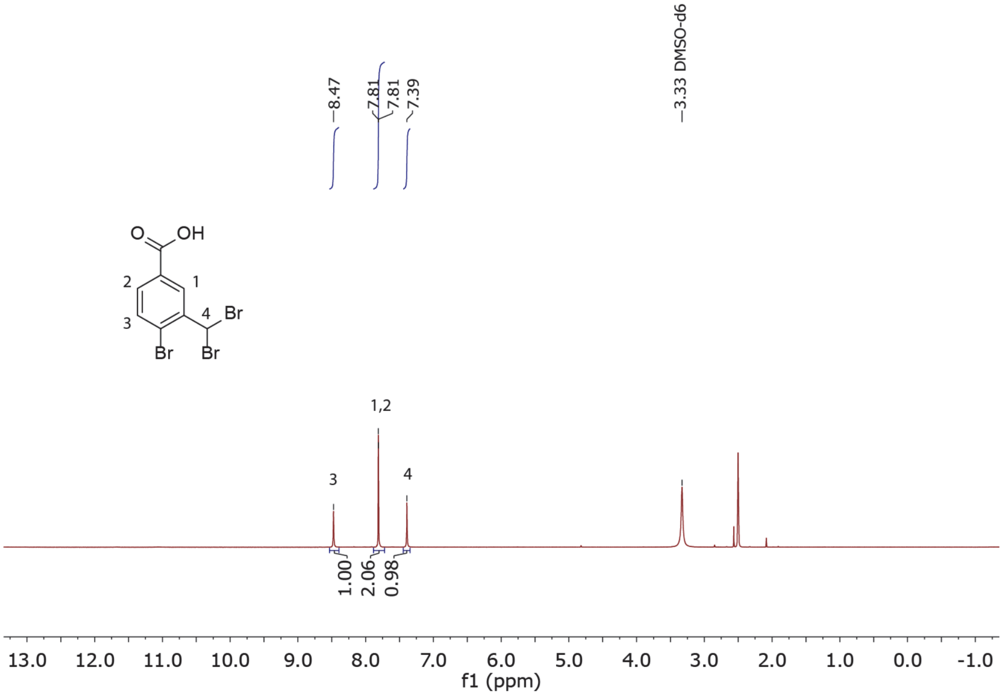


Figure 113. ^{1}H NMR (DMSO-d_6, 400 MHz) spectrum of compound **27**

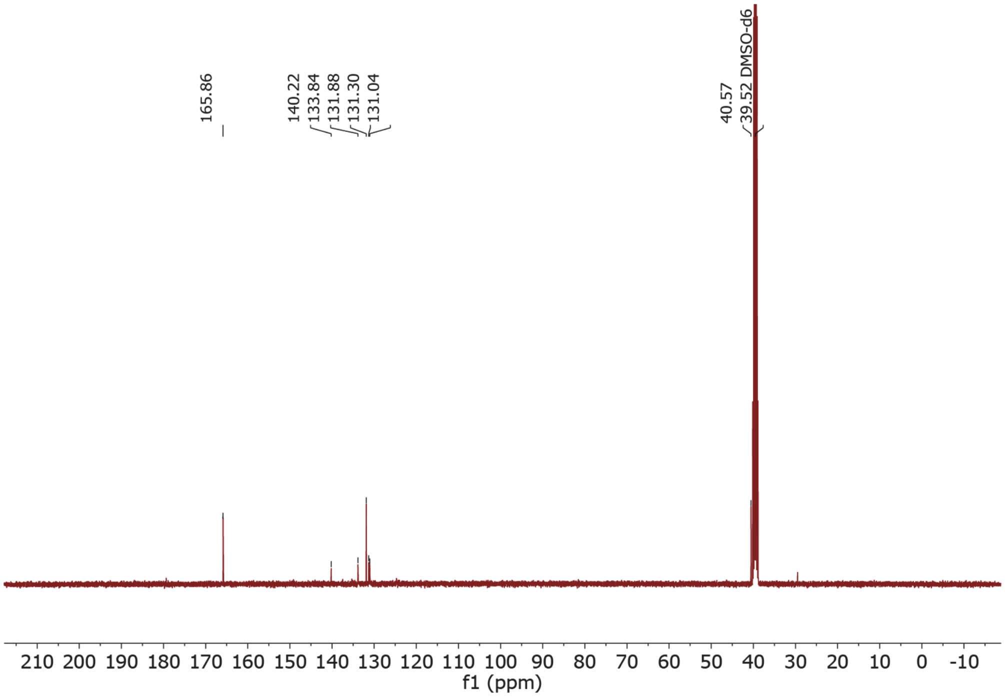


Figure 114. ^{13}C NMR (DMSO-d_6, 101 MHz) spectrum of compound **27**.

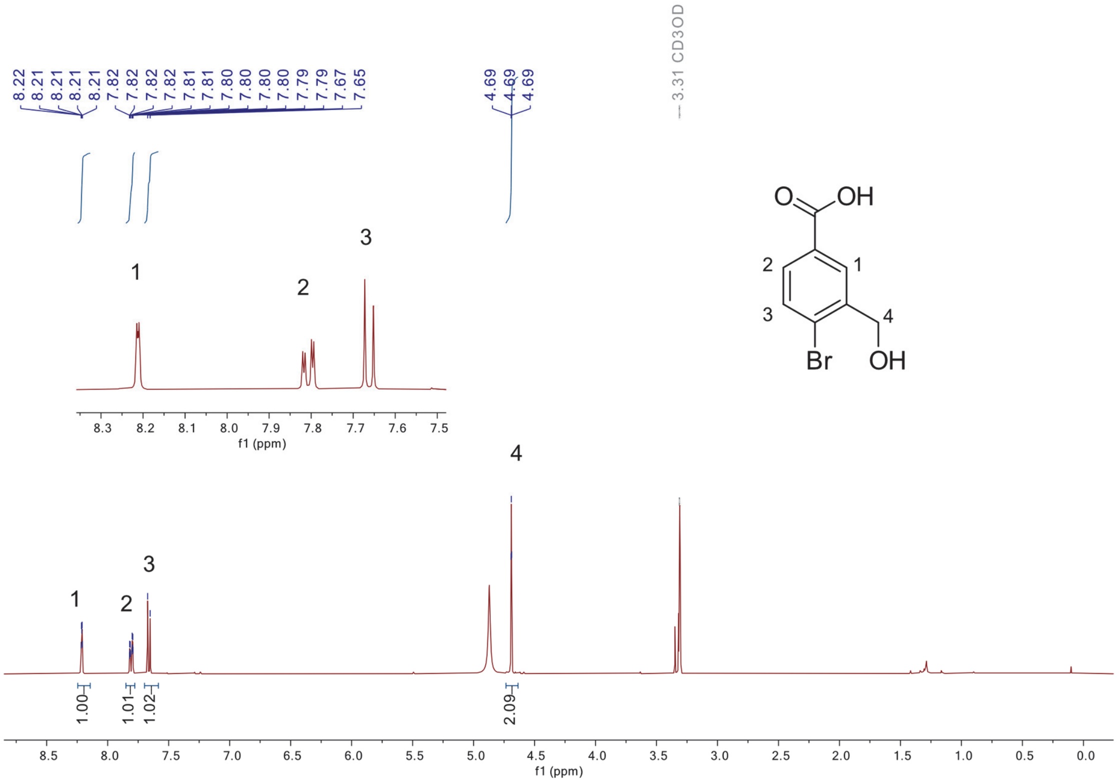


Figure 115. ^{1}H NMR (CD_3OD, 400 MHz) spectrum of compound **28**.

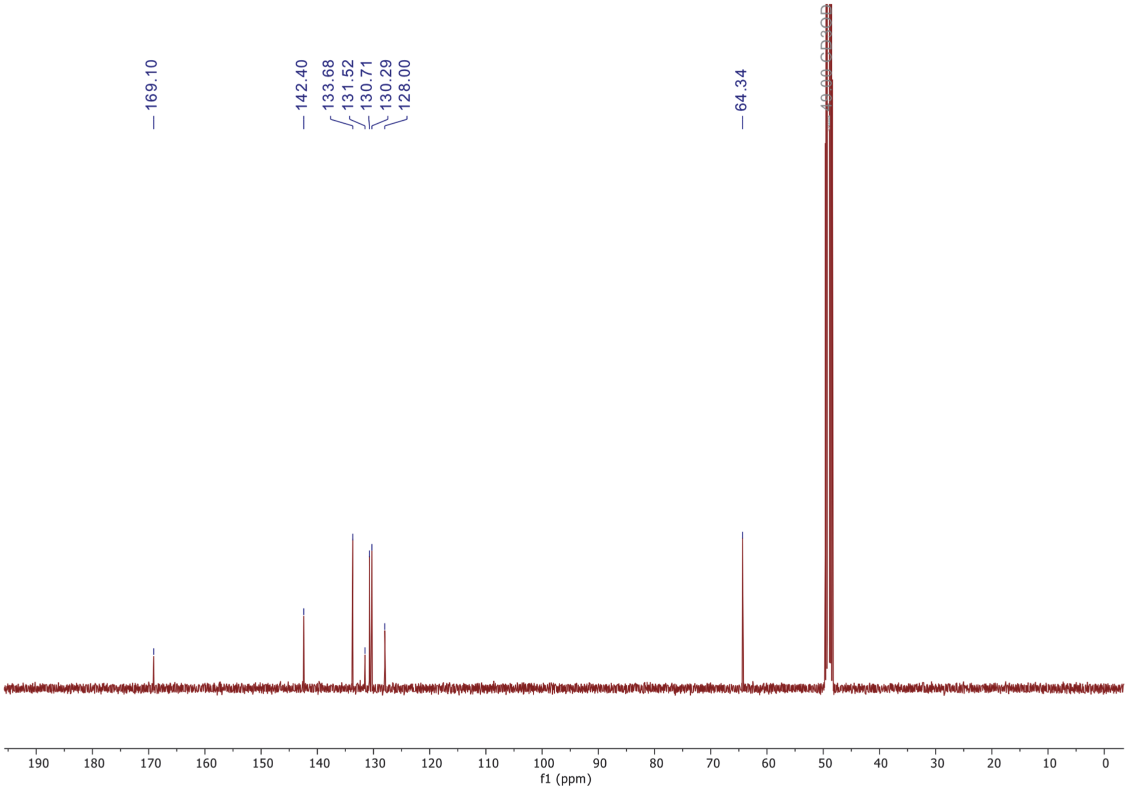


Figure 116. ^{13}C NMR (CD_3OD, 101 MHz) spectrum of compound **28**.

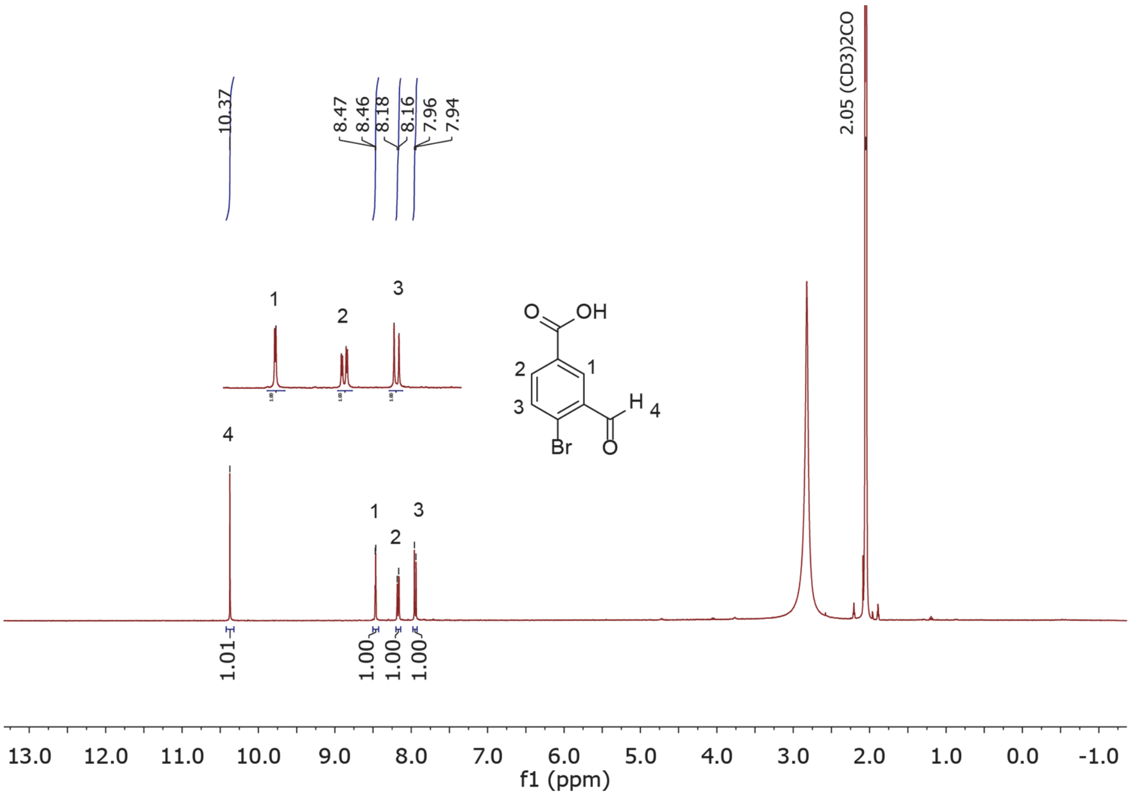


Figure 117. ^{1}H NMR ($(CD_3)_2CO$, 400 MHz) spectrum of compound **29**.

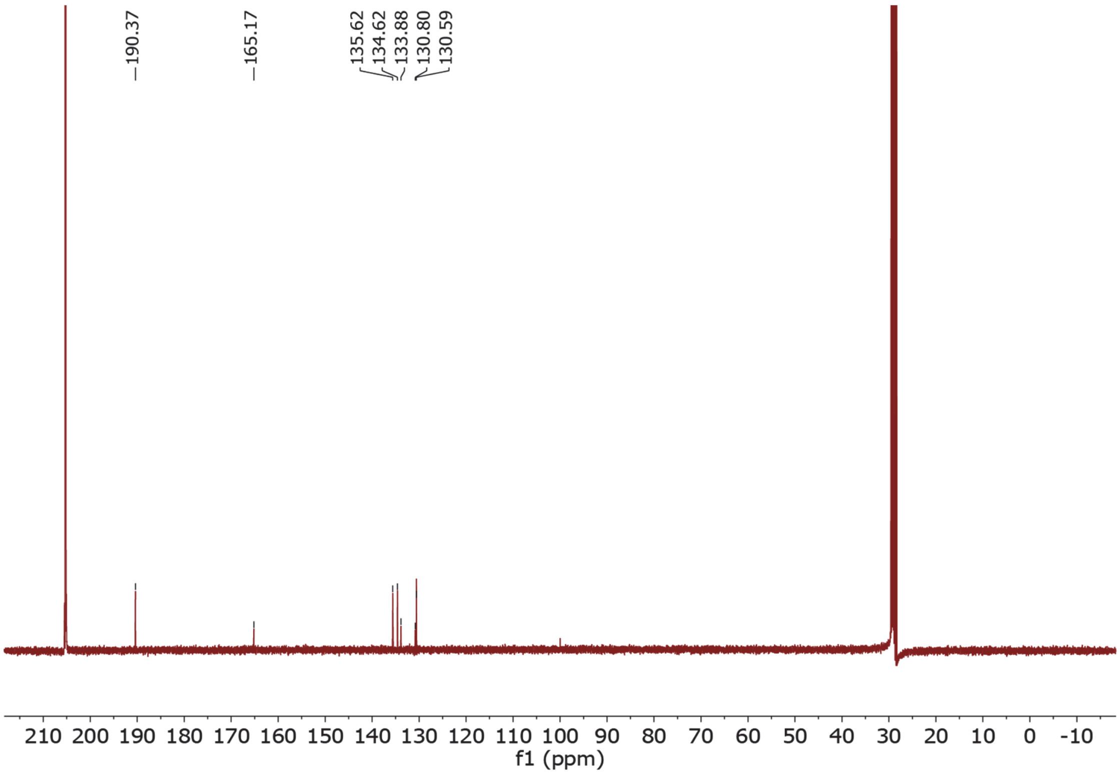


Figure 118. ^{13}C NMR ($(CD_3)_2CO$, 101 MHz) spectrum of compound **29**.

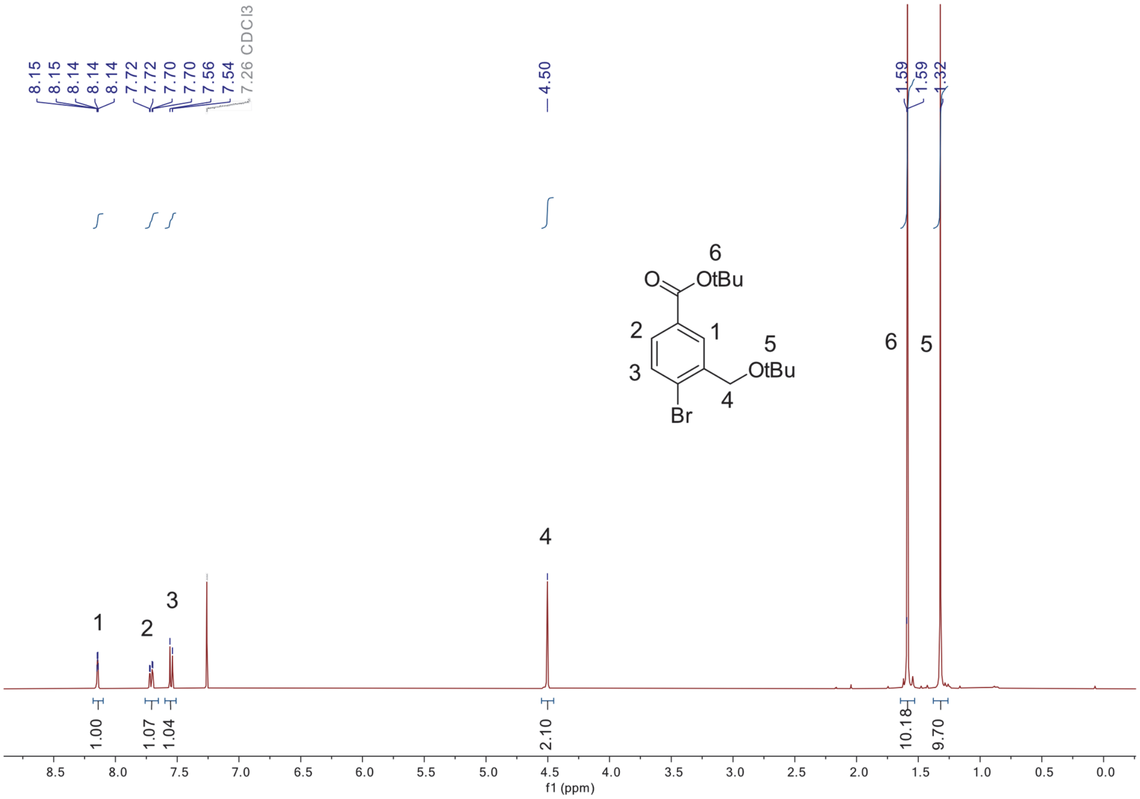


Figure 119. ^{1}H NMR ($CDCl_3$, 400 MHz) spectrum of compound **30.**

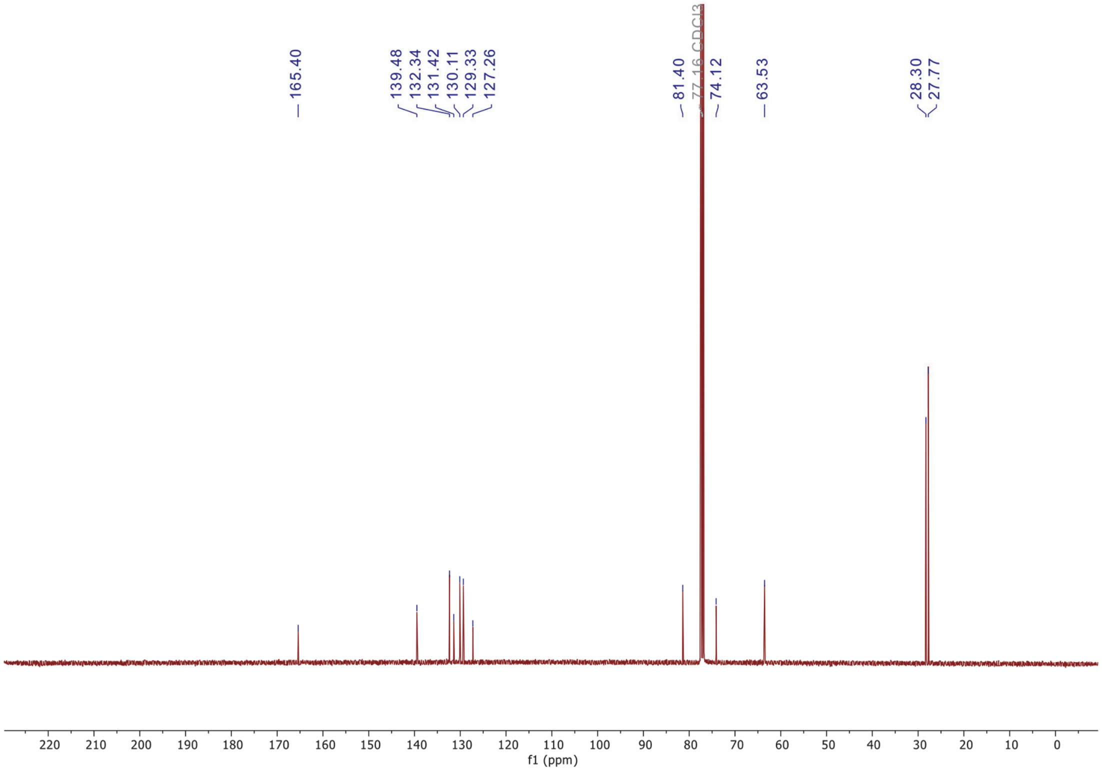


Figure 120. ^{13}C NMR ($CDCl_3$, 101 MHz) spectrum of compound **30.**

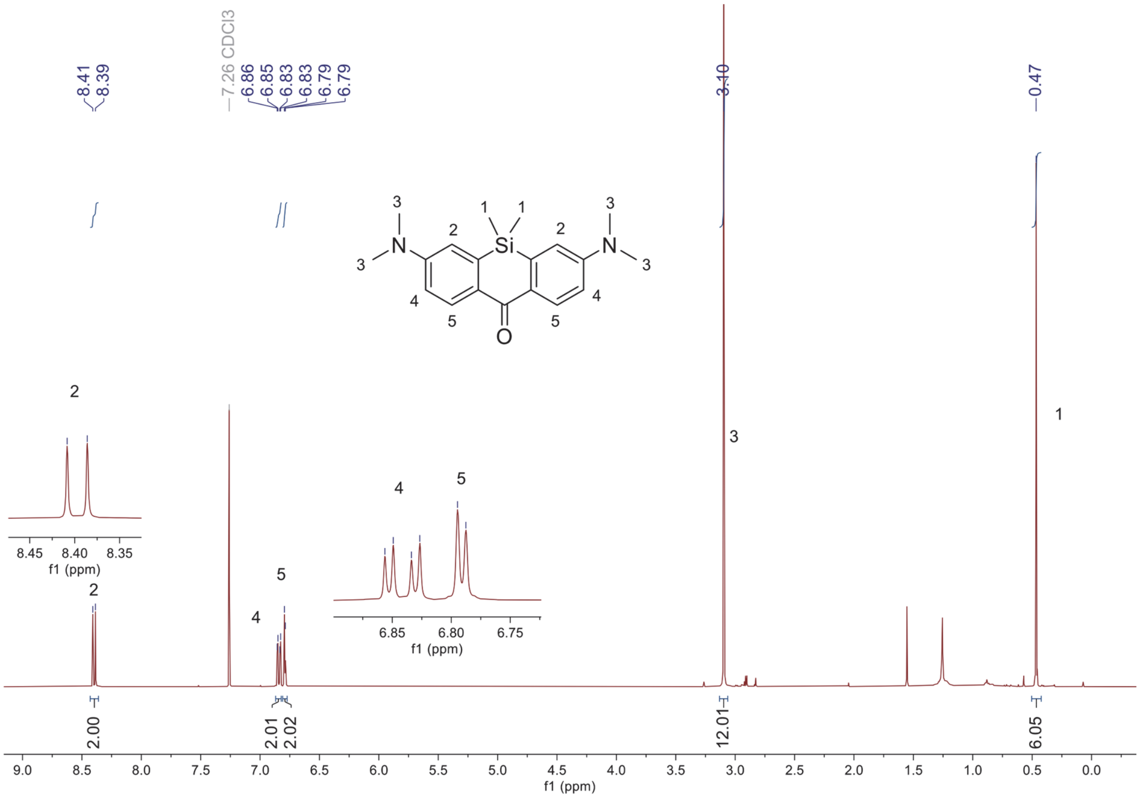


Figure 121. ^{1}H NMR (CD_3Cl, 400 MHz) spectrum of compound **31.**

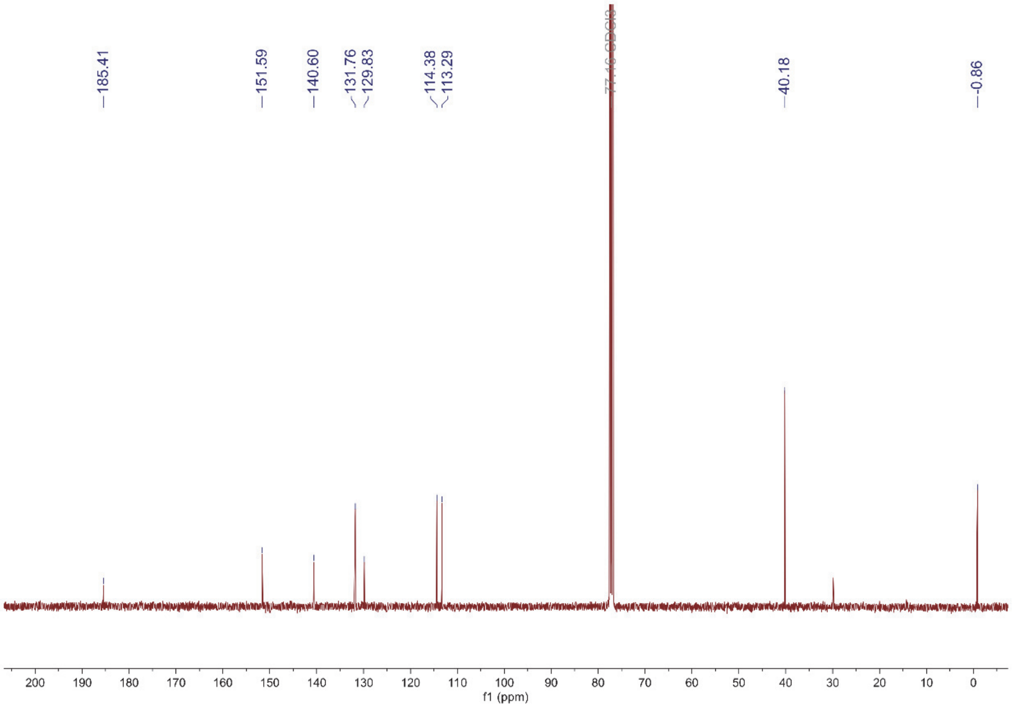


Figure 122. ^{13}C NMR (CD_3Cl, 100 MHz) spectrum of compound **31.**

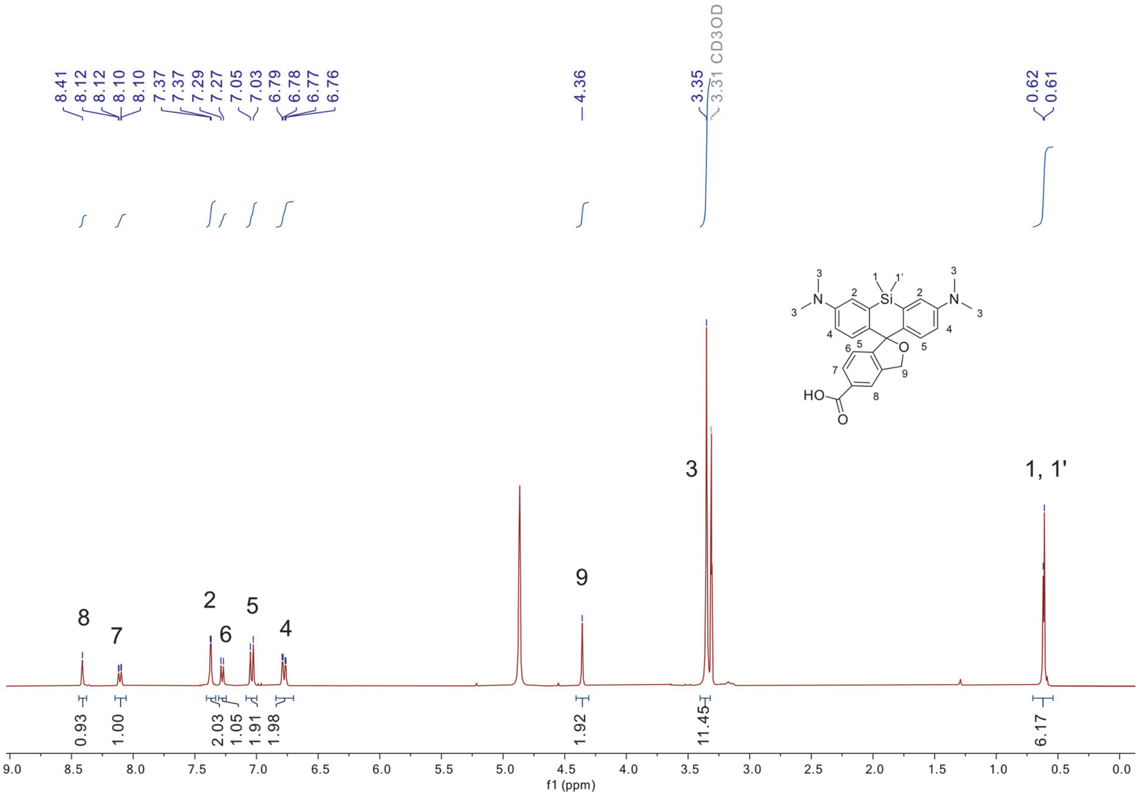


Figure 123. ^{1}H NMR (CD$_3$OD, 400 MHz) spectrum of compound **32**.

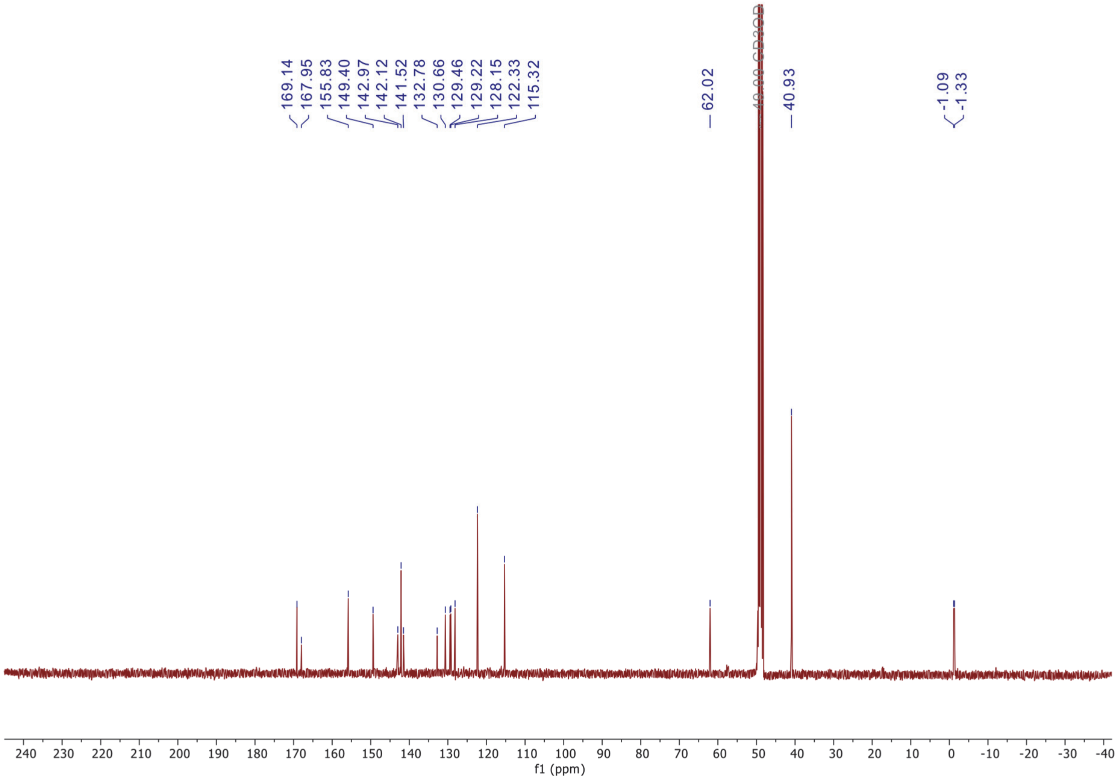


Figure 124. ^{13}C NMR (CD$_3$OD, 101 MHz) spectrum of compound **32**.

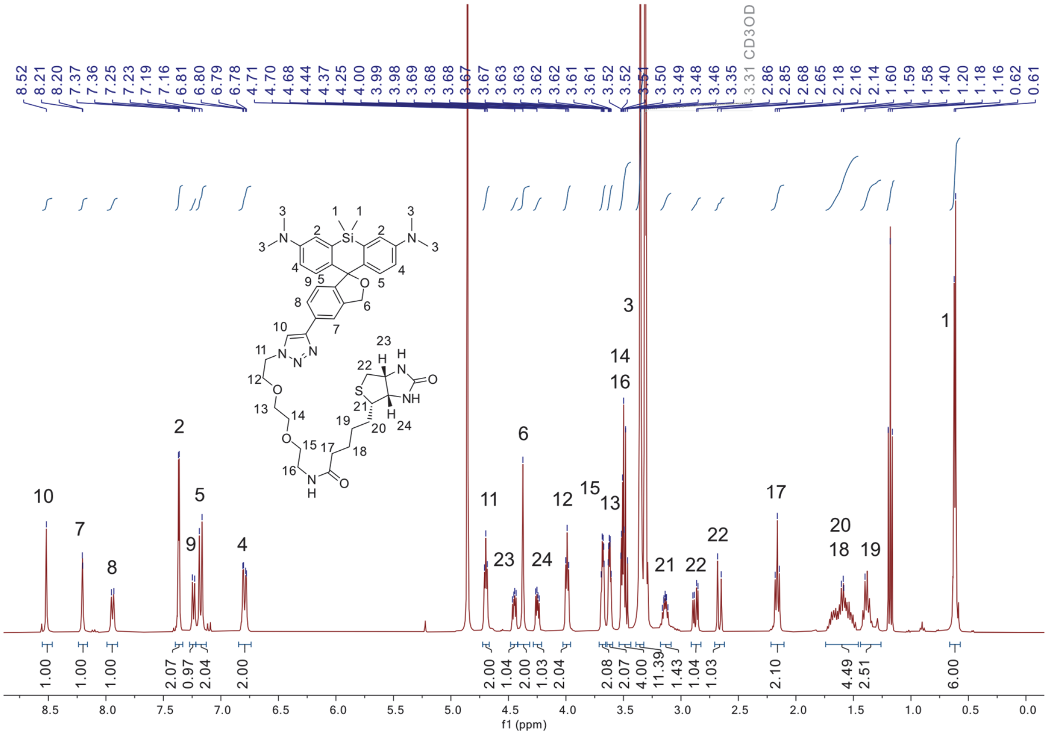


Figure 125. ^{1}H NMR (CD_3OD, 400 MHz) spectrum of compound **33**.

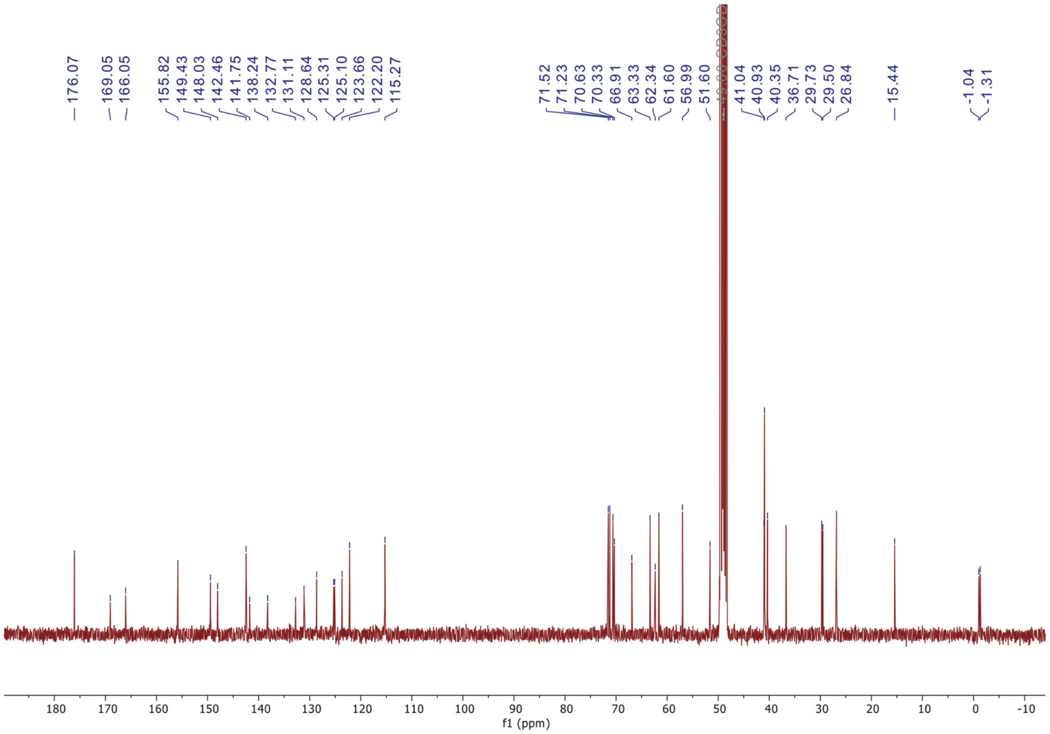


Figure 126. ^{13}C NMR (CD_3OD, 101 MHz) spectrum of compound **33.**

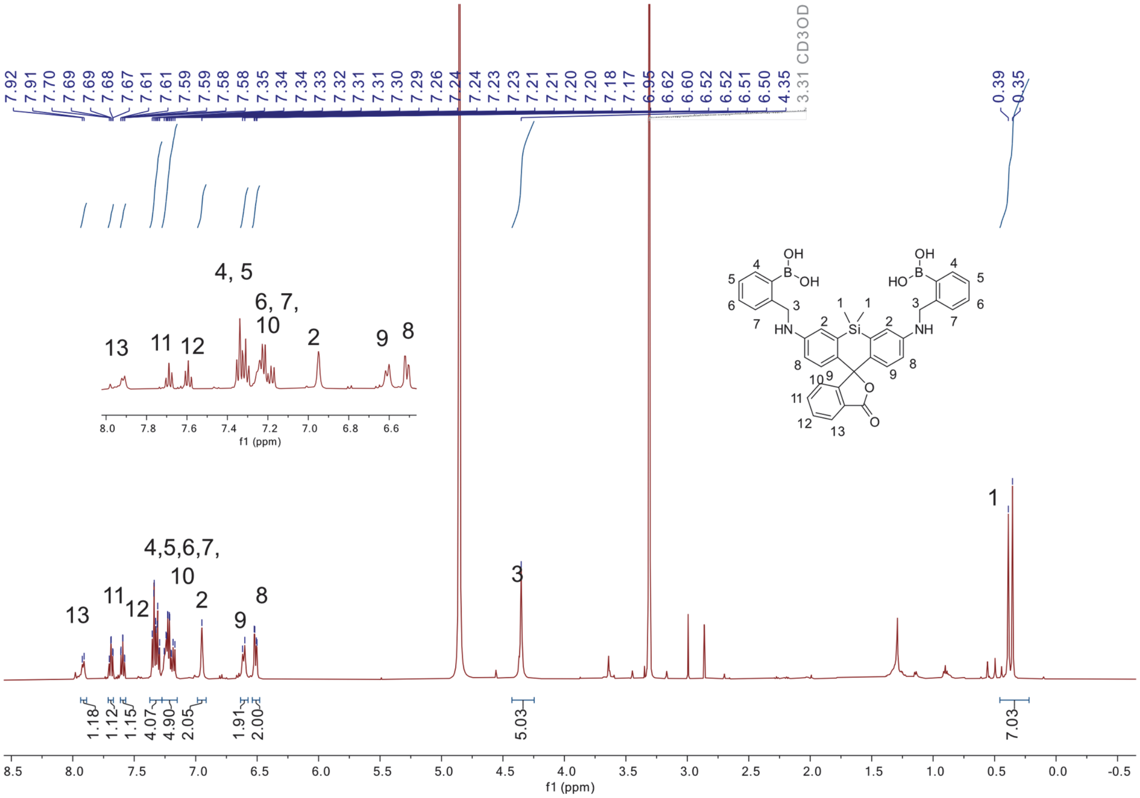


Figure 127. ^{1}H NMR (CD_3OD, 500 MHz) spectrum of compound **36**.

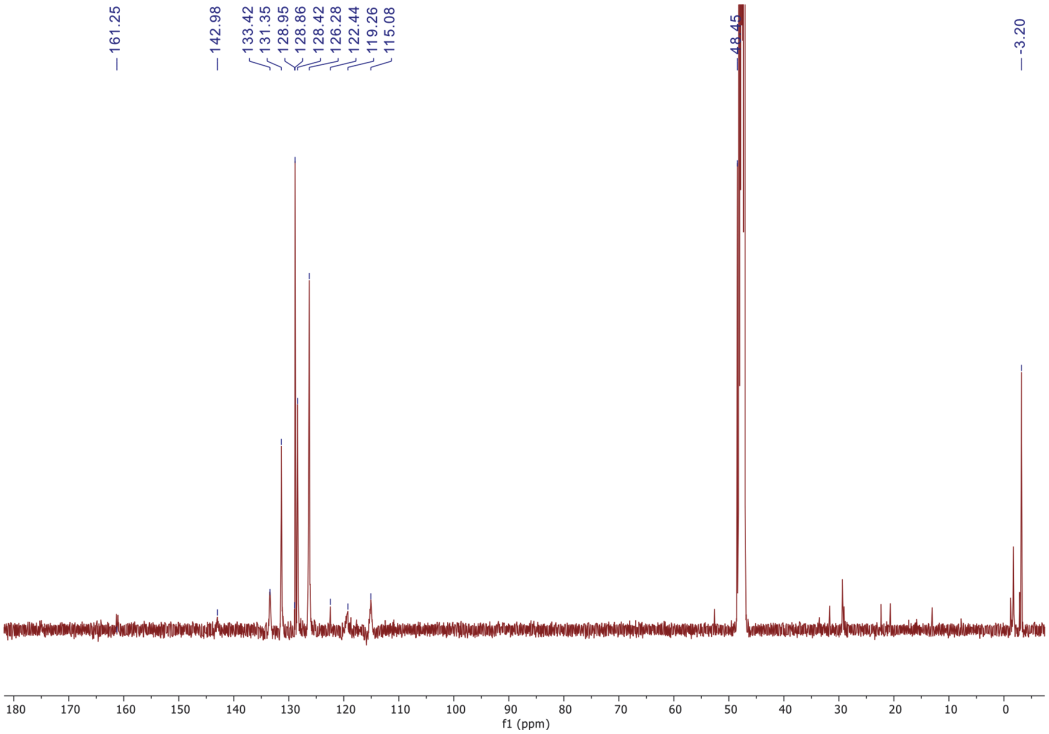


Figure 128. ^{13}C NMR (CD_3OD, 126 MHz) spectrum of compound **36**.

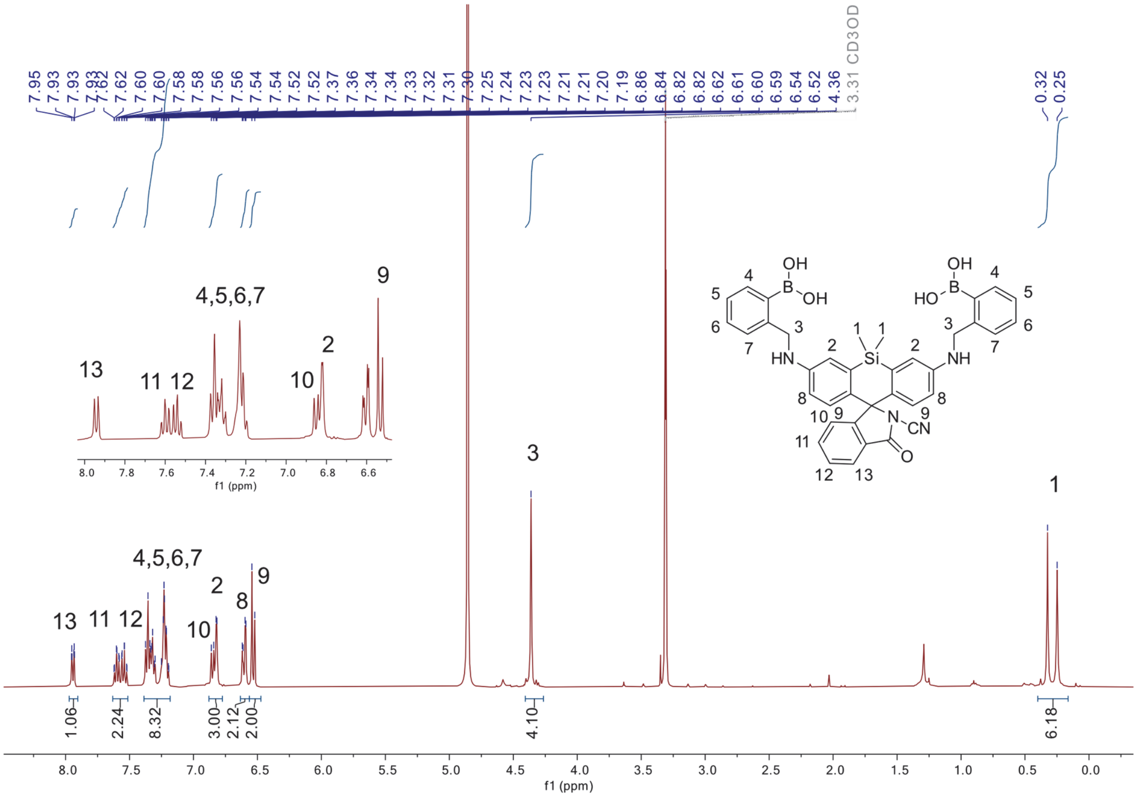


Figure 129. ^{1}H NMR (CD$_3$OD, 500 MHz) spectrum of compound **37**.

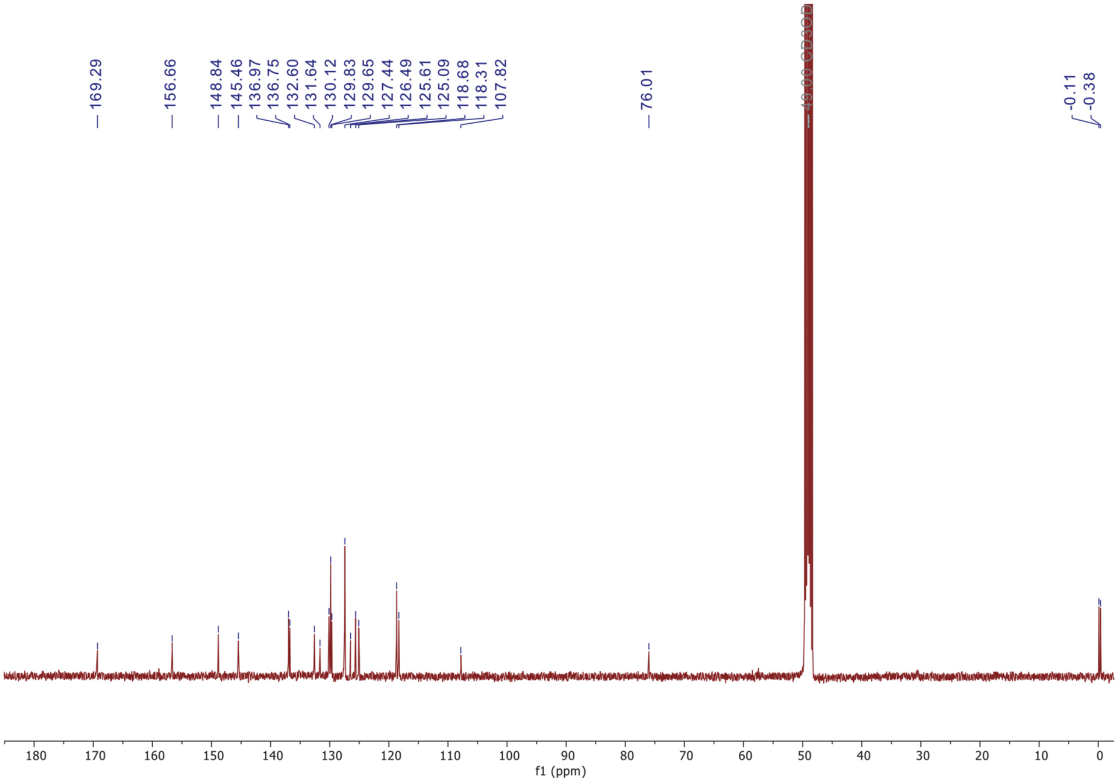


Figure 130. ^{13}C NMR (CD$_3$OD, 126 MHz) spectrum of compound **37**.

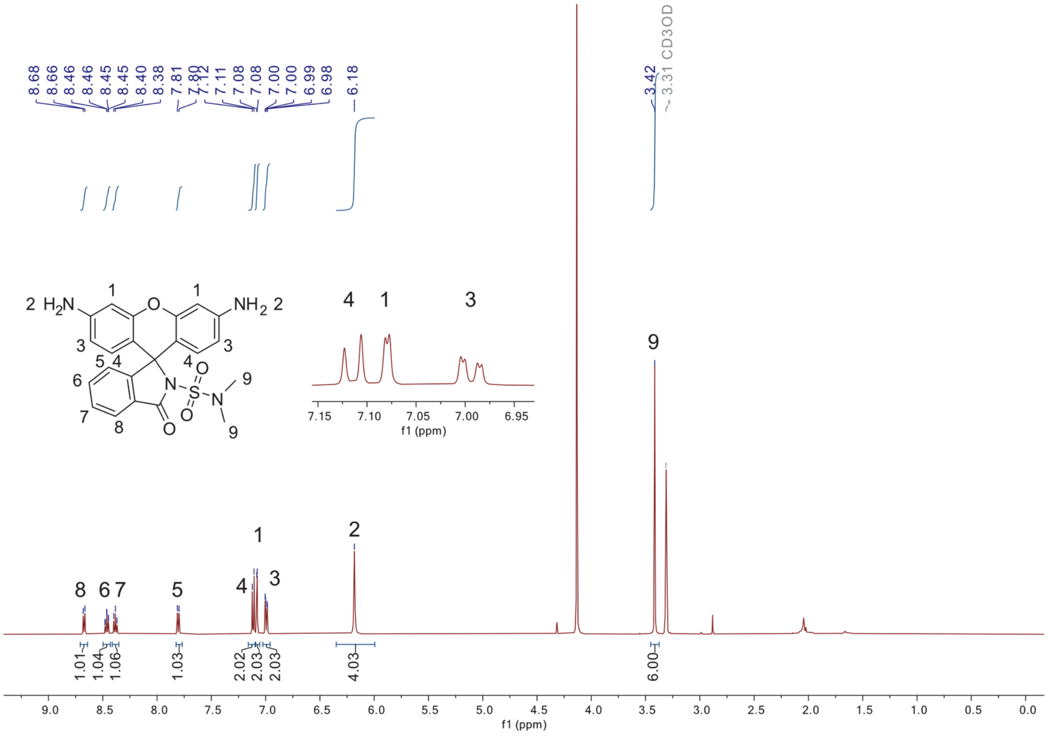


Figure 131. ^{1}H NMR (CD$_3$OD, 500 MHz) spectrum of compound **38**.

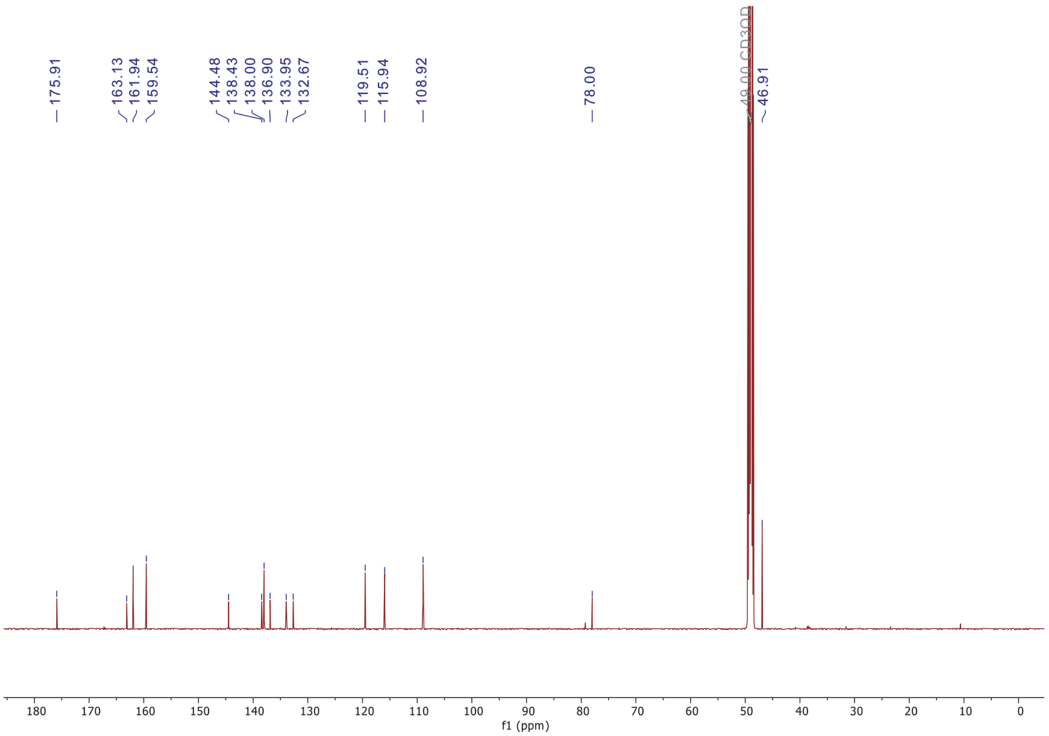


Figure 132. ^{13}C NMR (CD$_3$OD, 126 MHz) spectrum of compound **38.**

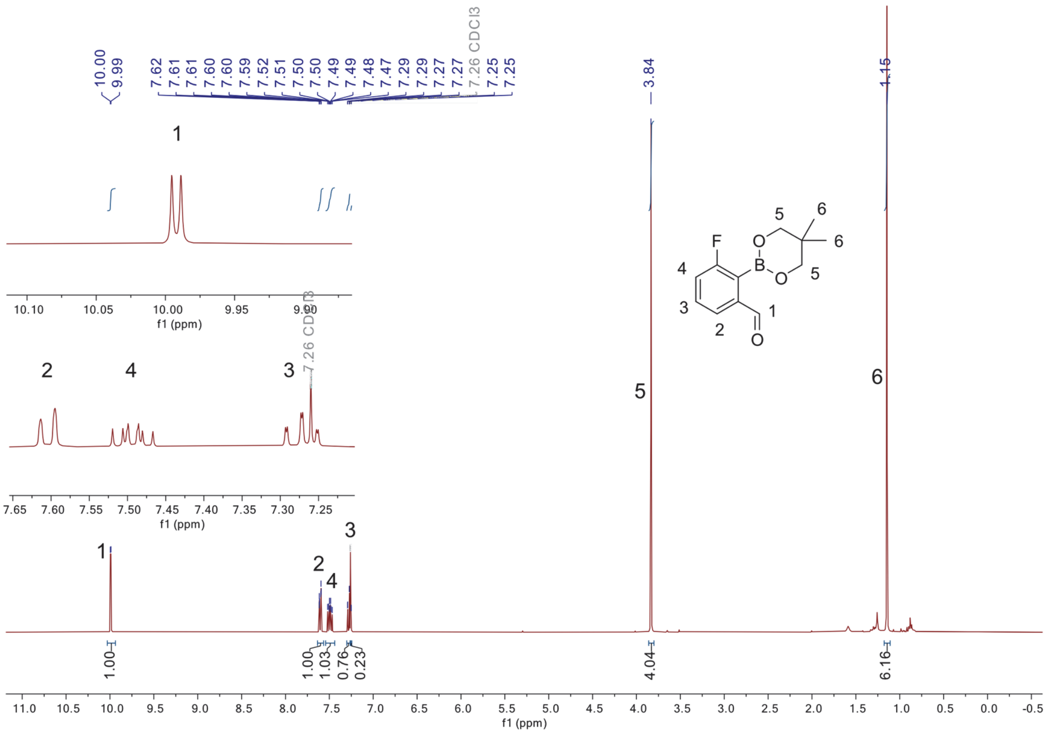


Figure 133. ^{1}H NMR (CDCl$_3$, 400 MHz) spectrum of compound **40**.

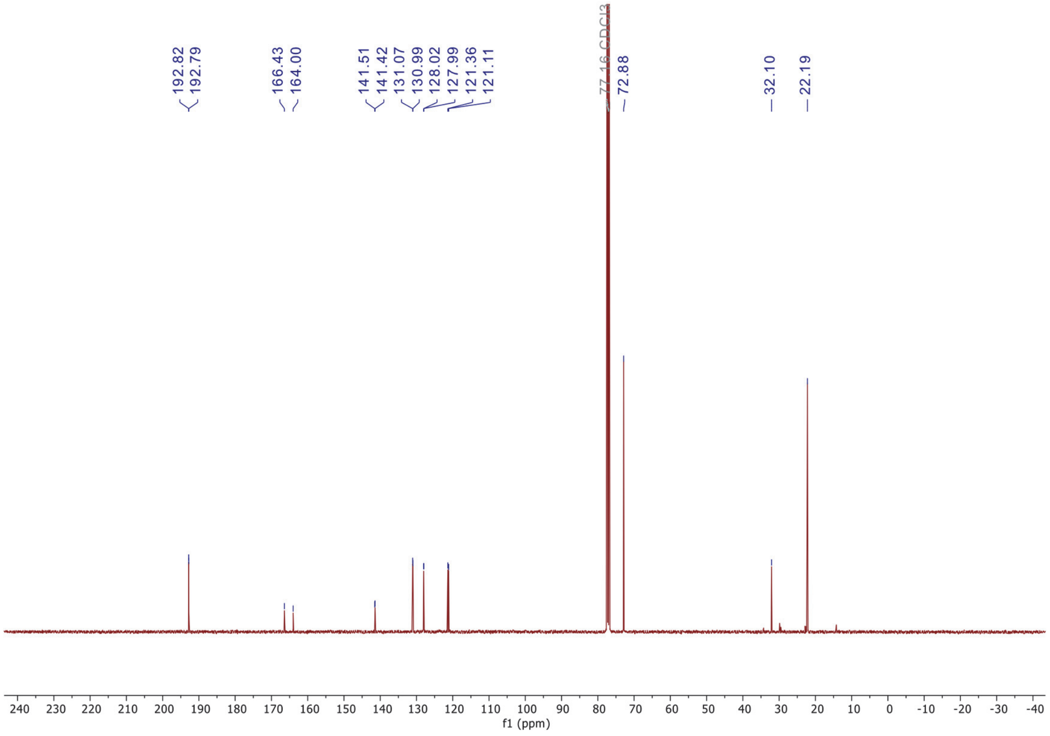


Figure 134. ^{13}C NMR (CDCl$_3$, 100 MHz) spectrum of compound **40.**

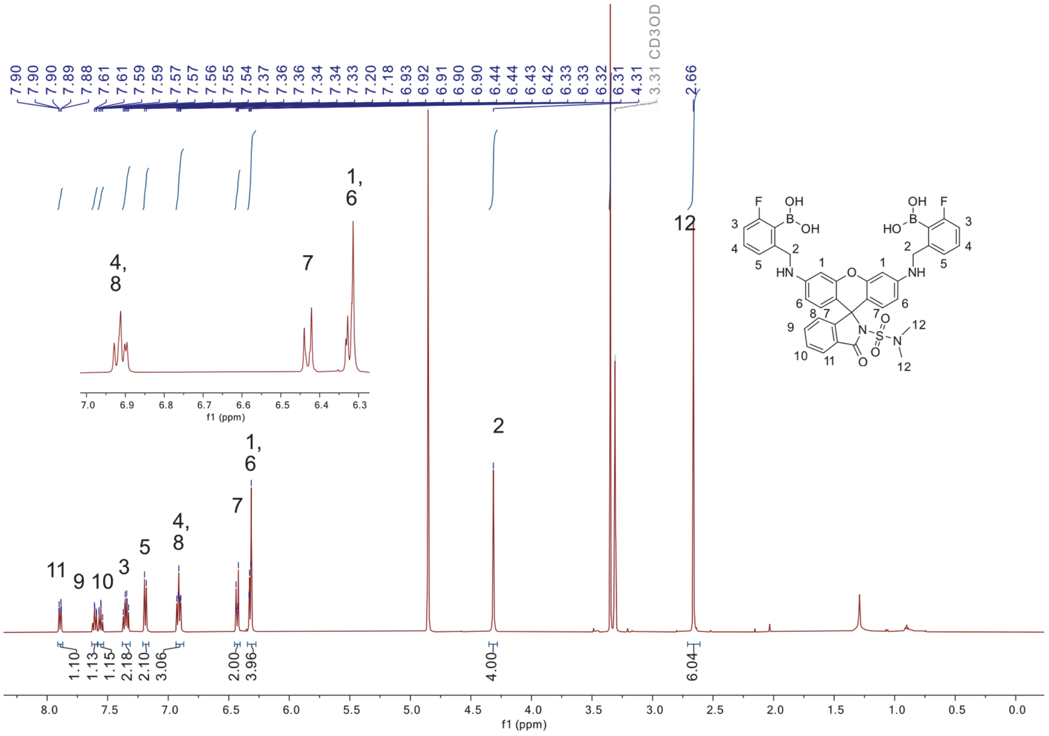


Figure 135. ^{1}H NMR (CD$_3$OD, 400 MHz) spectrum of compound **41**.

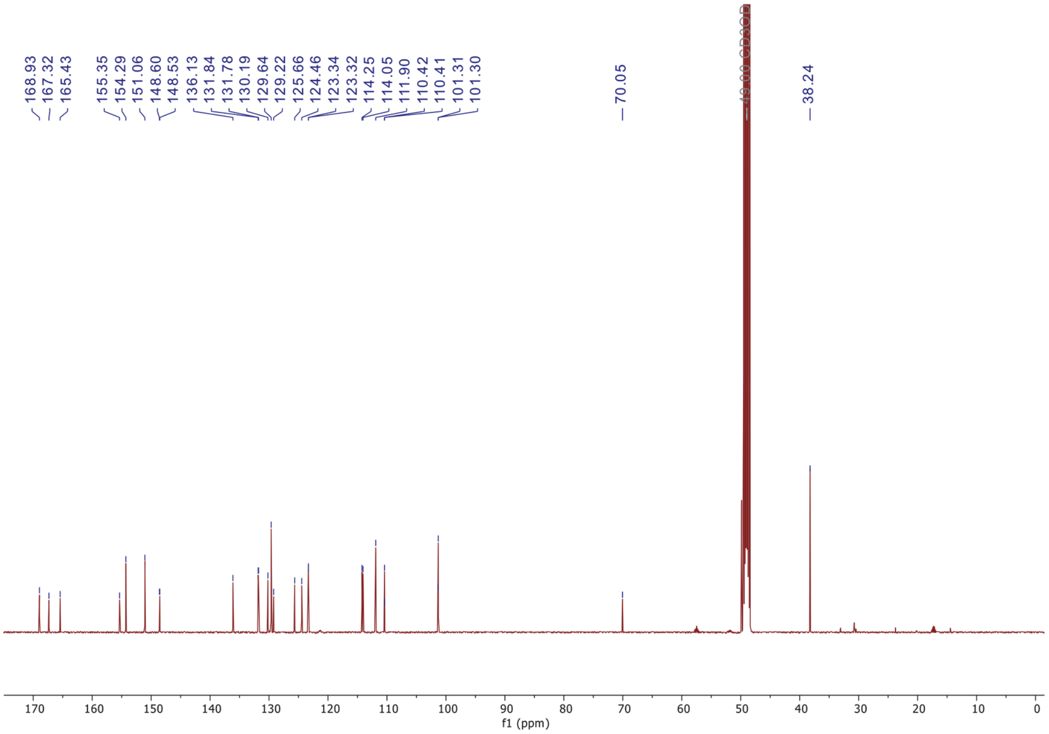


Figure 136. ^{13}C NMR (CD$_3$OD, 100 MHz) spectrum of compound **41.**

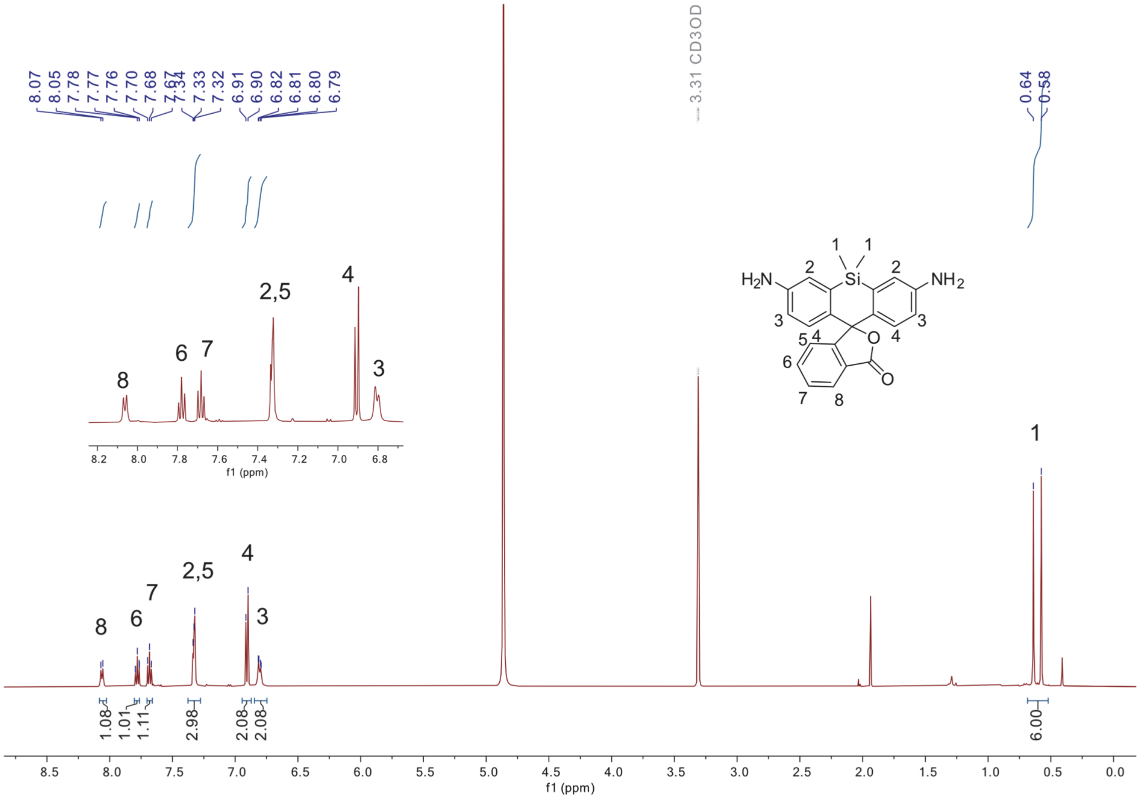


Figure 137. ^{1}H NMR (CD_3OD, 500 MHz) spectrum of compound **42**.

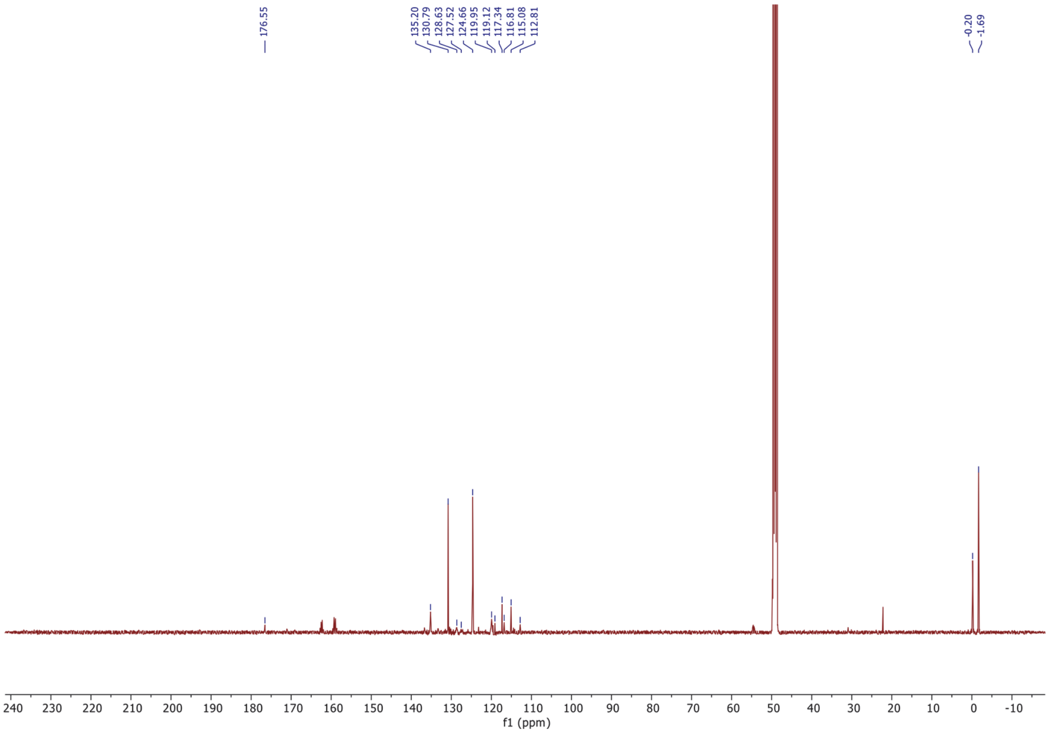


Figure 138. ^{13}C NMR (CD_3OD, 126 MHz) spectrum of compound **42.**

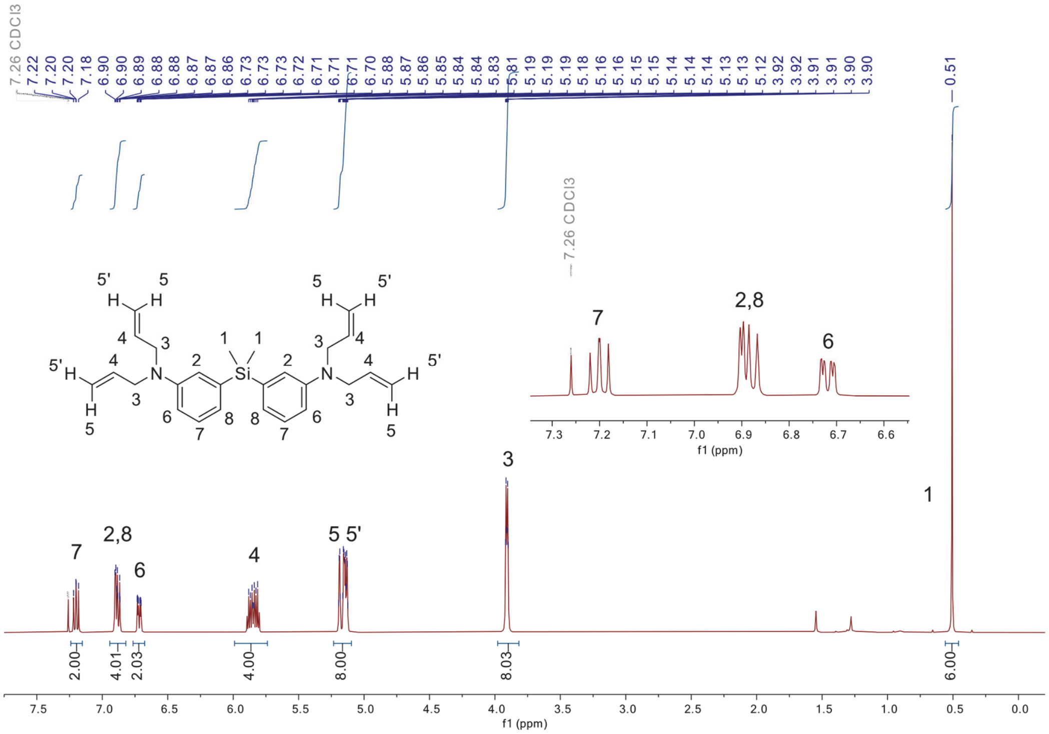


Figure 139. ^{1}H NMR (CD_3Cl, 400 MHz) spectrum of compound **44.**

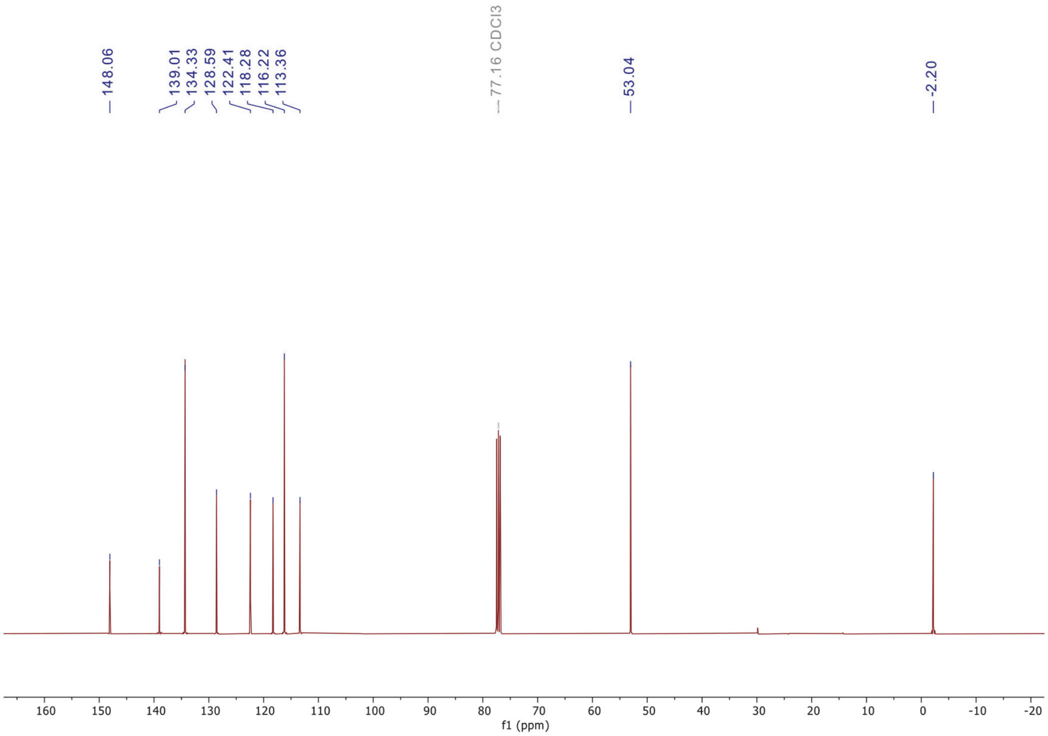


Figure 140. ^{13}C NMR (CD_3Cl, 101 MHz) spectrum of compound **44.**

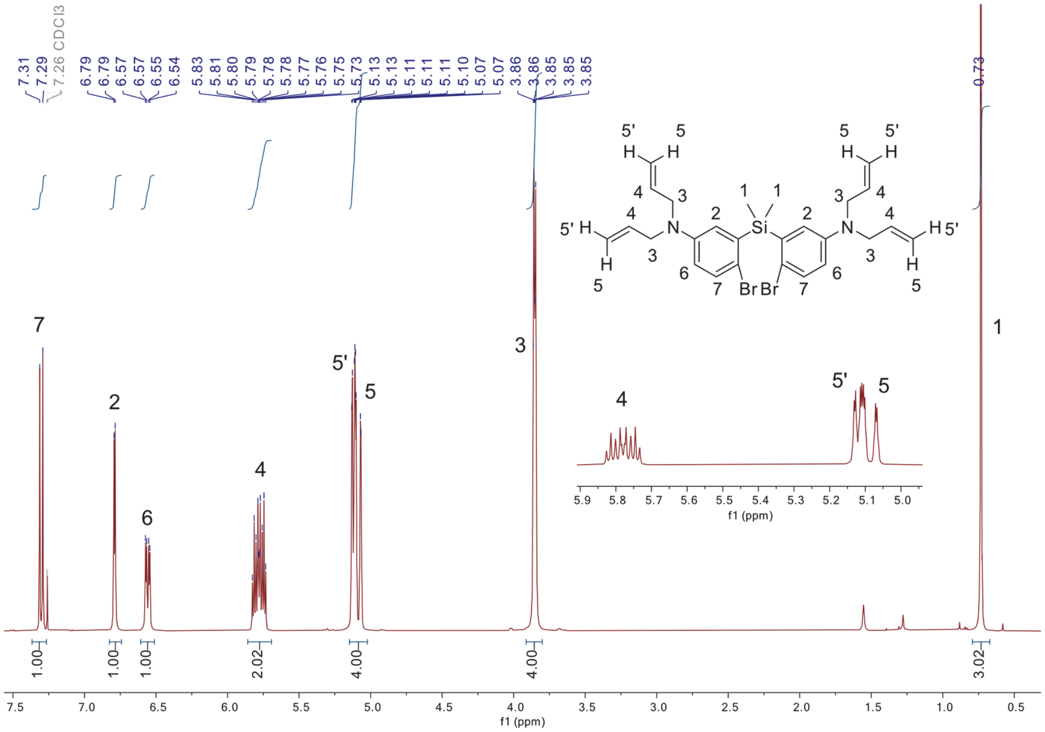


Figure 141. ^{1}H NMR (CD_3Cl, 400 MHz) spectrum of compound **45.**

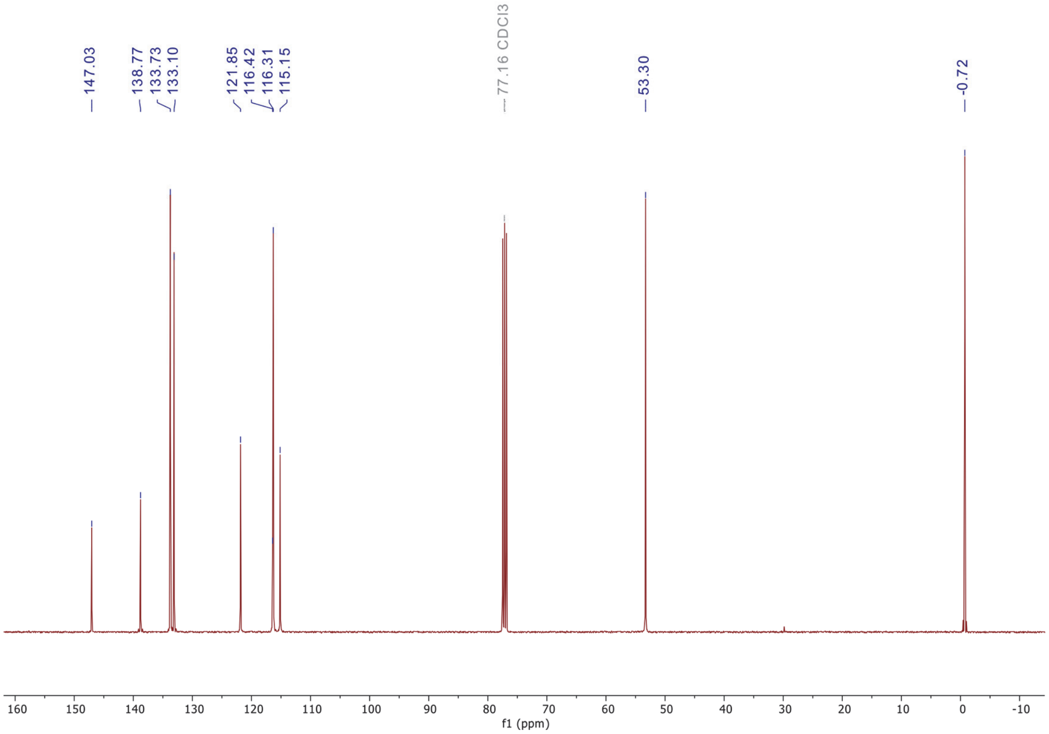


Figure 142. ^{13}C NMR (CD_3Cl, 101 MHz) spectrum of compound **45.**

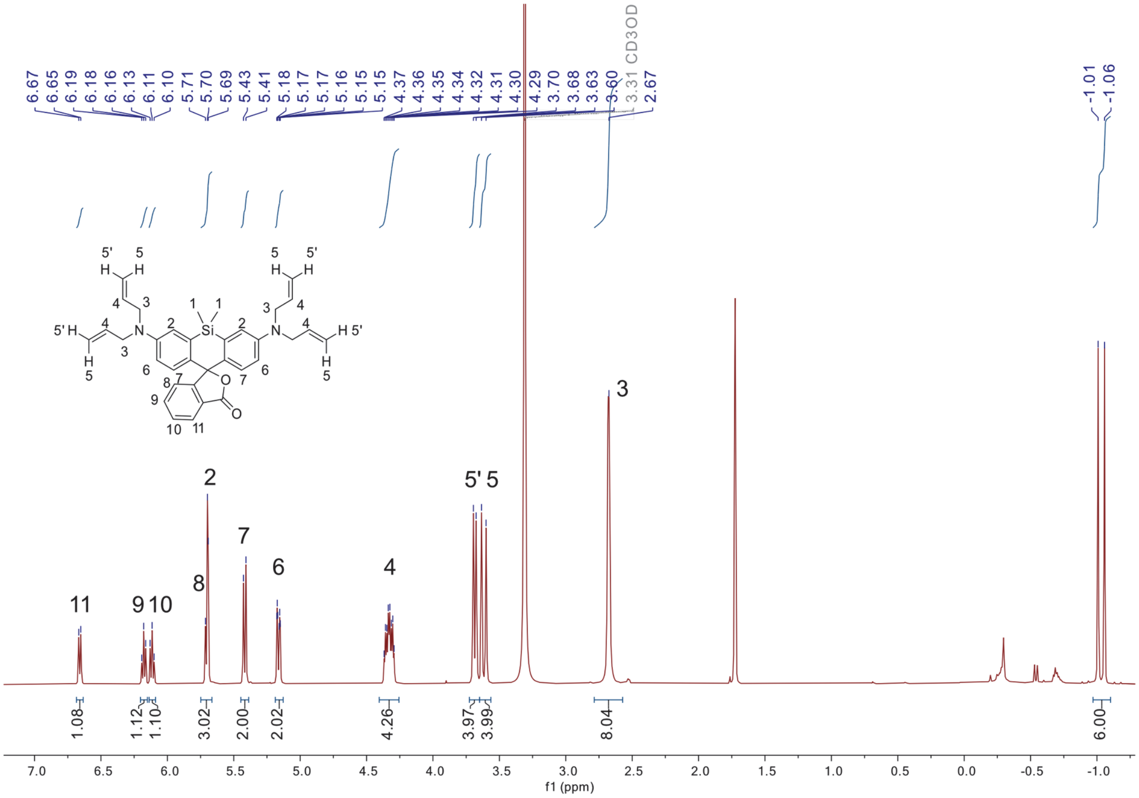


Figure 143. ^{1}H NMR (CD_3OD, 500 MHz) spectrum of compound **46.**

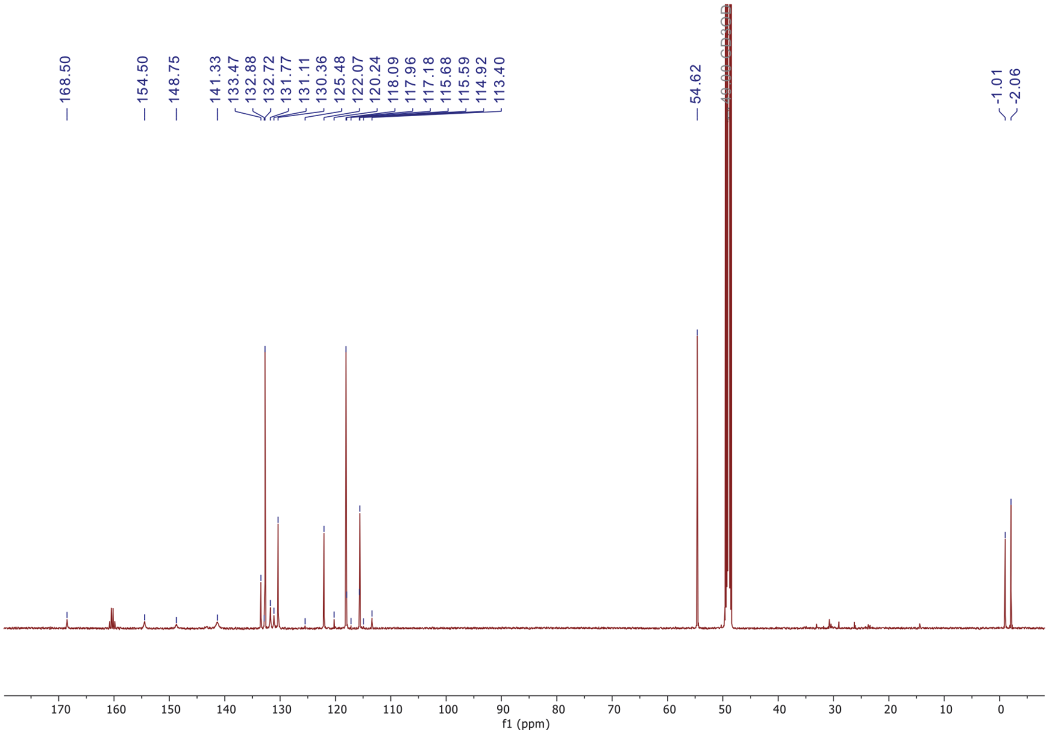


Figure 144. ^{13}C NMR (CD_3Cl, 126 MHz) spectrum of compound **46.**

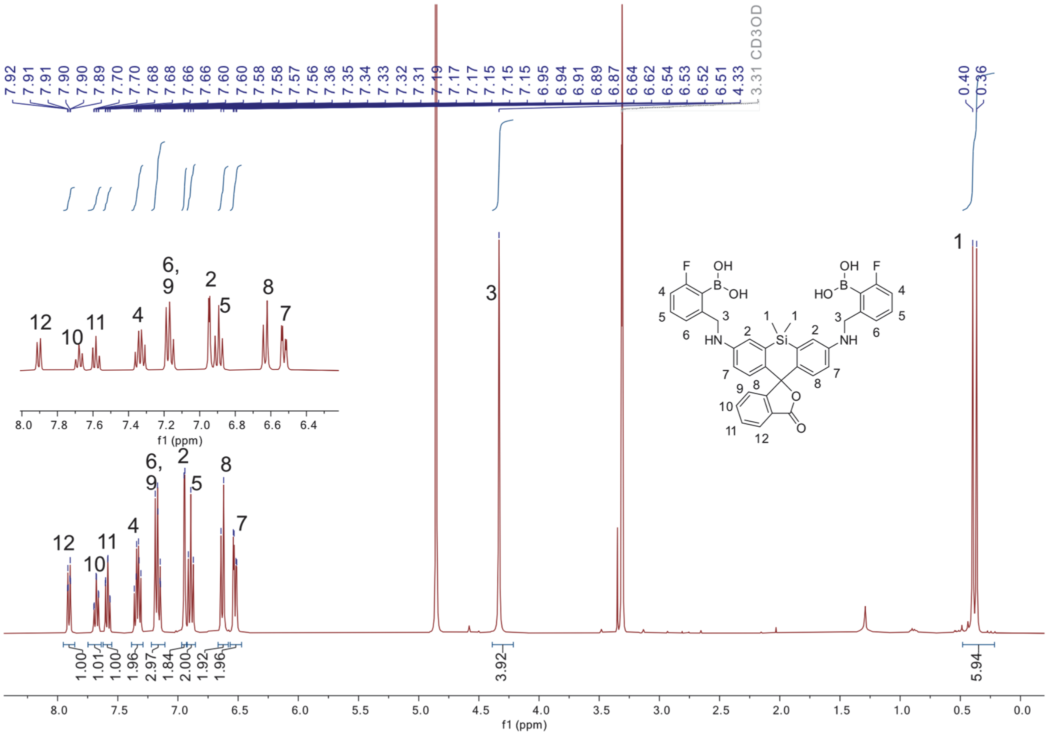


Figure 145. ^{1}H NMR (CD$_3$OD, 400 MHz) spectrum of compound **47**.

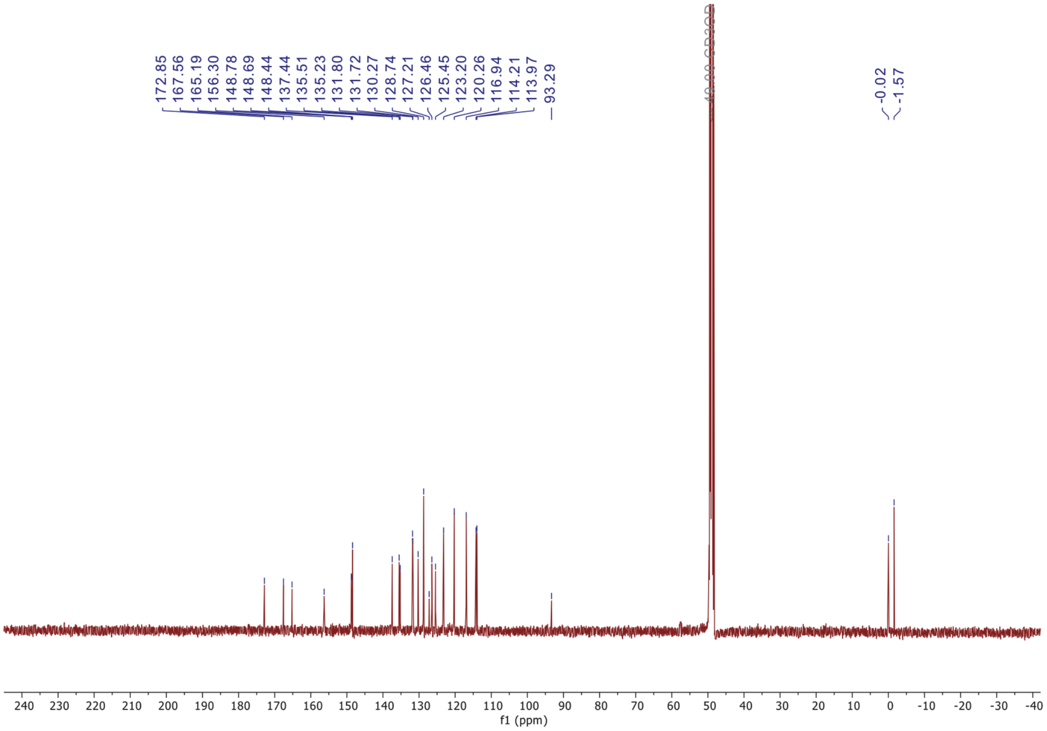


Figure 146. ^{13}C NMR (CD$_3$OD, 100 MHz) spectrum of compound **47.**

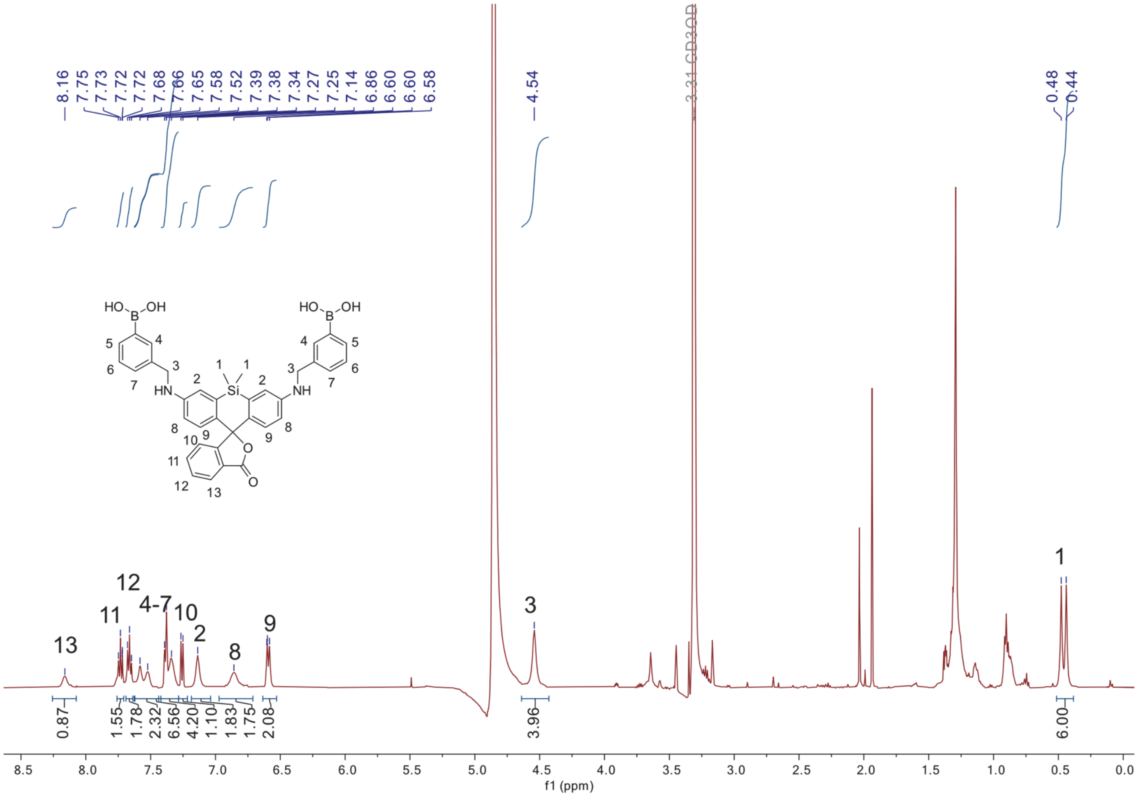


Figure 147. ^{1}H NMR (CD_3OD, 500 MHz) spectrum of compound **48**.

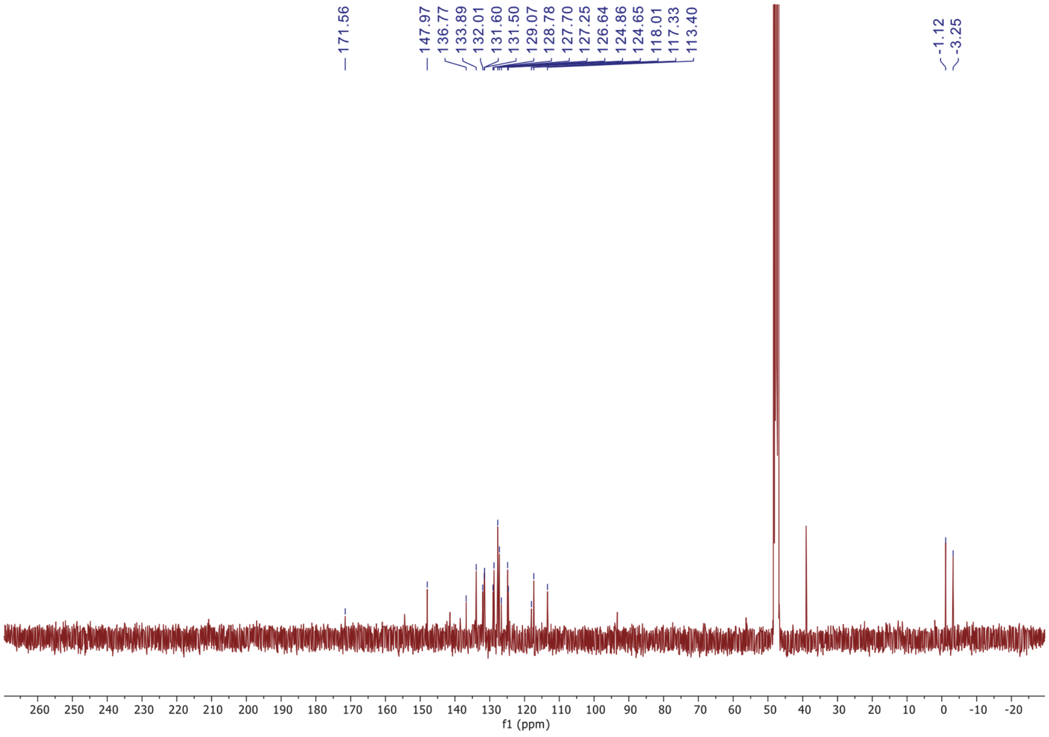


Figure 148. ^{13}C NMR (CD_3OD, 126 MHz) spectrum of compound **48.**

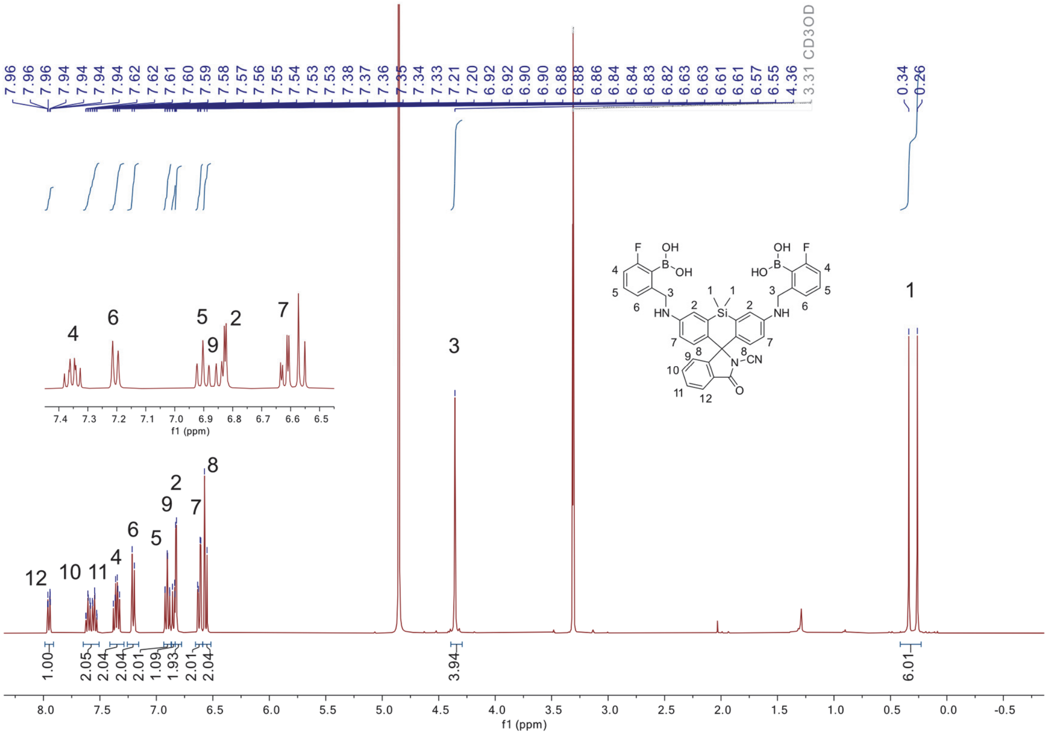


Figure 149. ^{1}H NMR (CD_3OD, 400 MHz) spectrum of compound **49**.

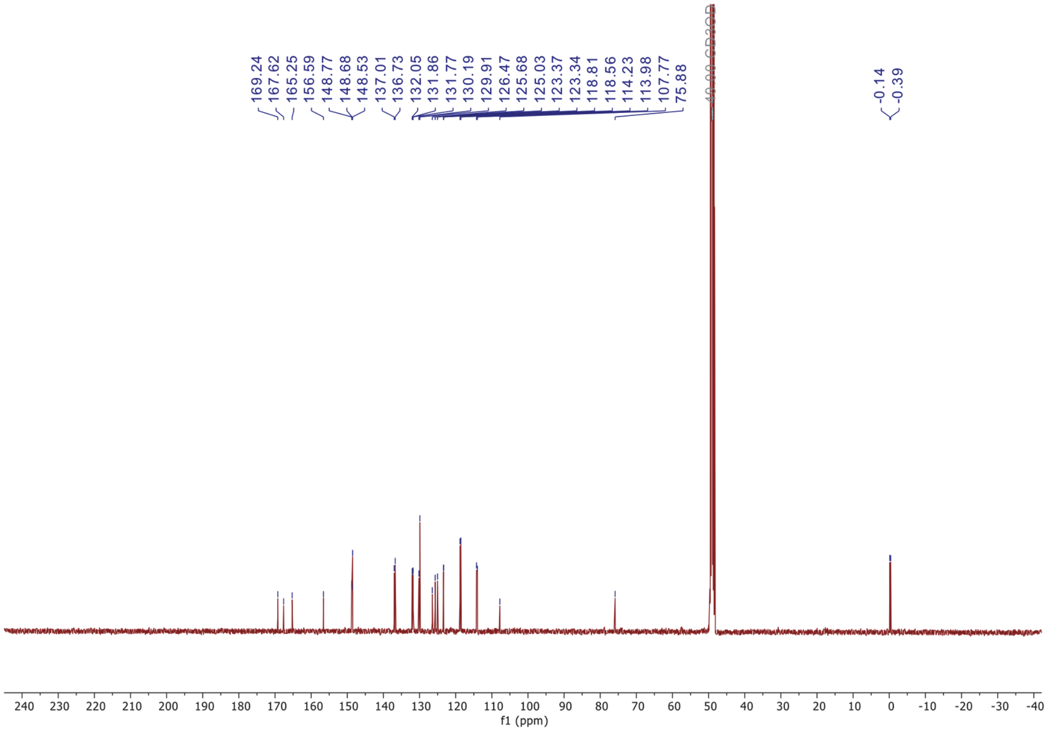


Figure 150. ^{13}C NMR (CD_3OD, 100 MHz) spectrum of compound **49.**

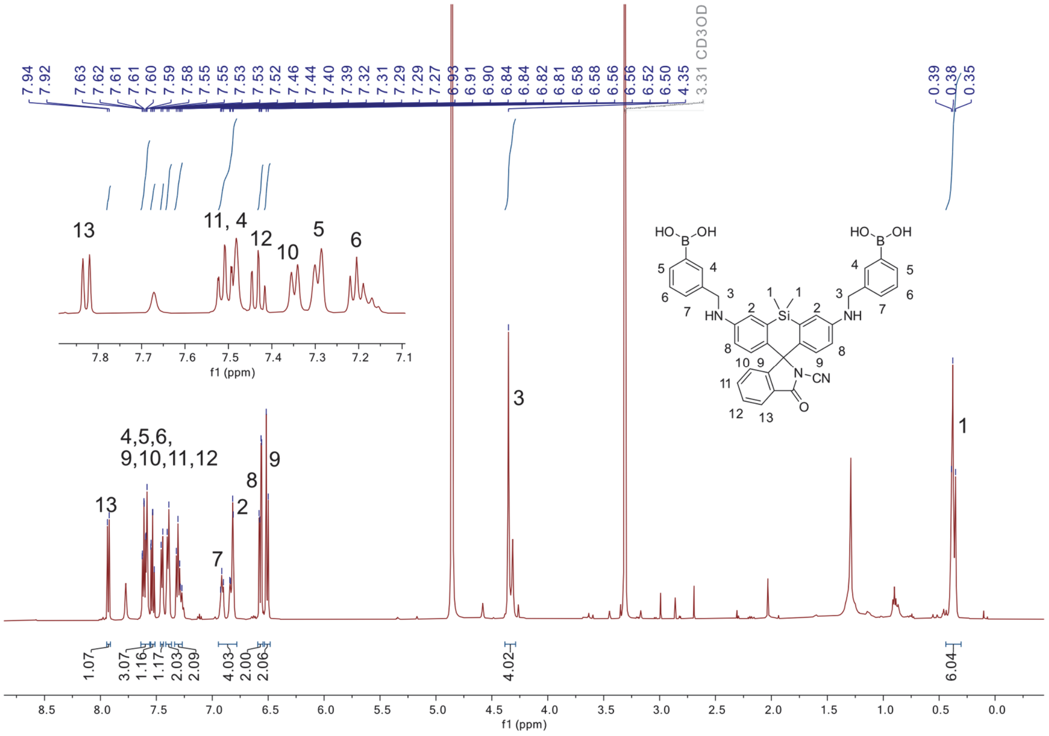


Figure 151. ^{1}H NMR (CD$_3$OD, 400 MHz) spectrum of compound **50**.

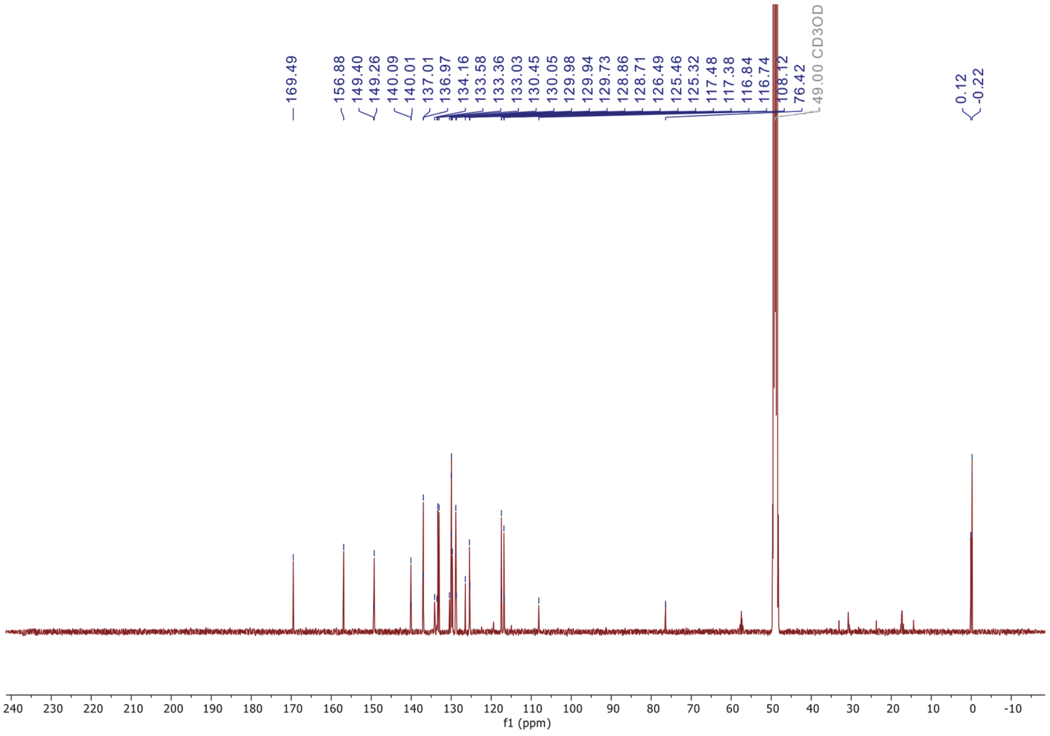


Figure 152. ^{13}C NMR (CD$_3$OD, 100 MHz) spectrum of compound **50.**

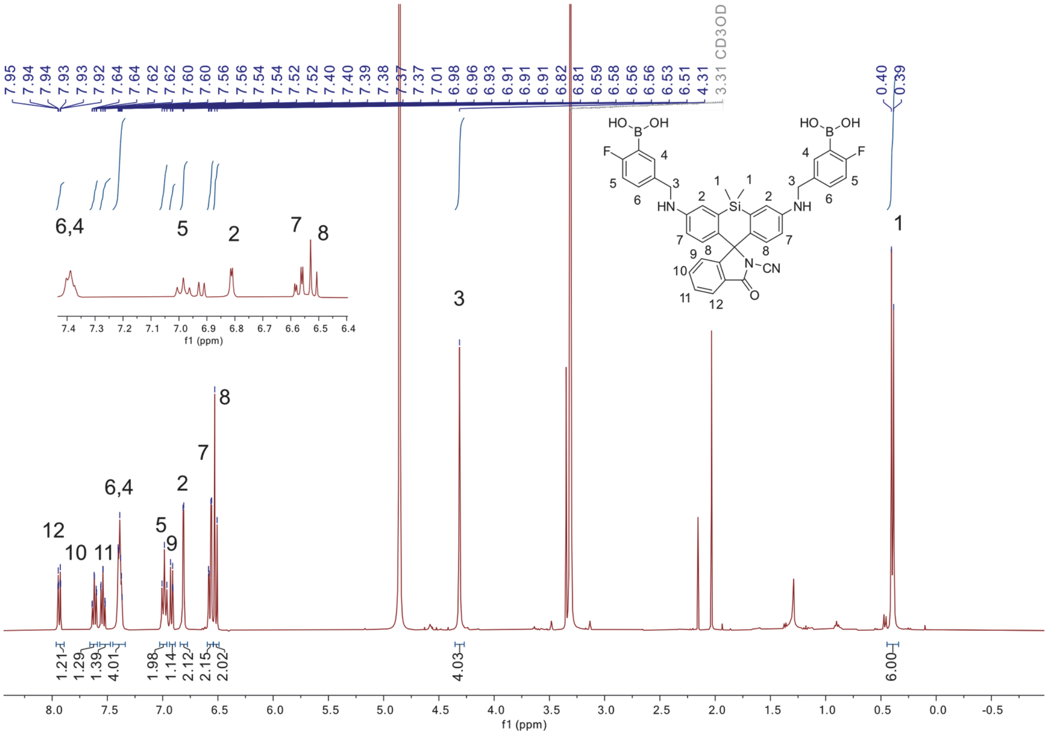


Figure 153. ^{1}H NMR (CD$_3$OD, 400 MHz) spectrum of compound **51**.

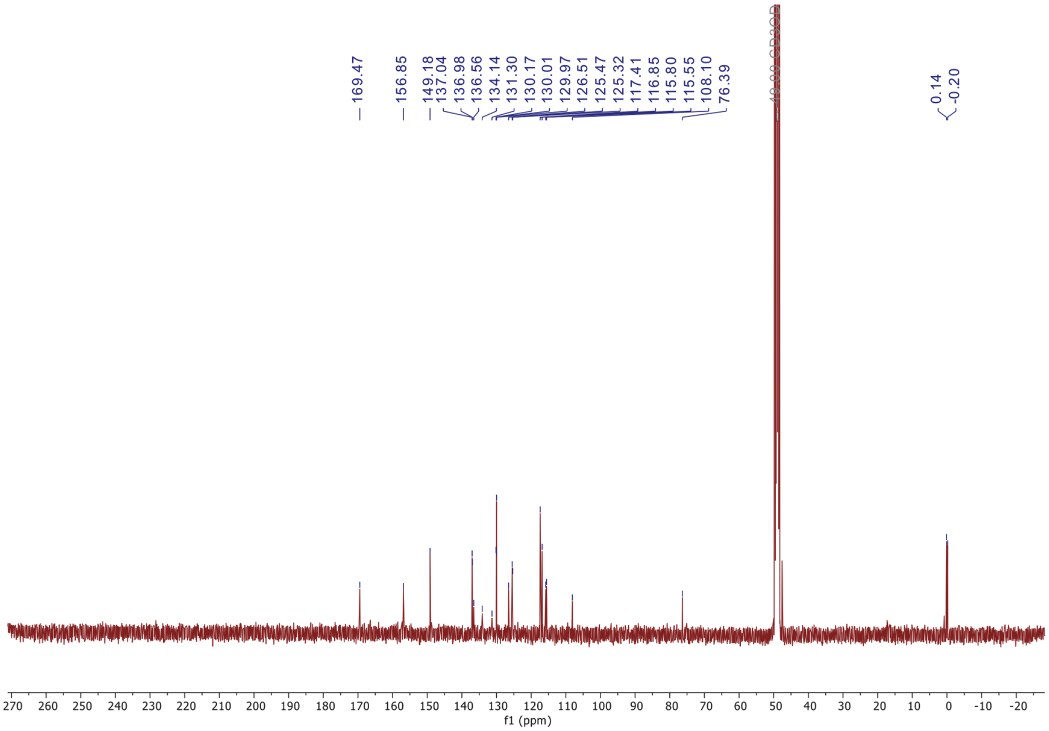


Figure 154. ^{13}C NMR (CD$_3$OD, 100 MHz) spectrum of compound **51**.

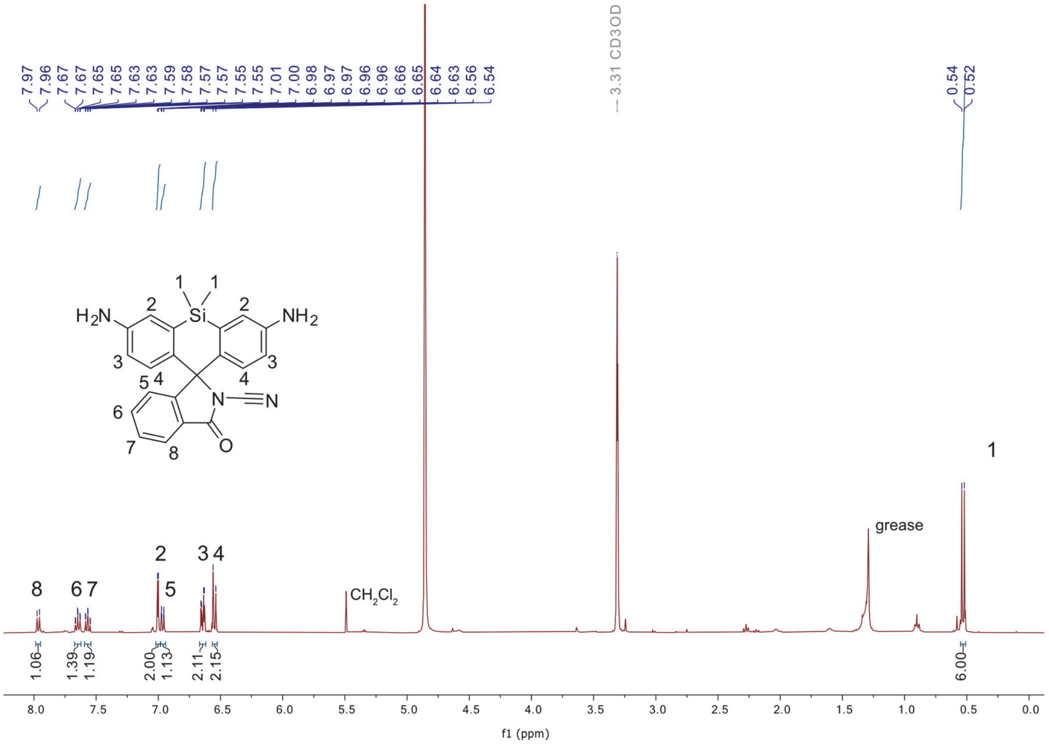


Figure 155. ^{1}H NMR (CD_3OD, 400 MHz) spectrum of compound **53**.

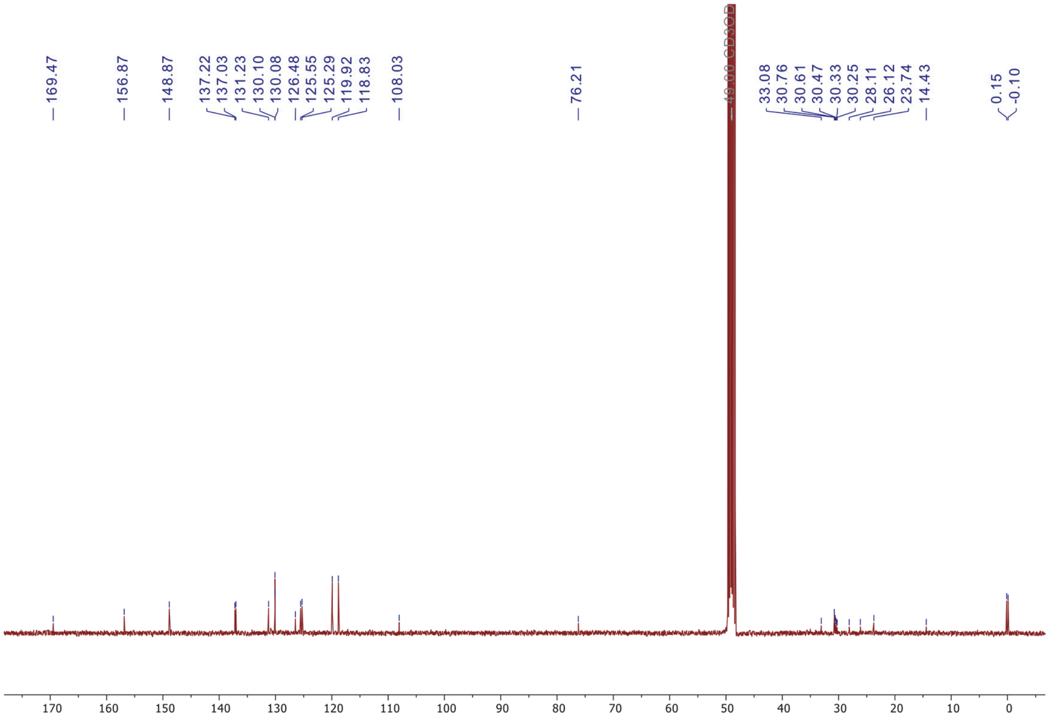
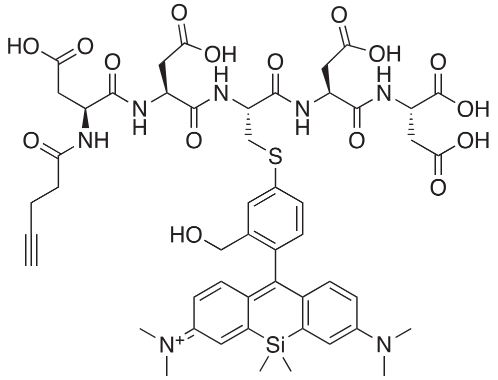


Figure 156. ^{13}C NMR (CD_3OD, 101 MHz) spectrum of compound **53**.

8.2.2 Peptides characterization



Pentynyl-DDC(HMSiR)DD-OH

HRMS (ESI/QTOF) $[M+2H]^{2+}$ calcd. for $[C_{50}H_{61}N_7O_{16}SSi]^{+2}$: 537.6827; found 537.6832.

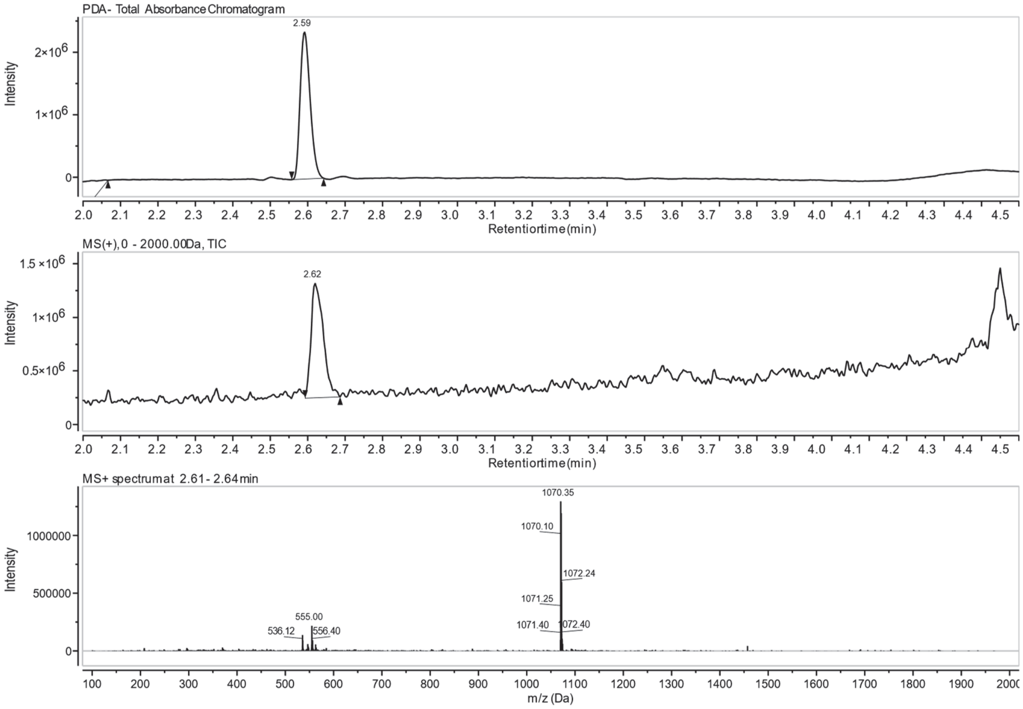
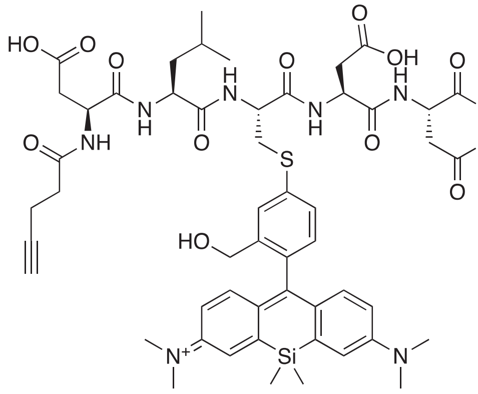


Figure 157. Structure, condensed formula, exact mass and LC-MS traces and extracted peak of peptide **C11**.



Pentynyl-DLC(HMSiR)DD-OH

HRMS (ESI/QTOF) $[M+2H]^{2+}$ calcd. for $[C_{52}H_{67}N_7O_{14}SSi]^{+2}$: 536.7113; found 536.7108.

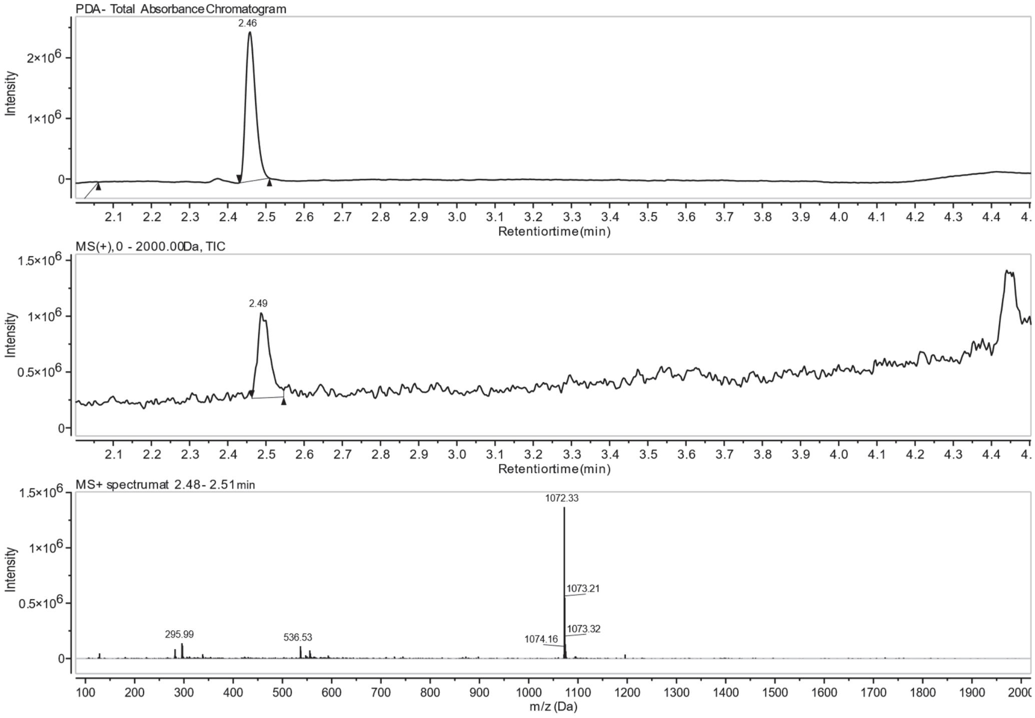
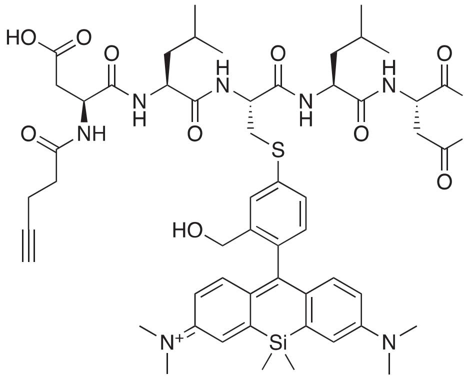


Figure 158. Structure, condensed formula, exact mass and LC-MS traces and extracted peak of peptide **C12.**



Pentynyl-DLC(HMSiR)LD-OH

HRMS (ESI/QTOF) $[M+2H]^{2+}$ calcd. for $[C_{54}H_{73}N_7O_{12}SSi]^{+2}$: 535.7398; found 535.7404.

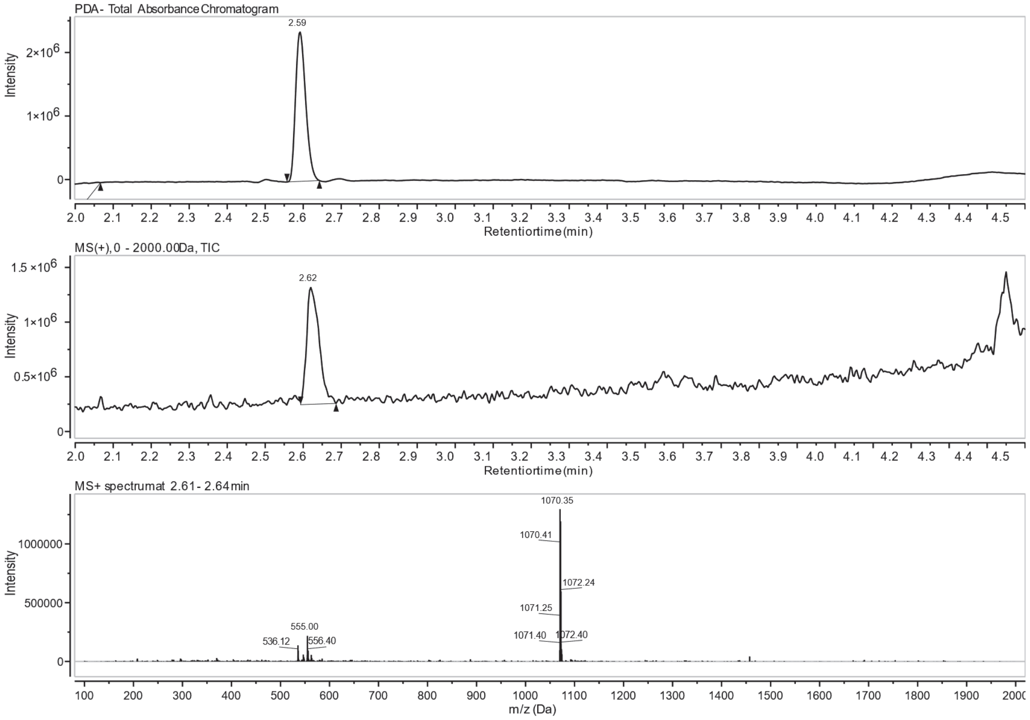
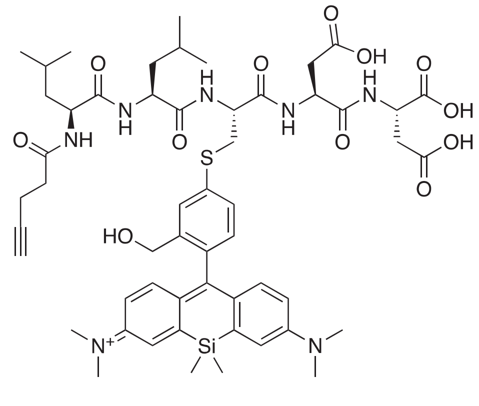


Figure 159. Structure, condensed formula, exact mass and LC-MS traces and extracted peak of peptide **C13**.



Pentynyl-LLC(HMSiR)DD-OH

HRMS (ESI/QTOF) $[M+2H]^{2+}$ calcd. for $[C_{54}H_{73}N_7O_{12}SSi]^{+2}$: 535.7398, found: 535.7408.

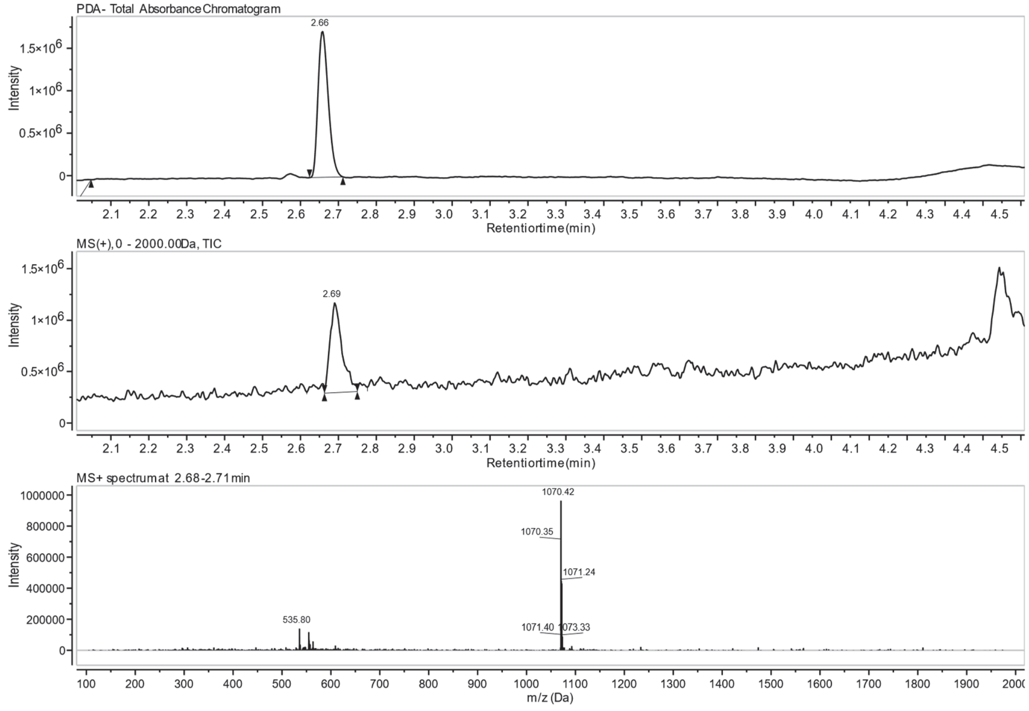
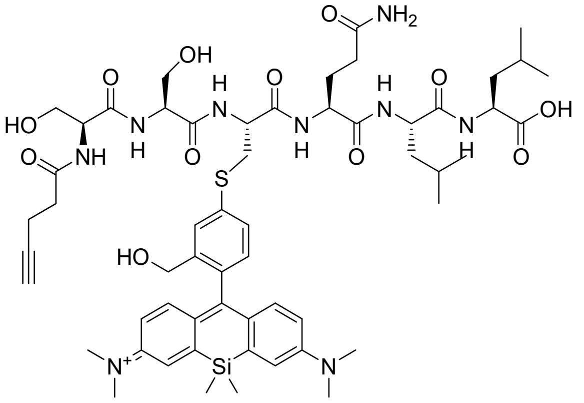


Figure 160. Structure, condensed formula, exact mass and LC-MS traces and extracted peak of peptide **C14.**



Pentynyl-SSC(HMSiR)QLL-OH

HRMS (ESI/QTOF) $[M+H]^+$ calcd. for $[C_{57}H_{80}N_9O_{12}SSi]^+$: 1142.5416, found: 1142.5408.

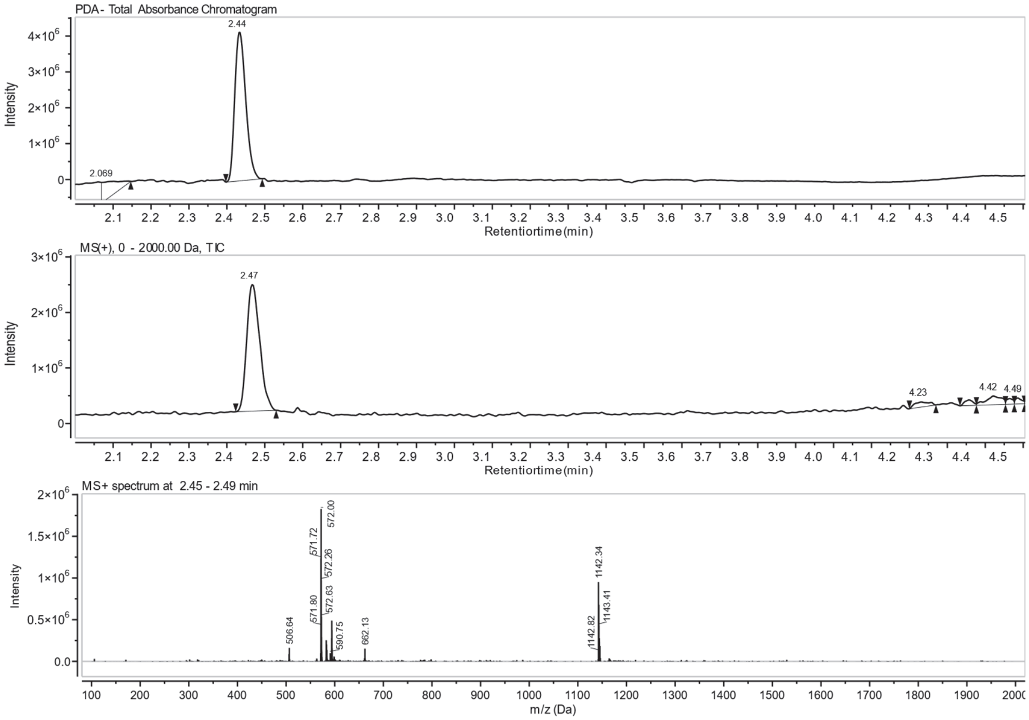
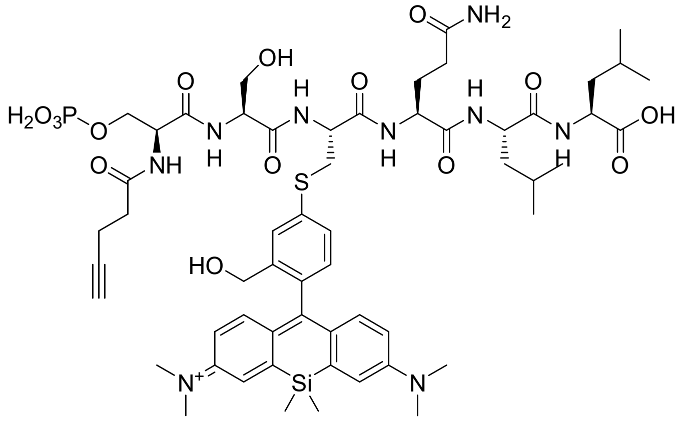


Figure 161. Structure, condensed formula, exact mass and LC-MS traces and extracted peak of peptide **P15**.



Pentynyl-pSSC(HMSiR)QLL-OH

HRMS (ESI/QTOF) $[M+2H]^{2+}$ calcd. for $[C_{57}H_{81}N_9NaO_{15}PSSi]^{+2}$: 622.7483; Found 622.7492.

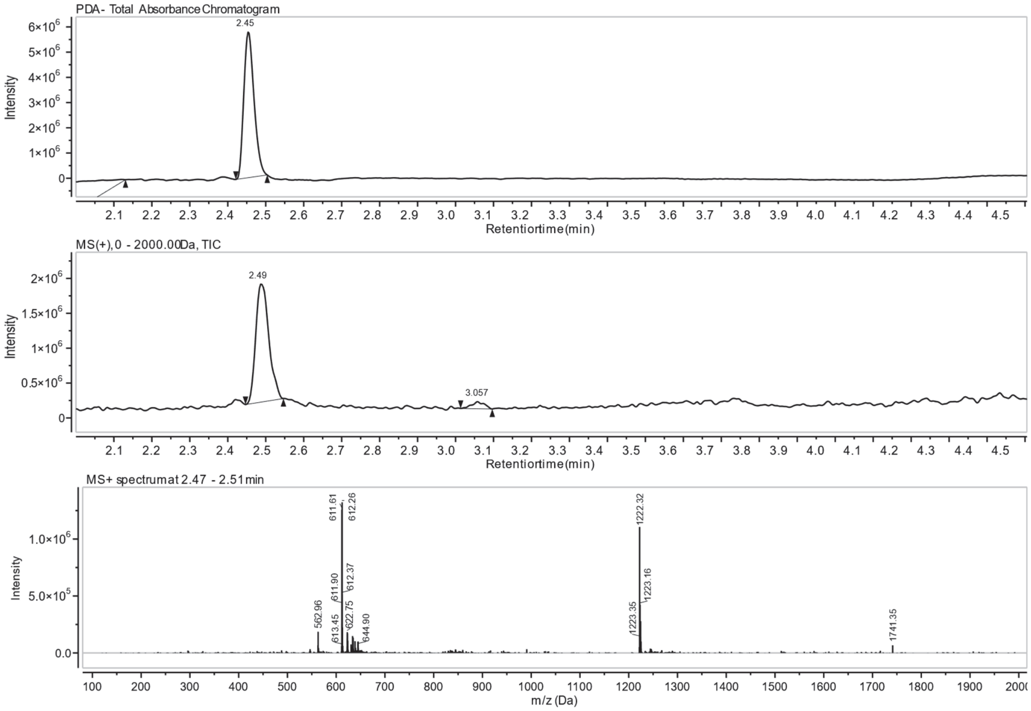
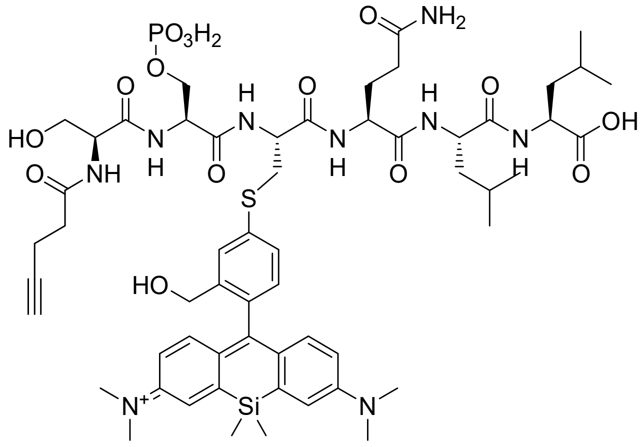


Figure 162. Structure, condensed formula, exact mass and LC-MS traces and extracted peak of peptide **P16**.



Pentynyl-SpSC(HMSiR)QLL-OH

HRMS (ESI/QTOF) $[M+2H]^{2+}$ calcd. for $[C_{57}H_{82}N_9O_{15}PSSi]^{+2}$: 611.7574, found: 611.7597.

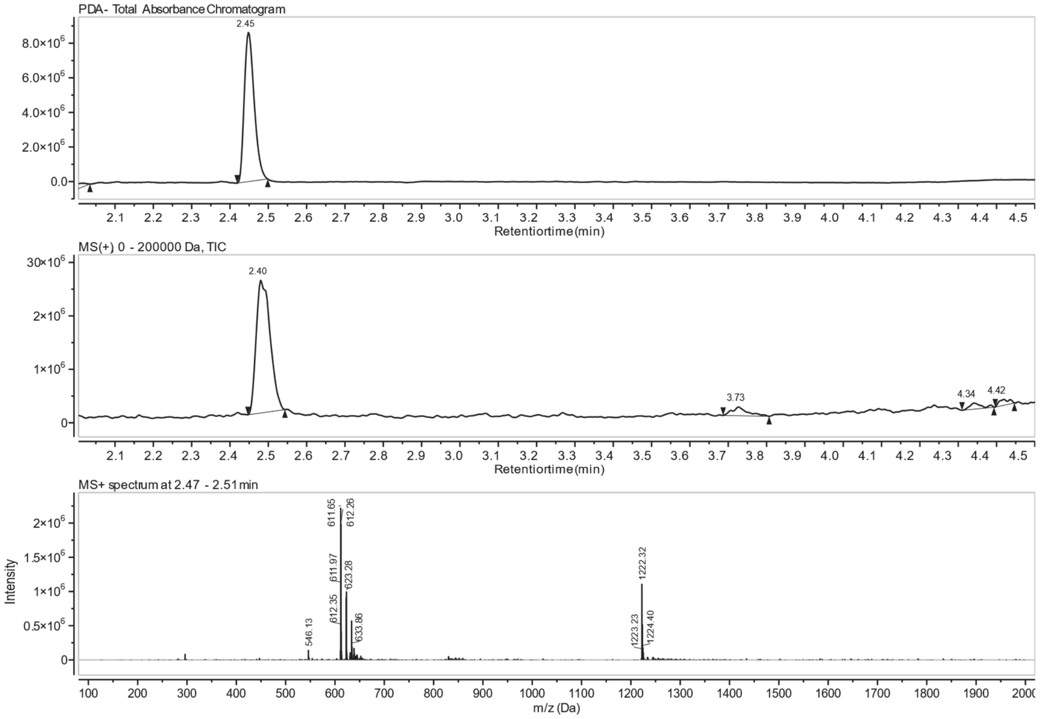
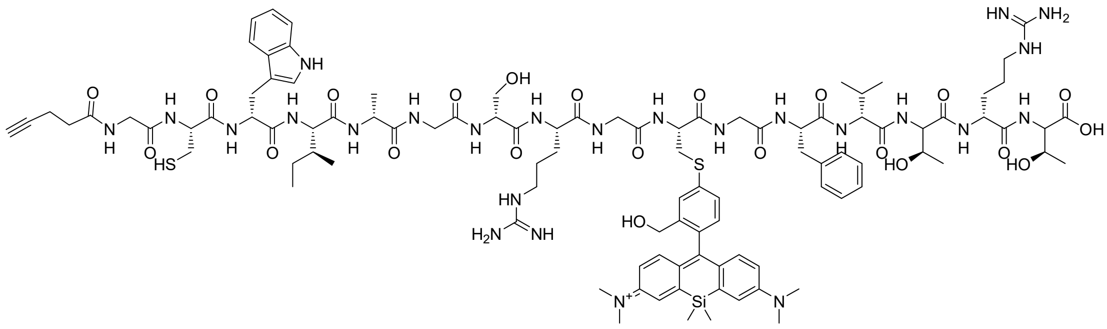


Figure 163. Structure, condensed formula, exact mass and LC-MS traces and extracted peak of peptide **P17**.



Pentynyl-GCWIAGSRGC(HMSiR)GFVTRT-OH

HRMS (ESI/QTOF) $[M+3H]^{3+}$ calcd. for $[C_{102}H_{147}N_{25}O_{22}S_2Si]^{+3}$: 722.0116; found: 722.0130.

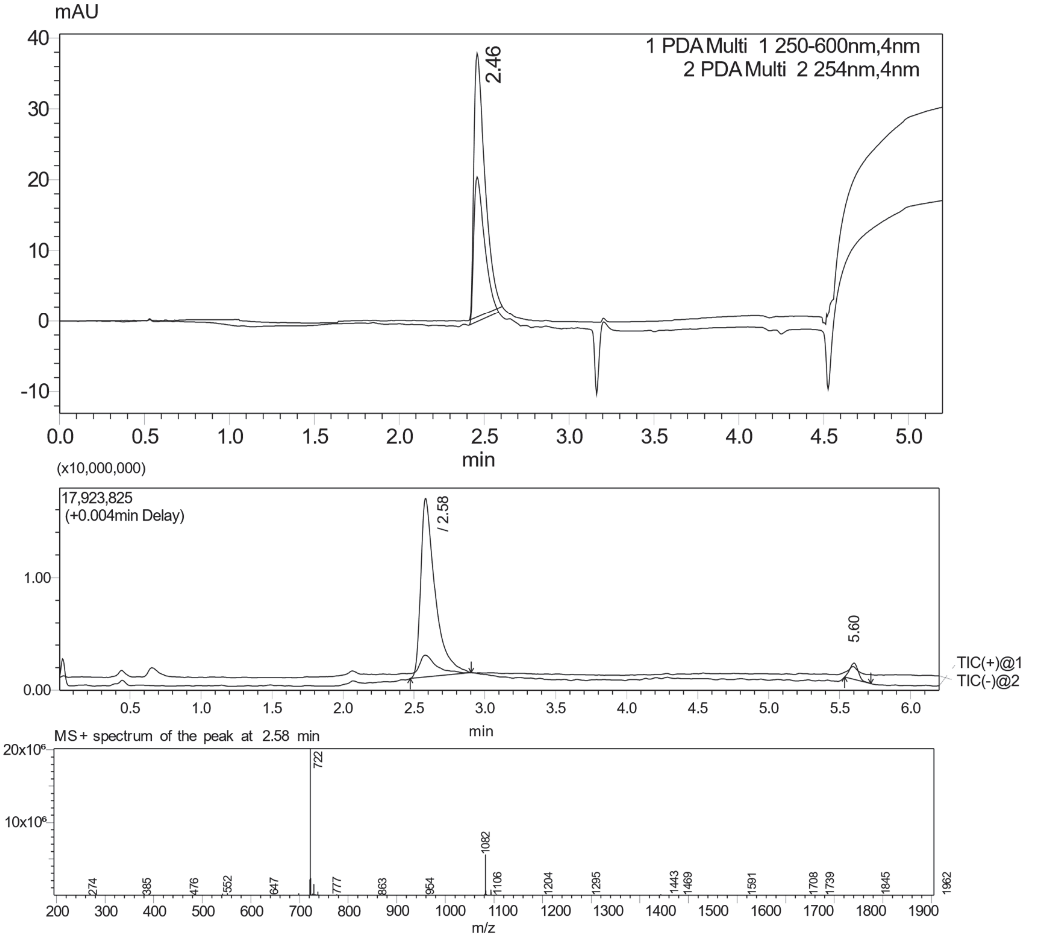
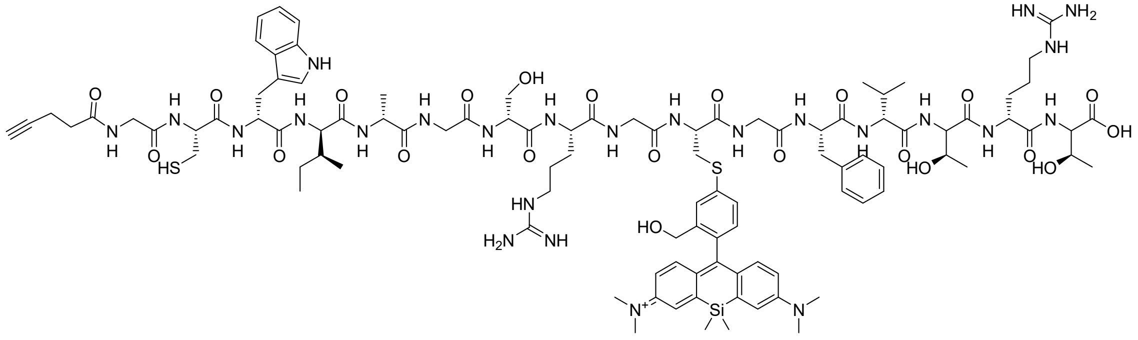


Figure 164. Structure, condensed formula, exact mass and LC-MS traces and extracted peak of peptide **E19.**



Pentynyl-GCWiAGSRGC(HMSiR)GFVTRT-OH

HRMS (ESI/QTOF) $[M+3H]^{3+}$ calcd. for $[C_{102}H_{147}N_{25}O_{22}S_2Si]^{+3}$: 722.0116; found: 722.0104.

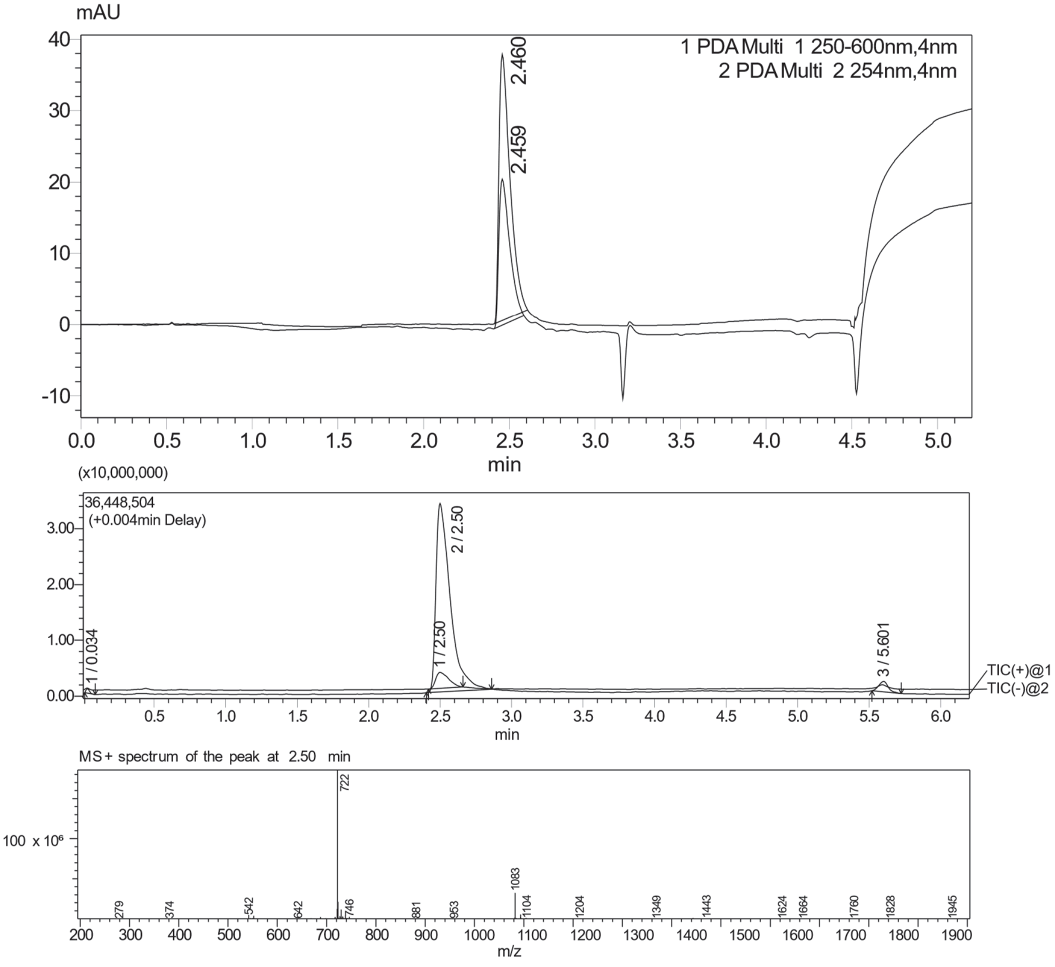
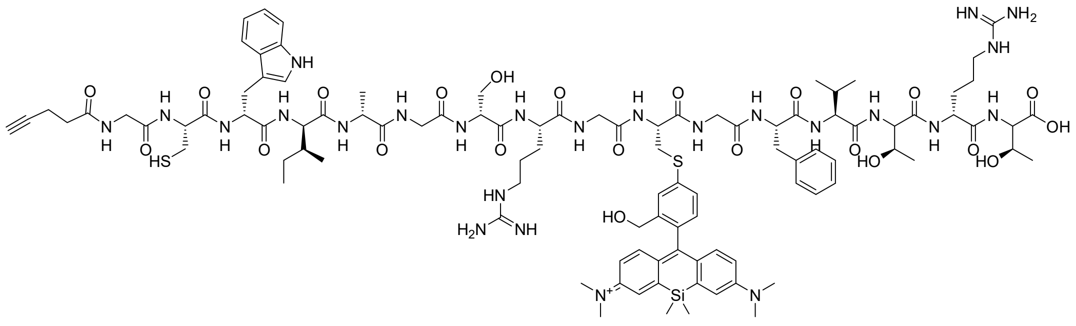


Figure 165. Structure, condensed formula, exact mass and LC-MS traces and extracted peak of peptide **E20**.



Pentynyl-GCWiAGSRGC(HMSiR)GFvTRT-OH

HRMS (ESI/QTOF) $[M+3H]^{3+}$ calcd. for $[C_{102}H_{146}N_{25}O_{22}S_2Si]^{+3}$: 721.6756; found: 721.6785.

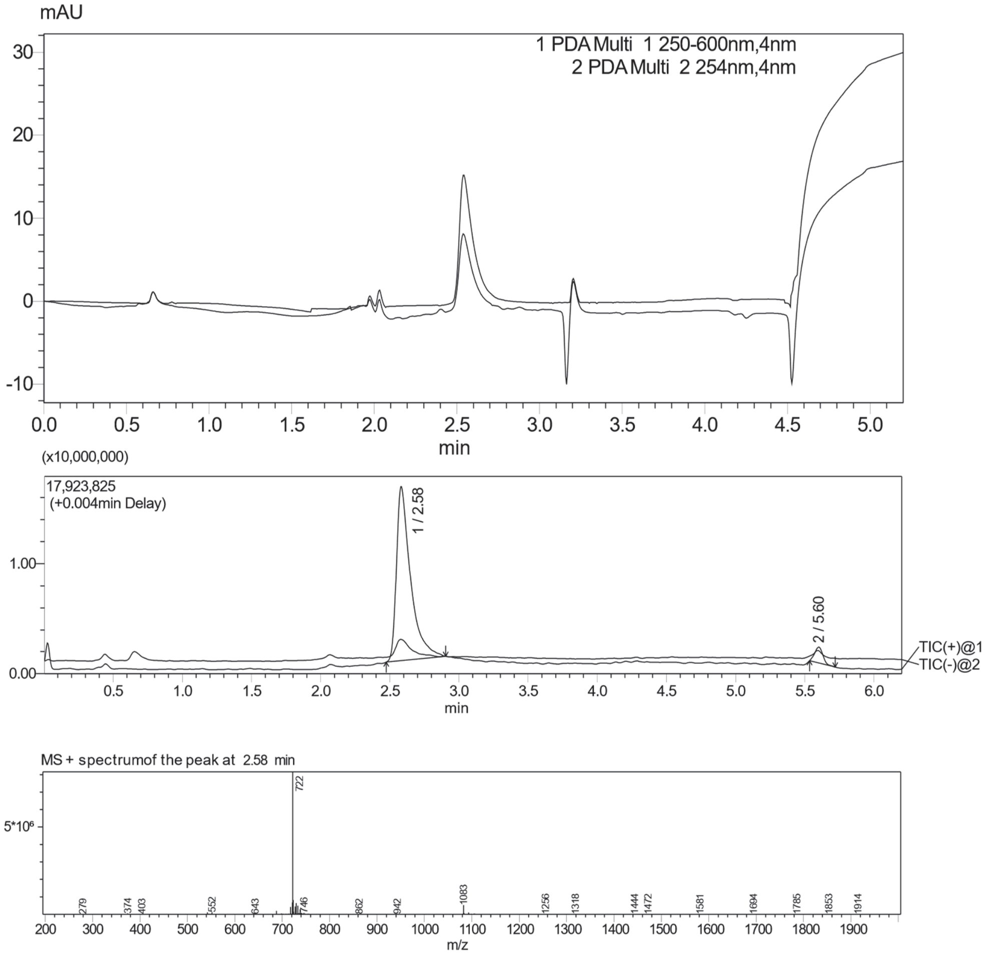


Figure 166. Structure, condensed formula, exact mass and LC-MS traces and extracted peak of peptide **E21**.

References

[1] S. J. Altschuler, L. F. Wu, *Cell* 2010, *141*, 559–563.

[2] J. A. Griffiths, A. Scialdone, J. C. Marioni, *Mol. Syst. Biol.* 2018, *14*, e8046.

[3] E. V. Koonin, *Biology Direct* 2012, *7*, 27.

[4] R. Aebersold, J. N. Agar, I. J. Amster, M. S. Baker, C. R. Bertozzi, E. S. Boja, C. E. Costello, B. F. Cravatt, C. Fenselau, B. A. Garcia, Y. Ge, J. Gunawardena, R. C. Hendrickson, P. J. Hergenrother, C. G. Huber, A. R. Ivanov, O. N. Jensen, M. C. Jewett, N. L. Kelleher, L. L. Kiessling, N. J. Krogan, M. R. Larsen, J. A. Loo, R. R. Ogorzalek Loo, E. Lundberg, M. J. MacCoss, P. Mallick, V. K. Mootha, M. Mrksich, T. W. Muir, S. M. Patrie, J. J. Pesavento, S. J. Pitteri, H. Rodriguez, A. Saghatelian, W. Sandoval, H. Schlüter, S. Sechi, S. A. Slavoff, L. M. Smith, M. P. Snyder, P. M. Thomas, M. Uhlén, J. E. Van Eyk, M. Vidal, D. R. Walt, F. M. White, E. R. Williams, T. Wohlschlager, V. H. Wysocki, N. A. Yates, N. L. Young, B. Zhang, *Nat. Chem. Biol.* 2018, *14*, 206–214.

[5] L. M. Smith, N. L. Kelleher, *Nat. Methods* 2013, *10*, 186–187.

[6] J. M. Heather, B. Chain, *Genomics* 2016, *107*, 1–8.

[7] S. M. Williams, A. V. Liyu, C.-F. Tsai, R. J. Moore, D. J. Orton, W. B. Chrisler, M. J. Gaffrey, T. Liu, R. D. Smith, R. T. Kelly, L. Pasa-Tolic, Y. Zhu, *Anal. Chem.* 2020, *92*, 10588–10596.

[8] M. J. MacCoss, J. A. Alfaro, D. A. Faivre, C. C. Wu, M. Wanunu, N. Slavov, *Nat. Methods* 2023, *20*, 339–346.

[9] F. Crick, *Nature* 1970, *227*, 561–563.

[10] L. Restrepo-Pérez, C. Joo, C. Dekker, *Nat. Nanotechnol.* 2018, *13*, 786–796.

[11] E. A. Ponomarenko, E. V. Poverennaya, E. V. Ilgisonis, M. A. Pyatnitskiy, A. T. Kopylov, V. G. Zgoda, A. V. Lisitsa, A. I. Archakov, *Int. J. Anal. Chem.* 2016, *2016*, e7436849.

[12] R. A. Zubarev, *Proteomics* 2013, *13*, 723–726.

[13] P. Edman, E. Högfeldt, L. G. Sillén, P.-O. Kinell, *Acta Chem. Scand.* 1950, *4*, 283–293.

[14] W. Timp, G. Timp, *Sci. Adv.* 2020, *6*, eaax8978.

[15] J. A. Alfaro, P. Bohländer, M. Dai, M. Filius, C. J. Howard, X. F. van Kooten, S. Ohayon, A. Pomorski, S. Schmid, A. Aksimentiev, E. V. Anslyn, G. Bedran, C. Cao, M. Chinappi, E. Coyaud, C. Dekker, G. Dittmar, N. Drachman, R. Eelkema, D. Goodlett, S. Hentz, U. Kalathiya, N. L. Kelleher, R. T. Kelly, Z. Kelman, S. H. Kim, B. Kuster, D. Rodriguez-Larrea, S. Lindsay, G. Maglia, E. M. Marcotte, J. P. Marino, C. Masselon, M. Mayer, P. Samaras, K. Sarthak, L. Sepiashvili, D. Stein, M. Wanunu, M. Wilhelm, P. Yin, A. Meller, C. Joo, *Nat. Methods* 2021, *18*, 604–617.

[16] J. Swaminathan, A. A. Boulgakov, E. M. Marcotte, *PLoS Comput. Biol.* 2015, *11*, e1004080.

[17] S. Ohayon, A. Girsault, M. Nasser, S. Shen-Orr, A. Meller, *PLoS Comput. Biol.* 2019, *15*, e1007067.

[18] C. V. de Lannoy, M. Filius, R. van Wee, C. Joo, D. de Ridder, *iScience* 2021, *24*, 103239.

[19] J. A. Owen, J. Punt, S. A. Stranford, P. P. Jones, J. Kuby, *Kuby Immunology*, W.H. Freeman New York, New York, 2013.

[20] L. Gu, C. Li, J. Aach, D. E. Hill, M. Vidal, G. M. Church, *Nature* 2014, *515*, 554–557.

[21] A. V. Ullal, R. Weissleder, in *Single Cell Protein Analysis: Methods and Protocols* (Eds.: A.K. Singh, A. Chandrasekaran), Springer, New York, NY, 2015, pp. 47–54.

[22] B. T. Chait, *Science* 2006, *314*, 65–66.

[23] R. Aebersold, M. Mann, *Nature* 2016, *537*, 347–355.

[24] N. Slavov, *Science* 2020, *367*, 512–513.

[25] B. Domon, R. Aebersold, *Nat. Biotechnol.* 2010, *28*, 710–721.

[26] N. Slavov, *Curr. Opin. Chem Biol.* 2021, *60*, 1–9.

[27] B. Bogdanov, R. D. Smith, *Mass Spectrom. Rev.* 2005, *24*, 168–200.

[28] M. S. Mansuri, K. Williams, A. C. Nairn, *Commun. Biol.* 2023, *6*, 1–6.

[29] B. Budnik, E. Levy, G. Harmange, N. Slavov, *Genome Biol.* 2018, *19*, 161.

[30] H. Specht, E. Emmott, A. A. Petelski, R. G. Huffman, D. H. Perlman, M. Serra, P. Kharchenko, A. Koller, N. Slavov, *Genome Biol.* 2021, *22*, 50.

[31] J. Derks, A. Leduc, G. Wallmann, R. G. Huffman, M. Willetts, S. Khan, H. Specht, M. Ralser, V. Demichev, N. Slavov, *Nat. Biotechnol.* 2023, *41*, 50–59.

[32] J. P. Savaryn, A. D. Catherman, P. M. Thomas, M. M. Abecassis, N. L. Kelleher, *Genome Med.* 2013, *5*, 53.

[33] A. D. Catherman, K. R. Durbin, D. R. Ahlf, B. P. Early, R. T. Fellers, J. C. Tran, P. M. Thomas, N. L. Kelleher, *Mol. Cell. Proteomics* 2013, *12*, 3465–3473.

[34] T. K. Toby, L. Fornelli, N. L. Kelleher, *Annu. Rev. Anal. Chem.* 2016, *9*, 499–519.

[35] V. Marx, *Nat. Methods* 2019, *16*, 809−812.

[36] J. Swaminathan, A. A. Boulgakov, E. T. Hernandez, A. M. Bardo, J. L. Bachman, J. Marotta, A. M. Johnson, E. V. Anslyn, E. M. Marcotte, *Nat. Biotechnol.* 2018, *36*, 1076−1091.

[37] X. Wei, T. Penkauskas, J. E. Reiner, C. Kennard, M. J. Uline, Q. Wang, S. Li, A. Aksimentiev, J. W. F. Robertson, C. Liu, *ACS Nano* 2023, *17*, 16369–16395.

[38] M. Di Ventra, M. Taniguchi, *Nat. Nanotechnol.* 2016, *11*, 117–126.

[39] M. Zwolak, M. Di Ventra, *Nano Lett.* 2005, *5*, 421–424.

[40] Y. Zhao, B. Ashcroft, P. Zhang, H. Liu, S. Sen, W. Song, J. Im, B. Gyarfas, S. Manna, S. Biswas, C. Borges, S. Lindsay, *Nat. Nanotechnol.* 2014, *9*, 466–473.

[41] T. Ohshiro, M. Tsutsui, K. Yokota, M. Furuhashi, M. Taniguchi, T. Kawai, *Nat. Nanotechnol.* 2014, *9*, 835–840.

[42] J. J. Kasianowicz, E. Brandin, D. Branton, D. W. Deamer, *Proc. Natl. Acad. Sci. U.S.A.* 1996, *93*, 13770–13773.

[43] D. Deamer, M. Akeson, D. Branton, *Nat. Biotechnol.* 2016, *34*, 518–524.

[44] Y. Wang, Y. Zhao, A. Bollas, Y. Wang, K. F. Au, *Nat. Biotechnol.* 2021, *39*, 1348–1365.

[45] Z.-L. Hu, M.-Z. Huo, Y.-L. Ying, Y.-T. Long, *Angew. Chem. Int. Ed.* 2021, *60*, 14738–14749.

[46] H. Ouldali, K. Sarthak, T. Ensslen, F. Piguet, P. Manivet, J. Pelta, J. C. Behrends, A. Aksimentiev, A. Oukhaled, *Nat. Biotechnol.* 2020, *38*, 176–181.

[47] M. Chinappi, F. Cecconi, *J. Phys.: Condens. Matter* 2018, *30*, 204002.

[48] P. Martin-Baniandres, W.-H. Lan, S. Board, M. Romero-Ruiz, S. Garcia-Manyes, Y. Qing, H. Bayley, *Nat. Nanotechnol.* 2023, *18*, 1335–1340.

[49] L. Yu, X. Kang, F. Li, B. Mehrafrooz, A. Makhamreh, A. Fallahi, J. C. Foster, A. Aksimentiev, M. Chen, M. Wanunu, *Nat. Biotechnol.* 2023, *41*, 1130–1139.

[50] J. Nivala, D. B. Marks, M. Akeson, *Nat. Biotechnol.* 2013, *31*, 247–250.

[51] J. Nivala, L. Mulroney, G. Li, J. Schreiber, M. Akeson, *ACS Nano* 2014, *8*, 12365–12375.

[52] H. Brinkerhoff, A. S. W. Kang, J. Liu, A. Aksimentiev, C. Dekker, *Science* 2021, *374*, 1509–1513.

[53] S. Zhang, G. Huang, R. C. A. Versloot, B. M. H. Bruininks, P. C. T. de Souza, S.-J. Marrink, G. Maglia, *Nat. Chem.* 2021, *13*, 1192–1199.

[54] L. Mereuta, M. Roy, A. Asandei, J. K. Lee, Y. Park, I. Andricioaei, T. Luchian, *Sci. Rep.* 2014, *4*, 3885.

[55] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, D. Baker, *Nature* 2023, *620*, 1089–1100.

[56] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, D. Baker, *Science* 2022, *378*, 49–56.

[57] K. Wang, S. Zhang, X. Zhou, X. Yang, X. Li, Y. Wang, P. Fan, Y. Xiao, W. Sun, P. Zhang, W. Li, S. Huang, *Nat. Methods* 2024, *21*, 92–101.

[58] M. Zhang, C. Tang, Z. Wang, S. Chen, D. Zhang, K. Li, K. Sun, C. Zhao, Y. Wang, M. Xu, L. Dai, G. Lu, H. Shi, H. Ren, L. Chen, J. Geng, *Nat. Methods* **2024**, 1–10.

[59] Y. Zhang, Y. Yi, Z. Li, K. Zhou, L. Liu, H.-C. Wu, *Nat. Methods* **2024**, *21*, 102–109.

[60] S. Schmid, P. Stömmer, H. Dietz, C. Dekker, *Nat. Nanotechnol.* **2021**, *16*, 1244–1250.

[61] C. Wen, E. Bertosin, X. Shi, C. Dekker, S. Schmid, *Nano Lett.* **2023**, *23*, 788–794.

[62] S. Schmid, C. Dekker, *iScience* **2021**, *24*, 103007.

[63] J. Fan, K. L. Gunderson, M. Bibikova, J. M. Yeakley, J. Chen, E. Wickham Garcia, L. L. Lebruska, M. Laurent, R. Shen, D. Barker, *Methods Enzymol.* **2006**, *410*, 57–73.

[64] T. D. Harris, P. R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, J. DiMeo, J. W. Efcavitch, E. Giladi, J. Gill, J. Healy, M. Jarosz, D. Lapen, K. Moulton, S. R. Quake, K. Steinmann, E. Thayer, A. Tyurina, R. Ward, H. Weiss, Z. Xie, *Science* **2008**, *320*, 106–109.

[65] P. W. Hook, W. Timp, *Nat. Rev. Genet.* **2023**, *24*, 627–641.

[66] B. C. Collins, R. Aebersold, *Nat. Biotechnol.* **2018**, *36*, 1051–1053.

[67] M. Filius, S. H. Kim, I. Severins, C. Joo, *Nano Lett.* **2021**, *21*, 3295–3301.

[68] M. Filius, R. van Wee, C. de Lannoy, I. Westerlaken, Z. Li, S. H. Kim, C. de Agrela Pinto, Y. Wu, G.-J. Boons, M. Pabst, D. de Ridder, C. Joo, *Nat. Nanotechnol.* **2024**, 1–8.

[69] J. D. Egertson, D. DiPasquo, A. Killeen, V. Lobanov, S. Patel, P. Mallick, **2021**, 2021.10.11.463967.

[70] T. Aksel, H. Qian, P. Hao, P. F. Indermuhle, C. Inman, S. Paul, K. Chen, R. Seghers, J. K. Robinson, M. D. Garate, B. Nortman, J. Tan, S. Hendricks, S. Sankar, P. Mallick, **2022**, 2022.05.02.490328.

[71] B. D. Reed, M. J. Meyer, V. Abramzon, O. Ad, P. Adcock, F. R. Ahmad, G. Alppay, J. A. Ball, J. Beach, D. Belhachemi, A. Bellofiore, M. Bellos, J. F. Beltrán, A. Betts, M. W. Bhuiya, K. Blacklock, R. Boer, D. Boisvert, N. D. Brault, A. Buxbaum, S. Caprio, C. Choi, T. D. Christian, R. Clancy, J. Clark, T. Connolly, K. F. Croce, R. Cullen, M. Davey, J. Davidson, M. M. Elshenawy, M. Ferrigno, D. Frier, S. Gudipati, S. Hamill, Z. He, S. Hosali, H. Huang, L. Huang, A. Kabiri, G. Kriger, B. Lathrop, A. Li, P. Lim, S. Liu, F. Luo, C. Lv, X. Ma, E. McCormack, M. Millham, R. Nani, M. Pandey, J. Parillo, G. Patel, D. H. Pike, K. Preston, A. Pichard-Kostuch, K. Rearick, T. Rearick, M. Ribezzi-Crivellari, G. Schmid, J. Schultz, X. Shi, B. Singh, N. Srivastava, S. F. Stewman, T. R. Thurston, P. Trioli, J. Tullman, X. Wang, Y.-C. Wang, E. A. G. Webster, Z. Zhang, J. Zuniga, S. S. Patel, A. D. Griffiths, A. M. van Oijen, M. McKenna, M. D. Dyer, J. M. Rothberg, *Science* **2022**, *378*, 186–192.

[72] J. Tullman, N. Callahan, B. Ellington, Z. Kelman, J. P. Marino, *Appl. Microbiol. Biotechnol.* **2019**, *103*, 2621–2633.

[73] F. Köhn, J. Hofkens, R. Gronheid, M. Van der Auweraer, F. C. De Schryver, *J. Phys. Chem. A* **2002**, *106*, 4808–4814.

[74] F. Köhn, J. Hofkens, U.-M. Wiesler, M. Cotlet, M. van der Auweraer, K. Müllen, F. C. De Schryver, *Chem. Eur. J.* **2001**, *7*, 4126–4133.

[75] W. P. Ambrose, P. M. Goodwin, J. C. Martin, R. A. Keller, *Phys. Rev. Lett.* **1994**, *72*, 160–163.

[76] T. D. M. Bell, A. Stefan, S. Masuo, T. Vosch, M. Lor, M. Cotlet, J. Hofkens, S. Bernhardt, K. Müllen, M. van der Auweraer, J. W. Verhoeven, F. C. De Schryver, *ChemPhysChem* **2005**, *6*, 942–948.

[77] J. N. Clifford, T. D. M. Bell, P. Tinnefeld, M. Heilemann, S. M. Melnikov, J. I. Hotta, M. Sliwa, P. Dedecker, M. Sauer, J. Hofkens, E. K. L. Yeow, *J. Phys. Chem. B* **2007**, *111*, 6987–6991.

[78] M. Levitus, S. Ranjit, *Q. Rev. Biophys.* **2011**, *44*, 123–151.

[79] T. Ha, P. Tinnefeld, *Annu. Rev. Phys. Chem.* **2012**, *63*, 595–617.

[80] K. Kawai, T. Koshimo, A. Maruyama, T. Majima, *Chem. Commun.* **2014**, *50*, 10478–10481.

[81] K. Kawai, A. Maruyama, *Chem. Commun.* **2015**, *51*, 4861–4864.

[82] K. Kawai, T. Miyata, N. Shimada, S. Ito, I. Miyasaka, A. Maruyama, H. Miyasaka, A. Maruyama, *Angew. Chem. Int. Ed.* 2017, *56*, 15329–15333.

[83] S. N. Uno, M. Kamiya, T. Yoshihara, K. Sugawara, K. Okabe, M. C. Tarhan, H. Fujita, T. Funatsu, Y. Okada, S. Tobita, Y. Urano, *Nat. Chem.* 2014, *6*, 681–689.

[84] W. Chi, D. Tan, Q. Qiao, Z. Xu, X. Liu, *Angew. Chem. Int. Ed.* 2023, *62*, e202306061.

[85] H. Takakura, Y. Zhang, R. S. Erdmann, A. D. Thompson, Y. Lin, B. McNellis, F. Rivera-Molina, S. Uno, M. Kamiya, Y. Urano, J. E. Rothman, J. Bewersdorf, A. Schepartz, D. Toomre, *Nat. Biotechnol.* 2017, *35*, 773–780.

[86] F. Deng, Q. Qiao, J. Li, W. Yin, L. Miao, X. Liu, Z. Xu, *J. Phys. Chem. B* 2020, *124*, 7467–7474.

[87] R. Tachibana, M. Kamiya, A. Morozumi, Y. Miyazaki, H. Fujioka, A. Nanjo, R. Kojima, T. Komatsu, T. Ueno, K. Hanaoka, T. Yoshihara, S. Tobita, Y. Urano, *Chem. Commun.* 2020, *56*, 13173–13176.

[88] W. Chi, Q. Qiao, C. Wang, J. Zheng, W. Zhou, N. Xu, X. Wu, X. Jiang, D. Tan, Z. Xu, X. Liu, *Angew. Chem. Int. Ed.* 2020, *59*, 20215–20223.

[89] R. W. Ramette, E. B. Sandell, *J. Am. Chem. Soc.* 1956, *78*, 4872–4878.

[90] K. N. Fish, "Total Internal Reflection Fluorescence (TIRF) Microscopy | Nikon's MicroscopyU," 2009.

[91] S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* 2023, *145*, 1441–1447.

[92] M. Vogel, W. Rettig, R. Sens, K. H. Drexhage, *Chem. Phys. Lett.* 1988, *147*, 452–460.

[93] J. B. Grimm, B. P. English, J. Chen, J. P. Slaughter, Z. Zhang, A. Revyakin, R. Patel, J. J. Macklin, D. Normanno, R. H. Singer, T. Lionnet, L. D. Lavis, *Nat. Methods* 2015, *12*, 244–250.

[94] M. Minoshima, K. Kikuchi, *J. Biol. Inorg. Chem.* 2017, *22*, 639–652.

[95] C. Wang, W. Chi, Q. Qiao, D. Tan, Z. Xu, X. Liu, *Chem. Soc. Rev.* 2021, *50*, 12656–12678.

[96] Y. Koide, Y. Urano, K. Hanaoka, T. Terai, T. Nagano, *ACS Chem. Biol.* 2011, *6*, 600–608.

[97] J. B. Grimm, A. N. Tkachuk, L. Xie, H. Choi, B. Mohar, N. Falco, K. Schaefer, R. Patel, Q. Zheng, Z. Liu, J. Lippincott-Schwartz, T. A. Brown, L. D. Lavis, *Nat. Methods* 2020, *17*, 815–821.

[98] M. Grzybowski, M. Taki, K. Senda, Y. Sato, T. Ariyoshi, Y. Okada, R. Kawakami, T. Imamura, S. Yamaguchi, *Angew. Chem. Int. Ed.* 2018, *57*, 10137–10141.

[99] J. Liu, Y.-Q. Sun, H. Zhang, H. Shi, Y. Shi, W. Guo, *ACS Appl. Mater. Interfaces* 2016, *8*, 22953–22962.

[100] G. Lukinavičius, K. Umezawa, N. Olivier, A. Honigmann, G. Yang, T. Plass, V. Mueller, L. Reymond, I. R. Corrêa Jr, Z.-G. Luo, C. Schultz, E. A. Lemke, P. Heppenstall, C. Eggeling, S. Manley, K. Johnsson, *Nat. Chem.* 2013, *5*, 132–139.

[101] G. Lukinavičius, L. Reymond, E. D'Este, A. Masharina, F. Göttfert, H. Ta, A. Güther, M. Fournier, S. Rizzo, H. Waldmann, C. Blaukopf, C. Sommer, D. W. Gerlich, H.-D. Arndt, S. W. Hell, K. Johnsson, *Nat. Methods* 2014, *11*, 731–733.

[102] Q. Zheng, A. X. Ayala, I. Chung, A. V. Weigel, A. Ranjan, N. Falco, J. B. Grimm, A. N. Tkachuk, C. Wu, J. Lippincott-Schwartz, R. H. Singer, L. D. Lavis, *ACS Cent. Sci.* 2019, *5*, 1602–1613.

[103] L. Wang, M. Tran, E. D'Este, J. Roberti, B. Koch, L. Xue, K. Johnsson, *Nat. Chem.* 2020, *12*, 165–172.

[104] N. Lardon, L. Wang, A. Tschanz, P. Hoess, M. Tran, E. D'Este, J. Ries, K. Johnsson, *J. Am. Chem. Soc.* 2021, *143*, 14592–14600.

[105] K. Umezawa, M. Yoshida, M. Kamiya, T. Yamasoba, Y. Urano, *Nat. Chem.* 2017, *9*, 279—286.

[106] J. B. Grimm, A. N. Tkachuk, R. Patel, S. T. Hennigan, A. Gutu, P. Dong, V. Gandin, A. M. Osowski, K. L. Holland, Z. J. Liu, T. A. Brown, L. D. Lavis, *J. Am. Chem. Soc.* 2023, *145*, 23000–23013.

[107] A. N. Butkevich, G. Yu. Mitronova, S. C. Sidenstein, J. L. Klocke, D. Kamin, D. N. H. Meineke, E. D'Este, P.-T. Kraemer, J. G. Danzl, V. N. Belov, S. W. Hell, *Angew. Chem. Int. Ed.* **2016**, *55*, 3290–3294.

[108] M. Fu, Y. Xiao, X. Qian, D. Zhao, Y. Xu, *Chem. Commun.* **2008**, 1780–1782.

[109] D. Si, Q. Li, Y. Bao, J. Zhang, L. Wang, *Angew. Chem. Int. Ed.* **2023**, *62*, e202307641.

[110] R. Gerasimaitė, J. Bucevičius, K. A. Kiszka, S. Schnorrenberg, G. Kostiuk, T. Koenen, G. Lukinavičius, *ACS Chem. Biol.* **2021**, *16*, 2130–2136.

[111] D.-B. Sung, J. Seok Lee, *RSC Med. Chem.* **2023**, *14*, 412–432.

[112] B. N. G. Giepmans, S. R. Adams, M. H. Ellisman, R. Y. Tsien, *Science* **2006**, *312*, 217–224.

[113] U. Schnell, F. Dijk, K. A. Sjollema, B. N. G. Giepmans, *Nat. Methods* **2012**, *9*, 152–158.

[114] A. A. Pakhomov, V. I. Martynov, *Chem. Biol.* **2008**, *15*, 755–764.

[115] R. Y. Tsien, *Angew. Chem. Int. Ed.* **2009**, *48*, 5612–5626.

[116] K. M. Marks, G. P. Nolan, *Nat. Methods* **2006**, *3*, 591–596.

[117] J. Wiedenmann, F. Oswald, G. U. Nienhaus, *IUBMB Life* **2009**, *61*, 1029–1042.

[118] G. V. Los, L. P. Encell, M. G. McDougall, D. D. Hartzell, N. Karassina, D. Simpson, J. Mendez, K. Zimmerman, P. Otto, G. Vidugiris, J. Zhu, *ACS Chem. Biol.* **2008**, *3*, 373–382.

[119] A. Juillerat, T. Gronemeyer, A. Keppler, S. Gendreizig, H. Pick, H. Vogel, K. Johnsson, *Chem. Biol.* **2003**, *10*, 313–317.

[120] A. Gautier, A. Juillerat, C. Heinis, I. R. Corrêia, M. Kindermann, F. Beaufils, K. Johnsson, I. R. Corrêa, M. Kindermann, F. Beaufils, K. Johnsson, *Chem. Biol.* **2008**, *15*, 128–136.

[121] L. W. Miller, Y. Cai, M. P. Sheetz, V. W. Cornish, *Nat. Methods* **2005**, *2*, 255–257.

[122] A. Keppler, S. Gendreizig, T. Gronemeyer, H. Pick, H. Vogel, K. Johnsson, *Nat. Biotechnol.* **2003**, *21*, 86–89.

[123] Z. Chen, C. Jing, S. S. Gallagher, M. P. Sheetz, V. W. Cornish, *J. Am. Chem. Soc.* **2012**, *134*, 13692–13699.

[124] L. Wang, M. Tran, E. D'Este, J. Roberti, B. Koch, L. Xue, K. Johnsson, *Nat. Chem.* **2020**, *12*, 165–172.

[125] M.-A. Plamont, E. Billon-Denis, S. Maurin, F. M. Pimenta, J. Shi, J. Qu'rard, B. Pan, J. Rossignol, Y. Chen, T. Le Saux, L. Jullien, A. Gautier, C. Gauron, M. Volovitch, S. Vriz, C. G. Specht, A. Triller, K. Moncoq, N. Morellet, E. Lescop, *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 497–502.

[126] W. Weber, V. Helms, J. A. McCammon, P. W. Langhoff, *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 6177–6182.

[127] J. Lotze, U. Reinhardt, O. Seitz, A. G. Beck-Sickinger, *Mol. BioSyst.* **2016**, *12*, 1731–1745.

[128] J. Wilhelm, S. Kühn, M. Tarnawski, G. Gotthard, J. Tünnermann, T. Tänzer, J. Karpenko, N. Mertes, L. Xue, U. Uhrig, J. Reinstein, J. Hiblot, K. Johnsson, *Biochemistry* **2021**, *60*, 2560–2575.

[129] K. O. Håkansson, J. R. Winther, *Acta Cryst D* **2007**, *63*, 288–294.

[130] Schrödinger, LLC, **2015**.

[131] H. Mao, S. A. Hart, A. Schink, B. A. Pollok, *J. Am. Chem. Soc.* **2004**, *126*, 2670–2671.

[132] S. Puthenveetil, D. S. Liu, K. A. White, S. Thompson, A. Y. Ting, *J. Am. Chem. Soc.* **2009**, *131*, 16430–16438.

[133] R. A. Scheck, A. Schepartz, *Acc. Chem. Res.* **2011**, *44*, 654–665.

[134] S. R. Adams, R. E. Campbell, L. A. Gross, B. R. Martin, G. K. Walkup, Y. Yao, J. Llopis, R. Y. Tsien, *J. Am. Chem. Soc* **2002**, *124*, 6063–6076.

[135] H. Cao, B. Chen, T. C. Squier, M. U. Mayer, *Chem. Commun.* **2006**, 2601–2603.

[136] B. A. Griffin, S. R. Adams, R. Y. Tsien, *Science* **1998**, *281*, 269–272.

[137] C. C. Spagnuolo, R. J. Vermeij, E. A. Jares-Erijman, *J. Am. Chem. Soc.* **2006**, *128*, 12040–12041.

[138] H. Cao, Y. Xiong, T. Wang, B. Chen, T. C. Squier, M. U. Mayer, *J. Am. Chem. Soc* 2007, *129*, 8672–8673.

[139] T. Wang, P. Yan, T. C. Squier, M. U. Mayer, *ChemBioChem* 2007, *8*, 1937–1940.

[140] B. Chen, M. U. Mayer, L. M. Markillie, D. L. Stenoien, T. C. Squier, *Biochemistry* 2005, *44*, 905–914.

[141] T. L. Halo, J. Appelbaum, E. M. Hobert, D. M. Balkin, A. Schepartz, *J. Am. Chem. Soc* 2009, *131*, 438–439.

[142] E. D'Este, G. Lukinavičius, R. Lincoln, F. Opazo, E. F. Fornasiero, *Trends Cell Biol.* 2024, DOI 10.1016/j.tcb.2023.12.001.

[143] X. Sun, T. D. James, E. V. Anslyn, *J. Am. Chem. Soc.* 2018, *140*, 2348–2354.

[144] M. Sibrian-Vazquez, J. O. Escobedo, M. Lowry, R. M. Strongin, *Pure and Applied Chemistry* 2012, *84*, 2443–2456.

[145] C. Cao, P. Magalhães, L. F. Krapp, J. F. Bada Juarez, S. F. Mayer, V. Rukes, A. Chiki, H. A. Lashuel, M. Dal Peraro, *ACS Nano* 2024, *18*, 1504–1515.

[146] J. O. Awoyemi, A. O. Adetunmbi, S. A. Oluwadare, in *Proceedings of the IEEE International Conference on Computing, Networking and Informatics*, Lagos, Nigeria, 2017, pp. 1–9.

[147] J. Cheng, A. N. Tegge, P. Baldi, *IEEE Rev. Biomed. Eng.* 2008, *1*, 41–49.

[148] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P. M. Agapow, M. Zietz, M. M. Hoffman, W. Xie, G. L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A. E. Carpenter, A. Shrikumar, J. Xu, E. M. Cofer, C. A. Lavender, S. C. Turaga, A. M. Alexandari, Z. Lu, D. J. Harris, D. Decaprio, Y. Qi, A. Kundaje, Y. Peng, L. K. Wiley, M. H. S. Segler, S. M. Boca, S. J. Swamidass, A. Huang, A. Gitter, C. S. Greene, *J. R. Soc. Interface.* 2018, *15*, 20170387.

[149] J. Padmanabhan, M. J. Johnson Premkumar, *IETE Technical Review* 2015, *32*, 240–251.

[150] A. Jain, R. Liu, Y. K. Xiang, T. Ha, *Nat. Protoc.* 2012, *7*, 445–452.

[151] H. C. Kolb, K. B. Sharpless, *Drug Discovery Today* 2003, *8*, 1128–1137.

[152] M. J. Rust, M. Bates, X. Zhuang, *Nat. Methods* 2006, *3*, 793–795.

[153] H. Labit, A. Goldar, G. Guilbaud, C. Douarche, O. Hyrien, K. Marheineke, *BioTechniques* 2008, *45*, 649–658.

[154] M. P. Nicholas, L. Rao, A. Gennerich, *Methods Mol. Biol.* 2014, *1136*, 137–169.

[155] J. L. Davis, B. Dong, C. Sun, H. F. Zhang, *J. Biomed. Opt.* 2018, *23*, 106501.

[156] C. F. Mallinson, J. E. Castle, J. F. Watts, *Surf. Sci. Spectra* 2015, *22*, 71–80.

[157] Z. Fan, C. Zhi, L. Wu, P. Zhang, C. Feng, L. Deng, B. Yu, L. Qian, *Coatings* 2019, *9*, 762.

[158] M. Zhu, M. Z. Lerum, W. Chen, *Langmuir* 2012, *28*, 416–423.

[159] A. Johnson-Buck, J. Li, M. Tewari, N. G. Walter, *Methods* 2019, *153*, 3–12.

[160] L. Britcher, T. J. Barnes, H. J. Griesser, C. A. Prestidge, *Langmuir* 2008, *24*, 7625–7627.

[161] S. Upadhyayula, T. Quinata, S. Bishop, S. Gupta, N. R. Johnson, B. Bahmani, K. Bozhilov, J. Stubbs, P. Jreij, P. Nallagatla, V. I. Vullev, *Langmuir* 2012, *28*, 5059–5069.

[162] S. D. Chandradoss, A. C. Haagsma, Y. K. Lee, J. H. Hwang, J. M. Nam, C. Joo, *J. Visualized Exp.* 2014, e50549.

[163] Z. Miao, Z. Cheng, *Bio-Protoc.* 2012, *2*, DOI 10.21769/bioprotoc.233.

[164] A. Krasovskiy, P. Knochel, *Synthesis* 2006, *5*, 890–891.

[165] D. Virág, B. Dalmadi-Kiss, K. Vékey, L. Drahos, I. Klebovich, I. Antal, K. Ludányi, *Chromatographia* 2020, *83*, 1–10.

[166] M. R. Larsen, M. B. Trelle, T. E. Thingholm, O. N. Jensen, *BioTechniques* 2006, *40*, 790–798.

[167] C. Reily, T. J. Stewart, M. B. Renfrow, J. Novak, *Nat. Rev. Nephrol.* 2019, *15*, 346–366.

[168] Y. Zhao, O. N. Jensen, *Proteomics* 2009, *9*, 4632–4641.

[169] A. Venerando, L. Cesaro, L. A. Pinna, *FEBS J.* 2017, *284*, 1936–1951.

[170] D. Altschuler, E. G. Lapetina, *J. Biol. Chem.* 1993, *268*, 7527–7531.

[171] M. Montalbán-López, T. A. Scott, S. Ramesh, I. R. Rahman, A. J. van Heel, J. H. Viel, V. Bandarian, E. Dittmann, O. Genilloud, Y. Goto, M. J. Grande Burgos, C. Hill, S. Kim, J. Koehnke, J. A. Latham, A. J. Link, B. Martínez, S. K. Nair, Y. Nicolet, S. Rebuffat, H.-G. Sahl, D. Sareen, E. W. Schmidt, L. Schmitt, K. Severinov, R. D. Süssmuth, A. W. Truman, H. Wang, J.-K. Weng, G. P. van Wezel, Q. Zhang, J. Zhong, J. Piel, D. A. Mitchell, O. P. Kuipers, W. A. van der Donk, *Nat. Prod. Rep.* **2020**, DOI: 10.1039/d0np00027b.

[172] A. L. Vagstad, T. Kuranaga, S. Püntener, V. R. Pattabiraman, J. W. Bode, J. Piel, *Angew. Chem. Int. Ed.* **2019**, *131*, 2268–2272.

[173] M. F. Freeman, C. Gurgui, M. J. Helf, B. I. Morinaka, A. R. Uria, N. J. Oldham, H. G. Sahl, S. Matsunaga, J. Piel, *Science* **2012**, *338*, 387–390.

[174] P. G. Arnison, M. J. Bibb, G. Bierbaum, A. A. Bowers, T. S. Bugni, G. Bulaj, J. A. Camarero, D. J. Campopiano, G. L. Challis, J. Clardy, P. D. Cotter, D. J. Craik, M. Dawson, E. Dittmann, S. Donadio, P. C. Dorrestein, K. D. Entian, M. A. Fischbach, J. S. Garavelli, U. Göransson, C. W. Gruber, D. H. Haft, T. K. Hemscheidt, C. Hertweck, C. Hill, A. R. Horswill, M. Jaspars, W. L. Kelly, J. P. Klinman, O. P. Kuipers, A. J. Link, W. Liu, M. A. Marahiel, D. A. Mitchell, G. N. Moll, B. S. Moore, R. Müller, S. K. Nair, I. F. Nes, G. E. Norris, B. M. Olivera, H. Onaka, M. L. Patchett, J. Piel, M. J. T. Reaney, S. Rebuffat, R. P. Ross, H. G. Sahl, E. W. Schmidt, M. E. Selsted, K. Severinov, B. Shen, K. Sivonen, L. Smith, T. Stein, R. D. Süssmuth, J. R. Tagg, G. L. Tang, A. W. Truman, J. C. Vederas, C. T. Walsh, J. D. Walton, S. C. Wenzel, J. M. Willey, W. A. Van Der Donk, *Nat. Prod. Rep.* **2013**, *30*, 108–160.

[175] J. Schnitzbauer, M. T. Strauss, T. Schlichthaerle, F. Schueder, R. Jungmann, *Nat. Protoc.* **2017**, *12*, 1198–1228.

[176] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G. L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, Y. Vázquez-Baeza, *Nat. Methods* **2020**, *17*, 261–272.

[177] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, O. Nieto, in *Proceedings of the 14th Python in Science Conference*, Austin, Texas, **2015**, pp. 18–24.

[178] C. Angermueller, T. Pärnamaa, L. Parts, O. Stegle, *Mol. Syst. Biol.* **2016**, *12*, 878.

[179] Y. Lecun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436–444.

[180] M. Reutlinger, G. Schneider, *J. Mol. Graphics Modell.* **2012**, *34*, 108–117.

[181] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

[182] M. A. Hedjazi, I. Kourbane, Y. Genc, in *Proceedings of the 25th Signal Processing and Communications Applications Conference*, Antalya, Turkey, **2017**, pp. 1–4.

[183] D. T. Jones, *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 659–660.

[184] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *Proc. IEEE* **1998**, *86*, 2278–2323.

[185] N. Sapoval, A. Aghazadeh, M. G. Nute, D. A. Antunes, A. Balaji, R. Baraniuk, C. J. Barberan, R. Dannenfelser, C. Dun, M. Edrisi, R. A. L. Elworth, B. Kille, A. Kyrillidis, L. Nakhleh, C. R. Wolfe, Z. Yan, V. Yao, T. J. Treangen, *Nat. Commun.* **2022**, *13*, 1728.

[186] A. Krizhevsky, I. Sutskever, G. E. Hinton, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., **2012**.

[187] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, B. Kingsbury, *IEEE Signal Processing Magazine* **2012**, *29*, 82–97.

[188] C. Farabet, C. Couprie, L. Najman, Y. LeCun, *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1915–1929.

[189] M. Helmstaedter, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, W. Denk, *Nature* **2013**, *500*, 168–174.

[190] H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. C. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, Q. Morris, Y. Barash, A. R. Krainer, N. Jojic, S. W. Scherer, B. J. Blencowe, B. J. Frey, *Science* **2015**, *347*, 1254806.

[191] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, *JMLR* **2011**, 2493–2537.

[192] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association For Computational Linguistics, Online, **2020**, pp. 38–45.

[193] I. Sutskever, O. Vinyals, Q. V. Le, in *Proceedings of the 27th International Conference on Advances in Neural Information Processing Systems*, Curran Associates, Inc., Quebec, Canada, **2014**.

[194] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583–589.

[195] Z. Yang, X. Zeng, Y. Zhao, R. Chen, *Signal Transduction Targeted Ther.* **2023**, *8*, 1–14.

[196] C. Shorten, T. M. Khoshgoftaar, *Journal of Big Data* **2019**, *6*, 60.

[197] L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A. S. Albahri, B. S. N. Al-dabbagh, M. A. Fadhel, M. Manoufali, J. Zhang, A. H. Al-Timemy, Y. Duan, A. Abdullah, L. Farhan, Y. Lu, A. Gupta, F. Albu, A. Abbosh, Y. Gu, *Journal of Big Data* **2023**, *10*, 46.

[198] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, D. J. Inman, *MSSP* **2021**, *151*, 107398.

[199] K.-S. Oh, K. Jung, *Pattern Recognit.* **2004**, *37*, 1311–1314.

[200] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, **2016**.

[201] E. Troullinou, G. Tsagkatakis, S. Chavlis, G. F. Turi, W. Li, A. Losonczy, P. Tsakalides, P. Poirazi, *TETCI* **2021**, *5*, 755–767.

[202] H. Ma, C. Chen, Q. Zhu, H. Yuan, L. Chen, M. Shu, *Computational and Mathematical Methods in Medicine* **2021**, *2021*, 1–10.

[203] S. Kiranyaz, T. Ince, M. Gabbouj, *TBME* **2016**, *63*, 664–675.

[204] M. L. Waskom, *Journal of Open Source Software* **2021**, *6*, 3021.

[205] S. Hochreiter, J. Schmidhuber, *Neural Comput.* **1997**, *9*, 1735–1780.

[206] S. Basodi, C. Ji, H. Zhang, Y. Pan, *Big Data Min. Anal.* **2020**, *3*, 196–207.

[207] “The Ultimate Showdown: RNN vs LSTM vs GRU – Which is the Best? - Shiksha Online,” can be found under https://www.shiksha.com/online-courses/articles/rnn-vs-gru-vs-lstm/, **2023** (accessed 22 December 2023)

[208] B.-H. Kim, J.-Y. Pyun, *Sensors* **2020**, *20*, 3069.

[209] T. O’Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, others, **2019**.

[210] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Y. Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, Xiaoqiang Zheng, **2015**.

[211] Plotly Technologies Inc., "Collaborative data science," can be found under https://plot.ly, **2015**.

[212] Y. Gal, Z. Ghahramani, in *Proceedings of the 33rd International Conference on Machine Learning*, New York, USA, **2016**, pp. 1050–1059.

[213] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, S. Nahavandi, *Inf. Fusion* **2021**, *76*, 243–297.

[214] Y. Gal, Uncertainty in Deep Learning, University of Cambridge, **2016**.

[215] J. Z. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax-Weiss, B. Lakshminarayanan, *arXiv*, Vancouver, Canada, **2020**.

[216] J.-C. Yen, F.-J. Chang, S. Chang, *IEEE Trans. Image Process.* **1995**, *4*, 370–378.

[217] S. N. Uno, M. Kamiya, A. Morozumi, Y. Urano, *Chem. Commun.* **2018**, *54*, 102–105.

[218] R. Tachibana, M. Kamiya, S. Suzuki, K. Morokuma, A. Nanjo, Y. Urano, *Commun. Chem.* **2020**, *3*, 1–8.

[219] J. Tyson, K. Hu, S. Zheng, P. Kidd, N. Dadina, L. Chu, D. Toomre, J. Bewersdorf, A. Schepartz, *ACS Cent. Sci.* **2021**, *7*, 1419–1426.

[220] P. J. Macdonald, S. Gayda, R. A. Haack, Q. Ruan, R. J. Himmelsbach, S. Y. Tetin, *Anal. Chem.* **2018**, *90*, 9165–9173.

[221] Q. Qiao, W. Liu, J. Chen, X. Wu, F. Deng, X. Fang, N. Xu, W. Zhou, S. Wu, W. Yin, X. Liu, Z. Xu, *Angew. Chem. Int. Ed.* **2022**, *61*, e202202961.

[222] M. Remmel, L. Scheiderer, A. N. Butkevich, M. L. Bossi, S. W. Hell, *Small* **2023**, *19*, 2206026.

[223] Y. Zheng, Z. Ye, Y. Xiao, *Anal. Chem.* **2023**, *95*, 4172–4179.

[224] A. Martin, P. Rivera-Fuentes, *Nat. Chem.* **2024**, *16*, 28–35.

[225] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.-A. Muller, *Data Min. Knowl. Disc.* **2019**, *33*, 917–963.

[226] A. Taherkhani, G. Cosma, T. M. McGinnity, *SN Comput. Sci.* **2023**, *4*, 832.

[227] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, L. Sun, **2023**.

[228] H. Turbé, M. Bjelogrlic, C. Lovis, G. Mengaldo, *Nat. Mach. Intell.* **2023**, *5*, 250–260.

[229] P. Cui, J. Wang, *Electronics* **2022**, *11*, 3500.

[230] J. Yang, K. Zhou, Y. Li, Z. Liu, **2024**, DOI 10.48550/arXiv.2110.11334.

[231] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. A. DePristo, J. V. Dillon, B. Lakshminarayanan, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, New York, USA, **2019**, pp. 14707–14718.

[232] Y.-C. Wang, M. D. Distefano, *Curr. Top. Pept. Protein Res.* **2014**, *15*, 1–23.

[233] C. J. Howard, B. M. Floyd, A. M. Bardo, J. Swaminathan, E. M. Marcotte, E. V. Anslyn, *ACS Chem. Biol.* **2020**, *15*, 1401–1407.

[234] P. H. Le-Khac, G. Healy, A. F. Smeaton, *IEEE Access* **2020**, *8*, 193907–193934.

[235] G. Allabadi, A. Lucic, P. Pao-Huang, Y.-X. Wang, V. Adve, **2023**.

[236] S. Zelger-Paulus, M. C. A. S. Hadzic, R. K. O. Sigel, R. Börner, *RNA Spectroscopy: Methods and Protocols* (Eds.: V. Arluison, F. Wien), Springer US, New York, NY, **2020**, pp. 1–16.

[237] S. Lin, X. Yang, S. Jia, A. M. Weeks, M. Hornsby, P. S. Lee, R. V. Nichiporuk, A. T. Iavarone, J. A. Wells, F. D. Toste, C. J. Chang, *Science* **2017**, *355*, 597–602.

[238] A. Bateman, M. J. Martin, S. Orchard, M. Magrane, R. Agivetova, S. Ahmad, E. Alpi, E. H. Bowler-Barnett, R. Britto, B. Bursteinas, H. Bye-A-Jee, R. Coetzee, A. Cukura, A. D. Silva, P. Denny, T. Dogan, T. G. Ebenezer, J. Fan, L. G. Castro, P. Garmiri, G. Georghiou, L. Gonzales, E. Hatton-Ellis, A. Hussein, A. Ignatchenko, G. Insana, R. Ishtiaq, P. Jokinen, V. Joshi, D. Jyothi, A. Lock, R. Lopez, A. Luciani, J. Luo, Y. Lussi, A. MacDougall, F. Madeira, M. Mahmoudy, M. Menchi, A. Mishra, K. Moulang, A. Nightingale, C. S. Oliveira, S. Pundir, G. Qi, S. Raj, D. Rice, M. R. Lopez, R. Saidi, J. Sampson, T. Sawford, E. Speretta, E. Turner, N. Tyagi, P. Vasudev, V. Volynkin, K. Warner, X. Watkins, R. Zaru, H. Zellner, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M. C. Blatter, J. Bolleman, E. Boutet, L. Breuza, C. Casals-Casas, E. de Castro, K. C. Echioukh, E. Coudert, B. Cuche, M. Doche, D. Dornevil, A. Estreicher, M. L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, N. Hyka-Nouspikel, F. Jungo, G. Keller, A. Kerhornou, V. Lara, P. Le Mercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. B. Neto, S. Paesano, I. Pedruzzi, S. Pilbout, L. Pourcel, M. Pozzato, M. Pruess, C. Rivoire, C. Sigrist, K. Sonesson, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang, L. S. Yeh, J. Zhang, *Nucleic Acids Res.* 2021, *49*, D480–D489.

[239] P. R. A. Zanon, F. Yu, P. Musacchio, L. Lewald, M. Zollo, K. Krauskopf, D. Mrdović, P. Raunft, T. E. Maher, M. Cigler, C. Chang, K. Lang, F. D. Toste, A. I. Nesvizhskii, S. M. Hacker, *chemRxiv* 2021, DOI 10.26434/chemrxiv-2021-w7rss-v2.

[240] R. Tessier, R. K. Nandi, B. G. Dwyer, D. Abegg, C. Sornay, J. Ceballos, S. Erb, S. Cianférani, A. Wagner, G. Chaubet, A. Adibekian, J. Waser, *Angew. Chem. Int. Ed.* 2020, *59*, 10961–10970.

[241] K. Deprey, J. A. Kritzer, *Bioconjugate Chem.* 2021, *32*, 964–970.

[242] M. Tarnawski, K. Johnsson, J. Hiblot, J. Kompa, 2022, DOI 10.2210/pdb7ZIW/pdb.

[243] Y. Wang, M. Gu, *Anal. Chem.* 2010, *82*, 7055–7062.

[244] Y. Wang, D. Kuehl, S. Simonoff, N. Zhang, Q. Zheng, C. Bioscience, 2023.

[245] “SAMMI – Cerno Bioscience,” can be found under https://cernobioscience.com/sammi/, 2024 (accessed 19 March 2024)

[246] B. Mollwitz, E. Brunk, S. Schmitt, F. Pojer, M. Bannwarth, M. Schiltz, U. Rothlisberger, K. Johnsson, *Biochemistry* 2012, *51*, 986–994.

[247] R. Iversen, P. A. Andersen, K. S. Jensen, J. R. Winther, B. W. Sigurskjold, *Biochemistry* 2010, *49*, 810–820.

[248] D. Montero, C. Tachibana, J. Rahr Winther, C. Appenzeller-Herzog, *Redox Biology* 2013, *1*, 508–513.

[249] J. I. Macdonald, H. K. Munch, T. Moore, M. B. Francis, *Nat. Chem. Biol.* 2015, *11*, 326–331.

[250] L. De Rosa, R. Di Stasi, A. Romanelli, L. D. D’Andrea, *Molecules* 2021, *26*, 3521.

[251] L. J. Barber, N. D. J. Yates, M. A. Fascione, A. Parkin, G. R. Hemsworth, P. G. Genever, C. D. Spicer, *RSC Chem. Biol.* 2023, *4*, 56–64.

[252] J.-R. Deng, N. C.-H. Lai, K. K.-Y. Kung, B. Yang, S.-F. Chung, A. S.-L. Leung, M.-C. Choi, Y.-C. Leung, M.-K. Wong, *Commun. Chem.* 2020, *3*, 1–9.

[253] A. M. Weeks, J. A. Wells, *Curr. Protoc. Chem. Biol.* 2020, *12*, e79.

[254] J.-P. Colletier, B. Chaize, M. Winterhalter, D. Fournier, *BMC Biotechnology* 2002, *2*, 9.

[255] G. B. Schober, S. Story, D. P. Arya, *Sci. Rep.* 2024, *14*, 2403.

[256] “What Is The Transition Temperature Of The Lipid?,” can be found under https://avantilipids.com/tech-support/faqs/transition-temperature, 2024 (accessed 26 February 2024)

[257] H. Cai, S. J. Wind, *Langmuir* 2016, *32*, 10034–10041.

[258] M.-L. Visnapuu, D. Duzdevich, E. C. Greene, *Mol. BioSyst.* 2008, *4*, 394–403.

[259] R. M. Pasternack, S. Rivillon Amy, Y. J. Chabal, *Langmuir* **2008**, *24*, 12963–12971.

[260] V. V. Naik, M. Crobu, N. V. Venkataraman, N. D. Spencer, *J. Phys. Chem. Lett.* **2013**, *4*, 2745–2751.

[261] V. V. Naik, R. Städler, N. D. Spencer, *Langmuir* **2014**, *30*, 14824–14831.

[262] M. Wang, K. M. Liechti, Q. Wang, J. M. White, *Langmuir* **2005**, *21*, 1848–1857.

[263] N.-P. Huang, J. Vörös, S. M. De Paul, M. Textor, N. D. Spencer, *Langmuir* **2002**, *18*, 220–230.

[264] T. Böcking, F. Aguet, S. C. Harrison, T. Kirchhausen, *Nat. Struct. Mol. Biol.* **2011**, *18*, 295–301.

[265] M. Zhao, P. R. Nicovich, M. Janco, Q. Deng, Z. Yang, Y. Ma, T. Böcking, K. Gaus, J. J. Gooding, *Langmuir* **2018**, *34*, 10012–10018.

[266] J. C. Bischof, X. He, *Ann. N. Y. Acad. Sci.* **2006**, *1066*, 12–33.

[267] A. Cook, F. Walterspiel, C. Deo, *ChemBioChem* **2023**, *24*, e202300022.

[268] J. Emonts, J. F. Buyel, *Comput. Struct. Biotechnol. J.* **2023**, *21*, 3234–3247.

[269] D. Kim, S. Stoldt, M. Weber, S. Jakobs, V. N. Belov, S. W. Hell, *Chemistry - Methods* **2023**, *3*, e202200076.

[270] S. M. Hacker, K. M. Backus, M. R. Lazear, S. Forli, B. E. Correia, B. F. Cravatt, *Nat. Chem.* **2017**, *9*, 1181–1190.

[271] C. Li, A. G. Tebo, M. Thauvin, M. Plamont, M. Volovitch, X. Morin, S. Vriz, A. Gautier, *Angew. Chem. Int. Ed.* **2020**, *132*, 18073–18079.

[272] W. F. An, N. Tolliday, *Mol Biotechnol* **2010**, *45*, 180–186.

[273] L. Wang, M. Tran, J. Roberti, B. Koch, L. Xue, K. Johnsson, *Nat. Chem.* **2020**, *12*, 165–172.

[274] C. L. Wardzala, Z. S. Clauss, J. R. Kramer, *Front. Cell Dev. Biol.* **2022**, *10*.

[275] E. M. Krauland, B. R. Peelle, K. D. Wittrup, A. M. Belcher, *Biotechnol. Bioeng.* **2007**, *97*, 1009–1020.

[276] R. E. Herman, D. Badders, M. Fuller, E. G. Makienko, M. E. Houston, S. C. Quay, P. H. Johnson, *J. Biol. Chem.* **2007**, *282*, 9813–9824.

[277] K. Y. Dane, L. A. Chan, J. J. Rice, P. S. Daugherty, *J. Immunol. Methods* **2006**, *309*, 120–129.

[278] R. Wolkowicz, G. C. Jager, G. P. Nolan, *J. Biol. Chem.* **2005**, *280*, 15195–15201.

[279] M. Mei, Y. Zhou, W. Peng, C. Yu, L. Ma, G. Zhang, L. Yi, *Microbiol. Res.* **2017**, *196*, 118–128.

[280] X. Zhao, M. Rahman, Z. Xu, T. Kasputis, Y. He, L. Yuan, R. C. Wright, J. Chen, *J. Agric. Food Chem.* **2023**, *71*, 8665–8672.

[281] B. Valldorf, S. C. Hinz, G. Russo, L. Pekar, L. Mohr, J. Klemm, A. Doerner, S. Krah, M. Hust, S. Zielonka, *Biol. Chem.* **2022**, *403*, 455–477.

[282] X. Yang, H. Tang, M. Song, Y. Shen, J. Hou, X. Bao, *Microb. Cell Fact.* **2019**, *18*, 85.

[283] C. Zhang, H. Chen, Y. Zhu, Y. Zhang, X. Li, F. Wang, *Front. Bioeng. Biotechnol.* **2022**, *10*.

[284] K. V. Teymennet-Ramírez, F. Martínez-Morales, M. R. Trejo-Hernández, *Front. Bioeng. Biotechnol.* **2022**, *9*.

[285] E. T. Boder, K. D. Wittrup, *Nat. Biotechnol.* **1997**, *15*, 553–557.

[286] S. R. Adams, R. E. Campbell, L. A. Gross, B. R. Martin, G. K. Walkup, Y. Yao, J. Llopis, R. Y. Tsien, *J. Am. Chem. Soc.* **2002**, *124*, 6063–6076.

[287] B. R. Martin, B. N. G. Giepmans, S. R. Adams, R. Y. Tsien, *Nat. Biotechnol.* **2005**, *23*, 1308–1314.

[288] S. C. Alexander, A. Schepartz, *Org. Lett.* **2014**, *16*, 3824–3827.

[289] L. Benatuil, J. M. Perez, J. Belk, C.-M. Hsieh, *Protein Eng., Des. Sel.* **2010**, *23*, 155–159.

[290] C. J. A. Sigrist, E. de Castro, L. Cerutti, B. A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, I. Xenarios, *Nucleic Acids Res.* **2013**, *41*, D344–D347.

[291] H. Ma, S. Kunes, P. J. Schatz, D. Botstein, *Gene* **1987**, *58*, 201–216.

[292] O. Kevin R., K. T. Vo, S. Michaelis, C. Paddon, *Nucleic Acids Res.* **1997**, *25*, 451–452.

[293] K. Ip, R. Yadin, K. W. George, in *DNA Cloning and Assembly: Methods and Protocols* (Eds.: S. Chandran, K.W. George), Springer US, New York, NY, **2020**, pp. 79–89.

[294] G. Chao, W. L. Lau, B. J. Hackel, S. L. Sazinsky, S. M. Lippow, K. D. Wittrup, *Nat. Protoc.* **2006**, *1*, 755–768.

[295] W. A. Marinaro, R. Prankerd, K. Kinnari, V. J. Stella, *J. Pharm. Sci.* **2015**, *104*, 1399–1408.

[296] C. A. Kettner, A. B. Shenvi, *J. Biol. Chem.* **1984**, *259*, 15106–15114.

[297] D. G. Hall, *Boronic Acids: Preparation and Applications in Organic Synthesis, Medicine and Materials*, Wiley-VCH Verlag GmbH & Co. KGaA., **2011**, pp. 1–132.

[298] D. B. Diaz, C. C. G. Scully, S. K. Liew, S. Adachi, P. Trinchera, J. D. St. Denis, A. K. Yudin, *Angew. Chem. Int. Ed.* **2016**, *55*, 12659–12663.

[299] J. D. St. Denis, A. Zajdlik, J. Tan, P. Trinchera, C. F. Lee, Z. He, S. Adachi, A. K. Yudin, *J. Am. Chem. Soc.* **2014**, *136*, 17669–17673.

[300] L. An, M. Said, L. Tran, S. Majumder, I. Goreshnik, G. R. Lee, D. Juergens, J. Dauparas, I. Anishchenko, B. Coventry, A. K. Bera, A. Kang, P. M. Levine, V. Alvarez, A. Pillai, C. Norn, D. Feldman, D. Zorine, D. R. Hicks, X. Li, M. G. Sanchez, D. K. Vafeados, P. J. Salveson, A. A. Vorobieva, D. Baker, **2023**, 2023.12.20.572602.

[301] N. R. Bennett, B. Coventry, I. Goreshnik, B. Huang, A. Allen, D. Vafeados, Y. P. Peng, J. Dauparas, M. Baek, L. Stewart, F. DiMaio, S. De Munck, S. N. Savvides, D. Baker, *Nat. Commun* **2023**, *14*, 2625.

[302] R. Krishna, J. Wang, W. Ahern, P. Sturmfels, P. Venkatesh, I. Kalvet, G. R. Lee, F. S. Morey-Burrows, I. Anishchenko, I. R. Humphreys, R. McHugh, D. Vafeados, X. Li, G. A. Sutherland, A. Hitchcock, C. N. Hunter, A. Kang, E. Brackenbrough, A. K. Bera, M. Baek, F. DiMaio, D. Baker, *Science* **2024**, *0*, eadl2528.

[303] H. Nonaka, S. Tsukiji, A. Ojida, I. Hamachi, *J. Am. Chem. Soc.* **2007**, *129*, 15777–15779.

[304] M. Sunbul, L. Nacheva, A. Jäschke, *Bioconjugate Chem.* **2015**, *26*, 1466–1469.

[305] T. Huang, Y. Li, *The Innovation* **2023**, *4*, 100446.

[306] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, *Science* **2021**, *373*, 871–876.

[307] S. Van Der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, *PeerJ* **2014**, DOI 10.7717/peerj.453.

[308] F. Chollet, others, **2015**.

[309] D. P. Kingma, J. Ba, *arXiv* **2017**.

[310] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, A. Bairoch, *The Proteomics Protocols Handbook* (Ed.: J.M. Walker), Humana Press, Totowa, NJ, **2005**, pp. 571–607.

Curriculum vitae

Salome Püntener
salome.puentener@gmx.ch
www.linkedin.com/in/salome-püntener-295561142
https://orcid.org/0000-0003-1877-025X

Education

04.2019 –05.2024 **Doctorate, EPF Lausanne**
Doctoral thesis, Prof. Dr. Pablo Rivera-Fuentes:
“Development of tools and methods for protein identification: From single molecules to in vivo applications”
02.2022 – 2024: University of Zürich
08.2019 – 02.2022: EPF Lausanne
04.2019 – 08.2019: ETH Zürich

09.2016 – 02.2019 **MSc in Interdisciplinary Sciences, ETH Zürich**
Main focus area: Chemistry and Biology
Master’s thesis, Prof. Dr. Pablo Rivera-Fuentes:
“Photoactivatable probes for in vivo cell tracking”

09.2013 – 02.2017 **BSc in Interdisciplinary Sciences, ETH Zürich**
Main focus area: Chemistry and Biology
Bachelor’s thesis, Prof. Dr. Jörn Piel:
“Exploring the substrate scope of epimerase OspD”

09.2006 – 05.2012 **Matura, Kantonsschule Zug**
Focus subjects: Biology and Chemistry

Practical Experience

08.2019 – 2024 **Research Assistant, EPF Lausanne and University of Zürich**
- Design and synthesis of fluorophores and fluorophore conjugates
- Protein expression and purification
- Yeast surface display for a peptide screen
- Single-molecule imaging
- Analysis of microscopy data using Fiji and Python
- Application of machine learning toward data analysis
- Supervised students in the laboratory and report writing
- Participating in organizing and conducting the move, setup, and organization of the laboratories as a safety officer

06.2018 – 12.2018 **Intern in Fragrance Discovery, Givaudan Schweiz**
- Synthesis of fragrance molecule candidate library
- Short natural product synthesis
- Preliminary olfactory evaluation of molecules

09.2017 – 12.2017 **Semester Project, ETH Zürich,** Group of Prof. Dr. Erick M. Carreira
- Ligand synthesis

- Iridium-catalysed enantioselective alkylation of allenylic carbonates

09.2016 – 12.2016 and 09.2015 – 12.2015 **Teaching Assistant, ETH Zürich,** Lecturer: Prof. Dr. Antonio Togni
Lecture: “General chemistry I: Inorganic chemistry”
- Preparation and teaching an exercise group
- Correction and discussion of problem sets

09.2015 – 12.2015 **Semester Project, ETH Zürich,** Group of Prof. Dr. François Diederich
- Optimization and exploration of the reaction conditions of the cycloaddition-retroelectrocyclization
- Synthesis of precursor molecules

Publications

2023 X. Yang, D. Chen, Q. Sun, Y. Wang, Y. Xia, J. Yang, C. Lin, X. Dang, Z. Cen, D. Liang, R. Wei, Z. Xu, G. Xi, G. Xue, C. Ye, L.-P. Wang, P. Zou, S.-Q. Wang, P. Rivera-Fuentes, S. Püntener, Z. Chen, Y. Liu, J. Zhang, Y. Zhao, *Cell Discov* **2023**, *9*, 1–26.

S. Püntener, P. Rivera-Fuentes, *J. Am. Chem. Soc.* **2023**, *145*, 1441–1447.

2020 E. A. Halabi, J. Arasa, S. Püntener, V. Collado-Diaz, C. Halin, P. Rivera-Fuentes, *ACS Chem. Biol.* **2020**, *15*, 1613–1620.

2019 A. L. Vagstad, T. Kuranaga, S. Püntener, V. R. Pattabiraman, J. W. Bode, J. Piel, *Angew. Chem. Int. Ed.* **2019**, *58*, 2246–2250.

2018 E. A. Halabi, S. Püntener, P. Rivera-Fuentes, *HCA* **2018**, *101*, e1800165.

2016 T. A. Reekie, E. J. Donckèle, G. Manenti, S. Püntener, N. Trapp, F. Diederich, *Org. Lett.* **2016**, *18*, 2252–2255.

Conferences

2023 *Oral presentation* at the EMBL Symposium: Seeing is Believing – Imaging the Molecular Processes of Life, Heidelberg
Oral presentation at the SCS Fall Meeting 2023, Bern
Poster at the GRC Chemical Imaging, Easton
Oral presentation at the CHanalysis 2023, Beatenberg
Poster at the 4th Swiss Industrial Chemistry Symposium 2023, Basel
Oral presentation at the Harvard Chemistry Future Leaders Symposium 2023, Cambridge

2022 *Participation* at the GRC Single-Molecule Approaches to Biology, Castelldefels

2021 *Poster* at the Advances in Chemical Biology 2021, online
Flash presentation at the French Swiss Photochemistry Symposium, online

Awards

2023 SCNAT/SCS *Travel Award* 2023
SNSF Scientific *Image Competition* 2023 Award
Poster prize at the 4th Swiss Industrial Chemistry Symposium

2021 *Best Flash presentation prize* at the French Swiss Photochemistry Symposium

2019 *ETH medal* for Outstanding Master’s thesis in Interdisciplinary Sciences

2018 *S. & N. Blank Preis* for the best Master’s thesis in organic, bioorganic, or materials science at ETH Zürich

Language Skills

German	Native
English	Proficient; Cambridge Advanced (CAE)
French	Advanced knowledge; DALF Level C1 in 2012, 08.2012 – 12.2012 Language stay in Bordeaux

IT Skills

Python intermediate level (experience with Pandas, Scikit-learn, Tensorflow), MS Office, ChemDraw, Fiji