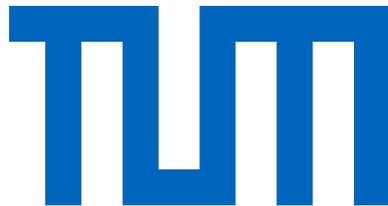


Technical University of Munich



Surface Modification and Topographical Analysis  
of Biomedical Interfaces

A thesis presented by

**Carolin Alexandra Rickert**

TUM School of Engineering and Design,  
Department of Materials Engineering

Professur für Biomechanik

# Surface Modification and Topographical Analysis of Biomedical Interfaces

**Carolin Alexandra Rickert**

Vollständiger Abdruck der von der TUM School of Engineering and Design der  
Technischen Universität München zur Erlangung einer

**Doktorin der Ingenieurwissenschaften (Dr.-Ing.)**

genehmigten Dissertation.

Vorsitz: Prof. Dr. Alisa Machner

Prüfer\*innen der Dissertation:

1. Prof. Dr. rer. nat. Oliver Lieleg
2. Prof. Dr. Thorsten Hugel
3. Prof. Dr.-Ing. Michael W. Gee

Die Dissertation wurde am 16.05.2023 bei der Technischen Universität München eingereicht  
und durch die TUM School of Engineering and Design am 21.10.2023 angenommen.



## Summary

Medical devices are indispensable tools in modern healthcare; they facilitate diagnosis, treatment, and patient care. Even though being carefully designed to serve specific purposes, however, those devices still pose the risk of inducing adverse side effects, which can potentially harm patients or even lead to life-threatening conditions. In this context, biopolymeric materials are promising candidates to tailor the material and surface properties of medical devices such that they integrate more seamlessly into the body environment. A highly interesting biopolymer is the multifunctional glycoprotein mucin, which combines several biomedically interesting properties such as excellent lubricity, good anti-biofouling properties, and high biocompatibility. In this thesis, mucin molecules are used to 3D-print objects that exhibit mechanical properties similar to soft tissue. It is shown that those printed objects exhibit high biocompatibility, good anti-biofouling properties, and immunomodulating activity – characteristics that hold great potential for a wide range of biomedical applications, including tissue engineering, wound healing, and soft robotics.

Furthermore, mucin macromolecules are used to create coatings on existing medical devices, aiming at transferring the beneficial properties of mucins to the respective surface. Taking contact lenses as an example, it is demonstrated that those coatings are indeed able to considerably improve the surface wettability and to prevent wear formation on tissue. These characteristics might be beneficial for mitigating the disruptive influence of a broad variety of medical devices and to improve the integration of synthetic materials into physiological environments. Remarkably, owing to their high transparency, the generated mucin coatings are well-suited to be used on optical devices. Here, the extraordinary anti-biofouling properties of the mucin layer can even help preserving the clearness of optical interfaces that are exposed to bodily fluids or tissues.

However, to meet the sterility required for a medical product, the mucins or the mucin-containing objects have to be treated with radiation, heat, or chemicals, and this could compromise the physico-chemical or structural properties of the biomolecule. In this thesis, it is shown that mucins can tolerate harsh decontamination treatments without suffering significant functional impairment or structural degradation. Whereas  $\gamma$  irradiation is identified as the most effective method for sterilizing lyophilized mucins, fumigation with ethylene oxide is the best method for maintaining the biochemical integrity and functionality of mucin coatings on medical materials. Overall, the presented results demonstrate an astonishing sturdiness of mucins towards selected sterilization strategies, which is a major step towards a clinical application of the glycoprotein.

An important aspect to be evaluated considering the safety of medical devices is their tribological performance on tissue. This not only includes determining friction coefficients, but, even more importantly, it involves evaluating putative wear formation. Assessing the surface quality of complex biological surfaces, however, requires time-consuming expert involvement and might still remain prone to errors and biases. To overcome these challenges, in this thesis, the potential ML holds to aid in the assessment of the surface conditions of a soft tissue is explored. It is shown that both, supervised and unsupervised ML methods, can achieve a meaningful surface classification using a set of topographical parameters as input features. Similar approaches could streamline material design processes, enabling precise customization of a material's surface behavior that would otherwise necessitate time-consuming trial-and-error investigations. To increase the accuracy of the ML model, a correlation-driven algorithm is developed to perform feature elimination – an essential step to

overcome the challenges encountered due to limited data availability when applying ML to experimental datasets. Among the different unsupervised ML models tested to automatically sort corneal samples without providing any output label, a  $k$ -Means clustering algorithm is demonstrated to be the most suitable for the task at hand. From a material science point of view, the algorithm indeed successfully generates four clusters that accurately sort the samples based on their surface appearance. Additionally, it is able to achieve a similar surface classification as conducted manually for the tribological evaluation of rigid contact lenses. Overall, it is shown that the presented ML-driven approach can guide and improve the analysis of datasets that – either owing to their sample size or their complexity – go beyond the cognitive capabilities of humans.

# Contents

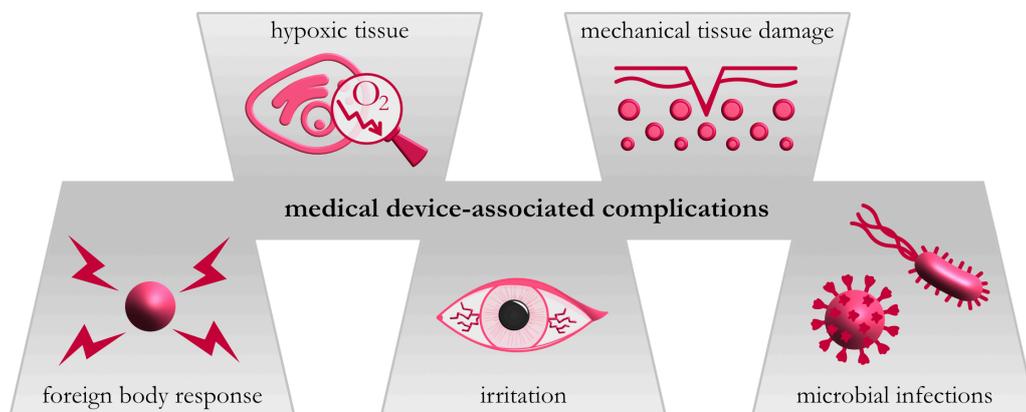
1.	Introduction .....	1
2.	Materials and methods.....	5
2.1.	Mucin macromolecules.....	5
2.2.	Producing and characterizing mucin-based bio-ink samples .....	6
2.2.1.	Functionalization of mucins with methacryloyl groups.....	6
2.2.2.	Preparation and curing of the bio-ink.....	7
2.2.3.	3D printing process .....	8
2.2.4.	Oscillatory shear rheometry .....	8
2.2.5.	Unconfined compression tests.....	9
2.2.6.	Swelling experiments .....	9
2.3.	Medical devices .....	10
2.4.	Generating and evaluating mucin coatings on polymeric surfaces .....	11
2.4.1.	Passively adsorbed mucin coatings.....	11
2.4.2.	Covalent mucin coatings.....	11
2.4.3.	Methods to detect mucin glycoproteins .....	12
2.4.4.	Wettability evaluation .....	14
2.4.5.	Transparency measurements .....	15
2.4.6.	Lipid deposition tests .....	16
2.5.	Tribology .....	17
2.5.1.	Rotational tribology measurements.....	18
2.5.2.	Ocular tribology measurements.....	19
2.6.	Laser scanning profilometry .....	20
2.7.	Sterilization and disinfection techniques .....	21
2.7.1.	Thermal treatments.....	22
2.7.2.	Autoclaving .....	22
2.7.3.	UV irradiation.....	22
2.7.4.	Treatment with $\gamma$ rays .....	23
2.7.5.	Ethylene oxide fumigation.....	23
2.8.	Supervised Machine Learning .....	24
2.8.1.	Sample datasets from biofilms and plant leaves.....	24
2.8.2.	Supervised Machine Learning algorithms .....	26
2.8.3.	Repeated stratified $k$ -fold cross-validation .....	28
2.9.	Correlation-based feature elimination.....	29
2.9.1.	Molecular dataset.....	29
2.9.2.	Network-based, correlation-driven redundancy elimination (NETCORE) .....	30

2.9.3.	Evaluation regarding multicollinearity.....	33
2.10.	Unsupervised Machine Learning.....	33
2.10.1.	Generation of a dataset containing corneal images.....	33
2.10.2.	Data preparation and preprocessing.....	34
2.10.3.	<i>k</i> -Means clustering.....	34
2.10.4.	Intrinsic validation metrics.....	35
2.10.5.	Extrinsic validation metrics.....	36
2.10.6.	Panel evaluation of the corneal dataset.....	36
2.11.	Statistical analysis and graphical representation.....	37
3.	Mucin macromolecules for biomedical applications.....	39
3.1.	UV-curable mucin gels for 3D bioprinting.....	39
3.2.	Covalent mucin coatings on silicone contact lenses.....	45
3.3.	Covalent mucin coatings on rigid, gas-permeable contact lenses.....	50
3.4.	Sterilization of mucin macromolecules.....	57
3.5.	Sterilization of covalent mucin coatings on medical devices.....	62
4.	Machine Learning to analyze complex surfaces.....	67
4.1.	Analyzing the surfaces of biological materials.....	68
4.2.	Correlation-based feature elimination.....	77
4.3.	Analyzing corneal tissue <i>via</i> unsupervised Machine Learning.....	86
5.	Summary and outlook.....	97
Appendix	.....	101
<b>Appendix A1:</b>	Additional experimental methods and results.....	101
<b>Appendix A2:</b>	Modifications of the covalent mucin coating process.....	116
<b>Appendix A3:</b>	Overview of topographical surface parameters.....	119
<b>Appendix A4:</b>	Characterization of the tribological setup.....	124
<b>Appendix A5:</b>	Details on supervised ML for biofilm and plant samples.....	125
<b>Appendix A6:</b>	Molecular datasets.....	134
<b>Appendix A7:</b>	NETCORE vs. other correlation-based approaches.....	139
<b>Appendix A8:</b>	Supporting material for unsupervised ML.....	145
List of abbreviations	.....	157
Bibliography	.....	159
Author contributions	.....	179
Full list of publications	.....	181
Acknowledgment.....	.....	183

# 1. Introduction

From pacemakers to insulin pumps, from endoscopes to pulse oximeters – medical devices are indispensable tools in modern healthcare, where they aid in the diagnosis and treatment of various diseases and health conditions. Endotracheal tubes, for example, are used to quickly and efficiently provide a clear airway and to ensure that a patient's breathing remains stable – a crucial action to prevent life-threatening respiratory failure. Similarly, catheters have revolutionized patient care by enabling the accurate delivery of medication into and the drainage of bodily fluids out of the body in a less invasive manner compared to traditional surgical procedures. Another (often-underestimated) class of medical devices are contact lenses. In addition to correcting even highly complex refractive errors (such as aphakia,<sup>1,2</sup> astigmatism,<sup>3,4</sup> or high anisometropia<sup>5</sup>), contact lenses can also be applied for the therapy of persistent epithelial defects,<sup>6</sup> to support post-refractive surgery rehabilitation,<sup>7</sup> or as sensors for ocular diagnostics.<sup>8</sup>

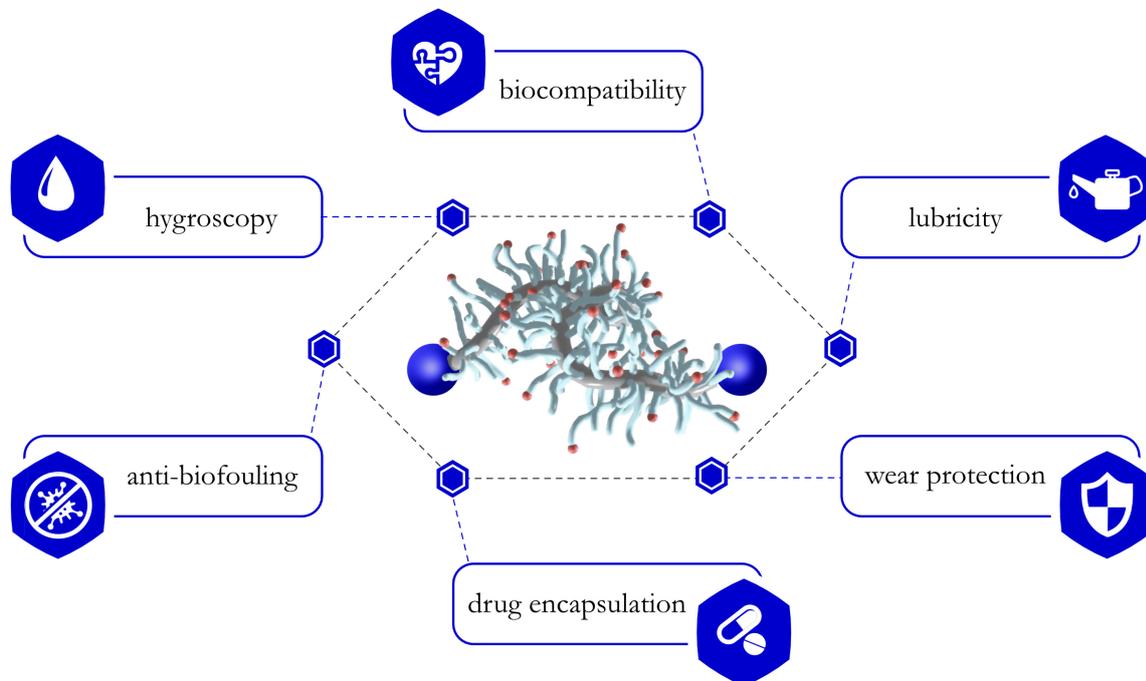
However, the great improvements achieved with such medical devices can be accompanied by severe, unintended side effects (**Figure 1.1**). An artificial material that comes into contact with human tissue poses the risk of entailing a foreign body response:<sup>9,10</sup> The immune system then reacts to the presence of the device as if it was an invading pathogen. Several immune reactions, such as pain, swelling, fever, or the formation of fluid-filled cysts around the devices can be triggered. This cannot only entail discomfort and scarring of the tissue, but it can also lead to the malfunction of the device or life-threatening further complications. In addition to triggering foreign body responses, the artificial materials medical devices mostly consist of differ from human tissue regarding their mechanical and physical properties – and this can negatively influence the physiological working principles of the body. The avascular cornea of the human eye, for instance, receives its oxygen supply through the tear film.<sup>11,12</sup> Contact lenses, however, constitute a physical barrier that limits the transport of oxygen towards the corneal surface – and this can entail hypoxic complications such as corneal neovascularization or limbal erythema.<sup>13-15</sup> Moreover, compared to tissue, artificial materials are usually quite hard, stiff, and often exhibit insufficient wettability.<sup>16,17</sup> Owing to this, those materials can impair the inherent lubrication mechanisms and induce friction, irritation, or even injured tissue and inflammations.<sup>18,19</sup> This disruptive influence can furthermore severely weaken the endogenous protective barriers of the body, leaving patients vulnerable to infections. Aggravatingly, an artificial material is typically much more prone to bacterial colonization.<sup>20,21</sup> Hence, such devices may bring pathogenic agents in close proximity to injured tissue – a direct pathway to microbial infections.



**Figure 1.1: Possible complications of medical devices.** Medical devices can entail several different complications, including foreign body responses, microbial infections, hypoxia, irritation, and mechanical damage.

## Introduction

In this context, biopolymeric materials that can integrate more seamlessly into the tissue environment have emerged as a promising solution to those challenges associated with medical device safety and effectiveness:<sup>22</sup> Many biopolymeric materials, such as proteins, carbohydrates, and lipids, that are derived from natural sources, combine a wide range of beneficial properties including good biocompatibility, lubricity, and anti-biofouling behavior (**Figure 1.2**).<sup>23-27</sup>



**Figure 1.2: Beneficial properties of biopolymers.** Biopolymers (exemplarily depicted is a mucin glycoprotein; see **chapter 2.1**) can exhibit several properties that can be beneficial for different biomedical applications. Those properties include good biocompatibility, hygroscopy, lubricity, wear protection, anti-biofouling behavior, and the possibility to be used for drug encapsulation.

Indeed, several medical products made from biopolymeric materials showed excellent performance in clinical applications. Collagen-based surgical sutures<sup>28,29</sup> and tissue bonding agents,<sup>30</sup> for instance, were shown to promote faster wound healing and induce less inflammations compared to commercial, synthetic sutures. Biodegradable stents made of poly(lactic acid) – a semi-synthetic biopolymer – are similarly effective for the treatment of coronary artery disease as their metallic counterparts<sup>31</sup> but they seem to pose lower risks of restenosis.<sup>32</sup> Other examples of biopolymeric medical devices include chitosan-based wound dressings with hemostatic and antibacterial properties,<sup>33,34</sup> hyaluronic acid-based intraocular implants that help preventing bacterial infections after ophthalmic surgery,<sup>35</sup> and resorbable silk-based fracture fixation screws that allow for incorporating antibacterial agents while maintaining a high mechanical stability of the screw.<sup>36,37</sup>

In addition to serving as a base material for biomedical products, biopolymers can be used to coat existing medical devices to tune their surface properties in terms of *e.g.*, wetting resistance, friction response, biocompatibility, or bio-fouling. Coatings with the polysaccharide heparin were, for example, shown to reduce encrustation and thromboembolic complications of various cardiovascular medical devices.<sup>38,39</sup> Similarly, orthopedic implants carrying polyglutamic acid coatings showed improved biocompatibility and strong antibacterial properties.<sup>40,41</sup> Other biopolymeric coatings, such as those generated with cyclodextrin<sup>42,43</sup> or chitosan,<sup>44,45</sup> were used for sustained drug release from medical surfaces while enhancing several other properties including mucoadhesion or osteoinductivity.

A biopolymer that has recently gained increasing attention in biomedical research is mucin – a densely glycosylated macromolecule that is produced by goblet cells of mammalian mucosa.<sup>46,47</sup> In the human body, mucin glycoproteins not only serve as lubricants and protective barriers on mucosal tissues, but they also ensure sufficient tissue hydration and act as a defense mechanism against bacteria and other pathogens.<sup>48-50</sup> In the first part of this thesis, the potential mucin molecules hold as a versatile material for different biomedical applications is investigated. Specifically, two different approaches are followed to integrate mucins into medical products: First, a mucin-based bio-ink is developed that allows for generating 3-dimensional mucin-based objects by the means of 3D bioprinting. Second, mucin macromolecules are used to create coatings on existing medical devices, aiming at transferring the beneficial properties of mucins to the respective surface. The generated objects and coated devices are then characterized in terms of various physical aspects, and the extent to which these new materials exhibit specific mucin-related properties is assessed. Additionally, as the mucin macromolecules are derived from animal sources (where sterility cannot be guaranteed even after purification), the stability of those mucins towards different sterilization techniques is investigated – a critical aspect for a material to be used in a medical context.

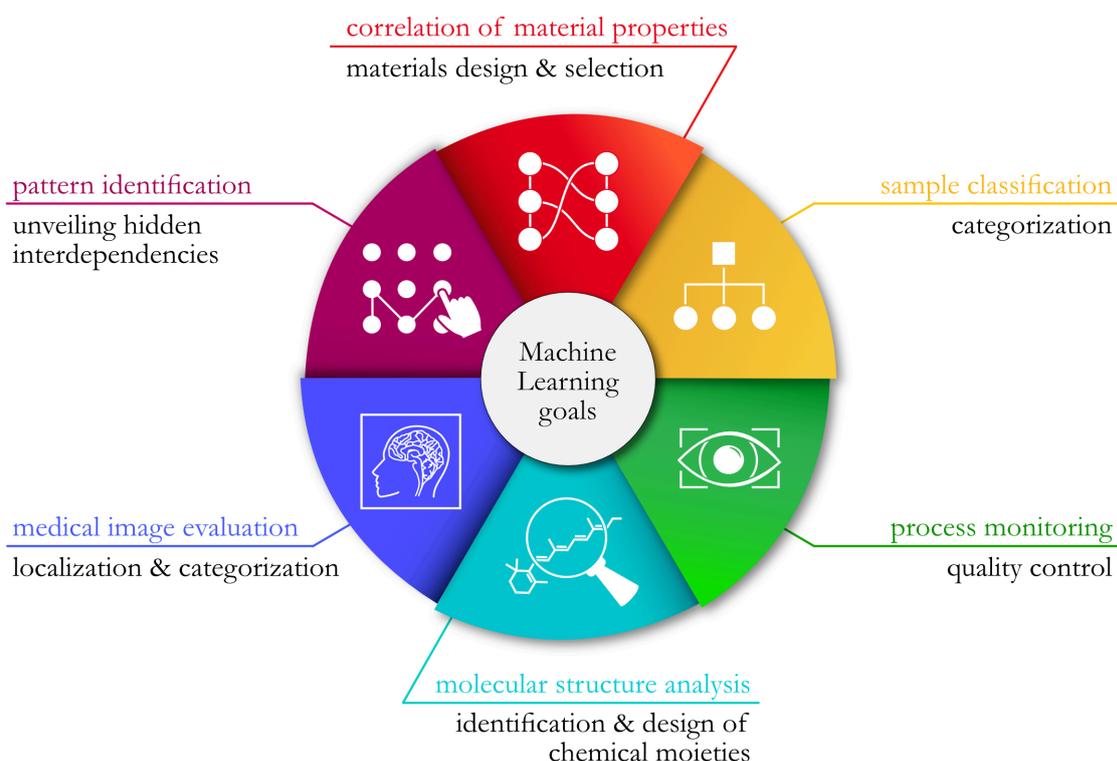
The second part of this thesis focusses more extensively on the performance analysis of mucin-coated medical devices. As mentioned above, one of the most notable characteristics of mucin macromolecules is their excellent lubricity. Hence, when investigating the functional improvements brought about by mucin coatings generated on medical devices, one important aspect to be evaluated is the tribological performance of such devices on tissue they come into contact with. This not only includes the assessment of the friction coefficients of the system, but – even more importantly – it involves assessing putative tissue damages inflicted by the applied stress (or the absence of those). However, for soft materials and complex damage patterns of different length scales, this is far from trivial: on the one hand, the initial conditions of biological tissue samples can vary considerably; and on the other hand, there is no direct and unambiguous method or metric to assess the quality of such multifaceted surfaces. One approach that has been used in previous studies<sup>51-53</sup> to evaluate the appearance of biological surfaces is to employ high-resolution imaging techniques, such as laser scanning profilometry, that return 3-dimensional topographical profiles. Visual inspections and surface descriptors obtained based on those profiles can then guide an estimation of the overall surface quality. Even with this approach, however, the assessment of the surface quality still requires time-consuming manual assessments by a trained expert and may still remain prone to errors and biases.

Here, Machine Learning-driven (ML-driven) decision processes may provide a remedy: As a sub-category of data science, ML focuses on developing algorithms that learn from provided data to make predictions or to draw conclusions. By analyzing multidimensional datasets, ML can create generalized, descriptive models that otherwise the human mind could hardly identify. The versatility of ML applications is virtually limitless, ranging from object detection<sup>54,55</sup> to process monitoring,<sup>56,57</sup> the identification of patterns or correlations within a dataset,<sup>58,59</sup> as well as data classification,<sup>60-62</sup> regression,<sup>63,64</sup> or clustering<sup>65,66</sup> (**Figure 1.3**).

Based on their working principles, ML approaches can be roughly subdivided into three categories: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, labelled data is used to train a model to map input features (*i.e.*, properties characterizing a sample) to a pre-defined output label. This principle is similar to how parents teach their children to name objects. Unsupervised learning, in contrast, does not rely on labelled data. Unsupervised learning is a data-driven approach in which the algorithm is only provided with unsorted input data; and this input data is then analyzed to identify trends, patterns, or clusters. Last, reinforcement-based learning is an action-based training approach. Here, the model is allowed to explore certain actions and it is trained to make the

## Introduction

'right' choice using a reward function. For all of those approaches, a broad variety of available algorithms and models exists (for examples, please refer to **chapters 2.8** and **2.10**). Some of those models come from the field of deep learning. Deep learning approaches – which can be all, supervised, unsupervised, or reinforcement-based – include algorithms that aim at mimicking the working principle of the human brain. Therefore, data is analyzed by multi-hierarchical structures of interconnected layers of digital neurons. As those constructs can handle data on different levels of abstraction, they are promising approaches for complex problems. This, however, comes at a prize: owing to the complex intrinsic structure of those models, the decision-making process is barely traceable.



**Figure 1.3: Typical objectives of ML approaches.** Typical goals of ML-driven analyses include the correlation of material properties, the classification of samples, the identification of patterns, process monitoring, molecular structure analysis, and the evaluation of medical images. *Adapted with permission from Rickert et al.<sup>67</sup>.*

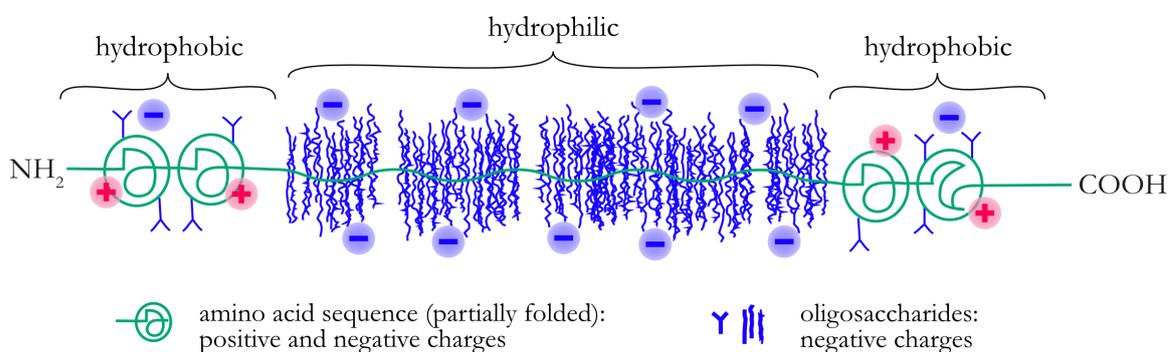
In this thesis, the potential ML holds to aid in the assessment of the surface conditions of a soft tissue is explored. In a first step, supervised ML models are employed to test whether complex biological surfaces can be accurately classified based on a set of surface parameters obtained from captured 3D profiles. For such supervised learning approaches, the samples of the dataset are not only characterized by the mentioned descriptors (features), but each sample additionally carries a predefined label. Afterwards, an unsupervised approach is employed to classify corneal tissue samples. Here, each sample is solely described by the input features and no predefined label is provided. The classification results achieved using the surface descriptors as input features are then compared to a manual classification conducted by humans.

## 2. Materials and methods

If not stated otherwise, all chemicals were obtained from Carl Roth GmbH & Co. KG (Karlsruhe, Germany). Double distilled water (ddH<sub>2</sub>O, pH 7.3) was freshly prepared for each experiment.

### 2.1. Mucin macromolecules

Mucin glycoproteins are the structural and functional key component of mammalian mucus, the viscoelastic hydrogel that lines all wet epithelia such as ocular surfaces,<sup>11</sup> the respiratory system,<sup>68</sup> and the gastrointestinal tract.<sup>50</sup> With regard to their functional and structural characteristics, mucin glycoproteins can be subdivided into either membrane-anchored or secreted mucins – the latter of which can additionally be distinguished into gel-forming and non-gel-forming molecules. The class of secreted, gel-forming mucins comprises five different isotopes (MUC2, MUC5AC, MUC5B, MUC6, MUC19),<sup>69,70</sup> which are expressed differently throughout the body but share a similar structure (**Figure 2.1**): mucin macromolecules have high molecular weights in the range of several MDa<sup>48</sup> and comprise a long protein backbone that is unfolded in the central region and partially folded in the terminal domains. The densely glycosylated central part of the molecule carries a large number of anionic sialic acid residues and sulfate groups,<sup>71,72</sup> which renders mucins highly hydrophilic and hygroscopic.<sup>73</sup> Additionally, the steric rigidity brought about by the densely attached oligosaccharide chains in combination with the repulsive electrostatic forces acting between the anionic groups is assumed to suppress polypeptide folding – and this leads to a bottle-brush-like structure of the glycoproteins.<sup>74</sup>



**Figure 2.1: Schematic of a mucin glycoprotein.** A densely glycosylated central region carrying a high number of anionic sulfate groups and sialic acid residues is framed by (partially) folded termini, which comprise hydrophobic and charged amino acids.

The glycosylated central region of the mucin molecules is framed by two (partially) folded termini, each comprising a sparsely glycosylated mixture of anionic, cationic and hydrophobic amino acids.<sup>48,75,76</sup> Those terminal regions are rich in cysteines, which imparts monomeric mucin molecules the ability to aggregate into oligomeric structures *via* intermolecular disulfide bonds.<sup>77,78</sup> Overall, the amphiphilic nature of mucin molecules not only enables their adsorption onto both, hydrophobic and hydrophilic surfaces, but also results in an excellent lubricating performance.<sup>72</sup> Furthermore, owing to the various different interaction sites provided by the complex microarchitecture of the mucin molecule, mucins constitute a unique barrier against bacteria and viruses, as they allow to trap or repel such pathogens.<sup>79,80</sup>

## Materials and methods

In this thesis, two different preparations of porcine gastric mucin MUC5AC were used: manually purified mucins (referred to as ‘MUC5AC’) as well as commercially available mucin from porcine stomach (Type II; Sigma-Aldrich, St. Louis, USA; referred to as ‘PGM’). The manual purification protocol followed to obtain MUC5AC is described in Schömig *et al.*<sup>81</sup> In brief, crude mucus was collected from pig stomachs (Schlacht- und Viehhof, München, Germany) by manually scraping the mucosal tissue surface with a lab spoon. The yield was diluted 1:5 in sodium phosphate buffer (10 mM, pH 7.0) containing 170 mM NaCl and 0.04 % (w/v) sodium azide, and it was thoroughly stirred at 4 °C overnight. To then remove cellular debris and mesoscopic impurities, two centrifugation steps (30 min at 17590 x g and 45 min at 158306 x g) were performed at 4 °C each. For the purification of mucins used for coatings on rigid, gas-permeable contact lenses (RGP lenses, see **chapter 3.3**) these two centrifugation steps were replaced by four filtration steps (mesh sizes: 1 mm, 500 µm, 200 µm, 125 µm) as described in Marczyński *et al.*<sup>82</sup> To further separate the mucins from other mucus constituents, size exclusion chromatography was conducted on an ÄKTA purifier system (GE Healthcare, Chicago, USA) equipped with an XK50/100 column packed with Sepharose 6FF resin (GE Healthcare). The obtained mucin-containing eluate was dialyzed against ultrapure water and concentrated by cross-flow filtration using an ultrafiltration hollow fiber cartridge with a molecular weight cut-off (MWCO) of 100 kDa (Xampler Ultrafiltration Cartridge; GE Healthcare). The concentrate was then lyophilized and stored at -80 °C until further usage.

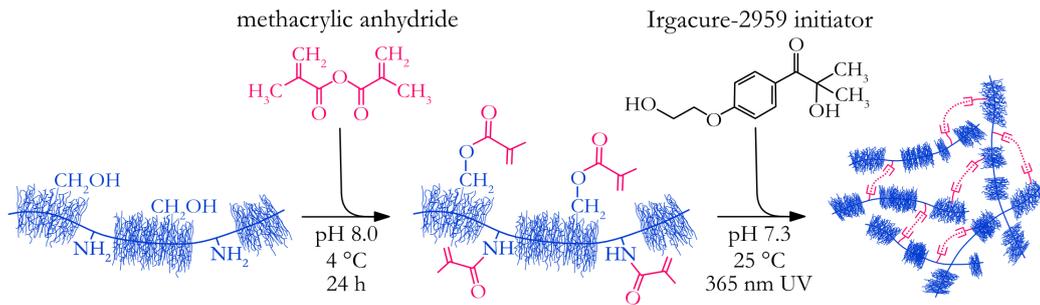
Both, MUC5AC and PGM, are usually received and stored as a lyophilized powder. To reconstitute manually purified MUC5AC, the lyophilized powder was dissolved in the desired buffer, vortexed for 5 min, and then homogenized on a shaking incubator (750 rpm) at 4 °C for 2 h. The commercially available PGM was first dissolved in 10 µL of dimethyl sulfoxide (DMSO) per 1 mg of PGM and vortexed for 5 min. Subsequently, the desired buffer was added, and the mixture was again vortexed for 5 min before it was kept on a shaking incubator (750 rpm) at 4 °C for 3 h.

## 2.2. Producing and characterizing mucin-based bio-ink samples

In this thesis, a mucin-based bio-ink was developed to be used for 3D printing, more precisely, for direct ink writing. From a manufacturing point of view, such bio-inks have to exhibit several characteristics, most importantly the ability to be extruded, to hold the shape during printing, and to develop inherent stability after being printed. To achieve all this, commercially available PGM was functionalized to enable UV-induced curing, and a stabilizing agent was added to tailor the rheological properties of the ink.

### 2.2.1. Functionalization of mucins with methacryloyl groups

Porcine gastric mucins inherently possess the ability to undergo sol-gel transformations, which can, for instance, be triggered by an alteration of the pH or the addition of certain ions. However, to achieve a controllable gel formation suitable for curing mucin-based materials in a 3D printing process, the PGM was functionalized by attaching methacryloyl groups (**Figure 2.2**). When supplemented with a photo-crosslinker, those methacryloyl groups then allow for photo-initiated covalent crosslinking of the mucin glycoproteins *via* free-radical polymerization. Even though lab-purified mucin may offer higher purity, here, commercial PGM was favored due to its easier availability and cost-effectiveness. Since the structural losses previously reported for commercial PGM molecules mainly affect the ability of mucins to passively adhere to surfaces and to undergo gelation at acidic pH,<sup>83</sup> it is expected that those mucins are sufficiently intact to be used as components of a bio-ink.



**Figure 2.2: Schematic illustration of the methacrylation procedure of PGM and the UV-driven curing of PGM-MA.** PGM macromolecules were methacrylated *via* the hydroxymethyl groups of their glycans and the amine groups of the protein backbone. Crosslinking by free-radical polymerization can then be achieved by addition of the photo-initiator Irgacure-2959 and subsequent UV irradiation.

For this purpose, 1.0 % (w/v) PGM was dissolved in ddH<sub>2</sub>O, cooled on ice and the pH was adjusted to ~8.0. Once cooled down, methacrylic anhydride (MA; Sigma-Aldrich) was added such that a ratio of 1000 MA molecules per PGM macromolecule was obtained. Since the initiated reaction induces a continuous decrease of the pH value, the pH of the solution was repeatedly readjusted to 8.0 by adding 5 M NaOH while gently stirring the solution. This pH control was conducted during the initial 6 h, after which the solution was stored at 4 °C overnight to allow for further reaction. To remove non-reacted MA and mucin precipitates, a centrifugation step (10 min, 5000 rpm; Centrifuge 5430; Eppendorf SE, Hamburg, Germany) was conducted followed by size exclusion chromatography of the supernatant using an ÄKTA purifier system and an XK50/100 column packed with Sepharose 6FF (both GE Healthcare). The collected fractions containing the methacrylated porcine gastric mucin (PGM-MA) were lyophilized and stored at -80 °C until further use.

### 2.2.2. Preparation and curing of the bio-ink

To prepare mucin-based bio-inks, 8.0 % (w/v) of lyophilized PGM-MA was dissolved in ddH<sub>2</sub>O, vortexed for 1 min, and thoroughly mixed with a magnetic stirrer (300 rpm) at 4 °C for 12 h. At the same time, 8.0 % (w/v) Laponite® XLG (BYK-Chemie GmbH, Wesel, Germany) was dissolved: Therefore, ddH<sub>2</sub>O was cooled on ice for 5 min and then agitated on a magnetic stirrer at 1500 rpm until a clear swirl was created. The Laponite powder was then slowly added, and the clay particles were allowed to swell for 16 h. Then, the PGM-MA solution and the Laponite solution were mixed in a 1:1 ratio and homogenized at 1200 rpm (high speed homogenizer FSH-2A; Vevor Corporation GmbH, Köln, Germany) for 2 min. For control solutions, that contained only one of these two components, the exact same procedure was conducted, but the respective stock solution was mixed in a 1:1 ratio with ddH<sub>2</sub>O.

For curing the bio-inks *via* free radical polymerization, the photo-initiator 2-hydroxy-4'-(2-hydroxyethoxy)-2-methylpropiophenone (Irgacure-2959; Sigma-Aldrich) was used to enable UV-induced crosslinking of the PGM-MA; this chemical was already shown to be a suitable agent that combines low cytotoxicity with efficient crosslinking.<sup>84</sup> Here, an aqueous solution of 80 % (v/v) ethanol containing 100 mg mL<sup>-1</sup> Irgacure-2959 was added to the prepared bio-ink such that a final Irgacure concentration of 1 mg mL<sup>-1</sup> was obtained. Immediately afterwards, the mixture was again homogenized at 1200 rpm for 2 min (to obtain a homogeneous dispersion of the photo-initiator) while avoiding any light exposure. Photo-crosslinking of the PGM-MA molecules (which is responsible for gelation of the PGM-MA solutions) was conducted by exposing the solutions to UV light (365 nm wavelength, ~10 mW cm<sup>-2</sup>; M365L2; Thorlabs GmbH, Lübeck, Germany) as described in the individual experiments.

## Materials and methods

### 2.2.3. 3D printing process

For 3D printing of the PGM-MA-based bio-inks, a direct ink writing bioprinter was used (nominal resolution along all axes: 1  $\mu\text{m}$ ; BioX<sup>TM</sup>; Cellink, Göteborg, Sweden). After filling the PGM-MA-based bio-ink into a UV-protected cartridge (to avoid premature photopolymerization) and attaching the needle (gauge: 25G), the cartridge was mounted into the printer and a manual calibration was conducted. All printing procedures were performed at room temperature (RT), and if not stated otherwise, an extrusion pressure of 80 kPa was applied and a moving speed of 20  $\text{mm s}^{-1}$  was selected. During the printing process, the printing area was permanently irradiated with UV light (*via* the integrated 365 nm UV module of the BioX<sup>TM</sup> printer) and a 365 nm UV curing head (BIO X<sup>TM</sup> Photocuring Toolhead 365 nm; Cellink) was used to cure each layer directly after extrusion for 30 s at a distance of 2 cm. Additionally, after printing of a complete object, the sample was irradiated for 5 more minutes using the curing head to ensure sufficient curing of all printed layers. Individual tests conducted to identify suitable printing parameters and to evaluate the printing performance can be found in **Appendix A1.1**.

### 2.2.4. Oscillatory shear rheometry

When undergoing deformation, most materials not purely behave like a solid or a fluid – instead, they unite both, viscous and elastic characteristics. Such viscoelastic materials can be characterized *via* rheological measurements by recording the material response  $\sigma^*$  to an applied oscillatory load  $\gamma^*$ . For such deformations, that oscillate with the angular frequency  $\omega$ , the corresponding shear modulus  $G^*(\omega)$  is a complex descriptor comprising a real and an imaginary part: The real part  $G'(\omega)$  – referred to as the storage modulus – represents the elastic material response whereas the imaginary part  $G''(\omega)$  – the loss modulus – describes the viscous material response:

$$G'(\omega) = \frac{\sigma_0}{\gamma_0} \cos(\delta) \qquad G''(\omega) = \frac{\sigma_0}{\gamma_0} \sin(\delta) \qquad (1, 2)$$

Here,  $\sigma_0$  and  $\gamma_0$  refer to the amplitude of the shear stress and the shear strain, respectively. Furthermore,  $\delta$  denotes the frequency-dependent phase shift between the oscillatory deformation of the material and its stress response.

In this thesis, the viscoelastic properties of mucin-based bio-ink formulations were determined using a shear rheometer (MCR102; Anton Paar, Graz, Austria) equipped with a plate-plate geometry. To allow for illumination of the samples (when desired), a transparent bottom plate (P-PTD200/GL; Anton Paar), that is permeable for the transmission of UV light, was chosen; furthermore, a UV lamp (M365L2; Thorlabs GmbH) was installed below this transparent bottom plate. With this experimental setup, the intensity of UV light in the sample plane was approximately 7.3  $\text{mW cm}^{-2}$ . All measurements were performed at a temperature of 25  $^\circ\text{C}$  with a light-proof liquid trap placed around the measuring chamber; this liquid trap not only protected the sample from drying but also prevented uncontrolled curing induced by environmental UV light. The gap size between the bottom plate and the measuring head (PP25; Anton Paar) was set to  $d = 300 \mu\text{m}$  for all experiments, and a sample volume of 250  $\mu\text{L}$  was used.

To determine the viscosity of the different material formulations before curing, logarithmic shear-rate ramps ranging from  $\dot{\gamma} = 10^3 \text{ s}^{-1}$  to  $\dot{\gamma} = 10^0 \text{ s}^{-1}$  were measured. To follow the time-dependent curing process of the materials,  $G'$  and  $G''$  were determined at a constant frequency of  $f = 1 \text{ Hz}$  while applying a constant torque of  $M = 5 \mu\text{Nm}$ . This torque-controlled procedure was chosen to avoid exceeding the linear response regime despite the major change in material properties during curing. After measuring a baseline for 5 min without UV exposure, the viscoelastic moduli were recorded for 15 min of UV

illumination. Next, to assess the viscoelastic behavior of the samples before and after the curing process in more detail,  $G'$  and  $G''$  were recorded in response to a constant shear strain that was applied with oscillation frequencies ranging from  $f = 10^{-1}$  Hz to  $f = 10^1$  Hz. To again ensure not to exceed the linear response regime, the shear strain applied for samples that were not cured yet was set to the 1.5-fold value of the averaged shear strain obtained when applying a minimal oscillatory torque of  $M = 0.5 \mu\text{Nm}$  (those values were obtained for every sample type in pre-tests). For cured samples, the constant strain was set to the strain values observed during the five last data points of the curing measurements (as described above).

### 2.2.5. Unconfined compression tests

Unconfined compression tests of cured bio-ink samples were performed on a commercial rheometer (MCR102; Anton Paar) equipped with a standard bottom plate (P-PTD200/AIR; Anton Paar) and a planar PP08 measuring head (PP08-5681; Anton Paar). The samples were generated in the wells of a 48-well microtiter plate to perfectly match the dimension of the measuring head. Therefore, 200  $\mu\text{L}$  of a bio-ink solution were filled into the wells and cured with UV light for 15 min. The cured samples were gently removed from the wells and placed onto the bottom plate of the rheometer. After resetting the normal force before each measurement, the measuring head was lowered until a gap height of  $d_{\text{initial}} = 4 \text{ mm}$  was reached (measured between the measuring head and the bottom plate). Then, the measuring head was further lowered at a constant speed of  $v = 10 \mu\text{m s}^{-1}$  and the resulting normal force  $F_N$  was recorded over time (one data point per second). For each sample, the gap size at which the first full contact between the measuring head and the sample occurred was noted – and this distance represents the individual sample height that is later on needed to calculate the axial strain.

Based on the obtained curves, the compression modulus was calculated as follows: First, the obtained normal force was converted into a normal pressure  $p$  according to  $p = F_N * (\pi * r^2)^{-1}$ , with the radius of the measuring head  $r = 7.965 \text{ mm}$ . Then, the continuously shrinking gap size  $\Delta d_g$  was converted into axial strain values according to  $\gamma = \Delta d_g * h_s^{-1}$ ; here,  $h_s$  denotes the individual sample height that was determined as described above. Afterwards, to only consider the linear regime of the obtained curve, the 20 points measured directly before sample failure were discarded, and a linear, least squares regression curve was fitted to the remaining data using the software GraphPad Prism (Prism 9 for macOS; Version 9.3.1 (350), December 7, 2021; GraphPad Software LLC; San Diego, USA). Finally, the slopes of those regression lines represented the compression moduli of the respective samples.

### 2.2.6. Swelling experiments

To observe the swelling behavior of the cured bio-ink formulations, the ink solutions were prepared as described before (including the addition of the photo-initiator), filled into the wells of a 96-well plate (100  $\mu\text{L}$  per well), and cured with UV light for 15 min (365 nm wavelength,  $\sim 10 \text{ mW cm}^{-2}$ ; M365L2, Thorlabs GmbH). This time span was found (based on macrorheological measurements, see **chapter 3.1**) to be more than sufficient to fully cure the hydrogel. Directly after curing, the specimens were weighed using a micro-balance (XSE205 Dual Range, Mettler Toledo, Columbus, USA). Then, they were stored in 5 mL of ddH<sub>2</sub>O supplemented with different concentrations of NaCl such that the samples were fully surrounded with liquid at all times. After an incubation time of 5 days (a time span that should be sufficient for the samples to reach an equilibrium state),<sup>84</sup> the hydrogel samples were gently blotted from each side to remove excess water, air-dried for 5 min, and again weighed.

## Materials and methods

### 2.3. Medical devices

Mucin coatings were generated and evaluated on several different medical devices. First, two different types of contact lenses were used: flexible silicone lenses and rigid, gas-permeable lenses. Additionally, mucin coatings were established and tested on urinary catheters and endotracheal tubes.

The silicone contact lenses used in this thesis were made from a highly biocompatible, aliphatic Pt-catalyzed liquid silicone based on Si-H- and Si-vinyl-polydimethylsiloxane (PDMS) without any further additives (Polymer Systems Technology, High Wycombe, UK). Additionally, samples of the same material but in the shape of a flat disk were used for contact angle measurements to facilitate measuring the water-enclosed angle between the droplet and the surface. For simplicity, those samples will be referred to as 'PDMS lenses' and 'PDMS disks', respectively. Those PDMS samples were produced by two-component vulcanization in a custom-made cast-molding procedure at a process temperature of 110 °C (conducted by Wöhlk Contactlinsen GmbH, Schönkirchen, Germany). If not stated otherwise, the PDMS lenses/disks were stored in phosphate-buffered saline (PBS) supplemented with 0.0005 % (v/v) polyhexamethylene biguanide (PHMB, pH 7.2 – 7.8; hereafter referred to as 'storage solution').

The RGP (rigid, gas-permeable) contact lenses used here were made from high-molecular, cross-linked, amorphous alkyl-siloxanyl-fluoromethacrylates with the material designation Fasifocon (available from the contact lens manufacturer Wöhlk Contactlinsen GmbH in four formulation variants). This methacrylate copolymer is essentially based on three types of monomers: the gas-permeable aliphatic siloxane-methacrylates of the type "TRIS" (CAS 17096-07-0), the fluoro-methacrylates of the type "TFPMA" (CAS 45102-52-1), and the stabilizing gas-impermeable alkyl-methacrylates of the type "MMA" (CAS 80-62-6). According to contact lens standard EN ISO 18369-1:2018 (chapter 4.6), these rather hydrophobic polymers belong to the material group 3. The two-step production process of the RGP lenses was performed by Wöhlk Contactlinsen GmbH: first, rod-shaped polymer blanks (containing a dye as a handling aid and a polymerized UV absorber methacrylate) were fabricated by radical-thermal copolymerization. From the prepared blanks, customized contact lenses were formed *via* CNC lathing (computerized numerical control lathing). With this production approach, very complex surface geometries can be realized to obtain both, high optical quality and durable, individual contact lenses. Target values for different physical and geometrical properties as presented in **chapter 3.3** were defined by the lens manufacturer; acceptable tolerance ranges for those characteristics were specified based on EN ISO 18369-2. If not stated otherwise, the contact lenses were stored in the storage solution provided by Wöhlk Contactlinsen (as for PDMS lenses).

Endotracheal tubes (Super Safetyclear, 10.0 mm, CH40) that were fabricated from latex-free polyvinyl chloride (PVC) were obtained from Radecker Notfallmedizin (Ammerbuch/Entringen, Germany). Urinary catheters (SpeediCath Standard, CH/FR 18/6.0 mm) made of polyurethane (PU) were purchased from Coloplast (Hamburg, Germany). Both devices are commercially available medical products with approval for clinical use on the European market (Conformité Européenne, CE labelled). To yield uniform samples of these two medical devices, a grommet press (IstaBreeze Germany GmbH, Bad Rappenburg, Germany) was used to cut circular specimens with a diameter of  $d = 7$  mm from the endotracheal tubes (curvature of the samples:  $0.1 \text{ mm}^{-1}$ ) and the catheters (curvature of the samples:  $0.17 \text{ mm}^{-1}$ ). Prior to any further use, those specimens were subjected to two washing steps (first with 70 % [v/v] ethanol and then with ddH<sub>2</sub>O) on a rolling shaker (rpm: 75; RS-TR 05; Phoenix Instrument GmbH, Garbsen, Germany) for 30 min each.

## 2.4. Generating and evaluating mucin coatings on polymeric surfaces

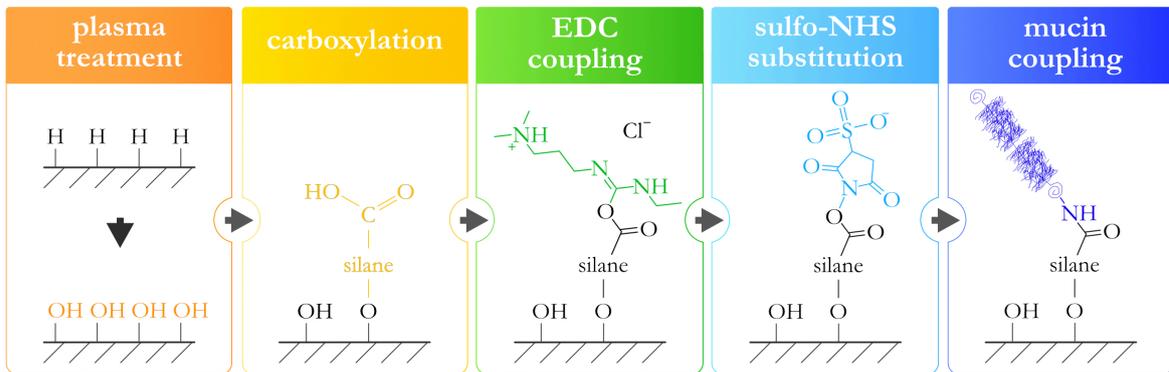
Two different coatings strategies, namely passive adsorption and covalent coupling, were employed to generate mucin coatings on the previously described medical devices. The impact those coatings have on several surface characteristics of the polymeric materials, such as wettability and tribological performance, was then evaluated.

### 2.4.1. Passively adsorbed mucin coatings

To generate passive mucin coatings on polymeric samples, the respective surfaces were gently cleaned with 80 % (v/v) ethanol and ddH<sub>2</sub>O, respectively. Afterwards, the surfaces to be coated were incubated with Dulbecco's phosphate-buffered saline (DPBS) containing 0.1 % (w/v) mucin macromolecules. Incubation was conducted at RT for 2 h while gently shaking the specimens. Finally, the surfaces were gently washed with 80 % (v/v) ethanol and stored in DPBS (if not stated otherwise).

### 2.4.2. Covalent mucin coatings

To covalently attach mucin macromolecules onto polymeric surfaces, carbodiimide coupling was employed (**Figure 2.3**). Importantly, the process parameters presented in the following section correspond to the coating strategy applied for samples made from PDMS. However, to account for the limited stability of Fasifocon (RGP lenses) towards heat or solvents, several steps of this procedure had to be modified (**Appendix A2**). Some of those modifications were also applied for PU and PVC samples. A full overview of the process parameters used for each material is listed in **Appendix A1.2**. For all materials, the samples were washed with 80 % (v/v) ethanol and ddH<sub>2</sub>O, respectively, before further treatment.



**Figure 2.3: Schematic illustration of the covalent mucin coupling process.** The polymeric surfaces were activated using plasma generated with ambient air before they were carboxylated with a silane coupling agent. Subsequently, carbodiimide coupling was employed to covalently attach the mucin glycoproteins onto the prepared surface. *This figure is adapted from Rickert et al. (2020)<sup>85</sup>.*

After gently blotting the sample surfaces and air drying them for several minutes, surface activation was conducted to replace surface-bound methyl groups with hydroxyl groups. This was achieved by exposing the specimens to plasma generated with ambient air for 90 s (pressure: 0.4 mbar, intensity: 30 W; plasma oven “SmartPlasma 2”; plasma technology GmbH, Herrenberg, Germany). Afterwards, to obtain a carboxylated surface, which is required for the final coupling step, the coupling agent

## Materials and methods

N-[(3-trimethoxysilyl)propyl]ethylenediamine triacetic acid trisodium salt (TMS-EDTA; abcr GmbH, Karlsruhe, Germany) was employed: 1.0 % TMS-EDTA was added to 10 mM acetate buffer (pH 4.5), and the samples were incubated in this solution at 60 °C for 5 h. To remove loosely bound silanes before performing the final coupling step, the samples were dipped into iso-propanol (>99.5 %) and washed in 96 % (v/v) ethanol on a rolling shaker (70 rpm; RS-TR 05; Phoenix Instrument GmbH) for 1 h. Afterwards, the previously generated siloxane bonds were stabilized by curing the specimens in a dry state at 80 °C for 2 h. To finally couple mucin macromolecules to the carboxylated surface, an EDC-NHS coupling reaction was employed. This was initiated by incubating the samples in 100 mM 2-(*N*-morpholino)ethanesulfonic acid buffer (MES buffer, pH 5.0; AppliChem GmbH, Darmstadt, Germany) supplemented with 5 mM 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride (EDC) and 5 mM sulfo-*N*-hydroxysuccinimide (sulfo-NHS, abcr GmbH) for 30 min while moderately shaking (35 rpm). Afterwards, the samples were immersed in DPBS containing 0.1 % (w/v) mucin macromolecules at 4 °C for 12 h. This coupling strategy targets amino groups of the mucin macromolecule,<sup>86</sup> that are mainly accessible in the non-glycosylated terminal regions of the glycoprotein.<sup>48</sup> The readily coated samples were washed in 80 % (v/v) ethanol and stored in DPBS (if not stated otherwise). Samples designated for sterilization experiments were washed in 80 % (v/v) ethanol, air-dried, and stored in sterilization bags (Medi Pack GmbH, Mönchengladbach, Germany).

### 2.4.3. Methods to detect mucin glycoproteins

To verify both, the presence of manually purified MUC5AC bound to a surface, and the structural integrity of the targeted motifs (antigens), two indirect, analytical methods were applied: an indirect enzyme-linked immunosorbent assay (ELISA) and a lectin depletion test. To visualize commercial PGM, the PGM molecules were fluorescently labelled and analyzed using a fluorescence microscope.

#### Indirect enzyme-linked immune-sorbent assay (Indirect ELISA)

An indirect ELISA performed with a MUC5AC-specific antibody constitutes a reliable method for the detection of the non-glycosylated, partially folded termini of the mucin macromolecules. The implementation of an indirect ELISA essentially involves three steps: First, the surface-bound antigen of interest is targeted with a primary antibody. Then, this first antibody is addressed by an enzyme-labelled secondary antibody. When finally adding a colorless substrate, that is able to react with the enzyme conjugate, a byproduct is formed that can be measured with colorimetric methods.

The protocol described in the following section was conducted with various different types of samples; for the exact volumes used for each sample, please refer to **Appendix A1.3**. In short, specimens carrying either passive or covalent mucin coatings were gently washed with DPBS-Tween (DPBS with 1 mg mL<sup>-1</sup> Tween 20, pH = 7.3) to obtain a clean, dust-free surface. To afterwards block all binding sites that would otherwise allow for unspecific attachment of the primary antibody, the samples were immersed in blocking buffer (DPBS-Tween supplemented with 5 % [w/v] milk powder) at 4 °C overnight. Likewise, this blocking step was conducted for the empty wells of a well plate (for details, see **Appendix A1.3**) using the same blocking buffer. After incubation, the coated specimens were relocated into the blocked wells, and everything was washed with DPBS-Tween for three times. Then, blocking buffer supplemented with a specific antibody for MUC5AC detection (ABIN966608 diluted 1:400; antibodies-online GmbH, Aachen, Germany) was added to each sample and it was allowed to incubate on a shaker (35 rpm) at RT for 1 h. Subsequently, after again washing the wells thrice with DPBS-Tween, blocking buffer comprising a horseradish peroxidase-conjugated goat anti-mouse (murine) IgG antibody

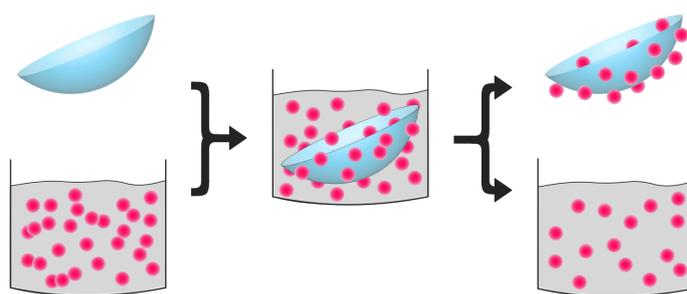
(ABIN237501 diluted 1:5000; antibodies-online GmbH) was added to each well, and incubation was conducted while gently shaking the well plate at 35 rpm and RT for 2 h. Afterwards, the wells were thoroughly washed with pure DPBS (Tween was omitted to avoid cross-reactions with the following Working Solution).

To finally initiate the colorimetric change, QuantaRed Working Solution (50 parts QuantaRed Enhancer Solution, 50 parts QuantaRed Stable Peroxide, and one part of QuantaRed ADHP Concentrate; QuantaRed Enhanced Chemifluorescent HRP Substrate Kit 15159; Thermo Fisher Scientific, Waltham, USA) was added to each well. After an incubation time of 30 min at RT, the specimens were removed from the wells, and the fluorescence signal created by the converted substrate was measured with a plate reader (ex.: 540 nm, em.: 590 nm; Fluoroskan Ascent; Thermo Fisher Scientific).

A slightly modified version of this protocol was also used to evaluate the integrity of mucins after sterilization. Therefore, mucin molecules were directly attached to the bottom of the wells of a 96-well microtiter plate by means of passive adsorption, and all steps were conducted directly in those wells.

### Lectin depletion assay

In addition to targeting the hydrophobic termini of the mucin molecules, the presence and integrity of surface-bound MUC5AC can be probed by targeting the glycosylated, central region of the glycoprotein. This was achieved by conducting a depletion assay (**Figure 2.4**) using fluorescently labelled lectins that specifically bind to sialic acids (more specifically, *N*-acetylneuraminic acid and *N*-glycolylneuraminic acid) and *N*-acetylglucosamine.<sup>87</sup> For this purpose, 300  $\mu\text{L}$  of DPBS supplemented with 12.5  $\mu\text{g mL}^{-1}$  fluorescently labelled lectins (fluorescein isothiocyanate [FITC]-conjugated lectin from *triticum vulgare*, wheat; Sigma-Aldrich) were injected into the wells of a 48-well microtiter plate. Then, the specimens of interest were washed thrice with DPBS and each placed into an individual well (containing the lectin solution). After an incubation time of 12 h under light exclusion during which the well plate was gently moved on an orbital shaker (35 rpm), 200  $\mu\text{L}$  of the liquid were removed from each well. The fluorescence intensity generated from those partially depleted lectin solutions was then quantified using a plate reader (ex.: 485 nm, em.: 538 nm; Fluoroskan Ascent; Thermo Fisher Scientific).



**Figure 2.4: Schematic illustration of a depletion assay.** A solution containing fluorescently labelled molecules is filled into the wells of a microtiter plate. Then, the specimen of interest is immersed in this solution and incubated for a defined duration. After removing the sample from the solution, the partially depleted solution is analyzed photometrically.

### Coating detection with labelled PGM

To allow for visualizing mucin coatings generated from commercial PGM, a fluorescent label was attached to the PGM molecules *via* carbodiimide coupling. Therefore, a stock solution comprising 10  $\text{mg mL}^{-1}$  carboxy-modified ATTO488 dye (ATTO-TEC GmbH, Siegen, Germany) in  $\text{ddH}_2\text{O}$  was diluted with 10 mM MES buffer (pH 5.0; AppliChem GmbH) such that a dye concentration of

## Materials and methods

$c_{\text{ATTO}} = 1 \text{ mg mL}^{-1}$  was reached. Then, 5 mM EDC and 5 mM sulfo-NHS (acbr GmbH) were added, and the mixture was incubated light excluded at RT for 3 h. This prolonged incubation time was chosen to fully hydrolyze excess EDC before adding the mucin molecules (to avoid cross-linking of the mucins). In parallel, 40 mg of PGM were dissolved in 19 mL of 10 mM PBS (pH 7.0). Then, 1 mL of the dye solution was added to the mucin solution, the mixture was thoroughly shaken and allowed to react at RT for 3 h. To afterwards remove unbound dye molecules, the solution was dialyzed against ddH<sub>2</sub>O (MWCO = 300 kDa). The labelled mucins were then lyophilized and stored at  $-80^\circ\text{C}$  until further use.

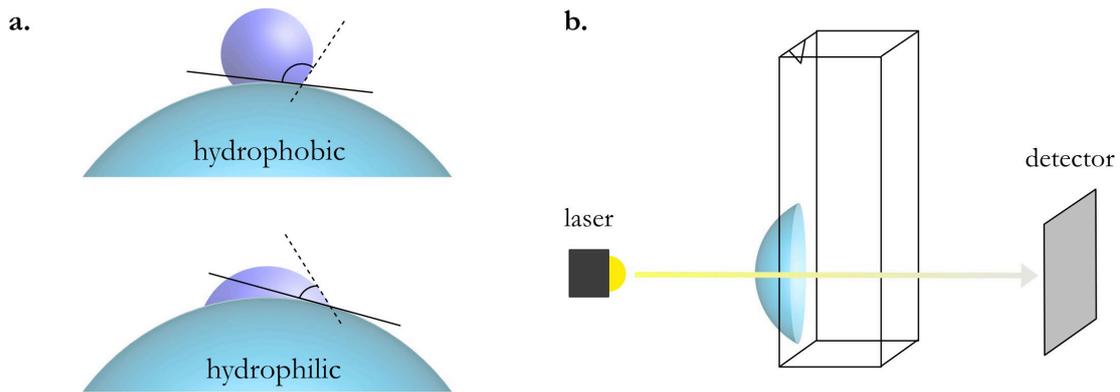
To finally visualize coatings generated with such labelled PGM, the coating procedures (passive or covalent) were conducted as described in **chapters 2.4.1** and **2.4.2**, respectively. However, to avoid bleaching of the fluorescent dye, special care was taken to avoid light exposure during all steps. Freshly coated samples were washed with 70 % (v/v) ethanol, gently blotted with a wipe, and placed onto a microscopy glass slide. Then, fluorescence images were obtained on a Leica DMI8 microscope (Leica, Wetzlar, Germany) using a digital camera (exposure time: 1 s, Orca Flash 4.0 C11440-22C, Hamamatsu Photonics, Hamamatsu, Japan). Five images of each sample were recorded using a 20 $\times$  objective (NA: 0.46; Leica) and a FITC green light filter (ex.: 460–500 nm, em.: 512–542 nm; Leica). For analysis, the light intensities of the images were determined and averaged using Matlab (version R2019a; MathWorks, Natick, USA).

### 2.4.4. Wettability evaluation

The surface wettability specifies the degree of contact between a solid surface and a liquid environment. It can depend on several factors, such as the surface topography, its chemical composition, and the different intermolecular forces acting between the liquid and the surface (such as van der Waals' forces or hydrogen bonding). A commonly used, quantitative measure of the surface wettability is given by the water contact angle. In addition, for contact lenses, a more application-oriented evaluation of the wettability is provided by the liquid break-up time.

#### Contact angle measurements

Contact angle measurements as described in the following were conducted either manually or by using a drop shape analyzing device equipped an automated water dosing unit (DSA25S; Krüss GmbH, Hamburg, Germany). For detailed information on the exact volumes and devices used for both techniques, please refer to **Appendix A1.4**. For both methods, the samples were first washed with 70 % (v/v) ethanol and ddH<sub>2</sub>O to remove dust or remnants from prior treatments. Afterwards, to obtain a dry surface that can be accurately analyzed, the samples were gently blotted and then blow-dried with oil-free pressurized air (Druckluft 67; Kontaktchemie, CRC Industries Europe BVBA, Zele, Belgium) for  $\sim 10$  s. Next, a small droplet of ddH<sub>2</sub>O was placed onto the surface and transversal images were acquired with a high-resolution camera. These images were then analyzed to identify both, the surface baseline and the contour of the droplet. Lastly, the static contact angles were determined as the water-enclosed angle between the tangent to the sample surface and the tangent to the edge of the droplet (**Figure 2.5a**). Surfaces showing contact angles below  $90^\circ$  are considered hydrophilic, whereas contact angles above  $90^\circ$  denote hydrophobic behavior. Moreover, contact angles that are higher than  $120^\circ$  represent superhydrophobic surface characteristics.



**Figure 2.5: Schematic illustration of contact angle and transparency measurements.** To measure the surface wettability, a droplet of ddH<sub>2</sub>O is placed on the surface, and the water-enclosed angle between the tangent to the sample surface and the tangent to the edge of the droplet is determined (a). Contact angles above 90° denote hydrophobic surface behavior, whereas contact angles below 90° denote hydrophilic surface behavior. To measure the transparency of a contact lens, the lens was attached or inserted into a cuvette such that it was located in the center of a spectrophotometric beam path (b).

### Assessment of break-up times

In analogy to the tear film break-up time of human eyes, which is an important measure for the diagnosis of dry eye symptomatology,<sup>88</sup> the liquid break-up time (BUT) of contact lenses can provide an indication for the application success of the lens. The BUT is defined as the time span during which a liquid film on a surface exposed to air remains intact without rupture, *i.e.*, the duration until first dry spots appear on the surface. To assess this parameter, contact lens samples were hydrated in DPBS for at least 24 h. For each individual measurement, a lens was taken out of the buffer, gently blotted with a laboratory paper towel to remove surplus water, and placed onto a microscopy glass slide in such a way that it protruded partially beyond its edge. This ensured that no reservoir of liquid was formed between the lens and the glass, which would otherwise falsify the results. Immediately after placing the lens, a stop watch was started, and it was stopped as soon as first ruptures of the liquid film appeared (typically at the edge of the contact lens).

### 2.4.5. Transparency measurements

To determine the transparency of the mucin coatings generated on contact lenses, UV/Vis-spectrophotometry was performed. Before conducting the trials, the contact lenses were immersed in storage solution for at least 24 h. After gently drying the lenses with oil-free compressed air (Druckluft 67, Kontaktchemie), they were attached to (for PDMS lenses) or inserted into (for RGP lenses) an empty cuvette in such a way that the lens was located directly in the beam path between the laser source and the detector of the spectrophotometer (Specord 210; Analytik Jena AG, Jena, Germany). Here, the convex side of the lens was always facing the light source (Figure 2.5b). The intensity of transmitted light was then measured for wavelengths ranging from 380 nm to 900 nm (step size: 1 nm; speed: 10 nm s<sup>-1</sup>). To compensate for the imperfect transparency of the respective lens material, the obtained results were normalized: spectra measured for coated lenses were pointwise divided by the mean values obtained for uncoated lenses of the same lens material. This adjustment ensured that the reported spectra accurately represented the transparency of the applied coatings.

## Materials and methods

### 2.4.6. Lipid deposition tests

Lipids are one of the major components of the human tear film. However, on contact lenses, unfavorable depositions of lipids can severely impair the lens transparency and even lead to considerably lower comfort. To evaluate the ability of mucin coatings to prevent lipid deposition on a lens surface, different methods were applied: The first approach involves performing the previously described transparency measurements before and after 10 days of storing the lenses in buffer that is enriched with 1,2-dioleoyl-sn-glycero-3-phosphocholine (DOPC; Avanti Polar Lipids, Alabaster, USA). Therefore, the DOPC was dissolved in chloroform ( $\geq 99\%$ ) and a small amount of this lipid-containing solution was deposited into a small glass vial. By allowing the chloroform to evaporate overnight, a thin lipid film was generated on the bottom of the vial. To finally prepare an aqueous solution, the DOPC film was dissolved in DPBS so that a final lipid concentration of  $50\ \mu\text{M}$  was obtained. This concentration approximates the abundance of choline-containing lipids in the human lacrimal fluid.<sup>89</sup> The transmission spectra acquired after incubation of the lenses in the DOPC-rich liquid were normalized to the transmission values measured for the exact same lens before incubation. Importantly, before both spectrophotometric measurements, the lenses were gently washed with ddH<sub>2</sub>O to remove dust or loosely attached lipids that could otherwise strongly distort the transmission values.

The second method used to assess lipid deposition on contact lenses follows the principle of a depletion assay. For this purpose, the samples were gently washed in  $70\%$  (v/v) ethanol and ddH<sub>2</sub>O, respectively, and then placed into the wells of a 48-well microtiter plate. Then,  $300\ \mu\text{L}$  of DPBS containing  $25\ \mu\text{M}$  of 1,2-dioleoyl-sn-glycero-3-phosphoethanolamine carrying a fluorescent label (DOPE-Atto590, ATTO-TEC GmbH) were added to each sample. Those DOPE-Atto590 solutions were prepared in the same way as the DOPC solutions described above. After gently shaking the well plate at  $25\ \text{rpm}$  and  $4\ ^\circ\text{C}$  for 4 h,  $200\ \mu\text{L}$  of the lipid solution were taken from each well and the fluorescence signals were quantified (ex.:  $584\ \text{nm}$ , em.:  $620\ \text{nm}$ ; Fluorskan, Thermo Fisher Scientific). Then, the obtained values were normalized to the mean value obtained for a lipid solution incubated in an empty well.

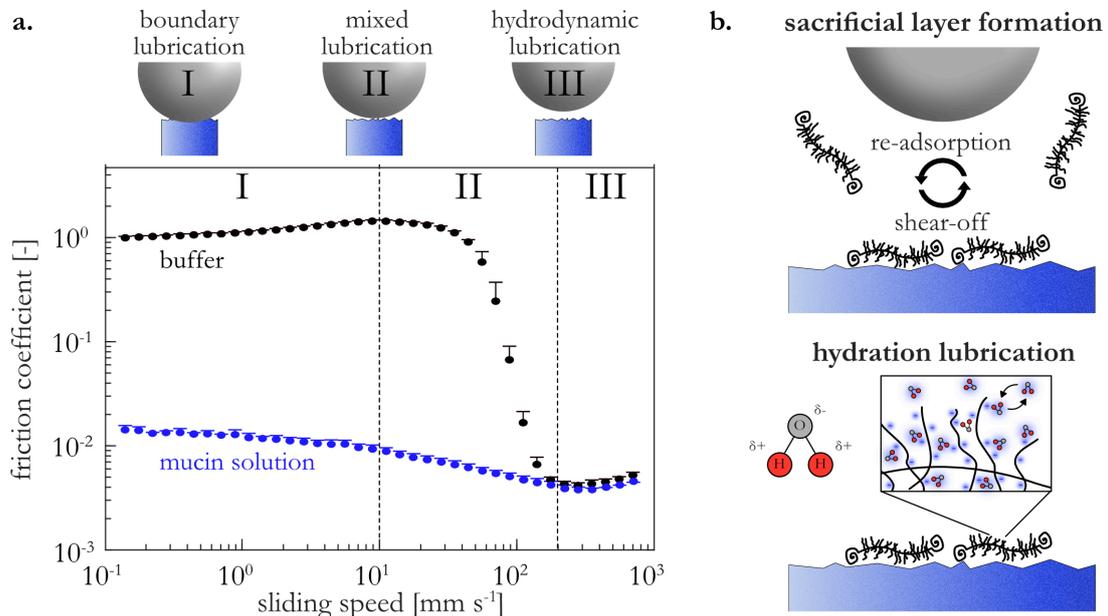
Lastly, a third method to assess lipid deposition was performed together with the lens manufacturer Wöhlk Contactlinsen GmbH. Therefore,  $0.05\%$  (w/v) cholesteryl palmitate (Sigma-Aldrich) and  $2.45\%$  (w/v) tetrahydrofurane were dissolved in cooled ddH<sub>2</sub>O ( $6\text{--}10\ ^\circ\text{C}$ ). For each measurement, an individual lens was first washed in ddH<sub>2</sub>O for 20 s. Afterwards, the lens was positioned in a beaker containing  $50\ \text{mL}$  of stirred lipid solution such that the convex side of the lens was facing the flow. After being kept in the flow for 30 s, the lens was washed twice in  $50\ \text{mL}$  ddH<sub>2</sub>O for 10 s and pictures of the treated lenses were recorded under a slit lamp. The images received from Wöhlk Contactlinsen GmbH were then analyzed with the software ImageJ2 (public domain, version 2.3.0/1.53q, 2021). For this analysis, the images were converted into 32-bit grayscale, and the analysis area was defined by manually segmenting the contour of the lens. Subsequently, a threshold filter (high-pass, threshold value: 100) was applied to identify areas covered with lipids (*i.e.*, white areas of the otherwise transparent lens), and the dimension of the such obtained area was compared to the complete lens surface. This setup challenges the lenses with higher lipid concentrations than usually present in a physiological environment to better demonstrate the lipid-repellency effects of the different surfaces.

## 2.5. Tribology

To investigate the lubricity of mucin solutions and mucin coatings and to probe the capability of such coatings to prevent wear formation on underlying tissue, tribological experiments were conducted. Therefore, a rotational load was applied, and the friction coefficient was obtained for varying sliding velocities  $v$  as follows:

$$\mu(v) = \frac{F_R(v)}{F_N} \quad (3)$$

Here,  $F_N$  represents the normal force applied to the probed surface, and  $F_R$  denotes the friction force as a function of  $v$ . Generally said, the friction force  $F_R$  describes the force opposing the relative motion of solid surfaces against each other. In the case of lubricated friction, *i.e.*, when a fluid separates the moving surfaces, the dependence of  $F_R$  on the frictional velocity  $v$  can be divided into three basic contact scenarios: boundary lubrication, mixed lubrication, and hydrodynamic lubrication (**Figure 2.6a**).



**Figure 2.6: Macromolecular lubrication mechanisms.** The classical Stribeck’s theory distinguishes between three different friction scenarios: hydrodynamic lubrication, mixed lubrication, and boundary lubrication (a). The extraordinary lubrication of mucin glycoproteins is mainly achieved by two mechanisms: first, the formation of a sacrificial layer that allows for a continuous shear-off and re-adsorption of the mucin molecules, and second, the ability of the mucins to supply a water film separating the opposing surfaces (*i.e.*, hydration lubrication; b). *Subfigure b is adapted from Marczynski (2022)<sup>90</sup>.*

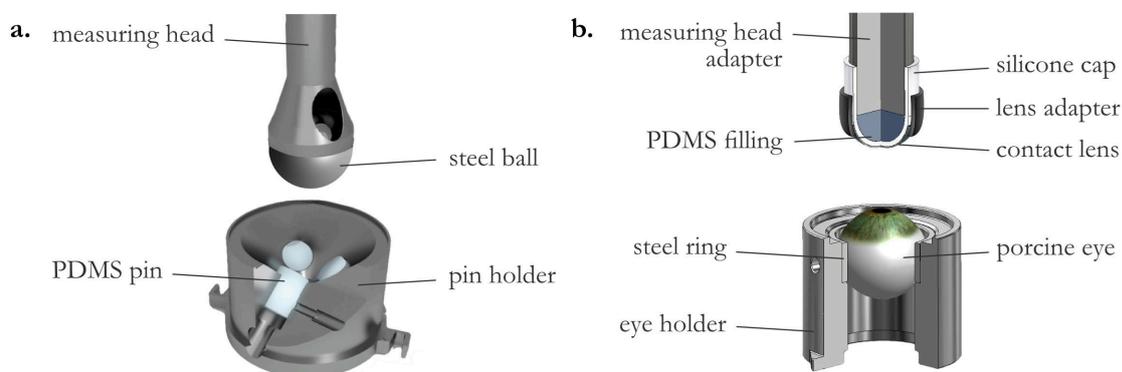
During boundary lubrication, the opposing surfaces are in direct contact to each other with hardly any lubricant in-between. This behavior is typical for slow sliding speeds and usually entails high friction coefficients. With increasing moving speed, the mixed lubrication regime is reached. Here, a significant drop of the friction coefficient is observed as the surfaces begin to separate due to the building hydrodynamic pressure of the liquid. With further increasing sliding velocities, a minimum of the friction coefficient is reached: This is the onset of the hydrodynamic lubrication regime. During hydrodynamic lubrication, the moving parts are fully separated by the liquid film. In this regime, the friction coefficients are low, but slightly increase with increasing speeds since friction within the fluid (*e.g.*, caused by turbulences) becomes stronger.

## Materials and methods

Certain biopolymers – including mucins – possess lubricating properties by which the principles of the classical Stribeck’s theory can (to some extent) be overcome. This is mainly attributable to two lubrication mechanisms on the molecular level, namely sacrificial layer formation and hydration lubrication (**Figure 2.6b**). The sacrificial layer mechanism describes the capability of mucin molecules to adhere to surfaces and, importantly, to reattach after having been sheared off. Those molecular shearing-off/re-attaching cycles lead to reduced friction coefficients as friction energy is dissipated.<sup>91</sup> Additionally, this property of being able to maintain a dynamic but complete biopolymeric layer on the surfaces also constitutes the basis for hydration lubrication: The polyanionic central region of mucin macromolecules acts strongly hygroscopic which allows for establishing a thin water film separating the two moving surfaces even at low sliding velocities. Furthermore, with applied pressure and acting shear forces, the exchange of water molecules between the polymeric layer and the surrounding fluid again dissipates frictional energy.<sup>92,93</sup>

### 2.5.1. Rotational tribology measurements

Friction measurements were carried out on a commercial shear rheometer (MCR 302; Anton Paar) equipped with a rotational tribology setup (T-PTD 200; Anton Paar) as introduced by Böttcher *et al.*<sup>94</sup> (**Figure 2.7a**). In brief, a ball-on-3-pins setup was created by combining a rotating steel sphere (1.4301,  $\varnothing = 12.7$  mm; Kugel Pompel, Wien, Austria) with three PDMS pins ( $\varnothing = 5.5$  mm, for preparation details see **Appendix A1.5**) as counterparts. Prior to each measurement, the pins were rinsed with 80 % (v/v) ethanol and placed into a pin holder. In this preparatory step, particular attention was paid to ensure a symmetrical arrangement of the pins (to achieve a centric rotation of the steel sphere on the three pins). Then, 600  $\mu$ L of lubricant were pipetted into the pin holder; this liquid volume ensured full coverage of the pins during the trial. If not stated otherwise, 20 mM HEPES buffer (4-[2-hydroxyethyl]-1-piperazineethanesulfonic acid buffer, pH = 7.3) was used as a lubricant and the temperature control was set to  $T = 20$  °C. Then, the measuring head was lowered until the steel ball was in contact with the PDMS pins. A normal force of  $F_N = 6$  N was imposed, which, according to Hertzian contact theory (with the Young’s moduli and Poisson’s ratios of PDMS and steel:  $E_{\text{PDMS}} = 2$  MPa,  $\nu_{\text{PDMS}} \approx 0.49$ ,  $E_{\text{steel}} = 210$  GPa,  $\nu_{\text{steel}} \approx 0.30$ ), corresponds to a contact pressure of  $\sim 0.35$  MPa applied to a contact area of  $\sim 8.1$  mm<sup>2</sup>.<sup>95</sup> Then, friction coefficients were measured for logarithmic speed ramps ranging from  $10^3$  to  $10^{-2}$  mm s<sup>-1</sup> using an acquisition time of 10 s per data point; decreasing velocities were chosen to minimize stick-slip effects.



**Figure 2.7: Tribological setups.** The classical setup (a) consists of a rotating steel sphere and a sample holder unit equipped with 3 cylindrical PDMS pins. The measuring head of the ocular tribology setup (b) carries a silicone cap, a contact lens, and an optional lens adapter. The ocular samples are pressed into a steel ring, that is then mounted into the sample holder on the bottom plate. *Subfigure b was adapted from Rickert et al. (2020)<sup>85</sup>.*

### 2.5.2. Ocular tribology measurements

To investigate the tribological performance of contact lenses on corneal tissue, the previously used shear rheometer (MCR302; Anton Paar) was equipped with a custom-made rotational tribology setup comprising a measuring head carrying a contact lens and a sample holder for porcine eyes (**Figure 2.7b**).

#### Preparation of porcine eyes

Fresh porcine eyes were obtained from a local butcher (Metzgerei Karl, Garching, Germany). To prevent tissue dehydration, the eyes were stored in DPBS and used within a maximum of 4 days. Particular care was taken to avoid damaging the corneal tissue. Right before each measurement, the eyeball was cleared of excess tissue and with slight pressure pushed into a steel ring. The applied gentle stretching forces helped to remove macroscopic wrinkles from the corneal tissue. The ring carrying the eye ball was then placed into a tribological specimen holder and secured with two screws. To ensure fair starting conditions for all measurements and to avoid erroneous results caused by the presence of initial damage features on corneal specimens even before being subjected to tribological stress, a three-step quality control was conducted with all prepared ocular samples: First, the corneal surface was subjectively evaluated with regard to apparent superficial, macroscopic damages. Second, topographical images of the tissue were acquired with a laser scanning microscope equipped with a 20× lens (see **chapter 2.6**) and screened for signs of defects on the microscopic level. Third, the initial surface roughness of those topographical profiles was assessed by the root mean square height  $S_q$  (see **Appendix A3**). Any ocular sample that exhibited an  $S_q > 0.4 \mu\text{m}$  or showed signs of surface damages in either of those controls was discarded.

#### Assembling the contact lens holder

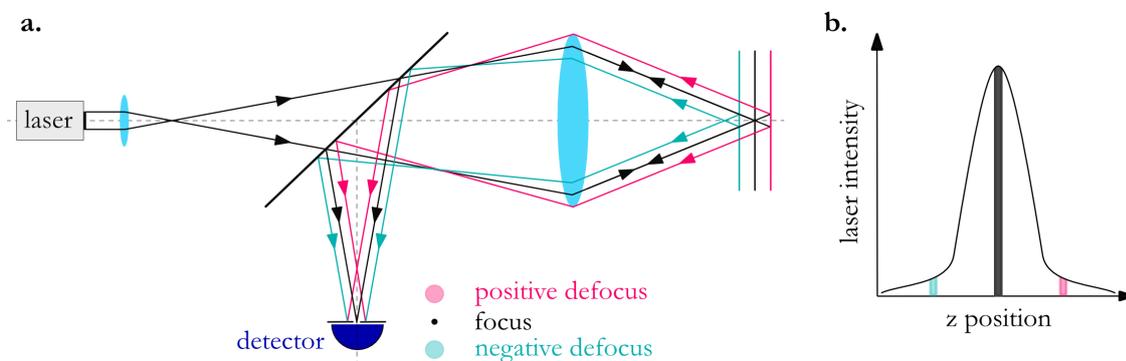
To allow for attaching a contact lens to the measuring head, a custom-made adapter was used and equipped with a spherical silicone cap (Elfeplastic GmbH, Viersen, Germany). To avoid deformation caused by the pressure imposed during the trials, the silicone cap was filled with 2 mL of PDMS (PDMS pre-polymer to cross-linker ratio: 10:1; cured at RT for several days; Sylgard 184, Dow Corning). For each trial, a fresh lens was gently washed with ddH<sub>2</sub>O, blow-dried with pressurized air (Druckluft 67, Kontaktchemie), and placed onto the silicone cap in a central position. To secure this position, PDMS lenses were fixed to the cap by applying instant glue (Krazy Glue; Elmer's Products Inc., Westerville, USA) around the edge, whereas RGP lenses were fixated through a custom-made adapter (**Figure 2.7b**).

#### Tribological measurements

To finally perform tribological measurements, the readily assembled contact lens holder was attached to the measuring system (D-CP/PP 7; Anton Paar), inserted into the shear rheometer (MCR302; Anton Paar), and a calibration of the normal force was conducted. Then, the sample holder carrying the porcine eye was mounted onto a tribology bottom plate (SN81114535; Anton Paar) which comprises a spring-based compensation system to reduce the undesired influence of putative lateral forces. Subsequently, the measuring head was lowered until a distance of 1 mm was established between the cornea and the lens. Then, 500  $\mu\text{L}$  of DPBS were injected into the gap, a liquid trap was installed around the setup to prevent evaporation of the lubricant, and the measurement was started. For all trials, a normal force of  $F_N = 0.14 \text{ N}$  was applied, which corresponds to a contact pressure of  $\sim 30 \text{ kPa}$  (see **Appendix A4**). For PDMS lenses, a rotational velocity of  $\nu = 0.125 \text{ mm s}^{-1}$  was applied, which represents the lower range of physiological friction speeds occurring during regular eye movements. For RGP lenses, a rotational speed of  $\nu = 50 \text{ mm s}^{-1}$  was applied, which corresponds to medium-to-high physiological friction speeds as reached during blinking. All measurements were conducted at 25 °C for a time span of 90 min each.

## 2.6. Laser scanning profilometry

Laser scanning profilometry is a contact-free technique to obtain micro-topographical profiles of surfaces. It is based on the principle of confocal microscopy, a contrast-enhancing technique that employs a spatial pinhole in front of the detector to block out-of-focus light. As depicted in **Figure 2.8a**, the working principle of laser scanning profilometry is as follows: A movable lens is used to focus laser light onto a point of the sample surface, and the reflected light is detected by a photodiode equipped with a pinhole. As the pinhole is at all times kept in a conjugated plane to the focal plane of the objective lens, out-of-focus light is efficiently blocked from the detector. By varying the lens position in  $z$ , a so-called confocal curve describing the intensity of the reflected light in relation to  $z$  is recorded (**Figure 2.8b**). At maximum reflection, the specimen surface is in alignment with the focus of the objective; the height value of the sample can then be derived from the recorded lens position. By scanning the sample surface in  $x$  and  $y$ , a full surface profile can be generated.

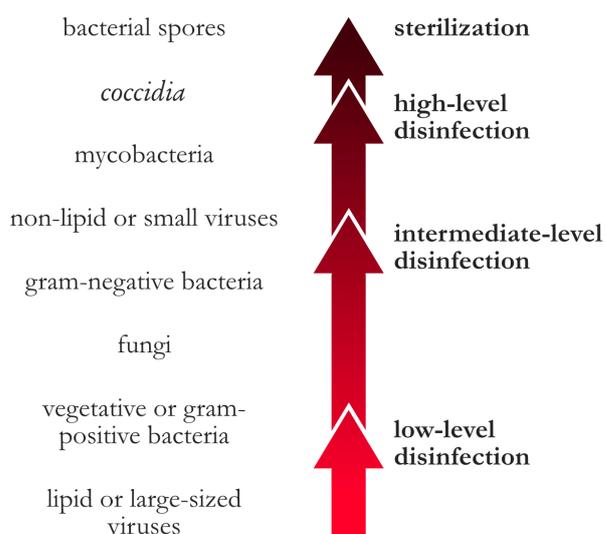


**Figure 2.8: Schematic illustration of the working principle of laser scanning profilometry.** (a) A laser is focused by an objective lens, reflected by the sample surface, and then detected by a photodiode equipped with a pinhole. Since the pinhole is located in a conjugate plane to the objective focus plane, off-focus light is mostly blocked. By varying the distance between the objective lens and the surface, a confocal curve (b) displaying the laser intensity in relation to the  $z$ -position is obtained. At maximum reflection, the sample surface is in the focus of the objective and the height value can be derived from the lens position.

In this thesis, profilometric analyses were conducted to evaluate corneal tissue before and after it was subjected to tribological stress. Therefore, the topography of the samples was imaged with a laser scanning microscope (VK-X1000; Keyence, Osaka, Japan) equipped with a  $20\times$  lens (CF Plan, NA = 0.46; Nikon, Chiyoda, Japan). As stated by the manufacturer, this device reaches resolutions of 0.5 nm in  $z$ -direction and 120 nm in  $x$ -/ $y$ -direction. For image acquisition, porcine eyes were prepared as described in **chapter 2.5.2**. Prior to each profilometric measurement, the corneal surfaces were carefully rinsed with physiological HEPES buffer (HEPES buffer supplemented with 154 mM NaCl) and then gently blow-dried with oil-free pressurized air. Images were then obtained from the center of the porcine cornea, which – when prepared in the used sample holder – corresponds to the highest point of the eye. Image analysis was then performed using the MultiFileAnalyzer software (v2.1.3.89, Keyence): To obtain profiles characterizing the microscopic roughness devoid of any macroscopic surface features, a wave form removal transformation was applied (cut-off wave length: 80  $\mu\text{m}$ ). From the preprocessed images, surface parameters were calculated according to ISO 25178-2 (for more detailed descriptions see **Appendix A3**).

## 2.7. Sterilization and disinfection techniques

Medical devices that come into direct contact with a patient's tissue carry a considerable risk of transmitting environmental pathogens, such as bacteria, viruses and spores, into the body. As this can result in life-threatening infections, a particular challenge in infection control is the decontamination of such medical products. Depending on their antimicrobial efficacy, available decontamination strategies can be subdivided into disinfection or sterilization methods: Whereas sterilization techniques are considered of being capable of killing all types of microorganisms, the antimicrobial efficacy of disinfection methods is limited (see **Figure 2.9**).



**Figure 2.9: Classes of decontamination strategies.** Based on their efficiency regarding different microorganisms, decontamination strategies can be classified as either sterilization or disinfection (low-, intermediate-, or high-level) methods.

The achieved degree of decontamination is quantified by the probability of viable microorganisms remaining on the object after the treatment; and this is referred to as the sterility assurance level (SAL). According to DIN EN 556-1, most medical devices including contact lenses, endotracheal tubes, and urinary catheters require sterility with an SAL of  $10^{-6}$ . To achieve such efficient decontamination, various standardized techniques from mainly three different categories, namely physical, chemical, or radiation-based sterilization, are commonly employed. The (standardized) sterilization methods applied in this thesis include autoclave steam sterilization,  $\gamma$  irradiation and fumigation with ethylene oxide. To systematically study the sturdiness of mucins towards heat up to temperature levels used for autoclavation, different thermal treatments are performed. Moreover, UV irradiation, which classifies as high-level disinfection, is conducted.

Three different sample types were subjected to those treatments: lyophilized MUC5AC powder, 1 % (w/v) MUC5AC solutions, and polymeric samples of medical devices carrying a covalent MUC5AC coating. If not stated otherwise, the mucin powder and the mucin solutions were treated in closed microtubes (volume: 2 mL; Eppendorf SE), and the coated samples of medical devices were air-dried and treated in commercial sterilization bags (Medi Pack GmbH).

## Materials and methods

### 2.7.1. Thermal treatments

Dry heat sterilization refers to exposing an item to conduction heat for a defined period of time. Simply put, microorganisms are destroyed by oxidizing essential proteins and cell constituents. For thermal treatments, the samples were placed into a preheated oven (60 °C, 80 °C, 100 °C, or 120 °C), and left there for 1 h. Even though these temperatures are not sufficient to actually sterilize the samples, they were chosen to systematically study the resilience of mucin molecules against thermal stress up to 120 °C, which approximates the typical temperature level used during autoclaving. For most proteins, temperatures above ~40 °C are already sufficient to induce denaturation.<sup>96</sup> However, the rather unfolded, densely glycosylated mucin might be more resistant to heat than the average protein.

### 2.7.2. Autoclaving

Compared to dry heat, a much more efficient transfer of thermal energy can be achieved by employing saturated steam as a sterilization agent. A steam-based technique widely used for the decontamination of medical devices or lab equipment is autoclavation: Here, the items to be sterilized are exposed to saturated vapor that is generated under pressure – a setup that leads to an increased thermal capacity of the steam. This moist heat then allows for rapidly destroying microorganisms by inducing irreversible denaturation and coagulation of structural proteins and enzymes.

For autoclaving of the mucin powder or the mucin solutions, the lids of the microtubes were only slightly closed so that a small gap remained (to allow the generated steam to exit the tube, thus avoiding overpressure). The coated specimens were packed into sterilization bags as mentioned before. Then, the samples were placed into an autoclave (Systec VX-150; Systec GmbH, Linden, Germany), and a standard sterilizing process (121 °C, 4 bar, 20 min) was applied. Once the process was completed, the tubes were removed from the autoclave and closed immediately. To compensate for lost liquid due to vapor generation (hence, restoring the initial MUC5AC concentration), sterile filtrated ddH<sub>2</sub>O was added to the solubilized samples. All other samples were left as they were until further usage.

### 2.7.3. UV irradiation

UV light is a common disinfection agent used for the decontamination of air, liquids, and nonporous surfaces. The applied UV-C rays (which range from 200 nm to 280 nm) inactivate microorganisms by disrupting nucleic acids and vital proteins in the living cells. Even though this does not directly kill the microorganisms, the damage induced to the genomic system prevents the cells from replicating and surviving. UV disinfection is very convenient and easy, but it comes with a major disadvantage: the UV rays are not able to penetrate most materials. Hence, only surfaces that are directly exposed to the light can be decontaminated. Owing to this limitation, and since several bacterial spores are resistant towards UV irradiation, this technique is mostly categorized as disinfection (instead of sterilization).

For disinfection *via* UV irradiation, opened microtubes containing either lyophilized or solubilized mucin samples were placed into a sterilization chamber (BLX-254; Vilber Lourmat GmbH, Eberhardzell, Germany) and exposed to UV light (254 nm, 5 × 8 W) for different durations (*i.e.*, 10 min, 30 min, or 60 min). The tubes were positioned in such a way, that the open part faced the source of the UV light. Afterwards, the lids were immediately closed. Coated specimens were treated in a petri dish and packed into sterilization bags right after irradiation for 30 min (BLX-254; 254 nm, 5 × 8 W; Vilber Lourmat GmbH). This procedure was chosen to prevent the UV light from being blocked by the packaging.

#### 2.7.4. Treatment with $\gamma$ rays

Compared to the previously introduced UV disinfection, the sterilization performance of irradiation-based decontamination methods can be drastically improved by performing them with electromagnetic radiation of higher energy. For this purpose,  $\gamma$  rays can be utilized, that are emitted during the self-disintegration of certain atomic nuclei. More specifically, short-wave, ionizing  $\gamma$  rays used for sterilization purposes are generated by the decay of the radioisotope Cobalt-60 ( $^{60}\text{Co}$ ). Owing to the high intensity of this radiation,  $\gamma$  rays can usually penetrate the entire (readily packed) product and achieve high sterilization efficiencies similar to those realized with autoclavation. Similar to decontamination with UV light,  $\gamma$  irradiation eliminates microorganisms by breaking down covalent bonds within DNA strands or other cellular compounds. These alterations at the molecular level lead to the death of the organisms or render them unable to reproduce.

For sterilization with  $\gamma$  rays, the samples (stored in tubes or sterilization bags as explained before) were placed into a paper cryo-box, and the whole box was then treated with  $\gamma$  rays using a dose of 25-50 kGy by applying a commercial standard process available at the company steripac GmbH (Calw, Germany; system type: JS9000; complied standards: EN ISO 9001, EN ISO 13 485, EN ISO 11137-1).

#### 2.7.5. Ethylene oxide fumigation

A widely used and very efficient example of chemical sterilization is fumigation with ethylene oxide (EO). During the sterilization procedure, the items are exposed to gaseous EO in a sealed vacuum chamber for several hours. The EO can penetrate air-permeable packaging and sterilizes all accessible surfaces of the product by alkylating the DNA of microorganisms and proteins responsible for cell reproduction. One of the main advantages is that EO sterilization is compatible with a broad range of materials that would be damaged by other sterilization methods. However, one major risk is that toxic EO residues can remain on or in the sterilized goods. To minimize the amount of such toxic residues, extensive aeration (up to several days) of the treated objects after sterilization is mandatory.

Ethylene oxide sterilization was conducted by employing a commercial standard process available at the company steripac GmbH (Calw, Germany; duration: 5 h, temperature: 45 °C, pressure: 610 mbar, average EO concentration: 700 mg L<sup>-1</sup>). This particular sterilization method was only carried out with the coated polymeric samples (packed in sterilization bags) but not with the lyophilized or solubilized mucins. This was mainly due to the following reasons: First, as a fumigation-based method, EO sterilization is not applicable for liquid samples. And second, the porous appearance of the mucin powder renders those samples too susceptible for the deposition of toxic residues that can afterwards not be easily removed from such mechanically instable samples (as the cotton candy-like sample would be destroyed by the required active aeration procedure).

### 2.8. Supervised Machine Learning\*

Supervised learning is a subcategory of ML (Machine Learning), in which models are developed based on labelled data – similar to how parents teach their children to name objects. Therefore, the algorithm needs to be provided a training dataset containing a sufficiently large number of samples. Each sample is represented by input data – *i.e.*, information (descriptors) that is likely to characterize the desired output – and corresponding output labels. Such datasets could, for example, comprise histological images of cancerous tissue (input) labelled with the name of the affected organ (output),<sup>97</sup> or they could link the composition of a biomaterial (input) to its mechanical behavior (output).<sup>98,99</sup> With such information offered, the ML models aim at identifying relationships between the input and the output and can then perform classification or prediction tasks for new data they were not confronted with before.

One goal of this thesis is to employ ML methods for an automated analysis of corneal tissue. This, however, is quite challenging: The topography of such surfaces is highly complex, and it is not easily possible to define an appropriate output label for training samples. Hence, to first gain a better understanding of the topographies of complex, biological surfaces in general and the possibility to analyze those with ML, sample datasets were generated from two types of biological surfaces that can be easily labelled: the surfaces of biofilms and plant leaves. Those datasets were then used to train and test the ability of five different ML algorithms (four model-based ones and one Deep Neural Network) to predict certain (predefined) output labels based on the sample topography provided as an input.

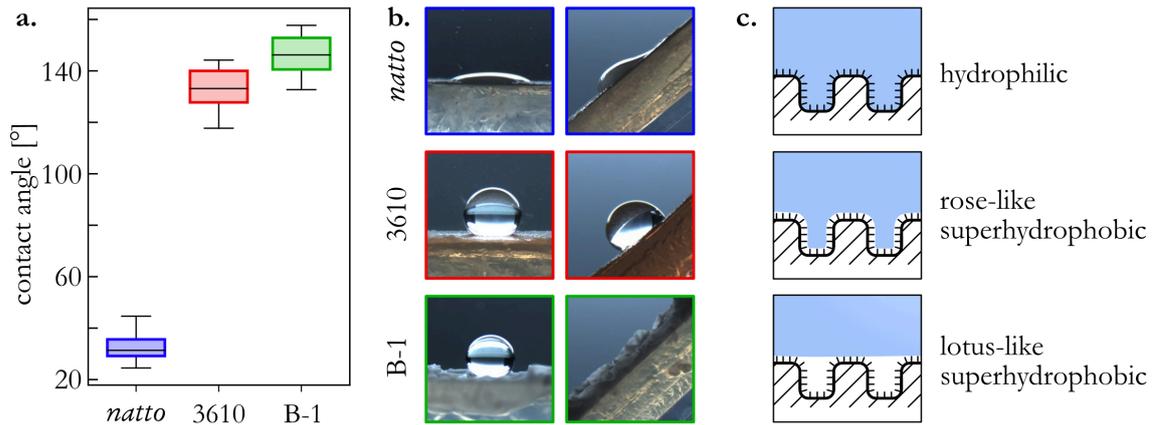
#### 2.8.1. Sample datasets from biofilms and plant leaves

Biofilms are surface-attached colonies of bacteria that form a protective matrix around themselves.<sup>100</sup> This particular biomaterial comes with the advantage that its surface properties can be broadly varied by selecting different bacterial strains and/or growth conditions for biofilm formation. In this thesis, biofilms generated by three different *Bacillus subtilis* variants – namely *natto*, NCIB 3610 (referred to as ‘3610’), and B-1 – were analyzed. A detailed description of the biofilm formation process and the growth media used can be found in **Appendix A5.1**. In brief, planktonic overnight cultures were prepared from bacterial glycerol stocks, and individual droplets of this solution were placed on agar plates enriched with either MSgg (minimal salt glycerol glutamate) medium,<sup>101</sup> or lysogeny broth medium (LB-medium; Luria/Miller). After an incubation time of 24 h at 37 °C, mature biofilms were formed that were then analyzed using a laser scanning microscope (VK-X1000; Keyence) equipped with a 20× lens (CF Plan, NA = 0.46; Nikon): the surface profiles were acquired by taking a 4 × 4 array of images from the central region of the individual biofilm colonies (this image size was chosen based on the spatial constraints of this same central region). Individual images were analyzed with the MultiFileAnalyzer software (v2.1.3.89, Keyence): After applying a tilt correction, 21 different metrological parameters were calculated according to ISO 25178-2 (see **Appendix A3**) and used as input features for each sample.

Importantly, when grown on agar plates enriched with MSgg medium, the biofilms generated from the three chosen bacteria not only differ in visual appearance, but they also exhibit a different surface wettability (characterized by the contact angles as described in **chapter 2.4.4**): Whereas biofilms generated by *natto* are clearly hydrophilic (CA < 90°), the biofilms formed by 3610 and B-1 behave superhydrophobic, *i.e.*, they exhibit contact angles above 120° (**Figure 2.10a**).

---

\* This section follows in part the publications Rickert & Lieleg, *Biophysics Reviews* (2022) and Rickert *et al.*, *ACS Biomaterials Science & Engineering* (2022)



**Figure 2.10: Wetting characterization of different biofilms grown on MSgg agar.** Contact angles (a) and exemplary images of tilt experiments (b) are shown as obtained for biofilms generated from the three *B. subtilis* strains *natto* (blue,  $n = 20$ ), NCIB 3610 (red,  $n = 18$ ), and B-1 (green,  $n = 18$ ) when grown on MSgg agar. Additionally, schematic illustrations of the wetting states observed on those biofilms are depicted (c).

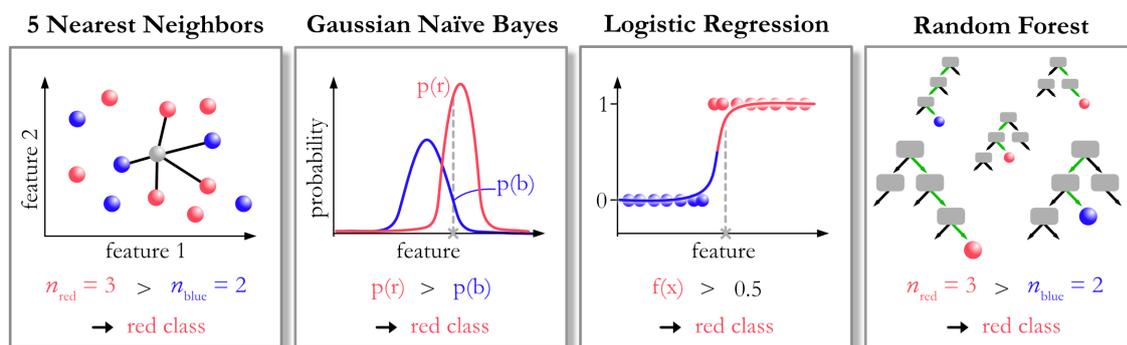
The superhydrophobic surfaces can be further differentiated based on how a droplet behaves on a tilted surface (**Figure 2.10b**). A water droplet placed on a B-1 biofilm directly rolls off (referred to as ‘lotus effect’), whereas a droplet sitting on a 3610 biofilm stays in place even when the surface is severely tilted (known as ‘rose petal effect’). These effects arise from a complex interplay between repulsion and adhesion forces acting on both, the nano- and the microscale (**Figure 2.10c**). On lotus-like surfaces, air pockets separate the surface and the water on all roughness scales, leading to high contact angles and close to zero adhesion forces acting between the two materials. In contrast, whereas air-pockets on the nano-scale lead to high contact angles of rose-like surfaces, microscopic roughness features can be impregnated by water such that adhesion forces are established that are strong enough to keep the drop in place even when the surface is tilted (or completely flipped). Hence, for the samples included in the generated dataset, both, the name of the bacterial strain used to grow the biofilm and the observed surface wettability (*i.e.*, ‘hydrophilic’ for *natto*, ‘rose-like’ for 3610, or ‘lotus-like’ for B-1) were assigned as output labels for each sample.

Additionally, the same set of *Bacillus subtilis* strains was also used to generate biofilms on agar that was enriched with LB medium. Here, the surface of biofilms generated from *natto* or B-1 exhibit the same wettability as that of biofilms from the same strain grown on MSgg agar (see **Appendix A5.2**). However, for biofilms grown from 3610, the wettability changes: whereas on MSgg agar, those biofilms behave rose-like, 3610 colonies grown on LB agar exhibit hydrophilic surfaces.

Last, a dataset containing three different plant samples was created such that each of the previously discussed surface wetting states is represented. Therefore, plant leaves of *Caladium praetermissum* (IPEN: XX-0-M-V/0926; kindly provided by the Botanical Garden Munich-Nymphenburg, Staatliche Naturwissenschaftliche Sammlungen Bayerns, Munich, Germany) were chosen to represent lotus-like surfaces, *rosa* ‘harlekin’ plants (Edeka Ernst, Garching, Germany) were included as rose-like surfaces, and *Rhododendron* leaves (obtained from a local garden) were added to represent hydrophilic surfaces (see **Appendix A5.2**). From each plant sample, pieces of approximately 3 cm × 2 cm size were freshly cut, gently rinsed with ddH<sub>2</sub>O and blow-dried with pressurized air before the profilometric measurements. To obtain a planar surface suitable for experiments, those pieces were attached to a glass slide on all four sides using fixation tape. Special care was taken to not damage those parts of the leaf/petal surfaces which were later analyzed. From those plant specimens, a 10 × 10 array of images was recorded from the ‘upper’ side of the leaves (for *Caladium* and *Rhododendron*) or the ‘outer’ side of the petal (for *Rosa*).

## 2.8.2. Supervised Machine Learning algorithms

Supervised ML algorithms were used to perform classification tasks, *i.e.*, to assign samples to discrete, predefined classes. Therefore, the algorithms were first trained on a provided dataset (training data) and they were then tasked to make predictions on unknown samples (query samples). In all those datasets, input variables that quantify individual measurable characteristics of a sample are referred to as ‘features’, and outputs assigned to samples are called ‘(class) labels’. If not stated otherwise, all supervised Machine Learning algorithms were implemented using python (Python Software Foundation; Python Language Reference, version 3.8.3; <http://www.python.org>)<sup>102</sup> with several extensions for data handling and visualization (Numpy<sup>103</sup> v1.19.4, Pandas<sup>104,105</sup> v1.2.2, Matplotlib<sup>106</sup> v3.2.2, Scipy<sup>107</sup> v1.6.0, and Seaborn<sup>108</sup> v0.11.1) as well as the Machine Learning toolbox scikit-learn<sup>109</sup> (v0.24.1). A schematic overview of how the four model-based algorithms used in this thesis make decisions is depicted in **Figure 2.11**.



**Figure 2.11: Schematic illustration of the four model-based, supervised ML algorithms used in this thesis.** The  $k$  Nearest Neighbors classifier (a) identifies the  $k$  samples that are closest to the query point in an  $n$  dimensional space ( $n$  equals the number of features) and assigns the query sample to that class which has the highest number of representatives (within those  $k$  samples). The Gaussian Naïve Bayes algorithm (b) calculates conditional probabilities and categorizes samples based on a ‘fits better’ principle. The Logistic Regression algorithm (c) models the probability of classes by assigning importance weights and fitting a sigmoidal function through all features. The Random Forest classifier (d) performs predictions in a popular-vote-like manner considering the decision of a large number of random, uncorrelated decision trees.

### $k$ Nearest Neighbors classifier

The simple but powerful  $k$  Nearest Neighbors ( $k$ NN) algorithm follows the assumption that similarity between samples is accompanied by proximity in the data space; in other words, similar samples are expected to come with similar inputs. Here, no internal model was developed; instead, the training data was memorized, and for each query sample, its  $k$  nearest neighbors (in the  $n$ -dimensional input feature space;  $n$  corresponding to the number of features) were determined based on the Manhattan distance<sup>110</sup> (**Figure 2.11a**). Each of those  $k$  nearest points was given a weight equal to the inverse of its distance to the point of interest. Considering those weights, the class of the query point was then determined in a popular vote-like manner based on the neighbors’ classes. For this algorithm, the optimal choice of the number of neighbors taken into account strongly depends on the data distribution. Overall, low  $k$  values render the classification more distinct, whereas higher  $k$  values lead to less noise. In this thesis,  $k = 5$  was selected as it was shown to entail the highest accuracies on the initial biofilm dataset (see **Appendix A5.3**). To find the nearest neighbors, a brute force search was chosen. Minmax feature scaling ( $x_{\text{scaled}} = (x - \min_{\text{feature}}) / (\max_{\text{feature}} - \min_{\text{feature}})$ ) was applied to each feature of both, the training and query data; this helps avoiding that parameters of larger magnitude influence the prediction overly strongly.

### Gaussian Naïve Bayes classifier

As one example of a probabilistic classifier, a Gaussian Naïve Bayes (GNB) algorithm<sup>111,112</sup> was selected (**Figure 2.11b**). Such an algorithm can – although being comparably simple – outperform even highly sophisticated classification methods.<sup>113</sup> The GNB classifier is based on Bayes’ theorem, which defines the relation between the conditional probability  $P(y|x_1, \dots, x_n)$  of a class  $y$  and a dependent feature vector  $(x_1, \dots, x_n)$  as follows:

$$P(y|x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n | y)}{P(x_1, \dots, x_n)}. \quad (4)$$

With the naïve assumption of conditional independence between all features, the resulting classifier can be mathematically formulated as (for a more detailed derivation, see **Appendix A5.4**):

$$\hat{y} = \arg \max_y P(y) \prod_{i=1}^n P(x_i | y) \quad (5)$$

With this approach, the *posterior* probability  $P(y | x_i)$  is determined using a prior assumption on the output distribution  $P(y)$  and a likelihood  $P(x_i | y)$  on the input data given the labels as described by **equation 6**. Here, it is assumed that the distributions of the features is Gaussian, where the expected value  $\mu_y$  and the variance  $\sigma_y^2$  are assessed with maximum likelihood.<sup>114</sup> The predicted class of  $y$  is the one which maximizes the *posterior* probability (**Figure 2.11b**).

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \quad (6)$$

### Multinomial Logistic Regression

As one example of a linear Machine Learning approach, Multinomial Logistic Regression (MLR, **Figure 2.11c**) was applied using an ‘lbfgs’ solver<sup>54, 55</sup> and L2 regularization.<sup>115</sup> Briefly, given the training data  $X$ , the probability  $P(y_j = c | X)$  that the output of the  $j^{\text{th}}$  query point would be class  $c$  (out of  $C$  total possible classes) was calculated using the following softmax function:<sup>56, 57</sup>

$$P(y_j = c | X) = \frac{e^{\beta_c x_j}}{\sum_{h=1}^C e^{\beta_h x_j}} \quad (7)$$

Therein,  $\beta_c$  denotes the coefficient weight vector obtained from the model fit and  $\mathbf{x}_j$  denotes the feature vector of the query point  $j$ . To compensate for a (putative) imbalance present in the dataset, class weights were applied that are inversely proportional to the class frequencies in the input data. Furthermore, standard scaling ( $x_{\text{scaled}} = (x - \text{mean}_{\text{feature}}) / \text{standard deviation}_{\text{feature}}$ ) was applied to prepare the data and to reduce the iteration depth.

## Materials and methods

### Random Forest classifier

The Random Forest (RF, **Figure 2.11d**) classifier is derived from the family of ensemble classifiers, which combine the predictions of multiple independent base estimators.<sup>58</sup> Here, those base estimators are decision trees – flow chart-like hierarchical decision-making models consisting of nodes (*i.e.*, decision stages) and branches connecting these nodes. Starting with a root node, the training data are (based on individual features) split in a stepwise manner by creating and answering simple true/false questions until all splits contain samples of one class only. A new (query) sample can then be classified by running through the tree using the input values of this new sample and the previously established decision rules. The ensemble decision is then made based on a popular vote-like principle by averaging the probabilistic prediction of each tree with equal contributions. To ensure that the individual trees are not correlated, each tree is trained using an individual bootstrap sample (*i.e.*, a random subset of the original dataset) and it can – at every node – only choose from a random set of features to split the data into subsets.

### Deep Neural Network

A Deep Neural Network (DNN) consists of multiple interconnected layers, each of which comprises a defined number of neurons, *i.e.*, units performing a linear sum of inputs. Such neurons can be activated with a typically non-linear activation function and then transform the input variables by applying trainable weights and a bias. All neurons of the same layer receive the previous layer's output as input, yet they perform a different transformation. For training, the model is first initialized with weights, and the input of the training data is processed. The output of the model is then compared to the true values of each data point and an error is calculated. By applying backpropagation, this error is then used to evaluate the effect of each weight to the final prediction error in the form of a gradient. Through this gradient, each weight is corrected, and the inputs are processed through the model to generate a new prediction. This process is recursively repeated until the error has converged to an acceptable minimum.

The hyperparameters of the DNN used in this thesis were optimized by a grid search-based approach, and the following ones were found to perform optimally: The network contains 3 hidden layers comprising 20, 12 and 8 neurons each, with the layers being activated by a ReLu,<sup>60</sup> ReLu, and tanh<sup>61</sup> activation function, respectively. The output layer was activated with a softmax function, and the loss function was chosen to be the categorical cross entropy.<sup>62</sup> For implementation, the TensorFlow library<sup>63</sup> (v2.5.0) was used.

### 2.8.3. Repeated stratified $k$ -fold cross-validation

To evaluate the performance of the implemented ML algorithms on a given set of data, repeated stratified  $k$ -fold cross-validation<sup>64-66</sup> was conducted. For this purpose, the data was split into  $k$  folds (if not stated otherwise,  $k = 5$ ), each of which was used as a test set once (*i.e.*, as a set of query samples to be classified); for each selection of a test set, the rest of the data was collectively used to train the algorithm. To avoid noisy estimates arising from this data separation, the repeated version ( $n_{\text{repeats}} = 10$ ) of the common cross-validation was chosen. Here, in each repetition, a different randomization was applied to the dataset. Moreover, the stratification (see **Appendix A5.5**) ensured an approximately equal sample distribution within each of the  $k$  folds compared to the whole training dataset, *i.e.*, that each fold was a good representative of the whole. This approach helped compensating for the imbalance of the training data (*i.e.*, the somewhat unequal numbers of training data obtained per biofilm strain).<sup>67</sup>

## 2.9. Correlation-based feature elimination\*

For many experimental research problems, such as those targeted in this thesis, the amount of data available for training a ML model is very limited. An important strategy to combat the problems arising from such data sparsity is feature elimination – a method that aims at reducing the dimensionality of an input feature space. Most of those strategies exclusively focus on analyzing pairwise correlations, or they eliminate features based on their relation to a selected output label or by optimizing performance measures of a certain ML model. However, those strategies do not necessarily remove redundant information from datasets and cannot be applied to certain situations, *e.g.*, to unsupervised learning models. Hence, in this thesis, a model- and output-independent feature elimination algorithm (named ‘NETCORE’) was developed, where the size of a feature vector is reduced by considering both, redundancy and elimination efficiency.

### 2.9.1. Molecular dataset

The NETCORE algorithm was developed and tested on a dataset created for an experimental drug loading study using different antibiotics. This rather small dataset contains 14 commonly used antibiotics, each of which is characterized by a feature vector summarizing 9 physico-chemical properties of the molecules. Since the NETCORE algorithm conducts an output-independent analysis, no further label was required. For all molecules included in this dataset, a 3D conformer representation was obtained from PubChem.<sup>116</sup> This 3D structure was then imported into MarvinSketch (version 21.17.0, 2021, Chemaxon, <http://www.chemaxon.com>), and the following 9 physico-chemical properties were determined using the Calculator Plugins (version 22.3, 2022) of MarvinSketch (for details regarding the molecules included in the dataset and the parameters used for calculations, please refer to **Appendix A6**):

- **rotatable bond count:** the number of bonds in a molecule that allow for a conformational change of the molecule geometry through rotation around the respective bond<sup>117</sup>
- **van der Waals volume:** the volume occupied by a molecule, *i.e.*, space impenetrable for other molecules with thermal energies at ordinary temperatures<sup>118</sup>
- **molecular weight:** the molecular mass calculated from standard atomic weights<sup>119,120</sup>
- **charge:** the total charge of a molecule at pH 7.4 calculated from the sum of its microspecies, weighted according to the microspecies distribution at pH 7.4 (the natural form was excluded)<sup>121</sup>
- **dipole moment:** the net molecular polarity specified as the electron density which is unequally distributed between the atoms of a molecule<sup>122</sup>
- **partition coefficient:** the logarithm (logP) of the concentration ratio of a chemical dissolved in two different phases (here: octanol and water), where both concentrations are at equilibrium;<sup>123</sup> in the case of multiple microspecies of a molecule, the non-ionic one was considered
- **aromatic ring count:** the number of aromatic rings in a molecule calculated from the smallest set of smallest aromatic rings (SSSAR)<sup>124</sup>
- **hydrogen bond acceptor sites** (average at pH 7.4): the sum of lone electron pairs in a molecule, which are available for establishing a hydrogen bond<sup>125</sup>
- **hydrogen bond donor sites** (average at pH 7.4): the sum of hydrogen atoms connected to atoms in the molecule that have hydrogen donor properties

---

\*This section follows in part the publication Rickert *et al.*, *APL Machine Learning* (2023)

### 2.9.2. Network-based, correlation-driven redundancy elimination (NETCORE)

Based on the database created as described above, a feature vector was generated, which quantitatively describes the physico-chemical properties of the set of molecules studied (**Figure 2.12a**). Then, a network-based feature elimination was performed to reduce the dimensionality of the feature space (**Figure 2.12b-f**). More precisely, the goal was to create a new feature vector, that contains features that are as distinct as possible from any other of the included features but still represent the eliminated features from the initial dataset in an optimal manner. Therefore, features from the input space that are strongly correlated with others were identified and removed as those features contain – at least to a certain extent – redundant information. If not stated otherwise, all implementations were conducted in python (Python Software Foundation; Python Language Reference, version 3.9.12; <http://www.python.org>).<sup>102</sup> Here, the following plugins were used: Pandas<sup>104,105</sup> (v1.4.2), NetworkX<sup>126</sup> (v2.7.1), Numpy<sup>103</sup> (v1.22.3), Seaborn<sup>108</sup> (v0.11.2), and Matplotlib<sup>106</sup> (v3.5.1). The working principle of the developed NETCORE algorithm is as follows:

#### Step 1: Correlation matrix

First, the feature space is analyzed by creating a Pearson's correlation matrix (**Figure 2.12b**): Here, a pairwise comparison of all  $n$  physico-chemical properties (= features) is conducted by calculating the linear correlation strength between feature  $x$  and feature  $y$  as described by the Pearson's correlation coefficient  $r$  (see **equation 8**).

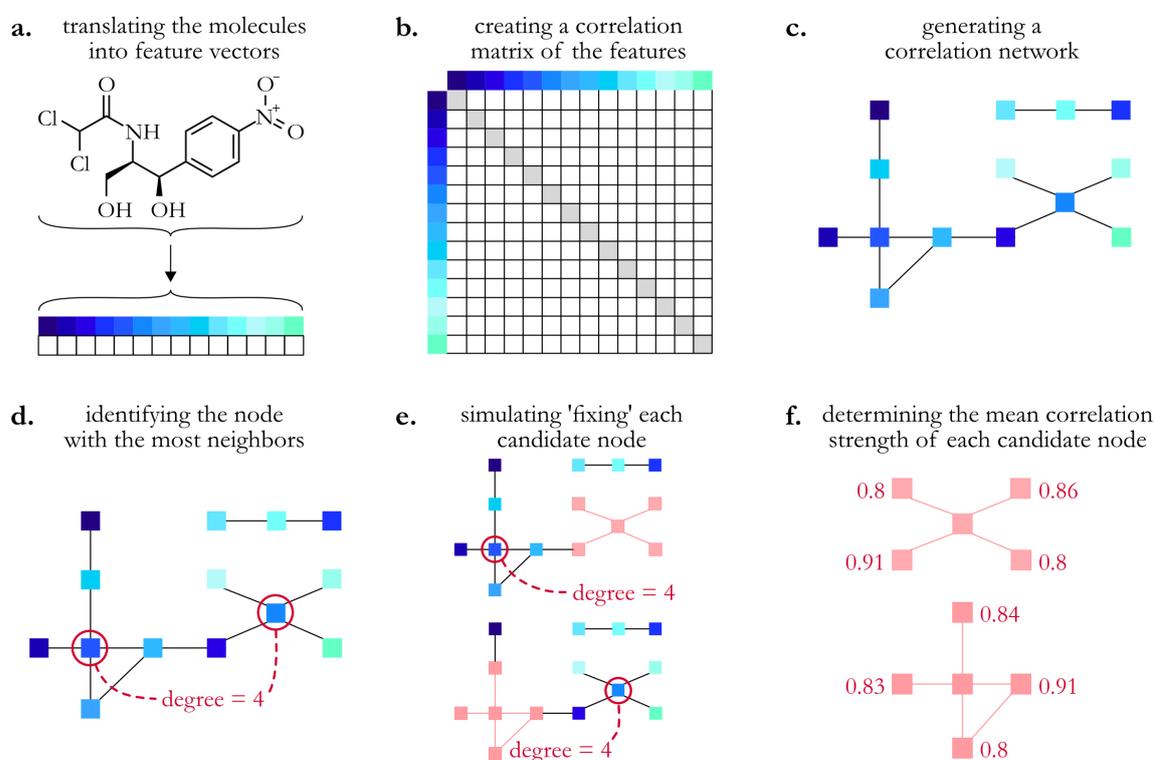
$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2]} \sqrt{[n \sum y^2 - (\sum y)^2]}} \quad (8)$$

Importantly, features for which no correlation coefficients can be calculated, are eliminated in this step without being added to the reduced feature vector. This mainly affects two types of features: features that are characterized by non-numerical inputs, and features that are represented by the exact same value for all samples. Whereas, for the latter, it is desirable to eliminate such features directly (as they contain no valuable information), non-numerical features might be important. Hence, datasets containing such non-numerical features should be preprocessed by performing a basic label encoding (as, for instance, provided by the Pandas<sup>104,105</sup> and scikit-learn<sup>109</sup> libraries) to convert non-numerical inputs to numerical ones; then, those features can be considered in the subsequent stages of the NETCORE algorithm. However, for the dataset analyzed in this thesis, such a conversion was not needed as all input features already were of numerical data types.

#### Step 2: Network generation

Based on the obtained  $n \times n$  correlation matrix calculated in the previous step, a correlation network is created (**Figure 2.12c**). This correlation network consists of  $n$  nodes, and each node represents one physico-chemical property (feature). Then, edges connecting the nodes are added only if the absolute value of the correlation coefficient between the two respective properties equals or exceeds a defined threshold  $t$  (if not stated otherwise,  $t = 0.6$  was selected). Last, each added edge is assigned a weight corresponding to the correlation coefficient of the connected feature pair.

In this step, the algorithm can be adjusted by changing the correlation threshold  $t$ , *i.e.*, the correlation strength needed to consider two features to be redundant. Here, since  $t$  is used to filter absolute values of correlation coefficients, only values between 0 and 1 are reasonable. A value of 1 would yield a correlation network in which edges are only present between fully correlated features. Since, in the further procedure of the algorithm, features can only be eliminated when they are connected to another feature *via* an edge (see **steps 3** and **4**), selecting a threshold of 1 entails the elimination of fully redundant features only. In contrast, a correlation threshold of 0 leads to a fully connected correlation network, *i.e.*, each feature is connected to every other feature. Hence, per the criteria described in the following sections, only one feature would be selected to represent the whole dataset. Thus, choosing the right correlation threshold is not trivial: smaller values for  $t$  lead to a stronger reduction of the feature vector, whereas larger values for  $t$  enforce that a stronger redundancy is required for elimination. As a rule of thumb, absolute correlation coefficients below 0.35 are considered to represent weak correlations, moderate correlations are indicated by coefficients between 0.36 and 0.67, and strong or very strong correlations lead to coefficients above 0.67 or 0.9, respectively.<sup>127</sup> As a default, a correlation threshold of  $t = 0.6$  was chosen for the feature elimination tasks performed in this thesis.



**Figure 2.12: Schematic overview of the different steps of the network-based feature elimination process.**

First, selected physico-chemical properties of molecules are translated into a feature vector (**a**) and a correlation matrix representing the correlation strength of the features is calculated (**b**). Using a pre-defined correlation threshold, this correlation matrix is converted into a network; here, each node represents a feature and edges denote correlation strengths that exceed the threshold value (**c**). The network is then reduced in a step-wise manner by identifying the feature of highest centrality, adding this feature to the (initially empty) new feature vector, and eliminating it together with all neighbors that are directly connected. To identify the features of highest centrality, the following steps are iteratively conducted: first, the feature of highest degree (*i.e.*, the highest number of connected neighbors) is identified (**d**). In case of parity, the same analysis is conducted with those networks, that would remain if the respective node and its connected neighbors were to be removed from the network (**e**). Finally, if still no unambiguous decision is possible, the mean correlation strength of the node to its direct neighbors is calculated and the feature with the highest average correlation strength is chosen.

## Materials and methods

### Step 3: Elimination of isolated nodes

After generating the correlation network, a first analysis step is performed to identify all independent features, *i.e.*, nodes that have no edge to any other node. Such isolated nodes represent features that exhibit no pairwise correlation to any other feature exceeding the previously defined correlation threshold. In other words, based on the correlation threshold chosen, features are considered to be independent when they are not sufficiently represented by any other feature in the dataset. As those features contain information that is not redundant, such features are identified first and directly added to the new, reduced feature vector. Importantly, this step is conducted during each iteration of the NETCORE algorithm – more specifically, before each iteration of the centrality analysis described in **step 4**.

### Step 4: Centrality analysis

To finally identify a reduced feature set that accurately represents the whole dataset without containing redundant information, the previously generated network is iteratively reduced by repeatedly analyzing the degree centrality of the nodes. Different from previously reported approaches that used eigenvector centrality<sup>128</sup> to calculate importance weights based on the ability of a feature to discriminate between classes,<sup>129</sup> the degree centrality criterion used here iteratively identifies and resolves local ‘clusters’ of features containing redundant information – and this process does not require any class labels. Another difference is that the degree centrality is directly derived from the graph without any spectral analysis. To assess the degree centrality, the number of direct neighbors a node has in the interconnected correlation network is determined (**Figure 2.12d**). If a unique node of highest degree (*i.e.*, a node with the highest number of connected neighbors) could be identified, the corresponding feature is added as the first entry to an initially empty vector (which thus becomes an  $n \times 1$  vector with  $n = 1$ ). At the same time, this node is removed from the network together with all of its directly connected neighbors (this whole procedure will hereafter be referred to as ‘fixing a parameter’). If such a unique node of highest degree could indeed be identified, the centrality analysis iteration is already complete, and a second iteration can be initiated with the remaining network. In the case of parity of multiple nodes with the same degree, however, the result of fixing each of those ‘candidate nodes’ is individually considered and the respective remaining networks are analyzed in more detail: now, the degree of all nodes remaining in the network (after eliminating the candidate node and its direct neighbors) is determined (**Figure 2.12e**). Then, in order to maintain a network for further analysis that is as well-connected as possible and to avoid unfavorable segmentation of the network into disjunct subnetworks, the particular candidate node leading to a network that contains the node of highest degree in all candidate networks is chosen to be fixed next. Finally, if no clear decision is possible after this step either, the mean correlation strength of each candidate node to all of its direct neighbors is calculated (**Figure 2.12f**). From those candidate nodes, the node showing the highest mean correlation strength is identified, added to the reduced feature vector and eliminated from the network together with its direct neighbors. However, in some cases where a unique node can still not be identified, the feature that was evaluated first (by random choice) is added to the feature vector. This is, for instance, the case for a connected pair of nodes which is otherwise isolated in the network (and thus, has no further neighbors as required for an algorithm-based selection). Overall, **step 3** and **step 4** are repeated until all nodes in the network were successfully eliminated; at this point, a maximally reduced feature vector is established. This iterative approach ensures that, even though the network is changed when a node and its neighbors are eliminated, the next node is again chosen based on optimal centrality.

### 2.9.3. Evaluation regarding multicollinearity

To test the success of the NETCORE algorithm regarding the elimination of correlating features, the variance inflation factor (VIF) was calculated – a common metric to quantify the severity of occurring multicollinearity.<sup>130</sup> In brief, the VIF describes how much the standard error of the regression coefficient of a predictor variable (feature) in a linear regression model is increased due to multicollinearity. In particular, the VIF of a certain feature  $i$  can be obtained according to the following **equation 9**, where  $R_i^2$  denotes the  $R^2$ -value obtained by regressing feature  $i$  based on all remaining features.

$$VIF_i = \frac{1}{1 - R_i^2} \quad (9)$$

## 2.10. Unsupervised Machine Learning\*

As stated before, one key objective of this thesis is to utilize ML techniques to automate the analysis of corneal tissue surfaces. Therefore, a dataset containing profilometric images of porcine corneas was created, in which each sample was characterized by a set of surface parameters obtained from 3D height profiles. However, as no predefined labels were available for corneal surfaces, supervised ML approaches – such as previously used for the prototype dataset of biofilm and plant samples – were not applicable. Instead, unsupervised ML methods had to be employed for the analysis of this dataset.

Unsupervised ML encompasses various data-driven techniques, where the algorithm is supplied with unsorted input data and given the freedom to independently draw its own conclusions; this can be achieved either by autonomously clustering the samples or by detecting trends, similarities, extreme points, or patterns within the data. In this thesis, different clustering algorithms were used to group corneal surfaces, and the results were compared to those obtained by a panel of untrained researchers. If not stated otherwise, all operations conducted for unsupervised ML were implemented using python (version 3.9.16)<sup>102</sup> with several extensions for data handling and visualization (Numpy<sup>103</sup> v1.22.4, Pandas<sup>104,105</sup> v1.5.3, Matplotlib<sup>106</sup> v3.7.1) as well as the ML toolbox scikit-learn<sup>109</sup> (v1.2.2).

### 2.10.1. Generation of a dataset containing corneal images

Fresh porcine eyes were prepared as described in **chapter 2.5.2** (but no sample was discarded due to quality constraints), and prior to any further treatment, profilometric images of each sample were acquired (see **chapter 2.6**). These untreated samples already showed quite diverse surface morphologies, as some surfaces exhibited signs of initial damage that was not inflicted in the lab (*e.g.*, as a consequence of the slaughtering procedure). To increase the sample size of damaged tissue, some samples were subjected to one of three disruptive mechanical treatments: applying tribological stress with a contact lens (see **chapter 2.5.2**); rubbing the surface with sand; applying gentle cuts with a scalpel. Those treated samples were again profilometrically imaged. However, no label (such as the damage cause) was added to any specimen for two reasons: first, for samples showing damage that was not created in the lab, the cause of surface damage was unknown; second, for actively damaged samples, the initial state of the tissue was not always comparable. Thus, a fully unlabeled dataset containing 2775 images was generated.

---

\* This section follows in part the publication Rickert *et al.*, ‘Unsupervised Machine Learning to Topographically Analyze Corneal Tissue Surfaces’, submitted 2023

### 2.10.2. Data preparation and preprocessing

From the previously generated dataset containing images of corneal samples, 21 metrological surface parameters were obtained for each sample (see **chapter 2.6**) to serve as an input feature vector for each corneal sample. To then prepare the dataset for an ML-based analysis, different preprocessing steps were conducted and evaluated. First, to normalize the features such that all features have the same range, Minmax feature scaling ( $x_{\text{scaled}} = (x - \min_{\text{feature}}) / (\max_{\text{feature}} - \min_{\text{feature}})$ ) was implemented using the scikit-learn toolbox.<sup>109</sup> Once this scaling was applied, each feature ranged from 0 to 1.

Second, to remove outliers from the dataset, the interquartile range (IQR) method was employed. Therefore, the IQR for each feature of the dataset was calculated as the difference between its third quartile (Q3) and its first quartile (Q1). Those calculations were conducted using the Numpy<sup>103</sup> quantile methods. Then, any data point that fell outside of the range defined by 1.5 times the IQR below Q1 and 1.5 times the IQR above Q3 of at least one feature was eliminated.

The third approach conducted for data preprocessing was a Yeo-Johnson Power Transformer<sup>131</sup> implemented using the scikit-learn library.<sup>109</sup> This method is especially suited to handle non-normal and heteroscedastic datasets, as it helps stabilize both variance and skewness by mapping the data to a distribution that is as close to Gaussian as possible. The Yeo-Johnson Power Transformer was selected over the more commonly used Box-Cox Transformer<sup>132</sup> due to its ability to handle non-positive inputs. To fine-tune the shape of the transformation, the Yeo-Johnson Transformer (**equation 10**) was parameterized by the hyperparameter  $\lambda$ , which was determined *via* a maximum likelihood estimation.

$$x_i^{(\lambda)} = \begin{cases} [(x_i + 1)^\lambda - 1]/\lambda & \text{if } \lambda \neq 0, x_i \geq 0 \\ \ln(x_i + 1) & \text{if } \lambda = 0, x_i \geq 0 \\ -[(-x_i + 1)^{2-\lambda} - 1]/(2 - \lambda) & \text{if } \lambda \neq 2, x_i < 0 \\ -\ln(-x_i + 1) & \text{if } \lambda = 2, x_i < 0 \end{cases} \quad (10)$$

### 2.10.3. $k$ -Means clustering

To group similar data points into clusters based on a given feature vector (comprising a set of surface parameters), a  $k$ -Means clustering algorithm was employed. This algorithm first randomly initialized  $k$  cluster centers and iteratively updated them based on the samples assigned to each cluster until sufficient convergence was reached. Therefore, each sample was assigned to the cluster with the closest centroid. After assigning all samples, the new centroid of each cluster was determined. This procedure was repeated until the variance of the samples within each cluster was minimized and the variance between the clusters was maximized.

The number of clusters  $k$  was predefined as stated in the individual tests. The initialization method for the cluster centers was set to ' $k$ -means++', and the maximum number of iterations for each run of the algorithm was set to 300. To increase the chances of finding the global optimum, the number of iterations the algorithm conducted with different centroid seeds was set to 10 with no seed provided for the random number generator (to allow for different initializations and reproducibility). In addition to the  $k$ -Means clustering algorithm, five other clustering methods were tested. All details regarding those algorithms can be found in **Appendix A8.1**.

#### 2.10.4. Intrinsic validation metrics

With missing ground truth labels, model evaluation has to be performed using model-intrinsic validation criteria. Therefore, three different metrics were used: the Silhouette coefficient, the Calinski-Harabasz index, and the Davies-Bouldin index. To achieve better comparability between clustering results obtained in different feature spaces, the coefficients were normalized as described in **Appendix A8.2**).

##### Silhouette coefficient

The Silhouette coefficient<sup>133</sup> (SIH) characterizes the achieved clustering performance by comparing the inter-cluster distances to the intra-cluster distances. In other words, when  $a$  denotes the mean distance between a sample and all other samples within the same cluster, and  $b$  denotes the mean distance between a sample and all other points in the nearest other cluster, the silhouette coefficient for this individual sample is given as:

$$SIH = \frac{b - a}{\max(a, b)} \quad (11)$$

For the evaluation of a full clustering result, the SIH is averaged over all individual samples. A value close to 1 relates to a model with well-defined clusters, values close to 0 indicate overlapping clusters, and negative values, especially those close to -1, indicate clusters with incorrectly assigned samples.

##### Calinski-Harabasz index

Since different from comparing distances in the cluster result, the Calinski-Harabasz index<sup>134</sup> (CH) quantifies the cluster dispersions, it is often referred to as the ‘variance ratio criterion’. In detail, for a dataset  $E$  containing  $n_E$  samples divided into  $k$  clusters, the CH is defined as the ratio between the sum of between-cluster dispersion and the within-cluster dispersion:

$$CH = \frac{\text{tr}(B_k)}{\text{tr}(W_k)} \times \frac{n_E - k}{k - 1} \quad (12)$$

Here,  $\text{tr}(B_k)$  is the trace of the between-group dispersion matrix and  $\text{tr}(W_k)$  is the trace of the within-cluster dispersion matrix given by **equations 13** and **14**, respectively. Therein,  $C_q$  represents a set of  $n_q$  points in cluster  $q$  with the cluster centroid  $c_q$ . Accordingly,  $c_E$  indicates the centroid of  $E$ .

$$B_k = \sum_{q=1}^k n_q (c_q - c_E)(c_q - c_E)^T \quad (13)$$

$$W_k = \sum_{q=1}^k \sum_{x \in C_q} (x - c_q)(x - c_q)^T \quad (14)$$

Per definition, higher numbers of the CH index correspond to dense and well-separated clusters. Other than the SIH, however, the CH is not limited to the range of -1 and 1.

## Materials and methods

### Davies-Bouldin index

Last, the Davies-Bouldin index<sup>135</sup> (DB) was calculated to quantify the average ‘similarity’ between clusters, by comparing the distance between clusters with the size of the clusters. Here, values close to zero indicate well-separated clusters and higher numbers imply higher (undesired) cluster similarity. The DB index is defined as

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} R_{ij}, \quad (15)$$

in which  $R_{ij}$  defines the similarity between two clusters based on the cluster diameters  $s_i$  and  $s_j$  (average distance of each point of a cluster to its centroid) and the distance between the cluster centroids  $d_{ij}$  (**equation 16**).

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \quad (16)$$

### 2.10.5. Extrinsic validation metrics

To compare different clustering results, the adjusted mutual information (AMI) score<sup>136</sup> was determined. This AMI score is an extended version of the mutual information (MI) score: whereas the MI quantifies the similarity of two clustering results solely based on the probabilities of their joint and marginal probability distributions, the AMI additionally accounts for chance. The AMI is normalized such that values of 1 indicate perfect agreement between two clustering results, and a score close to 0 describes that mutual information is due to chance alone. For two label assignments  $U$  and  $V$ , the AMI is defined according to **equation 17**.

$$AMI = \frac{MI - E[MI]}{\text{mean}(H(U), H(V)) - E[MI]} \quad (17)$$

Here,  $H(U)$  and  $H(V)$  quantify the entropy of  $U$  and  $V$ , respectively, and  $E[MI]$  indicates the expected value for the MI (for details, please refer to **Appendix A8.3**). The AMI is independent of the actual label names of a cluster and permutations of those do not have any effect. Additionally, the AMI is symmetric, *i.e.*, comparing result A to result B will deliver the same score as comparing result B to result A.

### 2.10.6. Panel evaluation of the corneal dataset

In addition to the ML-driven approaches employed, a panel of 10 individuals (4 female, 6 male) was tasked to group the corneal samples based on their surface appearance. Here, to keep the sample size manageable for humans, a reduced dataset containing every seventh image of the initial dataset was provided to the panel (in total 397 images). For better comparability, the number of clusters to be formed was predefined to  $k = 4$ . Each participant was given as much time as they wanted to visually inspect images of the 3D height profiles. During this time, the participants identified a ‘description’ for each of the four clusters. Once suitable clusters and their corresponding characteristics were determined, the participants were shown each image individually (in random order) and exclusively assigned those images to one cluster. Importantly, to briefly validate the consistency achieved by each participant, one duplicate sample (randomly chosen but identical for each panelist) was added to the dataset to be clustered.

## 2.11. Statistical analysis and graphical representation

Prior to testing the statistical significance between two samples, a normal data distribution was tested for each sample with a Lilliefors test, and a two-sample F-test was employed to check for equal variances. To test for significant differences between normally distributed samples, a two-sample  $t$ -test was applied when homogeneity of variances was met, whereas a Welch's  $t$ -test was performed for unequal variances. For samples that were not normally distributed, a Wilcoxon–Mann–Whitney test was performed. All statistical analyses were performed using Matlab (version R2019a, MathWorks, Natick, USA), and differences were considered statistically significant if a  $p$ -value below 0.05 was obtained.

All graphs shown in this work were generated using the software GraphPad Prism (Prism 9 for macOS; Version 9.3.1 (350), December 7, 2021; GraphPad Software LLC; San Diego, USA) together with Affinity Designer (version 1.10.4, Serif Europe Ltd., West Bridgford, UK). If results are presented in boxplots, the bottom and top of the box represent the first and third quartile, respectively. The line inside the box denotes the median. The length of the whiskers is based on data within the 1.5 interquartile range of the upper and lower quartile. Data points outside of this range are denoted as outliers (crosses or circles).



### 3. Mucin macromolecules for biomedical applications

To exploit the beneficial properties of mucins for biomedical applications, mucin macromolecules can either be used as a base material, or they can be used to generate surface coatings on existing devices. Both of those approaches are followed and evaluated in this chapter. First, functionalized mucin molecules that can be covalently crosslinked with UV light are used to develop a bio-ink for 3D printing. Objects fabricated with this bio-ink are then subjected to different environmental conditions and the material response is observed. Second, mucins are used to generate surface coatings on contact lenses, and the achieved surface properties, such as the wettability and the tribological performance of the lens on corneal tissue, are evaluated. Last, as sterilizability is of utmost important for a material to be used in a medical context, the sturdiness of mucins towards several sterilization methods is investigated.

#### 3.1. UV-curable mucin gels for 3D bioprinting\*

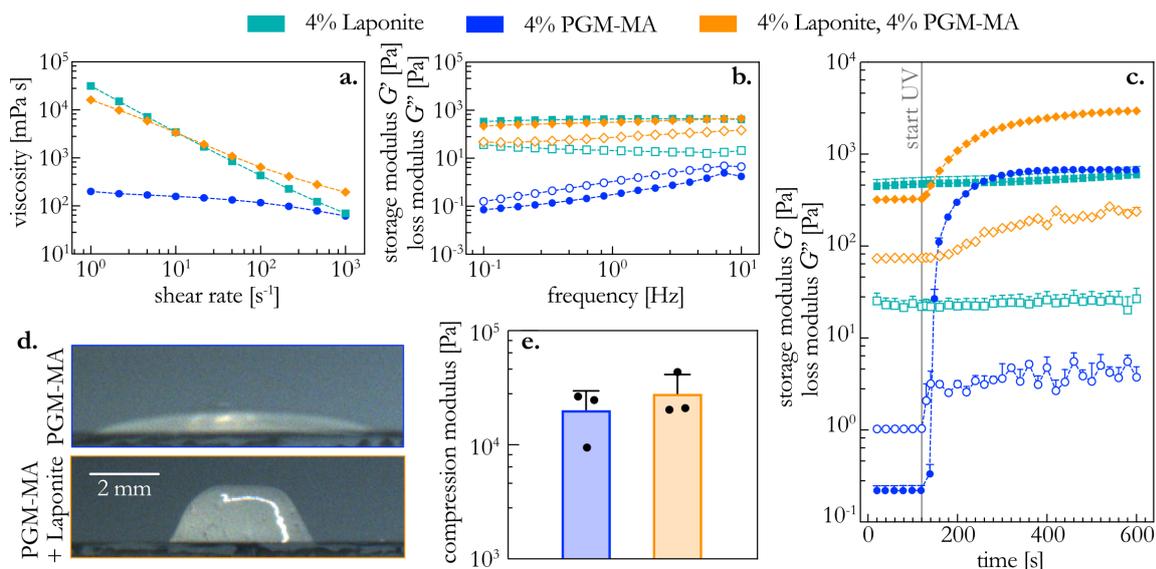
Owing to the potential to manufacture patient-tailored medical devices on-demand and at relatively low costs, 3D printing (3DP) is a very promising technology on the road to personalized medicine. Especially, extrusion based 3DP methods, such as fused deposition modeling of thermoplastics or direct ink writing of photosensitive polymer resins, have demonstrated a high potential for the production of customized (bio)medical and surgical devices or scaffolds based on tomography data or computer aided design.<sup>137-139</sup> In those techniques, the liquid ink – *i.e.*, a viscous polymer solution or a melt – is gently pushed through a nozzle that follows a predetermined path dictated by a virtual model; during this procedure, the ink is deposited onto a printing bed to build the desired object in a layer-by-layer fashion. To be suitable for such 3DP, the ink material must meet different specifications in terms of printability and stability.<sup>140-142</sup> This encompasses the ability to be extruded from a nozzle or printing head (which requires the material to flow under modest pressure), to be quickly curable upon deposition (*e.g.*, by photopolymerization, thixotropic effects, or temperature changes) and to exhibit mechanical properties that are sufficiently strong to support the following layers.<sup>143,144</sup> However, especially for hydrogel-based bio-inks, meeting all these requirements can be very difficult: they tend to spread on the printing bed immediately after deposition and their intrinsic mechanical properties are comparatively weak.<sup>145,146</sup> Hence, the rheological behavior of those inks needs to be tailored to achieve sufficient printability.

Throughout this thesis, the base component for all bio-ink formulations developed is commercial PGM functionalized with methacrylic anhydride (= PGM-MA, **Figure 2.2**). An aqueous solution comprising 4 % (w/v) PGM-MA shows shear-thinning properties with viscosities slightly above 100 mPa s and exhibits frequency-dependent characteristics typical for a viscoelastic fluid (**Figure 3.1a,b**). However, the presence of methacrylate groups on the PGM-MA enables photo-initiated covalent crosslinking of the mucins *via* free-radical polymerization:<sup>147</sup> upon addition of a suitable photo-crosslinker, polymerization of the PGM-MA molecules can be initiated by UV irradiation (**Figure 3.1c**), and this leads to a rapid and strong change of the material properties. Within 20 s of exposure, the initially liquid solution transitions into a gel state: now, the material response is dominated by elastic properties. After approximately 4 min of curing, the elastic modulus of this gel reaches a plateau value of around 0.8 kPa. Once the material has fully cured, this value is stable and quite constant over the whole range of shear frequencies tested (see **Appendix A1.6**), which is typical for a covalently cross-linked (bio)polymer gel.<sup>148</sup>

---

\* This section follows in part the publication Rickert *et al.*, ‘A Mucin-based Bio-Ink for 3D Printing of Objects with Anti-Biofouling Properties’, submitted 2023

## Mucin macromolecules for biomedical applications



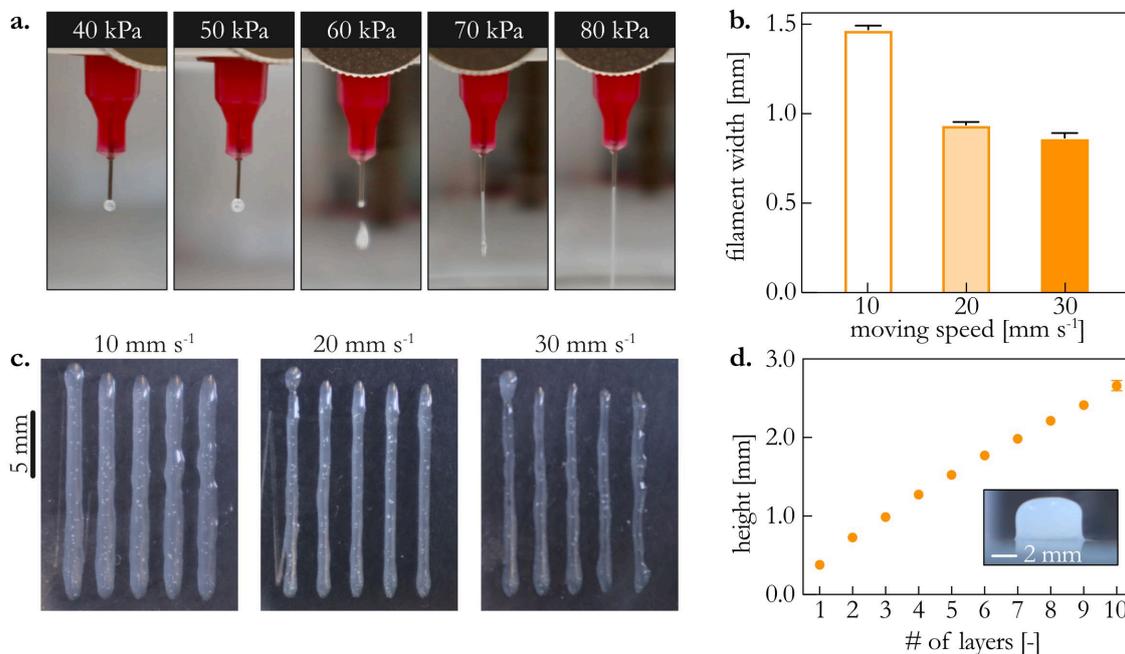
**Figure 3.1: Mechanical characterization of different bio-ink formulations.** The viscosity of the inks at different shear rates (a) and the viscoelastic moduli measured over a frequency spectrum (b) or at a set frequency during UV irradiation (c) are displayed. Moreover, images of 4 layers of bio-ink printed without any curing (d) and the compression moduli of the cured inks (e) are depicted. Error bars in (a,b,c,e) denote the standard error of the mean obtained from  $n \geq 3$  measurements (if not visible, they are smaller than the symbol size).

However, as mucin solutions are dominated by their viscous properties, they do not exhibit sufficient intrinsic stability before being cured. Hence, in a next step, those UV-curable mucin solutions are still used as a base component, but the rheological properties of the bio-ink are tuned to optimize them for extrusion-based additive manufacturing. For such a 3DP application, the viscosity of the mixture is an important parameter for two reasons: first, it influences the extrusion flow and defines the required extrusion pressure; second, the viscosity determines how fast a liquid material spreads on the printing bed, *i.e.*, how long the material can (at least to a certain degree) maintain its shape before covalent cross-links are introduced by UV illumination. Whereas the second aspect benefits from a high viscosity of the ink, a strongly viscous ink limits the achievable moving speeds during the printing process and hampers the smoothness of the printed filaments. Thus, to ensure good printability, the ink material should exhibit shear-thinning characteristics. Without additives, a bio-ink comprising PGM-MA only shows this desired shear-thinning behavior; yet, overall, its viscosity is rather low, which calls for adjustment. Moreover, after printing, the PGM-MA solution is not able to maintain its shape (Figure 3.1d), which further underscores the need to improve the viscous properties of the bio-ink.

There are several strategies to achieve control over the viscoelastic properties of a complex fluid, such as using a thermo-responsive additive or another mechanical adjuvans that acts as a stabilizing support. To avoid thermal stress on the biopolymeric PGM-MA, the strategy of choice is to add a temperature-independent stabilizing agent to the bio-ink. More specifically, the effect of adding either methylcellulose or nanoclay (Laponite® XLG) is evaluated. Indeed, either of those two additives successfully increases the viscosity of the bio-ink by several decades while maintaining the desired shear-thinning properties (Figure 3.1a and Appendix A1.7). Whatsoever, the nanoclay additive outperforms methylcellulose regarding several aspects: First, the Laponite nanoclay can be thoroughly mixed with the mucin, and this allows for higher additive concentrations to be used while still achieving a homogenous, mono-disperse ink formulation. Second, different from what is achieved with methylcellulose, the Laponite-containing bio-ink forms a viscoelastic gel even before curing (Figure 3.1b and Appendix A1.7). This particular property turns out to be important when comparing the extrusion performance of the two bio-ink

variants: a certain elasticity of the material is required to achieve a sufficiently high intrinsic stability of the printed product until the covalent cross-links generated by UV treatment are established (**Figure 3.1d**). Hence, for further experiments, only the Laponite-based formulation is considered. With this particular bio-ink formulation, the curing time of the material is very similar to that of pure PGM-MA solutions, and the elastic properties of the final, cured material are slightly increased to  $\sim 2$  kPa. As a material to be used in a biomedical context, those mechanical characteristics strongly influence the ability of the cured bio-ink to integrate well into soft tissue. Here, in addition to shear stress, compression is another important mechanical load to consider. Hence, unconfined compression tests are conducted with the bio-ink, revealing compression moduli of around 10-11 kPa (**Figure 3.1e**). Both, those compression moduli and the elastic moduli obtained during shear measurements, are in a similar range as values previously reported for mammalian soft tissue and organs.<sup>149-151</sup>

In a next step, the performance of this improved bio-ink formulation (containing 4 % Laponite and 4 % PGM-MA) is evaluated in greater detail using a direct in writing system (for details, see **Appendix A1.1**). The first objective is to identify a suitable pressure for the printing process to extrude the bio-ink into a continuous filament without interruption. A filament drop test indicates that pressure levels of 40 kPa and 50 kPa are insufficient to extrude the bio-ink, as only a small droplet of ink is formed at the needle tip without further extrusion (**Figure 3.2a**). At higher pressure levels of 60 kPa and 70 kPa, multiple individual droplets and small filament-like sections, respectively, are observed. Consequently, pressure levels of 80 kPa are required to obtain a continuous filament extrusion. Hence, for all further procedures, an extrusion pressure level of 80 kPa is chosen. When this pressure level is then used to extrude filaments at different moving speeds, a slow moving speed ( $10 \text{ mm s}^{-1}$ ) leads to thick, poorly defined filaments, whereas doubling this moving speed (to  $20 \text{ mm s}^{-1}$ ) improves both, the shape and reproducibility of the filament (**Figure 3.2b,c**). However, when the moving speed is further increased to  $30 \text{ mm s}^{-1}$ , the thickness of the filament only slightly decreases, but the shape of the filament appears more irregular and asymmetric. Hence, a moving speed of  $20 \text{ mm s}^{-1}$  is identified as best suited for the current setup.

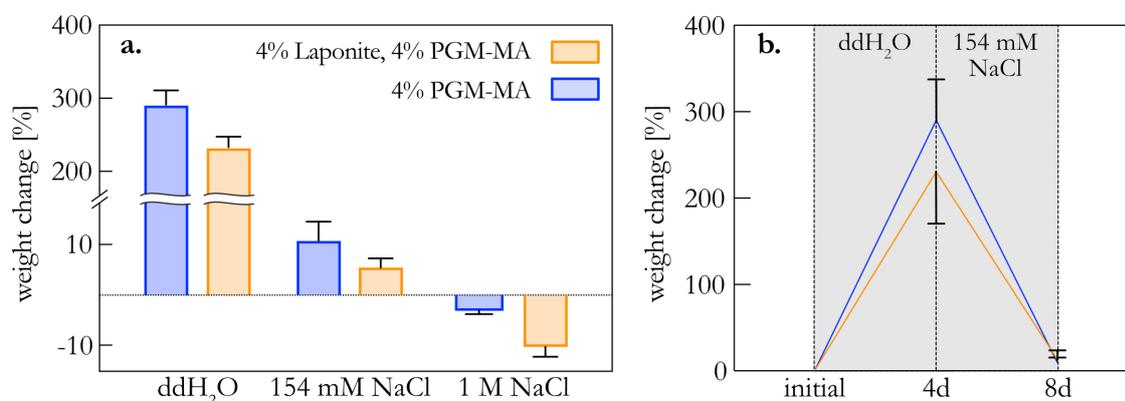


**Figure 3.2: 3D printing of the mucin-based composite bio-ink.** A drop test was conducted to identify a suitable pressure level for extrusion of the bio-ink (**a**). Then, filaments were printed at different moving speeds and their thickness was quantified (**b,c**). Last, the height increase per printed layer was monitored (**d**). Error bars denote the standard error of the mean obtained from  $n = 15$  (**b**) and  $n = 3$  (**d**) samples measured (if not visible, they are smaller than the symbol size).

## Mucin macromolecules for biomedical applications

Having identified suitable printing parameters for the developed bio-ink, a cylinder with a diameter of 5 mm is printed and the increase in sample height after extruding and curing each layer is monitored. As desired, when increasing the number of printed layers, a linear increase of the sample height is observed (**Figure 3.2d**), and a nicely shaped cylinder is generated (**Figure 3.2d**, inset). Thus, this first printing trial indicates that the developed ink material is indeed able to generate a 3D-printed sample of desired shape with sufficient intrinsic stability.

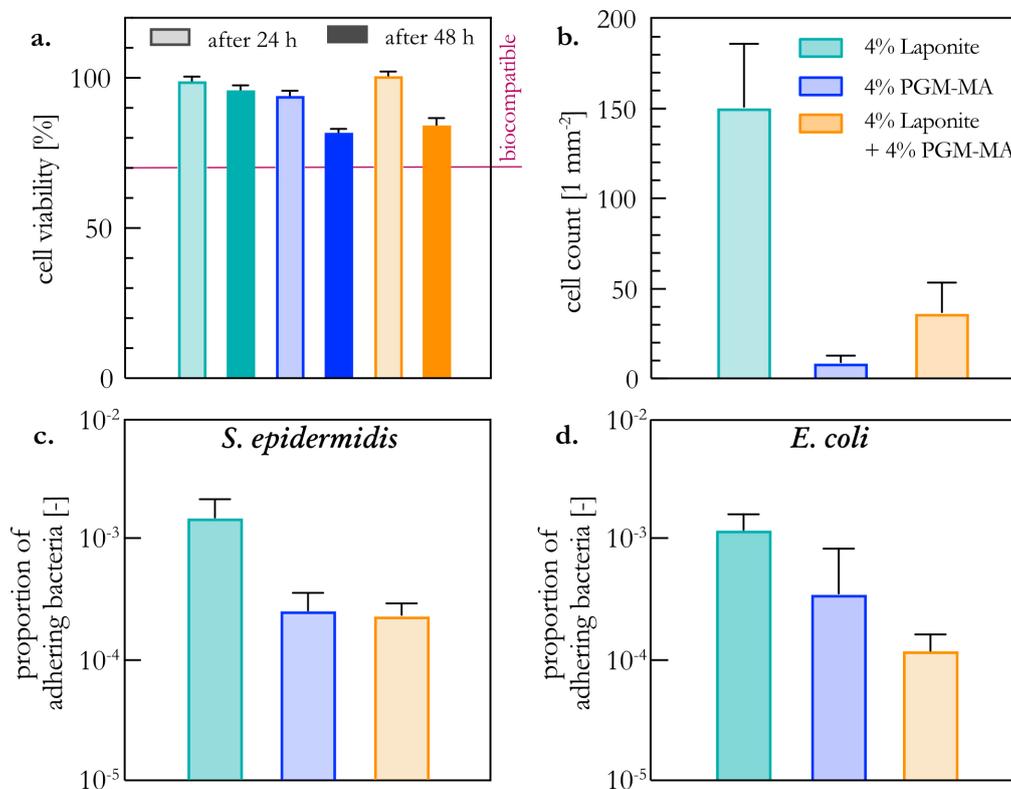
After evaluating the performance of the bio-ink in a 3DP setup, the material response of cured bio-ink samples towards different liquid environments is examined. To do so, cylindrical samples are created and are immersed into aqueous solutions supplemented with different amounts of salt. The change in weight of the samples after 5 days of storage is then determined (**Figure 3.3a**). Those swelling tests are only conducted with PGM-MA-containing samples, since pure Laponite specimens are not able to properly maintain their shape when immersed in a water-based fluid. For bio-ink samples stored in ddH<sub>2</sub>O, a strong swelling behavior is observed: An increase in weight of nearly 300 % is observed for pure PGM-MA specimens and of 250 % for samples containing both, PGM-MA and Laponite. This strong swelling is probably driven by the repulsive electrostatic forces acting between the strongly anionic mucin molecules: *Via* electrophoretic light scattering (**Appendix A1.8**), the zeta potential of the PGM-MA macromolecules is measured to be  $-(26.1 \pm 0.4)$  mV. With increasing salt concentrations, sample swelling is drastically reduced or even converted into shrinkage: weight increases of only 10 % are observed for samples immersed into a physiological saline solution, whereas samples incubated in a 1 M NaCl solution even shrink by up to 10 %. These findings agree with the notion that electrostatic repulsion forces drive sample swelling: With increasing NaCl content, the solubilized salt ions will increasingly shield the negatively charged groups on the mucin molecules (Debye-screening), which allows for a much more compact conformation of the bio-ink sample.



**Figure 3.3: Swelling behavior of different bio-ink formulations.** The weight change that cured bio-ink specimens undergo when exposed to liquid environments with different salt concentrations (**a**) and its reversibility (**b**) are displayed. Error bars denote the standard error of the mean obtained from  $n \geq 3$  samples.

Consistent with this picture, this sample swelling appears to be reversible: when a freshly cured sample (which was prepared in ddH<sub>2</sub>O as described in **chapter 2.2**) is first immersed into ddH<sub>2</sub>O (until full swelling is achieved) and then placed into water containing physiological NaCl concentrations (154 mM), the sample nearly fully recovers its initial volume. In other words, it is possible to adjust the volume of printed structures created with this mucin-based bio-ink by incubating them in aqueous solutions containing different ion concentrations. This characteristic might even enable the application of such objects as ion-sensitive actuators such as a valve that opens or closes depending on the surrounding ion concentration. Importantly, even in their swollen state, the printed samples are stable and do not break up into fragments.

One essential aspect for materials to be used in a biomedical context, is sufficient biocompatibility. Hence, to test for putative cytotoxic effects, a leaching test together with a WST-1 assay is performed following ISO 10993 using epithelial HeLa cells as a model cell line (for details, see **Appendix A1.9**). Within the time intervals tested (*i.e.*, 24 h and 48 h, respectively), no cytotoxic effects are detected as evidenced by cell viability values exceeding 70 % (**Figure 3.4a**). This outcome is consistent with prior studies that have reported the non-toxicity of both, crosslinked mucin hydrogels<sup>152,153</sup> and Laponite gels.<sup>154,155</sup> In addition to the biocompatibility of the bio-inks, the ability of cells to adhere to the surface of cured ink samples is assessed by incubating cells on the surface of the different samples (for details, see **Appendix A1.9**). As expected,<sup>155-157</sup> pure Laponite gels can be colonized by HeLa cells; in contrast, on pure PGM-MA gels, hardly any attached cells are found (**Figure 3.4b**). Consistently, the addition of mucins to the Laponite gel considerably reduces the number of cells adhering to the gel. This finding indicates the possibility to tune the cell adhesion behavior of the bio-ink by varying the mixture ratio of the two ink components. For the particular ratio studied in this thesis, the cell experiments demonstrate that mucin-based bio-inks exhibit a cell-repellent behavior while maintaining a high biocompatibility.



**Figure 3.4: Biocompatibility and adhesion of cells and bacteria to cured bio-ink samples.** The biocompatibility of the cured bio-ink formulations as tested with a leaching test according to ISO 10993 (a) and the adhesion of cells to the surfaces of cured bio-ink samples (b) are displayed. Subfigures (c) and (d) show the proportion of CFUs attached to the bio-ink samples after incubation of the specimens with bacterial suspensions. Error bars denote the standard deviation obtained from  $n \geq 3$  samples.

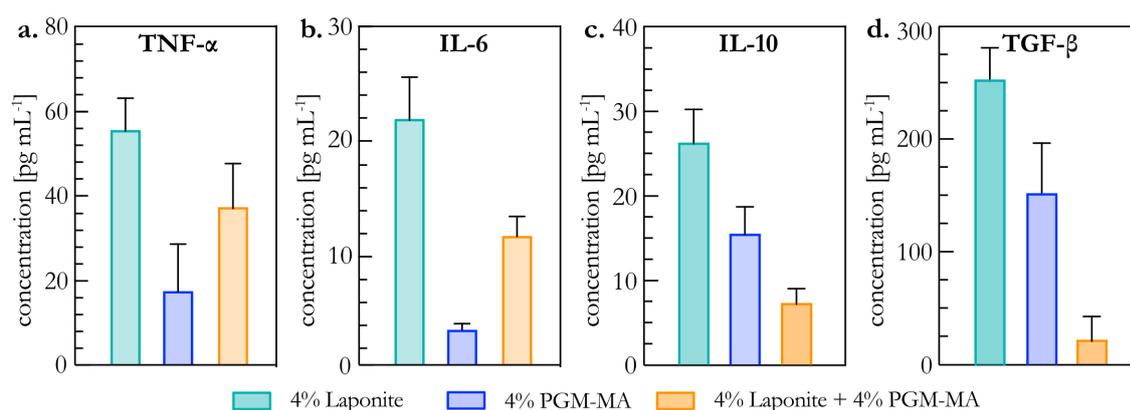
After demonstrating that mucin drastically reduces cell adhesion to the surfaces of bio-ink samples, next, the anti-bacterial behavior of those specimens is tested (for details, see **Appendix A1.9**). Several previous studies have suggested that surface coatings generated from mucin macromolecules can reduce bacterial adhesion,<sup>79,158</sup> and it seems possible that mucin-enriched gels might be able to show a similar effect. To test this hypothesis, the susceptibility of objects generated with the different bio-inks to become colonized by bacteria is investigated. In detail, such specimens are incubated with bacterial suspensions

## Mucin macromolecules for biomedical applications

of either *Staphylococcus epidermidis* or *Escherichia coli* and the proportion of bacteria adhering to the samples is determined. For specimens comprising Laponite only, approximately one out of 1,000 bacteria from the inoculation solution adheres to the samples. This already low number, however, is further reduced by a factor of 10 for PGM-MA-containing samples (**Figure 3.4c,d**). Conclusively, these results indicate that objects printed with the mucin-based bio-ink exhibit good bacteria-repellent properties.

However, even in the absence of pathogens, an artificial object that gets into contact with tissue surfaces elicits a foreign body response through its interaction with immune cells. This could entail several negative consequences resulting in rejection of the printed constructs. To evaluate the immune response triggered by each of the bio-ink components and the final bio-ink formulation, the concentration of different pro- and anti-inflammatory cytokines secreted by macrophages that are seeded onto the different materials is assessed (for details, see **Appendix A1.9**).

The observed cytokine secretion of macrophages seeded onto pure Laponite hydrogels is significantly higher than that detected for macrophages seeded onto PGM-MA hydrogels (**Figure 3.5**). This result is consistent with prior findings concerning the immediate and long-term reactions of macrophages that were placed on mucin-derived materials: In those studies, it was demonstrated that mucin-based gels have a broad dampening impact on the cytokine expression of macrophages.<sup>159,160</sup> However, for the composite bio-ink formulation, differences are observed between the pro- and anti-inflammatory cytokines: For the pro-inflammatory cytokines (*i.e.*, TNF- $\alpha$  and IL-6), the expression levels for the composite is in-between the secretion levels obtained for the individual materials. In contrast, for the anti-inflammatory cytokines (*i.e.*, IL-10 and TGF- $\beta$ ), the macrophages seeded onto the composite produce the lowest cytokine secretion. As stated before, mucin molecules broadly dampen the secretion of both, pro- and anti-inflammatory cytokines. For Laponite, such a dampening effect was previously only observed for anti-inflammatory cytokines,<sup>161</sup> whereas no effect (neither an increasing nor a dampening one) was observed for pro-inflammatory cytokines.<sup>162</sup> Hence, the intermediate values obtained for TNF- $\alpha$  and IL-6 likely result from the combination of two materials that influence the cytokine expression differently. In contrast, for IL-10 and TGF- $\beta$ , both individual components dampen this immune response, which is why the combination of the two appears to result in a synergistic effect. Together, these results indicate that printed constructs created from mucin-derived bio-ink formulation have an immunomodulatory effect.



**Figure 3.5: Cytokine expression on the different bio-ink formulations.** Macrophages were seeded onto cured bio-ink samples of the individual components (*i.e.*, Laponite hydrogels and PGM-MA hydrogels) and on the final bio-ink formulation. The secretion of the two pro-inflammatory makers TNF- $\alpha$  (a) and IL-6 (b) and the two anti-inflammatory cytokines IL-10 (c) and TGF- $\beta$  (d) are displayed. Error bars denote the standard deviation obtained from  $n = 8$  samples. For all four cytokines, all three samples are significantly different to each other based on a  $p$ -level of 0.05 (for simplicity, no asterisks were added to the figure).

### 3.2. Covalent mucin coatings on silicone contact lenses\*

In addition to directly creating 3D objects from mucin-based materials, it may also be beneficial to transfer the excellent properties of mucins to existing medical devices *via* surface coatings. One class of medical devices that could particularly benefit from the outstanding properties of mucins are contact lenses. A frequent problem many people wearing contact lenses complain about is ocular dryness and an irritating feeling in the eye; both of those symptoms can arise from a disruption of the tear film as caused by an insufficient wettability of a contact lens and the resulting increased friction.<sup>163,164</sup> The ensuing compromised lubrication of the eye can result in pain and tissue damage on either the cornea (as known for *keratoconjunctivitis sicca*)<sup>165</sup> or the eye lid (referred to as lid wiper epitheliopathy).<sup>166</sup> In addition to applying mechanical stress to the eye, contact lenses constitute a physical barrier which limits the transport of oxygen towards the eye surface. It was shown that a deficient oxygen supply of the cornea can lead to limbal hyperemia and increased invasion of limbal vessels into the cornea, hence promoting undesired neo-vascularities.<sup>167-169</sup> However, a major problem of newly developed oxygen-permeable contact lenses, such as those made from PDMS-based materials without any hydrogel additives, is their hydrophobic surface; and as outlined above, a poor wettability of contact lenses can lead to a disruption of the tear film and damage of the eye. Apart from unpleasant sensations, however, such tissue damage can also pave the way for microbial keratitis, a serious complication in which either bacteria, fungi, or amoeba are involved.<sup>170</sup>

In fact, wearing contact lenses was shown to be the main cause for microbial keratitis in the western society.<sup>171</sup> From a mechanistic point of view, the mechanical disruption of the tear film on the corneal epithelium caused by insufficient contact lens wettability is thought to impair the endogenous protective strategies present in healthy eyes, thus rendering the cornea more susceptible to infection.<sup>172-175</sup> Moreover, an artificial lens material is typically much more prone to bacterial colonization than healthy corneal tissue: freely floating microbes can adhere to the surface of a lens where they form contagious biofilms in close proximity to the epithelium, which cannot be easily cleared.<sup>176,177</sup> Ocular pain, corneal opacities, photophobia, scarring, decreased vision, and, if left untreated, blindness can be a consequence.

Previous studies have shown that surface coatings generated with mucin glycoproteins can effectively target both of those problems by reducing friction<sup>86</sup> and preventing bacterial adhesion on model surfaces.<sup>43</sup> Hence, in the following, the potential of mucin coatings to improve the surface characteristics of hydrophobic PDMS contact lenses is investigated. One approach to generate such coatings is to simply incubate the specimens in a mucin-rich solution; *via* their hydrophobic termini, the mucin molecules are then expected to spontaneously adsorb onto the hydrophobic lens surface. However, due to the transient nature of such binding, these passively adherent mucins can be easily removed under shear stress. To prevent such shearing off, covalent bonds are required between the mucins and the lens surface – and this can be achieved by employing a multi-step coupling strategy (see **chapter 2.4.2**).

In a first step, it is investigated if both, passive and covalent coating procedures, are similarly suitable to create a mucin layer on PDMS lenses. For those initial tests, both types of coatings are generated with manually purified MUC5AC and commercially available PGM, respectively. To verify the presence of mucins on the contact lens surfaces, two different detection methods are employed. For manually purified MUC5AC, detecting the fluorescence signal obtained from ELISA measurements is a commonly used approach. Indeed, when compared to uncoated reference samples, this method returns strong signals for both coating strategies (**Figure 3.6a**). This result suggests that a considerable amount of MUC5AC molecules is present on the surface of the coated lenses and that both coating strategies

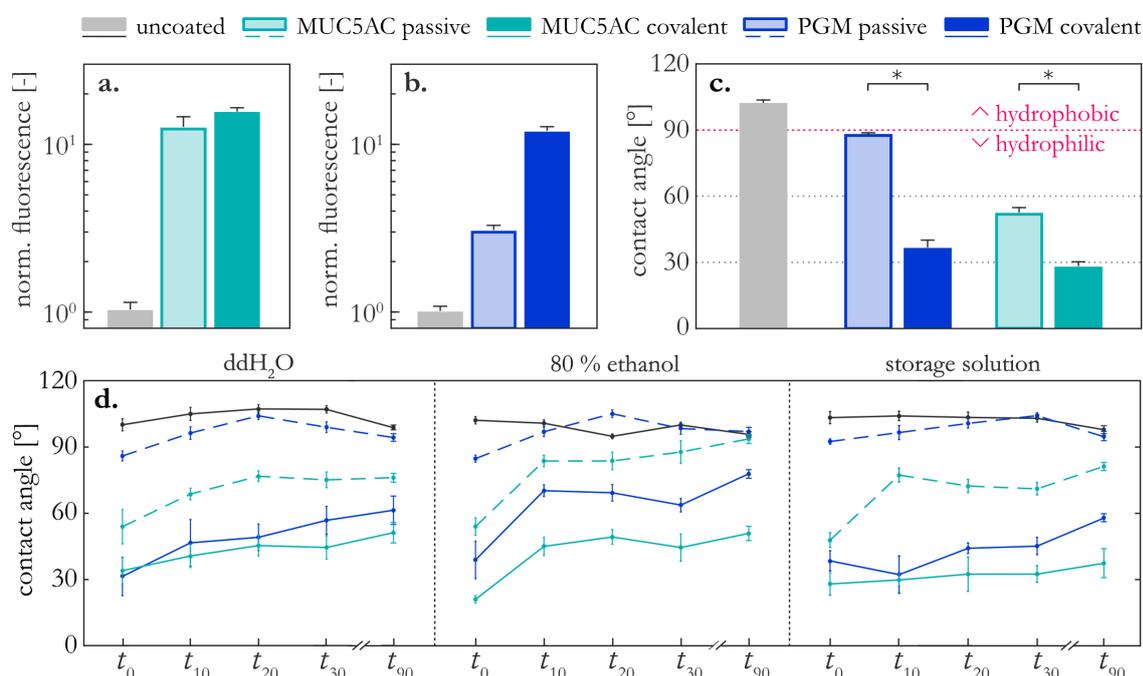
---

\* This section follows in part the publication Rickert *et al.*, *ACS Applied Materials & Interfaces* (2020)

## Mucin macromolecules for biomedical applications

yield similar coating densities. Considering the strong hydrophobic character of the lens material and the inherent ability of natively purified MUC5AC to adsorb very efficiently onto hydrophobic materials, this finding is in line with the expectations. However, this similarity in the ELISA signal makes it difficult to distinguish between lenses that were coated with MUC5AC by passive adsorption or covalent coupling.

Such a differentiation, however, is possible for lenses that were coated with commercially available PGM. Here, a detection assay based on fluorescently labelled mucins has to be employed since commercial PGM is not recognized well by the anti-MUC5AC antibody required for the ELISA. With this alternative detection method, considerably higher values are obtained for the covalently coated lenses compared to the passively coated ones (**Figure 3.6b**). This result agrees with findings reported by Marczynski *et al.*,<sup>83</sup> that demonstrated the inferior adsorption properties of this industrial PGM compared to in-lab-purified MUC5AC. However, the covalent coupling process applied here seems to be sufficient to compensate for this lack of PGM adsorption capabilities.



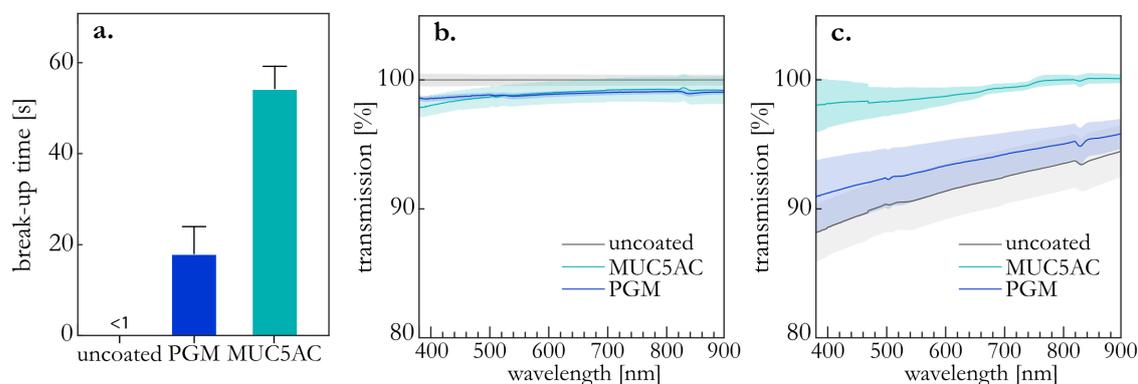
**Figure 3.6: Coating verification and wettability of coated and uncoated samples.** The presence of manually purified MUC5AC (a) and commercial PGM (b) on PDMS lenses was assessed *via* an ELISA and an assay using fluorescence microscopy, respectively. Error bars denote the standard error of the mean as obtained from  $n = 3$  samples. Contact angles as measured to quantify the surface wettability of the surfaces are displayed (c,d) for PDMS disks that were either left untreated or coated with mucins. Contact angles larger than 90° denote hydrophobic behavior, whereas values smaller than 90° represent hydrophilic surfaces. Error bars indicate the standard error of the mean as obtained from  $n = 9$  (c) and  $n = 3$  (d) samples. For the data depicted in c, statistical analysis was performed (based on a  $p$ -value of 0.05). All coated samples are significantly different from the untreated samples. Significant differences between the coating strategies are denoted with an asterisk.

Having verified the presence of mucins on the contact lens surfaces, in a next step, the impact of these coatings on the surface wettability of the lenses is determined by conducting contact angle (CA) measurements: here, CAs above 90° show hydrophobic behavior, whereas CAs smaller than 90° denote hydrophilic surface properties. To increase the accuracy of the measurements, the experiments are conducted with flat PDMS disks which have the same material composition as the PDMS lenses. The CA values obtained for uncoated disks are in the range of 100–110° (**Figure 3.6c**), which demonstrates the strong hydrophobic characteristics and poor wettability of the PDMS-based lens material. This,

however, drastically changes when using passive and covalent coatings of either MUC5AC or PGM: As depicted in **Figure 3.6c**, all coatings entail a decrease of the CA below  $90^\circ$ , which denotes an alteration of the surface characteristics from hydrophobic to hydrophilic. This improvement of the surface wettability can be mainly attributed to the large, glycosylated central region of the mucin molecules: the densely attached glycans carry a large number of anionic sialic acid residues and sulfate groups,<sup>71,72</sup> which renders mucins highly hydrophilic and hygroscopic.<sup>73</sup> Yet, the obtained values show that covalent coatings lead to a stronger reduction of the CA than those coatings generated by passive adsorption.

Next, the stability of the different coatings is evaluated by repeatedly reassessing the CA of coated PDMS disks that were incubated in different storage media (**Figure 3.6d**). As expected, the uncoated disks show a constant hydrophobic behavior at all time points and for all storage conditions, which confirms the good stability of the lens material. For the passively coated disks, however, the obtained CA values progressively increase as a function of storage time; many of the tested samples even return into the hydrophobic regime. In contrast, all covalently coated surfaces remain in the hydrophilic regime, even after 3 months of storage. The best results are obtained with covalently coated samples that were kept in a commercial storage solution provided by the lens manufacturer (DPBS with the antiseptic PHMB). Here, the observed changes in the CA values are very small, and excellent wettability is maintained during the whole test period. On the basis of those findings, only covalently coated PDMS lenses are further tested. Accordingly, for storing the lenses, from now on only the commercial storage solution is used.

So far, it was shown that covalent mucin coatings improve the wettability of PDMS-based contact lenses in an efficient and durable manner. Yet, the CA measurements give a rather local impression of the surface wettability. Hence, in a next step, a more application-oriented characterization of the contact lens wettability is performed by assessing the liquid film breakup time (BUT). This parameter is of high importance for the performance of a contact lens on the cornea, as the lens surface needs to be able to maintain a complete tear film coverage between two blinking events. For uncoated PDMS lenses, a very short BUT smaller than 1 s is found (**Figure 3.7a**): here, the liquid film is immediately disrupted as soon as the samples are removed from the storage solution. Mucin-coated contact lenses, in contrast, perform very well regarding this property: PGM-coated lenses maintain full liquid coverage for  $\sim 18$  s, and MUC5AC-coated lenses reach even larger BUTs in the range of 50 s (**Figure 3.7a**). Both of those values by far exceed the average duration of an inter-blinking period of humans, which is  $\sim 5$  s.<sup>178</sup> Together with the results discussed before, this application-oriented wetting test demonstrates that covalent mucin coatings are very well capable of changing the global surface properties of PDMS lenses from hydrophobic to hydrophilic, which is stable and more than sufficient for an application on ocular tissues.



**Figure 3.7: Break-up times and transparency of PDMS lenses.** Break-up times (a) and transmission intensities of blank and covalently coated lenses before (b) and after (c) incubation in a lipid solution. Error bars (a) and shaded areas (b,c) denote the standard error of the mean as obtained from  $n = 3$  samples.

## Mucin macromolecules for biomedical applications

Another essential property of contact lenses as an optical correction system is their transparency. Hence, it is of utmost importance that the applied coating does not disturb the optical properties of the lenses. Therefore, spectroscopy measurements are performed to quantify the transmission efficiency of light across mucin-coated PDMS lenses. Therefore, the intensities of transmitted light are measured for both, coated and uncoated lenses, and the value obtained for the latter is set to 100 % at each tested wavelength. As depicted in **Figure 3.7b**, excellent transmission efficiencies of >97 % are achieved for both covalent mucin coatings. These results are not only more than sufficient from an application point of view but also clearly outperform similar coatings conducted with the mucin-related glycoprotein lubricin/PRG4.<sup>179</sup> Conclusively, both mucin coatings maintain the transparency of the lenses very well.

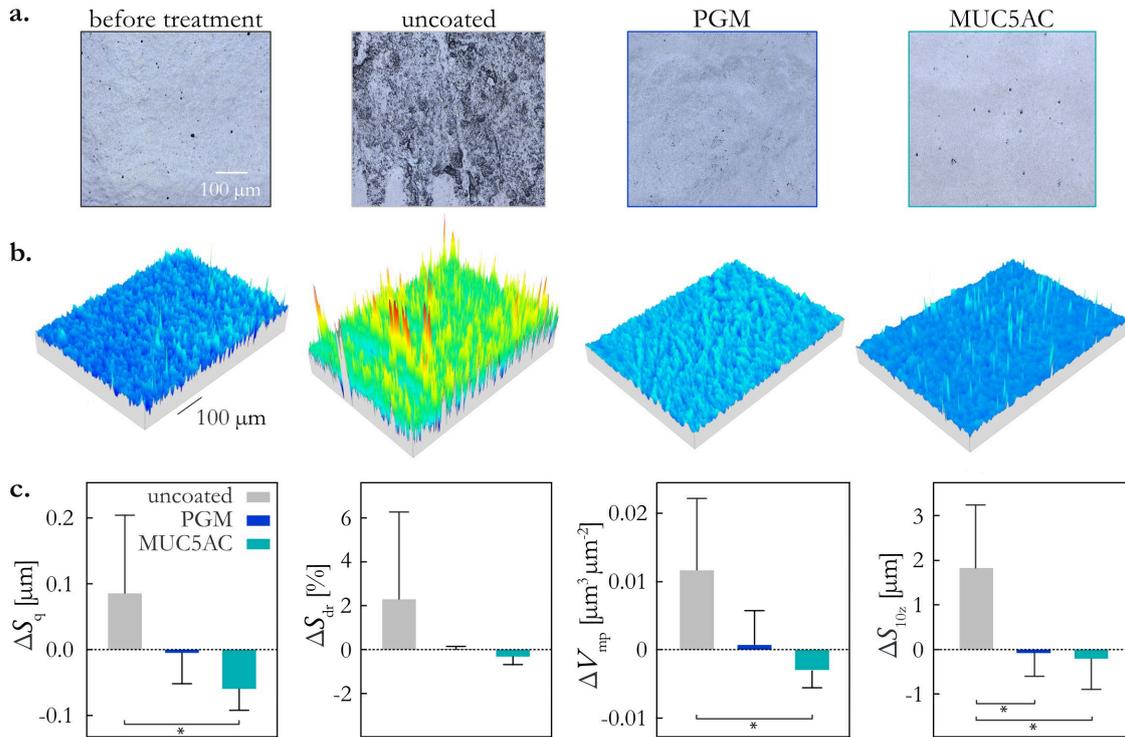
In addition to having a high initial transparency, it is also important that the contact lenses can maintain this transparency during usage. A physiological challenge that might reduce the lens transparency is the adsorption of lipids from the tear film. As mucin coatings have previously been shown to possess anti-biofouling properties,<sup>158</sup> it is next tested if this also applies to the coatings generated on the contact lenses when the lenses are exposed to lipids (**Figure 3.7c**). For uncoated PDMS lenses, a 6-12 % decrease of their transparency is found across all wavelengths of the measured spectrum. This reduction in transparency is similarly high for lenses that were coated with PGM. However, lenses carrying a covalent MUC5AC coating maintain their transparency very well. Hence, those mucin coatings are highly efficient in preventing lipid deposition.

To this point, it was demonstrated that a covalent coating using either PGM or MUC5AC improves the wettability of the contact lens material while maintaining its transparency. A good wettability of contact lenses is important as a break-up of the tear film due to poor wettability has, in addition to friction, been directly linked to discomfort.<sup>180</sup> In fact, a consistent and stable coverage of the corneal tissue by the ocular tear film is essential for preserving the integrity of the corneal tissue: If the epithelium is not efficiently protected by the viscoelastic tear fluid, the corneal surface can be injured by shear forces resulting from blinking or eye movements.<sup>181,182</sup> Thus, the next experiments aim at assessing the performance of the mucin coatings in terms of preventing tissue damage resulting from tribological stress. In detail, the mucin coatings are probed in a lens/cornea pairing and both, the friction response of the system and wear formation on the corneal surface of porcine eyes are determined. As shown in **Appendix A1.10**, similar friction coefficients for both, uncoated and coated PDMS lenses, are obtained. Even if these results differ from those reported for coatings with lubricin,<sup>179</sup> they agree with previous findings reported for passively adsorbed mucin coatings that were studied in a very similar setup.<sup>51</sup>

At this point, it is crucial to realize that ocular pain is not necessarily caused by high friction but rather by damage to the corneal tissue. Such tissue damage can be the result of the high mechanical stress induced by the contact lens materials when sliding across the eye surface. Importantly, several earlier studies on artificial materials,<sup>173,183,184</sup> tissues,<sup>185,186</sup> and cellular monolayers<sup>187</sup> showed that there is no direct correlation between friction and wear formation. Hence, surface damage resulting from tribological stress has to be determined independently. Here, the surface damage induced on the cornea tissue samples is assessed by comparing profilometric images obtained before and after exposure of the tissue to tribological stress. Indeed, when uncoated contact lenses were used, the tribological treatment induces severe alterations to the tissue surface (**Figure 3.8a** and **Appendix A1.10**). In contrast, no visual damage is observable for cornea samples that were treated with mucin-coated lenses, regardless of whether PGM or MUC5AC was used for the coating.

To quantify this visual impression, a set of metrological surface parameters is calculated from the profilometric 3D images (**Figure 3.8b**) and the change in those parameter values as a result of the tribological treatment is calculated (**Figure 3.8c**). To cover the broad morphological range of occurring surface alterations, four different parameters are selected, which all describe roughness features of a

surface but focus on different structural details: First, the root-mean-square height ( $S_q$ ), a classical height distribution parameter; second, the developed interfacial area ratio ( $S_{dr}$ ), a hybrid parameter that relates the total surface area of a region to a perfectly plain surface; third, the peak material volume ( $V_{mp}$ ), which represents the volume of material at an areal material ratio of 10 %; and fourth, the  $S_{10z}$  parameter, which describes the difference between the five highest peaks and the five lowest pits of the surface. Of course, none of those topographical parameters is ideally suited to quantify the broad range of wear morphologies observed on the cornea samples; thus, the error bars obtained from the different image quantification approaches are all relatively high.



**Figure 3.8: Wear formation on porcine cornea samples as quantified by optical profilometry.** Light microscopy (a) and 3D surface images (b) of corneal tissue samples and quantification of the latter by metrological parameters (c) are shown for samples that were tribologically treated with either an uncoated or a covalently mucin (PGM or MUC5AC)-coated contact lens. Error bars denote the standard deviation as obtained from  $n = 5$  measurements. Asterisks indicate statistical significance based on a  $p$ -value of 0.05.

Importantly, with each of these parameters (as well as 6 additional surface parameters) the same result is obtained (Figure 3.8c and Appendix A1.10): both mucin coatings prevent an increase in surface roughness after tribological treatment, and MUC5AC coatings perform slightly better than the PGM coating. In fact, for cornea treatment with mucin-coated lenses, sometimes even a slight reduction of the corneal surface roughness is detected, which can be attributed to mechanical smoothing of the tissue.

In conclusion, the presented results demonstrate that covalent mucin coatings may be an excellent choice to minimize the disruptive influence of contact lenses on the eye, as they are capable of stably generating wettable surfaces on otherwise hydrophobic PDMS lenses. With those mucin coatings, the lenses interact very well with liquid environments, such as the tear film, and do not generate shear-induced tissue damage on the cornea. These results open up great possibilities for a variety of other hydrophobic materials that are, to date, not suitable for a contact lens application.

### 3.3. Covalent mucin coatings on rigid, gas-permeable contact lenses\*

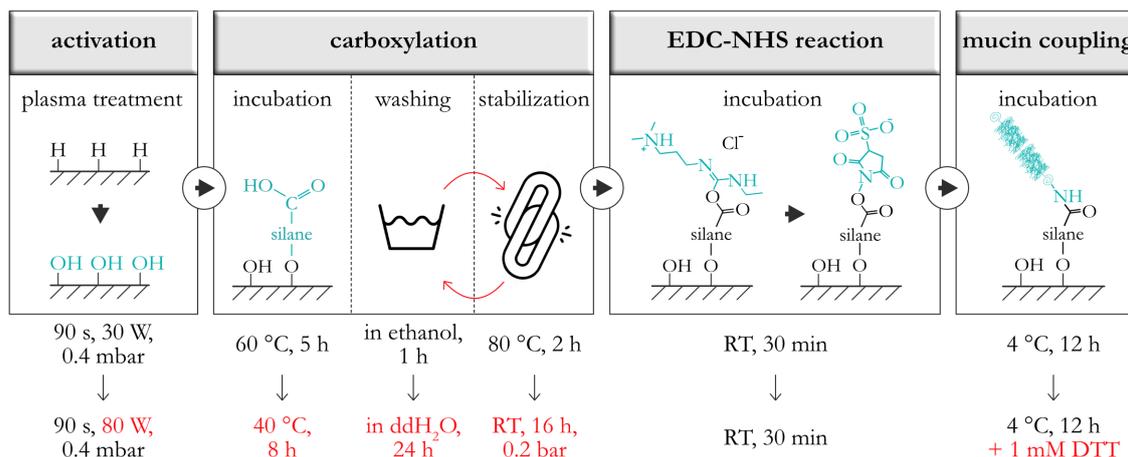
Rigid, gas-permeable lenses (RGP lenses) made from durable plastics with an excellent oxygen permeability are an important sub-category of contact lenses: As the avascular cornea receives its oxygen supply through the tear film, RGP lenses are usually favorable over silicone-hydrogel lenses as they reduce the risk of hypoxic complications such as corneal neovascularization or limbal erythema.<sup>188</sup> Additionally, owing to their good customizability and their inherent stiffness, those lenses are particularly suitable for the treatment of geometrically challenging visual impairments including severe or irregular astigmatism<sup>3,4</sup> as well as pathological eye conditions involving thinning and deforming corneas (as, for instance, common for keratoconus).<sup>189,190</sup> However, the perceived comfort of RGP lenses can be significantly lower than that of soft lenses, and this can – amongst other factors – be attributed to their insufficient surface wettability<sup>16</sup> or wear formation on the corneal tissue.<sup>17</sup> Those limitations not only discourage or even prevent many patients from wearing such rigid contact lenses, but, as discussed before, they can also entail complications such as the development of microbial keratitis.<sup>171</sup>

Hence, similar to what has been demonstrated in **chapter 3.2** for prototype PDMS lenses, also RGP lenses might profit from covalent mucin coatings. However, this particular class of contact lenses comes with several challenges: RGP lenses are geometrically very complex systems with precisely defined optical properties that should not be impaired by the coating itself or by the coating process. Problematically, though, the material tends to deform in organic solvents and shows limited stability when exposed to temperatures above 50 °C – two conditions that are essential factors in the previously described mucin coating strategy. Thus, to apply mucin coatings to such lenses while accounting for the specific requirements of the lens material, the covalent mucin coupling process used for PDMS lenses needs to be adjusted. This is, however, not trivial: Solvents are needed for the washing step to help creating transparent coatings. Moreover, increased temperatures during the incubation and stabilization steps are required for the formation of stable bonds between the applied layers. Hence, to develop a process yielding coatings combining both, excellent transparency and sufficient stability, all individual steps had to be iteratively adjusted and were evaluated regarding those two properties (for details see **Appendix A2**).

As described in the **chapter 2.4.2**, an established procedure to covalently attach mucin macromolecules to polymeric surfaces starts with a surface activation by exposure to plasma, followed by the attachment of a silane layer as a coupling agent to carboxylate the surface, and finally a carbodiimide coupling step to conjugate the mucin molecules to the silane layer. To be applicable to RGP lenses, the adapted process comprises the following major modifications (**Figure 3.9**). First, the plasma intensity is increased to achieve a more efficient surface activation. Second, the necessary reduction of the incubation temperature from 80 °C to RT (to stabilize siloxane bonds) is compensated by generating a vacuum and extending the stabilization time from 2 h to 16 h: Reducing the ambient pressure below the saturation vapor pressure of water (2340 Pa at 20 °C)<sup>191</sup> triggers a condensation reaction that converts weak hydrogen bonds into siloxane bonds<sup>192</sup> without the need of exposing the material to elevated temperature.<sup>193</sup> Third, loosely bound silanes are removed by washing the lenses in ddH<sub>2</sub>O instead of in 96 % (v/v) ethanol; however, this washing step is now conducted for 24 h instead of 1 h to counterbalance the lower detergency. Importantly, washing is conducted after stabilization to avoid removing convertible bonds. Last, the reducing agent DTT, which disrupts disulfide bonds between monomeric mucins,<sup>194</sup> is added to allow for a monomeric mucin layer to be formed (and to avoid the uncontrolled attachment of mucins *via* oligomerization).

---

\* This section follows in part the publication Rickert *et al.*, *Biomaterials Advances* (2022)



**Figure 3.9: Schematic representation of the adjusted process to create mucin coatings on RGP lenses.**

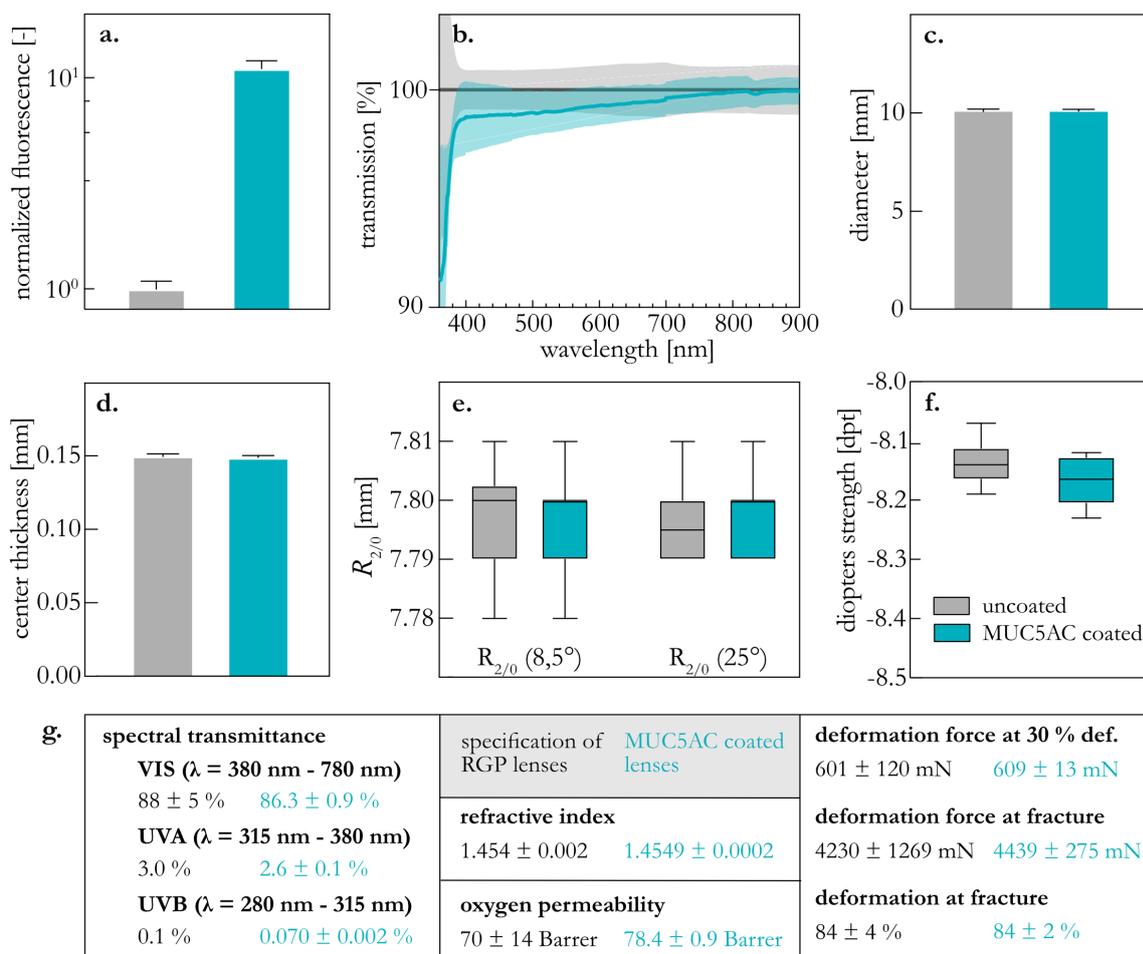
The contact lens surfaces were first activated using plasma generated with ambient air before they were carboxylated with a silane precursor. Subsequently, carbodiimide coupling was employed to covalently attach mucin glycoproteins onto the lens surface. For this process to become suitable for RGP lenses, the following modifications were applied: first, the plasma intensity was increased. Second, the temperatures for both, the silane incubation and silane stabilization step were decreased and the incubation/stabilization times were increased. Third, the order of the silane stabilization step and the silane washing step were switched (so that in the new process, stabilization is conducted before washing), and silane stabilization was conducted in an evacuated environment. Fourth, for washing the lenses, ethanol was substituted with ddH<sub>2</sub>O, and the washing time was strongly increased. Fifth, 1 mM DTT was added to the mucin solution to obtain monomeric mucins for the final coupling step.

Given the fact that, on PDMS lenses, coatings generated with MUC5AC outperformed those generated with PGM in several aspects, for RGP lenses, only covalent MUC5AC coatings are tested. As demonstrated *via* an ELISA (the same antibody-based MUC5AC detection method as used for PDMS lenses), the adapted coating strategy is suitable to attach a considerable amount of mucin macromolecules onto the surfaces of RGP contact lenses (**Figure 3.10a**). Compared to uncoated lenses, the observed increase in fluorescence intensity brought about by specific anti-body binding (and the subsequently enabled enzymatic reaction) is in a similar range as what was previously achieved for the covalent mucin coatings on PDMS lenses (**Figure 3.6a**). In addition to proving the presence of mucin molecules on the surface, it is also of interest to estimate the thickness of the generated layer. Assuming that the developed coupling procedure yields monomeric mucin coatings with a brush-like morphology, the theoretical maximal thickness of the mucin layer generated on the lens can be approximated based on the contour length of mucin monomers: considering the 5800 amino acids<sup>83</sup> of the protein backbone with an average length of 4 Å<sup>195</sup> each, the estimated maximal contour length reaches ~2332 nm. In reality, however, the mucin contains folded domains – at least in the terminal regions of the glycoprotein – and this may lead to somewhat thinner coatings than this estimate suggests. To test this hypothesis, the coating procedure is conducted with fluorescently labelled mucins, the generated mucin layer is visualized by confocal microscopy, and the layer thickness is then calculated based on the obtained fluorescence intensity profiles (see **Appendix A1.11**). With this technique, a fluorescent layer with a thickness of 3 μm is observed. Taking into account the inaccuracies caused by cross-illumination of fluorescent markers from different layers, this result agrees reasonably well with the previous estimate.

Since RGP lenses are complex, geometrically precisely defined optical systems, it is important that their properties are not negatively affected by either the mucin coating itself or by the coating procedure. Hence, the next experiments aim at evaluating the transparency and geometric stability of the coated

## Mucin macromolecules for biomedical applications

RGP lenses, their oxygen permeability and flexural rigidity, as well as their refractive properties (for details, please refer to **Appendix A1.12**). Importantly, with the applied MUC5AC coating, the transparency of the contact lenses is well preserved: for the complete spectrum of visible light (400 – 900 nm), the mucin coating exhibits transmission values above 98 % (normalized to blank lenses, **Figure 3.10b**). The same positive result is obtained when the spectral transmittance of the complete optical system (*i.e.*, the lens together with the applied coating) is determined according to standardized tests (EN ISO 18369-3). For all evaluated sub-spectra (VIS, UV-A, and UV-B), transmittance values are reached that perfectly agree with the required lens specifications (**Figure 3.10g**).



**Figure 3.10: Geometrical and physical properties of MUC5AC coated RGP lenses.** The presence of MUC5AC on the surface of coated RGP lenses was verified *via* an ELISA (**a**,  $n = 3$ ), and normalized transmission values of such mucin coated lenses are compared for a wavelength spectrum of 400 nm to 900 nm (**b**,  $n = 3$ ). Furthermore, geometrical lens parameters such as the lens diameter (**c**), the center thickness (**d**), the back surface radius of the optical zone (**e**), and the diopeters strength (**f**) of blank and coated lenses were quantified (all  $n = 10$ ). Last, the spectral transmittance ( $n = 5$ ), the oxygen permeability ( $n = 4$ ), the refractive index ( $n = 5$ ), and the deformation behavior ( $n = 10$ ) of coated lenses was determined (**g**, cyan) with standardized methods and compared to the lens specifications (**g**, black). Error bars in (**a,c,d**), shaded areas in (**b**), and the error intervals given for mucin coated lenses in (**g**, cyan) denote the standard error of the mean. Deviation intervals given for the lens specifications (**g**, black) denote the acceptable deviation range according to EN ISO 18369-2.

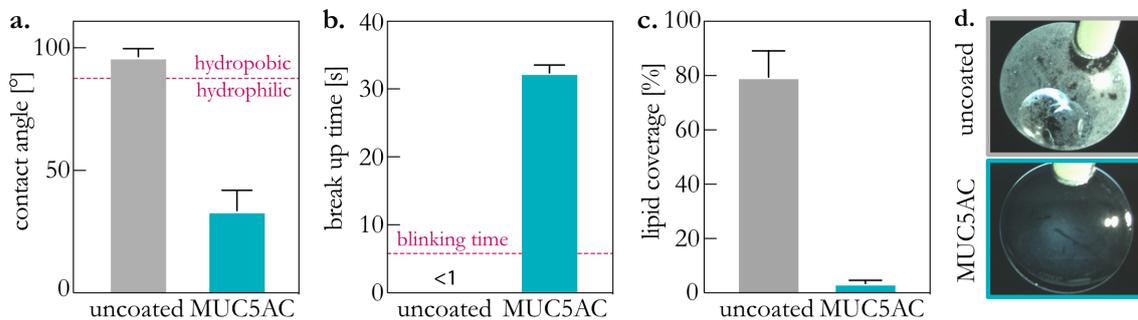
Next, different geometrical parameters of the RGP lenses are evaluated: this includes the lens diameter (**Figure 3.10c**), the center thickness of the lens (**Figure 3.10d**), and the back surface radius of the optical zone (measured at two different angles; **Figure 3.10e**). When assessing those parameters before and

after applying a covalent MUC5AC coating, no significant changes are observed. This verifies that all process parameters selected for the adapted coating procedure were appropriately chosen to prevent lens deformation. When evaluating the optical properties of coated specimens, neither the dioptric strength (**Figure 3.10f**), nor the refractive index (**Figure 3.10g**) of the lenses show considerable alterations. As those two lens characteristics represent the visual correction abilities of the lenses, their good preservation underscores the excellent suitability of the mucin coating for optical systems.

With contact lenses constituting a physical barrier that restricts the transport of oxygen towards the ocular surface, the lens permeability is another key characteristic that needs to be assessed. Sufficient oxygen supply is not only essential for maintaining a normal corneal metabolism (which, if impaired, could entail limbal hyperemia and promote undesired neovascularities on the corneal surface),<sup>196,197</sup> it was also shown to significantly influence bacterial binding to the cornea.<sup>198</sup> When determining the oxygen permeability of the lenses, it is observed that MUC5AC coated RGP lenses perfectly meet the permeability values of the lens specifications (**Figure 3.10g**). In other words, the covalent mucin coating applied on the lens surfaces does not hinder the transport of oxygen across the lens.

Last, the flexural rigidity of the coated lenses is assessed. A high lens flexure can become problematic when the lenses are supposed to neutralize visual impairments arising from a toric cornea – this is, for instance, the case for astigmatism. When placed onto such toric eyes, the contact lens might bend along the flattest curvature, and this in turn can limit its capability to sufficiently neutralize the refractive error.<sup>199,200</sup> To evaluate the lens flexure, three parameters are obtained: the maximal deformation occurring before breakage of the lens, the corresponding applied force, and the force needed for 30 % lens deformation (**Figure 3.10g**). For all three parameters, the RGP lenses coated with MUC5AC precisely meet the desired lens specifications.

Having verified that the mucin coating does not interfere with the geometric and optical properties of the RGP lenses, a set of experiments is conducted to assess the enhancement of the surface wettability brought about by the mucin coating. When determining CAs (contact angles), blank, untreated RGP lenses show contact angles in the range of 90 – 100° (**Figure 3.11a**), which represents hydrophobic surface properties and poor wettability. Again, such insufficient wetting behavior can lead to rupture of the tear film on the lens surface between two blinking events, and this would impede the physiologically important liquid coverage of the corneal tissue. A similar result is obtained, when the more application-oriented BUT (liquid break-up time) is evaluated. For blank, untreated RGP lenses without any coating, a very short BUT (<1 s) is measured (**Figure 3.11b**): The aqueous layer on the lens is disrupted as soon as the lens is removed from the fluid.



**Figure 3.11: Wettability and lipid adsorption resistance of coated and uncoated RGP lenses.** Contact angles (a), break-up times (b), and lipid adsorption results (c, d) are compared for MUC5AC coated and blank, untreated RGP lenses. Error bars denote the standard error of the mean as obtained from  $n = 3$  samples.

## Mucin macromolecules for biomedical applications

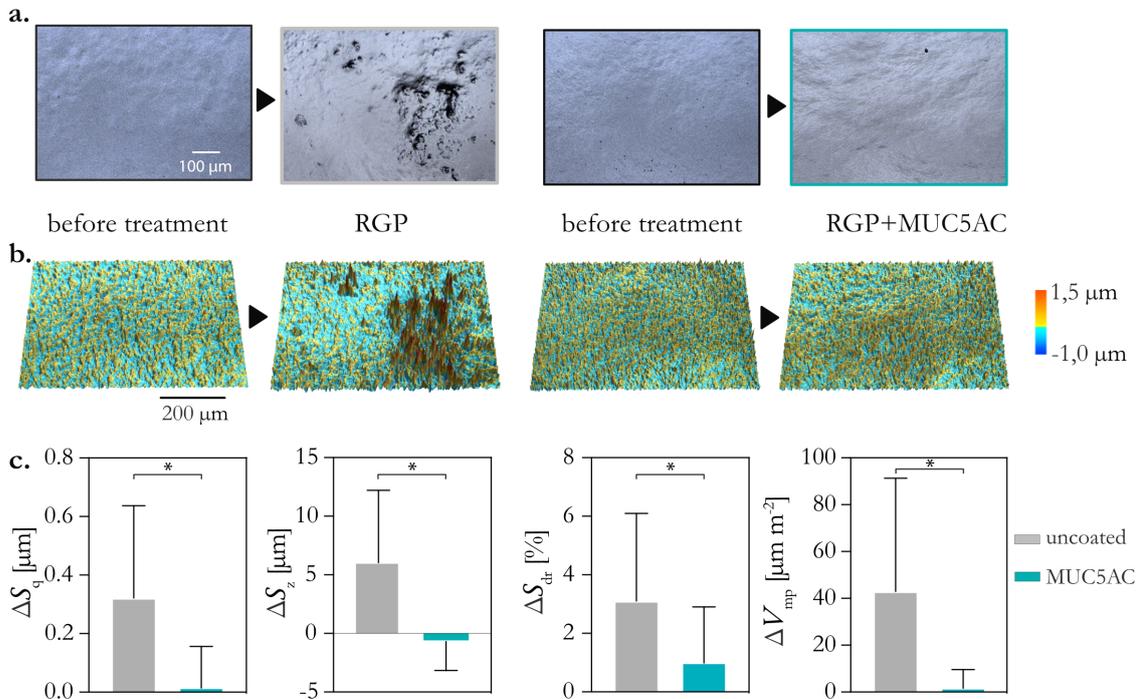
In contrast, for RGP lenses carrying a MUC5AC coating, a strong improvement of the surface wettability is observed with both of those wetting tests. First, as shown in **Figure 3.11a**, the mucin coating leads to a decrease of the CAs to approximately 40°; this strong change demonstrates an alteration of the surface wettability from hydrophobic to clearly hydrophilic. Remarkably, in a previous study, a combination of laser-induced surface microstructuring and plasma activation was necessary to achieve a comparable improvement of the surface wettability of RGP lenses.<sup>201</sup> Furthermore, the MUC5AC-coated RGP lenses are able to fully maintain liquid coverage for approximately 30 s (**Figure 3.11b**). Thus, in marked contrast to the result obtained for blank, untreated lenses, the BUTs of the coated lenses clearly exceed the typical duration of an inter-blinking period of humans (which is approximately 5 s).<sup>178</sup> In conclusion, these two tests nicely show that such covalent mucin coatings are well suited for transforming the wetting behavior of RGP lenses with a similar efficiency as previously shown for PDMS lenses. Starting from an initially hydrophobic surface, with an applied MUC5AC coating, hydrophilic levels are reached that are more than satisfactory for an application on eyes.

As mentioned before, sufficient transparency is a key requirement for contact lenses to fulfill their task as an optical correction system: in addition to possessing high initial transmittance values, it is equally important that RGP lenses maintain their excellent transparency during usage. A physiological challenge that can severely compromise this property is the deposition of lipids from the tear film to the lens surface: When exposing an uncoated RGP lens to a lipid-rich environment, a severe accumulation of lipids is observed on the lens surface, and this reduces the lens transparency by turning it opaque (**Figure 3.11c,d**). However, in agreement with results obtained during lipid deposition tests on PDMS lenses, on RGP lenses coated with MUC5AC, hardly any lipid adsorption is detectable (**Figure 3.11c,d**). Thus, it can be concluded that those mucin coatings are highly efficient in preventing lipid deposition on RGP contact lenses.

So far it was shown that mucin coatings on RGP lenses can drastically improve the lens wettability while maintaining critical lens characteristics such as the refractive performance. Again, a good lens wettability is key for maintaining the integrity of the corneal tissue: sufficient lubrication by the aqueous tear film is required to adequately protect the epithelium of the eye against shear stresses occurring due to eye movement and blinking.<sup>180-182,202</sup> This realization, combined with the tribological evaluation of coated PDMS lenses, leads to the expectation that the demonstrated major change in surface wettability of RGP lenses should also entail an improved tribological performance. To test this hypothesis, the previously used ocular tribology setup is employed to establish a friction pairing between a porcine cornea and an RGP lens, and possible damages inflicted to the cornea are analyzed (additionally, the friction response of the system was measured; but, as expected, the friction coefficients are again not significantly influenced by the mucin coating; see **Appendix A1.13**). And indeed, strong differences are detected between the surface texture of cornea samples that were tribologically treated with uncoated and coated RGP lenses, respectively: images captured from corneal tissue that was treated with a blank, untreated RGP lens show severe wear formation, whereas no obvious surface alterations are visible after similar treatments with MUC5AC-coated lenses (**Figure 3.12a** and **Appendix A1.13**).

To again quantify this subjective assessment, the relative change (before to after tribological treatment) of four different standardized metrological parameters is calculated from the topographic surface profiles (**Figure 3.12b,c**). As they were found to be meaningful before, the same parameters as presented in **chapter 3.2** were used with one exception: owing to a change in the analysis software, the ten-point-height  $S_{10z}$  had to be replaced by the individual maximum height  $S_z$  (for details regarding the surface parameters, please refer to **Appendix A3**). Importantly, all those four parameters (and 6 additional parameters; see **Appendix A2.13**) confirm the subjective evaluation of the profilometric images: The tribological treatment with uncoated RGP lenses entails a considerable increase in surface

roughness, which can be significantly reduced (or even prevented) with the applied MUC5AC coatings. Hence, it can be concluded that the observed increase in wettability brought about by the covalent MUC5AC coating on the RGP lens enables protection mechanisms that successfully prevent wear formation on the corneal tissue. This demonstrates that mucin coatings can significantly contribute to preserving ocular health during the usage of contact lenses. Preventing tissue damage not only leads to less perceived discomfort, but helps maintaining a properly functioning immune barrier – the endogenous protection strategy that is essential to avoid infections.



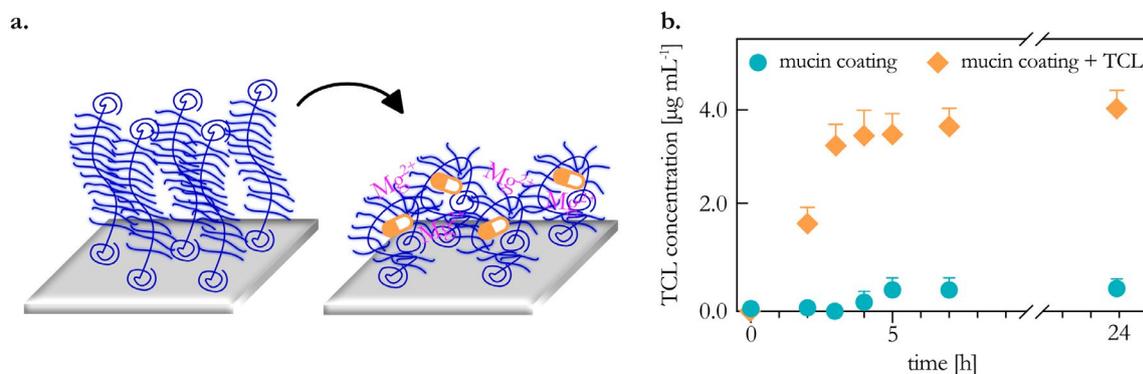
**Figure 3.12: Contact-lens induced wear formation on porcine cornea specimens as visualized by optical profilometry.** Light microscopy (a) and 3D surface images (b) of corneal tissue samples are shown for samples that were tribologically treated with either an uncoated or a covalently MUC5AC-coated RGP lens. A quantification of topographical images (similar to the exemplary ones shown in (b)) is depicted in (c). Error bars denote the standard deviation as obtained from  $n = 6$  independent measurements. Asterisks indicate statistical significance based on a  $p$ -value of 0.05.

Overall, these positive results agree with the findings reported in **chapter 3.2** for silicone-based prototype lenses. However, given the strong differences in material properties between silicone lenses and RGP lenses, this outcome is not trivial. Compared to silicone-based lenses, the hard, inflexible RGP lenses constitute a much more disruptive foreign body for the human eye. Additionally, as the RGP material is more sensitive towards elevated temperatures and solvents than silicone, the coating procedure had to be adjusted in several aspects. It was previously shown that even small changes in the chemistry of a material (for instance, when dealing with different polyurethane variants) can require major adjustments in the coupling procedure conducted – and this already starts at the plasma activation step;<sup>203</sup> hence, tailoring the coating process (that worked well for soft, silicone-based materials) such that it became applicable for the RGP material was key to leverage the potential such glycoprotein coatings hold for a broader variety of materials.

So far, it was shown that MUC5AC coatings neither compromise the optical and geometrical characteristics of RGP lenses, nor do they reduce their oxygen permeability. Instead, such coatings improve the surface properties of the lenses in terms of wettability, lipid repellency, and tribological

## Mucin macromolecules for biomedical applications

performance on corneal tissue. In addition to providing lubricity and anti-biofouling properties, another beneficial aspect of mucin coatings is their ability to undergo a conformational change from a brush-like, elongated state to a condensed conformation. This transition can be utilized to allow mucin coatings to be used as drug depots: when condensing the mucin layer in the presence of a drug, this drug is trapped but can be released again once the conformational change is reverted (**Figure 3.13a**).<sup>204</sup>



**Figure 3.13: Drug loading into and release from mucin coatings on RGP lenses.** (a) Schematic illustration of the drug encapsulation mechanism. (b) TCL release profile from MUC5AC-coated RGP lenses; data obtained for coated but unloaded (= drug-free) RGP lenses is shown for comparison. Error bars denote the standard error of the mean obtained from  $n \geq 3$  samples; if no error bars are visible, they are smaller than the symbol size.

And indeed, after attaching a covalent MUC5AC layer onto the surface of an RGP lens and encapsulating the broad-spectrum antibiotic tetracycline hydrochloride (TCL), a continuous drug release from this functionalized lens is observed over a time period of 24 h (**Figure 3.13b**; for details regarding the method, please refer to **Appendix A1.14**). This result indicates that, in addition to enhancing the surface properties of rigid contact lenses, a covalent mucin coating also offers the possibility to provide a depot for active pharmaceutical ingredients, which makes it possible to tailor such contact lenses for therapeutic properties.

In conclusion, the results obtained for both, prototype PDMS contact lenses and RGP lenses, demonstrate that mucin coatings can significantly contribute to preserving ocular health during the usage of contact lenses. In combination with the previously demonstrated ability of mucin coatings to reduce bacterial adhesion,<sup>158,205</sup> this suggests that such mucin coatings can be a highly beneficial agent to reduce the risk of sight-threatening infections, such as microbial keratitis. Additionally, the coupling process demonstrated in this study can be adapted to generate mucin coatings on a plethora of biomedical materials to entail a durable hydrophilization and to prevent wear formation on underlying tissue.

### 3.4. Sterilization of mucin macromolecules\*

As shown in the previous chapters, mucin macromolecules are a promising material for various biomedical applications. However, owing to their high molecular weight (human gastric mucin MUC5AC, for instance, comprises more than 5600 amino acids)<sup>206</sup> and the high complexity of their brush-like glycosylation pattern (which combines O-, N-, and C-glycosylation),<sup>207-209</sup> it is to date not possible to synthetically produce mucins. Hence, the only feasible way to obtain mucin molecules in substantial quantities is to purify them from animal sources. However, the harvested stomach mucus contains different microorganisms that can, to a certain extent, still be present in the purified mucin. This is, of course, in conflict with the requirements medical products must fulfill. Consequently, before using purified mucin molecules in real medical applications, they need to undergo a sterilization procedure.

However, each of the commonly used sterilization (or disinfection) methods, *e.g.*, UV or  $\gamma$  irradiation, fumigation with ethylene oxide, or autoclaving, may have a negative impact on the physico-chemical or structural properties of mucins. Thermal stress, for instance, often results in the perturbation of intramolecular forces stabilizing a protein conformation, such as hydrogen bonds or van der Waals forces;<sup>210</sup> very often, this results in denaturation, *i.e.*, changes in the secondary and tertiary structures of proteins. Similarly, also UV and  $\gamma$  irradiation might induce oxidation or cleavage of covalent bonds located in the protein backbone or in aromatic amino acid side chains.<sup>211,212</sup> Moreover, treatments with ethylene oxide are suspected to modify methionine and cysteine residues in proteins, and this, in turn, can affect their stability and agglomeration propensity.<sup>213,214</sup> To which extent such issues may impair the integrity of mucin glycoproteins has previously not been addressed. Hence, in the following, MUC5AC macromolecules are subjected to different decontamination strategies, and several methods are employed to test both, the structural integrity and the functionality of the sterilized molecules. All those treatments, namely thermal treatments, autoclaving and irradiation with either UV or  $\gamma$  rays, are conducted with both, mucin in its lyophilized form and mucin that is dissolved in buffer. For both of these sample types, fumigation with ethylene oxide is not practical: Whereas the penetration depth by such gassing is insufficient for the sterilization of liquids, *i.e.*, dissolved mucins, the porous structure of lyophilized mucin samples (which macroscopically have a cotton candy-like appearance) renders them too susceptible to the deposition of toxic residues.<sup>215</sup> Hence, this sterilization method was not applied to mucin solutions or powders.

A sensitive method to evaluate the functional intactness of mucin molecules is to assess the lubricity of mucin solutions and to compare the behavior of treated mucins to those of untreated mucins. Previous experiments have demonstrated that the lubricating potential of mucin solutions sensitively depends on the molecular integrity of the mucin glycoprotein: When using a steel-on-PDMS material pairing for friction measurements, the loss of the hydrophobic termini of the macromolecule,<sup>95</sup> severe damage to its glycosylated central region,<sup>73</sup> and even minor modifications, such as the removal of a single type of anionic residue (such as sulfate or sialic acid groups),<sup>72</sup> each leads to a significant and easily detectable loss of lubricity.

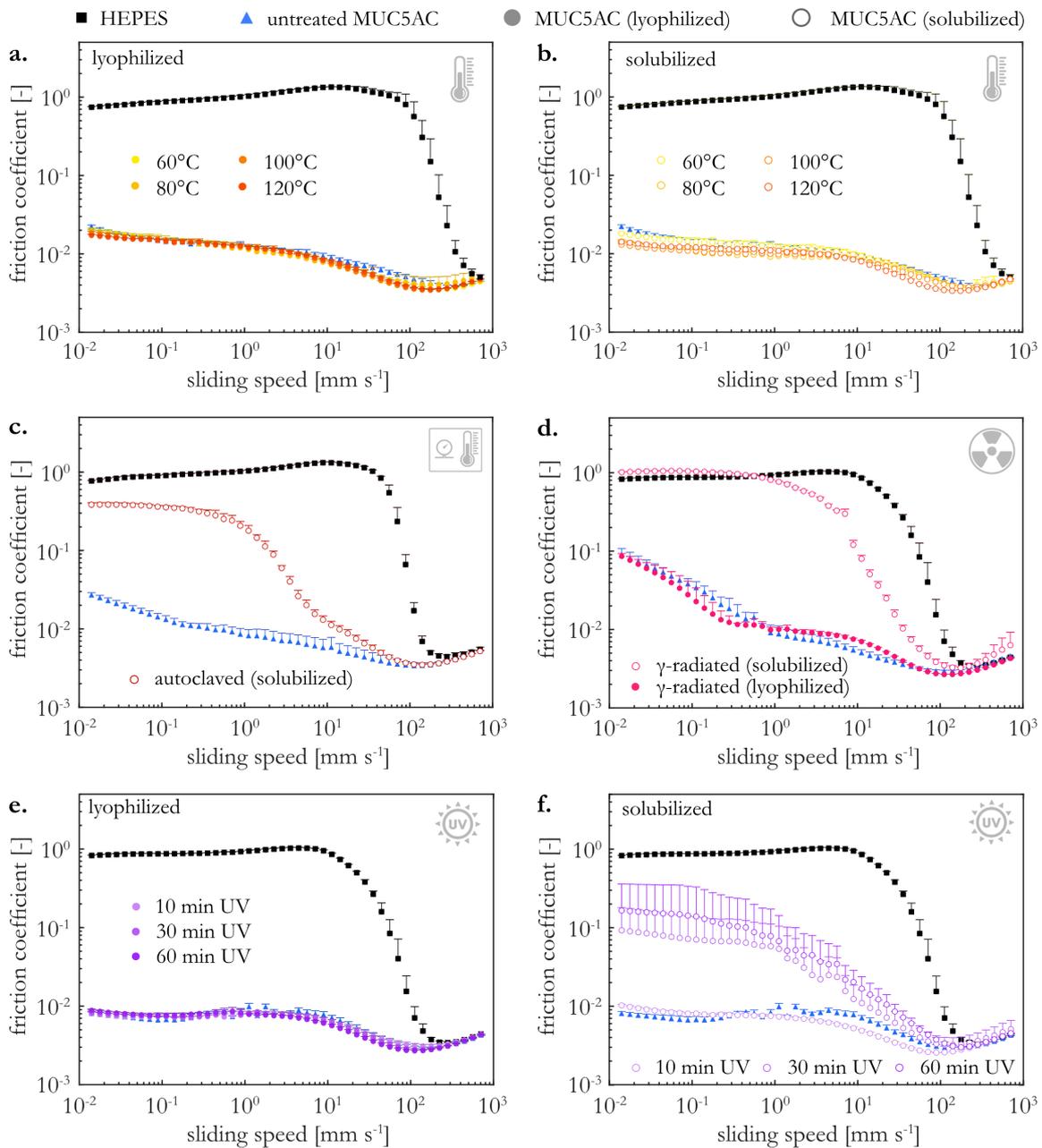
When gastric mucin MUC5AC is subjected to elevated temperatures, *i.e.*, *via* incubation of either mucin solutions or lyophilized mucin powder at 60 °C, 80 °C, 100 °C, or 120 °C for 1 h each, the lubricity of those thermally treated mucins is not compromised in any case (**Figure 3.14a,b**). All mucin samples show excellent lubricity, and the friction curves obtained with thermally treated mucins are virtually identical to those obtained with untreated mucins. This suggests that these thermal treatments do not

---

\* This section follows in part the publication Rickert *et al.*, *Macromolecular Bioscience* (2020)

## Mucin macromolecules for biomedical applications

induce any major damage in the mucin structure. Although such a behavior is not typical for proteins (which often tend to denature when exposed to temperatures above 40 °C), the high glycosylation density of mucins and low content of folded peptide sequences may be responsible for the unusual high sturdiness of this glycoprotein towards heat.



**Figure 3.14: Lubricating behavior of differently treated mucin samples.** Stribeck's curves obtained for solutions of mucins that have been exposed to heating/autoclaving, UV, or  $\gamma$  irradiation treatment, respectively. For data shown in (a,d,e), the treatment procedure was conducted with lyophilized mucin powder. For data shown in (b,c,d,f), the treatment was conducted with a mucin solution. All curves were measured in a rotational steel-on-PDMS pairing. Error bars denote the standard error of the mean as obtained from  $n = 3$  samples.

In contrast, a significant loss of functionality is observed when MUC5AC is thermally treated in an autoclave (Figure 3.14c). Here, after completion of a standard treatment cycle, the lyophilized mucin sample is not even soluble anymore. Such a loss of solubility is a hallmark for a protein denaturation

event<sup>216,217</sup> and renders those lyophilized mucins fully unusable for any further testing. However, the mucin samples treated in solubilized form remain a homogeneous liquid that still provides a reasonable level of lubricity – at least at medium and high sliding speeds, *i.e.*, in the mixed and hydrodynamic lubrication regime. At low sliding speeds, though, where boundary lubrication dominates, the lubricity of these autoclaved mucin solutions is clearly compromised; here, the measured friction coefficient is about one order of magnitude higher than what is obtained for solutions reconstituted from untreated MUC5AC.

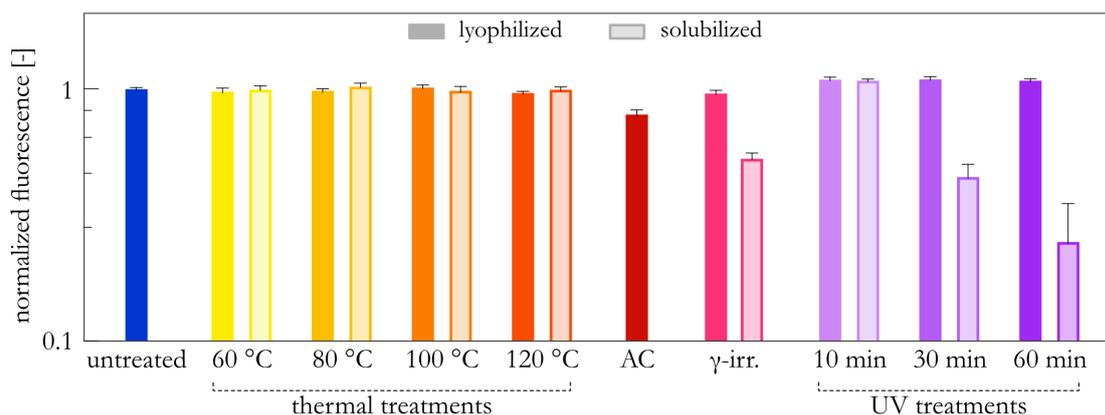
An even more pronounced decrease in lubricity is observed for MUC5AC that was  $\gamma$  irradiated in its solubilized form (**Figure 3.14d**). Here, the obtained friction curve is quite similar to that obtained with simple buffer (devoid of any mucins), which suggests that the molecular structure of mucins was severely compromised by the  $\gamma$  rays. The finding is supported by the fact that the solution showed discoloration after  $\gamma$  irradiation. This outcome, however, is in marked contrast to the results obtained with  $\gamma$ -treated mucin powder. Here, the lubricity of the reconstituted mucin solution is very similar to the result obtained with untreated mucins. This suggests that  $\gamma$  irradiation – although harmful for solubilized mucins – seems to leave the lyophilized MUC5AC macromolecule intact. A similar picture emerges for MUC5AC samples that were subjected to UV irradiation. When they were treated in their lyophilized state, those mucins still provide excellent lubricity independent of the treatment time (**Figure 3.14e**); in contrast, solubilized mucins (**Figure 3.14f**) appear to be more vulnerable to UV treatment. Here, a 30 min treatment with UV light leads to compromised lubricity, whereas shorter treatment times do not.

From this first set of experiments, it can be concluded that autoclaving is not a suitable treatment procedure for the sterilization of mucins. Moreover, in their lyophilized state, MUC5AC molecules appear to resist thermal and radiation/UV-treatment more efficiently than when they are treated as solutions.

As the tribology experiments showed impaired functionality for some of the treated mucin samples, further experiments are conducted to assess if any of the treatments induced structural damages to the mucin. First, an antibody-based detection method (ELISA) is employed that specifically targets the non-glycosylated, hydrophobic termini of the MUC5AC glycoproteins.<sup>218</sup> At this point, it is important to recall that most of the mucin structure comprises unfolded, heavily glycosylated regions. As such glycosylation patterns were previously observed to protect the protein backbone from proteolytic degradation,<sup>95</sup> it is mostly the hydrophobic termini of the mucin glycoprotein that could be vulnerable to denaturation. Since those termini are critically involved in the surface adsorption of mucins (which is mandatory for mucin lubricity), their integrity is crucial for many mucin functions. Owing to the amphiphilic character of the mucin molecule, it can adsorb onto both, hydrophobic and hydrophilic surfaces. As the used well plate surfaces are hydrophilic, the MUC5AC is expected to adsorb *via* the densely glycosylated central region. In this case, the C-termini of the mucins can be expected to be freely accessible for antibodies targeting this recognition site. Structural damages within those terminal regions of the mucin molecules will be visualized by decreased fluorescence intensities when compared to intact protein samples: insufficient integrity or accessibility of the C-terminus of the mucin molecule reduces the efficiency of the antibody binding reaction; hence, such molecules will not be recognized as well by the assay as untreated mucins.

Importantly, for all MUC5AC samples (with the exception of autoclaved mucin, of course, which was not soluble anymore) that were treated in their lyophilized form, no decreased ELISA signal compared to the untreated control sample is observed (**Figure 3.15**). This confirms the results from the previous tribology tests, which did not indicate any perceivable loss of functionality for those mucin molecules.

## Mucin macromolecules for biomedical applications



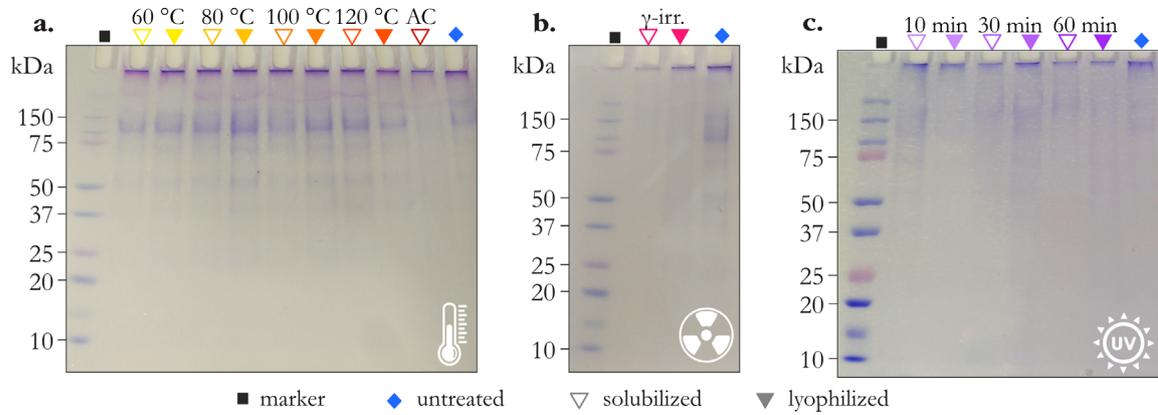
**Figure 3.15: Detection of adsorbed mucins *via* specific antibodies.** Normalized fluorescence intensities were obtained for mucin solutions that have been exposed to heating/autoclaving,  $\gamma$ , or UV irradiation treatments, respectively. The intensity of an untreated sample was used as a reference and set to 1, all other values were normalized accordingly. Signals obtained for mucins that were treated in their lyophilized state (full bars) are compared to results obtained for mucin samples that were treated as solutions (shaded bars). The error bars denote the standard error of the mean as obtained from  $n = 3$  samples.

For solubilized samples that were subjected to either a thermal treatment up to 120 °C or to a short, 10 min UV-exposure, the ELISA returns virtually identical intensity values as for the untreated control; again, this is consistent with the results obtained from tribology. In contrast, for the solubilized mucins that underwent an autoclaving procedure, longer UV-exposure (*i.e.*, for 30 min or 60 min) or a  $\gamma$  irradiation treatment, a decreased ELISA signal is detected. Also, this finding is in agreement with the results obtained from tribology and supports the notion that, for those treated mucin variants, the integrity of the hydrophobic peptide termini is compromised.

Of course, the termini of the mucin glycoprotein cannot only be damaged in terms of their conformation (*i.e.*, folding pattern) but they could also be cleaved from the mucin by hydrolysis. In this case, mucin fragments should be detectable in the treated samples. Testing for a fragmentation of mucins is important as other biopolymer fragments, *e.g.*, those from the extracellular matrix component laminin, have been shown to be cytotoxic even though the full biopolymer is not.<sup>219</sup> Thus, in a next step, the sterilized MUC5AC samples are subjected to a gel electrophoresis under denaturing conditions and a Coomassie staining is conducted (for method details see **Appendix A1.15**); this procedure allows for separating the (protein) components of the mucin samples according to their molecular weight and visualizing the different subpopulations. If mucin fragmentation occurs, it should be possible to detect additional protein bands in the band pattern, which are not present in the pattern of an untreated reference sample. However, no such additional bands are detected for any sample (**Figure 3.16**). Given that the antibody, which targets the mucin C-terminus, returns a similarly strong signal for each of the mucin samples that have been treated in the lyophilized state in the ELISA, this suggests that none of these treatments induced fragmentation of the lyophilized mucin.

It has been repeatedly reported that, owing to its high molecular weight, lab-purified MUC5AC hardly enters the matrix of a polyacrylamide gel,<sup>72,220-222</sup> and indeed, the majority of samples can be found in the pockets of the gel (**Figure 3.16**). These pronounced bands in the loading pockets of each channel indicate large amounts of high-molecular weight MUC5AC molecules. However, for mucin that was  $\gamma$  irradiated in its solubilized form, the intensity of this band is reduced. Since this mucin also failed to provide lubrication, this confirms the ELISA result which already indicated that this mucin variant was somehow damaged by this treatment. In addition, for almost all samples, weaker bands in the range of

10 kDa – 250 kDa are observed; there are only few exceptions where those additional bands appear to be absent, *i.e.*, when mucin was autoclaved or subjected to  $\gamma$  irradiation – either in its solubilized or lyophilized form. This suggests that, here, smaller protein impurities within the samples might have been broken down by thermal hydrolysis into such minuscule fragments that they are not detectable anymore.



**Figure 3.16: Molecular mass distribution of proteinous sample components as visualized by an SDS-PAGE.** Electrophoresis gels are displayed for thermally treated (a),  $\gamma$  irradiated (b), and UV treated (c) mucins. The blue signal originates from a Coomassie staining, which shows the presence of different polypeptides.

So far, the conducted experiments have shown that – in their lyophilized form – MUC5AC glycoproteins are able to withstand various disinfection or sterilization treatments without suffering significant lubrication impairment or structural degradation. The obtained results demonstrate a remarkable sturdiness of those mucins, which opens up broad possibilities for them to be further processed, *e.g.*, as a coating for medical devices.

### 3.5. Sterilization of covalent mucin coatings on medical devices\*

From an application point of view, instead of sterilizing the mucins first and then processing them under sterile conditions, it might be more practical to sterilize the final mucin-containing product as a whole. However, the robustness of mucins in a surface-bound state, *i.e.*, to which extent covalent mucin coatings are able to maintain their functionality after decontamination, needs to be investigated in more detail. For this purpose, covalent MUC5AC coatings are generated on samples of three medical devices: urinary catheters (made from PU), endotracheal tubes (made from PVC), and contact lenses (made from PDMS). Those coated samples are then subjected to four different decontamination strategies – namely autoclavation, ethylene oxide fumigation,  $\gamma$  irradiation, and UV irradiation – and the integrity and functionality of the mucin coatings after those procedures is evaluated. Thermal treatments with dry heat were omitted from the following experiments, as they are not commonly applied for sterilizing medical devices made from polymeric materials. Instead, ethylene oxide gassing is now included, as it is a widely used and efficient decontamination strategy for the surfaces of medical products.

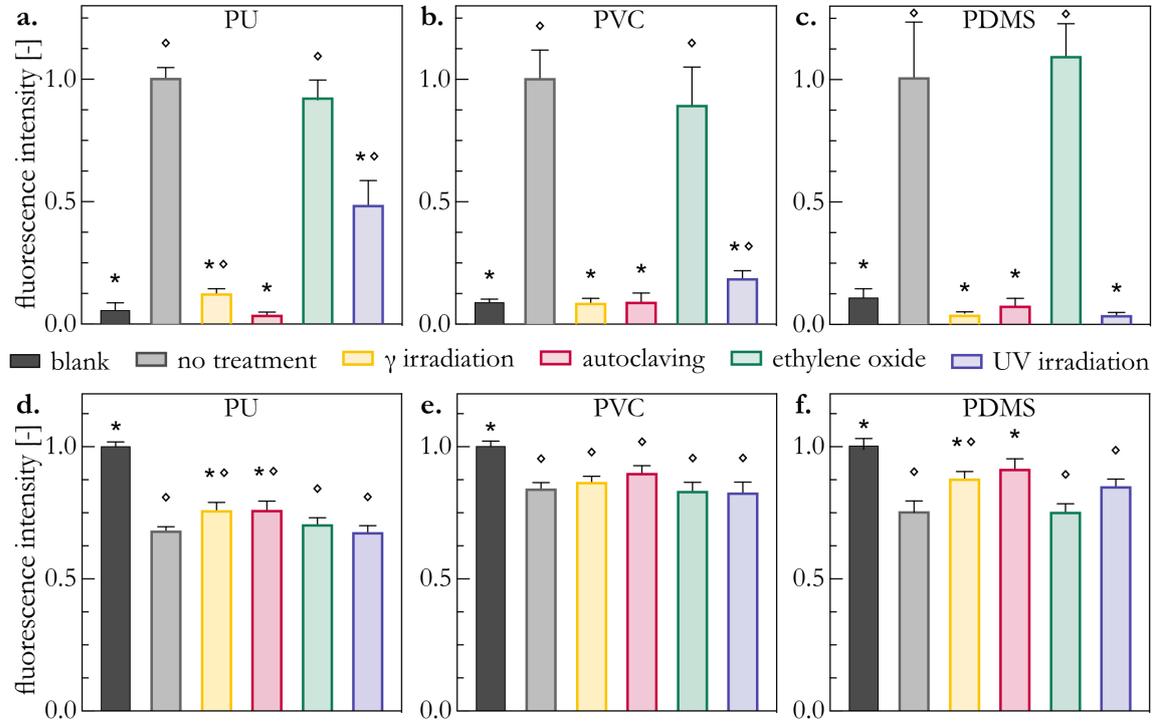
In a first set of experiments, the integrity of those covalent mucin coatings after exposure to different sterilization procedures is assessed with the ELISA that specifically targets the non-glycosylated, hydrophobic termini of the MUC5AC glycoproteins (**Figure 3.17a-c**). For all three tested materials, the fluorescence signals obtained after sterilization with either  $\gamma$  irradiation or autoclavation are significantly lower than those obtained for untreated mucin coatings. In fact, the measured values are comparable to those obtained for a blank sample. This indicates that these two sterilization approaches induce severe damage to or even full cleavage of the hydrophobic MUC5AC termini. The same picture arises for UV irradiated mucin coatings generated on PDMS or PVC. On PU, in contrast, somewhat higher values are obtained for UV irradiated samples; yet, also here, those values are still considerably lower than those obtained for untreated coatings. In marked contrast to those observations, samples that were subjected to ethylene oxide fumigation return fluorescent signals that are similarly high as those determined for untreated coatings – and this assessment applies to coatings generated on any of the three materials. From these tests, it can be concluded that the integrity of the hydrophobic termini of the surface-attached mucin molecules is impaired by autoclavation and irradiation with either  $\gamma$  or UV rays, respectively. In contrast, sterilization with ethylene oxide maintains the integrity of the terminal polypeptide chains. Given the high sturdiness of lyophilized mucins towards irradiation-based treatments (as demonstrated in **chapter 3.4**), such a pronounced irradiation-induced impairment of the integrity of the hydrophobic termini was not fully expected. However, it underscores the necessity to independently investigate the stability of covalent mucin coatings, since the stress generated by the different treatments appears to be different for mucins in a surface-bound state than for mucins in a bulk material.

So far, only the integrity of the non-glycosylated, hydrophobic terminal regions of mucins was tested; however, the glycosylated core region of the MUC5AC constitutes the largest part of the macromolecule and plays a key role for many of the molecule's important properties. Thus, in a next step, the presence and accessibility of this glycosylated part of surface-attached mucins is probed by employing a lectin binding assay that specifically detects a structural motif from the glycosylation pattern of the mucin glycoprotein (**Figure 3.17d-f**). Here, the coatings are incubated with a solution of fluorescent lectins, and the lectin solution is analyzed after this incubation step. Thus, low fluorescence intensity values represent a strong depletion of the lectin molecules and this, in turn, indicates the presence of a high number of glycosylated groups on the mucin coatings. Consistently, for almost all coatings, the obtained values are significantly lower than those obtained for blank, uncoated samples; only for autoclaved mucin

---

\* This section follows in part the publication Rickert *et al.*, *Advanced Materials Interfaces* (2021)

coatings generated on PDMS, the measured difference is not significant. Importantly, for all coatings that were sterilized by either ethylene oxide exposure or UV irradiation, the lectin depletion induced by the coatings is equally high as for untreated reference coatings. This is a good indication that, for those particular samples, the density of glycan groups (and thus the glycosylated area in general) in the treated coatings is not compromised by the sterilization treatment.

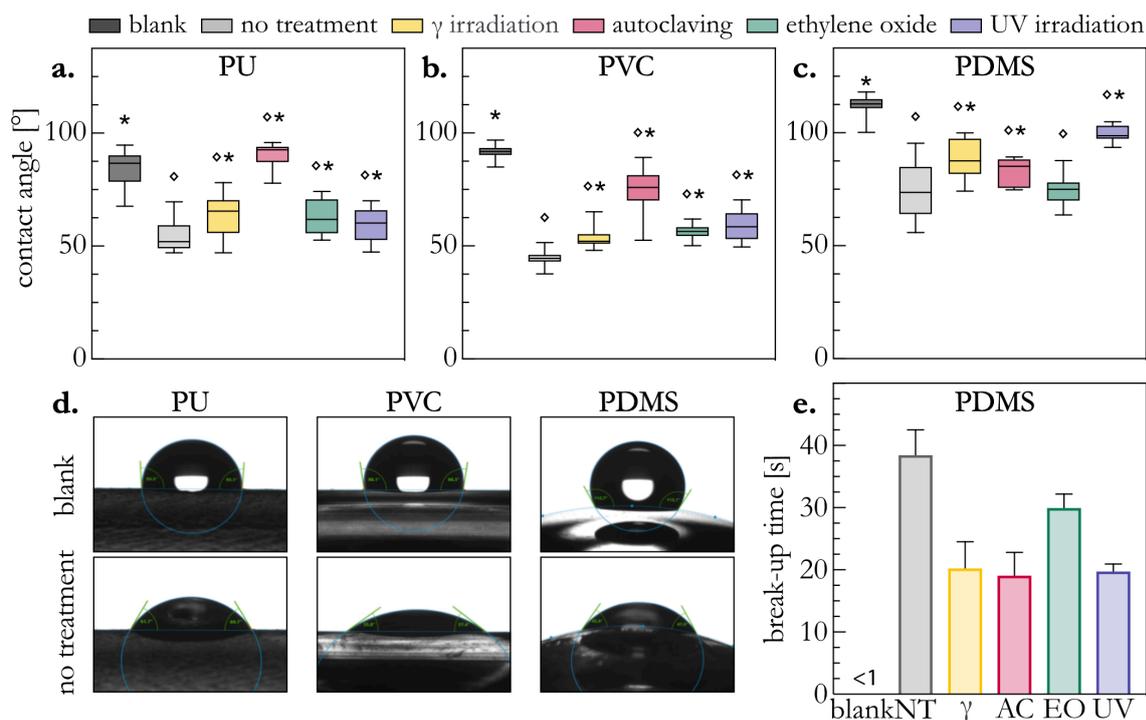


**Figure 3.17: Detection of surface-bound MUC5AC molecules *via* an ELISA and lectin-binding.** The normalized fluorescence intensities obtained with an ELISA (a–c) and a lectin-based depletion assay (d–f) are shown for different medical devices coated with mucins. The coated samples were either stored without any further treatment, or sterilized *via*  $\gamma$  irradiation, autoclaving, ethylene oxide fumigation, or UV irradiation. The error bars denote the standard error of the mean as obtained from  $n \geq 4$  samples. Asterisks and rhombi denote statistically significant differences between a treated sample group and the untreated references or the blank sample, respectively (based on a  $p$ -value of 0.05).

Together, the two assays show that the glycosylated regions of the mucin coatings seem to be more resistant towards the applied sterilization methods than the non-glycosylated terminal regions. This agrees with expectations based on previous studies, where the glycosylation pattern was already observed to protect the protein backbone from proteolytic degradation.<sup>95</sup> The terminal, non-glycosylated parts of the polypeptide backbone, in turn, are more vulnerable. Moreover, the results discussed so far suggest that both parts of the MUC5AC glycoprotein seem to survive a treatment with ethylene oxide gas very well. Here, with either assay, no significant differences were observed compared to untreated coatings.

Having probed the structural integrity of the covalent mucin coatings after subjecting them to the different sterilization techniques, the next goal is to test if selected functions of the coatings are compromised by the different treatments. More specifically, the wettability of the different samples is compared (Figure 3.18a–c), which plays a key role for the anti-biofouling and friction-reducing effects of mucin coatings.<sup>156,158</sup> To quantify the wetting behavior of the different samples, contact angles are measured. Exemplary images of droplets placed onto the different materials with and without covalent mucin coatings, respectively, are depicted in Figure 3.18.

## Mucin macromolecules for biomedical applications



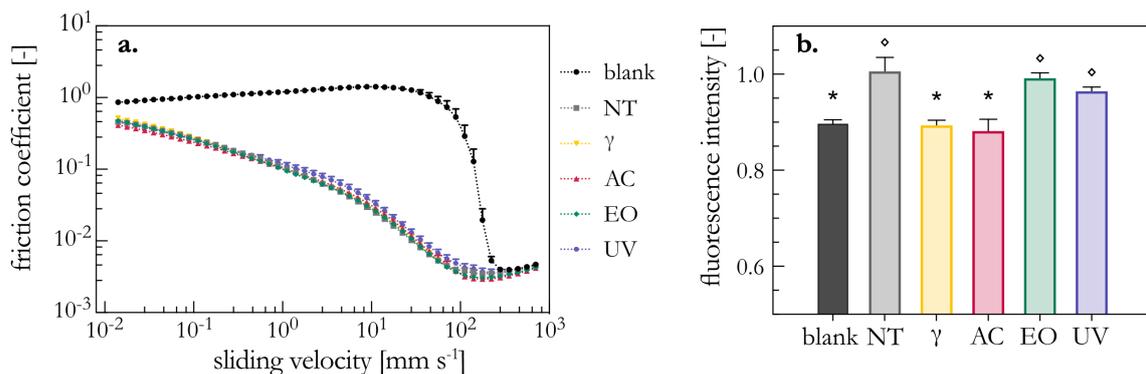
**Figure 3.18: Wettability of sterilized mucin coatings on different medical devices.** Contact angles ( $n \geq 10$ ) quantifying the surface wettability are displayed for medical specimens made from PU (a), PVC (b), and PDMS (c). Asterisks and rhombi denote statistically significant differences between a treated sample and the untreated reference or the blank sample, respectively (based on a  $p$ -value of 0.05). Exemplary images of the three blank device surfaces and the same set of surfaces carrying a mucin coating (untreated) are shown in (d). Break-up times were determined for PDMS-based contact lenses (blank, coated, or coated and sterilized; e). Error bars indicate the standard error of the mean obtained from  $n = 3$  samples.

The first observation is that, as expected, the wettability of the uncoated base materials differs. The CA values obtained for PU and PVC are located around the threshold between hydrophobic and hydrophilic behavior, whereas PDMS exhibits clear hydrophobic behavior with high contact angles around  $110^\circ$ . On all three materials, however, the mucin coatings decrease the contact angles by  $\sim 40^\circ$ – $50^\circ$ , which corresponds to an alteration of the surface properties into hydrophilic behavior. Remarkably, this strong alteration in the wetting properties is maintained for all materials after exposure to ethylene oxide; in addition, after irradiation with  $\gamma$  or UV rays, coatings generated on PU or PVC still provide clearly hydrophilic properties as well. Autoclaving, in contrast, leads to a strong increase of the contact angles, and this result agrees with the findings discussed above, which again underscores that this particular sterilization method induces severe damage to the mucin macromolecules.

Even though determining CA values is a standard approach in material science, in the context of contact lenses, the previously introduced, more application-oriented quantification of the liquid film BUT is conducted (Figure 3.18e). As demonstrated in chapter 3.2, the BUT for uncoated, blank PDMS lenses is below 1 s; in other words, the liquid film ruptures immediately after removing the sample from the liquid. Contact lenses that are covalently coated with mucins, in turn, perform way better: here, the measured BUT values are in the range of  $\sim 38$  s. After sterilization with ethylene oxide, the BUT still reaches 30 s, and the corresponding values are  $\sim 20$  s for AC,  $\gamma$ , or UV irradiation. Thus, all those values are much larger than the average duration of a human blinking period, which is  $\sim 5$  s.<sup>178</sup> Also, these results support the findings obtained from the CA measurements and demonstrate that all treatments maintain the wetting improvement brought about by the mucin coating – at least to a certain extent.

In addition to improving the wettability of surfaces, another important property brought about by mucins is providing lubricity. As explained in **chapter 2.5**, mucins typically achieve this *via* a combination of two effects: sacrificial layer formation and hydration lubrication. However, owing to the covalent coupling of the mucins to the surfaces of the medical devices (which prevents shearing off of the mucins), the sacrificial layer formation will be suppressed. In contrast, the second mechanism, hydration lubrication, should still be fully operable: as the densely glycosylated central region of the mucin glycoprotein can bind high amounts of water, it maintains a surface-bound lubricating liquid film even under tribological loads and thus reduces friction.

To assess the lubricating abilities of sterilized mucin coatings, rotational tribology is performed. For this particular trial, PDMS pins as described in **chapter 2.5** were used instead of actual contact lenses; this modification was necessary since the tribological setup requires this very sample geometry. For blank, uncoated PDMS samples (**Figure 3.19a**), a typical Stribeck's curve is obtained that shows low friction coefficients in the regime of hydrodynamic lubrication only (*i.e.*, at high sliding velocities, which correspond to blinking movements of the upper eye lid).<sup>182</sup> After a steep transition zone (mixed lubrication regime), the boundary lubrication regime is entered, which is most relevant for sliding speeds that are expected to occur between a contact lens and the cornea;<sup>223</sup> here, very high friction coefficients around 1 are obtained. For mucin-coated samples, however, reduced friction coefficients are recorded across almost the whole range of sliding speeds probed. Now, instead of a steep transition from low to high friction coefficients, a slow, gentle increase of friction with decreasing sliding speed is observed. Even at the slowest sliding speed probed, the coated samples still outperform the uncoated ones. Remarkably, none of the sterilization methods tested shows a measurable influence on the lubricity of the coating. With the previous results from the ELISA test and lectin assay in mind, this can be rationalized very well: As discussed above, the predominant lubrication mechanisms provided by covalent coatings is hydration lubrication, and this mechanism relies on the glycosylated parts of the mucin glycoprotein. As the results compiled in **Figure 3.17** showed, the glycan pattern of mucins is more resilient toward the sterilization methods tested here than the non-glycosylated, hydrophobic termini of mucin. Apparently, even with minor damages to this glycosylation pattern, the sterilized mucin layer can still bind sufficient amounts of water to provide hydration lubrication.



**Figure 3.19: Tribological behavior and lipid adsorption as observed for mucin-coated PDMS samples.** Stribeck's curves were obtained for PDMS samples that were either left uncoated (blank) or were covalently coated with mucin (**a**). The coated samples were either stored without any further treatment (NT), or they were treated with  $\gamma$  irradiation, autoclavation (AC), ethylene oxide fumigation (EO), and UV irradiation, respectively. The fluorescence intensities displayed in (**b**) were obtained in a lipid depletion assay. Higher values denote lower depletion of the lipids from the solution, hence, lower adsorption of the lipids onto the sample surfaces. Error bars in (**a**) and (**b**) denote the standard error of the mean as obtained from  $n \geq 4$  samples. Asterisks and rhombi in (**b**) denote statistically significant differences between a treated sample and the untreated reference or the blank sample, respectively (based on a  $p$ -value of 0.05).

## Mucin macromolecules for biomedical applications

In addition to providing lubricity, a second key function established by mucin coatings is to counteract biofouling events, *i.e.*, to reduce the undesired adsorption of molecules or cells onto surfaces.<sup>156,158</sup> For the medical devices studied here, this aspect is most relevant for contact lenses, which are optical devices that need to maintain a high transparency to allow for maximal light transmission. In **chapters 3.2 and 3.3**, it was shown that covalent mucin coatings strongly reduce the adsorption of lipids (a common constituent of the tear film) onto contact lenses, *i.e.*, such coatings help preserving the transparency of the optical device. Hence, in a last set of experiments, it is asked if this lipid-repellent effect is still present after sterilization of the mucin coatings, and a depletion assay is conducted to assess this question.

When exposing PDMS contact lenses to a lipid-rich liquid environment, a substantially higher depletion of lipids is observed for blank, uncoated samples than for unsterilized MUC5AC-coated contact lenses (**Figure 3.19b**). Importantly, for the latter samples, the measured fluorescence values suggest that, here, lipid adsorption onto the coated contact lens surface is negligibly low. This finding is in full agreement with previous results and it demonstrates the suitability of the employed depletion assay to study lipid deposition onto surfaces. Similarly good results as those obtained for untreated coatings are also reached with coated samples that were either subjected to ethylene oxide or UV sterilization. After  $\gamma$  irradiation or autoclavation, however, the outcome of this lipid deposition test is similar to that obtained for blank, uncoated PDMS lenses. Overall, these findings are consistent with the results obtained from the structural integrity tests shown in **Figure 3.17d-f**: Owing to the hydrophobic nature of PDMS, lipid adsorption can easily occur *via* hydrophobic interactions acting between the fatty acid chains of the lipids and the lens surface. For an intact mucin coating, the hydrophilic central region (which represents the largest part of the macromolecule) covers the surface and prevents the adsorption of hydrophobic objects. Accordingly, those sterilization methods that maintain the glycosylation pattern of the MUC5AC glycoprotein the best (*i.e.*, ethylene oxide exposure, and UV treatment), also preserve the lipid-resistance properties of the coating.

In summary, the results discussed here show that, among the sterilization methods investigated, MUC5AC coatings are most robust toward ethylene oxide exposure; here, the biochemical integrity of the mucins and the properties brought about by the coating were maintained the best. One major concern associated with an ethylene oxide-based sterilization process, however, is the putative retention of toxic residues in the material. To enable clinical usage of medical products that have been subjected to ethylene oxide fumigation, the amount of such toxic residues needs to be minimized, which is typically achieved by extensive aeration of the treated devices after sterilization. Yet, the efficiency of different aeration methods (such as air circulation under heat, pulsed vacuum postprocessing, or microwave desorption) and the necessary duration or intensity of such post-sterilization treatments needs to be individually studied for each medical device. Moreover, even though lab-purified mucin macromolecules were shown to be highly biocompatible,<sup>80</sup> assessing the biocompatibility of mucin coatings before and after sterilization should be tested following detailed ISO protocols (including endotoxin tests) so mucin-coated medical devices can enter the next stage towards medical application.

Overall, the results obtained here for sterilized medical devices carrying a mucin coating are very positive as they indicate that making use of the various beneficial properties established by such mucin coatings should be very well possible in a clinical context: The three medical devices tested in this thesis find broad usage in many medical disciplines. Moreover, they represent an even broader range of objects made from the same set of polymeric materials, which can equally profit from the hydrophilizing, anti-biofouling, or friction-reducing effects brought about by such mucin coatings.

## 4. Machine Learning to analyze complex surfaces

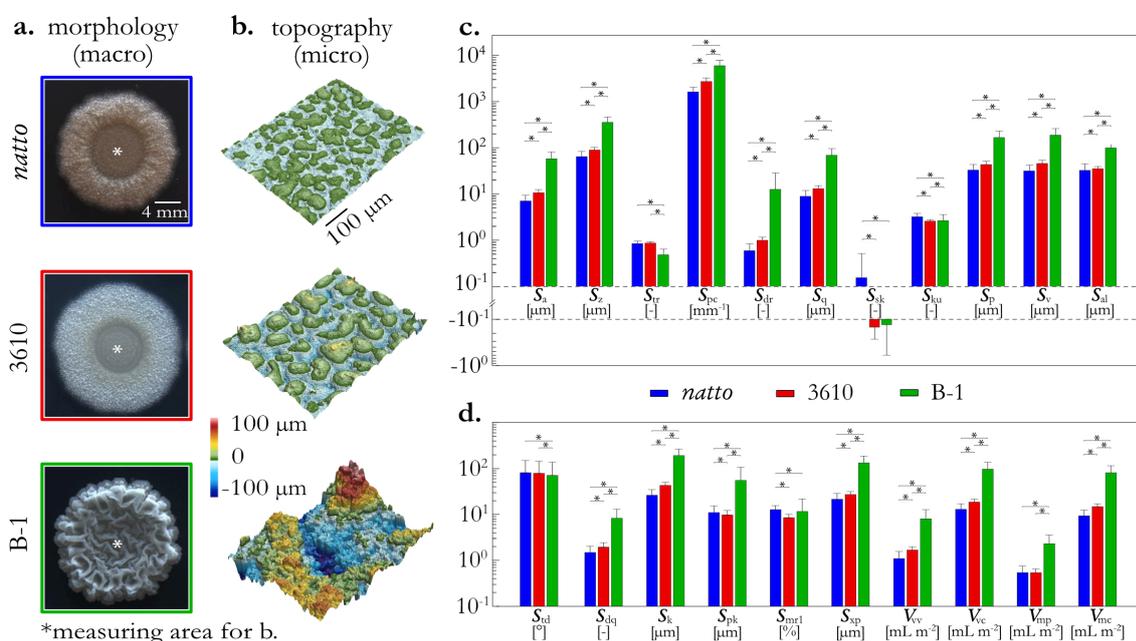
Assessing the quality of a surface by identifying and classifying damage features is a cornerstone in many fields of engineering. It helps engineers to evaluate the performance of a material and to make informed decisions about its suitability for specific applications or if modifications are required. Established methods to detect surface anomalies, such as cracks, pits, or scratches, include (macroscopic) visual inspections, optical microscopy, or scanning electron microscopy. Based on the acquired images, the dimensions, shape, and location of the respective defect can be measured to provide valuable insight regarding the type of damage the material suffered from. However, these techniques are mainly effective for simple and systematic damage morphologies. For more complex damage patterns, such as those observed in wear testing scenarios, such visual evaluation methods may not suffice. In such cases, an alternative approach is to determine the weight of the mass that was abraded during the test by comparing the weight of the material before and after subjecting it to tribological stress. This strategy provides a reliable measure to quantitatively evaluate the wear resistance of metals, ceramics, or hard polymer materials.

In addition to ‘classical’ wear studies conducted to understand and optimize the interaction of different components in a technical system, there is growing interest in investigating wear formation (or its prevention) in other systems, such as human tissue that is in contact with a medical device. As described in **chapters 3.2** and **3.3**, the wear caused by such devices can result in significant harm to the underlying tissue, potentially leading to adverse effects on the patient's health and well-being. Hence, understanding the tribological interactions between medical devices and human tissue is essential for designing and optimizing devices that minimize damage to surrounding tissues or organs. However, assessing the wear patterns on soft materials involved in such bio-tribological setups is often challenging: Owing to the complex surface morphologies of biological materials and the different length scales and heterogenous appearances of the (putatively) inflicted damages, the previously mentioned approaches mostly fall short.

In **chapters 3.2** and **3.3**, the assessment of tribology-induced damages inflicted by contact lenses on corneal tissue was manually conducted based on a small set of surface parameters along with a subjective evaluation of the profilometric images. However, this approach requires time-consuming manual assessments by a trained expert and may still remain prone to errors and biases. To overcome these challenges, in the following chapters of this work, the potential ML holds to aid in the assessment of the surface conditions of a soft tissue is explored. In a first step, supervised ML models are employed to test whether complex biological surfaces can be accurately classified based on a set of surface parameters obtained from captured 3D profiles. For such supervised learning approaches, the samples of the dataset are not only characterized by the mentioned descriptors (features) but each sample additionally carries a predefined label. For this purpose, a prototype dataset comprising complex biological surfaces (*i.e.*, bacterial biofilms and plant samples) that can be easily labelled is used. Afterwards, an unsupervised approach is employed to classify corneal tissue samples. Here, each sample is solely described by the input features and no predefined label is provided. The clustering results achieved using the surface descriptors as input features are then evaluated in detail and compared to a manual classification conducted by humans. One important step to be included in such ML pipelines is feature elimination – an efficient strategy to overcome the challenges encountered due to limited data availability when applying ML to experimental research problems. Accordingly, in the following, a correlation-driven feature elimination algorithm is developed and used to pre-process the dataset comprising the corneal surfaces.

## 4.1. Analyzing the surfaces of biological materials\*

As an initial step towards the goal of applying Machine Learning for analyzing complex surfaces of soft materials, first, a broad overview of the topographical features of bio-surfaces with different morphological characteristics is obtained. For this purpose, the central areas of surface-bound bacterial colonies (so-called biofilms) are analyzed. This particular biomaterial comes with the advantage that its surface properties can be broadly varied by selecting different bacterial strains and/or growth conditions for biofilm generation. In this first step, three different *Bacillus subtilis* variants that are closely related are cultivated on MSgg agar and analyzed after 24 h of incubation (this incubation time is customary when mature biofilms are desired;<sup>224</sup> **Figure 4.1a,b**). Macroscopically, the central areas of biofilm colonies generated by *B. subtilis natto* and *B. subtilis* NCIB 3610 bacteria are both fairly flat; their micro-topographies, however, appear different, as the latter biofilms exhibit somewhat bigger structures on this microscopic length scale. *B. subtilis* B-1 biofilms have a very diverse appearance, both on the macro- and microscale; they exhibit large-scale wrinkles and (locally) wavelike features.



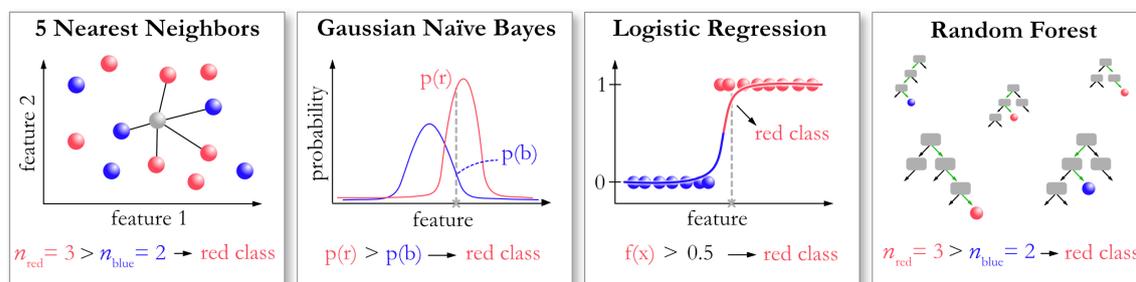
**Figure 4.1: Surface structure of different bacterial biofilms.** Macroscopic colony images (a), microscopic profilometric images (b), and metrological surface parameters (c,d) as determined from the latter are shown for biofilms generated by *B. subtilis natto* (blue,  $n = 705$ ), *B. subtilis* NCIB 3610 (red,  $n = 592$ ), and *B. subtilis* B-1 (green,  $n = 491$ ) bacteria cultivated on MSgg agar. The scale bars in (a,b) apply to all three images within the respective subfigure. The error bars in (c) and (d) denote the standard deviation as obtained from  $n \geq 491$  images each. Asterisks indicate statistical differences based on a  $p$ -value of 0.05.

For all biofilm variants mentioned above, profilometric images (approximately 500 per strain) acquired at the microscale are evaluated further. Similar as described in **chapters 3.2** and **3.3** for analyzing corneal tissue, different metrological surface parameters (a broad potpourri of height, spatial, hybrid, functional, and functional volume parameters according to EN ISO 25178, see **Figure 4.1c,d**) are calculated from those images to quantify different aspects of their surface topography (for mathematical definitions of the surface parameters, please refer to **Appendix A3**). For the majority of those 21 parameters, significant differences are obtained that allow for distinguishing the three different biofilm types. Overall,

\* This section follows in part the publication Rickert *et al.*, *ACS Biomaterials Science & Engineering* (2021)

for most parameters, the highest absolute values are found for B-1 biofilm surfaces; in contrast, *natto* biofilms tend to return the lowest values, yet this is only a general trend and does not provide a strict sorting criterion.

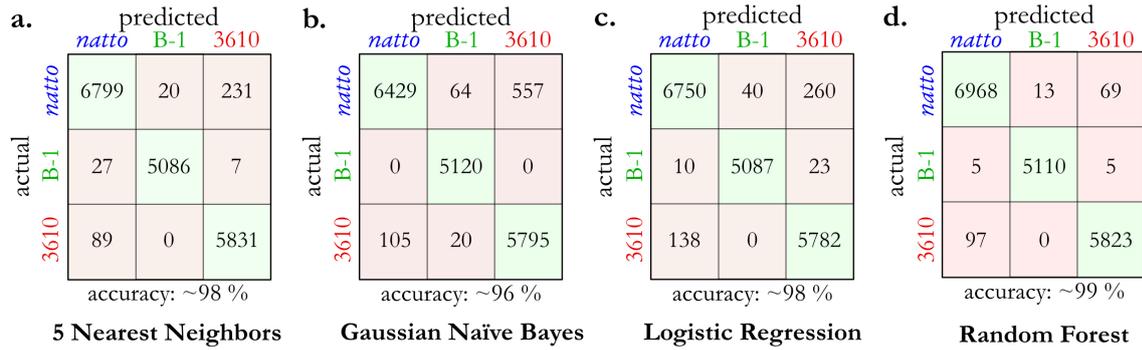
Thus, in the next step, an algorithm-based classification of the topographical biofilm images based on those metrological parameters is attempted. In detail, it is tested whether an automatized classification procedure can reliably decide what biofilm variant a parameter set obtained from a particular image belongs to. For this purpose, different Machine Learning methods of different complexity and abstraction levels are selected (**Figure 4.2**). Each of those algorithms is first trained and then tested with the dataset shown in **Figure 4.1**. For this purpose, the accuracies of the categorization decisions made by the four algorithms are determined by applying a repeated stratified 5-fold cross-validation (see **chapter 2.8.3**); in other words, the dataset (a total of  $\sim 1500$  images comprising  $>495$  images per biofilm variant) is repetitively separated into training data (80 %) and test data (20 %) and the obtained prediction accuracies are averaged.



**Figure 4.2: Schematic illustration of the four Machine Learning algorithms used in this study.** The  $k$  Nearest Neighbors classifier (a) identifies the  $k$  samples that are closest to the query point in an  $n$ -dimensional space ( $n$  equals the number of features) and assigns the query sample to that class which has the highest number of representatives (within those  $k$  samples). The Gaussian Naïve Bayes algorithm (b) calculates conditional probabilities and categorizes samples based on a “fits better” principle. The Logistic Regression algorithm (c) models the probability of classes by assigning importance weights and fitting a sigmoid function through all features. The Random Forest classifier (d) performs predictions in a popular-vote-like manner considering the decision of a large number of random, uncorrelated decision trees.

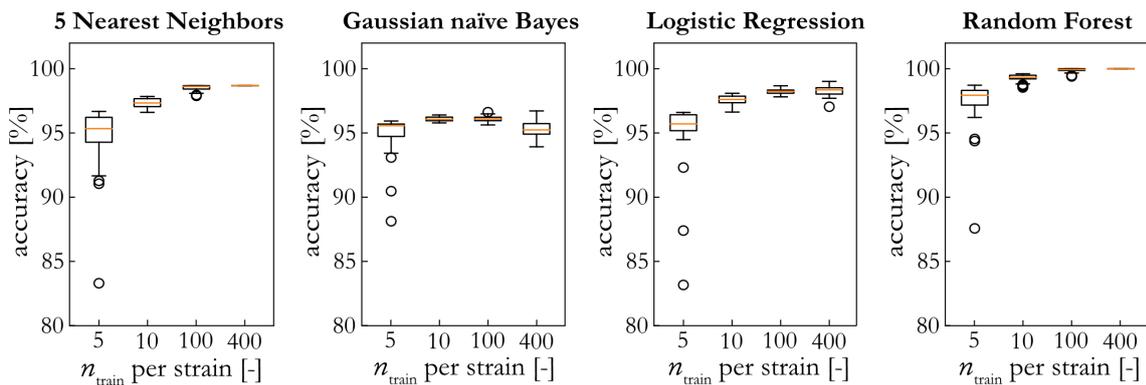
First, a rather basic approach is tested by applying a  $k$ NN ( $k$  Nearest Neighbors) classifier. Here, classification is performed based on a spatial comparison of the test data with the training data. When applying this method to the dataset described above, this  $k$ NN classifier achieves a prediction accuracy of approximately 98 % (**Figure 4.3a**). Similarly high accuracies are obtained with the other three algorithms. The Gaussian Naïve Bayes classifier (**Figure 4.3b**) calculates conditional probabilities based on the simplified assumptions that all features are independent, normally distributed and equally contributing. The Multinomial Logistic Regression (**Figure 4.3c**) determines the logarithmic odds of each class as a linear combination of the feature variables. Last, the Random Forest classifier (**Figure 4.3d**) makes predictions based on the most popular decision of a large number of randomly generated, uncorrelated decision trees. Moreover, in all cases tested, the calculated specificity and sensitivity values are above 90 % (see **Appendix A5.6**). From the very high classification accuracies obtained with each of the four algorithms, it can be concluded that the biofilms generated by the three different bacterial strains on MSgg agar can indeed be clearly classified with the set of topographical parameters calculated here.

## Machine Learning to analyze complex surfaces



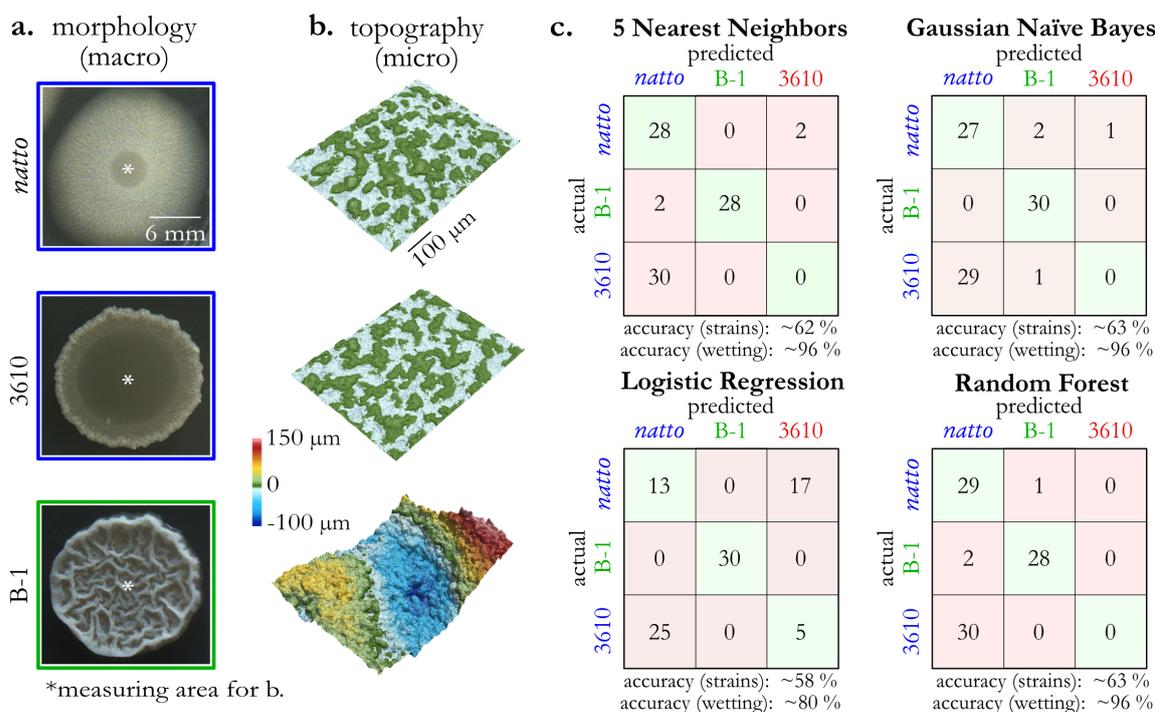
**Figure 4.3: Categorization results obtained with four different ML algorithms fed with biofilm data obtained on MSgg agar.** The confusion matrices (a-d) compare the actual classes (*i.e.*, the names of the bacterial strains from which the biofilms were grown) with the predicted classes when a 5 Nearest Neighbors classifier (a), a Gaussian Naïve Bayes model (b), multinomial logistic regression (c), and a Random Forest classifier (d) are applied to the data. The evaluation was performed by repeated (10 times) 5-fold cross validation with shuffling the dataset before each iteration; that is, each sample was used as a test sample 10 times with differently trained algorithms. The main diagonal (green boxes) contains the numbers of correctly classified samples whereas all other boxes represent wrongly assigned samples.

One major challenge when trying to apply Machine Learning approaches to experimentally obtained data is the rather small amount of available data points; typically, the number of such experimental repetitions is much smaller than the 500 topographical images captured here for the different biofilm variants. Thus, in the next step, the robustness of these ML approaches is tested when smaller, lab-scaled datasets are used. In detail, the same algorithms employed so far are challenged with a drastically reduced size of training data (Figure 4.4). Surprisingly, in all four cases even a much smaller number of training data entails categorization accuracies above 90 %. Whereas 5 training images per biofilm variant seem to be insufficient (here, all four algorithms return fluctuating decisions with accuracies between ~80–95 %), datasets calculated from 10 distinct images are enough to obtain stable and reliable results. This is remarkable considering that the dataset comprises images of biofilms grown on different days; it thus encompasses a significant level of (typical) biological variability.



**Figure 4.4: Prediction accuracies of the ML models for varying training set sizes.** The accuracy values of the four algorithms denote the total proportion of correctly classified samples as a function of the training set size (data obtained from  $n = 50$  runs in a 10 times repeated stratified 5-fold cross-validation).

Having verified that the selected algorithms can make reliable predictions based on the given set of topographical parameters, those classifiers are next challenged in a different way: they are asked to classify biofilms that were grown at different conditions than the biofilm samples with which the algorithms are trained. At this point, it is important to recapitulate that the three biofilm variants compared so far not only differ in terms of their surface characteristics but also with regard to a physical property, which is intimately related to the surface topography of a material: the surface wettability.<sup>225</sup> As described in **chapter 2.8.1**, the growth conditions for cultivating the three *B. subtilis* variants had been chosen such that the generated biofilms exhibit different wetting behaviors: *B. subtilis natto* biofilms have a hydrophilic surface; *B. subtilis* NCIB 3610 biofilms have superhydrophobic surfaces similar to rose petals; and *B. subtilis* B-1 biofilms have superhydrophobic surfaces similar to lotus leaves. The key difference between the latter two variants of superhydrophobic behavior can be identified when a water droplet is placed onto a tilted surface: on rose-like surfaces, the droplet sticks; in contrast, on lotus-like surfaces the droplet rolls off. In previous studies, these differences in surface wettability were found to mainly depend on two factors: the chemical composition and the surface topography of the sample.<sup>226,227</sup> With the same set of bacterial strains used so far, biofilms with different wetting profiles can be obtained by changing the growth medium from MSgg to LB (see **Appendix A5.2**). As a consequence of this alteration in nutrient availability, biofilms with a different macro-morphological appearance are created (**Figure 4.5a**).



**Figure 4.5: Surface structure and classification of biofilms grown on LB medium.** Macroscopic colony images (a) and microscopic profilometric records (b) of biofilms generated by *B. subtilis natto*, *B. subtilis* NCIB 3610, and *B. subtilis* B-1 bacteria cultivated on LB agar. The scale bars in (a,b) apply to all three images within the respective subfigure. The confusion matrices (c) show the categorization results of such samples as achieved by the four algorithms when fed the set of micro-topographical parameters: the actual classes (*i.e.*, the names of the bacterial strains the biofilms were grown from) are compared to the predicted classes when a 5 Nearest Neighbors classifier, a Gaussian Naïve Bayes model, Multinomial Logistic Regression, and a Random Forest classifier are used. All algorithms were trained with data obtained for biofilms grown on MSgg medium. The label colors indicate the wetting behavior of the respective sample: blue indicates hydrophilic behavior, red indicates rose-like hydrophobicity, and green denotes lotus-like hydrophobicity.

## Machine Learning to analyze complex surfaces

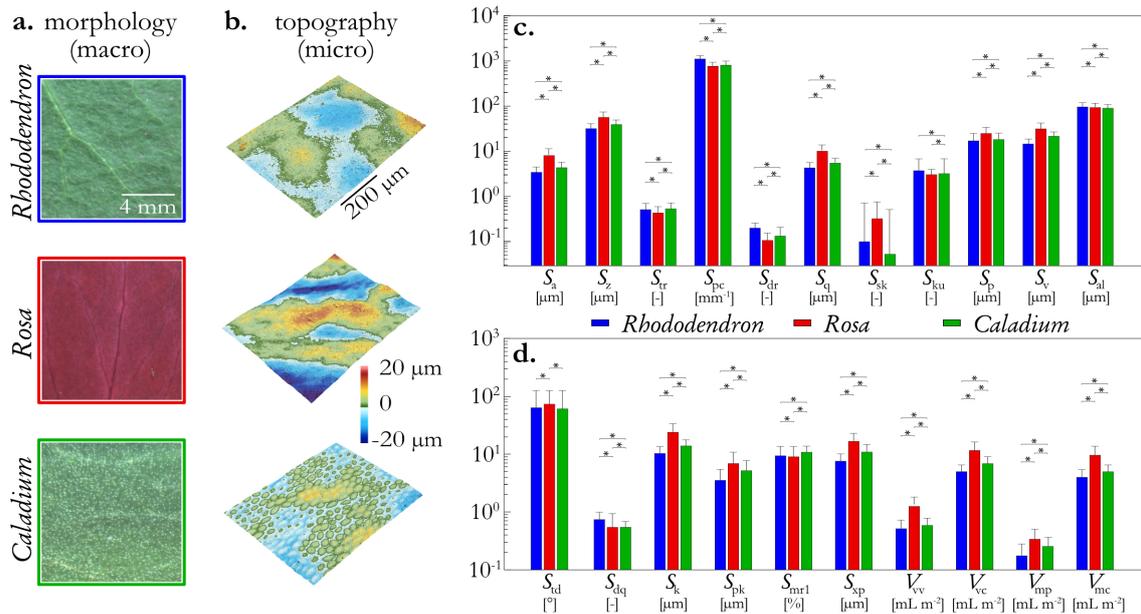
Also for this new set of biofilm colonies, profilometric images are acquired on the microscale (**Figure 4.5b**) and the same metrological parameters are calculated as for the biofilm samples discussed before (for an overview of this data, see **Appendix A5.7**). One interesting aspect is that for the strain 3610, not only the visual appearance of the biofilm colonies has changed when biofilm growth is conducted on LB agar, but also the wetting behavior of the biofilm colonies differs: now, hydrophilic surfaces are obtained (see **Appendix A5.2**) instead of rose-like superhydrophobic surfaces. In contrast, biofilm colonies grown from *B. subtilis natto* and B-1 bacteria exhibit the same wettability on LB agar as they do on MSgg agar, *i.e.*, they show hydrophilic and lotus-like behavior, respectively. Consequently, this new set of samples does not contain any biofilms with rose-like surface properties.

When the four algorithms trained with data obtained for biofilm colonies grown on MSgg agar are used to classify the biofilms grown on LB agar (**Figure 4.5c**), two aspects are observed. First, three out of four algorithms correctly and accurately classify *natto* and B-1 biofilms; only the logistic regression approach misinterprets the majority of images acquired from *natto* biofilms and assigns them to 3610 biofilms. This shows that those three classifiers are generalized enough to classify samples even if they differ from the training set in terms of macro-morphological and micro-topographical appearance. The second observation is that all four algorithms mainly assign data obtained from 3610 biofilms to *natto* biofilms. At first glance, this seems to be a mistake as indicated by the calculated low accuracies, which are only in the range of 60 %. However, when the detailed microbial species from which the biofilms are grown is ignored, and instead, the previously mentioned wetting properties of the biofilm colonies are considered (which are dictated by their surface topographies), this classification turns out to be highly meaningful: 3610 biofilms grown on LB medium exhibit the same wetting characteristics as *natto* biofilms grown on MSgg agar and both have hydrophilic surfaces. Thus, when the three variants of surface wetting properties are used as a class label, the  $k$ NN classifier, the Gaussian Naïve Bayes algorithm, and the Random Forest classifier correctly assign almost all data and reach accuracies close to those obtained for the biofilms grown on MSgg agar (approximately 96 % each). With this realization in mind, those wetting-based labels are now used for all data discussed in the remainder of this chapter. Importantly, even with those new labels used, the comparatively low accuracy of the logistic regression algorithm remains. Probably, this issue arises from an overfitting of the training data by the regression function, which is a common issue of this method. Hence, this particular classifier is not used further.

So far, three algorithms were identified that are capable of classifying biofilms according to their surface wettability by using a set of metrological parameters determined from the microtopography of the biofilms. In a next step, it is tested if the same approach can be applied to a different class of biomaterials. For this purpose, three different plant samples are selected (**Figure 4.6a**), which cover the same range of wetting properties as the biofilm samples analyzed above: *Rhododendron* leaves, that exhibit hydrophilic surfaces; *Rosa* “harlekin” petals that behave as rose-like superhydrophobic; and *Caladium praetermissum* leaves, which have superhydrophobic surfaces similar to lotus leaves (see **Appendix A5.2** for contact angles and tilt experiments). From those samples, profilometric images are again captured (**Figure 4.6b**) and the same set of surface parameters as before is calculated to quantify the microtopography (**Figure 4.6c,d**).

One major difference between this new dataset and the one obtained for biofilms is that, now, the distributions of calculated parameter values overlap more strongly; in other words, in terms of micro-topography, the plant surfaces are more similar to each other than the three biofilm variants discussed before. When the algorithms are trained with the plant data and the classification accuracy is evaluated *via* repeated cross-validation (again repeatedly using 80 % of the data for training and the remaining 20 % for testing), the following is observed: the Random Forest classifier performs best with a success rate of 92 %; the  $k$ NN classifier delivers acceptable results (87 % accuracy), and the Gaussian

Naïve Bayes algorithm returns the lowest yet still moderate accuracy (78 %, see **Appendix A5.8**). These findings indicate how the ability of the three classifiers to handle a dataset with overlapping classes differs. When applying the  $k$ NN, for instance, samples located at the intersection of two or more classes are prone to be misclassified due to the presence of numerous “wrong neighbors”. The Gaussian Naïve Bayes classifier has even stronger problems in this particular case. Here, samples that exhibit feature values far away from the mean of their actual class are very likely to be classified wrongly: Owing to the overlapping probability distributions, those samples are more likely interpreted to belong to another (= ‘wrong’) class. The Random Forest algorithm, in contrast, seems to handle such datasets quite well; and this can be attributed to the ability of this algorithm to narrowly subdivide the training samples when splitting the dataset.



**Figure 4.6: Surface topography of different plant samples.** Macroscopic images (a), microscopic profilometric images (b), and metrological surface parameters (c,d) as determined from the latter are shown for *Rhododendron* leaves ( $n = 600$ ), *Rosa* petals ( $n = 600$ ), and *Caladium* leaves ( $n = 500$ ). The scale bars in (a,b) apply to all three images within the respective subfigure. The error bars in (c,d) denote the standard deviation as obtained from  $n \geq 500$  images each. Asterisks indicate statistical differences assessed with a Wilcoxon–Mann–Whitney test applying a  $p$ -value of 0.05.

So far, training the algorithms with a subset of either biofilm or plant data allows for correctly sorting the rest of the datasets and using the wettability of the different samples as a classification criterion is a highly meaningful option. Naively, one might assume that training the algorithms with one set of data (e.g., using either biofilm or plant samples only) might also allow for classifying the other dataset. However, in none of the two possible directions (biofilms  $\rightarrow$  plants, plants  $\rightarrow$  biofilms) this is possible; very low prediction accuracies are obtained that are only slightly higher than the probability of simple guessing (see **Appendix A5.9**). In retrospect, this outcome is not surprising considering the very different ranges of absolute values that are obtained for the metrological parameters calculated for biofilm and plant samples, respectively. In other words, during training, the algorithms get adjusted to the feature scales of the respective material class (either biofilms or leaves) and thus, are not generalized enough to be applied to other surfaces with features that cover a much broader (or much narrower) range of length-scales.

## Machine Learning to analyze complex surfaces

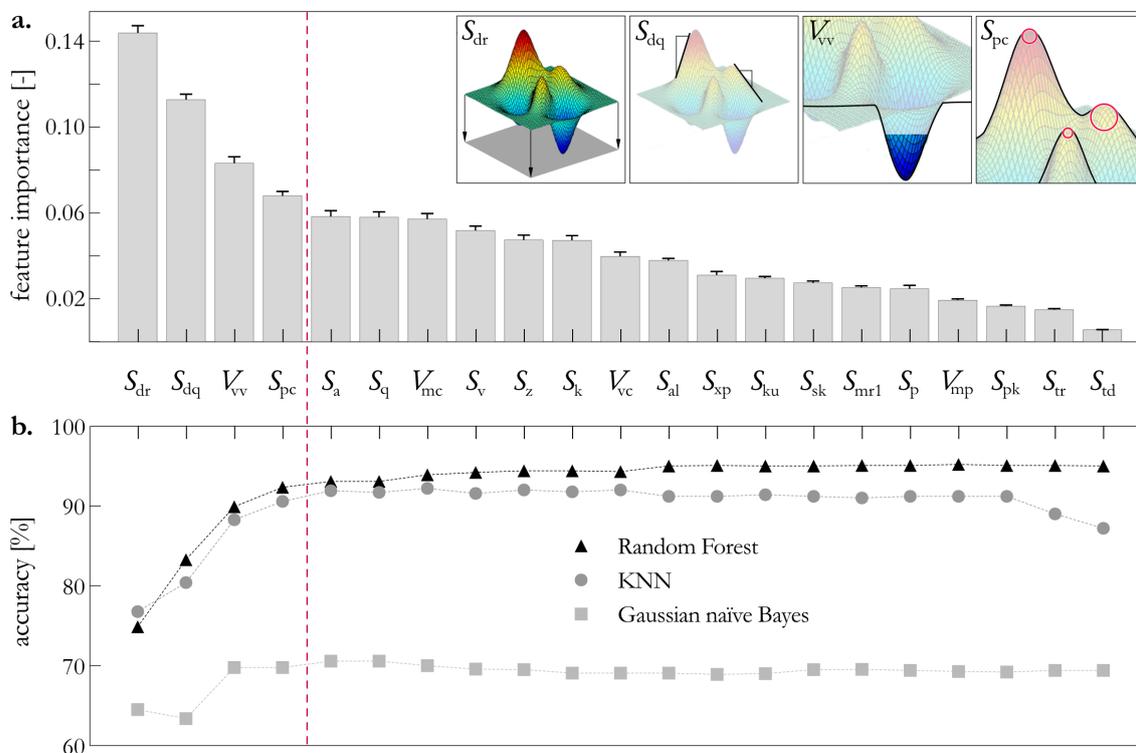
Thus, in the next step, the data obtained from both measurement series is pooled and a training/classification approach with this mixed dataset is tested. When now the classification results are evaluated *via* cross-validation (80 % training data and 20 % test data), it is observed that the three algorithms handle the pooled dataset differently well (see **Appendix A5.10**). Once more, the Random Forest algorithm performs best; here, very high classification accuracies of 95 % are achieved. The  $\epsilon$ NN classifier still reaches an acceptable accuracy of around 87 % (which is similar to the result obtained with this classifier when using plant data only), but the Gaussian Naïve Bayes classifier returns less accurate results than when applied to “biofilm-only” or “plant-only” data. This outcome is, however, not surprising considering that this algorithm assumes normally distributed parameter values, that is a condition that is clearly violated when the two datasets are pooled.

In a last step, it is asked which surface features of plant and biofilm samples are most suitable for a correct classification into objects with hydrophilic, rose-like, and lotus-like properties. For this purpose, only the Random Forest algorithm is further tested as this classifier has returned the best overall results for the data studied here. By tracing the subset of metrological parameters that this algorithm has used to make its predictions, specific surface characteristics are identified that influenced the decision-making process of the classifier the most. This method, also known as feature elimination, is a common approach in Machine Learning and can help reduce the dimensionality of datasets. If this is possible, not only the runtime of the algorithms can be shortened but also the accuracy of the algorithms can be improved by eliminating redundant or less relevant parameters.

To do so, the Gini importance is calculated, which describes the absolute decrease in node impurity induced by a feature weighted by the number of samples it splits.<sup>228</sup> For each feature, these values are then averaged over all trees of the random forest (here:  $n = 1000$ ) to obtain final importance scores. Then, all parameters are ranked according to this score (see **Figure 4.7a**). Interestingly, some of the top ranked parameters from this sorted list have been used previously to distinguish various samples with regard to their surface properties. The developed interfacial ratio  $S_{dr}$ , the peak curvature  $S_{pc}$ , and the root-mean-square gradient  $S_{dq}$ , for example, were shown to correlate well with the wettability of different synthetic surfaces.<sup>229-231</sup> Moreover, by making use of  $S_{dr}$  and the dale void volume  $V_{vv}$ , the wetting behavior of different biological surfaces could be discriminated,<sup>232,233</sup> and the peak curvature  $S_{pc}$  was discussed in context with the rose petal effect.<sup>234</sup> From a physical point of view, it is highly meaningful that the algorithm has identified this particular parameter subset to be relevant: The dale void volume together with the peak curvature, mean gradients, and the developed surface characterizes the size and shape of cavities in the surface structure of a given material; such surface cavities, in turn, were identified to crucially affect the wetting behavior of a material.<sup>235</sup> As the  $S_{dr}$  describes the increase of a surface area due to its texture, it quantifies the total area available for a surface-liquid interaction. The  $S_{pc}$ , in turn, hints toward the actual shape of some of those texture elements. A large surface area combined with smooth peak curvatures is typical for surfaces covered with ‘mushroom-like’ structures, and such a particular surface topography can be described best by the Cassie-Baxter model, which is associated with lotus-like wettability.<sup>236</sup> These findings indicate that the physical models that were established for regularly shaped, synthetic surfaces can also be applied to the irregularly structured surfaces of the biomaterials studied here.

If some of those metrological parameters are more important than others, is it really necessary to provide the algorithms with 21 different numbers characterizing each topographical image? This question is targeted by repeating the classification procedure with the pooled dataset but feeding the algorithms increasingly smaller parameter subsets. In detail, the Random Forest, the  $\epsilon$ NN, and the Gaussian Naïve Bayes classifier are challenged to sort images again and again; yet, in each sorting step the least important feature left from the sorted list shown in **Figure 4.7a** is eliminated until only  $S_{dr}$  values are left

(Figure 4.7b). Importantly, once fewer than the four top-scored parameters (*i.e.*,  $S_{dr}$ ,  $S_{dq}$ ,  $V_{vv}$ , and  $S_{pc}$ ) are considered, the accuracies achieved by the three algorithms are reduced (for specificity and sensitivity of the classification with the Random Forest classifier see Appendix A5.11). In other words, the four parameters discussed above seem to be both, necessary and sufficient to allow for a successful categorization of the different surfaces *via* model-based Machine Learning.



**Figure 4.7: Reduction of data dimensionality based on feature elimination.** Feature importance ranking as derived from the classification results obtained with the Random Forest classifier (a) as assessed by the Gini importance of each feature; that is, the total decrease of node impurity induced by a feature weighted by the number of samples the feature splits at the respective node. The error bars denote the standard error of the mean as obtained from  $n = 1000$  tree estimators of one Random Forest model. On the basis of this importance ranking, features are stepwise eliminated starting with the least important feature until only the  $S_{dr}$  is left. The mean classification accuracies ( $n = 50$ ) obtained during this feature elimination procedure are depicted in (b). Error bars depicting the standard error of the mean are smaller than the symbol size.

In a last set of trials, to complement the model-based algorithms applied so far, a Deep Neural Network (DNN) is employed to investigate whether deep learning methods could possibly outperform the prediction accuracy of the model-based ML approaches discussed so far. Therefore, a DNN is established, trained, and tested (*via* 10 times repeated stratified 5-fold cross-validation) with the three datasets, *i.e.*, those containing biofilms, plant samples, and a pooled dataset combining the two. Indeed, when classifying biofilms (after training on biofilms, Figure 4.8a) an exceptionally high accuracy of 99 % is obtained. When analyzing the plant samples with the DNN (after training the DNN with plant data only, Figure 4.8b), still a very good classification accuracy of 94 % is received. These results indicate that the DNN slightly outperforms the Random Forest classifier (which, among the model-based classifiers tested here, performed best). It is worth noting, that (similar to the model-based algorithms) the DNN cannot predict the wettability of plant samples when trained with biofilm data and *vice versa* (see Appendix A5.12). For the pooled dataset, however, an accuracy of 93 % is reached (Figure 4.8c). This is satisfactory, yet slightly below the accuracy reached by the Random Forest classifier.

## Machine Learning to analyze complex surfaces

		a. biofilms			b. leaves			c. biofilms & leaves		
		predicted			predicted			predicted		
		hydro- philic	lotus- like	rose- like	hydro- philic	lotus- like	rose- like	hydro- philic	lotus- like	rose- like
actual	hydro- philic	6977	13	60	5721	256	23	12540	261	249
	lotus- like	22	5098	0	143	4536	321	261	9218	641
	rose- like	65	0	5855	35	331	5634	202	715	11003
		accuracy: ~99 %			accuracy: ~94 %			accuracy: ~93 %		

**Figure 4.8. Categorization results obtained with a Deep Neural Network.** The confusion matrices compare the actual wetting behavior of datasets obtained from different biofilms (a), leaves (b), or pooled biofilm/leave samples (c) with the wetting behavior predicted by a DNN that was trained with 80 % of the respective dataset. The evaluation was performed by repeated (10 times) 5-fold cross validation, where the dataset was shuffled before each iteration, that is, each sample served 10 times as a test sample but each time was analyzed by a differently trained algorithm. The main diagonal (green boxes) contains the numbers of correctly classified samples whereas all other boxes represent wrongly assigned samples. The accuracy values below the confusion matrices denote the total proportion of correctly classified samples.

In summary, it was demonstrated that different ML algorithms are capable of classifying the surfaces of two very different types of biological materials based on their micro-topographical properties. This was even possible with relatively small, lab-scale training sets and as few as four different topographical parameters. Among the four model-based algorithms compared, the Random Forest classifier turned out to be the most robust one; this particular algorithm could reliably handle both, broad and narrow datasets of different sizes, densities, and distributions. The Gaussian Naïve Bayes algorithm was efficient for normally distributed datasets but struggled when fed with more diverse data distributions. The multinomial logistic regression seemed to be prone to overfitting, which limited its suitability. Finally, the  $k$ NN classifier delivered satisfactory accuracies in most trials and was especially suitable when low-dimensional input data was used. For the datasets containing samples of one biological material only (*i.e.*, either biofilms or plants), a DNN delivered comparable results as the Random Forest classifier. Whereas no classifier was able to perform cross-prediction (*i.e.*, predicting biofilms after being trained on plant samples as *vice versa*), the Random Forest classifier and the DNN delivered satisfactory results for the pooled dataset. Last, the Gini importance was used to reduce the feature set from 21 to 5 features that are sufficient to represent the whole feature space. For the  $k$ NN, this reduction of dimensionality even entailed a slight increase in prediction accuracy.

## 4.2. Correlation-based feature elimination\*

To a large extent, the success of ML approaches depends on the size and the quality of the database available for generating the models:<sup>237</sup> When samples are characterized by a vector containing quantitative measures of different properties (in **chapter 2.9** referred to as feature vector), a certain density of data points is needed to sufficiently cover the relevant region of the feature space. As shown for the  $k$ NN algorithm tested in **chapter 4.1**, even though providing more information for each sample, a larger feature vector can reduce the prediction accuracy of an algorithm.<sup>238</sup> This phenomenon is known as the ‘curse of dimensionality’; It often originates from a low data density in the multi-dimensional feature space on the one hand, and a distance concentration on the other hand (*i.e.*, the observation that pairwise distances of sample points in a high-dimensional feature space tend to converge to the same value).<sup>239</sup>

There are, of course, strategies to mitigate the problems associated with this curse of dimensionality, *e.g.*, increasing the sample size or reducing the dimensionality of the feature space. However, the former approach can be very costly and time consuming – especially for scientific problems that depend on experimental data acquisition. Thus, reducing the dimensionality of the dataset by selecting a subset of features while discarding all others – as in **chapter 4.1** achieved by ranking the features according to their Gini importance – is often the more feasible approach.<sup>240,241</sup> Feature selection techniques can be broadly subdivided into three categories: wrappers, embedded methods, and filters.<sup>242,243</sup> Wrapper methods use prediction results obtained with a specific ML model as a score to evaluate the usability of a given feature set.<sup>244</sup> Basic model performance measures are then repeatedly assessed to identify an optimal feature set following a greedy search approach.<sup>245,246</sup> One very common example is the sequential feature selection;<sup>247</sup> here, features are iteratively added (forward selection) or removed (backward selection) to establish a feature subset in a greedy fashion. In each iteration, the best feature to be added or to be removed is chosen based on the cross-validation score of a given ML model. However, since a full ML algorithm has to be executed several times, wrapper methods usually come with high computational costs and long runtimes.<sup>248</sup> Moreover, the optimality reached is model-specific (and not necessarily transferrable to other models) and this search for an optimum can even severely increase overfitting of the optimized model.<sup>249</sup>

Instead of eliminating features ‘outside’ the ML model, embedded methods directly integrate the feature selection into the learning process.<sup>250,251</sup> One of the most popular examples for embedded feature selection is the previously employed Random Forest classifier, that selects an individual feature to split the dataset in each step of the tree growth process.<sup>252,253</sup> Other commonly used embedded feature elimination approaches are the LASSO<sup>254</sup> (L1 penalty) and Ridge<sup>255</sup> (L2 penalty) regression for constructing a linear model. In these two methods, feature weights are on purpose reduced to (almost) zero, which basically corresponds to an elimination of those features. However, those methods usually do not analyze (putative) redundancy within the features, and the outcome of such ranking-based eliminations is only valid for the particular model that was used to conduct the elimination process.

Last, filter approaches analyze intrinsic properties of a dataset and conduct feature selection independent of any ML model.<sup>256</sup> Popular filter methods employ statistical quantifications to assess the impact a feature has on a given output label.<sup>257,258</sup> Even though the most popular feature selection approaches are used for supervised learning, *i.e.*, labelled data, filter methods can also be applied to analyze the feature space of unlabeled data. Since, here, no class or prediction label can be used to guide the search for important information, feature elimination has to be performed by solely evaluating intrinsic properties

---

\* This section follows in part the publication Rickert *et al.*, *APL Machine Learning* (2023)

## Machine Learning to analyze complex surfaces

of the dataset such as feature dependence,<sup>259,260</sup> the entropy of distances between data points,<sup>261</sup> or the Laplacian score.<sup>262</sup> Other examples of filter techniques to reduce the dimensionality of unlabeled data are principal component analyses (PCA),<sup>263</sup> factor analyses,<sup>264</sup> or projection pursuit.<sup>265,266</sup> However, rather than actually selecting some features while discarding others, those approaches perform feature transformations. Thus, even though a PCA, for instance, is easy to use, the interpretability of the selected principal components is usually rather low.<sup>267,268</sup>

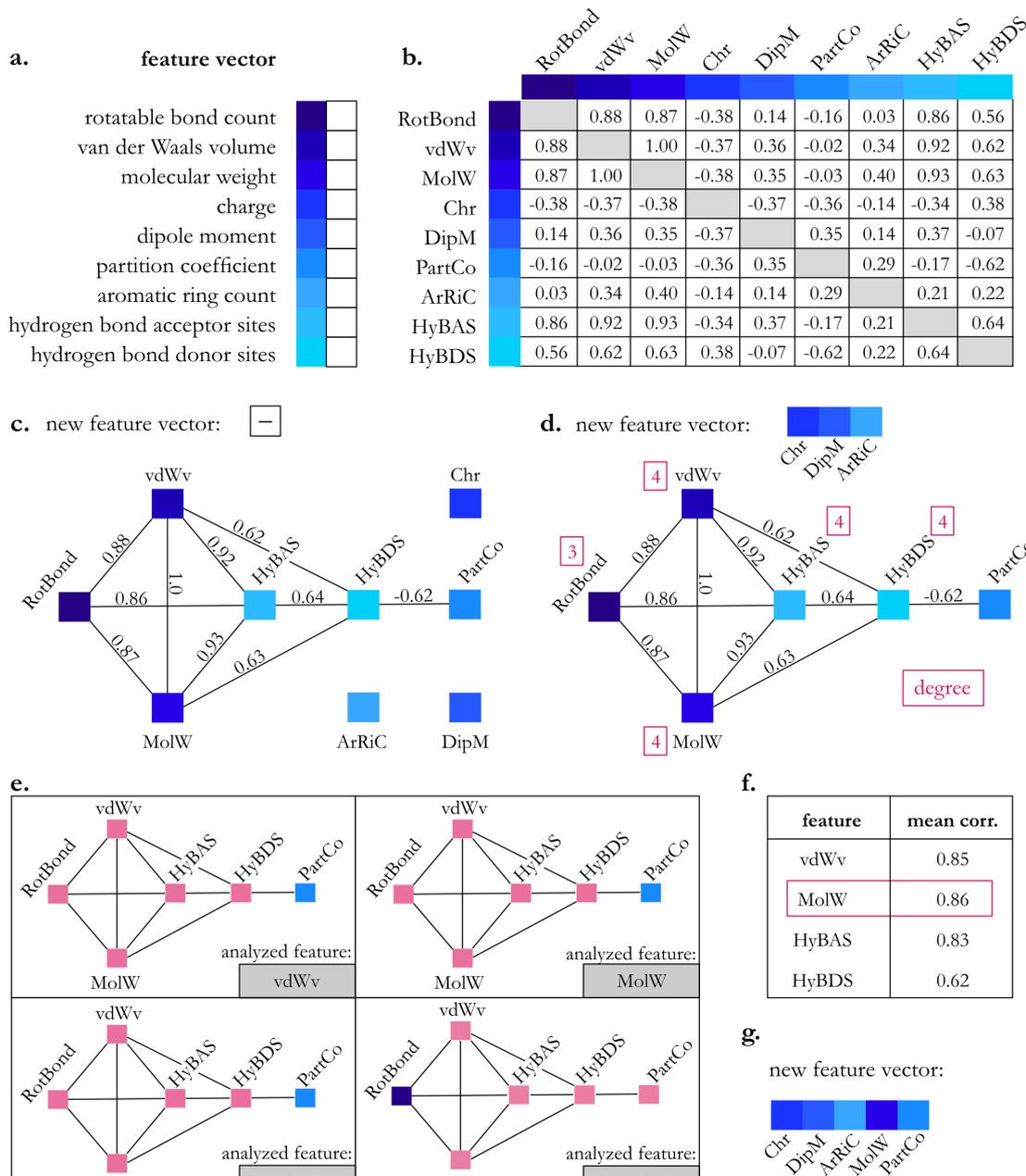
For the datasets used in **chapter 4.1**, basically all of the introduced feature elimination strategies could be applied, as the samples are labelled and predefined ML models were used. However, for the ML-driven analysis of an unlabeled dataset such as the targeted analysis of corneal images, a model- and output-independent feature elimination algorithm is needed. Hence, in this thesis, a correlation-driven algorithm to reduce the dimensionality of a feature space was developed that considers both, feature redundancy and elimination efficiency as described in **chapter 2.9.2**. This algorithm (named ‘NETCORE’) translates the dataset into a correlation network, which is then analyzed by conducting an iterative, three-step decision procedure. With this approach, the algorithm selects a subset of features that represents the full feature space on the basis of a (freely selectable) correlation threshold while taking into account the multi-connectivity of a feature to its neighbors in the correlation network.

To first achieve a step-by-step evaluation of the NETCORE algorithm, a test dataset created for an experimental drug loading study containing different antibiotics is used (see **Appendix A6**). This rather small dataset contains 14 commonly used antibiotics, each characterized by a feature vector summarizing 9 physico-chemical properties of the molecules (**Figure 4.9a**). Since the NETCORE algorithm conducts an output-independent analysis, no further label is required. As described in **chapter 2.9.2**, the available database is first translated into a correlation matrix based on the Pearson’s correlation coefficient  $r$  (**Figure 4.9b**). Here, correlation coefficients of  $r = 1$  denote maximal direct correlation, whereas coefficients of  $r = -1$  denote maximal inverse correlation. From the correlation coefficients calculated for the antibiotics dataset, a first overview over the relations between the features can be obtained. For instance, strong correlations ( $r > 0.8$ ) are present between the rotatable bond count, the van der Waals volume, the molecular weight, and the number of hydrogen bond acceptor sites. In contrast, the dipole moment and the aromatic ring count show no strong correlations to any of the other features.

By transforming the correlation matrix into a correlation network, a much better graphical representation of those relations can be achieved. To do so, a node is created for each of the features, and edges are added if the correlation coefficient between two features equals or exceeds a pre-defined correlation threshold. Selecting this hyperparameter, of course, is not always trivial as it fundamentally determines what correlation strength is regarded as sufficient to consider the information provided by two features to be redundant. For this first analysis, a correlation threshold of  $|r| \geq 0.6$  is chosen, which represents moderate correlation and leads to a well-connected correlation network as depicted in **Figure 4.9c**.

From the generated network, features showing no connection to any other feature can directly be identified as non-redundant. And indeed, for the analyzed antibiotics dataset, the features ‘charge’, ‘dipole moment’, and ‘aromatic ring count’ are not sufficiently represented by any of the other features. Consequently, those features need to be added to the initially empty new feature vector, and any subsequent analysis will be performed on the interconnected network only (**Figure 4.9d**). To identify the next feature to be added to the new feature vector, the node of highest degree centrality is searched; to do so, first, the degree of all nodes is determined (this parameter is defined as the number of neighbors connected to a particular node). However, for the generated network investigated at the moment, there is not a single node of highest degree; instead, four out of six nodes (*i.e.*, those representing the ‘van der Waals volume’, the ‘molecular weight’, the ‘hydrogen bond acceptor sites’, and the ‘hydrogen bond donor

sites) all exhibit a degree of four. When applying the second sorting criterium, the situation does not improve: for this very small network, removing one of the four candidate nodes would always lead to a network comprising one individual node only (see **Figure 4.9e** for a simulation of the respective outcome). Hence, the maximal degree of the nodes of all possibly remaining networks equals 0, which still does not allow for selecting a next feature in an unambiguous manner.



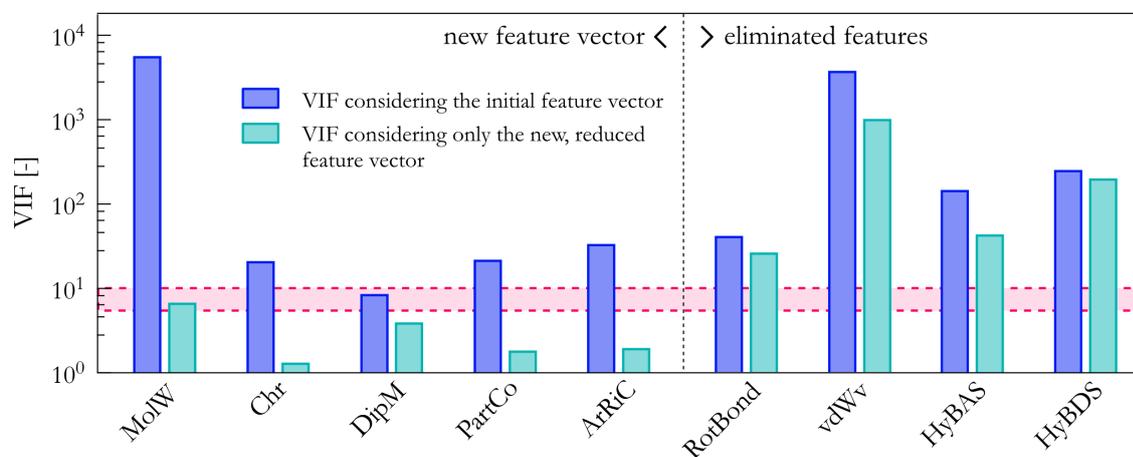
**Figure 4.9: Application of the NETCORE algorithm on a sample dataset generated from 14 antibiotics.**

The initial feature vector comprises 9 physico-chemical properties of the antibiotics (a), which exhibit different levels of correlation to each other as quantified by Pearson's correlation coefficients (b). Based on a correlation threshold of 0.6, a correlation network is generated (c) and iteratively reduced. First, four isolated nodes are eliminated, and the degree of all remaining nodes is determined (d). As four nodes ('vdWv', 'MolW', 'HyBAS', and 'HyBDS') share the highest degree, the four networks that would remain after fixing each of the candidate nodes are simulated (e). As no individual, simulated network comprising an overall node of highest degree can be identified, the mean correlation of each candidate node to its connected neighbors is calculated (f), 'MolW' is fixed, and the remaining node 'PartCo' is added to the new feature vector (g).

## Machine Learning to analyze complex surfaces

Since, in this particular case, neither the first nor the second selection criterion can provide an unambiguous answer, next, the average correlation coefficients of the candidate nodes are determined with respect to their directly connected neighbors (**Figure 4.9f**). Now, the algorithm finds that the feature ‘molecular weight’ has the highest average correlation strength among the candidate nodes. Consequently, the feature ‘molecular weight’ is added to the new feature vector, and all features directly connected to ‘molecular weight’ (*i.e.*, ‘rotatable bond count’, ‘van der Waals volume’, ‘hydrogen bond acceptor sites’, and ‘hydrogen bond donor sites’) are removed from the network. Having eliminated those four nodes, the network finally only comprises one last node (‘partition coefficient’). Similar to how the isolated nodes from the initial network were handled, also this single remaining node has to be added to the new feature vector (and thus, can be removed from the network). After this step, the remaining correlation network is empty, *i.e.*, all nodes were either added to the new feature vector or eliminated because another (redundant) node was added to the new feature vector. Once this state is reached, the feature elimination procedure is completed and the final, reduced feature vector is established. For the antibiotics dataset analyzed thus far, the reduced feature vector comprises five features: ‘charge’, ‘dipole moment’, ‘aromatic ring count’, ‘molecular weight’, and ‘partition coefficient’ (**Figure 4.9g**). In other words, the four features ‘rotatable bond count’, ‘van der Waals volume’, ‘hydrogen bond acceptor sites’, and ‘hydrogen bond donor sites’ were eliminated as the information they carry was determined to be redundant to those provided by the selected features.

To evaluate the success of the NETCORE algorithm with regard to reducing the redundancy of the feature set, the variance inflation factor (VIF) – a regression-based descriptor that quantifies the multicollinearity of a feature to a set of other features – is calculated (**Figure 4.10**). A VIF of 1 represents the absence of multicollinearity, whereas increasing values indicate increasing levels of multicollinearity. Even though there is no universal cutoff-value for this problem, there is a rule of thumb:<sup>269</sup> the features contained in a feature vector are sufficiently uncorrelated when no feature exhibits a VIF > 5-10.

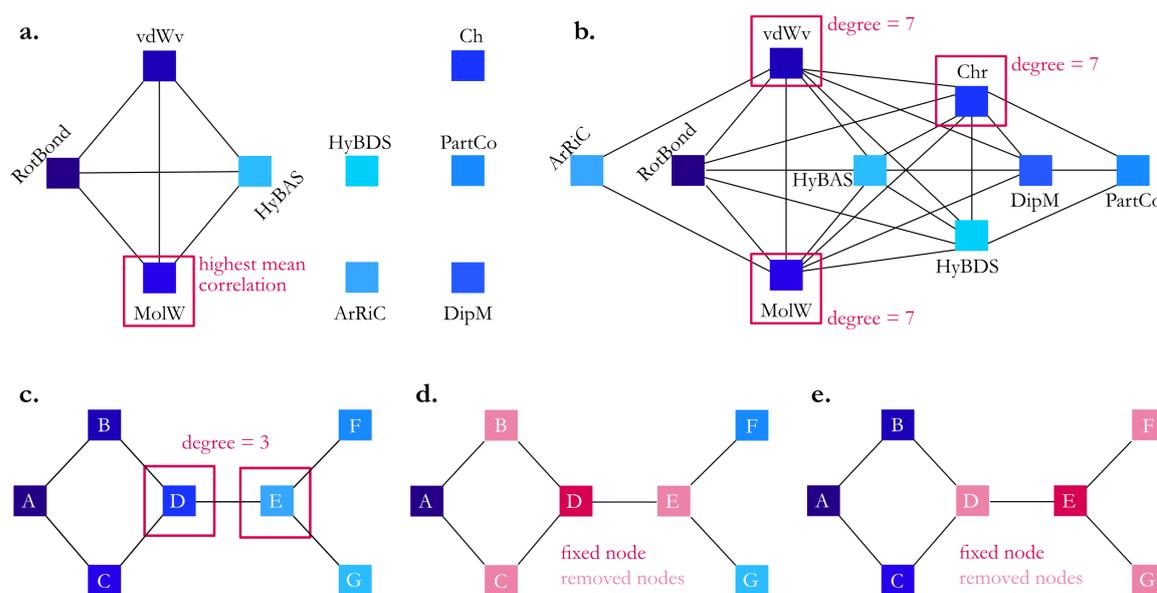


**Figure 4.10: Variance inflation factor (VIF) of all analyzed features before and after reducing the feature vector.** The VIF determined for all features of the initial feature vector (blue bars) is compared to the VIF of all features in the new, reduced feature vector (cyan bars, left side). Additionally, the VIF of the eliminated features was calculated based on the new, reduced feature vector (cyan bars, right side). The red zone denotes the threshold between moderate and strong multicollinearity as defined by Craney *et al.*<sup>269</sup>

In contrast, after conducting the NETCORE procedure (**Figure 4.10**, cyan bars), a completely different picture is obtained: now, the five features included in the new feature vector show low-to-moderate multicollinearity between each other. This confirms that, indeed, a set of uncorrelated features was selected by the NETCORE algorithm. Importantly, the four eliminated features exhibit high

multicollinearity to the features contained in the new, reduced feature vector. This underscores that the eliminated features are sufficiently represented by the selected features of the reduced feature vector. In other words, two targeted properties were indeed achieved: first, a feature vector was successfully created that contains as little redundancy as possible; second, this reduced feature vector still adequately represents the eliminated information.

Having shown that the NETCORE algorithm is indeed capable of condensing an initial feature vector to a set of rather uncorrelated features, next, the three individual decision criteria applied by the algorithm are inspected in more detail. Therefore, networks with different grades of interconnectivity are created by adjusting the correlation threshold  $t$  to either  $t = 0.8$  or to  $t = 0.3$ ; with this approach, a sparsely and a strongly connected network, respectively, is obtained (see **Figure 4.11a,b**).



**Figure 4.11: Illustration of the three decision criteria employed by the NETCORE algorithm.** Based on the previously introduced correlation matrix (**Figure 4.9b**), two correlation networks were generated by applying a correlation threshold of  $t = 0.8$  (**a**) and  $t = 0.3$  (**b**), respectively. The sparsely connected network (**a**) illustrates the importance of decision criterion 3 (highest mean correlation), whereas the densely connected network (**b**) demonstrates the application of decision criterion 1 (highest degree). Additionally, a third correlation network was artificially created (**c**) to illustrate a situation where decision criterion 2 (highest degree in the remaining networks) is required. Here, when eliminating candidate node D, three individual nodes remain (**d**), whereas three connected nodes remain when candidate node E is eliminated (**e**).

For the sparsely connected network (**Figure 4.11a**), all isolated nodes can be directly added to the new feature vector. In this particular scenario, however, the important step is the analysis of the four connected nodes: here, the algorithm detects a cluster of correlating nodes and selects the one with the highest mean correlation strength to the other three connected nodes (criterion 3; for details, see **chapter 2.9.2**). This decision strategy prevents a random feature to be picked; instead, the feature that best represents the eliminated nodes is selected and added to the new feature vector.

In contrast, for the strongly interconnected network (**Figure 4.11b**), the key elimination criterion is a different one: Here, to ensure that the most efficient reduction of the network is achieved, the node with the highest degree, *i.e.*, the node with the highest number of connected neighbors in the network, is identified (criterion 1). In fact, there are three candidate nodes for this step (*i.e.*, ‘vdWv’, ‘molW’, and

## Machine Learning to analyze complex surfaces

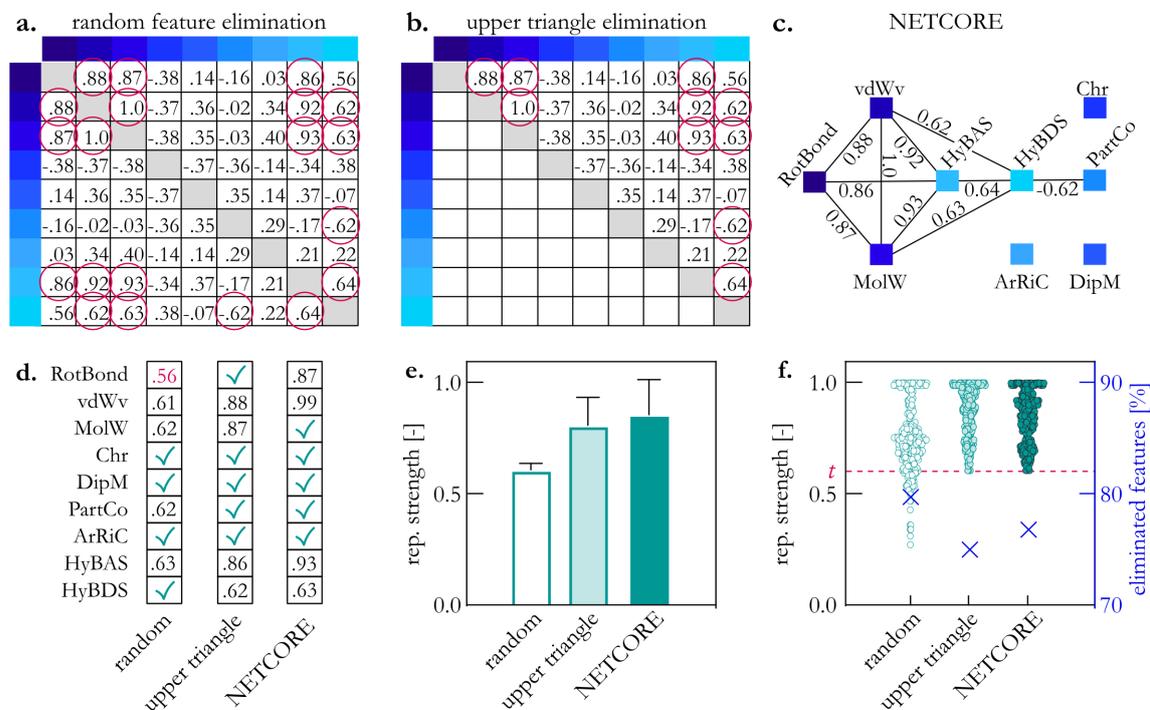
‘Chr’) that all exhibit a degree of 7. In other words, adding one of those features to the new feature vector would allow for removing 8 features (the candidate node and its directly connected neighbors) from the network. This degree-based selection ensures that as many features as possible are represented by the chosen feature. However, to unambiguously decide which of the identified candidate nodes to pick, the next two selection criteria still need to be applied: In each case, adding one of those candidate nodes to the new feature vector would lead to a network comprising one node only (criterion 2); thus, the mean correlation strength needs to be calculated (criterion 3). When doing so, ‘MolW’ shows the highest mean correlation among those three candidates, which is why this particular feature is added to the feature vector. Then, in a last step, only one node remains, which concludes the selection process.

So far, mainly the importance (and effect) of decisions made based on criteria 1 (highest degree) and 3 (highest mean correlation) were demonstrated. However, to demonstrate the significance of criterion 2 (highest degree in the remaining network), next, the artificially generated network depicted in **Figure 4.11c** is analyzed. In this network, there are two nodes that share the highest degree (= 3). However, fixing one of these nodes entails a completely different remaining correlation network: When adding node ‘D’ to the new feature vector, three individual nodes (‘A’, ‘F’, and ‘G’) are left (**Figure 4.11d**). Since those nodes do not have any connection to other nodes, it would be mandatory to include the corresponding features into the new feature vector. Hence, when choosing node ‘D’ to be fixed, the final feature vector inevitably contains 4 features to properly represent all initial features. This, however, can be avoided by fixing node ‘E’ instead (**Figure 4.11e**): by doing so, no isolated nodes remain. As the three remaining nodes (‘A’, ‘B’, and ‘C’) are connected to each other, they can all be represented by adding node ‘A’ to the new feature vector. With this choice, the final, reduced feature vector consists of 2 features only (compared to 4 entries which are obtained if node ‘D’ would have been fixed). Hence, analyzing the networks that would remain after eliminating a candidate node helps as it avoids an unfavorable segmentation of the network into disjunct subnetworks or isolated nodes; this procedure entails a more efficient reduction of the feature space, as more features can still be represented by their connected nodes.

The advantages brought about by the three-step decision procedure implemented in NETCORE become especially apparent when comparing the obtained elimination results to those obtained with simpler correlation-based strategies (such a comparison is described in detail in **Appendix A7**; an overview of the basic principles of the three compared methods is depicted in **Figure 4.12a-c**). All elimination strategies were tested with the previously introduced antibiotics dataset, 4 additional small molecular datasets and one big ‘inhibitors’ dataset obtained from Wu *et al.*<sup>270</sup> (containing 1513 molecular samples characterized by 590 features). In brief, the first tested rather simple correlation-based feature elimination method is to calculate a correlation matrix (the same as used for the NETCORE algorithm) and to search each column (corresponding to one feature) for correlation coefficients that equal or exceed a predefined correlation threshold (‘random feature elimination’, **Figure 4.12a**). If such coefficients are detected, the column (*i.e.*, the feature) and its corresponding row are removed from the correlation matrix. Once all columns have been checked, only those features remain in the correlation matrix, that do not exhibit a high-enough correlation to any other remaining feature.

To assess the quality of the feature elimination achieved with this simple algorithm, the ‘representation strength’ is calculated; the representation strength of an eliminated feature corresponds to the maximal correlation strength this particular feature has to any of the features contained in the new, reduced feature vector. When conducting the random feature elimination with an exemplary correlation threshold of  $t = 0.6$ , for each dataset, the representation strength of one or more eliminated features is lower than this applied threshold, *i.e.*, those features are not sufficiently represented anymore (results obtained for the antibiotics dataset and the big inhibitor dataset are depicted in **Figure 4.12d,e** and **Figure 4.12c**,

respectively; for results on all other datasets, please refer to **Appendix A6**). This problem mainly arises due to the fact that a feature that was previously chosen to represent an eliminated feature can afterwards be eliminated itself. In contrast, this problem is prevented by the NETCORE algorithm: here, a reduced feature vector is created that represents all eliminated features with representation strengths that equal or exceed the predefined correlation threshold (see **Figure 4.12a-c** and **Appendix A7**).

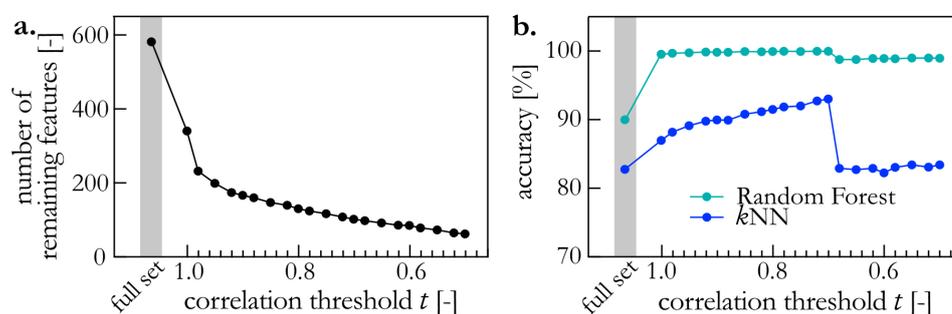


**Figure 4.12: Comparison of different correlation-based feature elimination strategies.** The decision bases of three tested correlation-driven feature elimination strategies are displayed (**a-c**): for the ‘random’ feature elimination, all feature columns of the correlation matrix are searched for correlation coefficients that exceed a certain threshold (**a**); for the ‘upper triangle’ method, the same procedure is conducted but only the upper triangle of the correlation matrix is used (**b**); for the NETCORE algorithm, a correlation network is generated, which is used to eliminate features based on feature centrality (**c**). The detailed elimination result achieved with these three approaches applied to the antibiotics dataset is displayed (**d**): here, green ticks denote features that are maintained in the new, reduced feature vector, and numbers quantify the representation strength of the eliminated features (*i.e.*, the maximal correlation strength this particular feature has to any of the features contained in the new, reduced feature vector). Mean representation strengths achieved for the three strategies are depicted in subfigure (**e**). Error bars denote the standard deviation obtained from the different numbers  $n$  of the eliminated features (according to **d**,  $n \geq 4$  for all bars). Last, the elimination results (*i.e.*, representation strengths and proportion of eliminated features) obtained with the three strategies applied to the big ‘inhibitor’ dataset are displayed (**f**).

A slight modification to the random correlation-based feature elimination can help to avoid dropping features that are needed to represent others: Instead of analyzing the whole correlation matrix, searching only the upper triangle of the correlation matrix ensures that features are only eliminated if they are sufficiently correlated to a feature that is kept in the reduced feature vector (**Figure 4.12b**). And indeed, this method is able to identify a reduced feature vector that properly represents all eliminated features. However, even though this ‘upper triangle’ method can identify a suitable feature vector, its selection result is sub-optimal: either the mean representation strength of the eliminated features is lower than

that achieved with the NETCORE algorithm (**Figure 4.12e** and **Appendix A7**), or the number of features needed to represent all other features is higher, *i.e.*, the elimination efficiency is lower (**Figure 4.12f**). This sub-optimality probably arises due to certain limitations of the upper triangle method: A feature can only be eliminated if it is redundant to a feature that is described by a column of the correlation matrix which is located lefthand of the feature to be eliminated. This, in turn, entails two major complications: First, the feature elimination process is subject to a certain bias. Features that are located in the ‘beginning’ of the correlation matrix tend to stay in the feature vector, whereas features that are located in ‘later’ columns are more likely to be discarded. Second, the reduced feature vector obtained from the upper triangle method strongly depends on the order by which the features appear in the correlation matrix. Thus, certain constellations in the correlation matrix also allow only for certain eliminations, and this limits the identification of an ‘optimal’ reduced feature vector. In contrast, neither of those limitations applies to the NETCORE algorithm, which is why here, a more efficient feature elimination is achieved.

Having shown the efficiency of the NETCORE algorithm, in a next set of trials, the impact of the achieved feature reduction on the sorting accuracy achieved with two common ML methods (a Random Forest classifier and a  $k$  Nearest Neighbors classifier) is tested. As a labelled dataset is required to allow for evaluating the obtained accuracies, the previously introduced ‘inhibitors’ dataset was used that contains binary labels of binding results for 1513 (putative) inhibitors of human  $\beta$ -secretase 1; here, each inhibitor molecule is characterized by 590 physico-chemical features. When applying the NETCORE strategy to this data set, the feature vector can be gradually reduced by adjusting the correlation threshold (**Figure 4.13a**): starting with a correlation threshold of 1.0 (here, only fully correlating features and those features that cannot be correlated are excluded) the number of features in the new feature vector continuously decreases until only 61 features are left when a correlation strength of 0.5 is chosen.

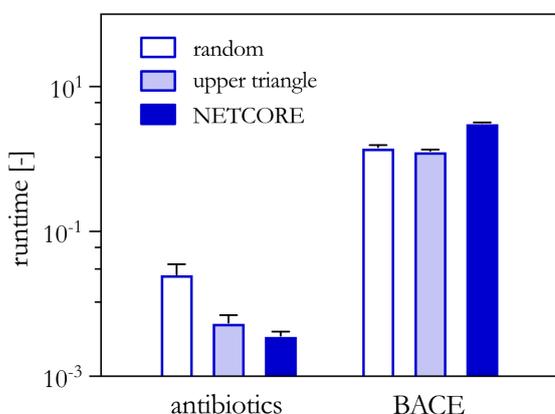


**Figure 4.13: NETCORE to preprocess data for supervised ML.** The NETCORE algorithm was applied to a dataset comprising 590 features (full set). By decreasing the correlation threshold  $t$ , the number of features in the new feature vector is reduced (**a**). The new, reduced feature vectors are then used to train and test an RF classifier and a  $k$ NN classifier; both classifiers were tasked to predict the binary label of a molecule (**b**). The displayed values denote the mean accuracy obtained from a repeated ( $n = 10$ ) stratified  $k$ -fold ( $k = 5$ ) cross-validation. Error bars represent the standard error of the mean (as determined from those 50 total runs) but are smaller than the symbol size. The mean VIF of features of the initial dataset (containing only features that can be correlated; *i.e.*, features that have the same value for all samples were already discarded) is compared to the mean VIF of the features included in reduced feature vectors that were generated by performing the NETCORE algorithm with a correlation threshold of either  $t = 0.6$  or  $t = 0.7$  (**c**).

When using this molecular dataset for training and testing the RF and the  $k$ NN classifier with the aim to predict the binary binding label of each sample based on the provided feature vector, a change in the accuracy of both classifiers is observed (**Figure 4.13b**). The RF model mainly profits from removing fully redundant information (and those features that cannot be correlated): When applying a feature

elimination based on a correlation threshold of 1.0 (by which the size of the feature vector is reduced from 592 to 340 features), the prediction accuracy increases from approximately 90 % to slightly over 99 %. For the  $k$ NN (which is a classifier that is known to perform badly when challenged with data of high dimensionality) an almost linear increase in accuracy is observed as the feature vector becomes smaller. However, once a correlation threshold of 0.7 is reached (which is commonly used as a threshold indicating a sufficiently strong level of correlation), a maximum in the accuracy is obtained. In other words, applying the NETCORE algorithm can considerably improve the performance of either ML classifier.

In conclusion, the NETCORE algorithm is able to outperform basic correlation-based feature elimination strategies as it more efficiently reduces the feature space while optimizing the representation strength of the eliminated features. Additionally, the obtained results also nicely demonstrate, that the NETCORE algorithm can be easily scaled to datasets that contain both, higher numbers of samples and higher feature dimensionality (as it was successfully used to analyze a dataset containing 1513 samples characterized by 590 physico-chemical properties). With an increasing size of the dataset, the only limiting factor might become the runtime. The runtimes required to analyze the small antibiotics dataset studied here is in the range of several milliseconds only (measured with the ‘times’ module of python, **Figure 4.14**). With an increasing size of the dataset, this runtime increases, of course. However, even for the big inhibitor dataset, the full NETCORE procedure is conducted within approximately 3 seconds (when running the NETCORE script on a MacBook Pro 2017 equipped with a 3.1 GHz Dual-Core Intel Core i5 processor), which can be considered very reasonable for such a feature elimination task. Compared to the simple correlation-based methods, the runtime of the NETCORE algorithm required to analyze the big dataset is slightly higher than that of the other two approaches. For a small dataset, however, the NETCORE algorithm is even the fastest among the three tested methods (**Figure 4.14**).

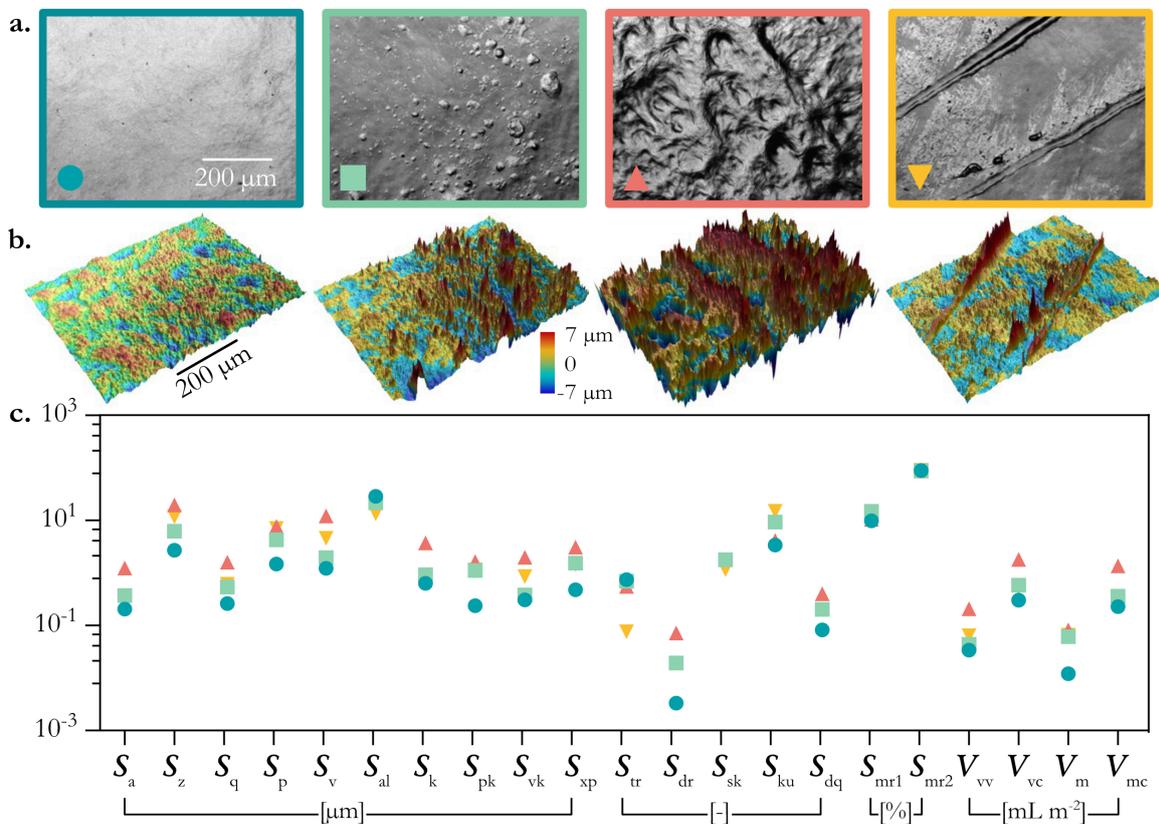


**Figure 4.14: Runtimes of different correlation-based algorithms.** The antibiotics dataset and the big BACE dataset were analyzed with the ‘random’ and the ‘upper triangle’ method, as well as with the NETCORE algorithm, and runtimes were measured using the ‘times’ model of python.

Overall, in addition to providing a fast and straight-forward method for reducing the dimensionality of a feature space, the NETCORE algorithm can help gaining deeper insights into the analyzed data: The generated new feature vector basically provides a ranking of the remaining features according to the centrality of a feature in the correlation network, and the decision-making process can be traced very easily. As a result, redundant information is not only eliminated from the feature vector, but can also be uncovered and interpreted. Thus, the NETCORE strategy developed here has the potential to be a highly beneficial component for various Machine Learning pathways.

### 4.3. Analyzing corneal tissue *via* unsupervised Machine Learning\*

Having shown that a meaningful surface classification is indeed possible based on a provided set of surface parameters (**chapter 4.1**) and having described a correlation-driven strategy for feature elimination (**chapter 4.2**), next, unsupervised ML models are employed to analyze a dataset comprising corneal images. In a first step towards an ML-based assessment of the surface quality of those corneal samples, digital representations of the corneal surfaces are generated by capturing profilometric images as described in the **chapter 2.6** (exemplary images are depicted in **Figure 4.15a,b**). From those 3-dimensional, topographical images (**Figure 4.15b**), descriptive features quantifying the surface appearance of the samples are obtained by calculating 21 metrological parameters according to ISO 25178-2 (**Figure 4.15c**). This set of parameters covers a broad range of surface characteristics including the height distribution, spatiality, texture, and volumetric properties. From the nearly 2800 acquired corneal images, four exemplary 2D images, the corresponding 3D height profiles, and the surface parameters calculated from the latter are depicted in **Figure 4.15**.



**Figure 4.15: Exemplary 2D images, 3D profiles, and surface parameters of corneal surfaces with different surface appearances.** Microscopic, two-dimensional images of corneal surfaces (a), the corresponding profilometric height profiles (b), and metrological surface parameters (c) as determined from the latter are shown. The scale bars in (a,b) apply to all four images within the respective subfigure.

From those four examples, the first sample (dark green circles) represents an intact corneal surface with no obvious imperfections, whereas the other three tissue surfaces show multiple local defects (light green squares), strong global damage (red upside triangles), and textured damage (yellow downside triangles),

\* This section follows in part the publication Rickert *et al.*, ‘Unsupervised Machine Learning to Topographically Analyze Corneal Tissue Surfaces’, submitted 2023

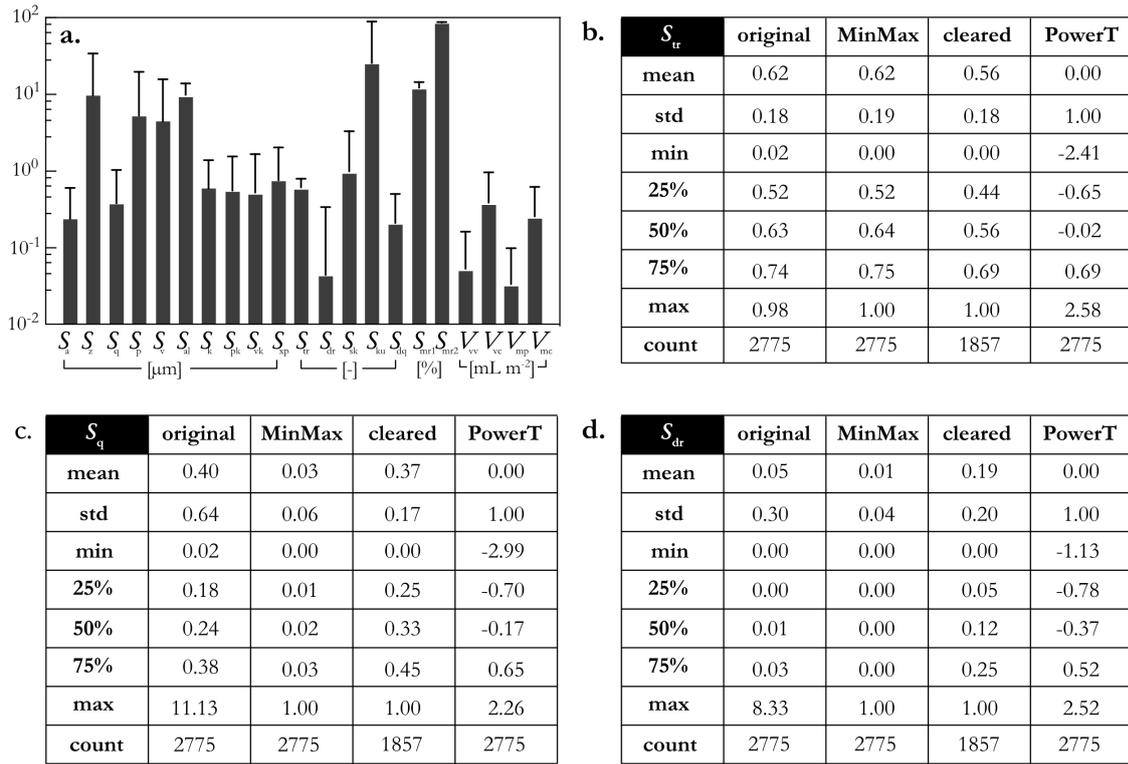
respectively. To a certain degree, this difference in surface appearance is also reflected by several topographical parameters: when quantifying the undamaged tissue surface, it seems that smaller values are obtained for most parameters than for the globally damage samples. However, for less distinct damage patterns, such as local defects (light green) or textured damage (yellow), the parametrical characterization result becomes more ambiguous.

This ambiguity becomes even stronger when the entire dataset is inspected: the visual appearance of the corneal surfaces is very heterogeneous with a much broader range of surface morphologies than what is covered by the four examples; accordingly, there is no obvious correlation between the metrological surface parameters calculated from the images and certain damage patterns. Owing to this high level of variability, a simple, manual categorization of the surface quality of the tissue samples based on the visual damage pattern (and this, of course, includes the absence of any damage) is far from trivial. Therefore, in a next step, unsupervised clustering algorithms are employed to automatically group the samples based on the set of metrological surface parameters. In detail, a cluster-to-label approach is attempted using the 21 surface parameters listed in **Figure 4.15c** as input features. As the  $k$  Nearest Neighbors algorithm was found to be well-suited for classifying samples based on those surface parameters (see **chapter 4.1**), its unsupervised counterpart, the  $k$ -Means clustering algorithm, is chosen as the starting point.

Before doing so, however, basic data curation has to be conducted as the dataset was *per se* not created to be analyzed with ML methods but rather contains experimental raw data; accordingly, as already observed for the four exemplary samples shown in **Figure 4.15**, the sample set exhibits a high degree of variability (which is typical when dealing with a biomedical problem). Hence, to ensure that the clustering algorithm can actually produce meaningful results, the full dataset is first inspected in more detail (**Figure 4.16**). As anticipated based on the four examples shown in **Figure 4.15**, the scales of the individual features exhibit a considerable degree of heterogeneity (for simplicity, units are neglected in the following): for instance, for the developed interfacial ratio  $S_{dr}$ , the average over all analyzed samples returns a value below 0.1; in contrast, for several other parameters, such as the surface curtosis  $S_{ku}$  or the upper material ratio  $S_{mr1}$ , mean values above 10 and even close to 100 are obtained (**Figure 4.16a**). This is problematic for most clustering algorithms, as features with a larger scale will dominate the distance calculation and thus will also dominate the clustering result.

To overcome this, a common approach is to normalize the features such that they have a similar scale before applying clustering algorithms; this can, for instance, be achieved by employing a Minmax-scaler. Here, the original feature values are mapped onto a range from 0 to 1, which then represent the minimal and maximal value of the respective feature, respectively. To evaluate the result of this transformation and to get a better understanding of the features included into the dataset, different statistical descriptors of the feature distributions are calculated and compared: the mean and the standard deviation, the minimal and maximal values of each feature, and the 25<sup>th</sup>, 50<sup>th</sup>, and 75<sup>th</sup> percentile (**Figure 4.16b-d** and **Appendix A8.4**). When applying a Minmax-scaler to the raw features, some features (such as the  $S_{tr}$  parameter shown in **Figure 4.16b**) indeed exhibit a desirable distribution: the mean is located somewhere around the middle of the data range and the standard deviation indicates a reasonable level of variance. This ‘desirable’ distribution is also reflected by the lower, median, and upper percentiles where the normalized feature values fill a significant proportion of the available range from 0 to 1. However, for the majority of features, an anomaly is observed: As exemplarily shown for the root mean square height  $S_q$  (**Figure 4.16c**), the mean values and the percentile values are often very low – even though they were scaled to a predefined range. In detail, for the  $S_q$  feature, the observed mean value is only 0.03; and more importantly, the 75<sup>th</sup> percentile is also only 0.03 – *i.e.*, 75 % of the  $S_q$  values are equal or below this value. An even more pronounced example is the  $S_{dr}$  feature (**Figure 4.16d**), where the majority of values is located very close to zero.

## Machine Learning to analyze complex surfaces

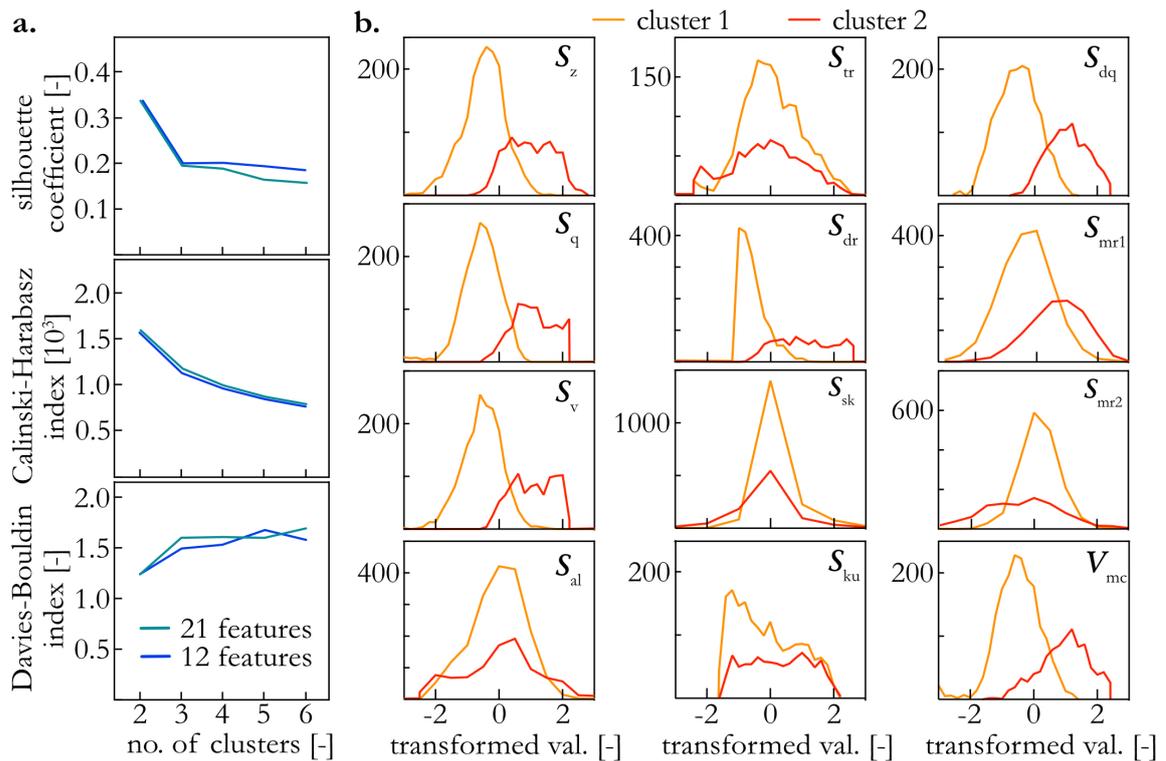


**Figure 4.16: Preprocessing results of the dataset.** The mean values of all features (surface parameters) of the dataset are depicted in subfigure (a). Error bars denote the standard deviation obtained from  $n = 2775$  samples. More detailed information about the data distribution of the original dataset and after applying three different preprocessing strategies is displayed for the individual features ‘surface texture  $S_{tr}$ ’ (b), the ‘mean square height  $S_q$ ’ (c), and the ‘developed interfacial area  $S_{dr}$ ’ (d). Here, ‘MinMax’ corresponds to a Minmax scaler that scales the feature distribution to a range of 0 to 1 based on the maximal and minimal values present in the original feature distribution. Next, ‘cleared’ refers to a Minmax scaler applied to a dataset that was previously cleared of outliers (based on an inter quantile range [IQR] filter). Last, ‘PowerT’ refers to the distributions obtained by applying a Yeo-Johnson Power Transformer.

Distributions like those found for  $S_q$  and  $S_{dr}$  mainly occur due to the presence of outliers, that are so far away from the mean that they artificially compress the actual value range of the feature. To mitigate this problem, two different variants of data preprocessing are evaluated: First, the same Minmax scaler as used before is employed, but – before applying the scaler – all outliers are eliminated based on the inter quantile range of the feature distribution (for details, see the **chapter 2.10.2**). As depicted in **Figure 4.16b-d** (‘cleared’), this strategy indeed improves the data distribution of the features (for all other parameters, see **Appendix A8.4**). However, this improvement comes at high cost: 918 out of 2775 samples have to be discarded from the dataset, which corresponds to approximately one third of the datapoints. This high number of ‘outliers’ indicates that the dataset itself is imbalanced. As eliminating such a significant proportion of the samples is not an option, a second approach is chosen that avoids removing outliers from the dataset: a Yeo-Johnson Power Transformer. This method employs a parametric, monotonic transformation to map any initial distribution onto a nearly Gaussian distribution with a mean value of zero and homoscedasticity between all features. As depicted in **Figure 4.16b-d** (‘PowerT’), this preprocessing strategy indeed leads to a desired shape of the data distributions: now all features are of the same scale and the feature distributions sufficiently cover the provided value range; moreover, all 2775 samples are kept in the dataset. Hence, for all further analysis, this transformed dataset is used.

Having found a suitable method to preprocess the imbalanced dataset, now the clustering of the samples is commenced. One pre-requisite of a  $k$ -Means clustering algorithm is that the number  $k$  of clusters to be formed needs to be selected before the algorithm can be run. A commonly used approach to identify the optimal  $k$  value is the ‘elbow method’:<sup>271</sup> Here,  $k$  is varied from 1 to 10, and the within-cluster sum of squares (WCSS) is calculated. With increasing  $k$  value, the number of samples contained in each cluster decreases, and the samples are located closer to their respective cluster centroid; consequently, the WCSS decreases. By assuming that the obtained WCSS plot resembles an arm, the ideal number of clusters is then given as the point of highest inflection (*i.e.*, the location of the ‘elbow’). When calculating and analyzing the WCSS curve obtained for a basic  $k$ -Means clustering algorithm applied to the introduced dataset (**Appendix A8.5**, cyan), the highest inflection is observed for a  $k$ -value of 2. Hence, according to this method, the ideal number of clusters appears to be  $k = 2$ .

To confirm this result, three different validation metrics are chosen: first, the silhouette coefficient (SIH), which quantifies the separation quality achieved by the clustering; second, the Calinski-Harabasz index<sup>134</sup> (CH), which describes the cluster dispersions; and third, the Davies-Bouldin index<sup>135</sup> (DB), which quantifies the average ‘similarity’ between  $k$  clusters. These 3 validation metrics are then calculated for the clustering results achieved with the previously identified ideal number of  $k = 2$  clusters, as well as for clustering results obtained for 3 to 6 clusters, respectively (**Figure 4.17a**, cyan). Interestingly, for all of those parameters, the same tendency is shown: subdividing the dataset into 2 clusters returns the best values for all metrics (*i.e.*, the highest values for the SIH and the CH, and low values for the DB). Hence, based on the metrics chosen, forming 2 clusters indeed returns the best clustering result.



**Figure 4.17:  $k$ -Means clustering and its validation.** Three different validation metrics were calculated to analyze the clustering results achieved with a  $k$ -Means clustering algorithm (**a**). All data is presented for clustering conducted either in the full 21-dimensional feature space (cyan) or in the reduced feature space (12 dimensions, blue). The frequency distribution of the features (per cluster) as obtained from a  $k$ -Means clustering algorithm employed to form 2 clusters in the 12-dimensional feature space is depicted in (**b**). All  $y$ -axes represent the frequency (starting from 0). The displayed  $x$ -axes apply to all graphs of the respective column.

## Machine Learning to analyze complex surfaces

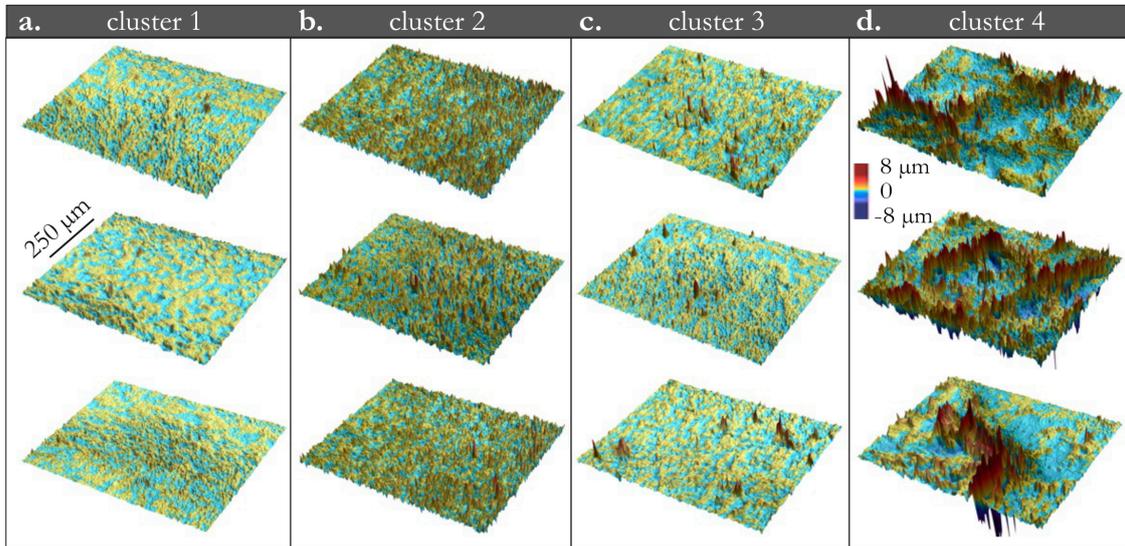
One last aspect to consider is that the currently analyzed 21 descriptors by definition comprise a certain degree of multicollinearity; and it is a known phenomenon that unnecessary dimensions of the feature space can reduce the accuracy of a ML model.<sup>272</sup> Hence, it is next tested if the feature space can be reduced while maintaining or even improving the clustering performance. To eliminate redundant features from the input space, the previously developed, correlation-driven NETCORE algorithm is employed (for the generated correlation matrix, please refer to **Appendix A8.6**). With this elimination strategy, the feature space is reduced from 21 surface parameters to 12 remaining features (a detailed overview of the features and the new WCSS curve can be found in **Appendix A8.7** and **A8.5**, respectively). Interestingly, when recalculating the previously used set of evaluation metrics, similar or even slightly better values are obtained for the reduced feature set (**Figure 4.17a**, blue). Thus, as the removed features seem to be unnecessary for the clustering result, from now on, the reduced feature set is used to characterize the samples. However, also when using this reduced feature set, forming 2 clusters still seems to give the best results (based on the given metrics).

Having prepared the dataset (by preprocessing it using a Power Transformer and reducing the feature space from 21 to 12 features) and having identified a suitable number of clusters ( $k = 2$ ), next, the  $k$ -Means clustering algorithm is employed to group the samples of the cornea dataset, and the obtained frequency distributions of the different features are inspected (**Figure 4.17b**). The obtained curves reflect the different cluster cardinalities: cluster 1 contains approximately 2/3 of the samples, and cluster 2 holds the remaining 1/3 of the samples. Additionally, the distributions give insights about how the algorithm decides which cluster a sample is assigned to. Whereas some features, such as  $S_{al}$ ,  $S_{tr}$ , or  $S_{sk}$  show no obvious differences for samples from the two clusters, for several other features, such as  $S_z$ ,  $S_{dr}$ , and  $S_{dq}$ , cluster 2 contains samples with higher values than samples from cluster 1. Hence, these results indicate that the algorithm automatically performed a basic subdivision into “smooth” and “rough” surfaces (in other words, into “intact” and “damaged”) similar to how it was manually conducted to distinguish between good and damaged tissues in **chapters 3.2** and **3.3**. Interestingly, the surface parameters used to sort samples into these two clusters are mainly those that are most often used by researchers to quantify the surface roughness of a sample.

Although the subdivision into two clusters may seem simple, it already represents a significant improvement over the manual classification conducted before: the clustering algorithm always considers the full ensemble of provided surface parameters, and no actual expert involvement (hence, no subjective assessment of which range of surfaces values qualifies a sample to be ‘smooth’) was necessary to achieve this result. It is important to note, however, that even though the calculated metrics suggest that the optimal outcome is achieved with two clusters, this may not hold true from a materials science perspective, where sorting the data into more categories might provide deeper insights. Also, the cluster metrics are good first indicators to assess the cluster quality, but they assume round-shaped and well-separated clusters rather than overall densely or asymmetrically populated dataspace. Hence, it is next investigated if a higher number of clusters could deliver more distinct results.

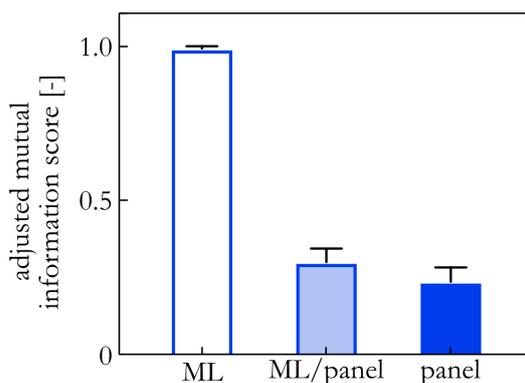
For this purpose,  $k = 4$  is selected, which corresponds to the second small ‘elbow’ observed in the WCSS curve (**Appendix A8.5**, blue). When doing so, the parameter distributions of the formed clusters reach a complexity level that is much more difficult to interpret than it was for 2 clusters (see **Appendix A8.8**). Thus, to evaluate the new clustering result, it seems more feasible to inspect the actual topographical pictures of the clustered corneal tissue samples, *i.e.*, to identify and compare the most central samples of each cluster. And indeed, the four clusters formed based on 12 different surface parameters comprise samples with very similar surface appearances: samples from cluster 1 have a rather flat appearance with shallow micro-wrinkles (**Figure 4.18a**); cluster 2 comprises tissue surfaces with a pronounced but rather uniform micro-roughness (**Figure 4.18b**); the tissue surfaces of cluster 3 are similarly smooth as those

of cluster 1 but they exhibit several local defects (**Figure 4.18c**); surfaces with pronounced roughness features (indicating severely damaged tissue samples) accumulate in cluster 4 (**Figure 4.18d**). Additional images of each cluster are provided in **Appendix A8.9**.



**Figure 4.18: Clustering result achieved with a  $k$ -Means clustering algorithm.** A  $k$ -Means clustering algorithm was employed to group the corneal samples into 4 clusters. For each cluster, the three ‘most central’ samples of each cluster, *i.e.*, those samples that have the smallest Euclidean distance to the respective cluster centroid, are displayed (**a-d**). The scale bars apply for all images.

Having confirmed the meaningfulness of the provided results achieved with the clustering algorithm, the next aim is to evaluate the robustness of this outcome. Therefore, the  $k$ -Means clustering algorithm is run multiple times ( $n = 10$ ) and the clustering results are compared by calculating the adjusted mutual information index (AMI). This measure quantifies the level of agreement between two label assignments: an AMI of 1 indicates perfect agreement between the two clustering results, whereas a score close to 0 describes that mutual information is a result of chance alone. When running the  $k$ -Means clustering algorithm 10 times on the same dataset (but with no predefined centers or given random seeds), an average AMI of 0.99 is achieved (**Figure 4.19**, ML). This result indicates that the cluster assignments are very robust and reproducible with negligible variation.



**Figure 4.19: Comparison of cluster results.** The mean adjusted mutual information scores obtained for the following three scenarios are displayed: first, comparing multiple runs of the  $k$ -Means clustering algorithm (ML); second, comparing those  $k$ -Means clustering results to the assignments provided by the panel (ML/panel); and third, comparing the panel assignments to each other (panel). Error bars denote the standard deviation obtained for  $n \geq 45$  datapoints.

## Machine Learning to analyze complex surfaces

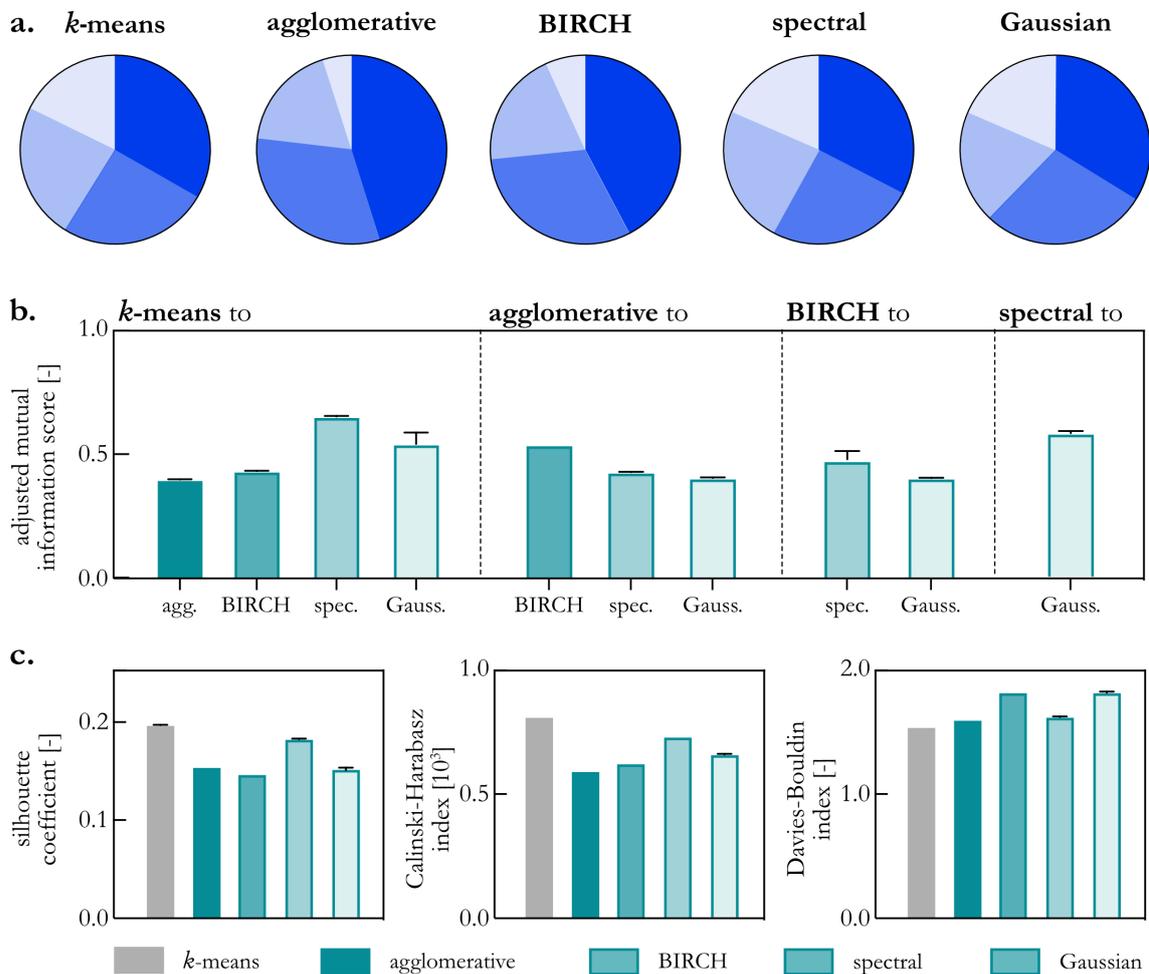
Based on the inspected central samples of each cluster and the demonstrated reproducibility of the results, one might assume that the outcome of the clustering is somewhat trivial. To test this assumption, a panel of 10 participants is enlisted to manually perform the same task, *i.e.*, to cluster the provided dataset into four groups. To compare the label assignments of the panel to those of the clustering algorithm, once again, the AMI is determined. Interestingly, the obtained AMI values are slightly below 0.3 (**Figure 4.19**, ML/panel), which is comparably low given that a score of 0 indicates mutual information achieved based on chance alone. This suggests that there is only moderate agreement between the results delivered by the clustering algorithm and those provided by the panel.

A potential explanation for this discrepancy could be that the algorithm might have failed to perceive patterns that are easily recognizable by humans. However, even among the panel participants, the clustering results show a high inconsistency: as indicated by AMI values of around 0.24, the agreement between the members of the panel is even lower than that between the panel and the algorithm (**Figure 4.19**, panel). Conclusively, there seems to be a significant amount of subjectivity involved in the human process of clustering the corneal surfaces and there is no clear consensus among the panelists on how to group these samples. An additional difficulty could arise from putative inconsistencies of each individual's evaluation: Even though the panel was presented a severely reduced dataset (397 of 2775 samples), 2 out of 10 participants categorized the two included duplicate samples into different clusters.

Overall, these findings indicate that, for complex datasets like the one analyzed in this study, an ML-based strategy can outperform a human assessment: the  $k$ -Means clustering approach offers a dependable and reproducible classification, whereas the manual clustering conducted by panelists is heavily reliant on subjective interpretation. This highlights the potential of Machine Learning approaches to guide and improve the analysis of datasets that – either owing to their sample size, or their complexity – go beyond the cognitive capabilities of humans.

Of course, the  $k$ -Means clustering algorithm used so far is only one example of many clustering strategies available to perform such a task. A similarly simple approach is the Mean Shift clustering algorithm that iteratively shifts cluster centroids towards the densest areas of the dataset. One advantage of the Mean Shift algorithm is that the number of clusters to be formed is automatically determined. Additionally, the clustering result can be optimized by adjusting the bandwidth (BW) hyperparameter, which describes how far the cluster centroids are shifted based on individual samples: A larger BW results in fewer and larger clusters, whereas a smaller BW results in multiple smaller clusters. However, when analyzing the cornea dataset with a Mean Shift algorithm, it is observed that the algorithm fails at generating meaningful clusters for all BW-values tested (for details, please refer to **Appendix A8.1**, ‘Mean Shift clustering’): for a small BW of 0.1, each sample constitutes an individual cluster; and for a large BW of 10, all samples are sorted into one cluster only. Furthermore, for intermediate BW values, either still too many clusters are formed (BW: 2, number of clusters formed: 64), or an acceptable number of clusters is achieved (10 and 4 clusters for BW values of 3 and 4, respectively), but more than 95 % of the samples are assigned to the same cluster. At this point, it is important to realize that, even though  $k$ -Means clustering and Mean Shift clustering might appear to be closely related, they differ in one important aspect: whereas the  $k$ -Means algorithm performs clustering based on distances, the Mean Shift algorithm clusters based on densities. Given the rather low silhouette coefficients obtained (**Figure 4.17a**), it can be assumed that the dataset exhibits a fairly uniform distribution across the feature space rather than having prominent density peaks. Hence, the density-based Mean Shift algorithm fails at identifying sufficient criteria to create well-separated clusters. Consequently, the Mean Shift algorithm is not considered further to analyze the cornea dataset. This example highlights the impact of the underlying data distribution on the algorithm performance and underscores the necessity of carefully selecting an appropriate clustering algorithm for the specific problem at hand.

However, four other clustering methods return reasonable cardinalities when tasked to form 4 clusters from the dataset (**Figure 4.20a**): First, Agglomerative Clustering, which follows a bottom-to-top approach by first assigning an individual cluster to each sample and then merging similar clusters until the desired number of clusters is reached ('agglomerative'); second, the BIRCH algorithm, a hierarchical algorithm that organizes data into tree-like structures with defined decision criteria and forms clusters by merging branches (BIRCH); third, a Spectral Clustering technique that uses information from the eigenvalues and eigenvectors of a similarity matrix to group data points into clusters ('spectral'); and fourth, a Gaussian Mixture algorithm – a probabilistic clustering technique that models the distribution of data points as a mixture of Gaussian distributions, and then estimates the parameters of these distributions to cluster the data ('Gaussian'). Further details regarding all those algorithms can be found in **Appendix A8.1**.



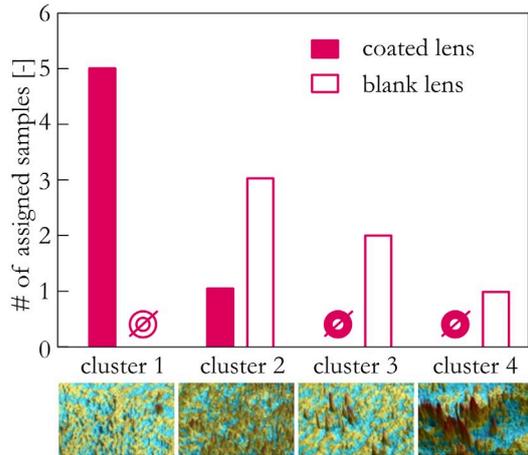
**Figure 4.20: Clustering results obtained with 5 different clustering algorithms.** The corneal dataset (2775 samples, 12 features) was divided into 4 clusters using 5 different clustering algorithms. The obtained label distributions of the 2775 categorized samples (**a**), the adjusted mutual information score between the clustering results (**b**), and the internal calibration metrics obtained for each individual cluster method (**c**) are depicted. Since the AMI is a symmetric score, it is only displayed in one direction. Error bars denote the standard deviation from  $n \geq 45$  (**b**) and  $n = 10$  (**c**) values, respectively. If no error bars are visible, the repetitions of the computations had no or too little variance to be displayed.

## Machine Learning to analyze complex surfaces

To compare the different clustering results returned by those different algorithms, the adjusted mutual information index between the label assignments is calculated (**Figure 4.20b**). As can be expected based on the cardinality charts (**Figure 4.20a**), the highest similarity is observed between the results of the  $k$ -Means algorithm and the Spectral Clustering algorithm. This outcome is highly reasonable, as both algorithms are distance-based: whereas the  $k$ -Means algorithm directly aims at minimizing within-cluster distances while maximizing between-cluster distances, Spectral Clustering groups samples based on an affinity score that is derived using Euclidean distances. Interestingly, the corresponding AMI values are around 0.65, which is rather low considering that the cardinality distributions are virtually identical. This indicates that, even though a similar overarching structure was identified in the dataset, the cluster centroids and the decision criteria still differ to a certain extent. A similar picture emerges when comparing Agglomerative Clustering and the BIRCH algorithm: Even though the working principles of both techniques are quite similar and, also here, the formed clusters are of similar size, the resulting AMI is only slightly above 0.5.

Overall, the moderate agreement observed between all algorithms tested here confirms the previous assumption that there is no unambiguous cluster result. This realization once again demonstrates the necessity to evaluate the cluster results in detail, *e.g.*, by calculating a set of intrinsic validation metrics, by visually inspecting the data distribution of the individual features, and by manually evaluating the most central samples of each cluster. Furthermore, comparing different clustering algorithms and their results can improve the understanding of the underlying data structure, guide the development of new clustering methods, or help to identify the most suitable method from the set of available algorithms. To evaluate if one of the four additional clustering algorithms outperforms the previously employed  $k$ -Means clustering approach in terms of cluster separation, the same set of intrinsic validation metrics mentioned above (*i.e.*, the silhouette coefficient, the Calinski-Harabasz index and the Davies-Bouldin index) is calculated again. Importantly, the obtained results are consistent across all metrics: The best-separated clusters are obtained with the  $k$ -Means clustering algorithm (**Figure 4.20c**). In other words, for this particular dataset, the seemingly simple  $k$ -Means algorithm indeed provides the most efficient and most meaningful result (at least among the options tested here assessed by the given set of evaluation metrics).

In a last step, the same  $k$ -Means model used so far is now tasked to assign 12 additional corneal images derived from the tribological measurements conducted with RGP contact lenses (see **chapter 3.3**) to the four clusters created above (none of those 12 additional samples was used to form these clusters). From those 12 samples, 6 corneal surfaces were subjected to tribological stress created with a blank contact lens, whereas the other 6 samples were subjected to tribological stress induced with a lens carrying a lubricating mucin coating (coated lens). In **chapter 3.3**, the quality of those corneal surfaces was manually assessed based on a subset of surface parameters and a visual inspection of the surface profiles. Based on this assessment, it was concluded that, whereas the coated lenses left the corneal samples mostly intact, the blank lenses induced severe damage to the corneal tissues. Interestingly, a similar picture emerges when automatically classifying these samples using the pre-trained  $k$ -Means clustering algorithm (**Figure 4.21**; for details, please refer to **Appendix A8.10**): the majority of samples treated with a coated lens is sorted into cluster 1 (flat surfaces) and only one sample is assigned to cluster 2 (micro-rough). In contrast, none of the samples treated with the blank lens was classified as ‘flat’. Instead, those corneal surfaces were classified as either micro-rough (cluster 2), locally damaged (cluster 3) or even globally damaged (cluster 4). Hence, for this small subset of corneal tissues, the  $k$ -Means clustering algorithm was able to achieve a similar quality assessment as the manual expert review conducted before. This result underscores the suitability of this ML-approach to guide and support human decision and assessment processes in a reliable and non-subjective manner.



**Figure 4.21: Classification of corneal samples subjected to tribological stress.** 12 corneal samples were subjected to tribological stress induced with a contact lens (6 of those with a blank RGP lens, and the 6 other ones with a lens carrying a lubricating coating, see **chapter 3.3**). The number of samples per condition assigned to each cluster of the pre-trained  $k$ -Means clustering model is displayed. The thumbnails under the cluster names correspond to the most central images of each cluster (as already presented in **Figure 4.18**).

In conclusion, unsupervised clustering methods are suitable approaches to analyze and cluster surfaces based on the provided set of metrological surface parameters. Therefore, different preprocessing strategies can be applied to overcome the challenges encountered due to the sample distribution of the dataset. For the imbalanced cornea dataset, basic Minmax-scaling was not sufficient, and an outlier removal would lead to a too high number of samples to be discarded. Instead, a Power Transformer was able to achieve a desired data distribution for all features. The  $k$ -Means clustering algorithm applied to the preprocessed dataset was then able to sort samples in a meaningful and reproducible way, and it outperformed other unsupervised clustering methods. Overall, the presented results showcase that by combining different strategies from the broad repertoire of ML-related algorithms (including methods for data preprocessing, feature elimination, and clustering) ML-pipelines can be tailored to analyze even complex datasets with the goal to support or even replace human decision processes.



## 5. Summary and outlook\*

The human body, with its remarkable abilities to protect and restore its own health and its intricate mechanisms to respond to environmental stressors, is a true masterpiece of nature. The immune system, for instance, fights against harmful pathogens, the musculoskeletal system can repair and regenerate damaged tissue, and the brain demonstrates an astonishing capacity for adaptation and learning. Yet, when intrinsic functionalities fail due to, *e.g.*, age, injury, or disease, medical intervention may become necessary. Although medical devices, such as implants or instruments, are carefully developed to serve specific beneficial purposes, they may also interfere with the endogenous protective mechanisms; and this can, for instance, trigger immune responses or lead to life-threatening infections. Hence, it remains crucial to devise strategies to reduce putative device-associated complications to enhance the success of patient care.

In this context, natural materials and systems provide a wealth of inspiration to guide the design and adaptation of medical devices such that they work in harmony with the body. One approach is to design materials that mimic biological principles, while another strategy involves directly integrating biological agents into the products. The latter approach is followed in the first part of this thesis using the multifunctional bio-macromolecule mucin. First, a mucin-based bio-ink to be used for 3D printing was developed, and objects generated from this ink were shown to exhibit mechanical properties similar to that of soft tissue (**chapter 3.1**). Moreover, those objects exhibited strong anti-biofouling properties, good biocompatibility, tunable cell adhesion, and immunomodulating behavior. This unique combination of characteristics renders this bio-ink a promising candidate for a wide range of applications, such as tissue engineering and regenerative medicine. Moreover, the ink's tunable cell adhesion and anti-biofouling properties might allow for creating patient-tailored, biocompatible interfaces between living tissue and artificial objects, *e.g.*, prosthetics or implantable medical devices. Additionally, the application potential of such a mucin-based bio-ink could extend to drug delivery and wound healing: not only can the integrated stabilizing agent Laponite promote wound healing by releasing magnesium ions,<sup>154,273</sup> but, even more importantly, previous studies have pinpointed the potential of mucin-based materials as promising vehicles for drug delivery.<sup>152,274</sup> The negative net charge of the mucin molecules allows for creating drug depots with charge-dependent release kinetics: whereas the acting electrostatic forces lead to a quick release of anionic molecules, the release of positively charged cargos is retarded.<sup>152,274</sup> Moreover, some mucin-based drug carriers were successfully designed to release their cargo in response to specific stimuli, such as the exposure to a physiological salt concentration or disease-specific nucleic acids.<sup>204,275,276</sup>

Following the successful development of a mucin-based bulk material, the thesis proceeded to explore the potential benefits of mucin coatings generated on medical devices (**chapters 3.2 and 3.3**). In detail, contact lenses were chosen as one example and it was shown that mucin coatings considerably improve the ability of the lenses to interact with liquid environments (such as the tear film). Remarkably, shear-induced tissue damage inflicted on corneal tissue was strongly reduced or even fully prevented. These results showcase the efficiency of mucin coatings in mitigating the disruptive influence of synthetic medical materials – and this suggest a promising avenue for enhancing the biocompatibility of a broad spectrum of medical devices including tracheal tubes, catheters, or endoscopes. Furthermore,

---

\* This section follows in part the publications Rickert *et al.*, 'A Mucin-based Bio-Ink for 3D Printing of Objects with Anti-Biofouling Properties', submitted 2023; Rickert *et al.*, *Biomaterials Advances* (2022); Rickert *et al.*, *Macromolecular Bioscience* (2020); Rickert *et al.*, *ACS Biomaterials Science & Engineering* (2021); Rickert *et al.*, *APL Machine Learning* (2023); and Rickert *et al.*, 'Unsupervised Machine Learning to Topographically Analyze Corneal Tissue Surfaces', submitted 2023

## Summary and outlook

mucin-coatings could expand the range of materials available for biomedical applications by overcoming the limitations of insufficient surface wettability. Importantly, while improving interactions between medical surfaces and tissue, those mucin coatings still show excellent transparency – and owing to the extraordinary anti-biofouling properties of the mucin layer, this transparency can be maintained even in challenging environments. This particular characteristic can be highly valuable for optical devices that are susceptible to becoming opaque upon contact with tissue or bodily fluids. Hence, in addition to basic contact lenses, these kinds of coatings can also provide benefits for intraocular lenses, such as those used after cataract surgery, or even for other optical devices used in medicine: the functionality of endoscopes, for instance, crucially depends on the transparency of their lenses. On capsule endoscopes, mucin coatings might help maintaining the required transparency of those devices even though they are exposed to a variety of challenging environments when travelling throughout the gastrointestinal system.

Overall, owing to the successful application-oriented integration of mucin molecules into medical devices – including those presented in this work – mucins have gained increasing attention as a versatile and multifunctional material for biomedical purposes. This development is further propelled by the continuously growing understanding of the structure-function relations of mucins and their influence on physiological processes,<sup>72,277,278</sup> recent advancements in mucin purification,<sup>82,279</sup> and the ability to conduct a variety of chemical conjugation and modification methods.<sup>95,280,281</sup> However, to be used in a clinical context, strict requirements regarding sterility have to be met; and for a material that is derived from animal origin, this can be a challenging aspect. To circumvent this problem, recent efforts aimed at synthetically recreating mucin-like polymers to achieve better control over the material.<sup>282</sup> These approaches typically combine small, basic monomers to create synthetic mucin substitutes that imitate the physical characteristics and functions of natural mucin glycoproteins (or specific domains of those). Even though, to a certain degree, individual mucin functionalities can be mimicked with those synthetic polymers,<sup>283-286</sup> it is to date not possible to fully synthesize mucin surrogates that replicate the whole multifunctionality of the glycoprotein. Hence, so far, the only possibility to obtain mucin molecules in significant amounts remains the purification from animal sources. However, those purified mucins – that, even after purification, might still contain microorganisms – need to undergo a sterilization procedure, which could harm the physico-chemical or structural properties of the mucin molecules.

Remarkably, the experiments carried out in the framework of this thesis revealed that mucins are indeed able to withstand several of those harsh decontamination treatments without suffering significant functional impairment or structural degradation (**chapters 3.4** and **3.5**). For the sterilization of mucins in their lyophilized form,  $\gamma$  irradiation was identified to be the most suitable among the tested treatments. The applied  $\gamma$  rays can permanently damage DNA and exhibit a high penetration efficiency across biological matter; thus, they cannot only inactivate bacteria and yeast cells but also (very sturdy) fungal spores can be efficiently killed. From an application point of view, however, instead of sterilizing the mucins first and then processing them under sterile conditions, it might be more practical to sterilize the final mucin-containing product as a whole. As evaluated for mucin coatings generated on three different medical devices, the biochemical integrity and the functionality were maintained best by fumigation with ethylene oxide. This very commonly used chemical sterilization method kills microorganisms by denaturing molecules that are responsible for cell reproduction. Importantly, to minimize the risk of toxic by-products remaining in the material, a suitable aeration strategy should be identified for each individual medical device. Overall, the presented results demonstrate an astonishing sturdiness of mucins towards selected sterilization strategies – a major step towards a clinical application of the glycoprotein.

Whereas the first part of this thesis focused on exploiting natural materials for medical products, it is worth noting that, in addition to guiding materials design, nature also serves as a rich source of inspiration for many other fields of science. Mimicking natural systems that have evolved over billions of years to

solve complex problems has, for instance, revolutionized the development of computational algorithms and models.<sup>287,288</sup> Impressive examples thereof are algorithms that mimic the behavior of ant colonies: models inspired by the ability of ants to optimize complex tasks through collective decision-making emerged as effective tools for solving difficult optimization problems including transportation routing, resource allocation, and network optimization.<sup>289,290</sup> The probably most famous class of bio-inspired algorithms are neural networks, that aim at simulating the structure and functionality of the human brain. Owing to their ability to learn, process, and make predictions on complex data, neural networks have recently gained popularity for various Machine Learning applications. However, in ML, which aims at analyzing datasets to unveil patterns or make inferences on unknown samples, neural networks are just one of many techniques available. Importantly, the key to success is not necessarily using the most complex model (which often comes at the cost of good interpretability), but rather selecting the appropriate technique for the problem at hand.

In the second part of this thesis, such ML methods were applied to analyze the surfaces of complex biological materials. These efforts were driven by the desire for a reproducible and efficient way to assess and compare the surface quality of tissue that was subjected to tribological stress induced by a medical device. In a first step, by employing supervised ML methods to sort a labelled dataset of biological surfaces, it was pinpointed that a meaningful surface classification is indeed possible using a set of topographical parameters as input features (**chapter 4.1**). Based on the dataset provided, the ML algorithms not only learned to distinguish the surfaces of different biofilm variants and plant samples, but they also correctly classified the samples according to their wettability. Similar approaches could streamline material design processes, enabling a precise customization of the surface behavior that would otherwise necessitate time-consuming trial-and-error investigations. Speculatively, other topography-dependent surface properties could include friction responses, biofouling susceptibility, surface reflectance, adhesion, or biocompatibility. However, by including additional input features, *e.g.*, the chemical composition of a material or process parameters of the fabrication procedure, the number of predictable characteristics could be considerably extended. Furthermore, whereas the trained models generated in this work were limited to making predictions on samples from the same class of materials, providing multifaceted datasets with samples from other experimental techniques or new classes of materials might help to generalize the models for predictions on a broader range of materials.

Generally, the reliability of ML-driven data analysis heavily relies on the dataset provided. Here, not only the quality and the distribution play a major role, but also the number of provided sample points is key: The multi-dimensional feature space has to be populated with a sufficient number of samples to allow for robust and accurate predictions. For most experimental research problems, however, data availability is very limited; and paradoxically, having more features to describe each sample can even result in reduced prediction accuracies (known as the 'curse of dimensionality'). Hence, one essential step to be integrated into ML pipelines is feature elimination – a method that aims at reducing the dimensionality of an input feature space. Most available strategies eliminate features based on their relation to a selected output label, by optimizing performance measures of a certain ML model, or they exclusively focus on analyzing pairwise correlations. However, those strategies do not necessarily remove redundant information from datasets and often cannot be applied to certain situations, *e.g.*, to unsupervised learning models. Neither of these limitations applies to the correlation-driven redundancy elimination (NETCORE) algorithm introduced in this work (**chapter 4.2**). The proposed algorithm translates the dataset into a correlation network, which is then analyzed by conducting an iterative, three-step decision procedure. With this approach, the algorithm selects a subset of features that represent the full feature space on the basis of a (freely selectable) correlation threshold while taking into account the multi-connectivity of a feature to its neighbors in the correlation network. It was shown that feature elimination conducted with NETCORE can indeed help to improve the prediction accuracy of a ML model.

## Summary and outlook

Furthermore, the NETCORE algorithm offers the possibility to be adapted regarding several aspects that were not tested in this thesis. For instance, the metric used to calculate the correlation coefficient could be varied to increase the robustness of the coefficients and to capture a broader range of possible relations between the parameters (including non-linear relations). Moreover, the decision criteria used to assess centrality could be adjusted. However, the impact those modifications have on both, the elimination result and the computational runtime, need to be carefully evaluated. Overall, the NETCORE algorithm provides an interpretable solution to identify and remove redundant features from datasets, which could streamline the analysis of experimental data and ultimately help reducing the amount of experimental effort required to generate the dataset.

In addition to providing an adequate number of samples when generating experimental datasets to be analyzed with ML methods, another significant challenge is to identify appropriate output labels for each sample; this can be expensive, time-consuming, or even outright impossible. In such cases, where no labels are accessible, supervised ML models cannot be applied as they rely on a predefined ground truth to be trained on. This is where unsupervised methods come into play, that draw conclusions solely based on the sample characteristics. Interestingly, such approaches often reveal hidden structures and associations within the data, which may not be apparent from a manual inspection. In this work, unsupervised ML was employed to analyze a dataset of corneal tissue surfaces (**chapter 4.3**). Among the different models tested, a  $k$ -Means clustering algorithm was demonstrated to be the most suitable for the task at hand. From a material science point of view, the algorithm indeed successfully generated four clusters that accurately sorted the samples based on their surface appearance. Additionally, it was able to achieve a similar surface classification as conducted manually for the tribological evaluation of rigid contact lenses (**chapter 3.3**). Overall, it was shown that the presented ML-driven approach can guide and improve the analysis of datasets that – either owing to their sample size, or their complexity – go beyond the cognitive capabilities of humans. Of course, more complex algorithms could equally be employed to support or replace human decision making: Instead of using a set of surface parameters that has to be calculated based on the 3D height profiles, deep learning algorithms could be used to directly analyze the profiles. An example of such an approach is presented in **Appendix A8.11**, in which a pre-trained neural network is used to extract features from the height profiles; and those features are then used as input features for the  $k$ -Means clustering approach. Additionally, if even larger amounts of surface data were acquired, or if the sample size would be increased by data augmentation, neural networks could be directly trained to encode corneal images. However, as discussed in **chapter 4.3**, each of those approaches has to be individually evaluated regarding meaningfulness. Ultimately, nature-inspired or nature-based materials design combined with ML-driven data analysis has the potential to revolutionize the way we approach materials development, enabling to create more sustainable, efficient, and effective solutions to face the challenges of a rapidly changing world.

# Appendix

## Appendix A1: Additional experimental methods and results\*

### A1.1 Tests to evaluate the 3D printing performance of the bio-ink

#### Filament drop test

To determine the required pressure to be applied to extrude the bio-ink during the printing process, a filament drop test was conducted. Here, extrusion pressure was applied in a static printer configuration, *i.e.*, without any movement of the nozzle. The distance between the nozzle and the printing bed was set to 40 mm. Lateral images of the formed filament (or fragments of a filament) were captured using an EOS M50 system camera (Canon, Tokyo, Japan) in high-speed mode.

#### Filament shape evaluation

The identified pressure was then applied to print filaments at different moving speeds of the nozzle. Therefore, the nozzle height was calibrated to the printing bed and 5 straight lines (2 cm each) were printed. This procedure was repeated 3 times for each moving speed and images were captured using an EOS M50 system camera (Canon). Those images were analyzed using imageJ2 (version 2.3.0/1.53q) by measuring the width of each filament at 5 random positions (resulting in 75 datapoints per moving speed: 3 repetitions of printing 5 filaments and measuring them at 5 different positions).

#### Height increase measurements

To assess the increase in sample height achieved when printing several layers of the PGM-MA-based bio-ink, cylinders with a diameter of 5 mm were printed (Bio X™, Cellink) with a 25 G nozzle moving at a lateral speed of 20 mm s<sup>-1</sup>. An extrusion pressure of 80 kPa was applied, a layer height of 0.3 mm was targeted, and an infill density of 45 % with a rectangular pattern was selected. After extrusion of each layer, the sample was cured for 30 seconds with the in-built UV module for irradiation with 365 nm (Bio X™ Photocuring Toolhead 365 nm, Cellink). After printing each layer, a lateral photo was captured from the printed object, and the such obtained images were analyzed using ImageJ2 (version 2.3.0/1.53q). To do so, two lines were manually fitted to approximate the bottom and top edge of the printed samples, respectively. The distance of those lines was then measured at the horizontal middle point of the sample.

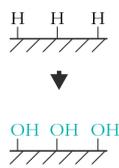
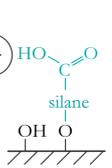
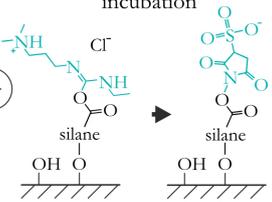
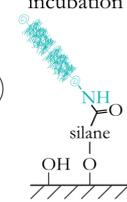
---

\* This section follows in part the publications Rickert *et al.*, 'A Mucin-based Bio-Ink for 3D Printing of Objects with Anti-Biofouling Properties', submitted 2023; Rickert *et al.*, *Biomaterials Advances* (2022), and Rickert *et al.*, *ACS Applied Materials & Interfaces* (2020)

## Appendix

### A1.2 Process parameters used to covalently attach mucins on different polymeric materials

The process carried out to generate covalent mucin coatings on surfaces as presented in **chapter 2.4.2** was adapted for each polymeric carrier material tested. The process parameters used for samples made from PDMS (silicone lenses), PU (urinary catheters), PVC (endotracheal tubes), and Fasifocon (RGP lenses) are depicted in **Figure A1.1**.

	activation	carboxylation			EDC-NHS reaction	mucin coupling
	plasma treatment 	incubation 	washing 	stabilization 	incubation 	incubation 
PDMS				80 °C, 2 h		
PU	90 s, 30 W, 0.4 mbar	60 °C, 5 h	in ethanol, 1 h	RT, 24 h	RT, 30 min	4 °C, 12 h
PVC						
Fasifocon	90 s, 80 W, 0.4 mbar	40 °C, 8 h	in ddH <sub>2</sub> O, 24 h	RT, 16 h, 0.2 bar		4 °C, 12 h + 1 mM DTT

**Figure A1.1: Covalent mucin coating procedure for different polymeric materials.** The process parameters applied to generate covalent mucin coatings on PDMS, PU, PVC, and Fasifocon (RGP lenses) are displayed.

### A1.3 Incubation volumes used for ELISA measurements

The ELISA described in **chapter 2.4.3** was performed with different types of samples and different sample holders. For all those sample types tested, the volumes of the chemicals used in each step are listed in the following table.

**Table A1.1: Incubation volumes used for the ELISA.** During the four incubation steps of the ELISA, different sample holders and liquid volumes were used to adapt the process to the respective sample shapes and sizes.

sample type	sample holder	blocking step	1 <sup>st</sup> antibody incubation	2 <sup>nd</sup> antibody incubation	color development
PDMS lenses	24-well plate	800 µL	500 µL	400 µL	300 µL
RGP lenses	24-well plate	800 µL	500 µL	400 µL	350 µL
PU & PVC samples	48-well plate	500 µL	300 µL	200 µL	150 µL
Passive mucin coatings in a 96-well plate	-	300 µL	300 µL	200 µL	150 µL

## A1.4 Detailed information regarding contact angle measurements

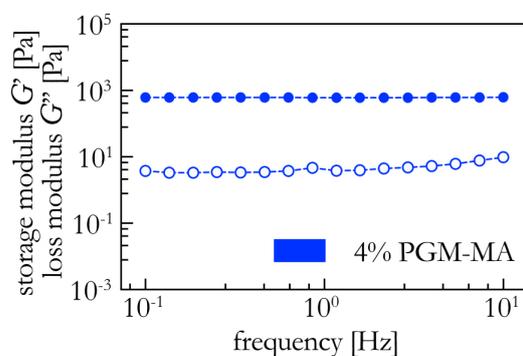
Contact angle measurements on PDMS disks, biofilms, and plant samples were conducted manually, whereas all other measurements were performed using a drop shape analyzing device equipped with an automated water dosing unit (DSA25S, Krüss GmbH, Hamburg, Germany). For the manual approach, a droplet of 10  $\mu\text{L}$  ddH<sub>2</sub>O was placed onto the cleaned and dried surfaces using a calibrated, mechanical pipet. The lateral images were then captured with a digital camera (Flea3, Point Gray Research, Richmond, Canada). Image processing was performed with the software ImageJ (public domain, version 1.52k, April 2019), and the static contact angles were determined with the “drop snake” plug-in of ImageJ.

For the automated approach using the drop shape analyzing device, a small droplet of 2  $\mu\text{L}$  ddH<sub>2</sub>O was placed onto the surface and transversal images were recorded with an integrated high-resolution camera (acA1920, Basler, Ahrensburg, Germany). These images were analyzed with the software ADVANCE (AD4021 v1.13, Krüss GmbH) using a manual drop shape fitting method with either a planar or a curved baseline (depending on the sample shape).

## A1.5 Preparation of PDMS pins

PDMS is a silicone-based, chemically inert, non-toxic, biocompatible, and transparent elastomer that finds broad application in various medical and industrial areas. In this thesis, cylindrical PDMS pins were used to realize a ball-on-three-pins setup for tribological experiments. Those PDMS pins were produced using a commercially available two-component system comprising a PDMS pre-polymer and the corresponding crosslinker (PDMS, Sylgard 184, Dow Corning, Wiesbaden, Germany). After thoroughly mixing the PDMS base with the curing agent in a ratio of 10:1, the mixture was degassed in a vacuum chamber for 1 h, poured into custom-made molds, and finally cured at 80 °C for 4 h. After curing, the pins were tempered at 100 °C for 2 h to release unreacted residues.<sup>291</sup>

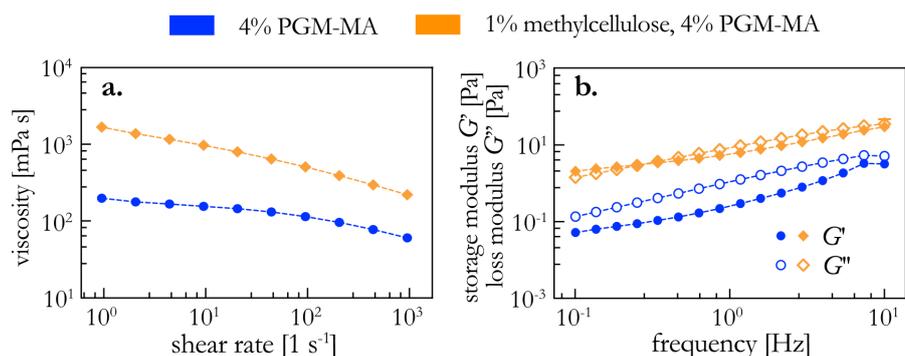
## A1.6 Viscoelastic properties of cured PGM-MA solutions



**Figure A1.2: Rheological characterization of PGM-MA solutions after curing.** The viscoelastic moduli measured for cured PGM-MA solutions over a frequency spectrum are displayed. Error bars denote the standard error of the mean obtained from  $n = 3$  measurements. If not visible, error bars are in the size of the symbols.

## A1.7 Viscoelastic properties of PGM-MA and methylcellulose

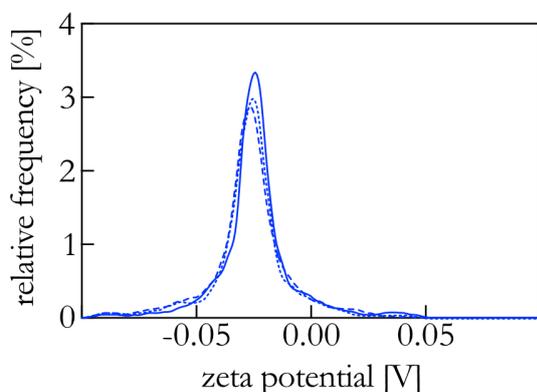
When exploring alternative formulations for a mucin-based bio-ink, methylcellulose (viscosity 4000 cP, Sigma-Aldrich) was used as an additive to adjust the viscoelastic properties of the formulation (see **chapter 3.1**). Methylcellulose powder was added to the lyophilized PGM-MA and the mixture was co-dissolved; for those samples, the mixing time was extended to 12 h to ensure a well-blended formulation.



**Figure A1.3: Rheological characterization of mucin-based of a bio-ink stabilized with methylcellulose.** The viscosity of the materials at different shear rates (a) and the viscoelastic moduli measured over a frequency spectrum (b) are displayed. Error bars denoting the standard error of the mean obtained from  $n = 3$  measurements are mostly smaller than the symbol size.

## A1.8 Zeta potential measurements

The zeta potential of the PGM-MA macromolecules was measured *via* electrophoretic light scattering using a Litesizer 500 (Anton Paar) equipped with a 35 mW laser diode ( $\lambda = 658 \text{ nm}$ ). Therefore, PGM-MA was dissolved at a concentration of  $0.1 \mu\text{g mL}^{-1}$  in ddH<sub>2</sub>O, injected into capillary cuvettes (Omega cuvette, Anton Paar), and inserted into the device. Then, the zeta potential was determined in automatic mode ( $T = 22 \text{ }^\circ\text{C}$ , equilibration time: 1 min) using the Smoluchowski approximation of the Henry equation (Henry factor: 1.5).



**Figure A1.4: Zeta potential of PGM-MA.** The frequency distributions of PGM-MA as obtained during electrophoretic light scattering measurements are displayed.

## A1.9 Additional methods to characterize mucin-based bio-inks

### Cytotoxicity tests

For cytotoxicity tests, an established human epithelial cell line (HeLa) was used as a model system. Those HeLa cells were cultivated in Minimum Essential Medium Eagle (MEM; Sigma-Aldrich) supplemented with 10 % (v/v) fetal bovine serum (FBS; Sigma-Aldrich), 2 mM L-glutamine (Sigma-Aldrich), 1 % (v/v) non-essential amino acid solution (NEAA; Sigma-Aldrich) and 1 % penicillin/streptomycin (Sigma-Aldrich). The cells were then cultured at 37 °C and 5 % CO<sub>2</sub> in a humidified atmosphere.

To analyze cell/bio-ink interactions and to investigate possible cytotoxic effects of the different bio-ink formulations, samples were prepared as described for the swelling experiments; then, cell-based assays according to ISO 10993 were conducted. In brief, the cured bio-ink specimens were immersed in the corresponding cell culture medium for 24 h at 37 °C to allow loosely integrated gel components to leach out. Afterwards, a WST-1 (water soluble tetrazolium salt 1) assay was employed to assess the cytotoxicity of this 'leaching' medium. For this step, HeLa cells were seeded into the wells of a 96-well plate (5,000 cells per well) and were incubated for 24 h at 37 °C. Subsequently, the cell culture medium was removed from the 96-well plate and replaced with sterile filtrated 'leaching' medium (as a control group, cells were incubated with fresh cell culture medium). After 24 and 48 h of incubation, respectively, the leaching medium was replaced with a 2 % (v/v) WST-1 solution (Sigma-Aldrich). After an incubation time of 45 min, the absorbance values of the samples were obtained at a wavelength of 450 nm and at a reference wavelength of 630 nm using a microplate reader (ABSPlus, Molecular Devices, Wokingham, Berkshire, UK). The reference values were subtracted from the absorbance values obtained for 450 nm and the cellular viability was then calculated by normalizing the values obtained for test groups to that of the control group.

### Cell adhesion tests

The attachment propensity of cells onto the bio-inks was analyzed by confocal laser scanning microscopy (CLSM; SP8, Leica; equipped with a 10× lens). For cell staining, the dye CellTracker Red CMTPX (ex.: 577 nm, em.: 602 nm; Thermo Fisher Scientific) was chosen; cells were stained prior to seeding them onto the bio-inks. To do so, HeLa cells were cultivated as described above, harvested, and resuspended in a serum-free medium (containing 10 μM of the dye) for 30 min. Then, the cells were washed twice with DPBS (pH 7.4; Sigma-Aldrich) and resuspended in the corresponding standard cell culture medium. Subsequently, the cells were seeded (10,000 cells per well) into 8-well chambers (80826, Ibidi, Gräfelfing, Germany) containing one of the bio-inks and incubated for 24 h at 37 °C. For each bio-ink variant, three sample replicates were prepared; empty wells were used as a control group. Then, using CLSM, 3D image stacks of the samples were collected by moving the focal plane (pinhole size = 1 airy disk) from the bottom of the samples to their top (nominal  $z$ -step size: 20 μm). The software Leica Application Suite X (Leica) was used for processing the 3D image stacks and cells were counted using ImageJ2 (public domain, version 2.3.0/1.53q).

## Appendix

### Bacterial adhesion tests

The adhesion of bacteria to bio-ink samples was evaluated for the strains *S. epidermidis* ATCC 14990 and *E. coli* ATCC 25922 (both American Type Culture Collection ATCC, Manassas, USA). Bio-ink samples were prepared in well plates as described for the swelling experiments (**chapter 2.2.6**), sterilized by exposing them to UV light at a wavelength of 254 nm for 2 h, and subsequently washed thrice with 70 % (v/v) ethanol. The bacterial strains were kept frozen in cryovials and were reconstituted in sterile PBS (pH 7.4). Afterwards, these bacterial suspensions were inoculated at a concentration of  $10^8$  CFU mL<sup>-1</sup> (colony forming units per mL) on the sterilized bio ink samples and incubated at 37 °C for 3 h. After incubation, the samples were thoroughly washed thrice with sterile PBS to remove the non-adherent bacteria. To detach the adherent bacteria from the bio-ink samples, the samples were transferred from the well plate into tubes and vortexed in sterile PBS for 1 min. For each sample, the bacterial suspension generated with this method was serially diluted and plated onto Chapman Agar (Sigma-Aldrich) in duplicates using the Interscience easySpiral Dilute (Interscience, Saint Nom la Bretèche, France). Then, to determine the number of adherent bacteria, those agar plates were incubated at 37 °C for 72 h, and the number of grown colony forming units was counted using the Scan Colony Counter (Interscience). The proportion of adhering bacteria was obtained by dividing the measured concentration by the concentration of the initial bacterial suspension.

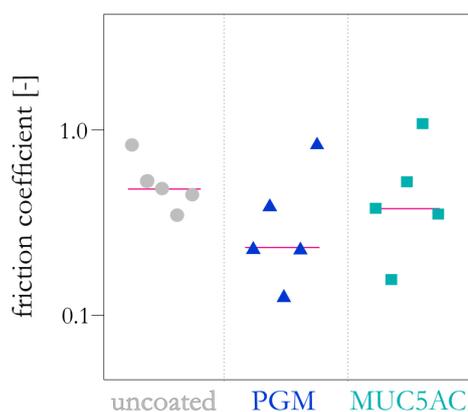
### Pro/anti-inflammatory cytokine release of macrophages

The immune response to the different bio-ink formulations was evaluated by determining the cytokine expression of monocyte-derived macrophages. The human cell line U937 (American Type Culture Collection ATCC) was kept in culture medium composed of RPMI 1640 medium (Gibco Life Technologies, Paisely, UK) supplemented with 10 % (v/v) fetal calf serum (Gibco),  $2 \times 10^{-3}$  M L-glutamine (Gibco),  $1 \times 10^{-3}$  M sodium pyruvate (Gibco), 1 U mL<sup>-1</sup> penicillin and 1 µg mL<sup>-1</sup> streptomycin (Gibco). The cells were then differentiated into M0 macrophages by culturing them in cell culture medium supplemented with  $100 \times 10^{-9}$  M phorbol-12-myristate-13-acetate for 72 h, and they were subsequently allowed to rest in fresh culture medium for 24 h. They were then seeded at a density of 900 000 cells cm<sup>-2</sup> onto the samples, and after 72 h of incubation at 37 °C with 5 % CO<sub>2</sub>, the cytokine concentration in the supernatant was evaluated. Quantification of the released proinflammatory (tumor necrosis factor- $\alpha$  and interleukin-6) and anti-inflammatory (interleukin-10 and transforming growth factor- $\beta$ ) cytokines was performed with the DuoSetELISA kit (R&D Systems, Minneapolis, USA) following the manufacturer's instructions using a multimode microplate reader (Spark, Tecan, Männedorf, Switzerland).

## A1.10 Additional results from ocular tribology with PDMS lenses

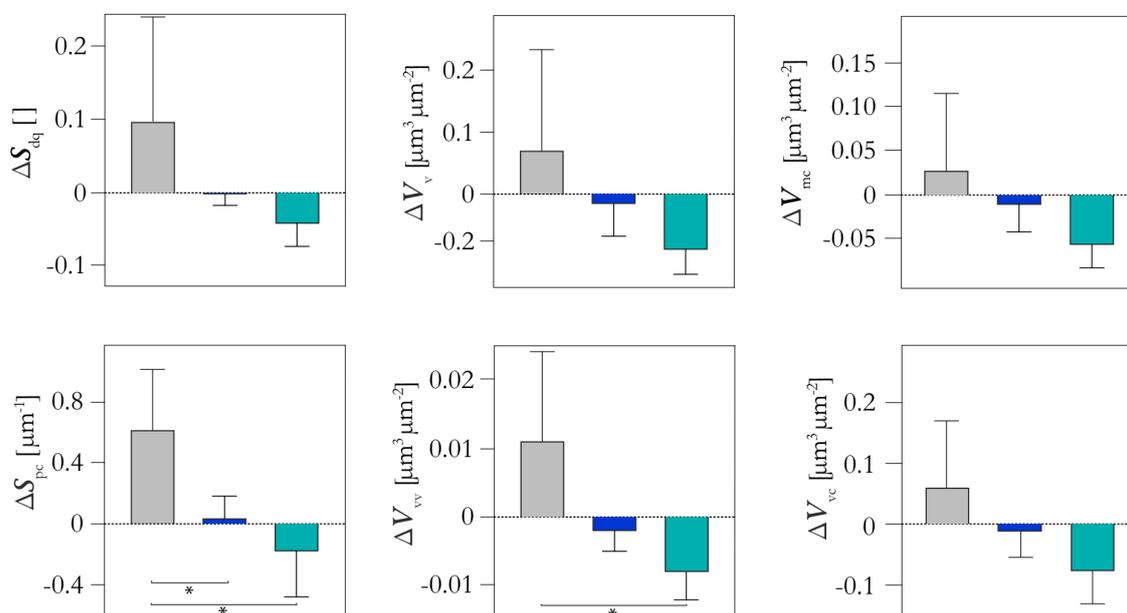
## Friction coefficients

Friction values were obtained during tribological measurements as described in **chapter 2.5.2**. Rotational stress was applied for a total duration of 90 min, during which a data point was acquired every minute. For each individual measurement, the data points of minute 15 to minute 75 were averaged.



**Figure A1.5: Friction coefficients obtained between PDMS lenses and porcine corneas.** The individual friction coefficients obtained for PDMS lenses that were either left blank and untreated (grey) or that were covalently coated with either PGM (blue) or MUC5AC (cyan) are displayed. Red lines indicate the median as obtained from  $n = 5$  samples.

## Quantification of wear formation by additional ISO parameters



**Figure A1.6: Relative surface roughness parameters of corneal tissue.** Six relative surface parameters were determined according to ISO 25178-2 (see **Appendix A3**) to support the parameters shown in **chapter 3.2**. The error bars denote the standard deviation as obtained from  $n = 5$  independent samples. The asterisks indicate significance at a level of  $p = 0.05$ .

## Appendix

### Profilometric images of corneal samples

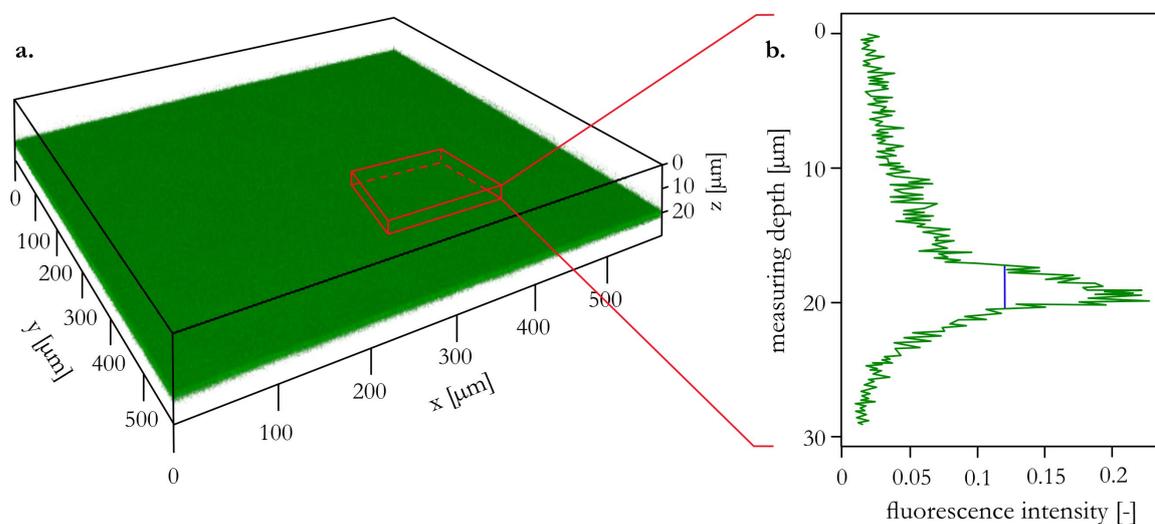


**Figure A1.7: Profilometry images of corneal tissue before and after tribological treatment.** The images were obtained before (black boxes) and after tribological treatments using either an uncoated PDMS lens (grey boxes) or PDMS lenses that were covalently coated with commercial PGM (blue) and manually purified MUC5AC (cyan), respectively.

## A1.11 Measuring the mucin layer thickness

To attach a fluorescent label to the MUC5AC molecules, carbodiimide coupling was employed as described in **chapter 2.4.3** for the protein detection of PGM. To then generate fluorescent, covalent mucin coatings on RGP lenses, those fluorescently labelled were dissolved in DPBS, and the covalent coating procedure was conducted as described in **chapter 2.4.2**. However, during all steps involving mucins, special care was taken to avoid light exposure.

To then visualize the generated mucin layer, confocal microscopy was employed using a Leica TCS SP5 II setup (Leica, Wetzlar, Germany) equipped with a DMI6000 microscope corpus (Leica) and a 40 $\times$  objective (HCX PL APO; NA = 1.25; Leica). The fluorescently labelled mucins were excited at 488 nm and emission was detected in a window ranging from 500 to 527 nm. A  $z$ -stack of images was then recorded with a step-size of 0.1  $\mu\text{m}$  (**Figure A1.8a**); from this stack, the thickness of the mucin layers was determined as follows: after converting the image stack from rgb to grayscale, the mean intensity was calculated for each individual image. From the obtained signal (which represents the average fluorescence intensity over the measuring depth; **Figure A2.9b**), the full width at half maximum (FWHM) was determined and used as an approximation of the thickness of the mucin layer. Importantly, to avoid any distortion of the intensity curve arising from the curved and tilted lens surface, this analysis was performed on 25 randomly selected 3 $\times$ 3 pixel patches of each image. All calculations were conducted in python (Python Software Foundation; Python Language Reference, version 3.9.12; <http://www.python.org>)<sup>102</sup> with the Numpy (v1.20.3)<sup>103</sup> extension for data handling and the scipy (v1.7.3)<sup>107</sup> plugin for data analysis.



**Figure A1.8: Visualization of the mucin coating and estimation of the coating thickness.** A confocal image obtained from a coating generated with fluorescently labelled MUC5AC is depicted (a). From this  $z$ -stack of fluorescence images, an intensity curve (b) was determined and the thickness of the coating was approximated as the full peak width at half maximum.

### A1.12 Validation of geometrical and physical lens properties

Various geometrical and physical properties of contact lenses were characterized according to standardized tests. These measurements were conducted by the lens manufacturer Wöhlk Contactlinsen GmbH as described in the following.

#### **Spectral transmittance**

To evaluate the spectral transmittance of coated RGP contact lenses according to EN ISO 18369-3, a UV/VIS photo-spectrometer (Lambda 14; PerkinElmer, Waltham, USA) was used. Importantly, lenses with a low diopter strength were used to minimize the deflection of light rays through the lens. Prior to conducting measurements, the lenses were stored in PBS (pH 7.2 – 7.4) for at least 24 h. After drying the lenses, they were inserted into a cuvette, which was then placed into the instrument such that the lenses were aligned with the beam path (with the convex side facing the light source). Then, transmission values were measured for wavelengths ranging from 280 to 800 nm and normalized to the values obtained for an empty cuvette. According to EN ISO 18369-3, the measured values were averaged over the following three spectral regions: UV-A (280 – 315 nm), UV-B (315 – 380 nm), and VIS (380 - 800 nm).

#### **Geometrical characterization**

All geometrical parameters were determined according to EN ISO 18369-3. In brief, the lens diameter was measured using a V-groove diameter gauge (custom-made by Wöhlk Contactlinsen GmbH according to norm specifications; resolution: 0.05 mm). The radius of the central optical zone was assessed on the back surface of the lens at two different angles (8.5° and 25°) using a commercial ophthalmometer (CL 110; Zeiss, Jena, Germany; resolution: 0.005 mm). The thickness of the central area of the lens was determined with a classical dial gauge (Mitutoyo, Kawasaki, Japan; resolution: 0.005 mm). Lastly, the diopter strength was measured with a vertex refractometer (Auto Lensmeter SLM-4000; Shin-Nippon, Chiyoda, Japan; resolution: 0.01 dpt).

#### **Refractive index**

The refractive index (RI) of RGP lenses was determined according to EN ISO 18369-4 using a digital, automated refractometer (CLR 12-70; Index Instruments Ltd., Huntingdon, UK) that measures light reflection at a wavelength of 589 nm. As the RI is temperature-dependent, a water bath (Lauda CS, Lauda DR. R. Wobser GmbH & CO. KG, Lauda, Germany) was used to cool the measuring system to 20 °C. Moreover, the system was calibrated before each measurement by assessing the RI of a plate made from fused silica as a reference. To do so, the plate was placed onto the prism of the instrument and a few drops of 1-methylnaphthaline were added in between the two materials to serve as a contact liquid. The RI was then measured after 30 s, and the system was considered calibrated well when the obtained values were within a defined tolerance range. To determine the RI of a contact lens, the lens was attached to a sample holder that was wetted with 1-methylnaphthaline. After adding a few drops of 1-methylnaphthalin onto the outer surface of the attached contact lens, the sample holder was mounted into the refractometer such that the lens was positioned on the prism. Then, the RI was again measured after 30 s.

### Oxygen permeability quantification

The thickness-independent oxygen permeability  $Dk$  was determined with a polarographic technique according to EN ISO 18369-4. In brief, a climatic chamber was lined with moistened paper towels to create a vapor-saturated atmosphere, and the temperature within the chamber was equilibrated to 35 °C. During equilibration, a current/temperature measuring instrument ( $O_2$  permeometer™ model 201T, Createch, Rehder Development, San Francisco, USA) with a measuring cell (radius of curvature of the electrode:  $r = 8.1$  mm; diameter of the gold cathode:  $d = 4$  mm) was placed into the chamber. Once atmospheric equilibrium was achieved, four lenses with different center thicknesses (ranging from 0.1 to 0.4 mm) were analyzed. Therefore, a piece of cigarette paper was placed crease-free on the measuring electrode, moistened with PBS, and fixed in the holder. This paper serves as a 'liquid bridge' needed to measure anhydrous contact lenses. After 20 min (when a constant current and temperature should be measured), the first lens was placed onto the electrode; after another 20 min, the reached current was measured. Then, the lens was removed and the procedure was repeated with the other three lenses. Based on those measurements, the uncorrected  $Dk$  values were calculated as described in EN ISO 18369-4.

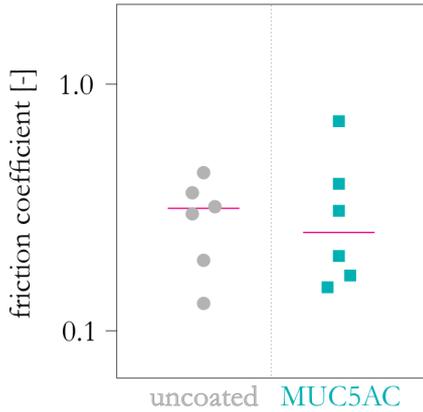
### Flexural rigidity

To evaluate the flexural rigidity of the lenses, deformation tests were conducted according to EN ISO 18369-4 using a universal testing device equipped with a vertical jig system (EZ S; Shimadzu, Kyoto, Japan). Before each measurement, the contact lenses were immersed in freshly autoclaved PBS at 2-25 °C for 48 h and then dried with pressurized air. After adjusting the jig according to the lens dimensions (9.5 mm) and resetting the force gauge and the odometer, the contact lens was vertically inserted into the compression module of the instrument and the measurement was started: compression stress was applied with a constant speed of 200 mm min<sup>-1</sup> (test mode: single; test type: compression), and the deformation force and the corresponding strain (% deformation with respect to the total lens diameter of 9.5 mm) were measured until fracture of the lens occurred. The key parameters derived from the resulting compression/strain curve are the compression force at a deformation of 30 %, the compression force at the time of breakage, and the maximum deformation before fracture occurs.

### A1.13 Additional results from ocular tribology with RGP lenses

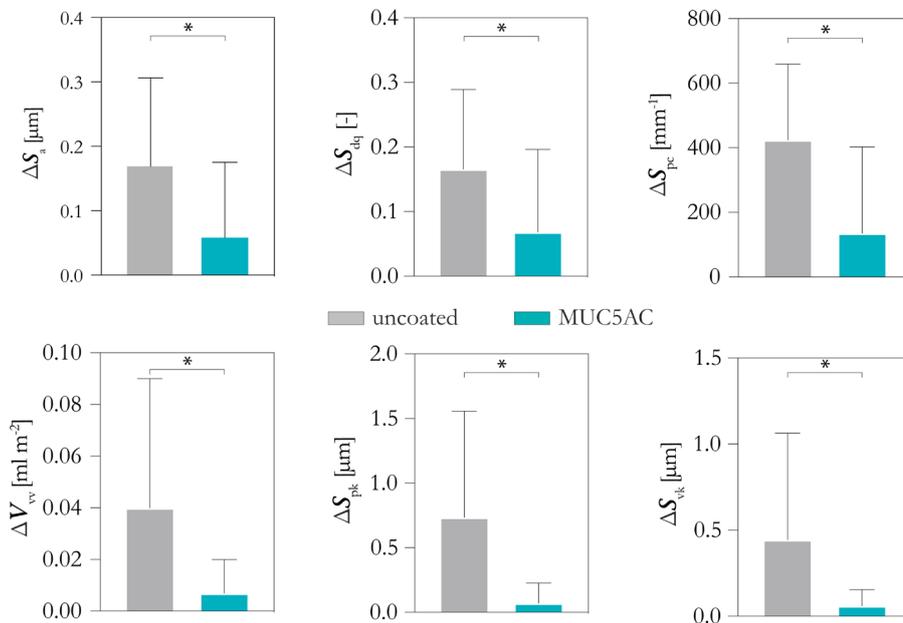
#### Friction coefficients

Friction values were obtained during tribological measurements as described in **chapter 2.5.2**. Rotational stress was applied for a total duration of 90 min, during which a data point was acquired every minute. For each individual measurement, the data points of minute 15 to minute 75 were averaged.



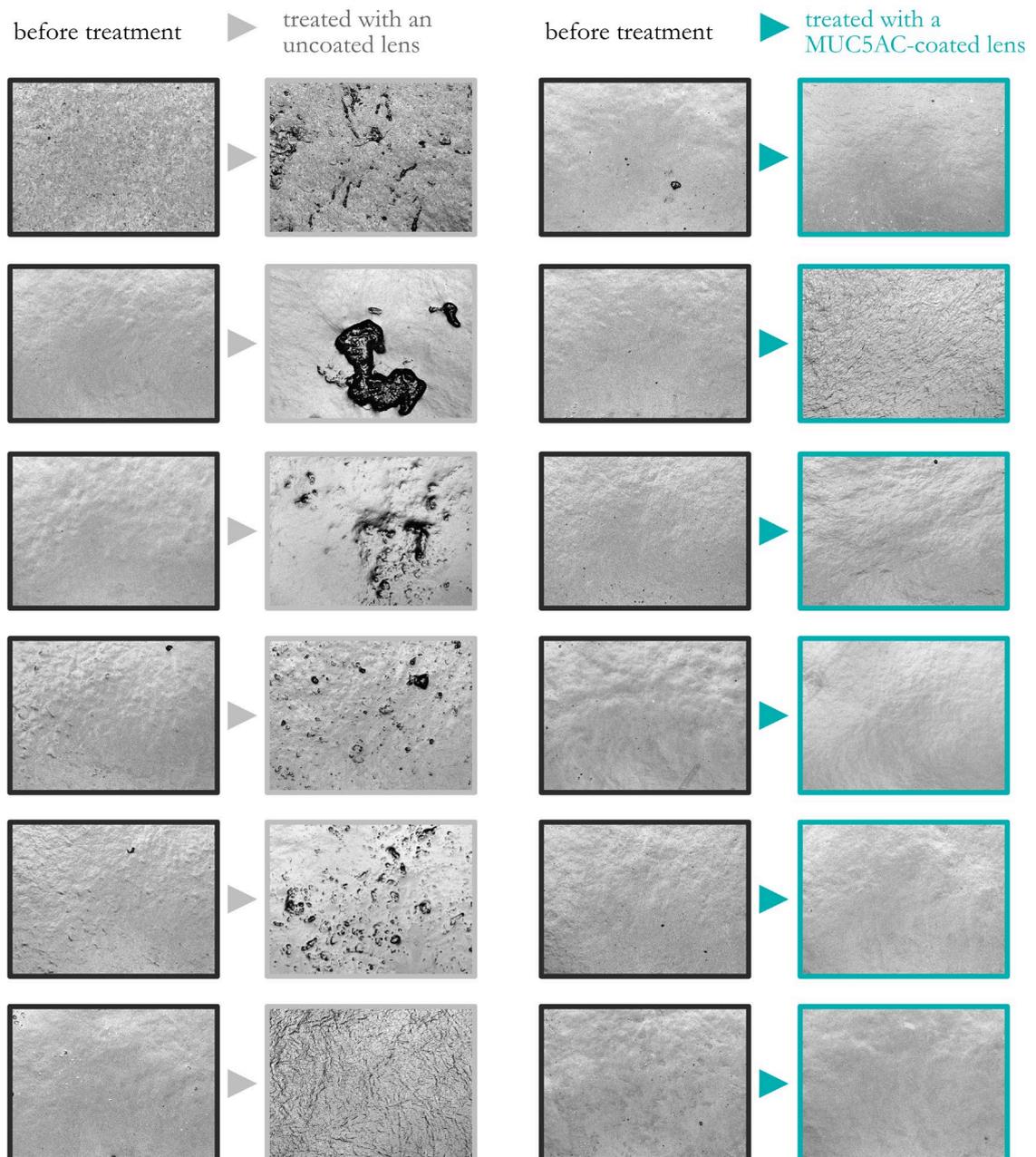
**Figure A1.9: Friction coefficients obtained between PDMS lenses and porcine corneas.** The individual friction coefficients obtained for PDMS lenses that were either left blank and untreated (grey) or that were covalently coated with either PGM (blue) or MUC5AC (cyan) are displayed. Red lines indicate the median as obtained from  $n = 6$  samples.

#### Quantification of wear formation by additional ISO parameters



**Figure A1.10: Relative surface roughness parameters of corneal tissue.** Six relative surface parameters were determined according to ISO 25178-2 to support the parameters shown in **chapter 3.3**. The error bars denote the standard deviation as obtained from  $n = 6$  samples. The asterisks indicate significance at a level of  $p = 0.05$ .

## Profilometric images of corneal samples

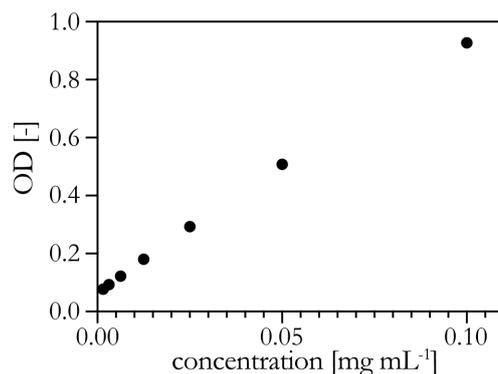


**Figure A1.11: Profilometry images of corneal tissue before and after tribological treatment.** The images were obtained before (black boxes) and after tribological treatments using either an uncoated RGP lens (grey boxes) or RGP lenses that were covalently coated with manually purified MUC5AC (cyan).

## A1.14 Antibiotic release experiments from mucin coatings

To test the release of drugs encapsulated in the mucin layer, the covalent coating process was conducted as described above, but  $1 \text{ mg mL}^{-1}$  tetracycline hydrochloride (TCL) was added to the mucin solution used during the final incubation step of the coating protocol. To subsequently trap the co-incubated drugs in the mucin layer, a conformational change from elongated mucin molecules to a condensed state was achieved by immersing the coated lenses in  $1 \text{ mL}$  ddH<sub>2</sub>O supplemented with 30 % (v/v) glycerol at  $4 \text{ }^\circ\text{C}$  for 16 h.<sup>292</sup> This particle-like conformation was ionically stabilized by incubating the lenses in  $1 \text{ mL}$  ddH<sub>2</sub>O containing 30 % (v/v) glycerol and  $100 \text{ mM}$  MgCl<sub>2</sub> at  $4 \text{ }^\circ\text{C}$  for 2 h.<sup>204</sup> To remove unbound remnants from the preparation steps, the specimens were gently washed with ddH<sub>2</sub>O before further experiments were conducted.

For the release experiments, each lens was placed into a well of a 48-well plate and  $300 \text{ }\mu\text{L}$  of DPBS were added on top. Then, the release of TCL was spectroscopically observed at  $360 \text{ nm}$ . Therefore,  $100 \text{ }\mu\text{L}$  of the DPBS were taken from each sample well, transferred to a 96-well plate, and absorbance values were measured (ABSPlus; Molecular Devices, San José, USA). Afterwards, the liquid was returned to the wells containing the lens samples. Between the measurements, the well plate with the specimens was gently moved on an orbital shaker at  $40 \text{ rpm}$  and  $37 \text{ }^\circ\text{C}$ . Last, the obtained optical density values were converted to drug concentrations based on a calibration curve measured for TCL in DPBS (Figure A1.12).



**Figure A1.12: Calibration curve obtained for tetracycline hydrochloride solutions.** Optical densities are shown for different concentrations of tetracycline hydrochloride dissolved in DPBS. Error bars indicating the standard error of the mean were obtained from  $n = 3$  samples but were smaller than the symbol size.

### A1.15 SDS-PAGE and Coomassie staining

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS PAGE) was conducted to visualize the molecular mass distribution of proteinous sample components in differently treated mucin solutions. Therefore, each mucin solution (ddH<sub>2</sub>O containing 1 % [w/v] mucin) was mixed in a ratio of 1:1 with 2x sample buffer (0.4 M triethanolamine [TEA], 2 % [w/v] SDS, 20 % [v/v] glycerol, 2 mM ethylenediaminetetraacetic acid, and 0.02 % (w/v) bromophenol blue; pH 7.8). After boiling at 95 °C for 5 min, 20 µL of each mixture were then loaded into separate lanes of a pre-cast polyacrylamide gel (4-20 % TruPAGE™ Precast Protein Gels; Sigma-Aldrich). Furthermore, 4 µL of a protein marker solution (Marker Precision Plus Protein Kaleidoscope Standards; Bio-Rad Laboratories Inc., Hercules, USA) were loaded into a separate lane of the gel. The gel electrophoresis was run in TEA-tricine-SDS buffer (60 mM TEA, 40 mM tricine 0.1 % (w/v) SDS; pH 8.3) at 120 V for 1 h.

A Coomassie staining (0.1 % [w/v] Coomassie R250, 10 % [v/v] glacial acetic acid, 40 % [v/v] methanol) was used to stain the gels for protein detection. Therefore, the gels were incubated in the staining solution while gently shaking at RT for 1 h. Afterwards, the gel was de-stained in 10 % (v/v) acetic acid overnight and then imaged.

### Appendix A2: Modifications of the covalent mucin coating process\*

As described in **chapters 2.4.2** and **3.3**, various steps of the covalent coating process used for silicone contact lenses were adjusted to meet the requirements of the RGP lens material (*i.e.*, its limited stability towards solvents and temperatures above 50 °C). However, altering some of the process parameters can considerably impair both, the lens transparency and the stability of the coating. For instance, replacing 96 % (v/v) ethanol with ddH<sub>2</sub>O in the 1 h washing step resulted in turbid lenses, and decreasing the temperature applied during the 4 h of silane stabilization reduced the coating stability. To compensate for those effects, further alterations had to be implemented.

#### Turbidity evaluation

During the coating optimization process, turbidity was visually (and qualitatively) evaluated by subjectively classifying lenses as “turbid” or “clear”. After having developed the “final” process, the lens transparency was quantitatively evaluated as stated in **chapter 2.4.5**.

#### Stability measurements

To analyze the stability of the mucin coatings, the coated contact lenses were subjected to intensive cleaning; contact angles were measured to indirectly assess the integrity of the coating. Each cleaning cycle involved rubbing the lens surface with glove-protected fingertips using a conventional contact lens cleaner (CL44; Optosol Chemische Produkte GmbH, Miesbach, Germany) for 10 s. Contact angle measurements were conducted on the lens surfaces before the first cleaning cycle, and after every fifth cleaning cycle. Prior to each contact angle measurement, to remove contact lens cleaner residues and any particles from the lens surface, the contact lenses were washed with ddH<sub>2</sub>O, rinsed with 80 % (v/v) ethanol, and then washed again in ddH<sub>2</sub>O. The lenses were then air dried for at least 10 min. If necessary, excess water was removed from the edge of the lenses with a paper towel (KIMTECH Science, Kimberly-Clark, Dallas, United States). To measure contact angles, a 7 µL droplet of ddH<sub>2</sub>O was applied centrally to the outer surface of each contact lens and lateral images were acquired using a digital camera (Point Grey, Flea 3 FL3-U3-32S2C-CS, lens: F1.8/9-90mm). The contact angles were evaluated using the “drop snake” application of the ImageJ (open-source software, v152.p). Subsequently, the measured contact angles were computationally reduced by 18° to compensate for the curvature of the contact lens surface, which is assumed by the software to be planar by default.

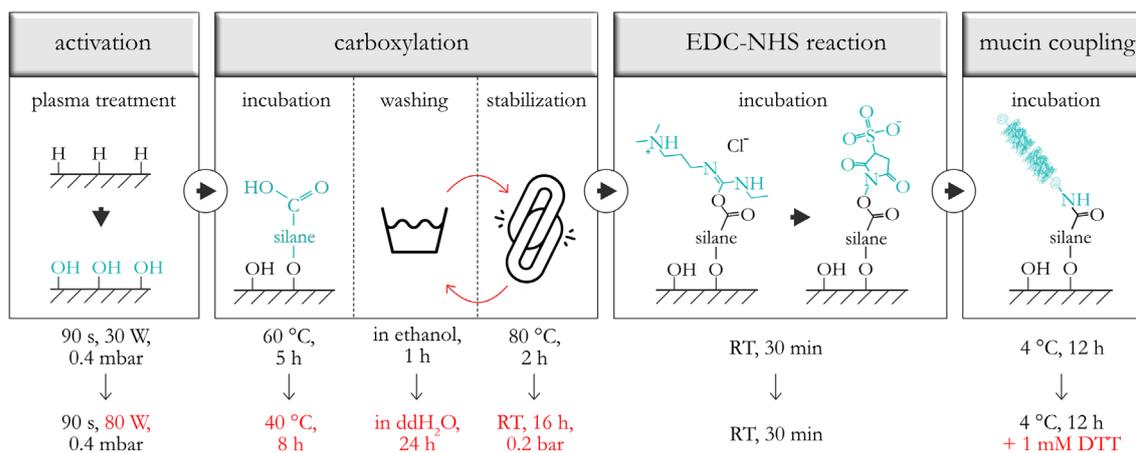
#### Original coating process

As described in **chapter 2.4.2**, the procedure conducted for generating covalent MUC5AC coatings on PDMS contact lenses consists of the following steps (see **Figure A2.1**): After cleaning the fresh lenses with 80 % (v/v) ethanol and ddH<sub>2</sub>O, the lens surfaces were activated *via* plasma activation using a pressure of 0.4 mbar and an intensity of 30 W for 90 s. This step is required to substitute the methyl groups on the surface with hydroxyl groups. To obtain a carboxylated surface (which is needed for the subsequent carbodiimide coupling), the lenses were incubated in 10 mM acetate buffer (pH 4.5) supplemented with 1.0 % (v/v) of the coupling agent TMS-EDTA at 60 °C for 5 h. To remove loosely

---

\* This section follows in part the publication Rickert *et al.*, *Biomaterials Advances* (2022)

bound silanes that could do both, induce turbidity on the lenses or destabilize the coating, the samples were dipped into isopropanol (>99.5 %) and were then washed in 96 % (v/v) ethanol on a rolling shaker at 70 rpm for 1 h. To stabilize the previously generated siloxane bonds, the specimens were cured at 80 °C for 2 h. To attach the MUC5AC macromolecules onto the surface *via* carbodiimide coupling, the carboxyl groups on the contact lenses were activated by incubating the samples in 100 mM MES buffer (pH 5.0) comprising 5 mM EDC and 5 mM sulfo-NHS at RT for 30 min during gentle shaking (35 rpm). Finally, the samples were immersed into DPBS supplemented with 0.1 % (w/v) mucin molecules at 4 °C for 12 h. The coated samples were washed in 80 % (v/v) ethanol and stored in DPBS.



**Figure A2.1: Schematic representation of the covalent mucin coating process and the adjusted process parameters.** Lens surfaces were activated using plasma generated with ambient air before they were carboxylated with a silane precursor. Subsequently, carbodiimide coupling was employed to covalently attach mucin molecules onto the lens surface. For the process to become suitable for RGP lenses, the following modifications were applied: first, the plasma intensity was increased. Moreover, the temperature for both, silane incubation and silane stabilization, was decreased and the corresponding incubation/stabilization times were increased. Furthermore, the silane stabilization step and the silane washing step were switched, and silane stabilization was conducted in an evacuated environment. For washing the lenses, ethanol was replaced with ddH<sub>2</sub>O, but the washing time was strongly increased. Lastly, 1 mM DTT was added to ensure that monomeric mucins were coupled onto the lens surface.

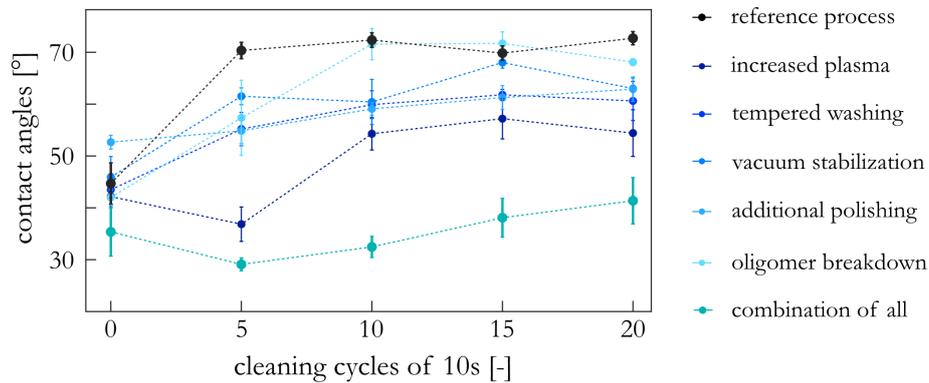
### Process modifications

First, all steps containing organic solvents were changed such that both, ethanol and isopropanol, were replaced with ddH<sub>2</sub>O. Owing to the lower detergency of water compared to those solvents, however, this change initially resulted in turbid lenses. To compensate for this, the efficiency of the mechanical washing step was strongly increased by changing the washing time from 1 h to 24 h. With this alteration, coatings without noticeable turbidity were achieved.

Next, all steps involving increased temperature levels were adjusted: The silane incubation step was now performed at 40 °C for 8 h (instead of at 60 °C for 5 h), and the silane stabilization step was conducted at RT for 16 h (instead of at 80 °C for 2 h). Furthermore, to ensure that covalent silane bonds were established before thoroughly washing the samples (and thereby possibly removing still weakly bonded silanes), the silane stabilization step was now performed before sample washing.

## Appendix

Yet, the coatings achieved with this process were still lacking stability: Contact angles measured on such coated lenses strongly increased after 5 cleaning cycles only (**Figure A2.2**, ‘reference process’). Therefore, additional modifications were tested (**Figure A2.2**): the plasma intensity was increased to 100 W, the washing procedure was performed at 40 °C, the silane stabilization was conducted in a vacuum chamber, an additional polishing step (employing the commercial lens cleaner used for the cleaning cycles) was added before washing the samples, and DTT was added to the mucin solution in the final incubation step (to break down mucin oligomers into monomers). Interestingly, all of those alterations improved the stability of the coatings – at least to a certain extent. Especially when combined, considerable improvements with regard of the coating stability were achieved (**Figure A2.2**, ‘combination of all’). Based on these findings and in consideration of good practicability in the laboratory, the final coating process described in detail in **chapter 3.3** and **Appendix A1.2** was developed.



**Figure A2.2: Wettability of differently coated lenses after various numbers of cleaning cycles.** Contact angles obtained for lenses coated with different process modifications after different numbers of cleaning cycles are shown. Error bars denote the standard error of the mean as obtained for  $n = 3$  samples each.

## Appendix A3: Overview of topographical surface parameters\*

### A3.1 Height parameters

#### Arithmetical mean height: $S_a$

The  $S_a$  represents the average of the absolute height values at each point of the analyzed area.

$$S_a = \frac{1}{A} \iint_A |z(x, y)| dx dy \quad (\text{A3.1})$$

#### Root mean square height: $S_q$

The root mean square height  $S_q$  quantifies the standard deviation of the height considering all individual points within the analyzed area.

$$S_q = \sqrt{\frac{1}{A} \iint_A z^2(x, y) dx dy} \quad (\text{A3.2})$$

#### Skewness: $S_{sk}$

The skewness  $S_{sk}$  denotes the cubic (normalized) average of the height. It indicates an asymmetry of the height distribution with regard to the reference plane:  $S_{sk} < 0$  and  $S_{sk} > 0$  refer to the predominance of pits or peak structures, respectively.

$$S_{sk} = \frac{1}{S_q^3} \left[ \frac{1}{A} \iint_A z^3(x, y) dx dy \right] \quad (\text{A3.3})$$

#### Kurtosis: $S_{ku}$

The kurtosis  $S_{ku}$  is a measure for the sharpness of the height profile, which is calculated as the average fourth power of the height normalized by the fourth power of  $S_q$ . An  $S_{ku}$  of 3 denotes a height distribution similar to a normal distribution, whereas  $S_{ku} < 3$  or  $S_{ku} > 3$  indicate an arched or spiked shape, respectively.

$$S_{ku} = \frac{1}{S_q^4} \left[ \frac{1}{A} \iint_A z^4(x, y) dx dy \right] \quad (\text{A3.4})$$

#### Maximal individual values: $S_p$ , $S_v$ and $S_z$

These individual values denote the maximum peak height  $S_p$  (*i.e.*, the height of the highest point in the analyzed area), the maximum pit depth  $S_v$  (*i.e.*, the absolute height value of the lowest point within the analyzed area), and the maximum height  $S_z$  calculated as the sum of both.

$$S_p = \max_A z(x, y) \quad (\text{A3.5})$$

$$S_v = \left| \min_A z(x, y) \right| \quad (\text{A3.6})$$

$$S_z = S_p + S_v \quad (\text{A3.7})$$

---

\* This section follows in part the publication Rickert *et al.*, *ACS Biomaterials Science & Engineering* (2021)

## Appendix

### Ten-point-height: $S_{10z}$

The  $S_{10z}$  takes into account the height of the five highest local maxima above the mean surface and the depth of the five lowest local minima below the mean height.

$$S_{10z} = \frac{1}{10} \left( \sum_{i=1}^5 Z_{i,max} + \sum_{j=1}^5 Z_{j,min} \right) \quad (A3.8)$$

### A3.2 Spatial parameters

Two of the following spatial parameters are calculated based on the autocorrelation function  $f_{ACF}$ :

$$f_{ACF}(t_x, t_y) = \frac{\iint_A z(x, y)z(x - t_x, y - t_y) dx dy}{\iint_A z(x, y)^2 dx dy} \quad (A3.9)$$

#### Autocorrelation length: $S_{al}$

The autocorrelation length represents the horizontal distance of the autocorrelation function with the fastest decay to a pre-defined value  $s$  (here 0.2). High  $S_{al}$  values indicate the dominance of long-wavelength components, whereas small values indicate that small-wavelength features are dominant.

$$S_{al} = \min_{t_x, t_y \in R} \sqrt{t_x^2 + t_y^2} \quad \text{where} \quad R = \{(t_x, t_y): f_{ACF}(t_x, t_y) \leq s\} \quad (A3.10)$$

#### Texture aspect ratio: $S_{tr}$

The texture aspect ratio indicates the uniformity of the surface texture by dividing  $S_{al}$  by the horizontal distance with the slowest decay to the predefined value  $s$ .  $S_{tr}$  can range from 0 to 1:  $S_{tr} > 0.5$  indicates an isotropic surface, whereas  $S_{tr} < 0.3$  means that the surface texture has a certain orientation.

$$S_{tr} = \frac{\min_{t_x, t_y \in R} \sqrt{t_x^2 + t_y^2}}{\max_{t_x, t_y \in Q} \sqrt{t_x^2 + t_y^2}}, \quad \text{where} \quad \begin{aligned} R &= \{(t_x, t_y): f_{ACF}(t_x, t_y) \leq s\} \\ Q &= \{(t_x, t_y): f_{ACF}(t_x, t_y) \geq s\} \end{aligned} \quad (A3.11)$$

#### Texture direction: $S_{td}$

The texture direction  $S_{td}$  denotes the value of the angle  $s$ , where the angular spectrum  $f_{APS}$  is maximized. With this parameter, no meaningful values can be obtained on isotropic surfaces (surfaces with  $S_{tr} > 0.5$ ). The angular spectrum is defined as:

$$f_{APS}(s) = \int_{R_2}^{R_1} r |F[r \sin(s - \theta), r \cos(s - \theta)]|^2 dr \quad (A3.12)$$

with

$$F(p, q) = \iint_A z(x, y) e^{-(ipx + ipy)} dx dy \quad (A3.13)$$

### A3.3 Hybrid parameters

#### Root mean square gradient: $S_{dq}$

The  $S_{dq}$  value is calculated as the root mean square of the gradient at all points of the definition area. It is particularly suitable to distinguish surfaces that have the same arithmetical height  $S_a$ . There, small  $S_{dq}$  values indicate surfaces that are dominated by long-wavelength components, whereas the  $S_{dq}$  will increase for surfaces that are dominated by small-wavelength features.

$$S_{dq} = \sqrt{\frac{1}{A} \iint_A \left[ \left( \frac{\partial z(x,y)}{\partial x} \right)^2 + \left( \frac{\partial z(x,y)}{\partial y} \right)^2 \right] dx dy} \quad (\text{A3.14})$$

#### Developed interfacial ratio: $S_{dr}$

The  $S_{dr}$  quantifies the percentage of additional surface area contributed by the texture as compared to a completely planar surface. It is zero if the surface is completely flat and perpendicular to the orientation of the height. The  $S_{dr}$  parameter is sensitive towards both, feature amplitude and wavelength.

$$S_{dr} = \frac{1}{A} \left[ \iint_A \left( \sqrt{\left[ 1 + \left( \frac{\partial z(x,y)}{\partial x} \right)^2 + \left( \frac{\partial z(x,y)}{\partial y} \right)^2 \right]} - 1 \right) dx dy \right] \quad (\text{A3.15})$$

#### Arithmetic mean peak curvature: $S_{pc}$

$S_{pc}$  quantifies the arithmetic mean of the principal curvatures of peaks on the surface. Due to the inability to perform Wolf pruning (*i.e.*, to merge the peaks of micro-hill areas with adjoining areas) with the software used here, this parameter does not conform with ISO 25178-2: 2012.

$$S_{pc} = -\frac{1}{2} \frac{1}{n} \sum_{k=1}^n \left( \frac{\partial^2 z(x,y)}{\partial x^2} + \frac{\partial^2 z(x,y)}{\partial y^2} \right) \quad (\text{A3.16})$$

### A3.4 Functional volume parameters

Functional volume parameters can be calculated based on the Abbot Firestone curve, which displays the areal material ratio  $S_{mr}$  at each height  $h$ . From this curve, two basic metrics can be calculated: first, the void volume (**equation A3.17**) which is defined as the volume of open space per unit area; second, the material volume (**equation A3.18**), which describes the volume of actual substance per unit area. In those equations,  $S_{mc}$  refers to the inverse areal material ratio, which represents the height  $h$  that meets the areal material ratio.  $K$  is a constant introduced for converting the units to mL mm<sup>-2</sup>. From these definitions, four functional volume parameters were determined as depicted in **Figure A3.1**.

$$V_{v(p)} = \frac{K}{100\%} \int_p^{100\%} [S_{mc(p)} - S_{mc(q)}] dq \quad (\text{A3.17})$$

$$V_{m(p)} = \frac{K}{100\%} \int_0^p [S_{mc(q)} - S_{mc(p)}] dq \quad (\text{A3.18})$$

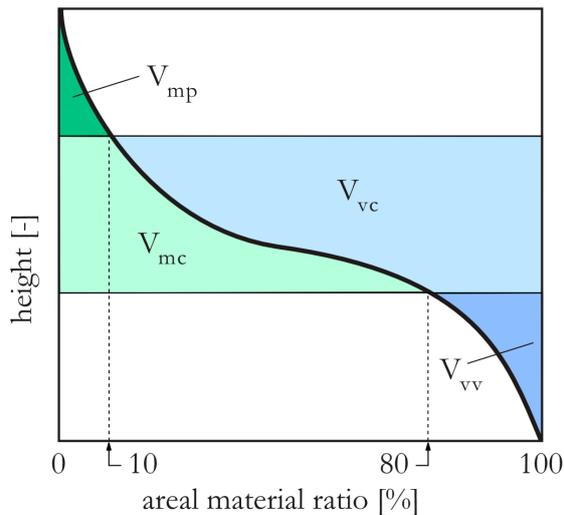
## Appendix

### Dale void volume ( $V_{vv}$ ) and core void volume ( $V_{vc}$ )

The dale void volume indicates a void volume in the range of  $p = 80\%$  to  $100\%$  of the areal material ratio. The core void volume is the void volume between the areal material ratios of  $p = 10\%$  and  $q = 80\%$ .

### Peak material volume ( $V_{mp}$ ) and core material volume ( $V_{mc}$ )

The peak material volume is the material volume when the areal material ratio equals  $p = 10\%$ . The core material volume denotes the material volume between  $p = 10\%$  and  $q = 80\%$  of the areal material ratio.



**Figure A3.1: Exemplary Abbot Firestone curve explaining the meaning of different functional volume parameters.** The areal material ratio  $S_{mr}$  is displayed for each height  $h$ . Based on this curve, four different functional volume parameters were determined here.

## A3.5 Functional parameters

The following functional parameters were calculated using an Abbot Firestone curve as mentioned above. To do so, a regression line was determined as a secant on the curve for which first the slope and then the square sum of the deviation in the vertical axis direction were reduced to a minimum (**Figure A3.2**). From that, the core surface is defined as the remaining surface, that is left, when those surface regions are removed, which do not contribute to the areal material ratio of the regression line.

### Core height ( $S_k$ ), reduced peak height ( $S_{pk}$ ), and reduced dale height ( $S_{vk}$ )

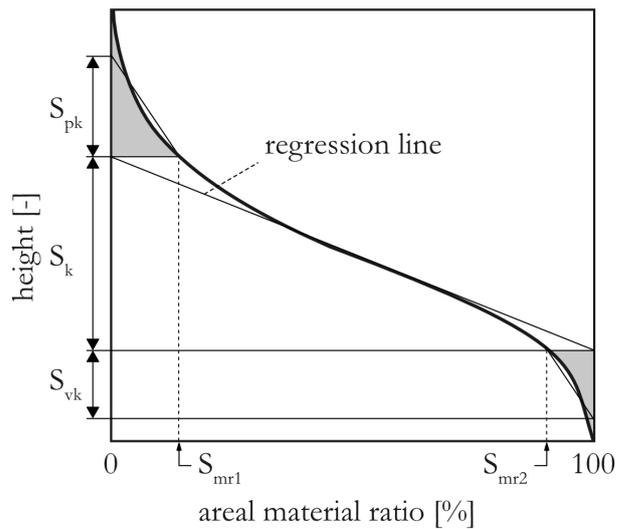
The core height describes the difference between the maximum core height and the minimum core height as determined from the regression line.  $S_{pk}$  denotes the average height of the reduced peak section, and  $S_{vk}$  describes the average depth of the reduced dale section. They are individually given by the height of the right-angled triangle whose area is equivalent to those grey peak/dale areas of the areal material ratio curve, that do not belong to the core section.

### Upper material ratio ( $S_{mr1}$ ) and lower material ratio ( $S_{mr2}$ )

$S_{mr1}$  and  $S_{mr2}$  refer to the areal material ratio at the intersection of the material ratio curve and the core's upper-part height and lower-part height, respectively.

### Peak extreme height: $S_{xp}$

The extreme peak height is defined as the difference in height between the areal material ratios of  $p = 2.5\%$  and  $q = 50\%$ .



**Figure A3.2: Exemplary Abbot Firestone curve explaining the meaning of different functional parameters.** A regression line is drawn to reach a minimum square sum of the deviations in vertical direction. From that, the core surface can be defined, and different functional parameters can be calculated.

### Appendix A4: Characterization of the tribological setup\*

For the tribological experiments performed with ocular samples, a custom-made assembly was used comprising a silicone-based contact lens holder (carrying the respective contact lens) and a full porcine eye that was fixed in a customized holder (see **chapter 2.5.2**). Considering the induced normal force of  $F_N = 0.14$  N, the resulting contact pressure  $p_0$  between the contact lens and the eye surface can be estimated based on the Hertzian model<sup>293</sup> as described in **equation A4.1**:

$$p_0 = \frac{2}{3} p_{max} = \frac{2}{3\pi} \sqrt[3]{6F_N * (E' \rho)^2} \quad \text{with} \quad \frac{1}{E'} = \frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2} \quad \text{and} \quad \rho = \frac{1}{r_1} + \frac{1}{r_2} \quad (\text{A4.1})$$

Here, the radii of the contact lens holder ( $r_1 = 7.35$  mm) and the eye ( $r_2 = 10.25$  mm) were used to describe the setup curvatures. Moreover, the Young's modulus  $E_1$  and the Poisson's ratio  $\nu_1$  of the contact lens holder were approximated with the respective values of pure PDMS ( $E_{\text{PDMS}} = 2$  MPa,  $\nu_{\text{PDMS}} = 0.49$ ).<sup>294</sup> Then, together with the Young's modulus and Poisson's ratio of porcine eyes ( $E_{\text{eye}} = 0.2$  MPa,  $\nu_{\text{eye}} = 0.5$ ),<sup>295</sup> the Hertzian contact theory returned an average contact pressure of approximately 30 kPa and a radius of the contact area of  $r_c = 1.23$  mm – the latter of which agrees well with the visual impression from the tribological measurements. With an average sliding velocity of  $v_s = 0.125$  mm s<sup>-1</sup> (estimated from the rotational speed used in the measurements at a radius of 0.615 mm), a lubricant viscosity approximated as  $\eta_{\text{buffer}} = 1$  mPas and the contact pressure determined above, a Sommerfeld number of  $S = \frac{v_s * \eta_{\text{buffer}}}{p_0} = 4.11 \times 10^{-12}$  can be estimated for the system.

Additionally, as the custom-made setup was used on a commercial rheometer, the measured friction coefficients were by default calculated under the assumption of a ball-on-three-pins geometry. To transfer those values to the ocular setup, a correction factor was applied to the determined friction coefficients according to **equation A4.2**. Here,  $\alpha$  denotes the inclination angle of the normal through the planar surface of a PDMS pin, which was defined by the rheometer manufacturer.

$$\mu_{\text{corrected}} = \mu_{\text{measured}} * \frac{3 * r_{\text{ball}} * \sin(\alpha)}{2 * r_c * \cos(\alpha)} \quad (\text{A4.2})$$

---

\* This section follows in part the publication Rickert *et al.*, *ACS Applied Materials & Interfaces* (2020)

## Appendix A5: Details on supervised ML for biofilm and plant samples\*

### A5.1 Bacterial growth media and biofilm formation process

If not stated otherwise, all chemicals were purchased from Carl Roth GmbH (Karlsruhe, Germany).

#### **MSgg (Minimal Salt Glycerol Glutamate) medium** (according to Branda *et al.*<sup>101</sup>)

5 mM potassium phosphate ( $K_2HPO_4$  [Carl Roth] and  $KH_2PO_4$  [AppliChem]),

100 mM 3-(*N*-Morpholino)propanesulfonic acid,

2 mM  $MgCl_2 \cdot 6H_2O$ ,

700  $\mu M$   $CaCl_2$ ,

50  $\mu M$   $MnCl_2 \cdot 4H_2O$ ,

50  $\mu M$   $FeCl_3 \cdot 6H_2O$  (Sigma-Aldrich),

1  $\mu M$   $ZnCl_2$ ,

2  $\mu M$  thiamine hydrochloride (AppliChem),

0.5 % (v/v) glycerol,

0.5 % (w/v) glutamate (Sigma-Aldrich),

50  $\mu g$   $mL^{-1}$  L-tryptophan (AppliChem),

50  $\mu g$   $mL^{-1}$  L-phenylalanine (AppliChem), and

50  $\mu g$   $mL^{-1}$  threonine (AppliChem).

#### **LB (Luria/Miller) medium** (purchased)

5  $g$   $L^{-1}$  yeast extract,

10  $g$   $L^{-1}$  tryptone, and

10  $g$   $L^{-1}$  sodium chloride (NaCl).

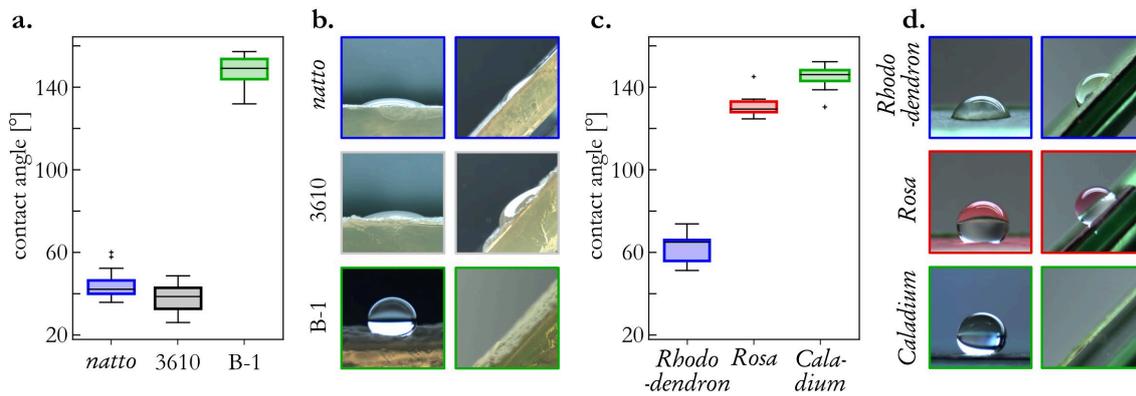
#### **Biofilm formation process**

*B. subtilis* NCIB 3610 (also referred to as “3610”) and B-1 strains were kindly provided by Roberto Kolter and Masaaki Morikawa, respectively; the *B. subtilis natto* (27E3) strain was obtained from the Bacillus Genetic Stock Center (BGSC, Columbus, USA). To prepare planktonic cultures of each strain, a small piece from a frozen bacterial glycerol stock was inoculated in 10 mL of 2.5 % (w/v) LB medium at 37 °C overnight while shaking at a speed of 200 rpm (Certomat BS-1; Sartorius AG, Göttingen, Germany). To obtain biofilm colonies from these three strains, seven separate 5  $\mu L$  droplets of a 16 h old bacterial overnight culture were placed onto 1.5 % (w/v) agar plates enriched with either 2.5 % (w/v) LB or MSgg medium. All biofilm variants were grown for 24 h at 37 °C and ~23 % humidity. The three *Bacillus subtilis* strains (*natto*, 3610, and B-1) used here to cultivate biofilms are closely related; nevertheless, the generated biofilms differ in terms of their physical behavior.<sup>296</sup>

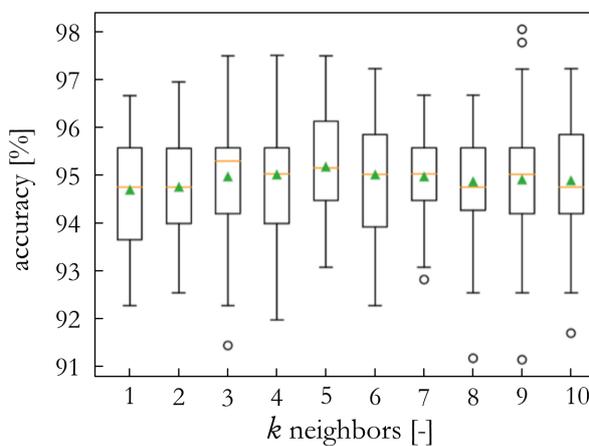
---

\* This section follows in part the publication Rickert *et al.*, *ACS Biomaterials Science & Engineering* (2021)

## A5.2 Wettability of biofilms (grown on LB agar) and plant samples



**Figure A5.1: Wetting characterization of *B. subtilis* strains grown on LB agar and different plant samples.** Contact angles (a) and exemplary images of tilt experiments (b) are shown as obtained for biofilms generated from the three *B. subtilis* strains *natto* (blue,  $n = 20$ ), NCIB 3610 (grey,  $n = 20$ ), and B-1 (green,  $n = 20$ ) when grown on LB agar. The contact angles were measured on the central regions of the colonies. Furthermore, contact angles (c) and exemplary images of tilt experiments (d) are shown as obtained for *Rhododendron* leaves (blue,  $n = 12$ ), rose petals (red,  $n = 12$ ), and *Caladium* leaves (green,  $n = 10$ ).

A5.3 Selection of  $k$  for the  $k$  Nearest Neighbors classifier

**Figure A5.2: Whisker-boxplots showing the prediction accuracy of the  $k$  Nearest Neighbors classifier with increasing values of  $k$ .** For these trials, the  $k$ NN classifier was applied to the initial dataset (*Bacillus subtilis* NCIB 3610, *B. subtilis* B-1, and *B. subtilis natto* biofilms, each cultivated on MSgg agar,  $n = 1788$ ). A repeated (10 times) stratified 5-fold cross validation (see **chapter 2.8.3**) was used to evaluate the accuracies. The green triangles indicate the mean obtained from the 50 (= 10 x 5) runs per trial.

### A5.4 Naïve Bayes' theorem

The Bayes theorem defines the relation between the conditional probability of a class  $y$  and a dependent feature vector  $(x_1, \dots, x_n)$  as follows:<sup>111,112</sup>

$$P(y|x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n | y)}{P(x_1, \dots, x_n)}. \quad (\text{A5.1})$$

Applying the naïve assumption of conditional independence, which is

$$P(x_i | y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i | y) \quad (\text{A5.2})$$

for all  $i$ , the following simplification results:

$$P(y|x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^n P(x_i | y)}{P(x_1, \dots, x_n)} \quad (\text{A5.3})$$

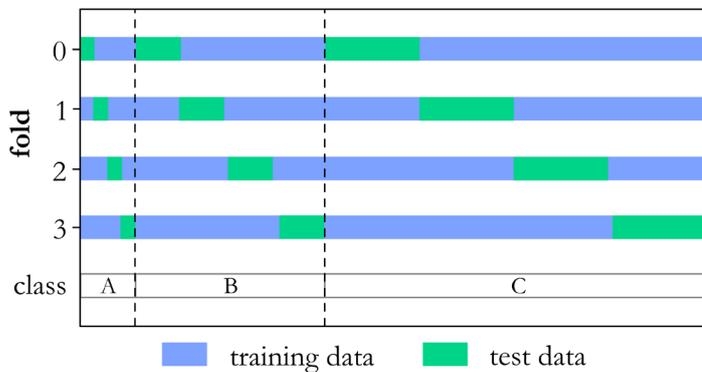
As  $P(x_1, \dots, x_n)$  is constant according to the input, a general classification rule as given by **equations A5.4** and **A5.5**, respectively, is obtained:

$$P(y | x_1, \dots, x_n) \propto P(y) \prod_{i=1}^n P(x_i | y) \quad (\text{A5.4})$$

$$\hat{y} = \arg \max_y P(y) \prod_{i=1}^n P(x_i | y) \quad (\text{A5.5})$$

### A5.5 Stratification of Data Splits

The models used in this thesis were evaluated using repeated  $k$ -fold cross-validation. Since these evaluations tend to have issues with imbalanced datasets,<sup>297</sup> a stratified version was used. This ensured that all created folds were good representatives of the whole dataset, *i.e.*, that the proportions of each class found in the original dataset was respected in all the folds. A visual explanation of the stratified sample split is depicted in **Figure A5.3**.



**Figure A5.3: Stratification of a dataset.** To compensate for imbalances of the class distribution within a dataset, the folds are created such that the class distribution in the respective training and test sets equals the class distribution of the whole dataset.

## A5.6 Sensitivity and specificity of the classifiers for predicting biofilms grown on MSgg agar

The *B. subtilis* strains *natto*, NCIB 3610, and B-1 were grown on MSgg agar. Data evaluation was conducted *via* repeated (10 times) stratified 5-fold cross validation – *i.e.*, 80 % of the data were used to train the algorithm, while the other 20 % were used for testing. The sensitivity and specificity were then calculated according to the following equations:

$$\text{sensitivity} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \quad (5.6)$$

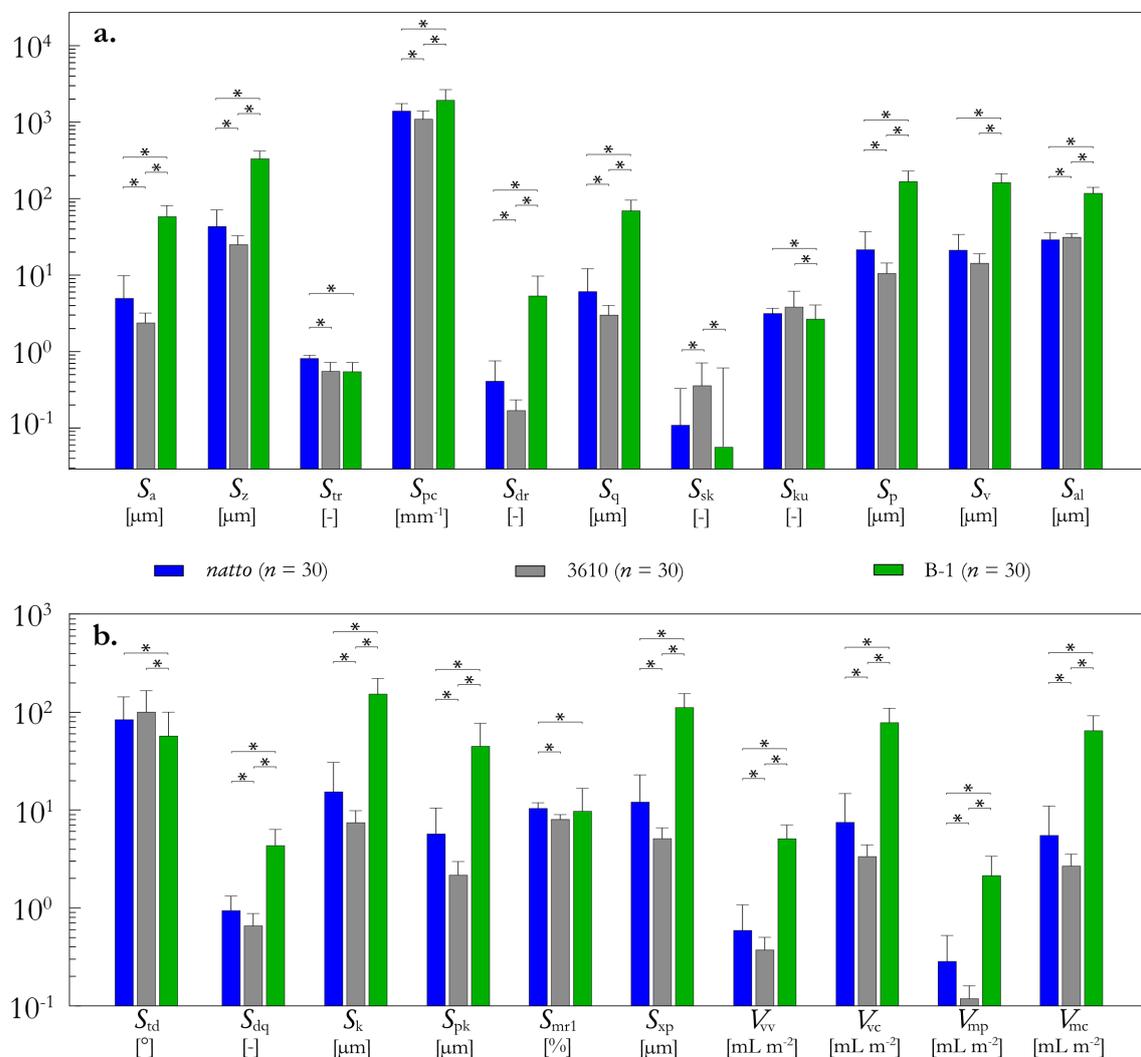
$$\text{specificity} = \frac{\text{true negative}}{\text{true negative} + \text{false positive}} \quad (5.7)$$

	5 Nearest Neighbors predicted			Gaussian Naïve Bayes predicted			Logistic Regression predicted			Random Forest predicted		
	<i>natto</i>	B-1	3610	<i>natto</i>	B-1	3610	<i>natto</i>	B-1	3610	<i>natto</i>	B-1	3610
actual <i>natto</i>	6799	20	231	6429	64	557	6750	40	260	6968	13	69
actual B-1	27	5086	7	0	5120	0	10	5087	23	5	5110	5
actual 3610	89	0	5831	105	20	5795	138	0	5782	97	0	5823
sensitivity	96.4 %	99.3 %	98.5 %	91.2 %	100 %	97.9 %	95.7 %	99.4 %	97.7 %	98.8 %	99.8 %	98.4 %
specificity	98.9 %	99.8 %	98.0 %	99.0 %	99.3 %	95.4 %	98.7 %	99.7 %	97.7 %	99.1 %	99.9 %	99.4 %

**Figure A5.4: Categorization results obtained with four different ML algorithms fed with biofilm data obtained on MSgg agar.** The confusion matrices compare the actual classes (*i.e.*, the names of the bacterial strains the biofilms were grown from) with the predicted classes when a 5 Nearest Neighbors classifier (**a**), a Gaussian Naïve Bayes model (**b**), Multinomial Logistic Regression (**c**), and a Random Forest classifier (**d**) are applied to the data. Evaluation was performed by repeated (10 times) 5-fold cross validation including data shuffling before each iteration – *i.e.*, each sample served 10 times as a test sample but was each time analyzed by a differently trained algorithm. The main diagonal (green boxes) contains the numbers of correctly classified samples whereas all other boxes represent wrongly assigned samples. Additionally, sensitivity and specificity for each prediction are depicted.

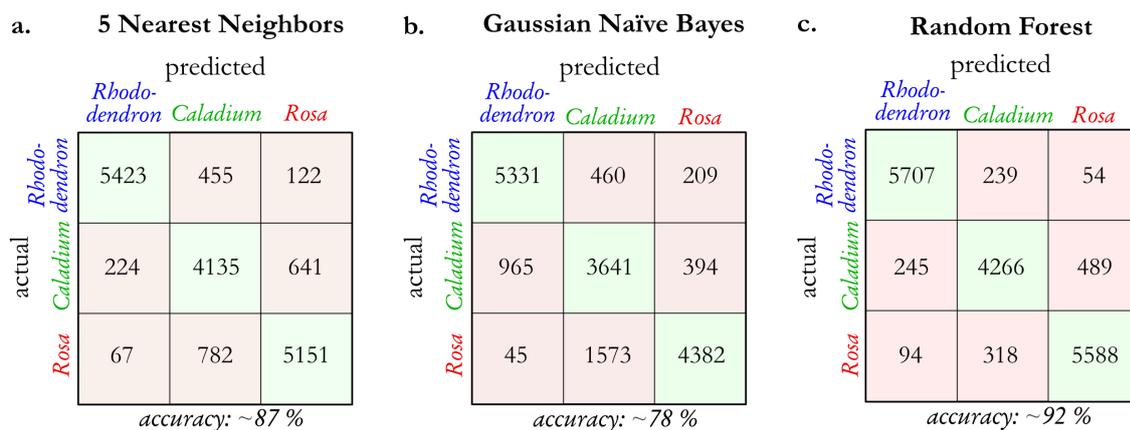
## A5.7 Topographical surface parameters of biofilms grown on LB agar

The *B. subtilis* strains *natto*, NCIB 3610, and B-1 were grown on LB agar as described in **Appendix A5.1**. From the central part of the colonies, 16 profilometric images were obtained from which the topographical parameters shown in **Figure A5.5** were determined.



**Figure A5.5: Micro-topographical parameters of biofilms generated by three *Bacillus subtilis* strains grown on LB agar.** The topographical parameters were calculated from profilometric images. Error bars denote the standard deviation obtained from  $n = 30$  images. Asterisks indicate statistical differences as assessed with a Wilcoxon-Mann-Whitney test applying a  $p$ -value of 0.05.

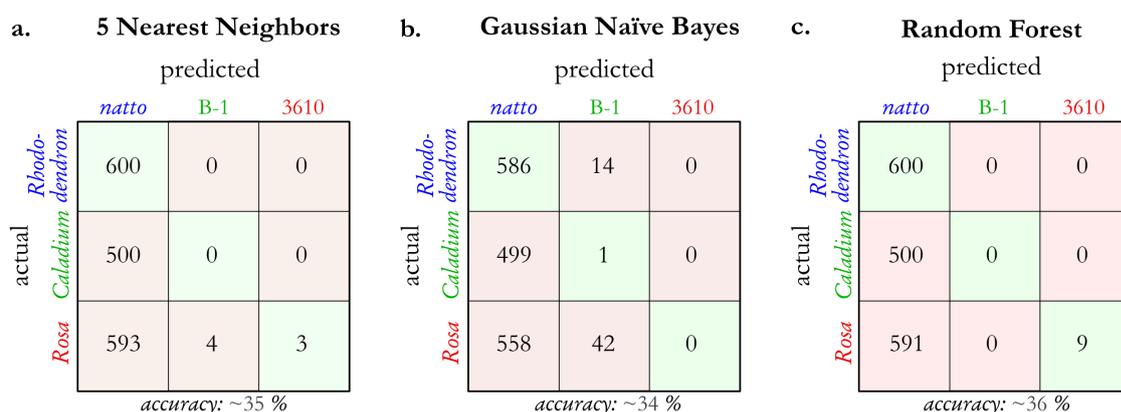
## A5.8 Validation of the classifiers applied on plant samples



**Figure A5.6: Categorization results obtained with three different ML algorithms fed with data obtained from plant samples.** The confusion matrices compare the actual classes (*i.e.*, the names of the plants) with the predicted classes when a 5 Nearest Neighbors classifier (a), a Gaussian Naïve Bayes model (b), and a Random Forest classifier (c) are applied to the data. The evaluation was performed by repeated (10 times) 5-fold cross validation with shuffling the dataset before each iteration – *i.e.*, each sample served 10 times as a test sample but was each time analyzed by a differently trained algorithm. The main diagonal (green boxes) contains the numbers of correctly classified samples whereas all other boxes represent wrongly assigned samples. The accuracies denote the ratio of correctly classified samples to all tested samples.

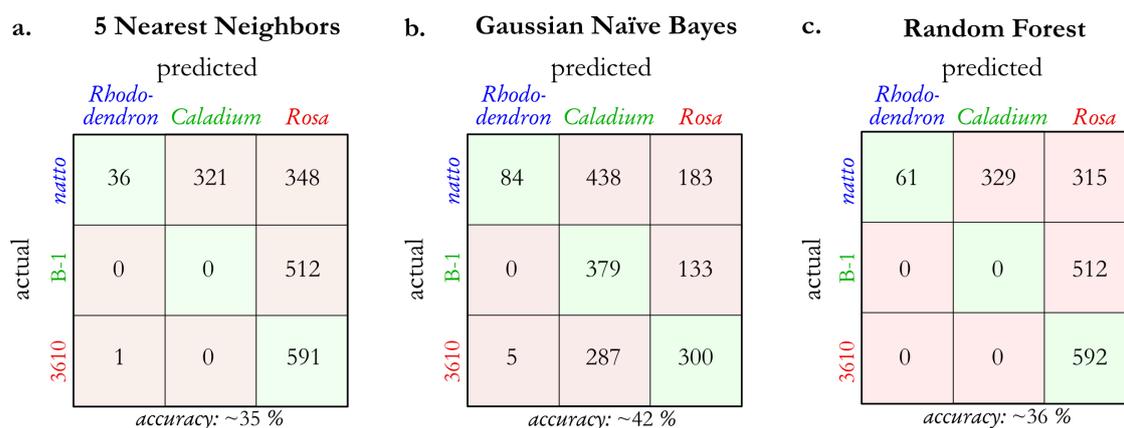
## A5.9 Evaluation of cross-trained algorithms

## Trained on biofilms, tested with plant samples



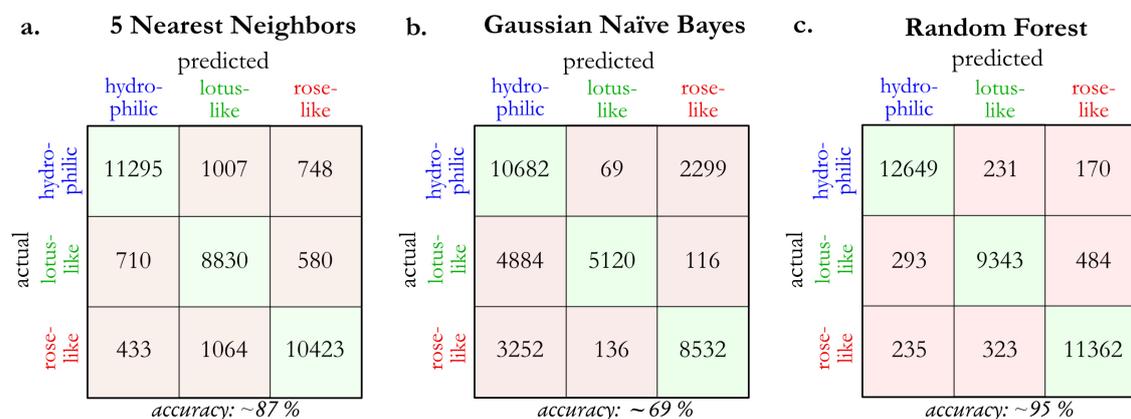
**Figure A5.7: Categorization results of cross-trained algorithms.** The confusion matrices compare the actual classes (*i.e.*, the plant names with the respective wetting behavior) with the predicted classes (*i.e.*, the names of the bacterial strains with the respective wetting behavior) when a 5 Nearest Neighbors classifier (a), a Gaussian Naïve Bayes model (b), and a Random Forest classifier (c) are applied to the data. The algorithms were trained on biofilms grown on MSgg agar and tested with plant samples. The colors indicate the wetting behavior of the surfaces: blue denotes hydrophilic behavior, green displays lotus-like hydrophobicity, and red indicates rose-like hydrophobicity. The main diagonal (green boxes) contains the numbers of correctly classified samples whereas all other boxes represent wrongly assigned samples. The accuracies denote the ratio of correctly classified samples to all tested samples.

## Trained on plant samples, tested with biofilms



**Figure A5.8: Categorization results of cross-trained algorithms.** The confusion matrices compare the actual classes (*i.e.*, the names of the bacterial strains with the respective wetting behavior) with the predicted classes (*i.e.*, the names of the plant names with the respective wetting behavior) when a 5 Nearest Neighbors classifier (a), a Gaussian Naïve Bayes model (b), and a Random Forest classifier (c) are applied to the data. The algorithms were trained on plant samples and tested with biofilms grown on MSgg agar. The colors indicate the wetting behavior of the surfaces: blue denotes hydrophilic behavior, green displays lotus-like hydrophobicity, and red indicates rose-like hydrophobicity. The main diagonal (green boxes) contains the numbers of correctly classified samples whereas all other boxes represent wrongly assigned samples. The accuracies denote the ratio of correctly classified samples to all tested samples.

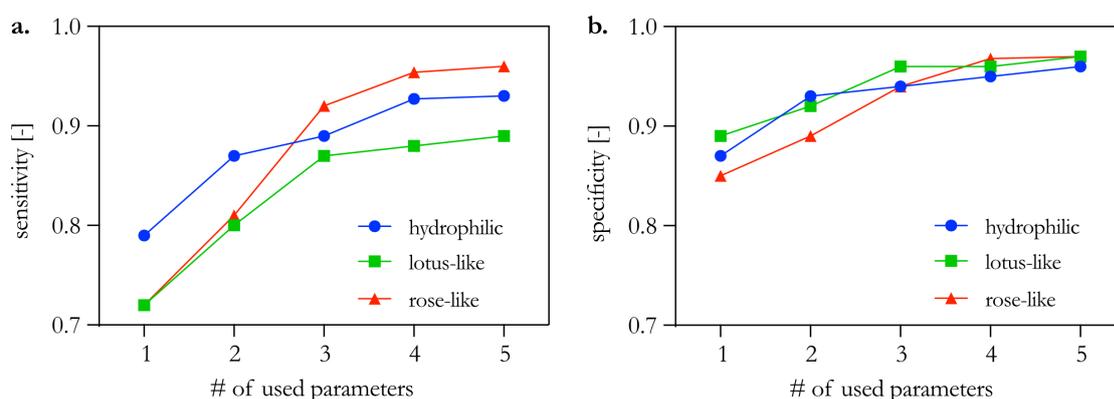
## A5.10 Evaluation of pooled datasets



**Figure A5.9: Categorization results of three different algorithms when fed a pooled biofilm/plant sample dataset.** The confusion matrices of three different classifiers used on the pooled dataset are shown. The dataset contains all data obtained for both, biofilms generated by three different bacterial strains grown on MSgg agar (*natto*: hydrophilic, 3610: hydrophobic rose-like, B-1: hydrophobic lotus-like) and plant samples (*Rhododendron*: hydrophilic, *Rosa*: hydrophobic rose-like, *Caladium*: hydrophobic lotus-like). All available metrological parameters were used. The evaluation was performed *via* repeated (10 times) 5-fold cross validation. The main diagonal (green boxes) contains the numbers of correctly classified samples whereas all other boxes represent wrongly assigned samples. The accuracies denote the ratio of correctly classified samples to all tested samples.

## A5.11 Sensitivity and specificity with reduced input sets

Here, the feature importance ranking described in **chapter 4.1** is analyzed in more detail. Therefore, additional accuracy measures are calculated for the Random Forest classifier. Instead of the 21 available parameters, reduced input parameter sets are selected based on the importance analysis depicted in **Figure 4.7**: the initial input set contains the five highest-ranked parameters (*i.e.*,  $S_{dr}$ ,  $S_{dq}$ ,  $V_{vv}$ ,  $S_{pc}$ , and  $S_a$ ). Based on this, the sensitivity and the specificity are calculated for each class (*i.e.*, for hydrophobic, lotus-like, and rose-like samples). In addition to the overall accuracy shown in **Figure 4.7b** of **chapter 4.1**, those two measures provide information if samples from certain classes are more difficult to classify than others. Both, sensitivity and specificity are then re-assessed while further reducing the number of input parameters, *i.e.*, by eliminating surface parameters in a stepwise manner (in full analogy to the accuracy calculation depicted in **Figure 4.7b**): starting with the  $S_a$  parameter, also  $S_{pc}$ ,  $V_{vv}$ , and  $S_{dq}$  are removed until only  $S_{dr}$  remains. Overall, it is observed that the sensitivity is lower than the specificity, and the former suffers more strongly than the latter when the parameter range is reduced (**Figure A5.10**). When comparing the three sample classes, the decrease of sensitivity is strongest for rose-like samples. This can be attributed to the intermediate surface roughness properties of this particular biofilm class,<sup>232</sup> which probably results in a higher probability of misclassification than for the two other classes.



**Figure A5.10: Sensitivity and specificity of the Random Forest classifier when fed with reduced input sets.** The accuracy measures were calculated (as described in **Appendix A5.6**) based on a repeated (10 times) 5-fold cross validation of the Random Forest classifier that was applied to the pooled dataset (including both, biofilm samples and plant samples). The range of five surface parameters ( $S_{dr}$ ,  $S_{dq}$ ,  $V_{vv}$ ,  $S_{pc}$ , and  $S_a$ ) was reduced in a stepwise manner: starting with the  $S_a$  parameter, also  $S_{pc}$ ,  $V_{vv}$  and  $S_{dq}$  were eliminated one after each other until only  $S_{dr}$  was left.

## A5.12 Evaluation by a cross-trained Deep Neural Network

		a. trained: biofilms tested: plants			b. trained: plants tested: biofilms		
		predicted			predicted		
		hydro- philic	lotus- like	rose- like	hydro- philic	lotus- like	rose- like
actual	hydro- philic	600	0	0	0	0	705
	lotus- like	500	0	0	0	0	512
	rose- like	600	0	0	0	0	592
		accuracy: ~35 %			accuracy: ~33 %		

**Figure A5.11: Categorization results of a Deep Neural Network trained and tested on different datasets.**

The confusion matrices of the Deep Neural Network either trained with biofilm data and tested with plant samples (a) or *vice versa* (b) are displayed. The biofilm dataset contains biofilms grown from three different bacterial strains on MSgg agar (*natto*: hydrophilic, 3610: hydrophobic rose-like, B-1: hydrophobic lotus-like). The plant set contains three different plant leaves/petals (*Rhododendron*: hydrophilic, *Rosa*: hydrophobic rose-like, *Caladium*: hydrophobic lotus-like). All available features were used. The main diagonal (green boxes) contains the numbers of correctly classified samples whereas all other boxes represent wrongly assigned samples. The accuracies denote the ratio of correctly classified samples to all tested samples.

## Appendix A6: Molecular datasets\*

### A6.1 Calculation parameters

**Table A6.1: Physicochemical properties determined for analyzed molecules.** A set of 9 physico-chemical properties was used to characterize each molecule of the created datasets. Those properties were determined using the Calculator Plugins of MarvinSketch (see **chapter 2.9.1**) applying the listed calculation parameters.

Feature	Calculation parameters
rotatable bond count	decimal places: 2; aromatization method: general; single fragment mode: no
van der Waals volume	energy unit: kcal/mol; decimal places: 2; set MMFF94 optimization: no; set projection optimization: no; calculate for lowest energy conformer: if molecule is in 2D; optimization limit: normal
molecular weight	recognize formula in pseudo labels: yes; use D / T symbols for deuterium / tritium: yes; single fragment mode: no
charge	decimal places: 2; show charge distribution: yes; pH step size: 0.1; keep explicit hydrogens: no; consider tautomerization / resonance: yes
dipole moment	-
partition coefficient	method: consensus; Cl <sup>-</sup> concentration (mol/dm <sup>3</sup> ): 0.1; Na <sup>+</sup> /K <sup>+</sup> concentration (mol/dm <sup>3</sup> ): 0.1; consider tautomerization / resonance: no
aromatic ring count	decimal places: 2; aromatization method: general; single fragment mode: no
hydrogen bond acceptor sites (average at pH 7.4)	decimal spaces: 2; donor: yes; acceptor: yes; exclude sulfur atoms from acceptors: yes; exclude halogens from acceptors: yes; show microspecies data by pH: yes; pH lower limit: 7; pH upper limit: 8; pH step size: 0.1; display major microspecies: no
hydrogen bond donor sites (average at pH 7.4)	decimal spaces: 2; donor: yes; acceptor: yes; exclude sulfur atoms from acceptors: yes; exclude halogens from acceptors: yes; show microspecies data by pH: Yes; pH lower limit: 7; pH upper limit: 8; pH step size: 0.1; display major microspecies: no

\* This section follows in part the publication Rickert *et al.*, *APL Machine Learning* (2023)

## A6.2 Pubchem references of the analyzed molecules

**Table A6.2 Molecules included in the molecular datasets.** Four different datasets were created containing 14 antibiotics, 10 antioxidants, 12 fluorophores, and 12 vitamins, respectively.

Antibiotics	PubChem CID	URL
Ampicillin	6249	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/6249">https://pubchem.ncbi.nlm.nih.gov/compound/6249</a>
Bacitracin	6714010	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/6714010">https://pubchem.ncbi.nlm.nih.gov/compound/6714010</a>
Bambermycin	6433554	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/6433554">https://pubchem.ncbi.nlm.nih.gov/compound/6433554</a>
Chloramphenicol	5959	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/5959">https://pubchem.ncbi.nlm.nih.gov/compound/5959</a>
Erythromycin	12560	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/12560">https://pubchem.ncbi.nlm.nih.gov/compound/12560</a>
Friulimicin B	56842195	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/56842195">https://pubchem.ncbi.nlm.nih.gov/compound/56842195</a>
Gentamicin	3467	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/3467">https://pubchem.ncbi.nlm.nih.gov/compound/3467</a>
Kanamycin	6032	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/6032">https://pubchem.ncbi.nlm.nih.gov/compound/6032</a>
Piperacillin	43672	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/43672">https://pubchem.ncbi.nlm.nih.gov/compound/43672</a>
Rifampicin	135398735	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/135398735">https://pubchem.ncbi.nlm.nih.gov/compound/135398735</a>
Spectinomycin	15541	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/15541">https://pubchem.ncbi.nlm.nih.gov/compound/15541</a>
Streptomycin	19649	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/19649">https://pubchem.ncbi.nlm.nih.gov/compound/19649</a>
Tetracycline	54675776	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/54675776">https://pubchem.ncbi.nlm.nih.gov/compound/54675776</a>
Vancomycin	14969	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/14969">https://pubchem.ncbi.nlm.nih.gov/compound/14969</a>

Antioxidants	PubChem CID	URL
N-Acetyl-L-Cysteine	12035	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/12035">https://pubchem.ncbi.nlm.nih.gov/compound/12035</a>
Bendazac	2313	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/2313">https://pubchem.ncbi.nlm.nih.gov/compound/2313</a>
Bilirubin	5280352	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/5280352">https://pubchem.ncbi.nlm.nih.gov/compound/5280352</a>
Butylhydroxytoluene	31404	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/31404">https://pubchem.ncbi.nlm.nih.gov/compound/31404</a>
4-Butyl-resorcinol	205912	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/205912">https://pubchem.ncbi.nlm.nih.gov/compound/205912</a>
Caffeic acid	689043	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/689043">https://pubchem.ncbi.nlm.nih.gov/compound/689043</a>
Capsanthin	5281228	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/5281228">https://pubchem.ncbi.nlm.nih.gov/compound/5281228</a>
Carnosic acid	65126	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/65126">https://pubchem.ncbi.nlm.nih.gov/compound/65126</a>
Trolox	40634	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/40634">https://pubchem.ncbi.nlm.nih.gov/compound/40634</a>
Uric acid	1175	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/1175">https://pubchem.ncbi.nlm.nih.gov/compound/1175</a>
Hesperidin	10621	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/10621">https://pubchem.ncbi.nlm.nih.gov/compound/10621</a>
L-Cysteine	5862	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/5862">https://pubchem.ncbi.nlm.nih.gov/compound/5862</a>

Fluorophores	PubChem CID	URL
ATTO 390	16218729	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/16218729">https://pubchem.ncbi.nlm.nih.gov/compound/16218729</a>
ATTO 425-2	25164096	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/25164096">https://pubchem.ncbi.nlm.nih.gov/compound/25164096</a>
ATTO 488	102227067	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/102227067">https://pubchem.ncbi.nlm.nih.gov/compound/102227067</a>
ATTO 495	124202518	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/124202518">https://pubchem.ncbi.nlm.nih.gov/compound/124202518</a>
Calcein	65079	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/65079">https://pubchem.ncbi.nlm.nih.gov/compound/65079</a>
Calcein AM	4126474	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/4126474">https://pubchem.ncbi.nlm.nih.gov/compound/4126474</a>
5-Carboxy-fluorescein	123755	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/123755">https://pubchem.ncbi.nlm.nih.gov/compound/123755</a>
4',6-Diamidino-2-phenylindole	2954	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/2954">https://pubchem.ncbi.nlm.nih.gov/compound/2954</a>
Doxorubicin Hydrochloride	443939	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/443939">https://pubchem.ncbi.nlm.nih.gov/compound/443939</a>
Fluorescein	91886391	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/91886391">https://pubchem.ncbi.nlm.nih.gov/compound/91886391</a>
Sybr green I	10436340	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/10436340">https://pubchem.ncbi.nlm.nih.gov/compound/10436340</a>
Oregon Green 488	10292443	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/10292443">https://pubchem.ncbi.nlm.nih.gov/compound/10292443</a>

<b>Vitamines</b>	<b>PubChem CID</b>	<b>URL</b>
Alpha-Tocopherol	14985	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/14985">https://pubchem.ncbi.nlm.nih.gov/compound/14985</a>
Ascorbic acid	54670067	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/54670067">https://pubchem.ncbi.nlm.nih.gov/compound/54670067</a>
Beta-Carotene	5280489	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/5280489">https://pubchem.ncbi.nlm.nih.gov/compound/5280489</a>
Cholecalciferol	5280795	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/5280795">https://pubchem.ncbi.nlm.nih.gov/compound/5280795</a>
Ergocalciferol	5280793	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/5280793">https://pubchem.ncbi.nlm.nih.gov/compound/5280793</a>
Folic acid	135398658	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/135398658">https://pubchem.ncbi.nlm.nih.gov/compound/135398658</a>
Menadione	4055	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/4055">https://pubchem.ncbi.nlm.nih.gov/compound/4055</a>
Menatetrenone	5282367	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/5282367">https://pubchem.ncbi.nlm.nih.gov/compound/5282367</a>
NHS-Biotin	6710714	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/6710714">https://pubchem.ncbi.nlm.nih.gov/compound/6710714</a>
Phylloquinone	5284607	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/5284607">https://pubchem.ncbi.nlm.nih.gov/compound/5284607</a>
Retinyl acetate	638034	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/638034">https://pubchem.ncbi.nlm.nih.gov/compound/638034</a>
Thiamine hydrochloride	6202	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/6202">https://pubchem.ncbi.nlm.nih.gov/compound/6202</a>

## A6.3 Molecular datasets

## Feature vector

- RotBd rotatable bond count [-]
- vdWv van der Waals volume [ $\text{\AA}^3$ ]
- MolW molecular weight [ $\text{g mol}^{-1}$ ]
- Chr total charge at pH 7.4 [e]
- DipM dipole moment [Debye]
- PartCo partition coefficient ( $\log P$ ) of non-ionic species [-]
- ArRiC aromatic ring count [-]
- HyBAS hydrogen bond acceptor sites [-]
- HYBDS hydrogen bond donor sites [-]

**Table A6.3: Molecular datasets.** Each molecule is characterized by 9 physico-chemical features.

Antibiotic	RotBd	vdWv	MolW	Chr	DipM	PartCo	ArRiC	HyBAS	HyBDS
Piperacillin	6	437.89	517.56	-1.0	1.75	-0.26	1	15	2
Tetracycline	2	379.21	444.440	-1.19	8.61	-3.82	1	18.84	5.16
Erythromycin	7	726.07	733.937	0.98	10.94	2.80	0	24.02	5.89
Chloramphenicol	6	249.14	323.13	-0.05	6.95	0.88	1	11.10	2.95
Gentamicin	7	458.68	477.603	4.39	4.62	-3.14	0	14.55	15.45
Vancomycin	13	1214.40	1449.27	0.94	8.47	-1.18	5	45.11	21.89
Kanamycin	6	429.99	484.503	3.35	1.38	-7.06	0	22.59	18.41
Streptomycin	9	502.21	581.580	2.83	5.81	-6.76	0	28.18	18.82
Rifampicin	5	758.71	822.953	-0.5	8.09	3.62	2	25.46	5.54
Spectinomycin	2	296.47	323.353	1.07	5.08	-1.40	0	14.90	6.10
Bacitracin	31	1314.37	1422.71	0.03	3.79	-2.52	2	35.91	20.09
Ampicillin	4	298.75	349.41	-0.60	4.41	0.26	1	9.60	3.40
Friulimicin B	23	1198.18	1303.480	-2.11	9.39	-4.58	0	41.11	14.89
Bamermycin	39	1416.16	1597.609	-2	7.81	-1.77	0	65	16

Fluorophore	RotBd	vdWv	MolW	Chr	DipM	PartCo	ArRiC	HyBAS	HyBDS
390 carboxy <sup>a</sup>	4	325.7	343.423	-1	13.83	4.08	2	8	0
425 carboxy <sup>a</sup>	7	370.58	401.459	-1	16.19	4.10	2	10	0
488 carboxy <sup>a</sup>	7	452.94	576.57	-2	19.82	0.34	4	24.95	4.05
495 carboxy <sup>a</sup>	6	338.66	352.457	0	1.28	-0.93	3	7	0
Sybr green I <sup>b</sup>	9	491.69	509.74	1.99	3.32	4.83	4	2.01	0.99
Fluorescein (sodium salt) <sup>b</sup>	0	273.17	332.311	-0.07	5.66	3.88	3	6.05	1.95
DAPI <sup>b</sup>	3	250.39	277.331	2	5.13	1.48	3	2	9
Calcein <sup>b</sup>	12	510.57	622.539	-3.2	6.90	1.68	3	27.25	2.75
5-Carboxy-fluorescein <sup>b</sup>	1	300.63	376.320	-1.07	3.08	3.54	3	11.05	1.95
Calcein AM <sup>c</sup>	32	835.10	994.865	0	13.07	1.52	3	24	0
Oregon Green 488 <sup>c</sup>	0	282.72	368.292	-0.62	7.10	4.17	3	6.47	1.53
Doxorubicin hydrochloride <sup>d</sup>	5	461.53	543.525	0.78	3.75	1.50	2	22.0	7.80
390 carboxy <sup>a</sup>	4	325.7	343.423	-1	13.83	4.08	2	8	0
425 carboxy <sup>a</sup>	7	370.58	401.459	-1	16.19	4.10	2	10	0

<sup>a</sup> ATTO-Tec, <sup>b</sup> Sigma-Aldrich, <sup>c</sup> Thermo Fisher, <sup>d</sup> AppliChem

Vitamin	RotBd	vdWv	MolW	Chr	DipM	PartCo	ArRiC	HyBAS	HyBDS
alpha-Tocopherol	12	476.24	430.717	0	0.67	10.51	1	4	1
Ascorbic acid	2	139.85	176.124	-0.16	4.47	-1.91	0	11	3
beta-Carotene	10	598.29	536.888	0	0.06	11.12	0	0	0
Cholecaliferol	6	423.60	384.648	0	1.59	7.13	0	2	1
Folic acid	9	365.61	441.404	-2	12.79	-0.91	3	19	5
NHS-Biotin	7	287.01	341.38	0	4.25	-0.31	0	8	2
Retinyl acetate	7	353.07	328.496	0	1.42	5.14	0	2	0
Thiamine hydrochloride	4	237.79	265.35	1.12	7.29	-3.10	2	4.99	3.01
Vitamin K1	4	486.44	450.707	0	4.76	9.70	1	4	0
Vitamin K2	11	468.48	444.659	0	4.79	4.48	1	4	0
Menadione	0	154.74	172.183	0	1.05	1.89	1	4	0
Ergocalciferol	5	434.87	396.659	0	1.61	7.05	0	2	1

Antioxidant	RotBd	vdWv	MolW	Chr	DipM	PartCo	ArRiC	HyBAS	HyBDS
Hesperidin	7	514.06	610.565	-0.06	4.72	-0.31	2	30.06	7.94
L-Cystein	2	103.40	121.15	-0.02	1.62		0	5.02	3.98
N-Acetyl-L-Cystein	3	139.71	163.19	-1	3.58	-0.71	0	7	2
Trolox	1	233.60	250.294	-1	3.49	3.66	1	9	1
Uric acid	0	121.49	168.112	-0.23	4.18	-1.54	2	7.22	3.39
Bendazac	5	248.16	282.299	-1	4.75	3.06	3	9	0
Bilirubin	12	531.94	584.673	-2	8.79	3.12	2	14	4
4-Butyl-resorcinol	3	166.58	166.220	-0.01	1.58	3.21	1	4.01	1.99
Butylhydroxytoluene	2	243.21	220.356	0	2.81	5.27	1	2	1
Caffeic acid	2	154.68	180.159	-1.01	12.60	1.53	1	9.01	1.99
Capsanthin	11	624.05	584.885	0	42.08	8.51	0	6	2
Carnosic Acid	2	324.54	332.44	-1.01	3.01	5.14	1	9.01	1.99
Hesperidin	7	514.06	610.565	-0.06	4.72	-0.31	2	30.06	7.94
L-Cystein	2	103.40	121.15	-0.02	1.62		0	5.02	3.98

## Appendix A7: NETCORE vs. other correlation-based approaches\*

The performance of the NETCORE algorithm is compared to two basic correlation-based feature selection approaches: For the first method ('random' selection), a correlation matrix is calculated based on the Pearson's correlation coefficient as described for the NETCORE algorithm. Then, each column (corresponding to one feature) is searched for correlation coefficients that equal or exceed a predefined correlation threshold. If such coefficients are detected, the column (*i.e.*, the feature) and its corresponding row are removed from the correlation matrix. Once all columns have been checked, only those features remain in the correlation matrix, that do not exhibit a high-enough correlation to any other remaining feature. The second method ('upper triangle' method) is a more advanced feature elimination strategy that filters features based on pairwise correlations while making sure to not drop features that are needed to represent others. Here, a very similar elimination strategy as described above is followed; however, instead of considering the whole correlation matrix, only the upper right triangle of the matrix is evaluated.

The following comparison between the two simple correlation-based algorithms and the NETCORE method is conducted with 6 different datasets: In addition to the dataset introduced in **chapter 4.2**, that contains 14 antibiotics characterized by 9 physico-chemical characteristics, similar datasets describing the same set of molecular properties were generated for three additional molecular classes, *i.e.*, fluorophores, antioxidants, and vitamins (see **Appendix A6**). Furthermore, another dataset was generated by pooling all entries of all four molecular classes. Last, a bigger (in terms of both, sample size and dimensionality) dataset was obtained from Wu *et al.*<sup>270</sup>. This dataset (in the remainder referred to as 'inhibitors dataset') contains binary labels of binding results for 1513 (putative) inhibitors of human  $\beta$ -secretase 1; here, each inhibitor molecule is characterized by 590 physico-chemical features.

When analyzing this big 'inhibitors' dataset with the random correlation-based elimination method (applying a correlation threshold of  $t = 0.6$ ), a reduced feature vector containing 74 features is obtained. For all eliminated features, the representation strength, *i.e.*, the maximal correlation strength these features have to a feature from the reduced feature vector, is calculated. It is observed that this random elimination strategy based on the correlation matrix alone is not able to provide a reduced feature vector that sufficiently represents all eliminated features according to the predefined correlation threshold (some features are only represented with a correlation of 0.25 even though a correlation threshold of  $t = 0.6$  was applied; a detailed overview of the result is provided in **Table A7.1**). This problem mainly arises from the fact, that a feature that was previously chosen to represent an eliminated feature, can afterwards be eliminated itself. In contrast, the NETCORE algorithm creates a reduced feature vector that represents all eliminated features with representation strengths that exceed the predefined correlation threshold.

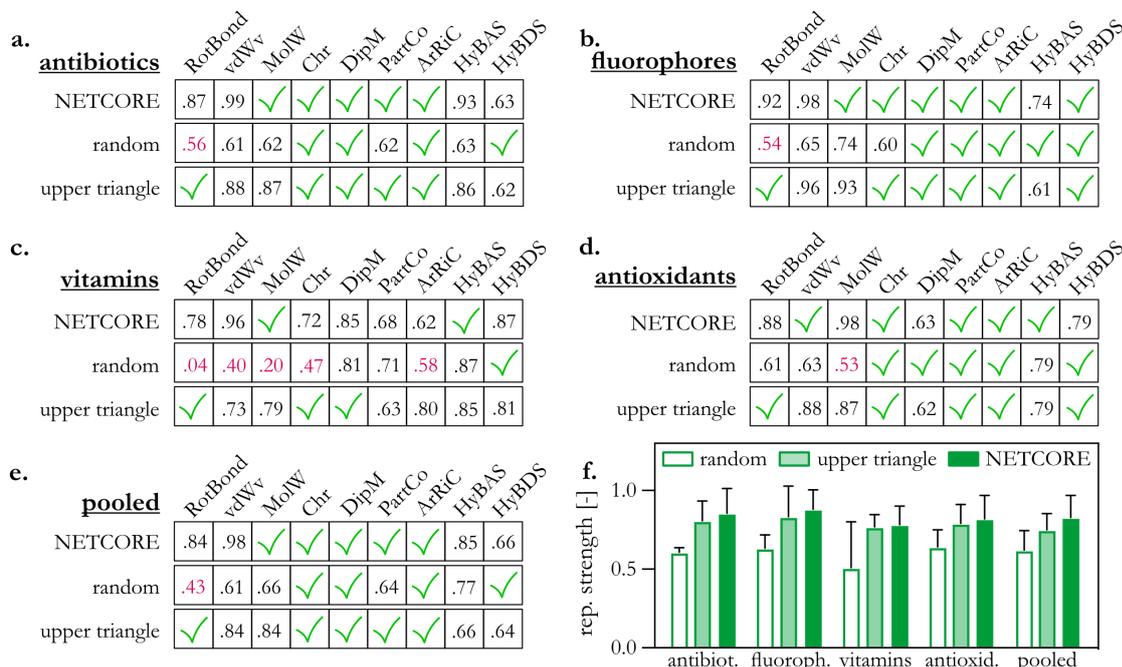
To make sure to not drop features that are needed to represent others, it is possible to analyze only the upper right triangle of the correlation matrix. When applying such an 'upper triangle' method to the inhibitors dataset, a feature vector is obtained that indeed properly represents all eliminated features. However, the such generated reduced feature vector contains 91 features, whereas the NETCORE algorithm is able to reduce the original feature vector to only 84 features (when applying the same correlation threshold of  $t = 0.6$ ; **Table A7.1**). At this point, it is important to reemphasize that the primary goal of NETCORE is to identify the smallest possible feature vector; then, for this minimal number of features, the representation strength is optimized. Hence, even though the upper triangle method can identify a suitable feature vector, its selection result is sub-optimal in terms of elimination efficiency. Most likely, this occurs since a feature can only be eliminated if it is redundant to a feature

---

\* This section follows in part the publication Rickert *et al.*, *APL Machine Learning* (2023)

that is described by a column of the correlation matrix which is located lefthand of the feature to be eliminated; this entails two major complications: First, the feature elimination process is subject to a certain bias: features that are located in the ‘beginning’ of the correlation matrix tend to stay in the feature vector, whereas features that are located in ‘later’ columns are likely to be discarded. Second, the reduced feature vector obtained from the upper triangle method strongly depends on the order by which the features appear in the correlation matrix. Thus, certain constellations in the correlation matrix also allow only for certain eliminations, and this limits the identification of an ‘optimal’ reduced feature vector.

When applying these two correlation-based feature elimination strategies to the 5 small molecular datasets (antibiotics, fluorophores, antioxidants, vitamins, and pooled), the outcome is sub-optimal as well (the reduced feature vectors created by the different algorithms are depicted in the supporting information **Figure A7.1**): For all 5 small molecular datasets, the random feature selection based on the full correlation matrix delivers clearly unsatisfactory results (**Figure A7.1** ‘random’): The representation strength of all eliminated features is considerably lower than those of the feature sets identified by the NETCORE algorithm – for several features, this value is even far below the initially selected correlation threshold. As mentioned above, this problem mainly arises from the fact, that a feature that was previously chosen to represent an eliminated feature, can afterwards be eliminated itself. A particularly pronounced example of this effect is observed for the ‘vitamins’ dataset. Here, only one feature is kept in the reduced feature vector, but 5 of the eliminated features are not sufficiently represented: the feature ‘rotatable bonds’, for instance, only exhibits a representation strength of 0.04. In contrast, when using the NETCORE algorithm, all features are sufficiently represented by the reduced feature vector – for all molecular datasets tested.



**Figure A7.1: Reduced feature vectors and representation strengths obtained for different small molecular datasets using different correlation-based feature elimination strategies.** Three different correlation-based algorithms were employed to perform a feature elimination on the molecular datasets. For all tests and methods, a correlation threshold of  $t = 0.6$  was chosen. (a-e) Features that are kept in the new, reduced feature vector are marked with a green tick. For eliminated features, the maximal correlation strength those features exhibit to one of the features included in the reduced feature vector are displayed. (f) Mean representation strengths for eliminated features are displayed. Error bars denote the standard deviation obtained from the different numbers  $n$  of the eliminated features (according to a-e  $n \geq 4$  for all bars).

Different from the random feature selection algorithm, the ‘upper triangle’ approach does create a reduced feature vector that sufficiently represents all eliminated features (based on the predefined correlation threshold of  $t = 0.6$ ). However, even though the created reduced feature vector contains the same number of features as the one created by NETCORE, the representation strength of the eliminated features is lower when the ‘upper triangle’ approach is used. As mentioned above, here, the elimination is strongly influenced by the sequence of the features in the correlation matrix. Hence, the feature ‘rotatable bonds’, for instance, is kept in the reduced feature vectors of all 5 molecular classes – just because it is the first feature to be analyzed and has no other feature it can be considered to be redundant to. And this leads to a sub-optimal outcome: The feature vectors obtained when analyzing the pooled dataset with the NETCORE algorithm and the upper triangle method are very similar (four out of five features are identical). However, whereas the upper triangle method selects the rotatable bond count to be included into the feature vector (for reasons described before), the NETCORE algorithm instead chooses the more central ‘molecular weight’ feature. With this small but important difference in choice, the mean representation strength of the eliminated features is improved from 0.75 to 0.83.

In conclusion, the NETCORE algorithm is able to outperform basic correlation-based feature elimination strategies as it more efficiently reduces the feature space while optimizing the representation strength of eliminated features.

**Table A7.1: Elimination result achieved using three different correlation-based strategies on the ‘inhibitors’ dataset.** As described in the previous section, a random correlation-based approach (‘rand’) and a correlation-based elimination considering only the upper triangle of the correlation matrix (‘uptri’) were compared to the NETCORE algorithm. Features included into the reduced feature vector are marked with a ‘-’, whereas for eliminated features, the representation strength (*i.e.*, the maximum correlation the feature has to a feature that is included in the reduced feature vector) is displayed.

feature	rand	uptri	NETC	feature	rand	uptri	NETC
MW	0.74	-	0.92	tCH_Cnt	1.00	1.00	1.00
AlogP	-	-	-	dsCH_Cnt	0.96	0.95	0.96
HBA	0.61	0.61	-	aaCH_Cnt	0.71	-	0.97
HBD	0.70	0.76	0.80	sssCH_Cnt	0.75	0.71	0.79
RB	0.74	0.89	0.83	tsC_Cnt	0.86	0.94	0.94
HeavyAtomCount	0.74	0.99	0.83	dssC_Cnt	0.81	-	0.85
ChiralCenterCount	-	-	-	aasC_Cnt	0.74	-	-
ChiralCenterCountAllPossible	0.66	0.72	0.78	aaaC_Cnt	0.99	1.00	-
RingCount	0.90	-	-	ssssC_Cnt	0.68	0.82	0.82
PSA	0.71	0.77	0.78	sNH3_Cnt	1.00	1.00	1.00
Estate	0.71	0.77	0.78	sNH2_Cnt	0.66	0.70	0.70
MR	0.73	0.97	0.93	ssNH2_Cnt	0.78	0.99	0.63
Polar	0.75	0.95	0.90	dNH_Cnt	1.00	1.0	1.0
sCH3_Key	0.92	-	-	ssNH_Cnt	0.68	0.73	0.73
dCH2_Key	1.00	-	-	aaNH_Cnt	0.97	0.97	0.97
ssCH2_Key	0.30	-	0.75	tN_Cnt	0.99	0.69	0.69
tCH_Key	1.00	-	-	sssNH_Cnt	0.93	0.94	-
dsCH_Key	0.99	-	-	dsN_Cnt	0.59	0.96	0.96
aaCH_Key	0.35	-	0.67	aaN_Cnt	0.84	0.86	0.99
sssCH_Key	0.55	-	-	sssN_Cnt	0.93	0.94	-
tsC_Key	0.69	-	-	aasN_Cnt	0.99	1.0	1.0
dssC_Key	0.69	-	-	sOH_Cnt	0.66	0.77	0.72
aasC_Key	0.35	1.00	0.67	dO_Cnt	0.76	0.77	0.71
aaaC_Key	0.99	-	1.00	ssO_Cnt	0.81	0.82	-
ssssC_Key	0.54	-	-	aaO_Cnt	0.99	1.0	1.0
sNH3_Key	1.00	-	-	aOm_Cnt	0.79	0.87	0.92
sNH2_Key	0.66	0.70	0.70	sOm_Cnt	0.89	0.81	0.92

ssNH2_Key	0.78	-	0.64	sF_Cnt	0.70	0.79	1.00
dNH_Key	1.00	-	-	ssS_Cnt	0.98	0.98	-
ssNH_Key	0.52	-	-	aaS_Cnt	0.99	1.0	1.0
aaNH_Key	1.00	-	-	dssS_Cnt	0.60	1.0	1.0
tN_Key	1.00	0.69	0.69	ddssS_Cnt	0.66	1.00	0.99
sssNH_Key	1.00	-	-	sCl_Cnt	0.80	0.96	1.00
dsN_Key	0.61	-	-	sBr_Cnt	0.68	1.0	1.0
aaN_Key	0.99	-	0.87	sI_Cnt	0.87	1.00	0.99
sssN_Key	0.97	-	0.94	sCH3_Sum	0.76	0.99	0.67
aasN_Key	0.99	-	-	dCH2_Sum	1.00	1.00	1.00
sOH_Key	0.68	0.82	0.78	ssCH2_Sum	0.79	-	0.68
dO_Key	0.42	-	0.68	tCH_Sum	1.00	1.00	1.00
ssO_Key	0.98	-	0.82	dsCH_Sum	0.95	0.92	0.95
aaO_Key	0.99	-	-	aaCH_Sum	0.74	0.97	-
aOm_Key	0.90	-	0.81	sssCH_Sum	0.78	0.72	0.69
sOm_Key	1.00	0.89	0.89	tsC_Sum	0.92	0.88	0.88
sF_Key	0.63	-	0.81	dssC_Sum	0.61	0.85	-
ssS_Key	0.99	-	0.98	aasC_Sum	-	0.74	0.74
aaS_Key	0.99	-	-	aaaC_Sum	1.00	0.99	0.99
dssS_Key	0.60	-	-	ssssC_Sum	0.59	-	0.93
ddssS_Key	0.66	-	0.99	sNH3_Sum	1.00	1.00	1.00
sCl_Key	0.73	-	0.96	sNH2_Sum	0.65	0.70	0.70
sBr_Key	0.68	-	-	ssNH2_Sum	0.78	0.99	0.64
sI_Key	0.87	-	0.99	dNH_Sum	1.00	1.00	1.00
sCH3_Cnt	0.72	-	0.99	ssNH_Sum	0.70	0.72	0.72
dCH2_Cnt	1.00	1.00	1.00	aaNH_Sum	0.86	0.87	-
ssCH2_Cnt	0.78	0.98	0.70	tN_Sum	0.99	0.69	0.69
sssNH_Sum	1.0	0.99	0.99	ZM1	0.69	0.97	0.92
dsN_Sum	0.59	0.95	0.95	ZM1V	0.62	0.86	0.69
aaN_Sum	0.86	0.87	-	ZM2	0.67	0.94	0.89
sssN_Sum	0.93	0.89	0.97	ZM2V	0.59	0.79	0.71
aasN_Sum	1.00	0.99	0.99	Pol	0.67	0.92	0.84
sOH_Sum	0.67	0.79	0.75	NST	0.73	0.95	0.89
dO_Sum	0.78	0.79	0.74	NHT	0.58	0.76	0.72
ssO_Sum	0.82	0.80	0.99	NGT	0.61	0.72	0.72
aaO_Sum	1.00	0.99	0.99	TSC	0.68	0.89	0.84
aOm_Sum	0.78	0.87	0.92	W	0.75	0.89	0.85
sOm_Sum	0.88	0.81	1.00	MW	0.71	0.95	0.89
sF_Sum	0.67	0.81	-	Xu	0.73	0.99	0.92
ssS_Sum	1.00	0.99	0.99	QIndex	0.69	0.74	0.74
aaS_Sum	1.00	0.99	0.99	RC	0.61	0.83	0.76
dssS_Sum	1.00	0.60	0.60	MSDB	0.53	-	-
ddssS_Sum	0.68	0.99	1.00	SP	0.26	-	0.71
sCl_Sum	0.81	0.96	-	Har	0.74	0.98	0.91
sBr_Sum	0.65	0.65	1.00	LPRS	0.75	0.99	0.92
sI_Sum	0.92	0.99	-	Pog	0.74	0.98	0.90
sCH3_Avg	-	0.92	0.92	SMT	0.74	0.89	0.84
dCH2_Avg	-	1.00	1.00	SMTV	0.72	0.88	0.85
ssCH2_Avg	0.59	0.75	-	MDDD	0.75	0.96	0.87
tCH_Avg	-	1.00	1.00	Ram	0.63	0.90	0.86
dsCH_Avg	-	0.99	-	GMT	0.73	0.89	0.84
aaCH_Avg	0.85	0.85	-	GMTV	0.69	0.85	0.85
sssCH_Avg	0.58	-	0.62	AVDD	0.77	0.97	0.89
tsC_Avg	0.79	0.98	0.98	UP	0.76	0.96	0.90
dssC_Avg	-	0.69	0.69	CENT	0.75	0.90	0.81
aasC_Avg	-	-	0.85	VAR	0.75	0.94	0.83
aaaC_Avg	-	0.99	0.99	MEV	0.68	0.86	0.73
ssssC_Avg	0.62	0.93	-	MEPV	0.50	0.68	-
sNH3_Avg	-	1.00	1.00	MENV	0.74	0.74	0.75

sNH2_Avg	0.65	0.70	0.70	ECCc	0.75	0.95	0.88
ssNH2_Avg	0.77	1.00	0.64	ECC	0.76	0.95	0.87
dNH_Avg	-	1.00	1.00	AECC	0.68	0.90	0.83
ssNH_Avg	0.59	0.99	0.99	DECC	0.60	0.79	0.73
aaNH_Avg	-	1.00	1.00	vX0	0.75	0.97	0.96
tN_Avg	-	0.69	0.69	vX1	0.74	0.96	0.97
sssNH_Avg	-	0.99	0.99	vX2	0.80	0.92	-
dsN_Avg	0.60	0.99	0.99	vX3	0.74	0.89	0.95
aaN_Avg	-	0.99	0.86	vX4	0.74	0.80	0.92
sssN_Avg	-	0.97	0.93	vX5	0.73	0.73	0.88
aasN_Avg	-	0.99	0.99	AvX0	0.79	0.69	0.65
sOH_Avg	0.68	0.83	0.80	AvX1	0.79	0.80	0.71
dO_Avg	0.53	0.98	0.72	AvX2	0.80	0.73	0.68
ssO_Avg	-	0.98	0.81	AvX3	0.73	0.78	0.69
aaO_Avg	-	0.99	0.99	AvX4	0.72	0.78	0.67
aOm_Avg	0.90	1.00	0.81	AvX5	0.69	0.74	0.68
sOm_Avg	-	0.90	0.89	QW	0.75	0.87	0.87
sF_Avg	0.65	1.00	0.80	FM	0.65	0.82	0.77
ssS_Avg	-	0.99	0.98	SM	0.62	0.79	0.77
aaS_Avg	-	0.99	0.99	STN	0.70	-	-
dsS_Avg	-	0.60	0.60	KBLI	0.79	0.80	0.71
ddsS_Avg	0.68	0.99	-	TCI1	0.60	0.83	0.83
sCL_Avg	0.73	1.00	0.96	TCI2	0.56	0.73	0.64
sBr_Avg	0.65	1.00	1.00	TCI3	0.46	0.68	0.68
sI_Avg	0.92	0.99	1.00	Jhetp	0.55	-	-
TCI5	0.61	0.85	0.79	TD	0.66	0.88	0.79
TCI6	0.59	0.87	0.81	TR	0.66	0.87	0.79
TCI7	0.64	0.86	0.82	PJ2DS	-	-	-
TCI8	0.69	0.94	0.89	J	0.61	0.86	0.86
TCI9	0.70	0.93	0.88	SCIX0	0.75	0.99	0.92
TCI10	0.71	0.93	0.88	SCIX1	0.73	0.99	0.92
MTCI1	0.71	-	0.94	SCIX2	0.68	0.98	0.93
MTCI2	0.52	-	-	SCIX3	0.71	0.95	0.86
MTCI3	0.53	-	0.63	SCIX4	0.70	0.91	0.85
MTCI4	0.54	-	0.63	SCIX5	0.61	0.84	0.79
MTCI5	-	-	-	CIX0	0.75	0.99	0.92
MTCI6	-	-	-	CIX1	0.74	0.98	0.91
MTCI7	0.63	0.60	0.69	CIX2	0.70	0.98	0.94
MTCI8	0.53	-	0.63	CIX3	0.70	0.93	0.85
MTCI9	-	-	-	CIX4	0.69	0.89	0.83
MTCI10	-	-	-	CIX5	0.62	0.86	0.82
GTC	0.81	0.94	-	ACIX0	0.60	0.73	0.70
HDPI	0.67	0.76	0.93	ACIX1	0.72	-	-
RHDPI	0.71	0.97	0.91	ACIX2	0.55	0.72	0.72
SRDS	0.74	0.98	0.91	ACIX3	0.54	0.83	0.83
MRC	0.72	0.99	0.91	ACIX4	0.53	0.85	0.85
BC	0.33	-	-	ACIX5	0.51	0.79	0.79
LC	0.33	0.99	0.99	RDR	0.71	0.97	0.91
KHE	0.75	0.97	0.87	RDSR	0.72	0.97	0.90
STD(N N)	0.51	0.67	0.80	KAMS1	0.77	0.99	0.91
STD(N O)	0.61	0.71	-	KAMS2	0.78	0.95	0.87
STD(N S)	0.82	0.63	0.65	KAMS3	0.73	0.90	0.88
STD(N F)	0.57	0.72	0.88	KF	0.79	0.93	0.85
STD(N Cl)	0.78	0.88	0.94	RSIpw2	-	0.72	0.81
STD(N Br)	0.46	0.80	0.80	RSIpw3	0.49	0.66	0.66
STD(N I)	0.80	0.95	0.93	RSIpw4	0.59	0.60	0.60
STD(O O)	0.64	0.63	0.95	RSIpw5	0.58	0.68	0.68
STD(O S)	-	0.66	0.68	ETP	0.71	0.94	0.81
STD(O F)	0.59	0.74	0.78	RNGCNT3	-	-	-

STD(O Cl)	-	<b>0.73</b>	<b>0.81</b>	RNGCNT4	-	-	.
STD(O Br)	<b>0.81</b>	<b>0.91</b>	<b>0.91</b>	RNGCNT5	-	-	-
STD(O I)	<b>0.96</b>	<b>0.86</b>	<b>0.89</b>	RNGCNT6	<b>0.62</b>	<b>0.64</b>	<b>0.64</b>
STD(S S)	-	-	-	RNGCNT7	-	-	-
STD(S F)	<b>0.65</b>	-	-	RNGCNT8	-	-	-
STD(S Cl)	-	-	-	RNGCNT13	-	-	-
STD(S Br)	-	<b>0.68</b>	<b>0.68</b>	RNGCNT14	-	<b>0.70</b>	<b>0.70</b>
STD(F F)	-	-	<b>0.67</b>	RNGCNT15	-	-	-
STD(F Cl)	-	-	-	RNGCNT16	-	-	-
STD(F Br)	-	-	-	ATMCNT	<b>0.74</b>	<b>0.99</b>	<b>0.92</b>
STD(F I)	-	<b>0.87</b>	<b>0.92</b>	BNDCNT	<b>0.71</b>	<b>0.98</b>	<b>0.91</b>
STD(Cl Cl)	-	-	-	ATMRNGCN	<b>0.89</b>	<b>0.90</b>	<b>0.90</b>
STD(Cl Br)	-	-	-	BNDRNGCN	<b>0.97</b>	<b>0.94</b>	<b>0.94</b>
STD(Cl I)	-	-	-	CYCLONUM	<b>0.90</b>	<b>1.00</b>	<b>1.00</b>
WhetZ	<b>0.75</b>	<b>0.88</b>	<b>0.86</b>	NRS	<b>0.85</b>	-	<b>0.85</b>
Whete	<b>0.75</b>	<b>0.88</b>	<b>0.86</b>	NNRS	<b>0.97</b>	<b>0.83</b>	<b>0.97</b>
Whetm	<b>0.75</b>	<b>0.87</b>	<b>0.86</b>	RFD	-	-	-
Whetv	<b>0.75</b>	<b>0.88</b>	<b>0.86</b>	RNGPERM	<b>0.99</b>	<b>0.92</b>	<b>0.92</b>
Whetp	<b>0.75</b>	<b>0.88</b>	<b>0.86</b>	RNGBDGE	-	<b>0.90</b>	<b>0.90</b>
JhetZ	<b>0.60</b>	<b>0.92</b>	<b>0.92</b>	MCD	<b>0.63</b>	<b>0.69</b>	<b>0.69</b>
Jhete	<b>0.60</b>	<b>0.92</b>	<b>0.92</b>	RFDELTA	<b>0.62</b>	<b>0.70</b>	<b>0.82</b>
Jhetm	<b>0.60</b>	<b>0.92</b>	<b>0.92</b>	RCI	-	<b>0.85</b>	-
Jhetv	<b>0.57</b>	<b>0.99</b>	<b>0.99</b>				
TCI4	<b>0.60</b>	<b>0.83</b>	<b>0.78</b>	PEOE1	-	<b>0.62</b>	<b>0.62</b>
VSA	<b>0.78</b>	<b>0.96</b>	<b>0.95</b>	PEOE2	-	<b>0.80</b>	<b>0.60</b>
MR1	-	-	-	PEOE3	-	<b>0.63</b>	<b>0.69</b>
MR8	<b>0.79</b>	<b>0.95</b>	<b>0.95</b>	PEOE4	-	<b>0.65</b>	<b>0.65</b>
ALOGP1	<b>0.88</b>	<b>0.79</b>	<b>0.83</b>	PEOE5	-	-	-
ALOGP2	<b>0.80</b>	<b>0.84</b>	<b>0.87</b>	PEOE6	-	-	-
ALOGP3	<b>0.72</b>	<b>0.76</b>	<b>0.74</b>	PEOE7	-	-	-
ALOGP4	-	-	-	PEOE8	-	<b>0.79</b>	<b>0.80</b>
ALOGP5	<b>0.75</b>	<b>0.66</b>	<b>0.66</b>	PEOE9	-	<b>0.71</b>	<b>0.73</b>
ALOGP6	-	-	-	PEOE10	-	-	-
ALOGP7	-	-	-	PEOE11	-	-	-
ALOGP8	-	<b>0.67</b>	<b>0.67</b>	PEOE12	-	<b>0.74</b>	<b>0.68</b>
ALOGP9	<b>0.69</b>	<b>0.73</b>	-	PEOE13	-	-	-
ALOGP10	<b>0.88</b>	<b>0.83</b>	<b>0.92</b>	PEOE14	-	<b>0.78</b>	<b>0.61</b>

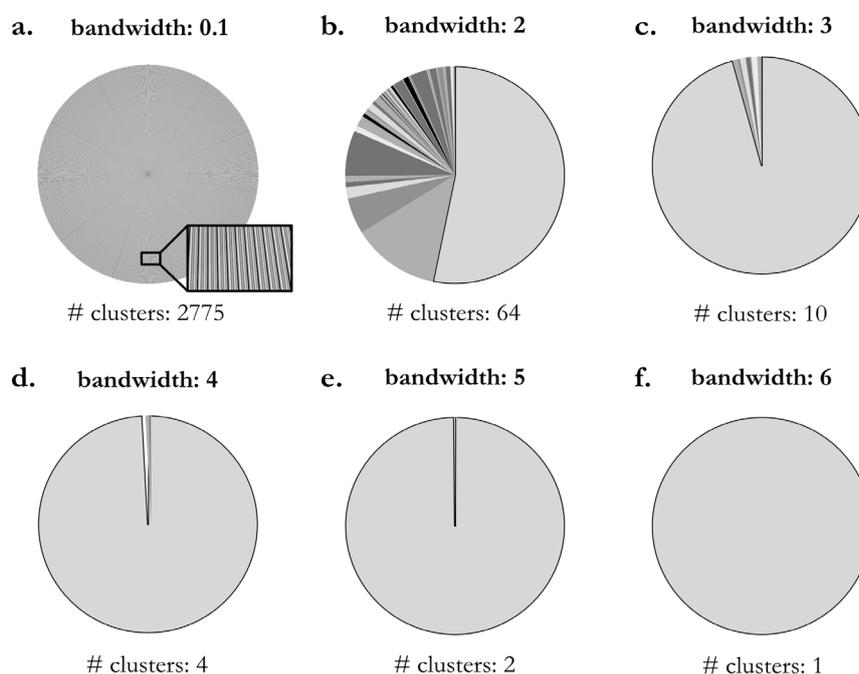
## Appendix A8: Supporting material for unsupervised ML\*

### A8.1 Other clustering algorithms

All clustering algorithms were implemented using python (Python Software Foundation; Python Language Reference, version 3.9.16; <http://www.python.org>)<sup>102</sup> with several extensions for data handling and visualization (Numpy<sup>103</sup> v1.22.4, Pandas<sup>104,105</sup> v1.5.3, Matplotlib<sup>106</sup> v3.7.1) as well as the Machine Learning toolbox scikit-learn<sup>109</sup> (v1.2.2).

#### Mean Shift clustering

The Mean Shift clustering algorithm iteratively shifts cluster centroids towards the densest areas of the dataset and merges clusters that are too similar. One advantage of the Mean Shift algorithm is that the number of clusters to be formed is automatically determined. Additionally, the clustering result can be optimized by adjusting the bandwidth hyperparameter, which describes how far the cluster centroids are shifted based on individual samples. A larger bandwidth results in fewer and larger clusters, whereas a smaller bandwidth results in more and smaller clusters. Mean shift clustering was conducted using different bandwidth values (see **Figure A8.1**). No seeds were used to initialize the kernels, bin seeding was disabled, the minimal bin frequency was set to 1, all points were considered for clustering (*i.e.*, isolated points were assigned to its nearest kernel), and the maximum iterations per seed point was set to 300.



**Figure A8.1: Cluster results of a Mean Shift clustering algorithm.** The cornea dataset containing 2775 samples characterized by 12 surface parameters was analyzed using a Mean Shift clustering algorithm. The bandwidth hyperparameter was varied (a-f) while all other cluster parameters were kept constant. The pie charts depict the number of clusters formed and the cluster cardinalities (*i.e.*, how many samples were assigned to each cluster).

\* This section follows in part the publication Rickert *et al.*, ‘Unsupervised Machine Learning to Topographically Analyze Corneal Tissue Surfaces’, submitted 2023

### Agglomerative Clustering

Agglomerative Clustering is a hierarchical clustering technique that first initializes each sample point as an own cluster and then iteratively merges the closest pairs of clusters until the desired number of clusters is reached (based on Euclidean distances).

Agglomerative Clustering was conducted to form a predefined number of 4 clusters ( $n\_cluster = 4$ ). Clusters were merged based on the 'ward' linkage criterion: pairs of clusters were merged such that that the variance of the new clusters is minimized (using Euclidean distances).

### Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)

The BIRCH clustering algorithm is a hierarchical approach that generates a tree data structure that stores the data points and their corresponding sub-clusters. In this tree, each node represents a sub-cluster and the nodes at the lowest level of the tree contain the actual data points.

The radius of the sub-cluster obtained during merging a new sample and the closest sub-cluster was set to be smaller than 0.5. If this criterion was not met, a new sub-cluster was started. The maximum number of sub-clusters combined in one node was set to 50. In case a new sample would enter this node such that the number of subclusters exceeds this threshold, that node was split into two nodes with the sub-clusters redistributed into each.

### Spectral Clustering

Spectral Clustering is a data clustering technique that uses information from the eigenvalues and eigenvectors of a similarity matrix to group data points into clusters. In simple words, the algorithm considers how data points are related to each other (in a connected graph) and uses that information to find groups that are similar to each other. It is especially useful for identifying clusters of complex shape.

Spectral Clustering was conducted to form 4 clusters ( $n\_cluster = 4$ ). The 'arpack' eigen-solver was used and the number of eigenvectors to use for the spectral embedding was set to equal the number of clusters. Labels were assigned to the data points using the 'discretize' method, which assigns the label of the closest cluster center to each data point. No random seed was provided and the affinity matrix was constructed using a radial basis function ('rbf') kernel with a kernel coefficient of 1.0.

### Gaussian Mixture model

The Gaussian Mixture clustering algorithm is a probabilistic clustering technique that models the distribution of data points as a mixture of Gaussian distributions, and estimates the parameters of these distributions to cluster the data.

Here, the number of mixture components ( $n\_components$ ) was set to 4, to obtain 4 clusters. For each of those 4 components, an individual covariance matrix was used ( $covariance\_type = full$ ). Furthermore, a convergence threshold  $1 \times 10^{-3}$  was used and  $1 \times 10^{-6}$  was used as a covariance regularization to assure that all covariance matrices are fully positive. A maximum of 100 iterations were performed, and one initialization of the weights, the means, and the precisions was conducted using the 'kmeans' method.

## A8.2 Coefficient normalization

To guarantee sufficient comparability of metrics calculated for clustering results that were generated in feature spaces of different dimensionality (here, either 21 or 12 features as displayed in **Figure 4.17**), the following procedure was applied: first, clustering was conducted in the desired feature space (*i.e.*, using either 21 or 12 features). Then, for each clustering result, all metrics were calculated in both feature spaces and the results were averaged. For clustering results generated in feature spaces with the same dimensionality (**Figure 4.20c**) no such normalizations had to be applied.

## A8.3 Adjusted mutual information index

To compare two clustering results, the adjusted mutual information score AMI was calculated as described by Vinh *et al.*<sup>136</sup> Therefore, for two clustering results (*i.e.*, assignments of labels)  $U$  and  $V$  of the same sample set of size  $N$ , the following metrics are calculated: First, the entropy of the label assignments is calculated as the amount of uncertainty for a partition set, as given by

$$H(U) = - \sum_{i=1}^{|U|} P(i) \log(P(i)) \quad (\text{A8.1})$$

and

$$H(V) = - \sum_{j=1}^{|V|} P'(j) \log(P'(j)). \quad (\text{A8.2})$$

Here,  $P(i) = |U_i|/N$  and  $P'(j) = |V_j|/N$  quantify the probability that a sample that was randomly picked falls into clusters  $U_i$  and  $V_j$ , respectively. The mutual information of  $U$  and  $V$  is then given by

$$MI(U, V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} P(i, j) \log \left( \frac{P(i, j)}{P(i)P'(j)} \right), \quad (\text{A8.3})$$

where  $P(i, j) = |U_i \cap V_j|/N$  denotes the probability that a randomly picked sample is assigned into both clusters  $U_i$  and  $V_j$ . Normalization of the MI is then achieved following **equation A8.4**.

$$NMI(U, V) = \frac{MI(U, V)}{\text{mean}(H(U), H(V))} \quad (\text{A8.4})$$

Next, the expected value of the MI is calculated as

$$E[MI(U, V)] = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \sum_{n_{ij}=(a_i+b_j-N)^+}^{\min(a_i, b_i)} \frac{n_{ij}}{N} \log \left( \frac{N * n_{ij}}{a_i b_i} \right) \frac{a_i! b_j! (N - a_i)! (N - b_j)!}{N! n_{ij}! (a_i - n_{ij})! (b_j - n_{ij})! (N - a_i - b_j + n_{ij})!}, \quad (\text{A8.5})$$

where  $a_i$  and  $b_i$  denote the number of elements in  $U_i$  and  $V_j$ , respectively. Finally, the adjusted mutual information score (AMI) is calculated as

$$AMI = \frac{MI - E[MI]}{\text{mean}(H(U), H(V)) E[MI]} \quad (\text{A8.6})$$

The mean values presented in **chapter 4.3** were obtained as follows: For comparing clustering results obtained by the same approach (*e.g.*, comparing multiple runs of the  $k$ -Means algorithm or comparing manual clustering results from 10 different individuals) the pairwise AMI of each sample to all other samples within the group was computed. Those pairwise AMIs were then averaged. To compare the two approaches to each other, the pairwise AMI of all results of the first algorithm to all results of the second approach was calculated. Once again, those values were then averaged.

## A8.4 Detailed description of the features of the dataset

**Table A8.1: Statistical descriptions of the dataset.** Different statistical descriptors are provided of each surface parameter included in the initial dataset (a), and the dataset after applying a Minmax-scaler (b), a Minmax-scaler after clearing the dataset using the inter quantile range (c), or a Power Transformation (d).

a. initial dataset								
parameter	count	mean	std	min	25 %	50 %	75 %	max
Sa	2775	0.25	0.36	0.02	0.13	0.17	0.27	7.79
Sz	2775	10.30	24.07	0.16	3.94	5.51	8.87	611.87
Str	2775	0.62	0.18	0.02	0.52	0.63	0.74	0.98
Sdr	2775	0.05	0.30	0.00	0.00	0.01	0.03	8.33
Sq	2775	0.40	0.64	0.02	0.18	0.24	0.38	11.13
Ssk	2775	1.00	2.36	-26.94	0.06	0.47	1.45	21.98
Sku	2775	26.26	65.50	3.11	5.09	8.89	22.75	1917.40
Sp	2775	5.52	14.17	0.09	2.43	3.45	5.22	390.68
Sv	2775	4.77	11.00	0.06	1.31	1.89	3.31	221.19
Sal	2775	9.92	4.09	2.09	7.29	9.67	12.03	38.83
Sdq	2775	0.21	0.30	0.01	0.10	0.15	0.24	6.09
Sk	2775	0.63	0.77	0.05	0.36	0.49	0.72	20.07
Spk	2775	0.58	0.98	0.02	0.23	0.34	0.55	15.90
Svk	2775	0.53	1.15	0.02	0.19	0.26	0.42	15.41
Smr1	2775	12.43	2.24	8.34	10.97	11.90	13.24	36.31
Smr2	2775	88.71	1.74	77.62	88.10	88.97	89.68	95.88
Sxp	2775	0.79	1.25	0.04	0.36	0.49	0.78	22.77
Vvv	2775	0.05	0.10	0.00	0.02	0.03	0.04	1.51
Vvc	2775	0.37	0.50	0.02	0.19	0.26	0.39	11.49
Vmp	2775	0.03	0.06	0.00	0.01	0.02	0.03	0.92
Vmc	2775	0.24	0.31	0.02	0.13	0.18	0.27	7.82

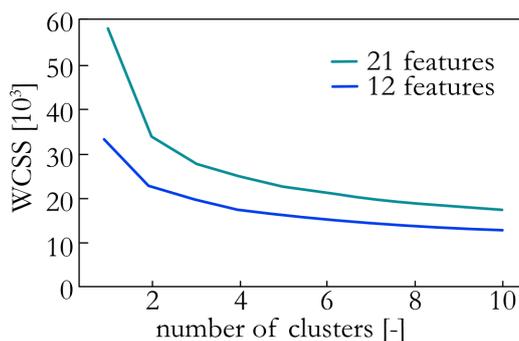
b. scaled with a Minmax-scaler								
parameter	count	mean	std	min	25 %	50 %	75 %	max
Sa	2775	0.03	0.05	0	0.01	0.02	0.03	1
Sz	2775	0.02	0.04	0	0.01	0.01	0.01	1
Str	2775	0.62	0.19	0	0.52	0.64	0.75	1
Sdr	2775	0.01	0.04	0	0.00	0.00	0.00	1
Sq	2775	0.03	0.06	0	0.01	0.02	0.03	1
Ssk	2775	0.57	0.05	0	0.55	0.56	0.58	1
Sku	2775	0.01	0.03	0	0.00	0.00	0.01	1
Sp	2775	0.01	0.04	0	0.01	0.01	0.01	1
Sv	2775	0.02	0.05	0	0.01	0.01	0.01	1
Sal	2775	0.21	0.11	0	0.14	0.21	0.27	1
Sdq	2775	0.03	0.05	0	0.01	0.02	0.04	1
Sk	2775	0.03	0.04	0	0.02	0.02	0.03	1
Spk	2775	0.04	0.06	0	0.01	0.02	0.03	1
Svk	2775	0.03	0.07	0	0.01	0.02	0.03	1
Smr1	2775	0.15	0.08	0	0.09	0.13	0.18	1
Smr2	2775	0.61	0.10	0	0.57	0.62	0.66	1
Sxp	2775	0.03	0.05	0	0.01	0.02	0.03	1
Vvv	2775	0.03	0.07	0	0.01	0.02	0.03	1
Vvc	2775	0.03	0.04	0	0.01	0.02	0.03	1
Vmp	2775	0.03	0.06	0	0.01	0.02	0.03	1
Vmc	2775	0.03	0.04	0	0.02	0.02	0.03	1

c. scaled with a Minmax-scaler on a cleared dataset								
parameter	count	mean	std	min	25 %	50 %	75 %	max
Sa	1857	0.36	0.17	0	0.24	0.32	0.44	1
Sz	1857	0.27	0.15	0	0.17	0.24	0.34	1
Str	1857	0.56	0.18	0	0.44	0.56	0.69	1
Sdr	1857	0.19	0.20	0	0.05	0.12	0.25	1
Sq	1857	0.37	0.17	0	0.25	0.33	0.45	1
Ssk	1857	0.43	0.17	0	0.30	0.37	0.50	1
Sku	1857	0.16	0.20	0	0.03	0.08	0.20	1
Sp	1857	0.29	0.16	0	0.18	0.27	0.37	1
Sv	1857	0.23	0.16	0	0.13	0.20	0.29	1
Sal	1857	0.42	0.19	0	0.28	0.41	0.54	1
Sdq	1857	0.34	0.21	0	0.17	0.29	0.45	1
Sk	1857	0.35	0.17	0	0.23	0.31	0.43	1
Spk	1857	0.28	0.15	0	0.18	0.25	0.35	1
Svk	1857	0.29	0.14	0	0.19	0.26	0.36	1
Smr1	1857	0.39	0.17	0	0.27	0.36	0.48	1
Smr2	1857	0.53	0.16	0	0.43	0.53	0.63	1
Sxp	1857	0.31	0.16	0	0.20	0.28	0.39	1
Vvv	1857	0.33	0.16	0	0.22	0.29	0.42	1
Vvc	1857	0.34	0.17	0	0.22	0.30	0.43	1
Vmp	1857	0.27	0.15	0	0.18	0.24	0.33	1
Vmc	1857	0.35	0.17	0	0.23	0.32	0.44	1

d. transformed with a Yeo-Johnson PowerTransformer								
parameter	count	mean	std	min	25 %	50 %	75 %	max
Sa	2775	0.00	1.00	-2.98	-0.71	-0.15	0.67	2.34
Sz	2775	0.00	1.00	-5.87	-0.64	-0.08	0.60	2.71
Str	2775	0.00	1.00	-2.41	-0.65	-0.02	0.69	2.58
Sdr	2775	0.00	1.00	-1.13	-0.78	-0.37	0.52	2.52
Sq	2775	0.00	1.00	-2.99	-0.70	-0.17	0.65	2.26
Ssk	2775	0.00	1.00	-13.94	-0.38	-0.20	0.21	8.06
Sku	2775	0.00	1.00	-1.69	-0.88	-0.10	0.86	2.19
Sp	2775	0.00	1.00	-4.88	-0.63	-0.03	0.61	3.12
Sv	2775	0.00	1.00	-3.86	-0.68	-0.12	0.62	2.17
Sal	2775	0.00	1.00	-2.62	-0.59	0.06	0.61	4.52
Sdq	2775	0.00	1.00	-2.51	-0.76	-0.10	0.72	2.34
Sk	2775	0.00	1.00	-3.38	-0.69	-0.09	0.69	2.80
Spk	2775	0.00	1.00	-2.82	-0.75	-0.14	0.66	2.21
Svk	2775	0.00	1.00	-2.71	-0.71	-0.17	0.61	2.22
Smr1	2775	0.00	1.00	-4.09	-0.70	-0.03	0.67	2.93
Smr2	2775	0.00	1.00	-3.79	-0.48	0.06	0.53	6.99
Sxp	2775	0.00	1.00	-3.12	-0.70	-0.13	0.66	2.32
Vvv	2775	0.00	1.00	-2.28	-0.70	-0.25	0.55	2.43
Vvc	2775	0.00	1.00	-3.07	-0.69	-0.14	0.67	2.38
Vmp	2775	0.00	1.00	-2.11	-0.73	-0.26	0.59	2.31
Vmc	2775	0.00	1.00	-3.07	-0.70	-0.12	0.68	2.46

### A8.5 Elbow method

To identify the optimal number of clusters to be formed with a  $k$ -Means clustering algorithm, the ‘elbow method’ was employed.<sup>271</sup> Here,  $k$  is varied from 1 to 10, the  $k$ -Means algorithm is conducted, and the within-cluster sum of squares (WCSS) is calculated using the  $k$ -Means inertia method of scikit-learn<sup>109</sup> (Figure A8.2). With increasing  $k$ -value, the number of samples contained in each cluster decreases, and the samples are located closer to their respective cluster centroid; consequently, the WCSS decreases. By assuming that the obtained WCSS plot resembles an arm, the ideal number of clusters is then given as the point of highest inflection (*i.e.*, the location of the ‘elbow’).



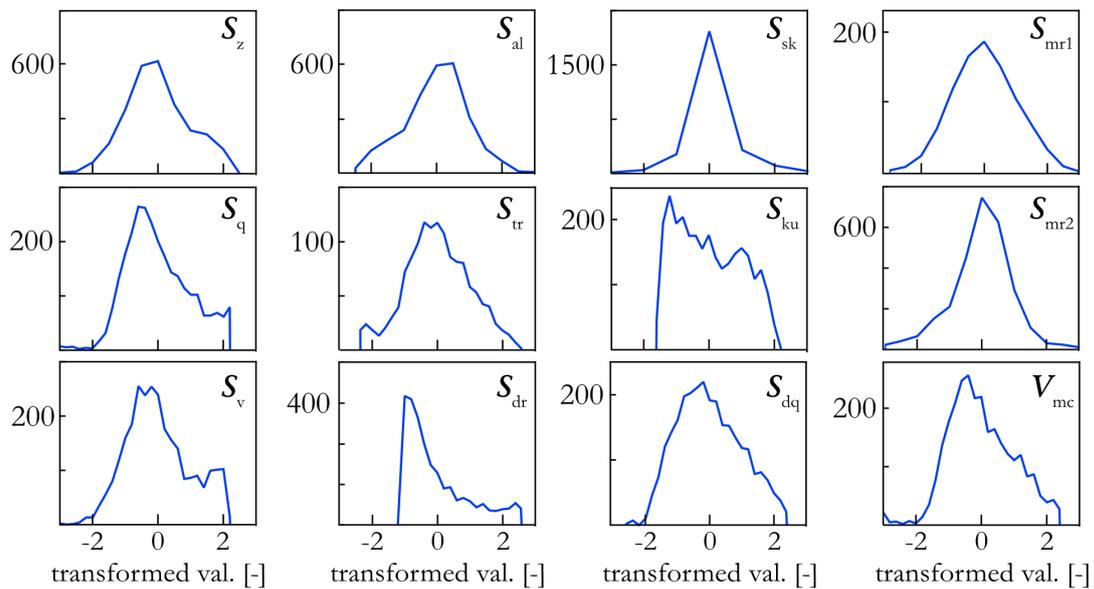
**Figure A8.2. Elbow curves of a  $k$ -Means clustering algorithm.** The within-cluster sum of squares (WCSS) achieved with a  $k$ -Means clustering algorithm applied to the cornea dataset is displayed for varying values of  $k$ . In the dataset, each sample ( $n = 2775$ ) is characterized by a feature vector containing either 21 (cyan) or 12 (blue) features.

### A8.6 Correlation matrix of the corneal dataset

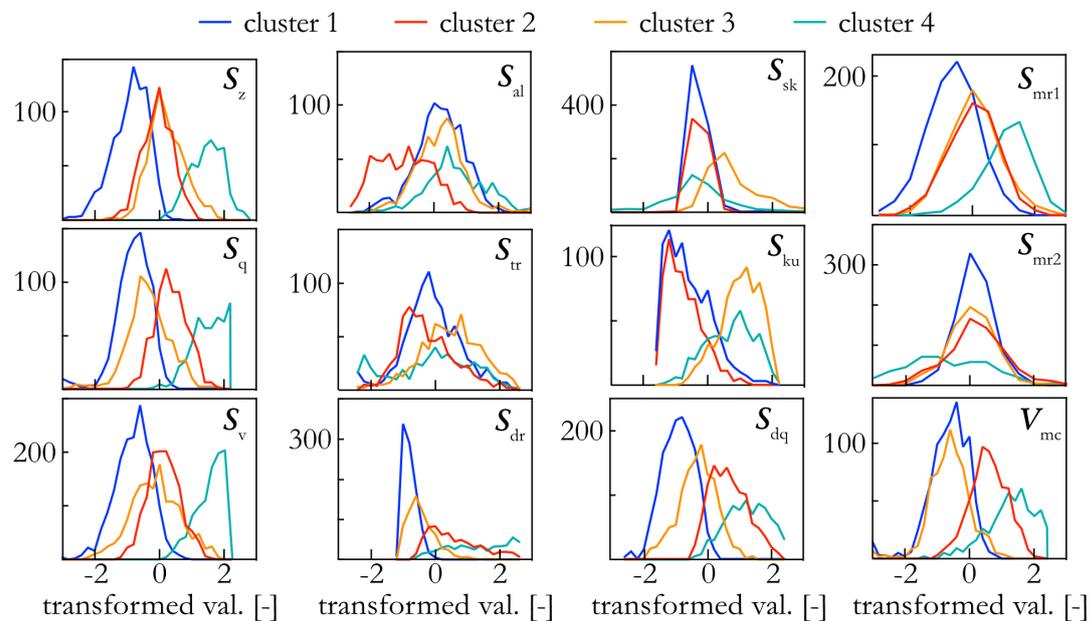
	$S_a$	$S_z$	$S_q$	$S_p$	$S_v$	$S_{al}$	$S_k$	$S_{pk}$	$S_{vk}$	$S_{xp}$	$S_{tr}$	$S_{dr}$	$S_{sk}$	$S_{ku}$	$S_{dq}$	$S_{mr1}$	$S_{mr2}$	$V_{vv}$	$V_{vc}$	$V_{mp}$	$V_{mc}$
$S_a$	1.0	0.9	0.9	0.8	0.8	0.1	0.9	0.9	0.9	0.9	0.0	0.5	-0.1	0.0	0.7	0.4	-0.3	0.9	0.9	0.9	0.9
$S_z$		1.0	0.9	0.9	0.9	0.2	0.8	0.9	0.8	0.8	0.0	0.4	0.0	0.1	0.6	0.3	-0.3	0.8	0.8	0.8	0.8
$S_q$			1.0	0.8	0.9	0.1	0.8	0.9	0.9	0.9	0.0	0.6	-0.1	0.0	0.8	0.4	-0.4	0.9	0.9	0.9	0.9
$S_p$				1.0	0.8	0.1	0.8	0.9	0.7	0.8	0.0	0.4	0.0	0.0	0.5	0.2	-0.2	0.7	0.8	0.8	0.8
$S_v$					1.0	0.2	0.7	0.8	0.8	0.8	0.0	0.4	-0.2	0.1	0.6	0.4	-0.3	0.8	0.8	0.8	0.8
$S_{al}$						1.0	0.1	0.2	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.2	0.1
$S_k$							1.0	0.8	0.7	0.8	0.0	0.4	-0.1	0.0	0.6	0.2	-0.2	0.7	0.9	0.7	0.9
$S_{pk}$								1.0	0.9	0.9	0.0	0.6	0.0	0.0	0.8	0.4	-0.4	0.9	0.9	0.9	0.8
$S_{vk}$									1.0	0.9	0.0	0.6	-0.1	0.0	0.7	0.5	-0.4	0.9	0.8	0.9	0.8
$S_{xp}$										1.0	0.0	0.6	-0.1	0.0	0.8	0.4	-0.3	0.9	0.9	0.9	0.9
$S_{tr}$											1.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
$S_{dr}$												1.0	0.0	0.0	0.8	0.2	-0.2	0.6	0.5	0.6	0.4
$S_{sk}$													1.0	0.2	-0.1	-0.1	0.0	-0.1	-0.1	0.0	-0.1
$S_{ku}$														1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$S_{dq}$															1.0	0.4	-0.3	0.8	0.7	0.8	0.6
$S_{mr1}$																1.0	-0.1	0.4	0.4	0.4	0.2
$S_{mr2}$																	1.0	-0.4	-0.3	-0.4	-0.2
$V_{vv}$																		1.0	0.9	0.9	0.8
$V_{vc}$																			1.0	0.8	0.9
$V_{mp}$																				1.0	0.8
$V_{mc}$																					1.0

**Figure A8.3: Correlation matrix of the full feature set.** The Pearson’s correlation coefficient between all 21 features of the initially obtained feature space is displayed. A value of 1 corresponds to full direct correlation, whereas a value of -1 denotes full antiproportional correlation.

## A8.7 Data distribution of the 12-dimensional feature space

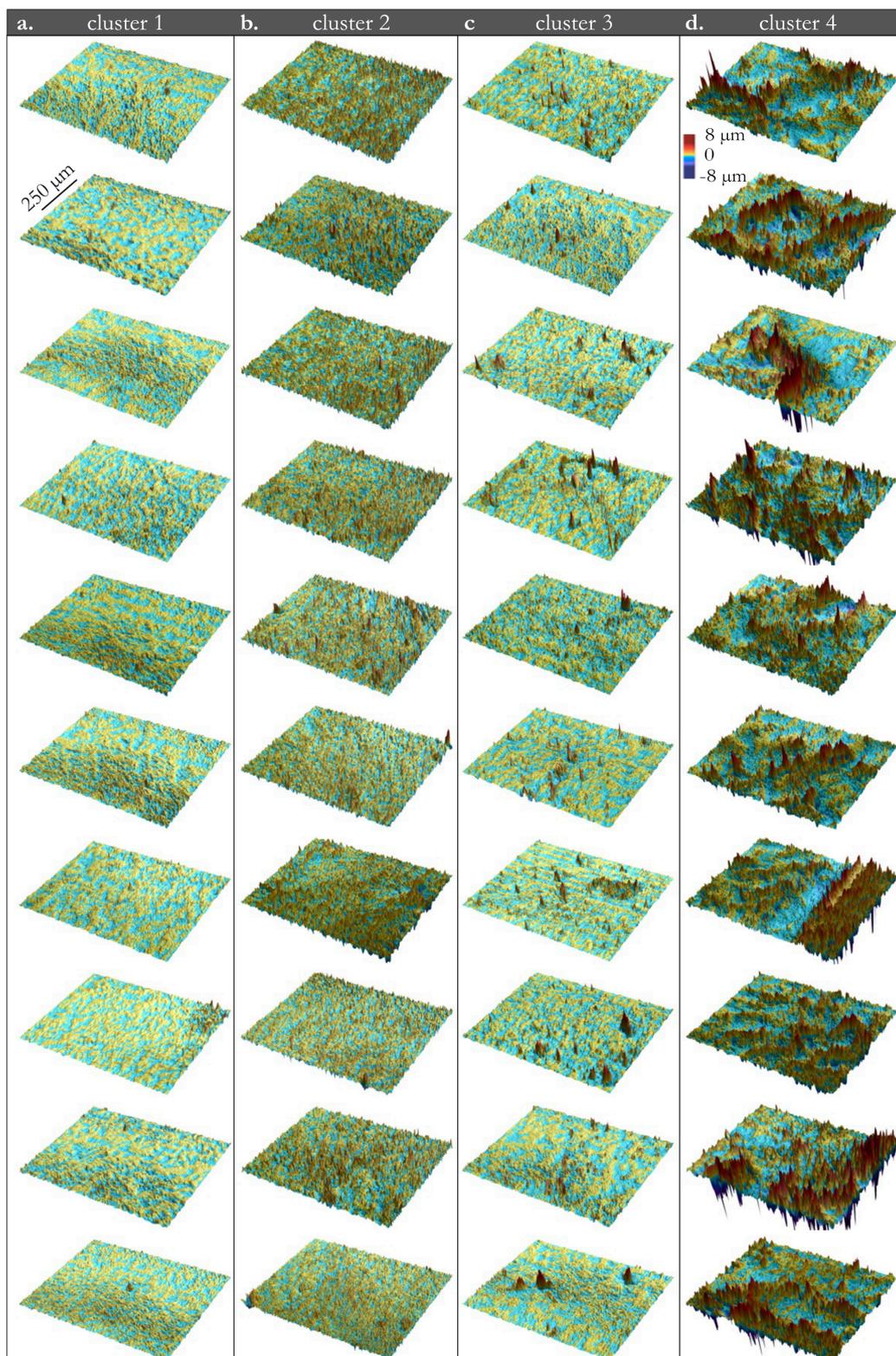


**Figure A8.4: Histograms of the 12 features of the feature vector used for clustering the samples.** The frequency distribution after applying the Yeo-Johnson Power Transformation is depicted for the features of the reduced feature vector comprising 12 surface parameters. All  $y$ -axes represent the frequency (starting from 0). The displayed  $x$ -axes apply to all graphs of the respective column.

A8.8 Data distribution of 4 clusters formed by a  $k$ -Means algorithm

**Figure A8.5: Frequency distributions of 4 clusters.** The frequency distribution of the features (per cluster) as obtained from a  $k$ -Means clustering algorithm employed to form 4 clusters in the 12-dimensional feature space is depicted. All  $y$ -axes represent the frequency (starting from 0). The displayed  $x$ -axes apply to all graphs of the respective column.

## A8.9 Central images of the clusters



**Figure A8.6: Clustering result achieved with a  $k$ -Means clustering algorithm.** A  $k$ -Means clustering algorithm was employed to group the corneal surfaces into 4 clusters (based on 12 surface properties). For each cluster, the ten 'most central' samples of each cluster, *i.e.*, those samples that have the smallest Euclidean distance to the respective cluster centroid, are displayed (a-d). The scale bars apply for all images.

## A8.10 Sample classification using the pre-trained $k$ -Means algorithm

To classify unknown samples with a  $k$ -Means clustering algorithm, the  $k$ -Means model was first used to form four clusters from the corneal samples as described in **chapter 2.10.3** (comprising 2775 surface samples characterized by 12 metrological parameters). The pre-trained model was then used to classify 12 additional corneal images derived from the tribological measurements conducted with RGP contact lenses (see **chapter 3.3**) to the four clusters created above (none of those 12 additional samples was used to form these clusters). Therefore, each of those samples was sorted into the cluster with the closest centroid to the sample in the 12-dimensional feature space.

## A8.11 Clustering based on features extracted by deep neural networks

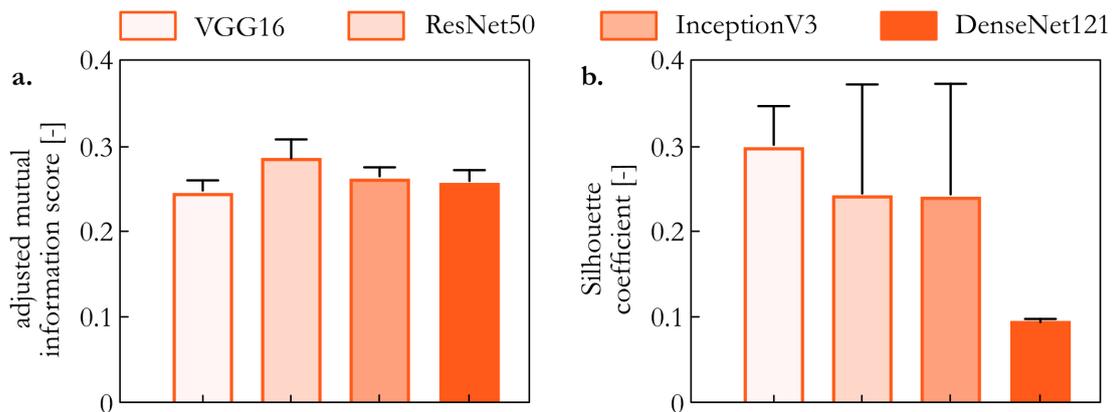
As presented in **chapter 4.3**, a  $k$ -Means clustering algorithm was successfully used to sort samples based on a set of metrological surface parameters. However, in addition to calculating this set of parameters, deep learning methods could be used to directly analyze the topographical surfaces. More specifically, Convolutional Neural Networks (CNNs) that were pre-trained for image classification (usually of everyday objects) can be used to extract features from the topographical profiles; and those extracted features can then be used as input for a  $k$ -Means clustering algorithm.

When given an image as an input, CNNs use trainable weights to assign importance gradings to different aspects of an image or to objects within the image. The networks can then be used to analyze or classify images, or to identify objects. For this purpose, CNNs mainly make use of three procedures: convolution, pooling, and flattening. For image convolution, filters are applied to each pixel. This can help the network to identify certain structures such as edges or peaks. Pooling can lower the computational cost by combining pixels from the same region into one, thus reducing the size of the image. After applying (multiple) convolution and pooling steps, the individual pixels of the resulting image matrix are fed into a standard neural network – a process, which is referred to as “flattening”. When taking a pre-trained CNN and removing the layers designated for classification while keeping only those that perform the feature extraction tasks, a newly created feature vector (directly derived from the surface appearance) can be obtained for each analyzed image.

To extract visual features from the topographical images (other than the calculated set of surface parameters), a pre-trained VGG16<sup>298</sup> model was employed, a Deep Convolutional Network for large-scale image recognition. The topographical profiles were imported from ISO-8859-1-encoded csv files, each containing a 768 x 1024 pixel matrix with the height value of every pixel. Missing height values were replaced by ‘0’ (the valid point density of the height profiles was usually >99.9 %). The individual height profiles were pre-processed inspired by Ghumman *et al.*:<sup>299</sup> To reduce the influence of measurement errors and statistical pixel outliers, the 1st and 99th percentile values of each individual height profiles were identified, and any pixel height outside this range was adjusted to the corresponding threshold. To then normalize each profile to a baseline of 0, the first percentile value was subtracted from each pixel. Last, the profile height values were adjusted to the typical RGB scale (ranging from 0 to 255) such that the maximum height of all profiles corresponds to 255.

Afterwards, the Python Imaging Library (version 8.4.0) was used to convert the (768 x 1024 x 1) arrays to grayscale images, to resize them to a shape of (224, 224, 1) and to triplicate the single layers to obtain a (224, 224, 3) image as required by the VGG16 model (this triplication simulates the RGB channels of a grayscale image). Subsequently, tensorflow keras<sup>300</sup> (version 2.12.0) was used to preprocess the images

according to the VGG16 requirements. Finally, all images were analyzed using the VGG16 network that was pre-trained on the ImageNet<sup>300</sup> dataset (a database containing more than 14 million images of more than 20,000 object categories). To obtain a feature vector instead of a classification result, the top three layers were dropped and all transformations were conducted with the first 16 layers. Layer 16 of VGG16 delivers a (7 x 7 x 512) feature matrix, that was flattened (*i.e.*, transformed into a one-dimensional array), pre-processed using a Yeo-Johnson Power Transformer, reduced to 100 features using a principal component analysis (from initially 25,088 features), and then used as an input feature vector for the  $k$ -Means clustering algorithm. Here, for better comparability, again 4 clusters were formed. To compare the results obtained with this strategy to that of the  $k$ -Means clustering based on surface parameters (**chapter 4.3**), the AMI (adjusted mutual information score) is calculated (**Figure A8.7a**). However, once again, the similarity of the two clustering results is rather low, as AMI values around 0.25 are obtained.



**Figure A8.7: Clustering based on features extracted using Convolutional Neural Networks.** Images of corneal surfaces were analyzed by four different CNNs to create a new feature vector. This new feature vector was then used to conduct  $k$ -Means clustering. The obtained clustering results were compared to those previously obtained with the  $k$ -Means clustering based on topographical surface parameters (a). Additionally, the different clustering results were intrinsically evaluated using the Silhouette coefficient (b). Error bars denote the standard deviation obtained from  $n = 100$  (a) and  $n = 10$  (b) datapoints, respectively.

However, the VGG16 network is only one of many available network architectures; to test, if other networks could possibly agree better with the parameter-based clustering, the same (224, 224, 3) images fed to the VGG16 network were also used to extract features using three other Convolutional Neural Networks: ResNet50,<sup>301</sup> DenseNet121,<sup>302</sup> and InceptionV3<sup>301</sup> that were all pre-trained on the ImageNet dataset. All of those networks were specifically designed for image recognition, but they differ in terms of network architecture. VGG16, for instance, comprises 16 layers of neurons, whereas ResNet50, DenseNet121, and InceptionV3 comprise 50, 121, and 48 layers, respectively. Furthermore, each network architecture comes with different key features: The VGG16 model uses narrow 3x3 filters in all convolutional layers, which gives it the ability to learn very detailed feature representations; ResNet50 uses residual connections between the layers, *i.e.*, each layer does not only feed the next layer but also 2-3 layers further down the network; DenseNet121 comprises an exceptionally dense architecture as each layer is connected to every other layer in a feed-forward fashion; and InceptionV3 uses factorization to reduce the dimensionality of the network.

For each of those networks, individual image preprocessing was conducted using the network-specific preprocessors provided by the keras library (*i.e.*, ResNet50 preprocessing for the ResNet50 network, DenseNet121 preprocessing for the DenseNet121 network, etc.). Then, feature extraction was conducted similar as previously described for the VGG16 network (*i.e.*, the classification layers were

dropped, the multidimensional feature vector was flattened to a 1-dimensional array, all features were transformed using a Yeo-Johnson Power Transformer, and the feature dimensionality was reduced to 100 using a principal component analysis). Among the tested networks, the ResNet50 delivers the highest AMI scores ( $\sim 0.3$ ). However, overall, the achieved scores are rather low. This once again underscores that there is no clear consensus or obvious sorting criterium for the provided dataset (**Figure A8.7a**).

As already explained in **chapter 4.3**, when no clear consensus of the clustering results is observed, each cluster result has to be individually evaluated regarding meaningfulness. For the new analysis based on the features extracted by the neural networks, a detailed analysis goes beyond the scope of this thesis. However, the Silhouette coefficient can be used as a first indicator. When calculating the Silhouette coefficient for the clustering result achieved based on features extracted by the four different networks, clustering based on the VGG16 features delivers the highest coefficients, whereas clustering based on DenseNet121 features delivers considerably lower results (**Figure A8.7b**). Hence, VGG16 indeed seems to be a good starting point for a deep-learning based feature extraction approach. However, also ResNet50 and InceptionV3 might be worth considering: their Silhouette score is below that obtained with VGG16-extracted features, but the obtained error is quite high. This suggests that by adjusting the  $k$ -Means clustering algorithm (in terms of iteration depth and repetitions) could stabilize those results on a higher level that obtained here.

Overall, the results presented here indicate, that clustering the corneal surfaces based on features extracted from deep neural networks should be very well possible. One aspect to keep in mind, however, is that the overall clustering pipeline becomes more and more difficult to interpret, once deep learning approaches are involved (as now, interpreting the features is outright impossible).

## List of abbreviations

3DP	3D printing
AMI	adjusted mutual information score
BW	band width
CA	contact angle
CFU	colony forming unit
CH	Calinski-Harabasz index
CLSM	confocal laser scanning microscopy
CNC	computerized numerical control
CNN	Convolutional Neural Network
DB	Davies-Bouldin index
ddH <sub>2</sub> O	double distilled water
DMSO	dimethyl sulfoxide
DNN	Deep Neural Network
DOPC	1,2-dioleoyl-sn-glycero-3-phosphocholine
DOPE	1,2-dioleoyl-sn-glycero-3-phosphoethanolamine
DPBS	Dulbecco's phosphate-buffered saline
EDC	1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride
ELISA	enzyme-linked immunosorbent assay
EO	ethylene oxide
FBS	fetal bovine serum
FWHM	full width at half maximum
GNB	Gaussian Naïve Bayes
HeLa	human epithelial cell line (named after the donor <u>Henrietta Lacks</u> ) <sup>303</sup>
HEPES	4-[2-hydroxyethyl]-1-piperazineethanesulfonic acid
IQR	inter quartile range
k <sub>1</sub> NN	k <sub>1</sub> Nearest Neighbors
LB	lysogeny broth
MA	methacrylic anhydride
MEM	Minimum Essential Medium Eagle

## List of abbreviations

MES	2-( <i>N</i> -morpholino)ethanesulfonic acid
MI	mutual information score
ML	Machine Learning
MLR	Multinomial Logistic Regression
MSgg	minimal salt glycerol glutamate
MUC5AC	manually purified porcine gastric mucin
MWCO	molecular weight cut-off
NEAA	non-essential amino acid solution
PBS	phosphate-buffered saline
PDMS	polydimethylsiloxane
PGM	porcine gastric mucin (commercially obtained from Sigma-Aldrich)
PGM-MA	porcine gastric mucin (PGM) functionalized with methacrylic anhydride
PU	polyurethane
PVC	polyvinyl chloride
RF	Random Forest
RGP	rigid, gas-permeable (type of contact lenses)
RI	refractive index
RT	room temperature
SAL	sterility assurance level
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SIH	Silhouette coefficient
sulfo-NHS	sulfo- <i>N</i> -hydroxysuccinimide
TCL	tetracycline hydrochloride
TMS-EDTA	N-[(3-trimethoxysilyl)propyl] ethylenediamine triacetic acid trisodium salt
VIF	variance inflation factor
WCSS	within-cluster sum of squares
WST-1	water soluble tetrazolium salt 1

## Bibliography

- (1) Lambert, S. R.; Kraker, R. T.; Pineles, S. L.; Hutchinson, A. K.; Wilson, L. B.; Galvin, J. A.; VanderVeen, D. K., Contact lens correction of aphakia in children: A report by the american academy of ophthalmology. *Ophthalmology* **2018**, *125* (9), 1452.
- (2) Duncan, G.; Wormstone, I.; Liu, C.; Marcantonio, J.; Davies, P., Thapsigargin-coated intraocular lenses inhibit human lens cell growth. *Nature Medicine* **1997**, *3* (9), 1026.
- (3) Jupiter, D. G.; Katz, H. R., Management of irregular astigmatism with rigid gas permeable contact lenses. *The CLAO Journal* **2000**, *26* (1), 14.
- (4) Michaud, L.; Bennett, E. S.; Woo, S. L.; Reeder, R.; Morgan, B. W.; Dinardo, A.; Harthan, J. S., Clinical evaluation of large diameter rigid-gas permeable versus soft toric contact lenses for the correction of refractive astigmatism. A multicenter study. *Eye & Contact Lens* **2018**, *44* (3), 164.
- (5) Li, J.; Hess, R. F.; Chan, L. Y.; Deng, D.; Yang, X.; Chen, X.; Yu, M.; Thompson, B., Quantitative measurement of interocular suppression in anisometropic amblyopia: A case-control study. *Ophthalmology* **2013**, *120* (8), 1672.
- (6) Vaidyanathan, U.; Hopping, G. C.; Liu, H. Y.; Somani, A. N.; Ronquillo, Y. C.; Hoopes, P. C.; Moshirfar, M., Persistent corneal epithelial defects: A review article. *Medical Hypothesis, Discovery and Innovation in Ophthalmology* **2019**, *8* (3), 163.
- (7) DiPasquale, S. A.; Uricoli, B.; DiCerbo, M. C.; Brown, T. L.; Byrne, M. E., Controlled release of multiple therapeutics from silicone hydrogel contact lenses for post-cataract/post-refractive surgery and uveitis treatment. *Translational Vision Science & Technology* **2021**, *10* (14), 5.
- (8) Kim, J.; Kim, M.; Lee, M.-S.; Kim, K.; Ji, S.; Kim, Y.-T.; Park, J.; Na, K.; Bae, K.-H.; Kyun Kim, H., Wearable smart sensor systems integrated on soft contact lenses for wireless ocular diagnostics. *Nature Communications* **2017**, *8* (1), 14997.
- (9) Zhang, D.; Chen, Q.; Shi, C.; Chen, M.; Ma, K.; Wan, J.; Liu, R., Dealing with the foreign-body response to implanted biomaterials: Strategies and applications of new materials. *Advanced Functional Materials* **2021**, *31* (6), 2007226.
- (10) Anderson, J. M.; Rodriguez, A.; Chang, D. T. Foreign body reaction to biomaterials. *Seminars in Immunology* **2008**, *20* (2), 86.
- (11) Davidson, H. J.; Kuonen, V. J., The tear film and ocular mucins. *Veterinary ophthalmology* **2004**, *7* (2), 71.
- (12) Uchino, M.; Tsubota, K., Tear film overview. in *Encyclopedia of the Eye, four-volume set*, **2010**, Elsevier, 263.
- (13) Pang, K.; Lennikov, A.; Yang, M., Hypoxia adaptation in the cornea: Current animal models and underlying mechanisms. *Animal Models and Experimental Medicine* **2021**, *4* (4), 300.
- (14) Sharif, Z.; Sharif, W., Corneal neovascularization: Updates on pathophysiology, investigations & management. *Romanian Journal of Ophthalmology* **2019**, *63* (1), 15.
- (15) Michaud, L.; Vincent, S., Scleral lenses and hypoxia: A balanced approach. *Contact Lens Spectrum* **2019**, *34*, 40.

## Bibliography

- (16) Guillon, M.; Dumbleton, K. A.; Theodoratos, P.; Wong, S.; Patel, K.; Banks, G.; Patel, T., Association between contact lens discomfort and pre-lens tear film kinetics. *Optometry and Vision Science* **2016**, *93* (8), 881.
- (17) Efron, N.; Jones, L.; Bron, A. J.; Knop, E.; Arita, R.; Barabino, S.; McDermott, A. M.; Villani, E.; Willcox, M. D.; Markoulli, M., The tfos international workshop on contact lens discomfort: Report of the contact lens interactions with the ocular surface and adnexa subcommittee. *Investigative Ophthalmology & Visual Science* **2013**, *54* (11), TFOS98.
- (18) Zhang, X.; Zhang, Y.; Jin, Z., A review of the bio-tribology of medical devices. *Friction* **2022**, *10* (1), 4.
- (19) Klaassen, M.; De Vries, E.; Masen, M., Friction in the contact between skin and a soft counter material: Effects of hardness and surface finish. *Journal of the Mechanical Behavior of Biomedical Materials* **2019**, *92*, 137.
- (20) Thorarinsdottir, H. R.; Kander, T.; Holmberg, A.; Petronis, S.; Klarin, B., Biofilm formation on three different endotracheal tubes: A prospective clinical trial. *Critical Care* **2020**, *24* (1), 1.
- (21) Yoo, B. Y.; Kim, B. H.; Lee, J. S.; Shin, B. H.; Kwon, H.; Koh, W.-G.; Heo, C. Y., Dual surface modification of pdms-based silicone implants to suppress capsular contracture. *Acta Biomaterialia* **2018**, *76*, 56.
- (22) Shabahang, S.; Kim, S.; Yun, S. H., Light-guiding biomaterials for biomedical applications. *Advanced Functional Materials* **2018**, *28* (24), 1706635.
- (23) Song, J.; Winkeljann, B.; Lieleg, O., Biopolymer-based coatings: Promising strategies to improve the biocompatibility and functionality of materials used in biomedical engineering. *Advanced Materials Interfaces* **2020**, *7* (17), 2000850.
- (24) Yadav, P.; Yadav, H.; Shah, V. G.; Shah, G.; Dhaka, G., Biomedical biopolymers, their origin and evolution in biomedical sciences: A systematic review. *Journal of Clinical and Diagnostic Research: JCDR* **2015**, *9* (9), ZE21.
- (25) Pattanashetti, N. A.; Heggannavar, G. B.; Kariduraganavar, M. Y., Smart biopolymers and their biomedical applications. *Procedia Manufacturing* **2017**, *12*, 263.
- (26) Biswas, M. C.; Jony, B.; Nandy, P. K.; Chowdhury, R. A.; Halder, S.; Kumar, D.; Ramakrishna, S.; Hassan, M.; Ahsan, M. A.; Hoque, M. E., Recent advancement of biopolymers and their potential biomedical applications. *Journal of Polymers and the Environment* **2021**, *30*, 51.
- (27) Kesel, S.; Mader, A.; Seeberger, P. H.; Lieleg, O.; Opitz, M., Carbohydrate coating reduces adhesion of biofilm-forming bacillus subtilis to gold surfaces. *Applied and Environmental Microbiology* **2014**, *80* (19), 5911.
- (28) Zheng, M.; Wang, X.; Chen, Y.; Yue, O.; Bai, Z.; Cui, B.; Jiang, H.; Liu, X., A review of recent progress on collagen-based biomaterials. *Advanced Healthcare Materials* **2022**, 2202042.
- (29) Hu, J.; Song, Y.; Zhang, C.; Huang, W.; Chen, A.; He, H.; Zhang, S.; Chen, Y.; Tu, C.; Liu, J., Highly aligned electrospun collagen/polycaprolactone surgical sutures with sustained release of growth factors for wound regeneration. *ACS Applied Bio Materials* **2020**, *3* (2), 965.
- (30) Pupkaite, J.; Ahumada, M.; Mclaughlin, S.; Temkit, M.; Alaziz, S.; Seymour, R.; Ruel, M.; Kochevar, I.; Griffith, M.; Suuronen, E. J., Collagen-based photoactive agent for tissue bonding. *ACS Applied Materials & Interfaces* **2017**, *9* (11), 9265.

- (31) Durand, E.; Sharkawi, T.; Leclerc, G.; Raveleau, M.; van Der Leest, M.; Vert, M.; Lafont, A., Head-to-head comparison of a drug-free early programmed dismantling polylactic acid bioresorbable scaffold and a metallic stent in the porcine coronary artery: Six-month angiography and optical coherence tomographic follow-up study. *Circulation: Cardiovascular Interventions* **2014**, *7* (1), 70.
- (32) Badal, R.; Mohapatra, S.; Bhoi, C.; Sharma, N.; Jena, S.; Panda, S. K., Multiphysics hemodynamic behavior of polylactic acid-based stent: A coupled simulation approach. *Journal of Engineering and Science in Medical Diagnostics and Therapy* **2021**, *4* (2), 021005.
- (33) Ong, S.-Y.; Wu, J.; Moochhala, S. M.; Tan, M.-H.; Lu, J., Development of a chitosan-based wound dressing with improved hemostatic and antimicrobial properties. *Biomaterials* **2008**, *29* (32), 4323.
- (34) Khan, M. A.; Mujahid, M., A review on recent advances in chitosan based composite for hemostatic dressings. *International Journal of Biological Macromolecules* **2019**, *124*, 138.
- (35) Kim, D. J.; Jung, M.-Y.; Park, J.-H.; Pak, H.-J.; Kim, M.; Chuck, R. S.; Park, C. Y., Moxifloxacin releasing intraocular implant based on a cross-linked hyaluronic acid membrane. *Scientific reports* **2021**, *11* (1), 24115.
- (36) Holland, C.; Numata, K.; Rnjak-Kovacina, J.; Seib, F. P., The biomedical use of silk: Past, present, future. *Advanced Healthcare Materials* **2019**, *8* (1), 1800465.
- (37) Shi, C.; Pu, X.; Zheng, G.; Feng, X.; Yang, X.; Zhang, B.; Zhang, Y.; Yin, Q.; Xia, H., An antibacterial and absorbable silk-based fixation material with impressive mechanical properties and biocompatibility. *Scientific Reports* **2016**, *6* (1), 37418.
- (38) Biran, R.; Pond, D., Heparin coatings for improving blood compatibility of medical devices. *Advanced Drug Delivery Reviews* **2017**, *112*, 12.
- (39) Chen, D., Heparin beyond anti-coagulation. *Current Research in Translational Medicine* **2021**, *69* (4), 103300.
- (40) Li, X.; Shi, H.; Pan, K.; Dai, M.; Wei, W.; Liu, X., Improved biocompatibility and antibacterial property of zinc alloy fabricated with  $\gamma$ -polyglutamic acid-g-dopamine/copper coatings for orthopedic implants. *Progress in Organic Coatings* **2022**, *173*, 107215.
- (41) Wu, W.; Sun, X.; Zhu, C.-L.; Zhang, F.; Zeng, R.-C.; Zou, Y.-H.; Li, S.-Q., Biocorrosion resistance and biocompatibility of mg–al layered double hydroxide/poly-l-glutamic acid hybrid coating on magnesium alloy az31. *Progress in Organic Coatings* **2020**, *147*, 105746.
- (42) Verza, B. S.; van den Beucken, J. J.; Brandt, J. V.; Junior, M. J.; Barao, V. A.; Piazza, R. D.; Tagit, O.; Spolidorio, D. M.; Vergani, C. E.; de Avila, E. D., A long-term controlled drug-delivery with anionic beta cyclodextrin complex in layer-by-layer coating for percutaneous implants devices. *Carbohydrate Polymers* **2021**, *257*, 117604.
- (43) Wankar, J.; Kotla, N. G.; Gera, S.; Rasala, S.; Pandit, A.; Rochev, Y. A., Recent advances in host–guest self-assembled cyclodextrin carriers: Implications for responsive drug delivery and biomedical engineering. *Advanced Functional Materials* **2020**, *30* (44), 1909049.
- (44) Sharif, S.; Abbas, G.; Hanif, M.; Bernkop-Schnürch, A.; Jalil, A.; Yaqoob, M., Mucoadhesive micro-composites: Chitosan coated halloysite nanotubes for sustained drug delivery. *Colloids and Surfaces B: Biointerfaces* **2019**, *184*, 110527.

## Bibliography

- (45) Tao, F.; Ma, S.; Tao, H.; Jin, L.; Luo, Y.; Zheng, J.; Xiang, W.; Deng, H., Chitosan-based drug delivery systems: From synthesis strategy to osteomyelitis treatment—a review. *Carbohydrate Polymers* **2021**, *251*, 117063.
- (46) Knoop, K. A.; Newberry, R. D., Goblet cells: Multifaceted players in immunity at mucosal surfaces. *Mucosal Immunology* **2018**, *11* (6), 1551.
- (47) Bansil, R.; Stanley, E.; LaMont, J. T., Mucin biophysics. *Annual Review of Physiology* **1995**, *57* (1), 635.
- (48) Bansil, R.; Turner, B. S., Mucin structure, aggregation, physiological functions and biomedical applications. *Current Opinion in Colloid & Interface Science* **2006**, *11* (2-3), 164.
- (49) Linden, S.; Sutton, P.; Karlsson, N.; Korolik, V.; McGuckin, M., Mucins in the mucosal barrier to infection. *Mucosal Immunology* **2008**, *1* (3), 183.
- (50) McGuckin, M. A.; Lindén, S. K.; Sutton, P.; Florin, T. H., Mucin dynamics and enteric pathogens. *Nature Reviews Microbiology* **2011**, *9* (4), 265.
- (51) Winkeljann, B.; Boettcher, K.; Balzer, B. N.; Lieleg, O., Mucin coatings prevent tissue damage at the cornea–contact lens interface. *Advanced Materials Interfaces* **2017**, *4* (19), 1700186.
- (52) Austin, R.; Giusca, C.; Macaulay, G.; Moazzez, R.; Bartlett, D., Confocal laser scanning microscopy and area-scale analysis used to quantify enamel surface textural changes from citric acid demineralization and salivary remineralization in vitro. *Dental Materials* **2016**, *32* (2), 278.
- (53) Boettcher, K.; Winkeljann, B.; Schmidt, T. A.; Lieleg, O., Quantification of cartilage wear morphologies in unidirectional sliding experiments: Influence of different macromolecular lubricants. *Biotribology* **2017**, *12*, 43.
- (54) Alam, M. M.; Islam, M. T., Machine learning approach of automatic identification and counting of blood cells. *Healthcare Technology Letters* **2019**, *6* (4), 103.
- (55) Yamanishi, C.; Parigoris, E.; Takayama, S., Kinetic analysis of label-free microscale collagen gel contraction using machine learning-aided image analysis. *Frontiers in Bioengineering and Biotechnology* **2020**, *8*, 1123.
- (56) Park, S.; Ahn, J. W.; Jo, Y.; Kang, H.-Y.; Kim, H. J.; Cheon, Y.; Kim, J. W.; Park, Y.; Lee, S.; Park, K., Label-free tomographic imaging of lipid droplets in foam cells for machine-learning-assisted therapeutic evaluation of targeted nanodrugs. *ACS Nano* **2020**, *14* (2), 1856.
- (57) Spanoudaki, V.; Doloff, J. C.; Huang, W.; Norcross, S. R.; Farah, S.; Langer, R.; Anderson, D. G., Simultaneous spatiotemporal tracking and oxygen sensing of transient implants in vivo using hot-spot mri and machine learning. *Proceedings of the National Academy of Sciences* **2019**, *116* (11), 4861.
- (58) Cunningham, J. M.; Koytiger, G.; Sorger, P. K.; AlQuraishi, M., Biophysical prediction of protein–peptide interactions and signaling networks using machine learning. *Nature Methods* **2020**, *17* (2), 175.
- (59) Jones, P.; Coupette, F.; Härtel, A.; Lee, A. A., Bayesian unsupervised learning reveals hidden structure in concentrated electrolytes. *The Journal of Chemical Physics* **2021**, *154* (13), 134902.

- (60) Clauser, J. C.; Maas, J.; Arens, J.; Schmitz-Rode, T.; Steinseifer, U.; Berkels, B., Automation of hemocompatibility analysis using image segmentation and supervised classification. *Engineering Applications of Artificial Intelligence* **2021**, *97*, 104009.
- (61) Chu, A.; Nguyen, D.; Talathi, S. S.; Wilson, A. C.; Ye, C.; Smith, W. L.; Kaplan, A. D.; Duoss, E. B.; Stolaroff, J. K.; Giera, B., Automated detection and sorting of microencapsulation via machine learning. *Lab on a Chip* **2019**, *19* (10), 1808.
- (62) Madiona, R. M.; Winkler, D. A.; Muir, B. W.; Pigram, P. J., Optimal machine learning models for robust materials classification using tof-sims data. *Applied Surface Science* **2019**, *487*, 773.
- (63) Barnett, J. W.; Bilchak, C. R.; Wang, Y.; Benicewicz, B. C.; Murdock, L. A.; Bereau, T.; Kumar, S. K., Designing exceptional gas-separation polymer membranes using machine learning. *Science Advances* **2020**, *6* (20), eaaz4301.
- (64) Liang, N.; Li, B.; Jia, Z.; Wang, C.; Wu, P.; Zheng, T.; Wang, Y.; Qiu, F.; Wu, Y.; Su, J., Ultrasensitive detection of circulating tumour DNA via deep methylation sequencing aided by machine learning. *Nature Biomedical Engineering* **2021**, *5* (6), 586.
- (65) Campano, C.; Lopez-Exposito, P.; Gonzalez-Aguilera, L.; Blanco, Á.; Negro, C., In-depth characterization of the aggregation state of cellulose nanocrystals through analysis of transmission electron microscopy images. *Carbohydrate Polymers* **2021**, *254*, 117271.
- (66) Ruggeri, F. S.; Flagmeier, P.; Kumita, J. R.; Meisl, G.; Chirgadze, D. Y.; Bongiovanni, M. N.; Knowles, T. P.; Dobson, C. M., The influence of pathogenic mutations in  $\alpha$ -synuclein on biophysical and structural characteristics of amyloid fibrils. *ACS Nano* **2020**, *14* (5), 5213.
- (67) Rickert, C. A.; Lieleg, O., Machine learning approaches for biomolecular, biophysical, and biomaterials research. *Biophysics Reviews* **2022**, *3* (2), 021306.
- (68) Thornton, D. J.; Rousseau, K.; McGuckin, M. A., Structure and function of the polymeric mucins in airways mucus. *Annual Reviews in Physiology* **2008**, *70*, 459.
- (69) Desseyn, J.-L.; Aubert, J.-P.; Porchet, N.; Laine, A., Evolution of the large secreted gel-forming mucins. *Molecular Biology and Evolution* **2000**, *17* (8), 1175.
- (70) Chen, Y.; Zhao, Y. H.; Kalaslavadi, T. B.; Hamati, E.; Nehrke, K.; Le, A. D.; Ann, D. K.; Wu, R., Genome-wide search and identification of a novel gel-forming mucin muc19/muc19 in glandular tissues. *American Journal of Respiratory Cell and Molecular Biology* **2004**, *30* (2), 155.
- (71) Baos, S.; Phillips, D. B.; Wildling, L.; McMaster, T.; Berry, M., Distribution of sialic acids on mucins and gels: A defense mechanism. *Biophysical Journal* **2012**, *102* (1), 176.
- (72) Marczynski, M.; Balzer, B. N.; Jiang, K.; Lutz, T. M.; Crouzier, T.; Lieleg, O., Charged glycan residues critically contribute to the adsorption and lubricity of mucins. *Colloids and Surfaces B: Biointerfaces* **2020**, *187*, 110614.
- (73) Crouzier, T.; Boettcher, K.; Geonnotti, A. R.; Kavanaugh, N. L.; Hirsch, J. B.; Ribbeck, K.; Lieleg, O., Modulating mucin hydration and lubrication by deglycosylation and polyethylene glycol binding. *Advanced Materials Interfaces* **2015**, *2* (18), 1500308.
- (74) Lutz, T. M.; Marczynski, M.; Grill, M. J.; Wall, W. A.; Lieleg, O., Repulsive backbone-backbone interactions modulate access to specific and unspecific binding sites on surface-bound mucins. *Langmuir* **2020**, *36* (43), 12973.

## Bibliography

- (75) Seregni, E.; Botti, C.; Massaron, S.; Lombardo, C.; Capobianco, A.; Bogni, A.; Bombardieri, E., Structure, function and gene expression of epithelial mucins. *Tumori Journal* **1997**, *83* (3), 625.
- (76) Perez-Vilar, J.; Hill, R. L., The structure and assembly of secreted mucins. *Journal of Biological Chemistry* **1999**, *274* (45), 31751.
- (77) Ambort, D.; van der Post, S.; Johansson, M. E.; MacKenzie, J.; Thomsson, E.; Kregel, U.; Hansson, G. C., Function of the cyst domain of the gel-forming muc2 mucin. *Biochemical Journal* **2011**, *436* (1), 61.
- (78) Ridley, C.; Kouvatso, N.; Raynal, B. D.; Howard, M.; Collins, R. F.; Desseyn, J.-L.; Jowitt, T. A.; Baldock, C.; Davis, C. W.; Hardingham, T. E., Assembly of the respiratory mucin muc5b: A new model for a gel-forming mucin. *Journal of Biological Chemistry* **2014**, *289* (23), 16409.
- (79) Shi, L.; Ardehali, R.; Caldwell, K. D.; Valint, P., Mucin coating on polymeric material surfaces to suppress bacterial adhesion. *Colloids and Surfaces B: Biointerfaces* **2000**, *17* (4), 229.
- (80) Lieleg, O.; Lieleg, C.; Bloom, J.; Buck, C. B.; Ribbeck, K., Mucin biopolymers as broad-spectrum antiviral agents. *Biomacromolecules* **2012**, *13* (6), 1724.
- (81) Schömig, V. J.; Käs Dorf, B. T.; Scholz, C.; Bidmon, K.; Lieleg, O.; Berensmeier, S., An optimized purification process for porcine gastric mucin with preservation of its native functional properties. *RSC Advances* **2016**, *6* (50), 44932.
- (82) Marczyński, M.; Rickert, C. A.; Fuhrmann, T.; Lieleg, O., An improved, filtration-based process to purify functional mucins from mucosal tissues with high yields. *Separation and Purification Technology* **2022**, *294*, 121209.
- (83) Marczyński, M.; Jiang, K.; Blakeley, M.; Srivastava, V.; Vilaplana, F.; Crouzier, T.; Lieleg, O., Structural alterations of mucins are associated with losses in functionality. *Biomacromolecules* **2021**, *22* (4), 1600.
- (84) Mironi-Harpaz, I.; Wang, D. Y.; Venkatraman, S.; Seliktar, D., Photopolymerization of cell-encapsulating hydrogels: Crosslinking efficiency versus cytotoxicity. *Acta Biomaterialia* **2012**, *8* (5), 1838.
- (85) Rickert, C. A.; Wittmann, B.; Fromme, R.; Lieleg, O., Highly transparent covalent mucin coatings improve the wettability and tribology of hydrophobic contact lenses. *ACS Applied Materials & Interfaces* **2020**, *12* (25), 28024.
- (86) Winkeljann, B.; Leipold, P. M. A.; Lieleg, O., Macromolecular coatings enhance the tribological performance of polymer-based lubricants. *Advanced Materials Interfaces* **2019**, *6* (16), 1900366.
- (87) Gallagher, J. T.; Harding, M.; Dale, R. E. Sialic acid as a cell surface binding site for wheat germ agglutinin. In *Proceedings of the Fifth Lectin Meeting Bern, May 31–June 5, 1982*, **2020**, De Gruyter, 311.
- (88) Lee, J. H.; Kee, C. W., The significance of tear film break-up time in the diagnosis of dry eye syndrome. *Korean Journal of Ophthalmology* **1988**, *2* (2), 69.
- (89) Rantamäki, A. H.; Seppänen-Laakso, T.; Oresic, M.; Jauhiainen, M.; Holopainen, J. M., Human tear fluid lipidome: From composition to function. *PLoS One* **2011**, *6* (5), e19553.
- (90) Marczyński, M., Structure-function relationships of mucin glycoproteins. Technische Universität München, **2022**.

- (91) Hsu, S. M.; Gates, R. S., Boundary lubricating films: Formation and lubrication mechanism. *Tribology International* **2005**, *38* (3), 305.
- (92) Gaisinskaya, A.; Ma, L.; Silbert, G.; Sorokin, R.; Tairy, O.; Goldberg, R.; Kampf, N.; Klein, J., Hydration lubrication: Exploring a new paradigm. *Faraday Discussions* **2012**, *156* (1), 217.
- (93) Jahn, S.; Klein, J., Hydration lubrication: The macromolecular domain. *Macromolecules* **2015**, *48* (15), 5059.
- (94) Boettcher, K.; Grumbein, S.; Winkler, U.; Nachtsheim, J.; Lieleg, O., Adapting a commercial shear rheometer for applications in cartilage research. *Review of Scientific Instruments* **2014**, *85* (9), 093903.
- (95) Käsdorf, B. T.; Weber, F.; Petrou, G.; Srivastava, V.; Crouzier, T.; Lieleg, O., Mucin-inspired lubrication on hydrophobic surfaces. *Biomacromolecules* **2017**, *18* (8), 2454.
- (96) Brandts, J. F., The thermodynamics of protein denaturation. I. The denaturation of chymotrypsinogen. *Journal of the American Chemical Society* **1964**, *86* (20), 4291.
- (97) Lu, M. Y.; Chen, T. Y.; Williamson, D. F.; Zhao, M.; Shady, M.; Lipkova, J.; Mahmood, F., Ai-based pathology predicts origins for cancers of unknown primary. *Nature* **2021**, *594* (7861), 106.
- (98) Gu, G. X.; Chen, C.-T.; Buehler, M. J., De novo composite design based on machine learning algorithm. *Extreme Mechanics Letters* **2018**, *18*, 19.
- (99) Ghouli, S.; Ayatollahi, M. R.; Bahrami, B.; Jamali, J., In-situ optical approach to predict mixed mode fracture in a polymeric biomaterial. *Theoretical and Applied Fracture Mechanics* **2021**, *118*, 103211.
- (100) Donlan, R. M., Biofilms: Microbial life on surfaces. *Emerging Infectious Diseases* **2002**, *8* (9), 881.
- (101) Branda, S. S.; Chu, F.; Kearns, D. B.; Losick, R.; Kolter, R., A major protein component of the bacillus subtilis biofilm matrix. *Molecular Microbiology* **2006**, *59* (4), 1229.
- (102) Van Rossum, G.; Drake, F. L. *Python 3 reference manual*; CreateSpace, **2009**.
- (103) Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J., Array programming with numpy. *Nature* **2020**, *585* (7825), 357.
- (104) McKinney, W., Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, **2010**, *445*, 51.
- (105) The pandas development team, Pandas-dev/pandas: Pandas, *Zenodo*, **2023**.
- (106) Hunter, J. D., Matplotlib: A 2d graphics environment. *IEEE Annals of the History of Computing* **2007**, *9* (03), 90.
- (107) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J., Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods* **2020**, *17* (3), 261.
- (108) Waskom, M.; Botvinnik, O.; Gelbart, M.; Ostblom, J.; Hobson, P.; Lukauskas, S.; Gemperline, D. C.; Augspurger, T.; Halchenko, Y.; Warmenhoven, J., Seaborn: Statistical data visualization. *Astrophysics Source Code Library* **2020**, 2012.2015.

## Bibliography

- (109) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research* **2011**, *12*, 2825.
- (110) Suwanda, R.; Syahputra, Z.; Zamzami, E., Analysis of euclidean distance and manhattan distance in the k-means algorithm for variations number of centroid k. *Journal of Physics: Conference Series* **2020**, *1566*, 012058.
- (111) Rish, I., An empirical study of the naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence* **2001**, *3* (22), 41.
- (112) Zhang, H., Exploring conditions for the optimality of naive bayes. *International Journal of Pattern Recognition and Artificial Intelligence* **2005**, *19* (02), 183.
- (113) VanderPlas, J., Python data science handbook: Essential tools for working with data, O'Reilly Media, Inc., **2016**.
- (114) Lou, W.; Wang, X.; Chen, F.; Chen, Y.; Jiang, B.; Zhang, H., Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and gaussian naive bayes. *PLoS One* **2014**, *9* (1), e86703.
- (115) Ng, A. Y., Feature selection, L1 vs. L2 regularization, and rotational invariance. *Proceedings of the twenty-first international conference on Machine learning* **2004**, 78.
- (116) Bolton, E. E.; Chen, J.; Kim, S.; Han, L.; He, S.; Shi, W.; Simonyan, V.; Sun, Y.; Thiessen, P. A.; Wang, J., PubChem3D: A new resource for scientists. *Journal of Cheminformatics* **2011**, *3* (1), 32.
- (117) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D., Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry* **2002**, *45* (12), 2615.
- (118) Bondi, A. v., Van der waals volumes and radii. *The Journal of Physical Chemistry* **1964**, *68* (3), 441.
- (119) Chemaxon. *Elemental analysis plugin*. <https://docs.chemaxon.com/display/docs/elemental-analysis-plugin.md#src-1806662-elementalanalysisplugin-standardatws> (accessed 23.02.2022).
- (120) Meija, J.; Coplen, T. B.; Berglund, M.; Brand, W. A.; Bièvre, P. D.; Gröning, M.; Holden, N. E.; Irrgeher, J.; Loss, R. D.; Walczyk, T.; et al., Atomic weights of the elements 2013 (iupac technical report). *Pure and Applied Chemistry* **2016**, *88* (3), 265.
- (121) Chemaxon. *Isoelectric point plugin*. <https://docs.chemaxon.com/display/docs/isoelectric-point-plugin.md> (accessed 25.02.2022).
- (122) Chemaxon. *Dipole moment calculation plugin*. <https://docs.chemaxon.com/display/docs/dipole-moment-calculation-plugin.md> (accessed 25.02.2022).
- (123) Schlosser, P.; Asgharian, B.; Medinsky, M., 1.04 inhalation exposure and absorption of toxicants. *Comprehensive Toxicology: General Principles* **2010**, *1*, 75.
- (124) Chemaxon. *Topology analysis*. Chemaxon, <https://chemaxon.com/webinar/topology-analysis> (accessed 23.02.2022).
- (125) Chemaxon. *Hydrogen bond donor acceptor plugin*. <https://docs.chemaxon.com/display/docs/hydrogen-bond-donor-acceptor-plugin.md> (accessed 25.02.2022).

- (126) Hagberg, A.; Swart, P.; S Chult, D. Exploring network structure, dynamics, and function using NetworkX; *Proceedings of the 7<sup>th</sup> Python in Science Conference* **2008**, 11.
- (127) Taylor, R., Interpretation of the correlation coefficient: A basic review. *Journal of Diagnostic Medical Sonography* **1990**, 6 (1), 35.
- (128) Ruhnau, B., Eigenvector-centrality – a node-centrality? *Social Networks* **2000**, 22 (4), 357.
- (129) Roffo, G.; Melzi, S. Ranking to learn. In *International Workshop on New Frontiers in Mining Complex Patterns*, **2016**, Springer, 19.
- (130) Thompson, C. G.; Kim, R. S.; Aloe, A. M.; Becker, B. J., Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. *Basic and Applied Social Psychology* **2017**, 39 (2), 81.
- (131) Yeo, I. K.; Johnson, R. A., A new family of power transformations to improve normality or symmetry. *Biometrika* **2000**, 87 (4), 954.
- (132) Box, G. E.; Cox, D. R., An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)* **1964**, 26 (2), 211.
- (133) Rousseeuw, P. J., Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **1987**, 20, 53.
- (134) Caliński, T.; Harabasz, J., A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* **1974**, 3 (1), 1.
- (135) Davies, D. L.; Bouldin, D. W., A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence* **1979**, PAMI-1 (2), 224.
- (136) Vinh, N. X.; Epps, J.; Bailey, J. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? *Proceedings of the 26th annual international conference on machine learning* **2009**, 1073.
- (137) Mishra, V.; Negi, S.; Kar, S.; Sharma, A. K.; Rajbahadur, Y. N. K.; Kumar, A., Recent advances in fused deposition modeling based additive manufacturing of thermoplastic composite structures: A review. *Journal of Thermoplastic Composite Materials* **2022**, 08927057221102857.
- (138) Alagoz, A. S.; Hasirci, V., 3d printing of polymeric tissue engineering scaffolds using open-source fused deposition modeling. *Emergent Materials* **2020**, 3 (4), 429.
- (139) Golcha, U.; Praveen, A.; Paul, D. B., Direct ink writing of ceramics for bio medical applications—a review. *IOP Conference Series: Materials Science and Engineering* **2020**; 912, 032041.
- (140) Bom, S.; Ribeiro, R.; Ribeiro, H. M.; Santos, C.; Marto, J., On the progress of hydrogel-based 3d printing: Correlating rheological properties with printing behaviour. *International Journal of Pharmaceutics* **2022**, 121506.
- (141) Elbadawi, M.; Gustaffson, T.; Gaisford, S.; Basit, A. W., 3D printing tablets: Predicting printability and drug dissolution from rheological data. *International Journal of Pharmaceutics* **2020**, 590, 119868.
- (142) Outrequin, T. C. R.; Gamonpilas, C.; Siriawatwechakul, W.; Sreearunothai, P., Extrusion-based 3d printing of food biopolymers: A highlight on the important rheological parameters to reach printability. *Journal of Food Engineering* **2022**, 111371.

## Bibliography

- (143) Yi, H.-G.; Kim, H.; Kwon, J.; Choi, Y.-J.; Jang, J.; Cho, D.-W., Application of 3d bioprinting in the prevention and the therapy for human diseases. *Signal Transduction and Targeted Therapy* **2021**, *6* (1), 177.
- (144) Habib, M. A.; Khoda, B., Rheological analysis of bio-ink for 3D bio-printing processes. *Journal of Manufacturing Processes* **2022**, *76*, 708.
- (145) Kim, M. K.; Jeong, W.; Lee, S. M.; Kim, J. B.; Jin, S.; Kang, H.-W., Decellularized extracellular matrix-based bio-ink with enhanced 3d printability and mechanical properties. *Biofabrication* **2020**, *12* (2), 025003.
- (146) Wu, W., Study on 3d printing technology and mechanical properties of a nano-enhanced composite hydrogel bio-ink. *Micro & Nano Letters* **2020**, *15* (13), 964.
- (147) Ouyang, L.; Highley, C. B.; Rodell, C. B.; Sun, W.; Burdick, J. A., 3d printing of shear-thinning hyaluronic acid hydrogels with secondary cross-linking. *ACS Biomaterials Science & Engineering* **2016**, *2* (10), 1743.
- (148) Lieleg, O.; Claessens, M. M. A. E.; Luan, Y.; Bausch, A. R., Transient binding and dissipation in cross-linked actin networks. *Physical Review Letters* **2008**, *101* (10), 108101.
- (149) Samani, A.; Bishop, J.; Luginbuhl, C.; Plewes, D. B., Measuring the elastic modulus of ex vivo small tissue samples. *Physics in Medicine & Biology* **2003**, *48* (14), 2183.
- (150) Egorov, V.; Tsyuryupa, S.; Kanilo, S.; Kogit, M.; Sarvazyan, A., Soft tissue elastometer. *Medical Engineering & Physics* **2008**, *30* (2), 206.
- (151) Sarvazyan, A., Elastic properties of soft tissues. *Handbook of elastic properties of solids, liquids and gases* **2001**, *3*, 107.
- (152) Duffy, C. V.; David, L.; Crouzier, T., Covalently-crosslinked mucin biopolymer hydrogels for sustained drug delivery. *Acta Biomaterialia* **2015**, *20*, 51.
- (153) Brand, J. S.; Forster, L.; Böck, T.; Stahlhut, P.; Teßmar, J.; Groll, J.; Albrecht, K., Covalently crosslinked pig gastric mucin hydrogels prepared by radical-based chain-growth and thiol-ene mechanisms. *Macromolecular Bioscience* **2021**, 2100274.
- (154) Tomás, H.; Alves, C. S.; Rodrigues, J., Laponite®: A key nanoplatform for biomedical applications? *Nanomedicine: Nanotechnology, Biology and Medicine* **2018**, *14* (7), 2407.
- (155) Ghadiri, M.; Chrzanowski, W.; Lee, W.; Fathi, A.; Dehghani, F.; Rohanizadeh, R., Physico-chemical, mechanical and cytotoxicity characterizations of laponite®/alginate nanocomposite. *Applied Clay Science* **2013**, *85*, 64.
- (156) Song, J.; Lutz, T. M.; Lang, N.; Lieleg, O., Bioinspired dopamine/mucin coatings provide lubricity, wear protection, and cell-repellent properties for medical applications. *Advanced Healthcare Materials* **2021**, *10* (4), 2000831.
- (157) Kimna, C.; Bauer, M. G.; Lutz, T. M.; Mansi, S.; Akyuz, E.; Doganyigit, Z.; Karakol, P.; Mela, P.; Lieleg, O., Multifunctional “janus-type” bilayer films combine broad-range tissue adhesion with guided drug release. *Advanced Functional Materials* **2022**, *32* (30), 2105721.
- (158) Winkeljann, B.; Bauer, M. G.; Marczyński, M.; Rauh, T.; Sieber, S. A.; Lieleg, O., Covalent mucin coatings form stable anti-biofouling layers on a broad range of medical polymer materials. *Advanced Materials Interfaces* **2020**, *7* (4), 1902069.

- (159) Yan, H.; Seignez, C.; Hjorth, M.; Winkeljann, B.; Blakeley, M.; Lieleg, O.; Phillipson, M.; Crouzier, T., Immune-informed mucin hydrogels evade fibrotic foreign body response in vivo. *Advanced Functional Materials* **2019**, *29* (46), 1902581.
- (160) Yan, H.; Hjorth, M.; Winkeljann, B.; Dobryden, I.; Lieleg, O.; Crouzier, T., Glyco-modification of mucin hydrogels to investigate their immune activity. *ACS Applied Materials & Interfaces* **2020**, *12* (17), 19324.
- (161) Mohammadi, S.; Ravanbakhsh, H.; Taheri, S.; Bao, G.; Mongeau, L., Immunomodulatory microgels support proregenerative macrophage activation and attenuate fibroblast collagen synthesis. *Advanced Healthcare Materials* **2022**, *11* (11), 2102366.
- (162) Bostan, L. E.; Clarkin, C. E.; Mousa, M.; Worsley, P. R.; Bader, D. L.; Dawson, J. I.; Evans, N. D., Synthetic nanoclay gels do not cause skin irritation in healthy human volunteers. *ACS Biomaterials Science & Engineering* **2021**, *7* (6), 2716.
- (163) Guillon, M.; Maissa, C., Dry eye symptomatology of soft contact lens wearers and nonwearers. *Optometry and Vision Science* **2005**, *82* (9), 829.
- (164) Korb, D. R.; Baron, D. F.; Herman, J. P.; Finnemore, V. M.; Exford, J. M.; Hermosa, J. L.; Leahy, C. D.; Glonek, T.; Greiner, J. V., Tear film lipid layer thickness as a function of blinking. *Cornea* **1994**, *13* (4), 354.
- (165) Goren, M. B.; Goren, S. B., Diagnostic tests in patients with symptoms of keratoconjunctivitis sicca. *American Journal of Ophthalmology* **1988**, *106* (5), 570.
- (166) Korb, D. R.; Herman, J. P.; Greiner, J. V.; Scaffidi, R. C.; Finnemore, V. M.; Exford, J. M.; Blackie, C. A.; Douglass, T., Lid wiper epitheliopathy and dry eye symptoms. *Eye Contact Lens* **2005**, *31* (1), 2.
- (167) Papas, E., On the relationship between soft contact lens oxygen transmissibility and induced limbal hyperaemia. *Experimental Eye Research* **1998**, *67* (2), 125.
- (168) McMonnies, C.; Chapman-Davies, A.; Holden, B., The vascular response to contact lens wear. *American Journal of Optometry and Physiological Optics* **1982**, *59* (10), 795.
- (169) Dixon, J. M.; Lawaczeck, E., Corneal vascularization due to contact lenses. *Archives of Ophthalmology* **1963**, *69* (1), 72.
- (170) Keay, L.; Edwards, K.; Naduvilath, T.; Taylor, H. R.; Snibson, G. R.; Forde, K.; Stapleton, F., Microbial keratitis: Predisposing factors and morbidity. *Ophthalmology* **2006**, *113* (1), 109.
- (171) Green, M.; Sara, S.; Hughes, I.; Apel, A.; Stapleton, F., Trends in contact lens microbial keratitis 1999 to 2015: A retrospective clinical review. *Clinical & Experimental Ophthalmology* **2019**, *47* (6), 726.
- (172) Wong, H.; Fatt, I.; Radke, C., Deposition and thinning of the human tear film. *Journal of Colloid and Interface Science* **1996**, *184* (1), 44.
- (173) Borruto, A.; Crivellone, G.; Marani, F., Influence of surface wettability on friction and wear tests. *Wear* **1998**, *222* (1), 57.
- (174) Fleiszig, S. M.; Kroken, A. R.; Nieto, V.; Grosser, M. R.; Wan, S. J.; Metruccio, M. M.; Evans, D. J., Contact lens-related corneal infection: Intrinsic resistance and its compromise. *Progress in Retinal and Eye Research* **2019**, *76*, 100804.

## Bibliography

- (175) Cheung, N.; Nagra, P.; Hammersmith, K., Emerging trends in contact lens-related infections. *Current Opinion in Ophthalmology* **2016**, *27* (4), 327.
- (176) Retuerto, M. A.; Szczołka-Flynn, L.; Mukherjee, P. K.; Debanne, S.; Iyengar, S. K.; Richardson, B.; Cameron, M.; Ghannoum, M. A., Diversity of ocular surface bacterial microbiome adherent to worn contact lenses and bacterial communities associated with care solution use. *Eye Contact Lens* **2019**, *45* (5), 331.
- (177) Datta, A.; Stapleton, F.; Willcox, M. D., Bacterial coaggregation among the most commonly isolated bacteria from contact lens cases. *Investigative Ophthalmology & Visual Science* **2017**, *58* (1), 50.
- (178) Carney, L. G.; Hill, R. M., The nature of normal blinking patterns. *Acta Ophthalmologica* **1982**, *60* (3), 427.
- (179) Korogiannaki, M.; Samsom, M.; Schmidt, T. A.; Sheardown, H., Surface-functionalized model contact lenses with a bioinspired proteoglycan 4 (prg4)-grafted layer. *ACS Applied Materials Interfaces* **2018**, *10* (36), 30125.
- (180) Pult, H.; Murphy, P. J.; Purslow, C., A novel method to predict the dry eye symptoms in new contact lens wearers. *Optometry and Vision Science* **2009**, *86* (9), E1042.
- (181) Rolando, M.; Zierhut, M., The ocular surface and tear film and their dysfunction in dry eye disease. *Survey of Ophthalmology* **2001**, *45*, S203.
- (182) Doane, M. G., Interactions of eyelids and tears in corneal wetting and the dynamics of the normal human eyeblink. *American Journal of Ophthalmology* **1980**, *89* (4), 507.
- (183) Wieleba, W., The statistical correlation of the coefficient of friction and wear rate of ptfе composites with steel counterface roughness and hardness. *Wear* **2002**, *252*, 719.
- (184) Rahaman, M. L.; Zhang, L.; Liu, M.; Liu, W., Surface roughness effect on the friction and wear of bulk metallic glasses. *Wear* **2015**, *332*, 1231.
- (185) Coles, J. M.; Zhang, L.; Blum, J. J.; Warman, M. L.; Jay, G. D.; Guilak, F.; Zauscher, S., Loss of cartilage structure, stiffness, and frictional properties in mice lacking prg4. *Arthritis & Rheumatology* **2010**, *62* (6), 1666.
- (186) Kienle, S.; Boettcher, K.; Wiegler, L.; Urban, J.; Burgkart, R.; Lieleg, O.; Hugel, T., Comparison of friction and wear of articular cartilage on different length scales. *Journal of Biomechanics* **2015**, *48* (12), 3052.
- (187) Dunn, A. C.; Cobb, J. A.; Kantzios, A. N.; Lee, S. J.; Sarntinoranont, M.; Tran-Son-Tay, R.; Sawyer, W. G., Friction coefficient measurement of hydrogel materials on living epithelial cells. *Tribology Letters* **2008**, *30* (1), 13.
- (188) Alipour, F.; Khaheshi, S.; Soleimanzadeh, M.; Heidarzadeh, S.; Heydarzadeh, S., Contact lens-related complications: A review. *Journal of Ophthalmic & Vision Research* **2017**, *12* (2), 193.
- (189) Zhang, X.-H.; Li, X., Effect of rigid gas permeable contact lens on keratoconus progression: A review. *International Journal of Ophthalmology* **2020**, *13* (7), 1124.
- (190) Saraç, Ö.; Kars, M. E.; Temel, B.; Çağıl, N., Clinical evaluation of different types of contact lenses in keratoconus management. *Contact Lens and Anterior Eye* **2019**, *42* (5), 482.

- (191) Huang, J., A simple accurate formula for calculating saturation vapor pressure of water and ice. *Journal of Applied Meteorology and Climatology* **2018**, 57 (6), 1265.
- (192) Arkles, B., Tailoring surfaces with silanes. *Chemtech* **1977**, 7, 766.
- (193) Plueddemann, E. P. Nature of adhesion through silane coupling agents. In *Silane coupling agents*, Springer, **1991**, 115.
- (194) Meldrum, O. W.; Yakubov, G. E.; Bonilla, M. R.; Deshmukh, O.; McGuckin, M. A.; Gidley, M. J., Mucin gel assembly is controlled by a collective action of non-mucin proteins, disulfide bridges, ca<sup>2+</sup>-mediated links, and hydrogen bonding. *Scientific Reports* **2018**, 8 (1), 5802.
- (195) Ainavarapu, S. R. K.; Brujić, J.; Huang, H. H.; Wiita, A. P.; Lu, H.; Li, L.; Walther, K. A.; Carrion-Vazquez, M.; Li, H.; Fernandez, J. M., Contour length and refolding rate of a small protein controlled by engineered disulfide bonds. *Biophysical Journal* **2007**, 92 (1), 225.
- (196) Dhallu, S. K.; Huarte, S. T.; Bilkhu, P. S.; Boychev, N.; Wolffsohn, J. S., Effect of scleral lens oxygen permeability on corneal physiology. *Optometry and Vision Science* **2020**, 97 (9), 669.
- (197) Liesegang, T. J., Physiologic changes of the cornea with contact lens wear. *Eye & Contact Lens* **2002**, 28 (1), 12.
- (198) Ren, D. H.; Petroll, W. M.; Jester, J. V.; Ho-Fan, J.; Cavanagh, H. D., The relationship between contact lens oxygen permeability and binding of pseudomonas aeruginosato human corneal epithelial cells after overnight and extended wear. *the CLAO Journal* **1999**, 25 (2), 81.
- (199) Harris, M. G.; Chu, C. S., The effect of contact lens thickness and corneal toricity on flexure and residual astigmatism. *Optometry and Vision Science* **1972**, 49 (4), 304.
- (200) Lin, M. C.; Snyder, C., Flexure and residual astigmatism with rgp lenses of low, medium, and high oxygen permeability. *International Contact Lens Clinic* **1999**, 26 (1), 5.
- (201) Tsai, H.-Y.; Hsieh, Y.-C.; Lin, Y.-H.; Chang, H.-C.; Tang, Y.-H.; Huang, K.-C., Fabrication of hydrophilic surface on rigid gas permeable contact lenses to enhance the wettability using ultraviolet laser system. *Micromachines* **2019**, 10 (6), 394.
- (202) Fonn, D., Targeting contact lens induced dryness and discomfort: What properties will make lenses more comfortable. *Optometry and Vision Science* **2007**, 84 (4), 279.
- (203) Bauer, M. G.; Reithmeir, R.; Lutz, T. M.; Lileg, O., Wetting behavior and stability of surface-modified polyurethane materials. *Plasma Processes and Polymers* **2021**, 18 (11), 2100126.
- (204) Kimna, C.; Winkeljann, B.; Song, J.; Lileg, O., Smart biopolymer-based multi-layers enable consecutive drug release events on demand. *Advanced Materials Interfaces* **2020**, 7 (19), 2000735.
- (205) Xu, X.; Jin, T.; Zhang, B.; Liu, H.; Ye, Z.; Xu, Q.; Chen, H.; Wang, B., In vitro and in vivo evaluation of the antibacterial properties of a nisin-grafted hydrated mucin multilayer film. *Polymer Testing* **2017**, 57, 270.
- (206) UniProtKB. *Muc5a\_human accession number: P98088*. <https://www.uniprot.org/uniprot/P98088> (accessed 10.10.2022).
- (207) Parry, S.; Hanisch, F. G.; Leir, S.-H.; Sutton-Smith, M.; Morris, H. R.; Dell, A.; Harris, A., N-glycosylation of the muc1 mucin in epithelial cells and secretions. *Glycobiology* **2006**, 16 (7), 623.

## Bibliography

- (208) Hang, H. C.; Bertozzi, C. R., The chemistry and biology of mucin-type o-linked glycosylation. *Bioorganic & Medicinal Chemistry* **2005**, *13* (17), 5021.
- (209) Perez-Vilar, J.; Randell, S. H.; Boucher, R. C., C-mannosylation of muc5ac and muc5b cys subdomains. *Glycobiology* **2004**, *14* (4), 325.
- (210) Ray, S.; Cooney, R. P., Thermal degradation of polymer and polymer composites. In *Handbook of environmental degradation of materials*, **2018**, Elsevier, 185.
- (211) Khoroshilova, E. V.; Repeyev, Y. A.; Nikogosyan, D. N., Uv photolysis of aromatic amino acids and related dipeptides and tripeptides. *Journal of Photochemistry and Photobiology B: Biology* **1990**, *7* (2-4), 159.
- (212) Reisz, J. A.; Bansal, N.; Qian, J.; Zhao, W.; Furdui, C. M., Effects of ionizing radiation on biological molecules—mechanisms of damage and emerging methods of detection. *Antioxidants & Redox Signaling* **2014**, *21* (2), 260.
- (213) Funatsu, K.; Kiminami, H.; Abe, Y.; Carpenter, J. F., Impact of ethylene oxide sterilization of polymer-based prefilled syringes on chemical degradation of a model therapeutic protein during storage. *Journal of Pharmaceutical Sciences* **2019**, *108* (1), 770-774.
- (214) Chen, L.; Sloey, C.; Zhang, Z.; Bondarenko, P. V.; Kim, H., Chemical modifications of therapeutic proteins induced by residual ethylene oxide. *Journal of Pharmaceutical Sciences* **2015**, *104* (2), 731.
- (215) da Cunha Mendes, G. C.; da Silva Brandao, T. R.; Miranda Silva, C. L., Ethylene oxide potential toxicity. *Expert Review of Medical Devices* **2008**, *5* (3), 323.
- (216) Xiong, Y. L. Protein denaturation and functionality losses. In *Quality in frozen food*, Springer, **1997**, 111.
- (217) Wagner, J.; Anon, M., Effect of frozen storage on protein denaturation in bovine muscle. II. Influence on solubility, viscosity and electrophoretic behaviour of myofibrillar proteins. *International Journal of Food Science & Technology* **1986**, *21* (5), 547.
- (218) Bara, J.; Chastre, E.; Mahiou, J.; Singh, R. L.; Forgue-Lafitte, M. E.; Hollande, E.; Godeau, F., Gastric m1 mucin, an early oncofetal marker of colon carcinogenesis, is encoded by the muc5ac gene. *International Journal of Cancer* **1998**, *75* (5), 767.
- (219) Aumailley, M.; Timpl, R.; Risau, W., Differences in laminin fragment interactions of normal and transformed endothelial cells. *Experimental Cell Research* **1991**, *196* (2), 177.
- (220) Biegler, M.; Delius, J.; Käs Dorf, B. T.; Hofmann, T.; Lieleg, O., Cationic astringents alter the tribological and rheological properties of human saliva and salivary mucin solutions. *Biotribology* **2016**, *6*, 12.
- (221) Mall, A. S.; Dent, D. M.; McLeod, H.; Kahn, D.; Taylor, K.; Hickman, R., Extraction, isolation, and sds-page analysis of purified gastric mucin in a patient with menetrier's disease. *American Journal of Gastroenterology* **2002**, *97* (3), 752.
- (222) Gibbins, H. L.; Proctor, G. B.; Yakubov, G. E.; Wilson, S.; Carpenter, G. H., Siga binding to mucosal surfaces is mediated by mucin-mucin interactions. *PLoS One* **2015**, *10* (3), 1.
- (223) Dunn, A. C.; Tichy, J. A.; Urueña, J. M.; Sawyer, W. G., Lubrication regimes in contact lens wear during a blink. *Tribology International* **2013**, *63*, 45.

- (224) Kesel, S.; von Bronk, B.; García, C. F.; Götz, A.; Lieleg, O.; Opitz, M., Matrix composition determines the dimensions of bacillus subtilis ncib 3610 biofilm colonies grown on lb agar. *RSC Advances* **2017**, 7 (51), 31886.
- (225) Werb, M.; García, C. F.; Bach, N. C.; Grumbein, S.; Sieber, S. A.; Opitz, M.; Lieleg, O., Surface topology affects wetting behavior of bacillus subtilis biofilms. *npj Biofilms and Microbiomes* **2017**, 3 (1), 1.
- (226) Shim, M. H.; Kim, J.; Park, C. H., The effects of surface energy and roughness on the hydrophobicity of woven fabrics. *Textile Research Journal* **2014**, 84 (12), 1268.
- (227) Erbil, H. Y.; Demirel, A. L.; Avci, Y.; Mert, O., Transformation of a simple plastic into a superhydrophobic surface. *Science* **2003**, 299 (5611), 1377.
- (228) Menze, B. H.; Kelm, B. M.; Masuch, R.; Himmelreich, U.; Bachert, P.; Petrich, W.; Hamprecht, F. A., A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* **2009**, 10 (1), 1.
- (229) Noro, A.; Kaneko, M.; Murata, I.; Yoshinari, M., Influence of surface topography and surface physicochemistry on wettability of zirconia (tetragonal zirconia polycrystal). *Journal of Biomedical Materials Research Part B: Applied Biomaterials* **2013**, 101 (2), 355.
- (230) Hosseinabadi, H. N.; Sajjady, S.; Amini, S., Creating micro textured surfaces for the improvement of surface wettability through ultrasonic vibration assisted turning. *The International Journal of Advanced Manufacturing Technology* **2018**, 96 (5), 2825.
- (231) Melentiev, R.; Fang, F.; Narala, S. K. R., Influence of different pretreatments on ti-6al-4v surface integrity and scratch-resistance of epoxy coating: Analysis of topography, microstructure, chemistry and wettability. *Surface and Coatings Technology* **2020**, 404, 126436.
- (232) García, C. F.; Stangl, F.; Götz, A.; Zhao, W.; Sieber, S. A.; Opitz, M.; Lieleg, O., Topographical alterations render bacterial biofilms susceptible to chemical and mechanical stress. *Biomaterials Science* **2019**, 7 (1), 220.
- (233) Hayta, E. N.; Lieleg, O., Biopolymer-enriched b. Subtilis ncib 3610 biofilms exhibit increased erosion resistance. *Biomaterials Science* **2019**, 7 (11), 4675.
- (234) Bhushan, B.; Nosonovsky, M., The rose petal effect and the modes of superhydrophobicity. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2010**, 368 (1929), 4713.
- (235) Sheng, Y.-J.; Jiang, S.; Tsao, H.-K., Effects of geometrical characteristics of surface roughness on droplet wetting. *The Journal of Chemical Physics* **2007**, 127 (23), 234704.
- (236) Kang, S. M., Role of wide tip of mushroom-like micropillar arrays to make the cassie state on superrepellent surfaces. *RSC Advances* **2016**, 6 (78), 74670.
- (237) Jain, A.; Patel, H.; Nagalapatti, L.; Gupta, N.; Mehta, S.; Guttula, S.; Mujumdar, S.; Afzal, S.; Sharma Mittal, R.; Munigala, V., Overview and importance of data quality for machine learning tasks. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* **2020**, 3561.

## Bibliography

- (238) Obaid, H. S.; Dheyab, S. A.; Sabry, S. S., The impact of data pre-processing techniques and dimensionality reduction on the accuracy of machine learning. *9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*, **2019**, IEEE, 279.
- (239) Aggarwal, C. C.; Hinneburg, A.; Keim, D. A., On the surprising behavior of distance metrics in high dimensional space. in *International Conference on Database Theory*, **2001**, Springer, 420.
- (240) Liu, H.; Motoda, H., Feature selection for knowledge discovery and data mining; Springer Science & Business Media, **2012**.
- (241) Liu, H.; Motoda, H. Computational methods of feature selection, **2007**, CRC Press,.
- (242) Beniwal, S.; Arora, J., Classification and feature selection techniques in data mining. *International Journal of Engineering Research & Technology* **2012**, 1 (6), 1.
- (243) Stańczyk, U., Feature evaluation by filter, wrapper, and embedded approaches. in *Feature selection for data and pattern recognition*, **2015**, Springer, 29.
- (244) El Aboudi, N.; Benhlima, L., Review on wrapper feature selection approaches. *2016 International Conference on Engineering & MIS (ICEMIS)*, **2016**, IEEE, 1.
- (245) Zebari, R.; Abdulazeez, A.; Zeebaree, D.; Zebari, D.; Saeed, J., A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends* **2020**, 1 (2), 56.
- (246) Huang, X.; Wu, L.; Ye, Y., A review on dimensionality reduction techniques. *International Journal of Pattern Recognition and Artificial Intelligence* **2019**, 33 (10), 1950017.
- (247) Rückstieß, T.; Osendorfer, C.; Smagt, P. v. d., Sequential feature selection for classification. in *Australasian Joint Conference on Artificial Intelligence*, **2011**, Springer, 132.
- (248) Hira, Z. M.; Gillies, D. F., A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics* **2015**, 2015.
- (249) Loughrey, J.; Cunningham, P., Overfitting in wrapper-based feature subset selection: The harder you try the worse it gets. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, **2004**, Springer, 33.
- (250) Wang, S.; Tang, J.; Liu, H., Embedded unsupervised feature selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, **2015**, 29.
- (251) Maldonado, S.; López, J., Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for svm classification. *Applied Soft Computing* **2018**, 67, 94.
- (252) Parmar, A.; Katariya, R.; Patel, V., A review on random forest: An ensemble classifier. in *International Conference on Intelligent Data Communication Technologies and Internet of Things*, **2018**, Springer, 758.
- (253) Fawagreh, K.; Gaber, M. M.; Elyan, E., Random forests: From early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal* **2014**, 2 (1), 602.
- (254) Ranstam, J.; Cook, J., Lasso regression. *Journal of British Surgery* **2018**, 105 (10), 1348.
- (255) McDonald, G. C., Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics* **2009**, 1 (1), 93.

- (256) Sánchez-Marroño, N.; Alonso-Betanzos, A.; Tombilla-Sanromán, M. Filter methods for feature selection—a comparative study. in *International Conference on Intelligent Data Engineering and Automated Learning*, **2007**, Springer, 178.
- (257) Vergara, J. R.; Estévez, P. A., A review of feature selection methods based on mutual information. *Neural Computing and Applications* **2014**, *24* (1), 175.
- (258) Estévez, P. A.; Tesmer, M.; Perez, C. A.; Zurada, J. M., Normalized mutual information feature selection. *IEEE Transactions on Neural Networks* **2009**, *20* (2), 189.
- (259) Talavera, L., Dependency-based feature selection for clustering symbolic data. *Intelligent Data Analysis* **2000**, *4* (1), 19.
- (260) Barbiero, P.; Squillero, G.; Tonda, A., Predictable features elimination: An unsupervised approach to feature selection. in *International Conference on Machine Learning, Optimization, and Data Science*, **2021**, Springer, 399.
- (261) Dash, M.; Choi, K.; Scheuermann, P.; Liu, H. Feature selection for clustering—a filter solution. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, **2002**, IEEE, 115.
- (262) He, X.; Cai, D.; Niyogi, P., Laplacian score for feature selection. *Advances in Neural Information Processing Systems* **2005**, 18.
- (263) Abdi, H.; Williams, L. J., Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* **2010**, *2* (4), 433.
- (264) Salas-Gonzalez, D.; Górriz, J.; Ramírez, J.; Illán, I.; López, M.; Segovia, F.; Chaves, R.; Padilla, P.; Puntonet, C.; Initiative, A. s. D. N., Feature selection using factor analysis for alzheimer's diagnosis using pet images. *Medical Physics* **2010**, *37* (11), 6084.
- (265) Jimenez, L.; Landgrebe, D. A. Projection pursuit in high dimensional data reduction: Initial conditions, feature selection and the assumption of normality. in *1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, **1995**, IEEE, 1, 401.
- (266) Jimenez, L. O.; Landgrebe, D. A., Hyperspectral data analysis and supervised feature reduction via projection pursuit. *IEEE Transactions on Geoscience and Remote Sensing* **1999**, *37* (6), 2653.
- (267) Mazlum, N.; MAZLUM, A. Ö. S., Interpretation of water quality data by principal components analysis. *Turkish Journal of Engineering and Environmental Sciences* **1999**, *23* (1), 19.
- (268) Zou, H.; Xue, L., A selective overview of sparse principal component analysis. *Proceedings of the IEEE* **2018**, *106* (8), 1311.
- (269) Craney, T. A.; Surles, J. G., Model-dependent variance inflation factor cutoff values. *Quality Engineering* **2002**, *14* (3), 391.
- (270) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V., Moleculenet: A benchmark for molecular machine learning. *Chemical Science* **2018**, *9* (2), 513.
- (271) Cui, M., Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance* **2020**, *1* (1), 5.
- (272) Köppen, M. The curse of dimensionality. in *5th online world conference on soft computing in industrial applications (WSC5)*, **2000**, 1, 4.

## Bibliography

- (273) Ghadiri, M.; Chrzanowski, W.; Lee, W.; Rohanizadeh, R., Layered silicate clay functionalized with amino acids: Wound healing application. *RSC Advances* **2014**, *4* (67), 35332.
- (274) Nowald, C.; Penk, A.; Chiu, H.-Y.; Bein, T.; Huster, D.; Lileg, O., A selective mucin/methylcellulose hybrid gel with tailored mechanical properties. *Macromolecular Bioscience* **2016**, *16* (4), 567.
- (275) Marczyński, M.; Kimna, C.; Lileg, O., Purified mucins in drug delivery research. *Advanced Drug Delivery Reviews* **2021**, *178*, 113845.
- (276) Kimna, C.; Lutz, T. M.; Yan, H.; Song, J.; Crouzier, T.; Lileg, O., DNA strands trigger the intracellular release of drugs from mucin-based nanocarriers. *ACS Nano* **2020**, *15* (2), 2350.
- (277) Marczyński, M.; Käs Dorf, B. T.; Altaner, B.; Wenzler, A.; Gerland, U.; Lileg, O., Transient binding promotes molecule penetration into mucin hydrogels by enhancing molecular partitioning. *Biomaterials Science* **2018**, *6* (12), 3373.
- (278) Wheeler, K. M.; Cárcamo-Oyarce, G.; Turner, B. S.; Dellos-Nolan, S.; Co, J. Y.; Lehoux, S.; Cummings, R. D.; Wozniak, D. J.; Ribbeck, K., Mucin glycans attenuate the virulence of pseudomonas aeruginosa in infection. *Nature Microbiology* **2019**, *4* (12), 2146.
- (279) Marczyński, M.; Winkeljann, B.; Lileg, O., Advances in mucin biopolymer research: Purification, characterization, and applications. *Biopolymers for Biomedical and Biotechnological Applications* **2021**, 181.
- (280) Jiang, K.; Yan, H.; Rickert, C.; Marczyński, M.; Sixtensson, K.; Vilaplana, F.; Lileg, O.; Crouzier, T., Modulating the bioactivity of mucin hydrogels with crosslinking architecture. *Advanced Functional Materials* **2021**, *31* (10), 2008428.
- (281) Balabushevich, N. G.; Kovalenko, E. A.; Filatova, L. Y.; Kirzhanova, E. A.; Mikhailchik, E. V.; Volodkin, D.; Vikulina, A. S., Hybrid mucin-vaterite microspheres for delivery of proteolytic enzyme chymotrypsin. *Macromolecular Bioscience* **2022**, *22* (7), 2200005.
- (282) Kwan, C.-S.; Cerullo, A. R.; Braunschweig, A. B., Design and synthesis of mucin-inspired glycopolymers. *ChemPlusChem* **2020**, *85* (12), 2704.
- (283) Canton, I.; Warren, N. J.; Chahal, A.; Amps, K.; Wood, A.; Weightman, R.; Wang, E.; Moore, H.; Armes, S. P., Mucin-inspired thermoresponsive synthetic hydrogels induce stasis in human pluripotent stem cells and human embryos. *ACS Central Science* **2016**, *2* (2), 65.
- (284) Wallert, M.; Nie, C.; Anilkumar, P.; Abbina, S.; Bhatia, S.; Ludwig, K.; Kizhakkedathu, J. N.; Haag, R.; Block, S., Mucin-inspired, high molecular weight virus binding inhibitors show biphasic binding behavior to influenza A viruses. *Small* **2020**, *16* (47), 2004635.
- (285) Kruger, A. G.; Brucks, S. D.; Yan, T.; Cárcamo-Oyarce, G.; Wei, Y.; Wen, D. H.; Carvalho, D. R.; Hore, M. J. A.; Ribbeck, K.; Schrock, R. R.; et al., Stereochemical control yields mucin mimetic polymers. *ACS Central Science* **2021**, *7* (4), 624.
- (286) Sharma, A.; Thongrom, B.; Bhatia, S.; von Lospichl, B.; Addante, A.; Graeber, S. Y.; Lauster, D.; Mall, M. A.; Gradzielski, M.; Haag, R., Polyglycerol-based mucus-inspired hydrogels. *Macromolecular Rapid Communications* **2021**, 2100303.
- (287) Del Ser, J.; Osaba, E.; Molina, D.; Yang, X.-S.; Salcedo-Sanz, S.; Camacho, D.; Das, S.; Suganthan, P. N.; Coello, C. A. C.; Herrera, F., Bio-inspired computation: Where we stand and what's next. *Swarm and Evolutionary Computation* **2019**, *48*, 220-250.

- (288) Kar, A. K., Bio inspired computing—a review of algorithms and scope of applications. *Expert Systems with Applications* **2016**, *59*, 20.
- (289) Deng, W.; Xu, J.; Zhao, H., An improved ant colony optimization algorithm based on hybrid strategies for scheduling problem. *IEEE access* **2019**, *7*, 20281.
- (290) Dorigo, M.; Stützle, T. *Ant colony optimization: Overview and recent advances*; Springer, **2019**.
- (291) Kim, J.; Chaudhury, M. K.; Owen, M. J., Hydrophobic recovery of polydimethylsiloxane elastomer exposed to partial electrical discharge. *Journal of Colloid and Interface Science* **2000**, *226* (2), 231.
- (292) Yan, H.; Chircov, C.; Zhong, X.; Winkeljann, B.; Dobryden, I.; Nilsson, H. E.; Lieleg, O.; Claesson, P. M.; Hedberg, Y.; Crouzier, T., Reversible condensation of mucins into nanoparticles. *Langmuir* **2018**, *34* (45), 13615.
- (293) Hertz, H., Über die berührung fester elastischer körper. *Journal für die reine und angewandte Mathematik* **1882**, *92*, 156.
- (294) Johnston, I.; McCluskey, D.; Tan, C.; Tracey, M., Mechanical characterization of bulk sylgard 184 for microfluidics and microengineering. *Journal of Micromechanical Microengineering* **2014**, *24* (3), 035017.
- (295) Pierscionek, B. K.; Asejczyk-Widlicka, M.; Schachar, R. A., The effect of changing intraocular pressure on the corneal and scleral curvatures in the fresh porcine eye. *British Journal of Ophthalmology* **2007**, *91* (6), 801.
- (296) Kesel, S.; Grumbein, S.; Gümperlein, I.; Tallawi, M.; Marel, A.-K.; Lieleg, O.; Opitz, M., Direct comparison of physical properties of bacillus subtilis ncib 3610 and B-1 biofilms. *Applied and Environmental Microbiology* **2016**, *82* (8), 2424.
- (297) Raeder, T.; Forman, G.; Chawla, N. V. Learning from imbalanced data: Evaluation matters. in *Data mining: Foundations and Intelligent Paradigms*, **2012**, Springer, 315.
- (298) Simonyan, K.; Zisserman, A., Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
- (299) Ghumman, U. F.; Chen, Q.; D'Angelo, V. E.; Clark, M.; Chen, J.; Shull, K. R.; Chen, W., Crack surface analysis of elastomers using transfer learning. *ACS Applied Materials & Interfaces* **2023**.
- (300) Chollet, F., *Keras*, <https://keras.io>, **2015**.
- (301) He, K.; Zhang, X.; Ren, S.; Sun, J., Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, **2016**, 770.
- (302) Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Q., Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, **2017**, 4700.
- (303) Masters, J. R., Hela cells 50 years on: The good, the bad and the ugly. *Nature Reviews Cancer* **2002**, *2* (4), 315.



## Author contributions

This thesis follows a total number of 8 research articles and 1 review article to which I contributed during my time as doctoral candidate. The publications on which the respective chapters of this dissertation are based are marked accordingly by footnotes. My individual contribution to these publications is listed below:

- 1) **Rickert, C. A.**; Wittmann, B.; Fromme, R.; Lieleg, O., Highly transparent covalent mucin coatings improve the wettability and tribology of hydrophobic contact lenses. *ACS Applied Materials & Interfaces* **2020**, *12* (25), 28024-28033.

**Individual contribution:** I contributed to the conception of this study, the design and performance of the experiments, the data analysis, and writing of the manuscript.

- 2) **Rickert, C. A.**; Lutz, T. M.; Marczynski, M.; Lieleg, O., Several sterilization strategies maintain the functionality of mucin glycoproteins. *Macromolecular Bioscience* **2020**, *20* (7), 2000090.

**Individual contribution:** I contributed to the conception of this study, the design and performance of the experiments, the data analysis, and writing of the manuscript.

- 3) **Rickert, C. A.**; Hayta, E. N.; Selle, D. M.; Kouroudis, I.; Harth, M.; Gagliardi, A.; Lieleg, O., Machine learning approach to analyze the surface properties of biological materials. *ACS Biomaterials Science & Engineering* **2021**, *7* (9), 4614-4625.

**Individual contribution:** I contributed to the conception of this study, the design and performance of the experiments, the data analysis, and writing of the manuscript.

- 4) **Rickert, C. A.**; Bauer, M. G.; Hoffmeister, J. C.; Lieleg, O., Effects of sterilization methods on the integrity and functionality of covalent mucin coatings on medical devices. *Advanced Materials Interfaces* **2021**, *9* (3), 2101716.

**Individual contribution:** I contributed to the conception of this study, the design and performance of the experiments, the data analysis, and writing of the manuscript.

- 5) **Rickert, C. A.**; Lieleg, O., Machine learning approaches for biomolecular, biophysical, and biomaterials research. *Biophysics Reviews* **2022**, *3* (2), 021306.

**Individual contribution:** I contributed to the conception, the literature review and writing of the manuscript.

- 6) **Rickert, C. A.**; Piller, I.; Henkel, F.; Fromme, R.; Lieleg, O., Multifunctional glycoprotein coatings improve the surface properties of highly oxygen permeable contact lenses. *Biomaterials Advances* **2022**, *145*, 213233.

**Individual contribution:** I contributed to the conception of this study, the design and performance of the experiments, the data analysis, and writing of the manuscript.

- 7) **Rickert, C. A.;** Henkel, M.; Lieleg, O., An efficiency-driven, correlation-based feature elimination strategy for small datasets. *APL Machine Learning* **2023**, *1* (1), 016105.

**Individual contribution:** I contributed to the conception of this study, the development of the algorithm, the design and performance of the computational experiments, the data analysis, and writing of the manuscript.

- 8) **Rickert, C. A.;** Mansi, S.; Fan, D.; Mela, P.; Lieleg, O., A mucin-based bio-ink for 3D printing of objects with anti-biofouling properties. *Advanced Materials Technologies*, *submitted*

**Individual contribution:** I contributed to the conception of this study, the design and performance of the experiments, the data analysis, and writing of the manuscript.

- 9) **Rickert, C. A.;** Henkel, F.; Lieleg, O., Unsupervised Machine Learning to Topographically Analyze Corneal Tissue Surfaces. *APL Machine Learning*, *submitted*

**Individual contribution:** I contributed to the conception of this study, the implementation of the algorithms, the design and performance of the computational experiments, the data analysis, and writing of the manuscript.

## Full list of publications

- 1) Marczynski, M.; **Rickert, C. A.**; Semerdzhiev, S.; van Dijk, W.; Segers-Nolten, I.; Claessens, M. M. A. E.; Lieleg, O.,  $\alpha$ -Synuclein penetrates mucin hydrogels despite its mucoadhesive properties. *Biomacromolecules* **2019**, *20* (12), 4332-4344.
- 2) **Rickert, C. A.**, Wittmann, B.; Fromme, R.; Lieleg, O., Highly transparent covalent mucin coatings improve the wettability and tribology of hydrophobic contact lenses. *ACS Applied Materials & Interfaces* **2020**, *12* (25), 28024-28033.
- 3) **Rickert, C. A.**; Lutz, T. M.; Marczynski, M.; Lieleg, O., Several sterilization strategies maintain the functionality of mucin glycoproteins. *Macromolecular Bioscience* **2020**, *20* (7), 2000090.
- 4) Jiang, K.; Yan, H.; **Rickert, C. A.**; Marczynski, M.; Sixtensson, K.; Vilaplana, F.; Lieleg, O.; Crouzier, T., Modulating the bioactivity of mucin hydrogels with crosslinking architecture. *Advanced Functional Materials* **2021**, *31* (10), 2008428.
- 5) Hayta, E. N.; **Rickert, C. A.**; Lieleg, O., Topography quantifications allow for identifying the contribution of parental strains to physical properties of co-cultured biofilms. *Biofilm* **2021**, *3*, 100044.
- 6) **Rickert, C. A.**; Hayta, E. N.; Selle, D. M.; Kouroudis, I.; Harth, M.; Gagliardi, A.; Lieleg, O., Machine learning approach to analyze the surface properties of biological materials. *ACS Biomaterials Science & Engineering* **2021**, *7* (9), 4614-4625.
- 7) **Rickert, C. A.**; Bauer, M. G.; Hoffmeister, J. C.; Lieleg, O., Effects of sterilization methods on the integrity and functionality of covalent mucin coatings on medical devices. *Advanced Materials Interfaces* **2021**, *9* (3), 2101716.
- 8) Marczynski, M.; **Rickert, C. A.**; Fuhrmann, T.; Lieleg, O., An improved, filtration-based process to purify functional mucins from mucosal tissues with high yields. *Separation and Purification Technology* **2022**, *294*, 121209.
- 9) **Rickert, C. A.**; Lieleg, O., Machine learning approaches for biomolecular, biophysical, and biomaterials research. *Biophysics Reviews* **2022**, *3* (2), 021306.
- 10) **Rickert, C. A.**; Piller, I.; Henkel, F.; Fromme, R.; Lieleg, O., Multifunctional glycoprotein coatings improve the surface properties of highly oxygen permeable contact lenses. *Biomaterials Advances* **2022**, *145*, 213233.
- 11) **Rickert, C. A.**; Henkel, M.; Lieleg, O., An efficiency-driven, correlation-based feature elimination strategy for small datasets. *APL Machine Learning* **2023**, *1* (1), 016105.

- 12) **Rickert, C. A.**; Mansi, S.; Fan, D.; Mela, P.; Lieleg, O., A mucin-based bio-ink for 3D printing of objects with anti-biofouling properties. *Advanced Materials Technologies*, submitted
- 13) **Rickert, C. A.**; Henkel, F.; Lieleg, O., Unsupervised Machine Learning to Topographically Analyze Corneal Tissue Surfaces. *APL Machine Learning*, submitted
- 14) Mansi, S.; Dummert, S.; Topping, G. J.; Hussain, M. Z.; **Rickert, C. A.**; Müller, K. M. A.; Schilling, F.; Fischer, R. A.; Lieleg, O.; Mela, P., Introducing metal-organic frameworks to melt electrowriting: multifunctional scaffolds with controlled microarchitecture for tissue engineering applications. *Advanced Functional Materials*, submitted

## Acknowledgment

An dieser Stelle möchte ich allen danken, die auf unterschiedliche Weise zum Gelingen dieser Doktorarbeit beigetragen haben.

Zunächst möchte ich meinem Doktorvater, Prof. Dr. Oliver Lieleg, für die Möglichkeit danken, meine Promotion in seiner Arbeitsgruppe und unter seiner ausführlichen Betreuung durchführen zu dürfen. Vielen Dank für viele lehrreiche Gespräche und fachliche Diskussionen, für Deine immer offene Tür, und für Deine unnachahmliche Unterstützung bei der Umsetzung meiner Projekte von der Konzeptionalisierung bis zur Veröffentlichung. Besonders danke ich Dir dafür, dass ich in meiner Zeit als Doktorandin ein breites Spektrum an Themen verschiedener wissenschaftlicher Bereiche bearbeiten und meinen eigenen Weg finden durfte, auch wenn uns dieser teilweise in sehr fremde Gewässer geführt hat. Von Herzen vielen Dank, dass Du mich durch diese wichtige Phase meiner Karriere begleitet hast.

Darüber hinaus danke ich den Mitgliedern der Prüfungskommission für die Übernahme ihrer jeweiligen Aufgaben als Prüfungsvorsitz und Zweitprüfer.

Ich bedanke mich außerdem herzlich bei allen aktuellen und ehemaligen Mitgliedern der Professur für Biomechanik für ihre Kollegialität und gute Zusammenarbeit. Insbesondere gilt mein Dank Dr. Matthias ‚Matsi‘ Marczynski für gemeinsame Teepausen, den ein oder anderen gemeinsamen Prosecco, das Übernehmen meiner Kuchenpflichten, unzählige Ratschläge und seine Aktivität als Care-Bear. Ich könnte nicht dankbarer sein, dich an meiner Seite gehabt zu haben. Ein großer Dank gilt auch Dr. Marvin ‚MG‘ Ertelt für die vielen Getränkefahrten, Lindor- und Kinderriegel-Pakete, deine unnachahmliche Hilfsbereitschaft und natürlich das regelmäßige Beliefern mit Pizza oder indischem Essen. Ich danke Bernardo ‚Berny‘ Miller Naranjo für unterhaltsame sprachliche Kuriositäten, gemeinsame Spaziergänge, die zuverlässige Versorgung mit Schokoladen-Eiern und -Riegeln und zahlreiche nicht-Dates. Ich danke Ceren Kımna für gemeinsames Fachsimpeln, die Einführung in die Grundlagen türkischer Schimpfwörter (in diesem Sinne auch einen herzlichen Dank an Ufuk Gürer) und ihre Unterstützung bei *drug release*-bezogenen Problemstellungen. Vielen Dank an Iris König-Decker und Tobias ‚Tofu‘ Fuhrmann für Unterstützung in administrativen und technischen Angelegenheiten. Ein großer Dank geht hierbei auch an alle meine Studenten: Babsi, Fabio, Julia, Daniel, Manuel und Silvia. Ich durfte bei der Zusammenarbeit mit euch viel lernen, und es ist toll, einige von euch als neue Doktoranden- und IT Support-Generation nachrücken zu sehen.

Darüber hinaus möchte ich von ganzem Herzen meinem emotionalen Support danken – meinem Freund Marcel. Deine Unterstützung während meines gesamten Studiums und während meiner Promotion hat wesentlich zum Erfolg dieser Zeit beigetragen. Es ist schwer in Worte zu fassen, wie wichtig deine Motivation, deine Geduld und dein unerschütterliches Vertrauen in mich sind und wie dankbar ich bin, dass du diesen Weg mit mir gegangen bist.

Zuletzt gilt ein ganz besonderer Dank meinem Bruder Till und meinen Eltern, die mir all‘ dies ermöglicht haben. Von Herzen danke dafür, dass ihr immer an mich glaubt, auch dann, wenn ich es selbst nicht mehr tue, dafür dass ihr mir immer Mut macht, jede Herausforderung anzunehmen, und dafür, dass ich mich immer auf euch verlassen kann. Ihr seid meine Vorbilder, meine Inspiration und der Wind in meinen Segeln und ich weiß, dass ich nichts von allem ohne eure Unterstützung geschafft hätte.