https://doi.org/10.15388/vu.thesis.367
https://orcid.org/0000-0001-8126-3459

VILNIUS UNIVERSITY

Tomas Raila

Computer modeling methods for phospholipid membrane damage assessment

DOCTORAL DISSERTATION

Natural Sciences,
Informatics (N 009)

VILNIUS 2022

This doctoral dissertation was written between 2017 and 2021 at Vilnius University. The research was partially funded by the Research Council of Lithuania, project "Quantitative assessment of the membrane damage by the pore-forming toxins" (P-MIP-19-394).

Academic supervisor - **Prof. Dr. Tadas Meškauskas** (Vilnius University, Natural Sciences, Informatics - N 009).

This doctoral dissertation will be defended in a public meeting of the Dissertation Defence Panel:

Chairman - **Prof. Dr. Olga Kurasova** (Vilnius University, Natural Sciences, Informatics - N 009).
Members:
Dr. Jolita Bernatavičienė (Vilnius University, Natural Sciences, Informatics - N 009),
Prof. Dr. Stefano Bonnini (University of Ferrara, Italy, Natural Sciences, Informatics - N 009),
Dr. Rima Budvytytė (Vilnius University, Natural Sciences, Biochemistry - N 004),
Prof. Dr. Arnas Kačeniauskas (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering - T 007).

The dissertation shall be defended at a public meeting of the Dissertation Defence Panel at 10 a.m. on September 12th, 2022 in meeting room 211 of the Institute of Computer Science of Vilnius University. Address: Didlaukio str. 47, LT-08303, Vilnius, Lithuania. Tel. +370 5 219 5040 ; e-mail: mif@mif.vu.lt

The text of this dissertation can be accessed at the Library of Vilnius University, as well as on the website of Vilnius University:
`www.vu.lt/lt/naujienos/ivykiu-kalendorius`

https://doi.org/10.15388/vu.thesis.367
https://orcid.org/0000-0001-8126-3459

VILNIAUS UNIVERSITETAS

Tomas Raila

Kompiuterinio modeliavimo metodai fosfolipidinių membranų pažaidos įvertinimui

DAKTARO DISERTACIJA

Gamtos mokslai,
informatika (N 009)

VILNIUS 2022

Disertacija rengta 2017-2021 metais Vilniaus universitete. Disertacijos rengimą iš dalies parėmė projektas „Kiekybinė fosfolipidinių membranų pažaidos poras formuojančiais toksinais detekcija", finansuojamas Lietuvos mokslo tarybos (projekto nr. P-MIP-19-394).

Mokslinis vadovas - prof. dr. Tadas Meškauskas (Vilniaus universitetas, gamtos mokslai, informatika - N 009).

Gynimo taryba:
Pirmininkė - **prof. dr. Olga Kurasova** (Vilniaus universitetas, gamtos mokslai, informatika – N 009).
Nariai:
dr. Jolita Bernatavičienė (Vilniaus universitetas, gamtos mokslai, informatika – N 009),
prof. dr. Stefano Bonnini (Feraros universitetas, Italija, gamtos mokslai, informatika – N 009),
dr. Rima Budvytytė (Vilniaus universitetas, gamtos mokslai, biochemija – N 004),
prof. dr. Arnas Kačeniauskas (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – T 007).

Disertacija ginama viešame Gynimo tarybos posėdyje 2022 m. rugsėjo 12 d., 10 val. Vilniaus Universiteto Matematikos ir informatikos fakulteto Informatikos instituto 211 auditorijoje. Adresas: Didlaukio g. 47, LT-08303, Vilnius, Lietuva. Tel. +370 5 219 5040; el. paštas: mif@mif.vu.lt

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir Vilniaus universiteto interneto svetainėje adresu:
`www.vu.lt/lt/naujienos/ivykiu-kalendorius`

Table of Contents

Notation

$\arg Y_{min}$ - minimum admittance (Y) phase value
f - frequency of alternating electric current
f_{min} - frequency at the minimum admittance phase ($\arg Y_{min}$) value
N - defect count
N_{def} - defect density
r_{def} - defect radius
R^2 - coefficient of determination
ρ_{sub} - specific resistance of the submembrane layer
Y - admittance

Abbreviations

AC - alternating current
AFM - atomic force microscopy
CHT - circular Hough transform
CNN - convolutional neural network
EEC - equivalent electrical circuit
EIS - electrochemical impedance spectroscopy
FEA - finite element analysis
KNN - K-nearest neighbors
MAE - mean absolute error
MAPE - mean absolute percentage error
PCR - principal component regression
PDE - partial differential equation
PLS - partial least squares
PFT - pore-forming toxin
tBLM - tethered bilayer lipid membrane

Introduction

Biological membranes are a major object of research in various life sciences due to their key role in many vital physiological processes taking place in the cells of living organisms. Due to their complex nature and difficulties associated with studying them in their natural environment, a variety of artificial membrane models have been developed over the last decades [78]. Such biomimetic membranes closely resemble the main workings of their natural counterparts while enabling their extensive study in a controlled laboratory environment by various experimental techniques. Tethered bilayer lipid membranes (tBLMs, Figure 1) is one particular type of such artificial membrane, notable in their versatility and stability under experimental conditions [88]. Applications of tBLMs include studies of membrane-protein or membrane-peptide interactions [27, 46, 68], biosensor development [9, 41, 91, 87], photocurrent generation [39] and others.

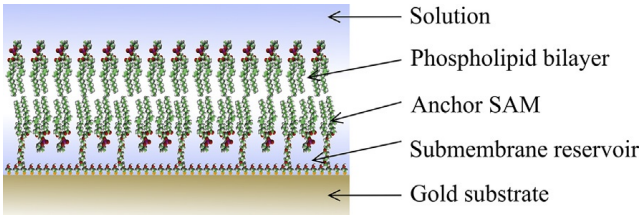


Figure 1: Schematic of the structure of a tethered bilayer lipid membrane [63].

One of the key experimental techniques used in tBLM research is the electrochemical impedance spectroscopy (EIS) [45]. This method is based on the measurement of the electrical response (impedance) to alternating current (AC) applied over a certain frequency range to the system under test. EIS has been applied to assess the tBLM membrane damage caused by interaction with pore-forming toxins (PFT), contributing to the research of tBLM-based biosensors [63, 38]. Although EIS can reveal basic physical properties of the examined object, in-depth interpretation of the spectra is only possible by modeling. A common approach is the equivalent electrical circuit (EEC) model, which involves matching the experimentally obtained EIS spectra with the ones obtained from electrical circuits consisting of simple elements (resistors, capacitors). Despite being extensively used in lipid membrane studies [27, 8] such a model proved to be insufficient for explaining certain parameters of more complex tBLM systems, such as the properties of membrane defects. One notable

work in alleviating this problem was the development of an analytical model for tBLMs with a special case of regularly-distributed defects [45]. Numerical modeling approaches for EIS such as finite element analysis (FEA) [49] have also been recently applied in tBLM research and related fields [37, 32, 85, 95].

Another crucial tool in studying tBLMs is atomic force microscopy (AFM). This is a nanometer-resolution non-optical imaging technique that is based on measuring the interaction between the sample and the microscope's mechanical probe [36]. In contrast to EIS, which is used to determine the overall physical properties of a tBLM sample (possibly spanning millimeters in surface area), AFM is typically used to probe a micrometer-size tBLM surface patch acquiring a localized but very detailed view. This can reveal various surface features, such as membrane defects caused by interaction with PFTs or other phenomena. AFM is sometimes used in conjunction with EIS in analyzing the same tBLM samples [71, 75].

Despite the well-established experimental procedures and sophisticated measurement equipment used in studying tBLMs, methods for analyzing and intepreting experimental EIS or AFM data are still limited. EIS spectra cannot directly reveal important membrane properties, such as density or size of membrane defects - although these properties can be estimated from AFM image data. However, analyzing AFM images of tBLMs often involves mostly manual work of annotating and quantifying objects of interest and using only basic image processing tools [68, 84].

Research objectives

The main goal of the research is to develop a methodology for modeling the electrochemical response of three-dimensional tBLM membranes with arbitrarily distributed defects and interpreting EIS data by using machine learning techniques in order to estimate qualitative and quantitative properties of the membrane damage. The key objectives in achieving this are the following:

1. Develop a three-dimensional numerical model capable of simulating EIS spectra of tBLM membrane with an arbitrary spatial defect distribution.

2. Create predictive models for estimating quantitative tBLM membrane characteristics from their EIS spectra by using machine learning techniques.

3. Create defect distribution models suitable for producing realistic computer-generated defect sets, define metrics for their comparison and investigate model effects on simulated EIS data.

4. Develop methods for automated defect detection in AFM image data and their accuracy evaluation, and examine the performance impact on the modeled EIS spectra.

5. Validate proposed methods and synthetically-generated modeling data against experimentally-obtained EIS and AFM measurement data.

Research methods and tools

The three-dimensional tBLM membrane model was implemented using the finite element method (FEM) and COMSOL Multiphysics FEM package (versions 5.3 - 5.4), and a special model preparation tool (using COMSOL API) was implemented in Java. Most scripts used for data analysis were implemented in Python (version 3.7), using core scientific libraries (Numpy, SciPy, Pandas, Matplotlib), machine learning libraries (scikit-learn, Tensorflow) and Jupyter Notebook environment. Calculations were performed in several different hardware environments:

- Workstation 1 (Intel Core i5-8600K 3.60 GHz CPU (6 cores), 64 GB of RAM, Ubuntu Linux 18.04 OS).

- Workstation 2 (4 x Intel Xeon Gold 6126 2.60GHz CPU (4 x 12 cores), 377 GB RAM, CentOS Linux 7).

- Thinkpad T470s notebook (Intel Core i5-7300 CPU (4 cores), 20 GB RAM, Debian Linux).

Scientific novelty

The presented methodology is novel in its capabilities of analyzing a broad range of tBLM membrane models having different properties and interpreting their electrochemical response to estimate various qualitative and quantitative characteristics of the membranes. It can be summarized as follows:

1. A three-dimensional model of a tBLM membrane was implemented using the finite element method which allows for modelling electrochemical impedance response with any given defect distribution, generated independently.

2. Novel defect clustering models were developed and demonstrated to be capable of generating realistic defect sets at varying clustering levels. In addition, metrics for clustering effect estimation from arbitrary defect sets were defined.

3. Algorithms for automated defect detection in AFM images were developed and the relationship between their accuracy and correspondingly modeled EIS spectra were investigated for the first time.

Practical significance of the results

The developed methodology of EIS data analysis can be applied for fast quantitative membrane damage assessment in tBLM-based impedance biosensors or other similar systems. It also enables the estimation of certain membrane properties (such as the specific resistance of submembrane reservoir or the clustering of membrane defects) which cannot be measured directly by using EIS or AFM techniques. Methods of automated defect detection in AFM images can be beneficial to researchers working in the domain area by making the process of AFM data analysis faster and more precise.

Statements to be defended

1. Quantitative properties of phospholipid membranes with defects can be estimated from their electrochemical impedance spectra by using finite element modeling and machine learning methods.

2. Defect clustering phenomena in phospholipid membranes can be described with computational models which enable the quantification of the clustering effect from atomic force microscopy images or electrochemical impedance spectra.

3. Computer vision techniques can be applied to automatically detect membrane defects in atomic force microscopy images at the accuracy levels sufficient for modeling purposes.

Approbation

The research results have been published in peer-reviewed periodical scientific journals by three publications [A1, A2, A5]. The thesis author's contributions

to each listed publication include numerical model development and implementation, conducting computational experiments, data validation and analysis, writing parts of manuscripts and LaTeX text formatting. In addition, two articles have been published in international conference proceedings [A4, A3].

The author has presented the research results in the following scientific conferences:

1. DAMSS 2018 (Druskininkai, Lithuania). Finite elements modeling of electrochemical impedance spectra of defected phospholipid membranes (poster presentation). Data Analysis Methods for Software Systems, 10th international workshop. November 29 - December 1, 2018.

2. JMK 2019 (Vilnius, Lithuania). Baigtinių elementų metodo taikymas modeliuojant defektuotų fosfolipidinių membranų elektrocheminio impedanso spektrus (poster presentation). 9th Conference for Lithuanian Junior Researchers, Interdisciplinary Applications of Physical and Technological Sciences. March 12, 2019.

3. NUMTA 2019 (Crotone, Italy). Computer modeling of electrochemical impedance spectra for defected phospholipid membranes: finite element analysis (oral presentation). Numerical Computations: Theory and Algorithms, 3rd international conference. June 15 - June 20, 2019.

4. ICCSA 2020 (online). Computational models of defect clustering for tethered bilayer membranes (oral presentation). International Conference on Computational Science and its Applications, 20th international conference. July 1 - July 4, 2020.

Structure of the thesis

The thesis consists of three main chapters. Chapter 1 presents the three-dimensional tBLM model, capable of simulating EIS response of a membrane with an arbitrary defect distribution. Finite element analysis is applied for various modeling cases, involving different membrane defect sizes, densities and distribution patterns. Membrane parameter estimation is performed by applying machine learning techniques both for modeled and experimental EIS and AFM data. Chapter 2 describes the membrane defect clustering phenomena and presents several algorithms applicable in generating realistic clustered defect sets. Clustering effect on modeled EIS spectra and the methods for its quantification are investigated, with the presented approach also being applied

for experimental AFM data. Chapter 3 covers the problem of automated defect detection in AFM images and presents several algorithms for this task, with the impact of defect detection accuracy on modeled EIS spectra being analyzed as well.

1. Electrochemical response modeling of damaged phospholipid membranes

1.1. Atomic force microscopy

Atomic force microscopy (AFM) is an imaging technique based on the measurement of the mechanical interaction between the instrument's probe and the sample surface. AFM microscope typically consists (Figure 2) of a cantilever with a sharp tip, feedback control of the detection system for measuring the cantilever's bending, sample movement system and the visualization system of the acquired data [83].

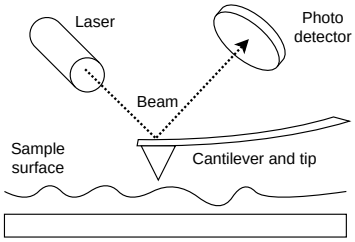


Figure 2: Schematic illustration of the working principle of AFM microscope.

The imaging process involves moving the cantilever with the tip over the sample and measuring its displacement which is caused by a repulsive or attractive force acting between the tip and the surface. The tip can interact with the sample in different operation modes, such as the contact mode (the tip is constantly in contact with the surface) or the oscillation mode (the cantilever with the tip oscillates with a certain frequency and amplitude). Collecting and processing the data allows the topography of the sample surface to be reconstructed, as well as additional information depending on the operation mode (i.e. amplitude and phase - Fig. 3).

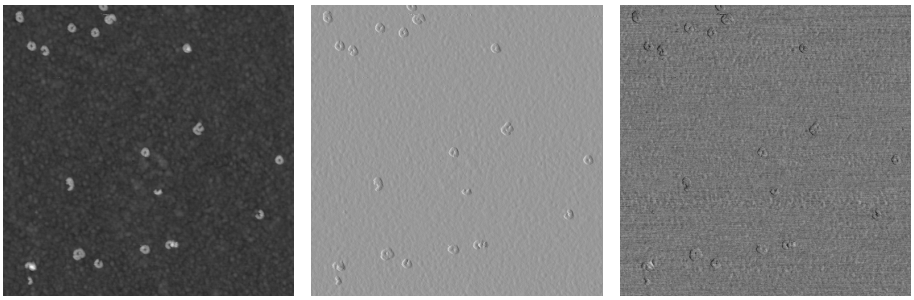


Figure 3: Example of tBLM surface images obtained by AFM. Left: height data; center: amplitude data; right: phase data.

AFM is capable of obtaining much higher image resolution compared to optical microscopy, which is limited by the wavelength of light being used. It also has the advantage of not requiring the sample to be conductive, in contrast to another powerful microscopy technique, the scanning tunneling microscopy (STM). Such properties make AFM a useful tool in life sciences and the study of cells, proteins, viruses and other biological objects [36, 33].

1.2. Electrochemical impedance spectroscopy

Electrochemical impedance spectroscopy (EIS) is a popular technique for characterizing electrochemical systems in terms of their conductive and dielectric properties. The basic working principle of EIS is the application of AC (alternating current) voltage on the system and measuring the current flowing through. This allows determining the impedance of the system, a physical quantity describing the ability to resist the flow of electric current. By measuring impedance at different AC frequencies, one can obtain a spectrum reflecting the physical properties of the object under investigation [30].

AC voltage with angular frequency ω and amplitude V_0 applied to the system can be expressed as a function of time t:

$$V(t) = V_0 \sin(\omega t). \tag{1}$$

The measured current has the same frequency but can differ in amplitude I_0 and phase ϕ:

$$I(t) = I_0 \sin(\omega t + \phi). \tag{2}$$

Following Ohm's law, the impedance Z is then defined as:

$$Z = \frac{V(t)}{I(t)} = \frac{V_0 \sin(\omega t)}{I_0 \sin(\omega t + \phi)} = Z_0 \frac{\sin(\omega t)}{\sin(\omega t + \phi)}. \tag{3}$$

Impedance can be expressed as a complex number with magnitude Z_0 and argument ϕ. Another physical quantity commonly used in place of impedance is the admittance Y - a reciprocal term indicating how easily the circuit lets the electric current flow through:

$$Y = \frac{1}{Z}. \tag{4}$$

The admittance or impedance of an EIS measurement is often visualized in the form of Bode plots (Fig. 4), which show the frequency response of the system:

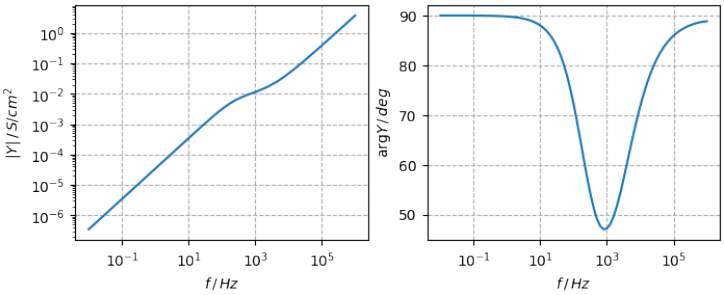


Figure 4: Example of Bode plots of modeled EIS data. Left: admittance magnitude; Right: admittance phase.

1.3. Three-dimensional membrane model

1.3.1. Model definition

The purpose of the three-dimensional membrane model presented in this thesis is to simulate the flow of alternating current through a membrane cell containing defects and determine its electrochemical response as a function of AC frequency. The model consists of four stacked layers, representing (in top-down order) the solution, membrane, submembrane reservoir and Helmholtz layers, with the addition of an arbitrary number of membrane defects (Figure 5, left). The solution and submembrane layers together with membrane defects are electrically conductive, while the membrane and Helmholtz layers are not and have dielectric properties. A similar structure of a tBLM membrane model was described earlier [37].

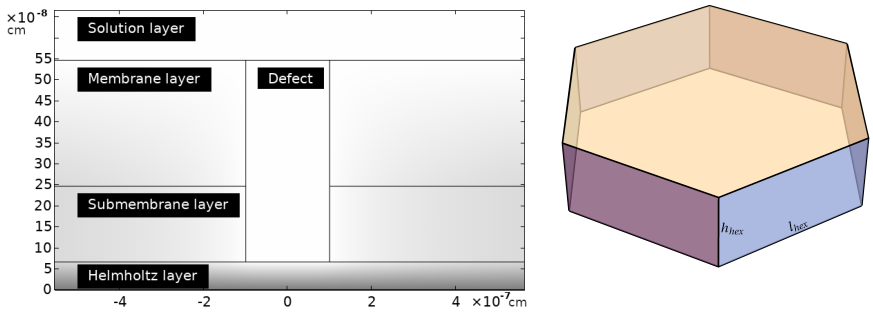


Figure 5: Schematic representation of the three-dimensional membrane model. Left: cross-section of the model within the vicinity of a defect. Right: hexagonal modeling domain.

The modeling domain consisting of the listed layers is shaped like a hexagonal prism (Figure 5, right) with l_{hex} side length. The height of the prism h_{hex} is the sum of the heights of Helmholtz (d_H), submembrane (d_{sub}), membrane (d_m) and solution (d_{sol}) layers:

$$h_{hex} = d_H + d_{sub} + d_m + d_{sol}. \tag{5}$$

The area of the hexagon is defined as:

$$S_{hex} = \frac{3\sqrt{3}(l_{hex})^2}{2}. \tag{6}$$

The motivation for the choice of this specific model geometry relates to the fact that any flat macroscopic 2D area can be filled by the microscopic hexagons with no voids. A hexagon also allows for the densest possible packing of equally-sized circles [35] which is advantageous for the comparison of this model and the radial symmetry approach [45] where each membrane defect is associated with a certain occupancy radius. Despite these preferences, the model geometry can also be derived from a different shape, such as a square, without changes in any other properties of the model.

The given instance of the model may contain any number N of arbitrarily positioned membrane defects. Each defect is represented as a cylinder with radius r_{def} intersecting the membrane and submembrane layers and having a height of $d_m + d_{sub}$. Defect density N_{def} is expressed as the number of defects per square micrometer (assuming the same units of S_{hex}):

$$N_{def} = \frac{N}{S_{hex}}. \tag{7}$$

The flow of electric current through the system is governed by Laplace's equation. This is an elliptic second-order partial differential equation applicable to steady-state (independent of time) problems and is commonly used in many areas of physics, such as electrostatics, fluid dynamics, heat conduction and others [40]. Its application in computing impedance of various electrochemical systems has been described in many works [32, 14, 31, 7, 42].

In the current modeling case the solution of this equation is the complex voltage Φ expressed as a function over the 3D modeling domain:

$$\nabla \cdot (\tilde{\sigma}(x,y,z)\nabla\Phi(x,y,z)) = 0. \tag{8}$$

Here $\tilde{\sigma}$ denotes the complex conductivity at any point of the system:

$$\tilde{\sigma}(x,y,z) = \sigma(x,y,z) + j\,\omega\varepsilon(x,y,z). \tag{9}$$

The real and imaginary parts (imaginary unit denoted by j) correspond to the conductivity and permittivity of different parts of the system. Specific conductivity σ is applicable for conductive layers of the model: the solution, submembrane layers and the membrane defects. In contrast, permittivity ε (dielectric constant) is defined for the membrane and Helmholtz layers. $\omega = 2\pi f$ denotes the angular AC frequency and f is the frequency in hertz (Hz). Specific resistance ρ can also be used to describe the conductive layers and is defined as the reciprocal of specific conductivity:

$$\rho = \frac{1}{\sigma}. \tag{10}$$

Electric current flow through the system is facilitated by the application of 1 V electric potential at the top of the modeling domain and 0 V at the bottom. This defines the following Dirichlet boundary conditions of the model:

$$\Phi(x,y,h_{hex}) = 1, \tag{11}$$

$$\Phi(x,y,0) = 0. \tag{12}$$

Each side wall of the hexagonal prism is assumed to be electrically insulating. This allows us to define the Neumann boundary conditions, where n denotes the normal vector of a side wall:

$$n \cdot \nabla\Phi(x,y,z) = 0. \tag{13}$$

Once the equation (8) is solved for Φ, the current density J can be calculated at any point of the modeling domain, following Ohm's law [40]:

$$J(x,y,z) = -\tilde{\sigma}(x,y,z)\,\nabla\Phi(x,y,z). \tag{14}$$

The final step is computing the admittance Y at the top of the modeling domain, where n is the normal vector of the surface:

$$Y = \frac{\iint\limits_{(x,y)\in\Gamma_{hex}} -n\cdot J(x,y,h_{hex})\,dx\,dy}{S_{hex}} \times \frac{1}{\Phi(x,y,h_{hex})}. \tag{15}$$

As the admittance is evaluated for a single AC frequency value selected in (9), obtaining the full EIS spectrum requires performing the described calculations with a certain range of discrete frequency values. All modeling cases presented in this thesis were computed with frequency values spanning a range from $10^{-}2$ Hz to 10^{6} Hz on a logarithmic scale with 10 points per decade, resulting in a total of 81 values.

Table 1: Parameters employed in modeling.

Description	Notation	Value	Dimension
Thickness of the Helmholtz layer	d_H	$6.6\cdot10^{-8}$	cm
Thickness of the submembrane reservoir	d_{sub}	$1.8\cdot10^{-7}$	cm
Thickness of the membrane hydrophobic core	d_m	$3\cdot10^{-7}$	cm
Thickness of the solution	d_{sol}	$50\cdot10^{-7}$	cm
Height of the model hexagonal prism	h_{hex}	$6.46\cdot10^{-7}$	cm
Side length of the base of the model hexagonal prism	l_{hex}	variable	cm
Relative dielectric constant of the Helmholtz layer	ε_H	4.0975	dimensionless
Relative dielectric constant of the phospholipid membrane	ε_m	2.2	dimensionless
Specific conductance of the submembrane reservoir	σ_{sub}	1×10^{-5}	$\mathrm{S\,cm^{-1}}$
Specific conductance of the water-filled channel*	σ_{def}	1×10^{-2}	$\mathrm{S\,cm^{-1}}$
Specific conductance of the solution	σ_{sol}	1×10^{-2}	$\mathrm{S\,cm^{-1}}$
Defect radius	r_{def}	variable	cm
Defect count	N	variable	dimensionless
Defect density	N_{def}	variable	$\mathrm{cm^{-2}}$

* it is assumed that the water-filled channel extends from the top of the membrane to the Helmholtz layer.

Figure 6 shows an example of a modeled EIS spectrum (admittance phase $\arg Y$ dependency on AC frequency f). Admittance Y is a complex quantity, thus its argument (phase) $\arg Y$ is expressed in degrees (deg), while the frequency is measured in hertz (Hz). Qualitatively similar curves have been observed both in experimental measurements and modeled data in earlier studies [71, 46]. A notable property of such EIS spectra is the presence of a single minimum point at mid-range frequencies. Previous research showed that the coordinates (16,17) of this particular point can be indicative of certain quantitative membrane parameters, such as the defect density or their size [45]. However, it is assumed in this thesis that such property may not necessarily apply to all modeling cases and some variations in the shape of EIS spectrum (i.e. more than one minimum point) might be encountered as well.

$$f_{min} \quad \text{– frequency } f \text{ at which } \arg Y(f) \text{ is lowest,} \tag{16}$$

$$\arg Y_{min} \text{ – admittance phase value at } f_{min}. \tag{17}$$

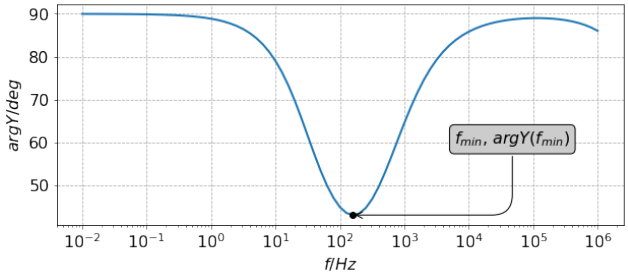


Figure 6: Bode plot of modeled EIS spectrum showing admittance phase dependency on the AC frequency. The annotated black dot indicates the minimum point of the curve. The admittance phase $\arg Y$ is expressed in degrees (deg), frequency f units are hertz (Hz).

A special case of the membrane model with zero defects could also be considered. Such a case results in a distinct EIS spectrum that does not have a distinguishable minimum qualitatively similar to the model cases containing defects. Figure 7 shows an example of such a spectrum which has constant $\arg Y$ values in the major part of the frequency range and indicates a decrease only in the higher frequencies ($f > 10^4$). Due to such properties, modeling cases with $N = 0$ are not further considered in this thesis.

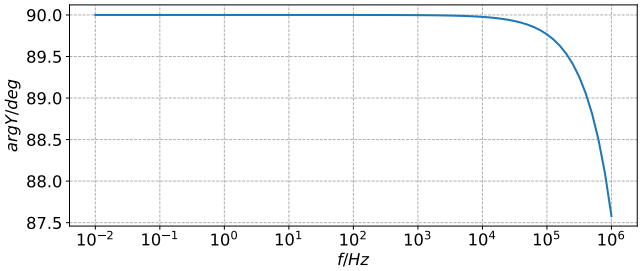


Figure 7: Example of EIS spectrum of a membrane model containing no defects.

1.3.2. Finite element analysis

Partial differential equations (PDE) describing various real-world phenomena are often impossible to be solved analytically, which requires the use of numerical techniques to find approximate solutions. The most widely used types of such numerical methods are the finite difference method (FDM), finite element method (FEM) and finite volume method (FVM) [50]. FDM is the oldest and relatively simplest approach which approximates the derivative terms of PDEs by finite differences, expressed over the problem domain discretized by a regularly-spaced grid [12]. This converts the problem defined by PDE into a system of linear equations, which can be solved efficiently by various linear algebra techniques to obtain an approximate solution with the desired accuracy. However, FDM is difficult to use with curved, irregular or otherwise complex geometries and is preferable for problem areas involving rectangular or box-shaped geometries, such as meteorological, seismological, astrophysical simulations and many others [80, 66].

Due to the non-trivial geometry of the described 3D membrane model, consisting of several layers with different physical properties and containing cylinder-shaped defects, FEM has been chosen over FDM to perform simulations. FEM [49] is based on a different approach of discretization where the solution is approximated as the sum of elementary (basis) functions defined over small discrete parts of the domain (elements). Originally developed for simulations in structural mechanics, FEM has since evolved into a highly versatile numerical technique applicable to various problems in solid and fluid mechanics, electromagnetics and other areas [48]. Solving this problem with FEM generally involves the following steps:

1. Weak formulation of the problem. The original (strong) PDE form is

converted to the weak form, relaxing continuity requirements for the solution. Such a concept is illustrated by the following classic example of Poisson's equation (generalization of Laplace's equation) with Dirichlet boundary conditions:

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \delta\Omega.$$

Weak formulation is derived by integrating over domain Ω, multiplying both sides of the equation by a test function v and integrating by parts:

$$\int_\Omega \nabla u \cdot \nabla v \, d\Omega = \int_\Omega f \cdot v \, d\Omega.$$

2. Selection of the shape functions. The solution $\tilde{u}(x)$ is approximated by a linear combination of the basis (shape) functions ϕ_i:

$$\tilde{u}(x) = \sum_{i=0}^{N} u_i \phi_i(x).$$

Shape functions interpolate the solution locally over each element. An important characteristic is that a single function ϕ_i is equal to 1 at i-th node and 0 at all other nodes (where nodes are points of intersection between discrete elements). Linear or low-order polynomial functions are usually used as shape functions.

3. Discretization of the modeling domain (mesh). The domain Ω is subdivided into non-overlapping elements of an arbitrary size, intersecting at N nodal points (nodes). In the case of 2D problems triangular or quadrilateral elements are commonly used, while 3D domains can be discretized by tetrahedra, hexahedra, prisms or pyramids.

4. Assembling the stiffness matrix. The contributions of all basis functions to the k-th element are defined by a local stiffness matrix:

$$A_{ij}^{(k)} = \int_{\Omega_k} \nabla \phi_i \cdot \nabla \phi_j \, dx.$$

The global stiffness matrix is constructed by summing all local stiffness matrices, leading to a linear system:

$$\mathrm{A}u = f.$$

Matrix A is a sparse, symmetric and positive-definite $N \times N$ square matrix.

5. Solving the system of linear equations. The assembled linear system is solved either by a direct or iterative method (solver) [51]. Direct solvers are robust and guaranteed to find the exact solution in a finite amount of calculations, although they typically require lots of computing resources and scale poorly as the size of the linear system increases. Iterative solvers gradually converge towards an approximate solution, are less stable and their performance for a specific problem largely depends on their configuration (such as convergence tolerance or the choice of preconditioner). However, they tend to require much less memory, making them applicable to large-scale problems, for which direct methods would be impractical. Most FEM software packages include efficient, parallelized implementations of both types of solvers [17, 5].

1.3.3. Defect distribution models

As the presented three-dimensional membrane model is flexible in its ability to represent an arbitrary number of membrane defects distributed in any way, it is necessary to define algorithms for generating such defect sets so that they would exhibit some desired properties. Given the defect count N and the defect density N_{def}, such an algorithm would generate a list of defects where each instance is defined by its center coordinates (X and Y) in the modeling plane and its radius. Although in principle both coordinates and radii can vary across the defect set, all defect distribution models in this thesis are defined with a simplifying assumption of constant defect radius r_{def}, which is provided as a parameter for the defect set generation algorithm.

The regular defect distribution model is the first one introduced in this thesis. This model defines a pattern where each defect has a constant occupancy radius and all defects are placed in the modeling domain with equal distances between each other, resembling a regular grid. From the computational point of view this can be considered a circle packing problem [29] with the aim of placing N equal-sized circles into a hexagonal container. Such a model closely resembles the modeling approach presented in the earlier study, where the EIS response of a tBLM membrane is computed analytically, using Hankel functions [45]. The algorithm for generating a defect set following the regular model is implemented with the limitation that the N value is chosen specifically to result in a maximum fill of the hexagonal area. If k denotes the number of defects along a single hexagon side, N is then derived as:

$$N = 3k(k+1)+1. \tag{18}$$

Figure 8 (left panel) shows an example of a defect set generated following this approach.

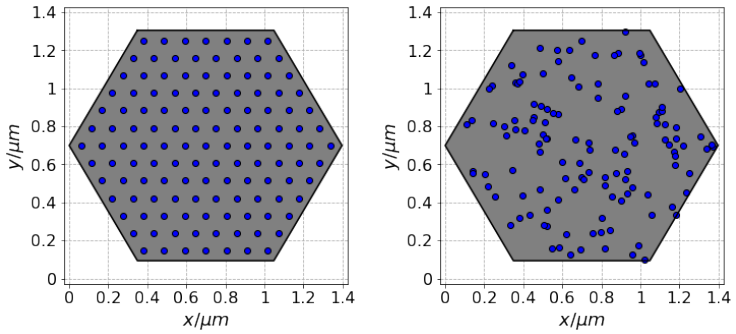


Figure 8: Examples of different spatial defect distributions on the modeled hexagonal membrane surface ($N = 127$). Left: regular defect distribution. Right: random defect distribution.

Although the regular defect distribution model enables a direct comparison of the FEA-based approach described in subsection 1.3.1 against the analytical solution [45], experimental evidence shows that such defect patterns are unlikely to be observed in real tBLM membranes [37, 71]. The next model, introduced as a more realistic yet simple alternative to the aforementioned regular model is the random defect distribution model. In this approach, the X and Y coordinates of each defect are sampled randomly from the uniform distribution, in which the low and high bounds match the dimensions of the modeling domain (which depends on the defect count N and the defect density N_{def} selected for the model). In the case of the hexagonal model domain shape, the sampled coordinates are also checked if they are inside the domain, otherwise the defect instance is discarded. Based on the assumption that there is no interaction between individual defects in the membrane, this procedure is repeated until the required number of defects inside the modeling domain are collected. An instance of the random defect distribution model is shown in Figure 8 (right panel).

1.3.4. Membrane parameter properties

The parameters N_{def}, r_{def} and ρ_{sub} (the reciprocal of specific conductance of the submembrane layer σ_{sub}) are the most important in quantifying the membrane damage and directly influence the EIS response of the system [63, 71]. However, it has been shown that these parameters are coupled and their effect on the EIS spectrum depends on the combination of their values [45]. This poses an issue that the three listed parameters cannot be unambiguously determined all at once from a given EIS spectrum and different parameter combinations can correspond to identical EIS spectra.

A proposed solution is presented further. The initial assumption (based on earlier research [45]) is that any triplet of $N_{def(0)}$, $r_{def(0)}$ and $\rho_{sub(0)}$ values is related to other N_{def}, r_{def} and ρ_{sub} triplet via some constant t:

$$\begin{aligned} N_{def} &= N_{def(0)} \cdot t^2, \\ r_{def} &= r_{def(0)} \ / \ t, \\ \rho_{sub} &= \rho_{sub(0)} \cdot t^2. \end{aligned} \tag{19}$$

Taking the logarithm of both sides of all equations:

$$\begin{aligned} \log N_{def} &= \log N_{def(0)} + 2\log t, \\ \log r_{def} &= \log r_{def(0)} - \log t, \\ \log \rho_{sub} &= \log \rho_{sub(0)} + 2\log t. \end{aligned} \tag{20}$$

Substituting $\log N_{def} = x$, $\log r_{def} = y$, $\log \rho_{sub} = z$ and $\log t = s$:

$$\begin{aligned} x &= x_0 + 2s, \\ y &= y_0 - s, \\ z &= z_0 + 2s. \end{aligned} \tag{21}$$

This can be interpreted as the parametric equations of a 3D line. As the direction vector $\vec{d} = (2, -1, 2)$ is constant, lines corresponding to different model parameters will be parallel to each other. Furthermore, each line can be characterized by a point at which it intersects a plane whose normal vector is equal to the direction vector $\vec{d}$. Given that such plane passes through the point $(0,0,0)$, its equation is the following:

$$2x - y + 2z = 0. \tag{22}$$

Substituting each component by its parametric expression (21) allows us to define the parameter s at which the line intersects the plane:

$$s = -\frac{2}{9}x_0 + \frac{1}{9}y_0 - \frac{2}{9}z_0. \tag{23}$$

The intersection point p of the line (expressed by $p_0 = (x_0, y_0, z_0)$) and this plane can then be derived from (21) and (23):

$$p = Qp_0 = \begin{bmatrix} \frac{5}{9} & \frac{2}{9} & -\frac{4}{9} \\ \frac{2}{9} & \frac{8}{9} & \frac{2}{9} \\ -\frac{4}{9} & \frac{2}{9} & \frac{5}{9} \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix}. \tag{24}$$

As these 3D points lie on the same plane, they can be equivalently represented by two coordinates instead of three. This can be achieved with a linear transformation (rotation) into a plane perpendicular to the z-axis (normal vector $\vec{n} = (0,0,1)$) and containing the point $(0,0,0)$. The rotation matrix can be derived from the angle θ between the normal vectors of the both planes ($\cos\theta = \frac{2}{3}$, $\sin\theta = \frac{\sqrt{5}}{3}$) and the intersection line $(-\frac{1}{\sqrt{5}}, -\frac{2}{\sqrt{5}}, 0)$:

$$R = \begin{bmatrix} \frac{11}{15} & \frac{2}{15} & -\frac{2}{3} \\ \frac{2}{15} & \frac{14}{15} & \frac{1}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \end{bmatrix}. \tag{25}$$

Applying this rotation on a given intersection point p (24) results in a point on the x-y plane with $z = 0$:

$$Rp = \begin{bmatrix} v_1 \\ v_2 \\ 0 \end{bmatrix}. \tag{26}$$

The described procedure can be applied in reverse to reconstruct the initial N_{def}, r_{def} and ρ_{sub} values when the coefficients v_1 and v_2 are known. However, this requires one of the three parameters to be fixed so that the other two can be unambiguously determined. The first step in the reverse procedure is calculating the point p_0 (24) by solving the following equation:

$$Qp_0 = R^T v. \tag{27}$$

Assuming that either x, y or z is known, the point on the line defined by s can then be determined from (21), leading to the remaining two coordinates of that point and, subsequently, the values of N_{def}, r_{def} and ρ_{sub}.

In summary, the coefficients v_1 and v_2 can be used to represent the set of all membrane parameter values (19) given some initial point $p_0 = (x_0, y_0, z_0)$. This also implies that EIS spectra obtained from multiple membrane models with the same spatial defect distribution but different sets of parameters satisfying the (19) condition should be equivalent and could be represented by a single unique pair of v_1 and v_2 values. To verify this assumption experimentally, a single membrane model was prepared with randomly distributed defects ($N = 200$) and specific initial parameters ($N_{def(0)} = 10$; $r_{def(0)} = 10$; $\rho_{sub(0)} = 10^{4.5}$). Then, a series of additional models were generated from the initial one, where each new set of N_{def}, r_{def} and ρ_{sub} parameters was derived by applying the coefficient t, which was adjusted from 0.4 to 2 with increments of 0.1. This resulted in a total of 17 models, each having the same coefficients $v_1 = -2.1(3)$ and $v_2 = 2.5(6)$. The changes of N_{def} were implemented by scaling the modeling domain and the center coordinates of each defect so that the same overall defect count and their relative positions would be retained.

Figure 9 shows the modeled EIS spectra of all models. The only noticeable differences exist in the highest frequency range ($f > 10^5$), while the remaining parts of the spectra look identical. The standard deviations of $\log f_{min}$ and $\arg Y_{min}$ values are 6.34×10^{-4} and 1.34×10^{-2} respectively. Such negligible variations can be attributed to the inherent approximation errors existing in the FEA modeling process.

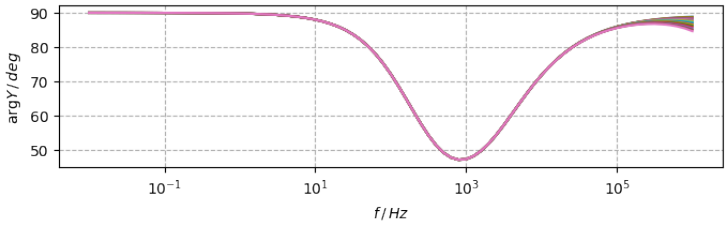


Figure 9: Series of EIS spectra corresponding to the same v_1 and v_2 values.

1.3.5. Model implementation

The presented three-dimensional membrane model was implemented with COMSOL Multiphysics FEA software. Two types of mesh elements were considered for the membrane models - tetrahedral elements and triangular prisms.

Both types of meshes were generated using built-in COMSOL mesh generation functionality. The assumption was made that current flux was the most intense inside and close to defects, so mesh generation parameters for the areas inside and outside defects were set separately (Table 2), ensuring that the defect areas were meshed significantly more densely. Both tetrahedral and prismatic meshes were generated in several density levels depending on the following parameters:

- k_d - ratio between defect radius and mesh element size inside the defect;
- k_h - ratio between the hexagonal prism side length and maximum mesh element size outside defects;
- k_s - number of swept mesh layers for defect and its surrounding submembrane and membrane layers (prismatic mesh only);
- r_0 - defect radius;
- l_h - hexagonal prism side length.

The ratio k_d was varied, while k_h was fixed and set to 20. All defects had the same radius r_0 of 1 nm. The hexagon side length l_h was also fixed in all cases and derived from defect count and density during defect distribution generation. Table 3 shows the dependency between the mesh generation parameters k_d and k_s and the degrees of freedom (DoF) of the resulting model.

Table 2: COMSOL mesh generation settings.

Element size parameter	**Value (defect areas)**	**Value (other areas)**
Maximum element size	r_0 / k_d	l_h / k_h
Minimum element size	r_0 / k_d	l_h / k_h
Maximum element growth rate	1.7	1.7
Curvature factor	0.5	0.5
Resolution of narrow regions	0.5	0.5

In order to estimate the effect of mesh density level (Table 3) on the solution accuracy expressed in terms of EIS spectral features, experiments were performed with the direct solver (MUMPS [17]), both mesh element types (prisms and tetrahedra) and varying mesh densities. In all modeling cases the same model geometry having 100 randomly scattered defects was used.

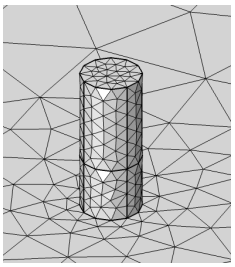


Figure 10: Example of tetrahedral mesh elements at density level #4 in and around the defect.

Table 3: DoF dependency on mesh type and the k_d ratio when the model contains 100 randomly distributed defects.

#	Ratio (k_d)	Swept layers (prisms)	DoF (prisms)	DoF (tetrahedra)
1	0.5	2	3.15E+05	5.05E+05
2	1.0	4	6.46E+05	7.87E+05
3	1.5	6	1.04E+06	9.99E+05
4	2.0	8	1.38E+06	1.27E+06
5	2.5	10	2.03E+06	1.82E+06
6	3.0	12	1.78E+06	2.38E+06
7	3.5	14	3.95E+06	3.25E+06

Results (Figure 11) indicate that for both tetrahedral and prismatic meshes increasing their density past level #3 (k_d = 1.5) does not result in significant changes in $argY(f_{min})$ values, although f_{min} still shows a slight increase or decrease, depending on the mesh element type. By considering such variations at higher mesh density levels as negligible, the majority of modeling instances examined in this thesis were computed with mesh density of levels 3 and 4.

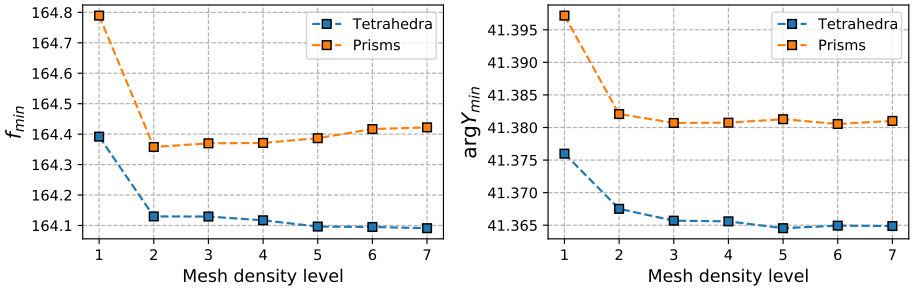


Figure 11: Solution dependency on mesh density levels for tetrahedral and prismatic meshes.

1.4. Regression models

1.4.1. Linear regression

Linear regression is a fundamental concept in statistics and the related fields which models linear relationships between quantitative variables [34]. Such

dependency is expressed between the dependent (target) variable y and one or more independent variables x_i, where the model is defined by its coefficients w_0 to w_d:

$$y = w_0 + \sum_{i=1}^{d} w_i x_i. \tag{28}$$

Variables x_i (features) can represent not only the original knowledge about the modeling domain but can also be derived from other variables by applying various transformations ($\log x_i$, $\sqrt{x_i}$), polynomial basis expansion (x_i^2, x_i^3), interactions between variables ($x_3 = x_1 \cdot x_2$) and more. This allows describing non-linear relationships between the variables while keeping the model itself linear.

Fitting the model to the data requires determining the coefficients $\mathbf{w} = (w_0, ..., w_d)$. The most common approach is the least-squares method, where the coefficients are chosen by minimizing the residual sum of squares (RSS):

$$RSS(\mathbf{w}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{29}$$

Here n denotes the number of training examples, y_i is the actual value of the target variable and $\hat{y}_i$ is the prediction of the model (28). This cost function is quadratic with respect to its coefficients $\mathbf{w}$ and can be minimized by solving the closed-form equations or using some optimization algorithm, such as the gradient descent.

One of the variations of the described linear model is Lasso regression, which introduces an additional regularization term to the cost function:

$$J(\mathbf{w}) = RSS(\mathbf{w}) + \lambda ||\mathbf{w}||_1 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=0}^{d} |w_j|.$$

This term penalizes the model by the coefficient magnitudes resulting in a sparse model, where some coefficients become equal to 0. Such property makes lasso applicable as a feature selection technique which simplifies a linear model by discarding the least informative features. The shrinkage effect is controlled by the regularization parameter λ with higher values increasing the effect.

1.4.2. Principal component regression

One of the regression techniques commonly used with spectral data is the principal component regression (PCR) [13]. This method is based on the combination of principal component analysis (PCA) and linear regression. In this approach, PCA (with selected k principal components) is first performed on the data matrix X to decompose it into score matrix S and loading matrix P:

$$X = SP^T + E. \tag{30}$$

Then, a linear regression model for the prediction of response variable y is trained using the selected principal components S. The advantage of PCR over multiple linear regression (MLR) is that the linear transformation performed by PCA produces uncorrelated variables and thus prevents multicollinearity, which is common in spectroscopy data and can make fitting a linear regression model numerically unstable and increase the risk of overfitting. Although PCR is a relatively lesser-known technique, it has been used with impedance data by some authors [10, 70, 19].

1.4.3. PLS regression

Another regression technique, popular in chemometrics and applicable to spectral data as well, is the partial least squares (PLS) regression [4]. This model represents a linear dependency between predictors and one or more response variables. Like PCR, the PLS method performs linear projection of data into a new space with fewer dimensions and fits a linear regression model on the projected variables. However, PLS works in a supervised approach and the projection is selected to maximize the covariance between the predictor and response variables (unlike PCR which projects only the predictor variables):

$$X = SP^T + E, \tag{31}$$

$$Y = UQ^T + F. \tag{32}$$

Here X is the $n \times m$ matrix of predictors, Y is the $n \times p$ matrix of response variables, S and U are $n \times k$ score (projection) matrices of X and Y, P and Q are loading matrices and E and F denote the error terms. The objective of the PLS algorithm is to maximize the covariance between S and U matrices.

PLS can also be applied for classification tasks where discrete classes are represented as unit vectors (i.e. by performing one-hot encoding), referring to such approach as PLS-DA (PLS discriminant analysis). Some applications of PLS in multivariate analysis of impedance spectra have been described in the literature [89, 44].

1.4.4. K-nearest neighbors regression

K-nearest neighbors (KNN) is a non-parametric supervised learning method applicable to non-linear classification and regression tasks [34]. KNN predictions are based on the selection of the k nearest points (relative to the provided query point) from the training dataset. The general working principle of KNN regression can be described by the following steps:

1. Given the input vector $\mathbf{x}_t$, find k examples $\{(\mathbf{x_i}, y_i) : i = 1..k\}$ from the training set ($\mathbf{x}_i$ and y_i denote the feature vector and target value of a single observation) which are most similar to $\mathbf{x}_t$ in terms of distance metric D.

2. Return the mean of selected y_i values as the prediction of the model.

The distance metric D can be chosen arbitrarily, with Euclidean distance being one of the most common options. The prediction can also be computed as a weighted average, such that the closer neighbors of the input point could have greater influence. Although KNN regression is less common in the analysis of impedance spectra, successful applications of this method have been demonstrated for other types of multivariate spectroscopic data [72, 52, 92].

1.4.5. Model evaluation

Different regression models can be evaluated and compared by using various quantitative metrics. The following list includes several well-known regression metrics [34] used further in this thesis:

- Coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{RSS}{Var(y)}, \tag{33}$$

  where $\bar{y}$ denotes the mean of the true values of y. R^2 metric indicates the fraction of explained variance in model predictions. The best possible

score is $R^2 = 1$, while a constant model which always predicts the same value would have the score $R^2 = 0$. It does not have a lower bound and the values can be negative and arbitrarily low.

- Mean absolute error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| . \tag{34}$$

MAE metric indicates how much, on average, the predictions of the model deviate from the true values. One of the benefits of MAE is its ease of interpretability due to it having the same units as the quantities it describes.

- Mean absolute percentage error (MAPE):

$$MAPE = \frac{100}{n}\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| . \tag{35}$$

MAPE represents the relative error between the true and predicted values in an intuitive way, although it has the drawback of being inapplicable to data with zero values.

1.5. Computational experiments

1.5.1. Defect distribution model comparison

To validate the three-dimensional membrane model implemented using the FEA approach, EIS responses of membrane models with regular defect distributions were first compared with the corresponding analytical solutions. EIS spectra were computed with defect densities N_{def} varying in a logarithmic scale from 10^{-1} to 10^{2} and using two defect size r_{def} options of 1 nm and 25 nm. Regular defect sets used in FEA modeling contained $N = 127$ defects. Analytical solutions were computed by using the most of physical constants and geometry parameters listed in Table 1 and additional specific parameters required by the analytical model: the solution impedance Z_{sol} was set to $0\,\Omega$ and defect resistance Z_{def} set to $1.13 \times 10^{9}\,\Omega$ and $8.08 \times 10^{6}\,\Omega$ for 1 nm and 25 nm defect sizes correspondingly.

Figure 12 shows the comparison of both sets of spectra. With both r_{def} options the resulting spectra are qualitatively similar at every N_{def} value, although some discrepancies in terms of minimum point coordinates $\log f_{min}$

and $\arg Y_{min}$ are apparent. As the defect density increases (in the case of $r_{def} = 1$nm), the absolute difference of $\log f_{min}$ values remains almost constant with a minor increase from 0.1 to 0.12 while the average shift of $\arg Y_{min}$ is 1.58. In the case of large (25 nm) defects the differences are relatively more significant with $\log f_{min}$ shift increasing from 0.09 to 0.14 and $\arg Y_{min}$ shift decreasing from 1.34 to 0.44.

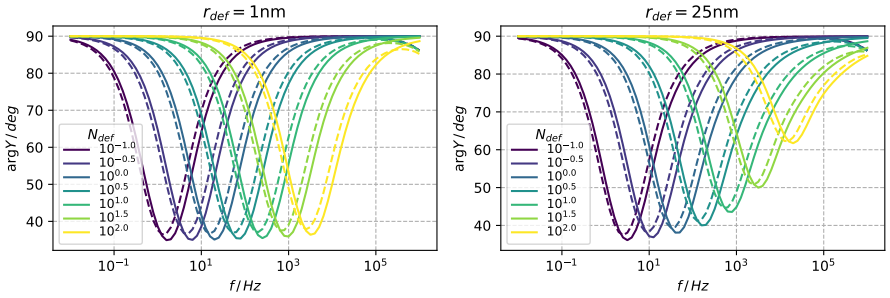


Figure 12: Comparison of EIS spectra obtained by the analytical approach (solid lines) and FEA modeling with regular defect distributions (dashed lines) at varying defect densities and two defect size options.

The next comparison was performed between the regular and random defect distribution models. For each combination of N_{def} and r_{def} a group of 10 membrane models with random defect distributions ($N = 100$) was generated and their EIS spectra were computed. Figure 13 shows the averaged curves of each group in comparison with the regular defect distribution cases. Results indicate much more significant differences between both defect distribution models relative to the previous comparison with the analytical solutions. For 1 nm defects the shift in $\log f_{min}$ values increases from 0.02 to 0.1 while the differences in $\arg Y_{min}$ values show a linear increase from 3.4 to 5.92. The deviations are even higher in the case of 25 nm defects: $\log f_{min}$ shift increases from 0.02 to 0.4 whereas the $\arg Y_{min}$ shifts display a non-linear trend of increasing from 5.62 to 8.11 at $N_{def} = 10^1$ and then dropping to 5.77 at $N_{def} = 10^2$.

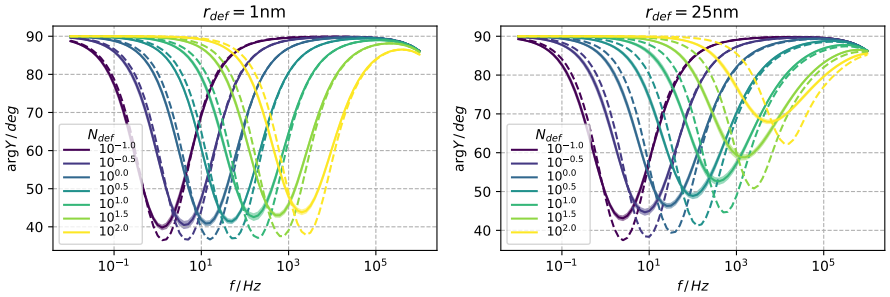


Figure 13: Comparison of averaged EIS spectra modeled using random defect distributions (solid lines) and regular defect distributions (dashed lines) at varying defect densities and two defect size options. The colored bands represent the standard deviations of $\arg Y$ values in each group of random cases.

1.5.2. Dependency on the modeled defect count

Although the defect density N_{def} is one of the main parameters determining the EIS response of the membrane, defect count N is also important to consider from the modeling perspective. There is a trade-off in the selection of specific N value, where the lower number of defects corresponds to a simpler and faster to compute but less representative model of the membrane, resulting in higher variability of EIS spectral features, while a higher defect count leads to more stable results at the expense of a significant increase in computation times and resource utilization.

In order to evaluate the effect of defect count on the properties of modeled EIS spectra and modeling process performance, a series of models with varying defect counts were generated by using the random defect distribution approach. Defect count N was adjusted on an approximately-logarithmic scale from 6 to 500, while the defect density was kept constant at $N_{def} = 10$. This was repeated with two options for defect size r_{def} values of 1 nm and 25 nm. For each unique parameter combination a total of 10 modeling cases with unique random defect distributions were prepared. Modeling was performed in the computing environment equipped with four Intel Xeon Gold 6126 2.60GHz CPUs (4 x 12 cores), 377 GB of RAM and CentOS Linux 7. COMSOL was run in distributed mode using 6 nodes (instances) and 3 CPU cores per node, resulting in a total of 18 cores being utilized in parallel.

Table 4 shows the results of the described experiment. Mean CPU time values indicate approximately linear dependency on the defect count. The standard deviations of $\log f_{min}$ and $\arg Y_{min}$ decrease by lower amounts as defect

count grows. A significant distinction between two r_{def} options can be observed, where the models with small (1nm) defects and $N \geq 25$ take almost twice as long to compute, relative to models with $r_{def} = 25$.

Table 4: Modeling results of random defect distribution model cases with increasing defect counts N.

N	r_{def}	$\log f_{min}$		$argY_{min}$		DoF		$CPU\ time(s)$	
		Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
6	1	2.117	0.074	41.336	1.547	149632	2197	96	1
12	1	2.166	0.037	42.209	1.606	198765	5239	117	3
25	1	2.184	0.023	41.894	1.216	348510	9314	196	6
50	1	2.190	0.013	42.012	0.872	639877	13167	346	9
100	1	2.199	0.018	42.076	0.815	1261131	13448	703	9
200	1	2.201	0.019	42.148	0.322	2524987	7433	1440	24
500	1	2.203	0.008	41.770	0.337	6136825	20407	4038	46
6	25	2.568	0.163	52.540	3.912	108713	1362	75	1
12	25	2.546	0.079	52.736	1.625	119655	2442	73	1
25	25	2.567	0.041	53.258	1.626	187673	5124	107	3
50	25	2.556	0.033	52.852	0.962	324837	5803	173	4
100	25	2.588	0.037	53.072	0.692	627343	17007	334	10
200	25	2.581	0.021	52.800	0.505	1225674	14614	681	42
500	25	2.577	0.016	52.604	0.396	3001465	23966	1800	50

1.5.3. Comparison with experimental AFM data

To evaluate how well the random defect distribution model corresponds to the positions of defects observed in actual tBLM membranes, EIS spectra of both computer-generated and experimentally-measured defect sets were compared. Figures 14 and 15 show two different AFM images of actual tBLM membranes with and without visible defect clusters. Membrane defects were manually annotated by domain expert and their coordinates (ones within the hexagonal modeling domain) were used to model EIS spectra. The first model (without clusters) contained $N = 74$ defects at $N_{def} = 12.66$ density, while the second model (with clusters) had $N = 41$ defects at $N_{def} = 15.78$ density. Additionally, random defect sets were generated (10 cases for each parameter combination) using the same defect counts N and densities N_{def} of the two images to obtain additional models differing only in the exact positions of the defects. Modeling of all cases was performed using four different defect radii $r_{def} = 1, 9, 17,$ 25 nm, representing most likely real defect sizes, according to earlier research

[22].

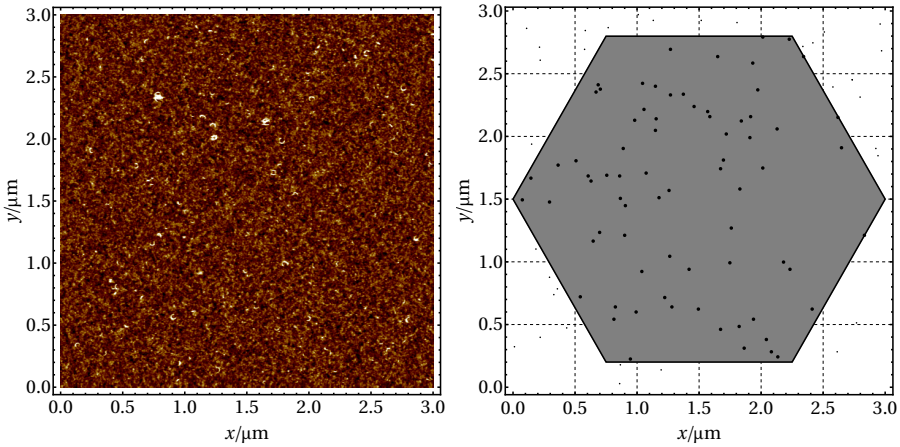


Figure 14: Example of the membrane sample containing no clearly distinguishable defect clusters. Left: AFM image of the membrane. Right: annotated defects in the hexagonal modeling domain.

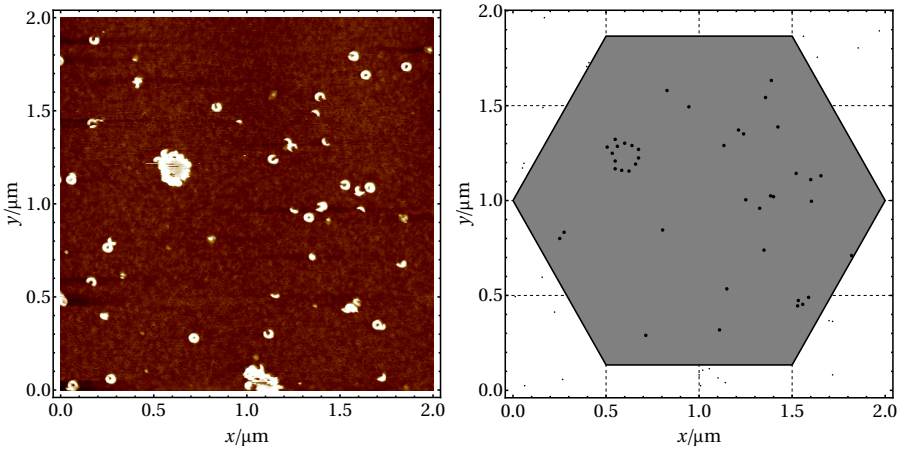


Figure 15: Example of the membrane sample containing defect clusters. Left: AFM image of the membrane. Right: annotated defects in the hexagonal modeling domain.

The comparison of EIS spectra obtained from the experimental and computer-generated random defect sets is presented in Figure 16. Spectra of non-clustered defect distribution indicate a good match with corresponding random cases at all r_{def} levels, where $\log f_{min}$ shift does not exceed 0.02 and $\arg Y_{min}$ difference ranges from 0.37 to 0.93 (comparing with the averaged

curves of random cases). However, an apparent mismatch between actual AFM-registered defect coordinates and the random defect distribution model is visible for the clustered case, where $\log f_{min}$ is shifted towards low frequencies by 0.17 to 0.30 and $\arg Y_{min}$ difference varies from 1.35 to -1.21 as r_{def} increases.

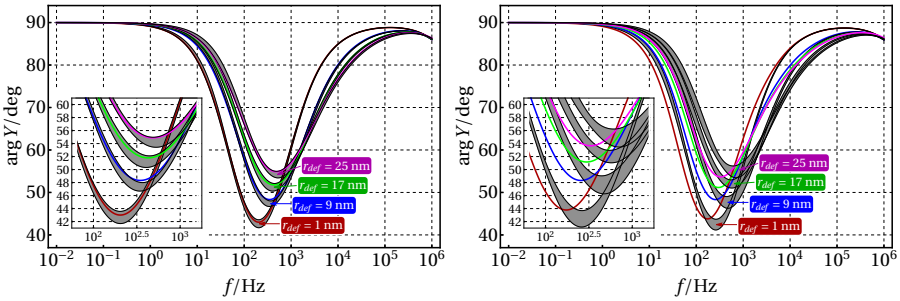


Figure 16: Modeled EIS spectra of computer-generated random defect distributions (gray bands) and experimentally registered distributions obtained from AFM images (colored curves). Left: non-clustered defect distribution. Right: clustered defect distribution

1.5.4. Membrane parameter prediction from EIS spectra

The dataset for the membrane parameter prediction experiments was collected by modeling EIS spectra with various combinations of N_{def}, r_{def} and ρ_{sub} parameters. In each case the coordinates of a fixed number of defects ($N = 200$) were sampled using the random defect distribution model. A total of 10 such cases were generated for each of 546 unique parameter combinations (Table 5), resulting in 5460 distinct models. Finite element modeling was then performed for each instance and EIS spectra were computed. This dataset was therefore used to compare different regression models and assess their prediction accuracy.

Table 5: Parameter values used for EIS dataset.

Parameter	Scale	Min	Max	Values
N_{def}	Log	10^{-1}	10^{2}	13
r_{def}	Linear	1	25	7
ρ_{sub}	Log	10^{4}	10^{5}	6

Each model was evaluated by performing 10-fold cross-validation on the described dataset. Treating each of 546 unique parameter combinations and its corresponding 10 model instances as a cohesive group of examples, training and validation sets for each fold were collected in such a way that examples of any given group would not be split among both sets. In each experiment two separate regression models were evaluated in predicting v_1 and v_2 coefficients, which represent a unique EIS spectrum (coefficients described in detail in subsection 1.3.4). The model accuracy was assessed in terms of the coefficient of determination (R^2), by computing the values for each fold and then aggregating them by calculating the mean and the standard deviation.

Table 6 shows the performance of linear regression models trained using two input features ($\log f_{min}$ and $\arg Y_{min}$) and their additional polynomial terms (up to the 3rd degree). The prediction accuracy of v_1 is very high regardless of the specific feature set being used, indicating a clear linear dependency between the minimum point of the EIS spectrum and the coefficient. However, v_2 predictions are less accurate, especially if only the two initial features are used. Adding 2nd or 3rd-degree polynomial features to the model results in a significant improvement and relatively high overall accuracy.

Table 6: Cross-validation results of linear regression models with polynomial features.

Poly. degree	Features	v_1		v_2	
		Mean	Stdev	Mean	Stdev
1	2	0.999	0.000	0.564	0.040
2	5	0.999	0.000	0.829	0.019
3	9	0.999	0.000	0.880	0.013

Lasso regression models were evaluated for the same prediction task as well, with the objective of simplifying the linear regression models by removing some less important features without a significant decrease in model accuracy. The initial feature set consisted of $\log f_{min}$ and $\arg Y_{min}$ values plus polynomial features of up to 3rd degree, comprising a total set of 9 input features. Regularization parameter λ was adjusted on a logarithmic scale from 10^{-2} to 10^2. Table 7 shows the cross-validation scores (mean and standard deviation values) and the counts of non-zero model coefficients at each λ value. The prediction accuracy of v_1 coefficient is not significantly influenced by λ and

this model does not show any clear benefits compared to its simpler alternative presented in Table 6. The results of v_2 model display a more interesting trend where the accuracy remains almost constant and relatively high for λ values below 10^0 while the number of non-zero coefficients of the model is decreased to 3. This suggests that the Lasso regression model with $\lambda = 10^0$ could be useful for v_2 prediction as a simpler alternative to previously examined linear regression models at the expense of slightly lower accuracy.

Table 7: Cross-validation results of Lasso regression models with varying regularization parameter values.

λ	v_1			v_2		
	Mean	SD	Coefs.	Mean	SD	Coefs.
$10^{-2.0}$	0.999	0.000	5	0.804	0.024	6
$10^{-1.5}$	0.999	0.000	5	0.801	0.025	5
$10^{-1.0}$	0.997	0.001	5	0.801	0.026	3
$10^{-0.5}$	0.984	0.005	4	0.801	0.026	3
$10^{0.0}$	0.982	0.005	4	0.798	0.025	3
$10^{0.5}$	0.968	0.005	2	0.773	0.025	3
$10^{1.0}$	0.968	0.005	2	0.533	0.043	3
$10^{1.5}$	0.967	0.004	2	0.468	0.037	2
$10^{2.0}$	0.959	0.004	2	0.446	0.028	2

While the described linear regression and lasso regression models show adequate performance in predicting v_2 values based on EIS minimum point coordinates, there remains a possibility that even better results could be achieved by using the whole EIS spectrum instead. To test this assumption, PCR and PLS regression models were evaluated by the same cross-validation procedure but using the full frequency range of the EIS response (consisting of 81 points) as the input feature set. The number of components for both model types was adjusted from 4 to 20 with the increments of 4. Tables 8 and 9 show the results of PCR and PLS models respectively. While v_1 prediction accuracy remains high in all cases, v_2 prediction is much worse compared to linear models discussed earlier with the mean R^2 scores not exceeding 0.26. Increasing the number of components does not indicate any significant improving trend for either of the models.

Table 8: Cross-validation results of the PCR model.

Components	v_1		v_2	
	Mean	Stdev	Mean	Stdev
4	0.983	0.005	0.083	0.045
8	0.994	0.001	0.210	0.043
12	0.995	0.001	0.259	0.091
16	0.995	0.001	0.249	0.098
20	0.995	0.001	0.246	0.104

Table 9: Cross-validation results of the PLS regression model.

Components	v_1		v_2	
	Mean	Stdev	Mean	Stdev
4	0.992	0.003	0.151	0.056
8	0.995	0.001	0.254	0.102
12	0.995	0.001	0.247	0.098
16	0.995	0.001	0.238	0.108
20	0.995	0.001	0.224	0.121

K-nearest neighbors regression was also evaluated as a non-linear alternative to PCR and PLS methods. The model was cross-validated in the same way as the aforementioned linear models by using all EIS spectral values as the input features. The experiment was repeated with several different settings of the number of neighbors and the Euclidean distance metric was used to measure similarity between examples. The predictions of v_1 and v_2 were computed by weighing all neighbourhood points equally. As indicated in Table 10, the KNN regression model showed much better performance in predicting v_2 compared to PCR or PLS models, although still somewhat worse than the linear regression models examined earlier. Mean prediction accuracy did not depend much on the chosen number of neighbors, while the standard deviations of R^2 scores show a slight decreasing trend, suggesting that model variance gets lower.

Table 10: Cross-validation results of the K-nearest neighbors regression

Neighbors	v_1		v_2	
	Mean	SD	Mean	SD
2	0.999	0.000	0.700	0.058
5	0.999	0.000	0.725	0.050
10	0.999	0.000	0.733	0.042
20	0.999	0.000	0.729	0.031

Although the described experiments illustrate how accurately different regression models can predict v_1 and v_2 coefficients, it does not reveal the prediction accuracy of the actual membrane parameters, which are derived from these coefficients. In order to estimate it for the best-performing linear regression models described in Table 6 their cross-validation with the modeled dataset was performed with the additional step of reconstructing membrane parameters (N_{def}, r_{def}, ρ_{sub}) using the predictions of v_1, v_2 and the actual known values of one of the three listed membrane parameters. The v_1 is predicted with a linear regression model using $\log f_{min}$ and $\arg Y_{min}$ as the inputs and no additional polynomial features, while the v_2 model used 3rd-degree polynomial features. Mean absolute error (MAE) and mean absolute percentage error (MAPE) were used in place of the coefficient of determination to provide more interpretable estimates. Although both N_{def} and ρ_{sub} parameters are preferred to be expressed in logarithmic scale, this is only performed for ρ_{sub}, as some of the $\log N_{def}$ values used in modeling were equal to 0 (Table 5), rendering MAPE estimates invalid in such cases.

Table 11 shows the MAE and MAPE values for each membrane parameter, where one of the remaining two parameters is assumed to be known. N_{def} and $\log \rho_{sub}$ predictions strongly depend on which of the other two parameters is fixed - selecting r_{def} in both cases results in significantly less accurate estimates in terms of relative error, while using ρ_{sub} to predict N_{def} and vice versa gives the best results. The prediction accuracy of r_{def} remains the lowest among all three parameters, where choosing between N_{def} and $\log \rho_{sub}$ as the fixed parameter makes no significant impact.

Table 11: Membrane parameter prediction with linear regression models using full feature sets.

Fixed parameter	N_{def}		r_{def}		$\log \rho_{sub}$	
	MAE	MAPE	MAE	MAPE	MAE	MAPE
N_{def}	-	-	2.607	28.320	0.034	0.762
r_{def}	6.028	77.580	-	-	0.246	5.507
ρ_{sub}	1.349	7.842	2.965	32.184	-	-

As was shown in Table 7, linear regression models can be simplified to use fewer input features at the expense of a relatively small reduction in their predictive accuracy. Equations (36) and (37) show such simplified models, where the v_1 equation represents the linear regression model with no additional polynomial features (Table 6) and the v_2 equation is expressed from the lasso regression model with $\lambda = 1$ (Table 7).

$$v_1 \approx 0.669 \log f_{min} - 0.005 \arg Y_{min} - 3.88; \tag{36}$$

$$\begin{aligned} v_2 \approx 5.427 \times 10^{-3} (\arg Y_{min})^2 - 1.448 \times 10^{-4} \log f_{min} (\arg Y_{min})^2 \\ -5.164 \times 10^{-5} (\arg Y_{min})^3 - 3.248. \end{aligned} \tag{37}$$

Table 12 shows the parameter prediction accuracy where the simplified models were used to predict v_1 and v_2 coefficients. As expected, the MAE and MAPE estimates are lower in all cases compared to the original results (Table 11). However, the reduction in accuracy is most significant in predicting r_{def}, while the MAPE value of N_{def} predictions (given that the exact ρ_{sub} value is known) increased by just 1% and the impact on $\log \rho_{sub}$ predictions is almost negligible.

Table 12: Membrane parameter prediction with simplified regression models.

Fixed parameter	N_{def}		r_{def}		$\log \rho_{sub}$	
	MAE	MAPE	MAE	MAPE	MAE	MAPE
N_{def}	-	-	3.731	42.100	0.040	0.890
r_{def}	9.613	135.245	-	-	0.332	7.405
ρ_{sub}	1.623	8.713	4.130	48.072	-	-

1.5.5. EIS spectral feature prediction from membrane parameters

In addition to the membrane parameter estimation from EIS spectra, a reverse task of predicting EIS spectral features ($\log f_{min}$ and $\arg Y_{min}$) from membrane parameters by regression models was attempted as well. Such models can be useful as a faster method to estimate the most informative properties of EIS spectra, compared to the time-consuming process of modeling the entire EIS spectrum with FEA. Linear regression models were fitted to the modeled dataset described in subsection 1.5.4. Table 13 shows the cross-validation results in terms of MAE and MAPE errors. The results indicate that simple linear models can adequately describe the relationship between the three listed membrane parameters and EIS spectral features. Adding 2nd or 3rd-degree polynomial features results in some error reduction, more significant for the $\arg Y_{min}$ model, compared to $\log f_{min}$.

Table 13: Performance of linear regression models with polynomial features

Poly. degree	Model coefs.	$\log f_{min}$		$\arg Y_{min}$	
		MAE	MAPE	MAE	MAPE
1	3	0.041	2.481	2.019	4.238
2	9	0.024	1.287	0.869	1.849
3	19	0.020	1.042	0.599	1.285

The equations (38) and (39) represent the fitted linear regression models for the prediction of both spectral features. The position of the minimum point in the frequency axis (expressed in $\log f_{min}$) strongly depends on $\log N_{def}$ and $\log \rho_{sub}$, while r_{def} has relatively little effect. The significance of $\log N_{def}$ is also evident in the regression equation of $\arg Y_{min}$, with r_{def} being the next important feature and $\log \rho_{sub}$ having the least influence.

$$\log f_{min} \approx 1.127 \log N_{def} + 0.013\ r_{def} - 0.997 \log \rho_{sub} + 6.139; \tag{38}$$

$$\arg Y_{min} \approx 4.603 \log N_{def} + 0.390\ r_{def} + 0.092 \log \rho_{sub} + 39.279. \tag{39}$$

1.5.6. Membrane parameter prediction from experimental EIS spectra

The described methodology for predicting quantitative membrane properties was validated using experimental EIS data. The dataset (Figure 17) was obtained in experimental conditions where the assembled tBLMs were exposed to a solution containing pore-forming toxin vaginolysin (VLY) [46] which induced defects in the membrane samples. EIS measurements were taken using an electrochemical workstation (Zennium, Zahner-Elektrik) at a frequency range from 0.1 Hz to 100 kHz, with 10 logarithmically distributed measurement points per decade. EIS spectra were registered at different periods after VLY exposure, ranging from 2 to 120 minutes. The experiment was conducted by Dr. Tadas Penkauskas, researcher at the Department of Bioelectrochemistry and Biospectroscopy, Life Sciences Center, Vilnius University. The true values of N_{def}, r_{def} or ρ_{sub} were not known due to the nature of the EIS measurement technique and the specific experimental setting in which no additional methods were used to independently measure these properties.

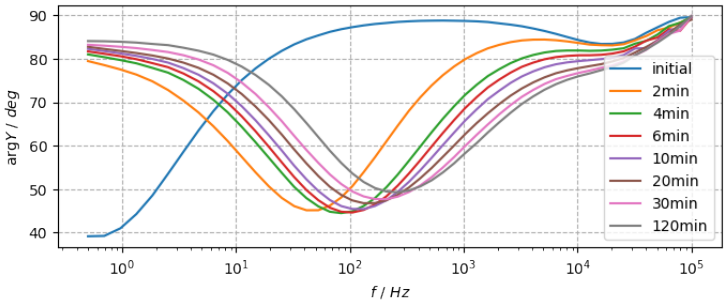


Figure 17: Experimental EIS data measured for tBLM sample at different times (listed in the legend) after its exposure to the pore-forming toxin.

For prediction of the quantitative membrane properties from experimental EIS spectra the regression models were trained using the model dataset described in subsection 1.5.4 and the specific models were selected according to their prediction accuracy presented earlier. The model for predicting the v_1 coefficient was a linear regression model using $\log f_{min}$ and $\arg Y_{min}$ as the input features, while the v_2 prediction model was also a linear regression model using both the listed two quantities and their 3rd-degree polynomial features. The predicted values of v_1 and v_2 were then used to compute the estimates of N_{def} and r_{def} while keeping the value of the remaining parameter ρ_{sub} fixed at $10^5\ \Omega \cdot cm$. Table 14 shows the estimated parameter values for each experi-

mental EIS spectrum (excluding the initial measurement).

Table 14: Membrane parameter estimates predicted from experimental EIS spectra by using linear regression models.

Time (min.)	N_{def}	r_{def}
2	2.019	10.795
4	3.696	6.300
6	4.482	6.036
10	4.990	7.750
20	5.458	11.956
30	6.601	14.225
120	7.863	19.305

Although the predicted values cannot be directly compared with the true ones, some trends in the data may suggest the validity of the applied methodology. N_{def} estimates display a monotonic increase which corresponds to the experimental conditions of membrane damage accumulating over time, due to the prolonged contact with the pore-forming toxin. The range of N_{def} predictions is on the same order of magnitude as the estimates presented in similar studies [71] and the values determined from experimental AFM data, as described in subsection 1.5.3. The r_{def} values show the pattern of an initial decrease followed by an increase - this can be attributed to a complex process of defect formation and varying amounts of both complete and incomplete pores present on the membrane surface at different times [53]. The highest predicted r_{def} value of 19.3 is also in agreement with the approximate maximum pore size of 25 nm (induced by a toxin similar to VLY), as described in another study [22].

1.6. Conclusions

- The three-dimensional membrane model implemented using FEA is capable of simulating EIS responses qualitatively similar to the ones obtained by using the analytical model [45], assuming the regular defect distribution. The versatility of the model relative to the previous approach is that it supports defect sets with arbitrary defect positions, counts and sizes.

- The series of EIS modeling tasks performed with the random defect distribution model, varying defect counts N and multiple levels of mesh density provided practical guidelines for choosing the suitable modeling parameters according to the available computing resources or the solution accuracy requirements.

- Comparison of EIS spectra modeled by using the AFM-recorded non-clustered defect set and randomly generated defect sets of equivalent density showed no significant differences. However, a similar comparison involving AFM-recorded clustered defect set demonstrated a clear discrepancy against equivalent random defect sets. This indicates that the random defect distribution model cannot adequately describe clustered defect distributions and the clustering effect causes quantitative changes in the EIS spectrum.

- A novel methodology of EIS spectral data analysis based on machine learning methods has been presented and evaluated using modeled EIS data. The methodology enables the estimation of defect density N_{def}, defect size r_{def} and submembrane specific resistance ρ_{sub} assuming that one of the three parameters is known. A comparison of several regression approaches showed that the linear regression models using the polynomial combinations of EIS minimum point coordinates $\log f_{min}$ and $\arg Y_{min}$ were the most accurate in predicting the aforementioned membrane parameters.

- The preliminary application of the developed membrane parameter prediction methodology on the experimental EIS data demonstrated qualitatively valid results which are in agreement with the experimental settings and within the ranges of expected values.

2. Defect clustering models

2.1. Clustering evaluation methods

Assuming that all defects distributed on the membrane surface have identical physical properties and dimensions, they can be characterized as points on a 2D plane. As the clustering effect is expressed more by the spatial relationship between a group of points rather than their exact coordinates, a useful tool for estimating it is the Voronoi diagram. This is a well-known computational geometry concept that has applications in a variety of different fields [15]. A Voronoi diagram is computed for a finite set of points on a 2D plane, where the plane is subdivided into regions assigned to each point and enclosing the area of the plane closest to the corresponding point. Membrane model planes (hexagonal or rectangular) containing defects are partitioned into such regions accordingly (Figure 18, left). Isolated defects correspond to larger regions, while closely packed groups (clusters) of defects produce a number of smaller regions, thus reflecting the defect clustering effect and enabling its quantification.

In order to estimate and compare the clustering strength for different defect sets, possibly having different amounts of defects, histograms of Voronoi diagram sector areas are used (Fig. 18, right). They are computed using a fixed number of bins with equal widths from the sector areas scaled with respect to defect density N_{def}. Such an approach enables direct analysis and comparison of relative sector areas across different defect distribution cases, although at the expense of omitting some spatial properties of each defect, such as the number of its neighboring defects.

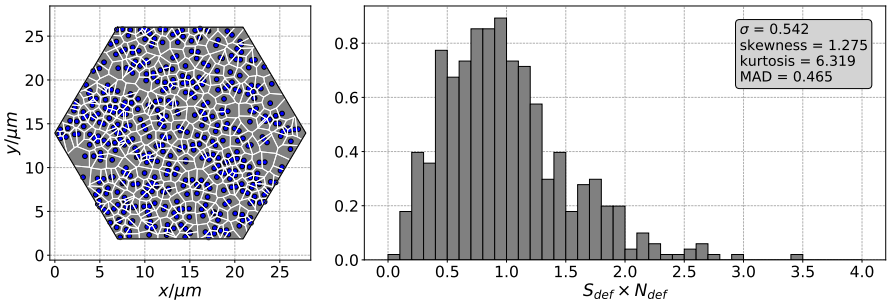


Figure 18: Example of computer-generated random defect distribution. Left: Voronoi diagram of defects distributed over the hexagonal modeling domain. Right: histogram of corresponding Voronoi sector areas.

As the probability distribution of Voronoi sector areas is unknown, statistical moments can be used to describe the general quantitative properties of these distributions, without assuming normality. The first four moments are mean, standard deviation, skewness and kurtosis. In addition, the median absolute deviation (MAD) is used as a more robust (in terms of outliers) alternative to standard deviation. As the sector areas are normalized with respect to defect density, their mean is always equal to 1, so this measure is discarded. The metrics are defined as follows:

- Standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(x_i - \mu)^2}. \tag{40}$$

- Median absolute deviation (MAD):

$$MAD = \text{median}(|x_i - \tilde{x}|). \tag{41}$$

- Skewness:

$$Skew[X] = \frac{1}{N}\sum_{i=0}^{N}\frac{(x_i - \mu)^3}{\sigma^3}. \tag{42}$$

- Kurtosis:

$$Kurt[X] = \frac{1}{N}\sum_{i=0}^{N}\frac{(x_i - \mu)^4}{\sigma^4}. \tag{43}$$

Here $X = (x_i, ..., x_N)$ denotes the vector of Voronoi sector areas, N - defect count, μ - area mean, $\tilde{x}$ - area median.

In order to measure the similarity between two defect sets, without making any statistical assumptions about the distribution of their Voronoi sector area values, the Earth mover's distance (EMD) metric was selected [16]. This measure represents the minimum cost required to transform one probability distribution to the other and will be used to compare the histograms of Voronoi sector areas of a given pair of defect sets.

2.2. Clustering models

2.2.1. Attraction model

The first method of generating clustered defect sets is based on the assumption that defects naturally attract one another and thus tend to cluster together. Such

type of object interaction is fundamental and common in nature (i.e. gravitational and electromagnetic forces) and also applicable in biological membrane models [43]. In this model attraction takes effect if the distance between two defects is below the predefined threshold d_T, which can be expressed in one of these two ways:

- Number of defect radiuses (of the attracting defect).
- Fixed distance in nanometers.

Generating a clustered defect distribution involves the following steps:

1. Coordinates of the first defect are picked randomly from uniform distribution.
2. For each of the subsequent defects:
   (a) Initial coordinates for the current defect with radius r_c are selected randomly from a uniform distribution.
   (b) Closest existing defect is selected and designated as the attractor with radius r_a.
   (c) Distance between the current and attractor defects is calculated.
   (d) If the distance is below the predefined threshold d_T and above the minimum distance of $1.5 \times (r_c + r_a)$, the current defect is shifted towards the attractor defect. The minimum distance is retained to avoid defect overlapping.
   (e) Otherwise, if the distance is below the minimum distance of $1.5 \times (r_c + r_a)$, the current defect is shifted away from the attractor defect until the distance between their centers matches the minimum distance.
   (f) If the updated coordinates of the current defect fall outside the hexagon area, the defect is discarded.

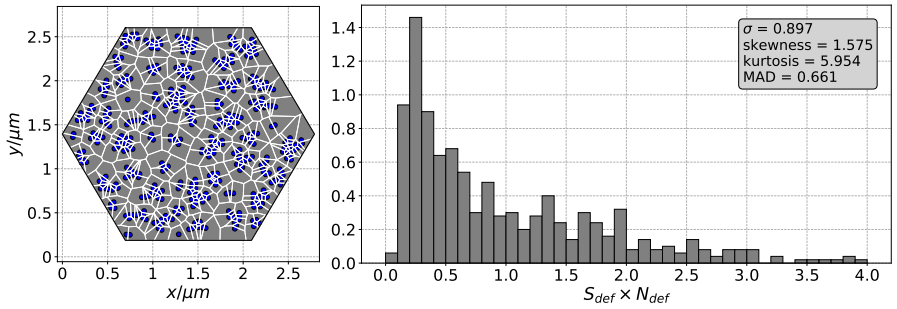


Figure 19: Example of synthetic defect distribution generated using the attraction model, where $d_T = 15$ (expressed in defect radii).

Assuming that d_T is expressed in defect radii, the clustering model has three parameters directly influencing the clustering effect (thus excluding the defect count): defect density N_{def}, defect radius r_{def} and attraction threshold d_T. Figure 19 shows one example of generated defect set, obtained by the described model. The distribution consists of 500 defects with equal radius (13nm), dispersed with a density of 100 defects per square micrometer. Such cases are characterized by tightly packed defect groups all containing a similar number of defects. This is reflected in the sector area histogram where the clustered defects represent a large number of small Voronoi sectors, in contrast with random distributions (Fig. 18).

2.2.2. LCN model

This model is inspired by the idea that membrane defect clusters tend to form complex structures of varying size and shape, visually resembling clouds. This concept is relevant in computer graphics where various algorithms are used to procedurally generate cloud or smoke textures. For the implementation of this model we chose the *lattice convolutional noise* (LCN) algorithm [20] and extended it by introducing two additional parameters by which the clustering effects are adjusted:

- Average relative cluster size: S,
- Minimal probability of defect appearance: P.

The parameter S is a positive real number which adjusts the scaling of the LCN-generated initial image - smaller values correspond to a larger amount of small clusters. P is selected from $[0,1]$ interval and represents the lower

bound of probability field values. Defect distribution generation consists of the following steps:

1. By using the LCN algorithm a probability field of fixed resolution (i.e. 4096×4096) is generated and clipped by the hexagonal model domain shape (Fig. 20). Each pixel in the field corresponds to the probability p_i of a defect appearing at that point.

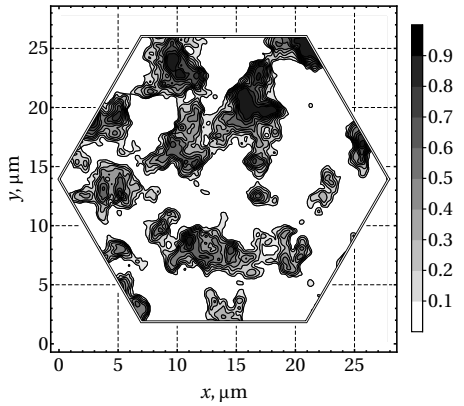


Figure 20: Example of probability field generated with the LCN algorithm.

2. Probability sum S_N is calculated for the field consisting of N points. Interval $[0; S_N]$ is divided into N subintervals, where each corresponds to the probability p_i of the respective pixel (Fig. 21).

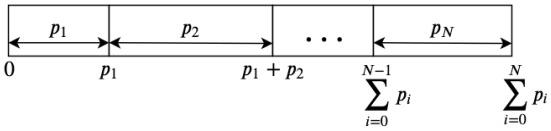


Figure 21: Weighted roulette wheel selection of probability field pixels

3. For each defect a random number is uniformly sampled from the interval $[0; S_N]$ and the corresponding pixel of the probability field is designated as the center of that defect.

This model produces clustered defect distributions (Fig. 22) which are visually distinct from the ones obtained by applying the attraction model (Fig. 19). Clusters exhibit different sizes and various irregularities which are also reflected by statistical properties of Voronoi sector areas, where the majority of small sectors are offset by a number of very large ones.

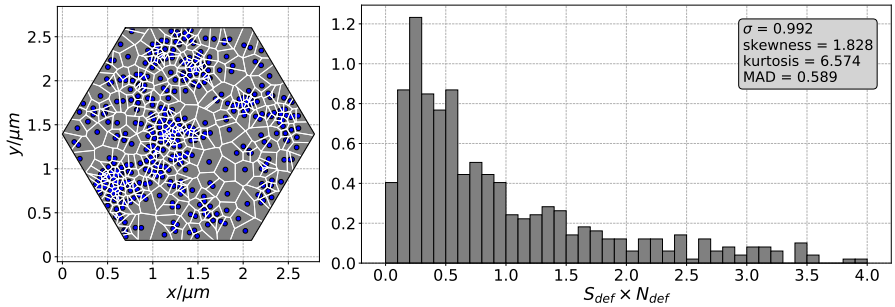


Figure 22: Example of synthetic defect distribution generated using the LCN model, where $S = 1$ and $P = 0.1$.

2.2.3. Point process model

One more approach for modeling the clustered placement of defects on the membrane surface is based on the theory of spatial point processes. Such models describe the random patterns of 2D or 3D points representing objects in a spatial context and are commonly used in various research areas, such as epidemiology, ecology, astronomy, geography and others [21, 56]. The specific model chosen as the basis for clustered defect set generation is the Thomas cluster point process, which extends the more general Poisson point process. The Thomas cluster process generates a random number of parent points (cluster centers), each of which is assigned a random number of offspring points (cluster members) randomly displaced from the center by a vector sampled from isotropic Gaussian distribution (with the same scale for each axis). Thus, the process is controlled by three parameters:

- Parent point rate κ
- Cluster scale r
- Offspring point rate α

The Thomas cluster point process will be referred to simply as the point process model further in this thesis. The algorithm for generating an instance of the point process model (a clustered defect set with density N_{def}) is defined as follows:

1. Number of parent points (cluster centers) is sampled from a Poisson distribution with the rate parameter κ.

2. Coordinates of all parent points are sampled from the uniform distribution with the interval $[-4r, 1+4r]$, corresponding to an extended simulation window (unit square $\Omega \in [0,1]^2$ with additional margins depending on r).

3. For each parent point:

   (a) Number of offspring points (belonging to a cluster) is sampled from a Poisson distribution with the rate parameter α.

   (b) Coordinates of the offspring points related to the current parent point are sampled from normal distributions $\mathcal{N}(x_i, r^2)$ and $\mathcal{N}(y_i, r^2)$, where x_i and y_i are the parent point coordinates.

4. Offspring points belonging to the original simulation window (unit square) are retained, while the ones outside it and the parent points are discarded.

5. The unit square and the point coordinates are scaled to match the specified defect density N_{def}.

Contrary to the previously described attraction and LCN models, the point process model does not generate defect sets with an exact defect count N, although this quantity is influenced by the parameters κ and α. Figure 23 shows an example of a defect set generated using this model.

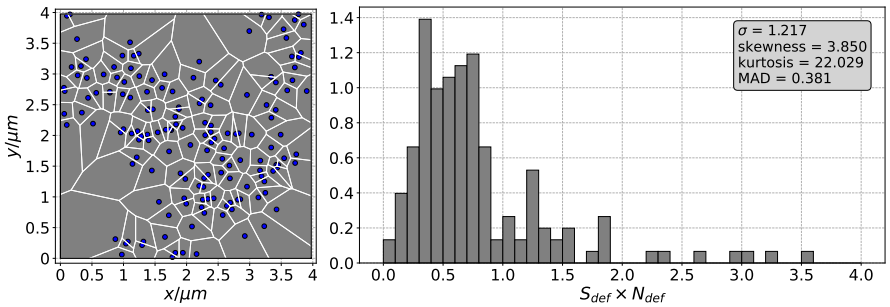


Figure 23: Example of a clustered defect distribution generated using the point process model, where $\kappa = 10$, $\sigma = 0.1$ and $\alpha = 10$.

A distinct advantage of the point process model over the attraction and LCN models is the possibility to infer model parameters directly from data. One well-known method is the minimum contrast [3], although other methods

have been developed as well [24, 60]. The minimum contrast method uses the so-called K-function [23] as a statistic of the clustering effect in point process data and finds the model parameters which result in the lowest discrepancy between the empirical K-function values directly computed from data and the theoretical K-function of the chosen spatial point process model. The empirical K-function represents the standardised average number of neighbours around a point within the t radius and is defined as follows [56]:

$$\widehat{K}(t) = \frac{|W|}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathbf{1}\{d_{ij} \leq t\} e_{ij}(t). \tag{44}$$

Here n is the number of points, d_{ij} is a pairwise distance between two points, $|W|$ is the area of the observation window and e_{ij} is the correction weight to compensate for edge effects. The theoretical K-function of the Thomas cluster point process is the following:

$$K(t) = \pi t^2 + \frac{1}{\kappa}\left(1 - \exp\left(\frac{-t^2}{4r^2}\right)\right). \tag{45}$$

The model parameters are then chosen to minimize:

$$\int_a^b |K(t) - \widehat{K}(t)|^p \mathrm{d}t, \tag{46}$$

with chosen $0 \leq a \leq b$ and $p > 0$.

2.3. Clustering model comparison

The described defect clustering models were first evaluated with respect to the random defect distribution model. For that purpose 100 instances of random defect distributions were independently generated, each consisting of $N = 500$ defects at $N_{def} = 10$ defect density. Voronoi diagrams were then computed for each defect set and Voronoi sector areas were normalized by the defect density, to enable a valid comparison with other defect sets, exhibiting different defect counts or densities. Figure 24 shows the histogram of all normalized Voronoi sector areas and the statistical properties of their distribution. By attempting to fit various well-known probability distributions to this dataset the Gamma distribution was determined to be the best fit (parameters listed in Figure 24 legend).

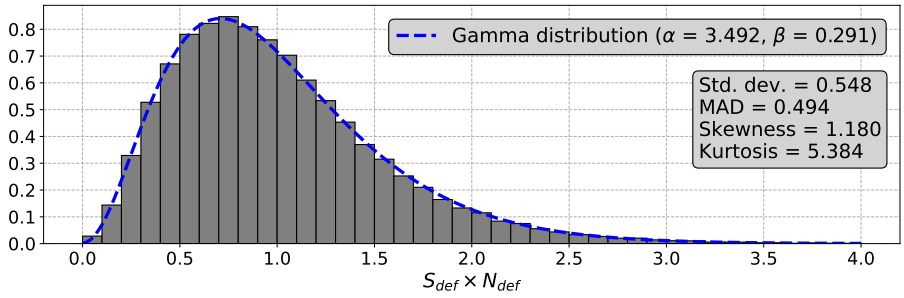


Figure 24: Histogram of normalized Voronoi sector areas of 100 independently generated random defect sets.

To compare the random and clustered defect distributions, a number of synthetic defect sets were generated by applying the described clustering models with different parameter combinations (Table 15). A total of 54, 48 and 60 combinations for attraction, LCN and point process models respectively were examined and summary statistics were computed for 100 independently generated cases for each option. In the case of attraction and LCN models each defect distribution instance consisted of 500 defects. Parameters of all models were selected to cover a wide range of visually different clustering cases, ranging from instances practically identical to random distributions to a strongly expressed clustering effect. Sector areas in all cases were normalized with respect to defect density.

Table 15: Clustering model parameter values used in synthetic defect set generation.

Clustering model	Model parameter	Values
Attraction	Defect density N_{def}	1; 10; 100
	Defect size r_{def} (nm)	0.5; 13; 25.5
	Attraction threshold[1] d_T	5; 10; 15; 20; 25; 30
LCN	Min. probability P	0; 0.03; 0.06; 0.09; 0.12; 0.15
	Cluster size S	0.25; 0.5; 0.75; 1; 1.25; 1.5; 1.75; 2
Point process	Parent rate κ	5; 10; 15
	Cluster scale r	0.03; 0.06; 0.09; 0.12; 0.15
	Offspring rate α	5; 10; 15; 20

[1] Expressed as the number of defect radii (r_{def}).

Summary statistics (described in subsection 2.1) were computed from Voronoi sector areas for each set of 100 defect set instances generated with particular parameter combinations. Figure 25 shows the statistics of each clustering model compared to the random defect distribution model. Although none of the metrics indicate unambiguous separation of the random and all of the clustered defect sets obtained from the three models, the standard deviation is the most indicative of the clustering effect as almost clustered cases (except for several attraction model cases) show values higher than approximately 0.54 (random defect distribution model). Due to this property and intuitive interpretation of this metric, the standard deviation of Voronoi sector areas will be used as a simple quantitative measure of defect clustering effect and will be referred to as the Voronoi σ further in this thesis.

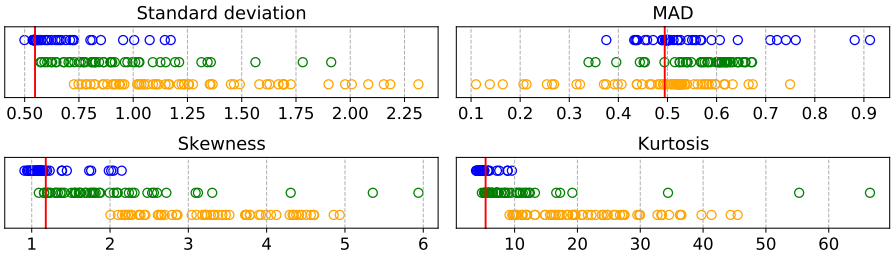


Figure 25: Summary statistics of Voronoi sector areas (averaged histogram of 100 cases) of clustered defect sets (coloured circles) compared against the random defect distribution model (vertical red lines). Attraction, LCN and point process models are represented by blue, green and yellow circles, respectively. Standard deviation is further referred to as Voronoi σ.

For each group of 100 defect sets generated by a specific clustering model and its parameter combination (Table 15) a best-fitting probability distribution was determined by the same approach as described earlier for the random defect distribution model (Figure 24). As shown in Table 16, the selected probability distributions vary depending on clustering model parameters. The majority of attraction model cases can be characterized by gamma, beta or generalized extreme value probability distributions, where gamma distribution fits mostly correspond to parameter combinations with $r_{def} = 0.5$ values. In the case of LCN instances, defect sets with $S < 1$ values tend to be more often characterized by a gamma distribution, while the larger S values mostly correspond to a lognormal distribution. Most of the generated point process model instances are characterized either by lognormal or generalized extreme value distributions.

Table 16: Summary of best-fitting probability distributions determined for clustered defect sets generated by different clustering models with varying parameter combinations.

Prob. distribution	Attraction	LCN	Point process
Beta	15	9	3
Exponential	1	1	2
Gamma	22	11	0
Generalized extreme value	10	5	16
Lognormal	5	18	28
Pareto	1	3	10
Student's t	0	1	1
Total cases	54	48	60

To better understand how each of the clustering model parameters is related to the overall clustering effect in resulting defect sets, Voronoi σ values were plotted against particular parameter values in the following figures. Attraction model results (Fig. 26) show very different Voronoi σ trends for various defect size (r_{def}) and density (N_{def}) values. In the case of small (0.5 nm) defects it is impossible to distinguish the clustered and random distributions. A similar issue applies for medium size (13 nm) defects with low density, although density increase introduces a clear trend of Voronoi σ growth depending on the attraction threshold.

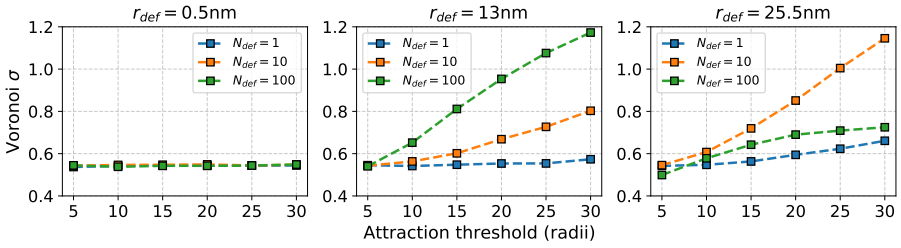


Figure 26: Voronoi σ values (averaged from 100 generated defect sets) of attraction model instances.

An interesting effect of the largest Voronoi σ growth with medium defect density can be observed in the largest (25.5 nm) defect case. This can be explained by the relationship between the defect density N_{def} and the range over

which the attraction effect takes place (defined by r_{def} and d_T). When the attraction range is relatively low concerning the overall dimensions of the model domain (depending on selected N_{def}), defect clusters tend to be numerous and contain small amounts of individual defects, leading to comparatively low variations in Voronoi sector areas and Voronoi σ values being close to the baseline of the random defect distributions. Due to the increase of either the attraction range or the defect density, the clusters tend to accumulate more defects while their overall count decreases and they get more isolated, causing an increase in Voronoi σ. However, when the attraction range approaches the dimensions of the modeling domain, the clusters merge and form irregular structures of non-overlapping defects covering the major part of the model area. This leads to more evenly-sized Voronoi sectors (relative to heavily-clustered instances) and a decrease in Voronoi σ. Such phenomena are also illustrated in Figure 27.

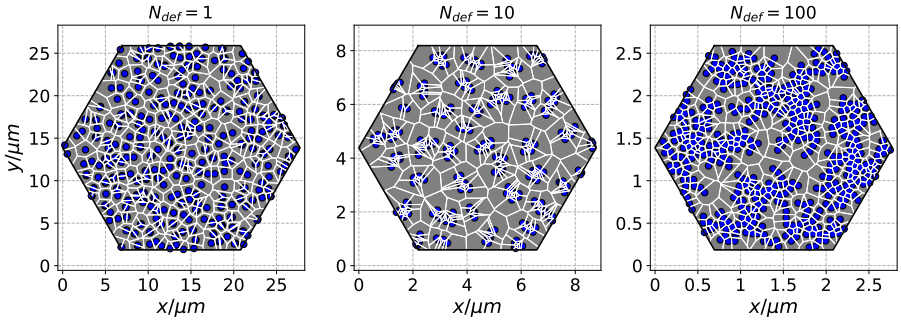


Figure 27: Examples of attraction model instances ($N = 500$; $r_{def} = 25.5\,\text{nm}$, $d_T = 30$) generated with different N_{def} values.

Figure 28 illustrates the dependency of Voronoi σ values by S and P parameters of LCN model. By decreasing the values of both parameters, standard deviation approaches the random distribution baseline, but does not cross that threshold (Voronoi $\sigma \approx 0.54$). Although a clear increasing trend can be observed in all P cases, the specific clustering parameter values for a given clustered distribution cannot be unambiguously estimated just from the standard deviation of its Voronoi sector areas. A similar trend can be observed for the point process model as well (Figure 29).

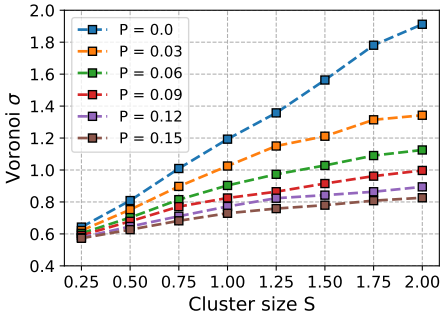


Figure 28: Voronoi σ values (averaged from 100 generated defect sets) of LCN model instances.

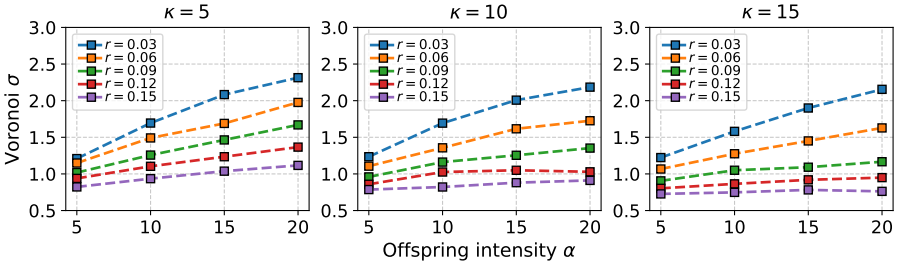


Figure 29: Voronoi σ values (averaged from 100 generated defect sets) of point process model instances.

Dependency between clustering model parameters and Voronoi σ was also evaluated by fitting linear regression models for predicting the aforementioned value. Models were cross-validated (10-fold) using the datasets described in Table 15, where each training example corresponded either to a group of 100 defect set cases generated with the specific clustering parameter combination (Figure 30) or an individual defect set (Figure 31). Models additionally had included 2nd degree polynomial features derived from the original clustering parameter values. The results obtained indicate that such regression models can accurately ($R^2 > 0.94$) describe the dependency for LCN and point process models in the averaged case, although the accuracy drops when individual defect sets are considered (Figure 31). This indicates that Voronoi σ varies significantly among the series of defect sets generated by using the same clustering parameter values.

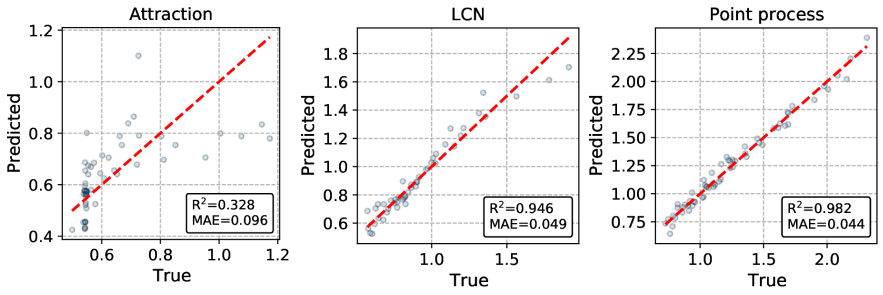


Figure 30: Linear regression model predictions of Voronoi σ by clustering model parameters (averaged instances).

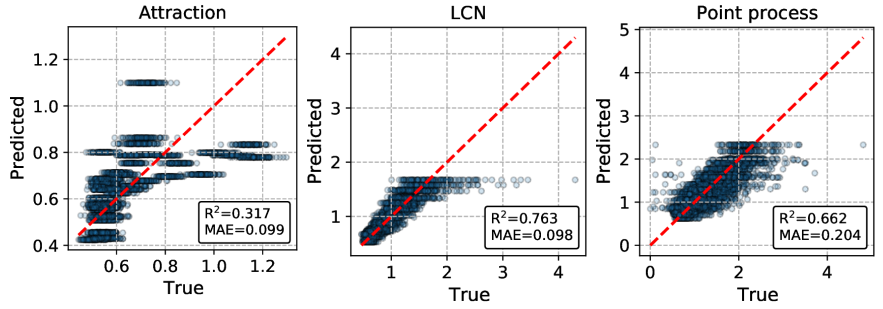


Figure 31: Linear regression model predictions of Voronoi σ by clustering model parameters (individual instances).

2.4. Clustering effect estimation from EIS spectra

2.4.1. EIS spectra of clustered defect distributions

As the methodology of EIS data analysis presented in Chapter 1 was developed solely with the random defect distribution model in mind, it is unclear whether the same approaches would be applicable to the qualitatively different cases of clustered defect distributions. To evaluate and compare predictive models for estimating clustering effect from EIS spectrum, a dataset of EIS spectra was collected by modeling various clustered defect set cases. The point process clustering model was used to generate membrane model instances exhibiting a varying degree of defect clustering effect, which was controlled by adjusting clustering model parameters as well as the defect density and defect size. Defect count was not explicitly defined and varied in the range from 4 to 477, depending on the clustering model parameters. Table 17 lists the individual parameter values, yielding 216 unique parameter combinations. Ten model

instances were independently generated for each combination, resulting in a total of 2160 FEA model cases.

Table 17: Parameters used to produce the modeled EIS dataset of clustered defect distributions.

Parameter	Values
N_{def}	0.1; 1; 10; 100
r_{def}	0.5; 25.5
Parent rate κ	5; 10; 15
Cluster scale r	0.05; 0.1; 0.15
Offspring rate α	5; 10; 15

The initial review of the modeled EIS curves of clustered defect sets revealed some important qualitative differences relative to the EIS data of random defect distributions examined in Chapter 1. This was assessed by the number of minimum and maximum points in the admittance phase curves as well as their first and second-order derivatives (calculated numerically using the forward difference scheme). All EIS spectra of random cases were found to satisfy the following conditions, related to the numbers of the extremum points:

- One minimum point and zero or one maximum points in EIS spectrum (the latter implied by the possible $\arg Y$ decrease in a high-frequency spectral region).
- One minimum and one maximum point in the first derivative of EIS spectrum.
- One or two minima points and one or two maxima points in the second derivative of EIS spectrum.

However, the described properties did not apply to a significant part of the clustered cases. Approximately 1% of them had clearly distinguishable double minima, while an additional 30% had other unusual spectral features reflected by higher counts of extrema in their first and second derivatives, compared to random cases. Figure 32 illustrates these phenomena where two clustered cases exhibit anomalous spectral shapes relative to an example of a random case.

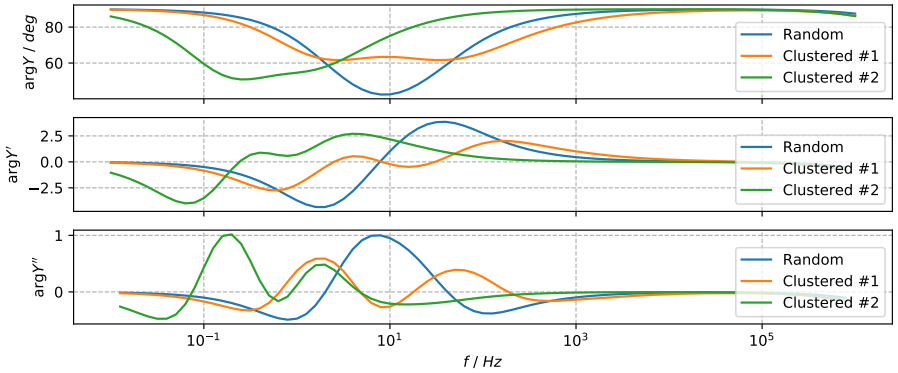


Figure 32: Examples of modeled EIS spectra obtained using random and clustered defect distributions. The top plot displays the raw spectra, middle and bottom plots show the first and second derivatives of the admittance phase, respectively.

Some insight into the occurrence of such spectral anomalies can be gained by partitioning and analyzing the clustered spectra dataset (Table 17) either by defect set properties (density and size) or the parameters of the clustering model. Figure 33 shows the fractions of anomalous spectra in the subsets represented by the specific N_{def} and r_{def} values. Although each subset contains a significant part of anomalies, ranging from 0.17 to 0.44, different trends can be observed for both defect size options, where the fraction for small (0.5 nm) defects grows with the increase of defect density and the opposite can be noticed for large (25.5 nm) defects.

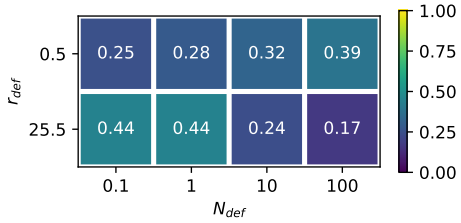


Figure 33: Heatmap of fractions of anomalous spectra in the clustered spectra dataset grouped by defect density and defect size.

Partitioning the dataset by clustering model parameters reveals a stronger effect on the fraction of anomalous spectra, compared to defect density and size parameters. As Figure 34 indicates, the fraction ranges widely from 0.03 to 0.70 depending on the parameter combination. Parent rate κ has a moderate

influence which is most distinct at $\kappa = 5$, while the larger values tend to reduce the effect. The combination of low cluster scale r and high offspring rate α values is more significant in increasing the fraction of anomalous spectra, which can be seen for all κ values. These observations suggest that defect distributions containing a relatively low number of tightly-packed clusters with high numbers of defects are most likely to result in unusual features in EIS spectra. This is also reflected by the mean values of standard deviations of Voronoi sector areas (Figure 35) which correlate well with the aforementioned fractions of anomalous spectra.

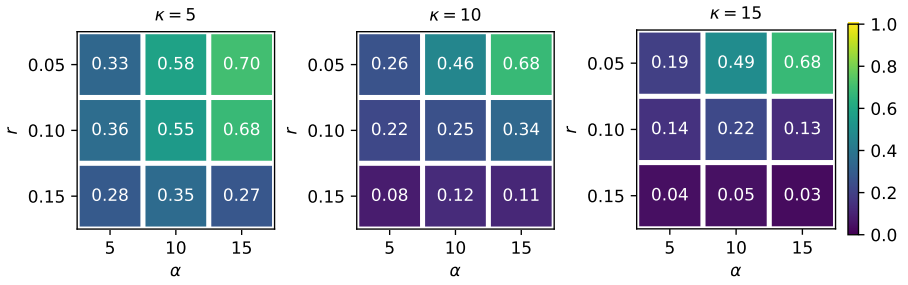


Figure 34: Heatmaps of fractions of anomalies in the clustered spectra dataset grouped by parameters of the point process clustering model.

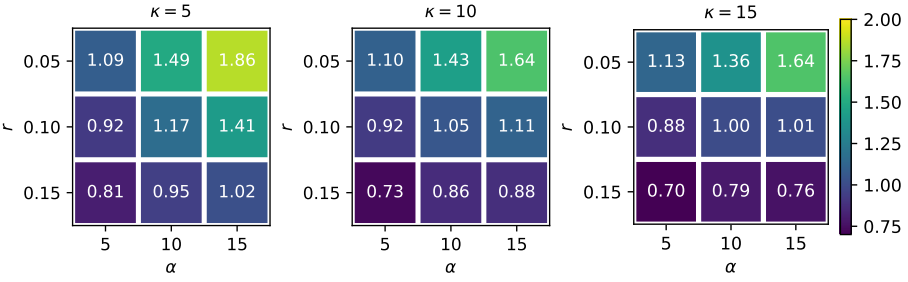


Figure 35: Heatmaps of mean Voronoi σ values in the clustered spectra dataset grouped by parameters of the point process clustering model.

2.4.2. Clustering effect prediction

As discussed earlier, the standard deviation σ of Voronoi sector areas can be used as a simple and interpretable metric for quantifying the clustering effect in a given defect set. While it is trivial to compute for the defect set itself, it is unclear how accurately it could be predicted from an EIS spectrum using a regression model. Ideally, such a model would require no additional knowledge

about the properties of the defect set, such as the defect density or size, and would be based solely on features extracted from EIS spectra.

To evaluate how the minimum EIS spectrum point coordinates are related to Voronoi σ values, $\log f_{min}$ and $\arg Y_{min}$ were determined for each EIS spectrum in the clustered set. In the case of anomalous spectra containing two minima (approx. 1% of the dataset), the point with a lower $\arg Y$ value is selected. Visualization of the data revealed the major influence of Voronoi σ values on the two listed spectral features, when the defect density and size parameters are constant. An example of this effect is presented in Figure 36, where the minimum points of EIS spectra from the clustered set (with varying Voronoi σ values) are compared to the equivalent points obtained using random defect distributions (averaged from 10 independently modeled cases). This observation points toward an alternative approach – extracting spectral features from EIS data by characterizing the peak shape rather than its exact position in the spectrum.

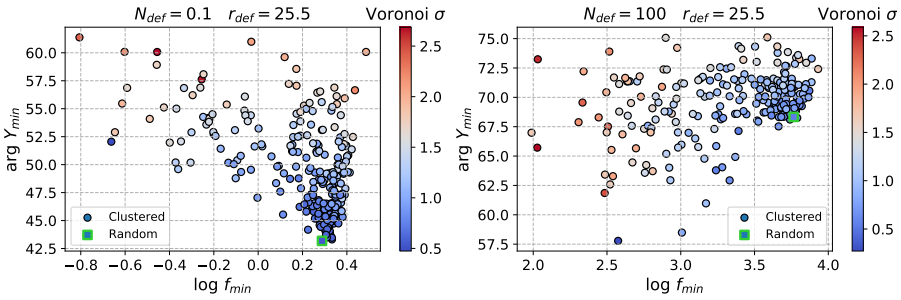


Figure 36: Voronoi σ dependency on $\log f_{min}$ and $\arg Y_{min}$ in the clustered EIS dataset at two defect densities $N_{def} = 0.1$ and $N_{def} = 100$. Green rectangles represent averaged minimum point positions of random defect sets whose EIS spectra were modeled with the same N_{def} and r_{def} values.

The full width at half maximum (FWHM) is a common metric used to characterize peak shapes in various types of spectral data [94, 73]. This measure defines the width of a spectral peak at half of its height from the baseline. While FWHM is commonly computed by fitting the spectrum with some analytical function (i.e. Gaussian), a simplified approach can be used in the case of EIS spectra, assuming that $\arg Y$ values are always in the range from 0 deg to 90 deg. FWHM is then defined as follows:

$$FWHM = \log_{10} f_2 - \log_{10} f_1. \tag{47}$$

Here frequencies f_1 and f_2 represent two points of EIS spectrum (illustrated in Figure 37) where:

$$\arg Y = 90 - \frac{90 - \arg Y_{min}}{2}. \tag{48}$$

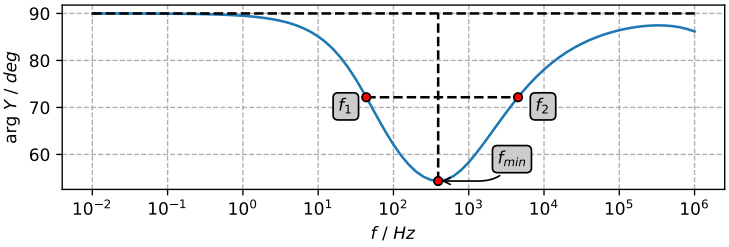


Figure 37: Frequency points in EIS spectrum used to compute FWHM.

Computing FWHM values for EIS spectra in the clustered dataset and comparing them to corresponding Voronoi σ values showed a moderate correlation of 0.62. While this suggests that FWHM can be used as a predictor variable for estimating Voronoi σ, additional features related to the overall peak shape could also be informative, as evident in Figure 32. They can be derived by performing peak fitting, which is a common practice employed in spectral data analysis [94, 79, 28]. Based on the visual evaluation of the peak shapes in EIS spectra, the Gaussian function was chosen, assuming that a higher mismatch between the curve and $\arg Y$ values should correspond to higher Voronoi σ values. Curve fitting is performed only for the spectral region around the minimum point ($\arg Y$ values not exceeding 87 degrees), as illustrated in Figure 38.

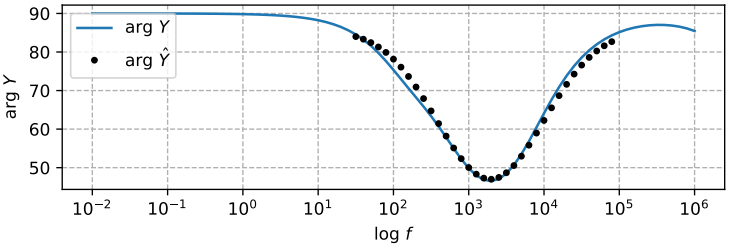


Figure 38: Example of a Gaussian curve fitted to EIS spectrum in the vicinity of the minimum point.

The mismatch between k actual ($\arg Y$) and fitted ($\arg \hat{Y}$) points (indices ranging from t_1 to t_2) can then be summarized by several quantities:

- Mean absolute difference:

$$p_{mean} = \frac{1}{k} \sum_{i=t_1}^{t_2} |\arg Y_i - \arg \hat{Y}_i|. \tag{49}$$

- Standard deviation of absolute differences:

$$p_{std} = \sqrt{\frac{1}{k} \sum_{i=t_1}^{t_2} (\arg \hat{Y}_i - p_{mean})^2}. \tag{50}$$

- Maximum absolute difference:

$$p_{max} = max(|\arg Y_i - \arg \hat{Y}_i|), \quad i = t_1, ..., t_2. \tag{51}$$

In addition to the described features, extrema counts in EIS spectrum or its first and second derivatives (discussed in subsection 2.4.1) are used with the following notation: m_0, m_1 and m_2 denotes the maxima counts in EIS spectrum, its first and second derivatives respectively, while n_0, n_1 and n_2 represent the minima counts in the same way. The full feature set selected for the regression task included FWHM, p_{mean}, p_{std}, p_{max} and the 6 extrema count features, comprising a total of 10 features. A linear regression model with L1 regularization (Lasso) was used to represent the relationship between the 10 listed features and Voronoi σ as well as distinguishing the most informative features. To investigate how the model describes the spectral variations related exclusively to defect clustering, rather than caused by varying N_{def} and r_{def} values, 8-fold cross-validation was performed using the leave-one-group-out approach, by excluding all EIS spectra of a specific N_{def} and r_{def} combination in each fold. Regularization parameter λ was varied from 10^{-3} to 10^{-1}.

Table 18: Lasso regression model performance (in terms of R^2) of predicting Voronoi σ from EIS spectra of clustered defect sets.

λ	Coefs.	Training		Validation	
		Mean	SD	Mean	SD
$10^{-3.0}$	9	0.603	0.024	0.295	0.372
$10^{-2.75}$	9	0.597	0.023	0.346	0.299
$10^{-2.5}$	7	0.582	0.023	0.383	0.244
$10^{-2.25}$	5	0.572	0.020	0.430	0.209
$10^{-2.0}$	6	0.559	0.021	0.449	0.187
$10^{-1.75}$	4	0.518	0.023	0.450	0.153
$10^{-1.5}$	3	0.463	0.023	0.432	0.110
$10^{-1.25}$	3	0.319	0.018	0.309	0.082
$10^{-1.0}$	2	0.176	0.014	0.170	0.071

Table 18 summarizes the training and validation scores (R^2) at each λ level and the counts of non-zero model coefficients. The high variance of the model is evident at lower λ values, while the higher λ leads to the opposite effect of high bias. An acceptable bias-variance tradeoff can be observed at $\lambda = 10^{-1.5}$ with 3 selected features: FWHM, p_{max} and m_2. MAE and MAPE errors of this specific model were equal to 0.22 and 24% respectively. The mediocre overall performance suggests that the engineered spectral features and a linear model cannot fully represent the effect of varying Voronoi σ on the spectrum, while discarding the similar variations caused by N_{def} and r_{def} parameters. To test this claim, a linear regression model using the three selected features (FWHM, p_{max}, m_2) was cross-validated with each of 8 subsets (corresponding to the specific combinations of N_{def} and r_{def}) of the clustered EIS dataset separately, instead of using the full dataset.

Table 19: Cross-validation results of predicting Voronoi σ for subsets of clustered EIS dataset.

N_{def}	r_{def}	Training		Validation	
		Mean	SD	Mean	SD
0.1	0.5	0.800	0.010	0.775	0.100
0.1	25.5	0.824	0.007	0.808	0.070
1	0.5	0.805	0.010	0.785	0.072
1	25.5	0.750	0.009	0.706	0.120
10	0.5	0.799	0.007	0.779	0.046
10	25.5	0.635	0.015	0.574	0.146
100	0.5	0.760	0.018	0.708	0.218
100	25.5	0.274	0.010	0.245	0.118

The results listed in Table 19 indicate significantly better model performance in most cases, except for the subset of $N_{def} = 100$ and $r_{def} = 25.5$. Still, a performance drop can be noticed with an increase in defect density, more evident for large defects (25.5 nm). Figure 39 illustrates the comparison of the cross-validation results of the linear regression model trained using the three selected features (FWHM, p_{max}, m_2) both for the full clustered EIS dataset and its specific subset ($N_{def} = 1$; $r_{def} = 25.5$). In the case of the full dataset, the model tends to overestimate low values of Voronoi σ and underestimate high values.

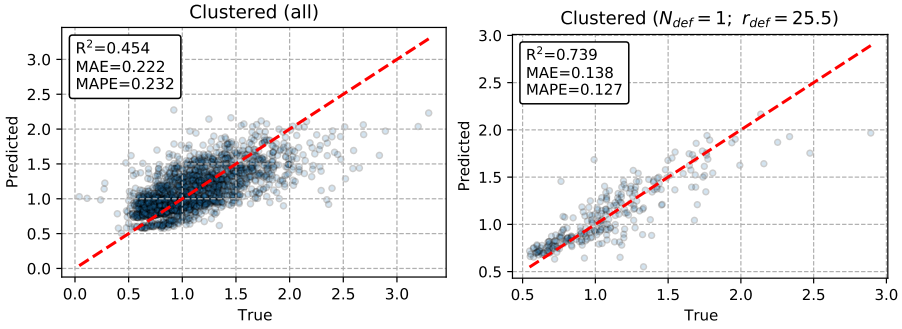


Figure 39: Cross-validation results of Voronoi σ prediction using different sets of EIS spectra. Left: entire clustered EIS dataset. Right: subset of clustered EIS dataset with $N_{def} = 1$ and $r_{def} = 25.5$ nm.

In addition to the regression models trained to predict Voronoi σ value, a

classification experiment was also performed using the clustered EIS dataset and another one, consisting of EIS spectra obtained from random defect distributions (dataset described in subsection 1.5.4). A logistic regression model was used for both datasets to predict whether the spectrum originated from a random defect set (negative class) or clustered one (positive class). The same binary classification task was performed twice, using only the Voronoi σ feature for the reference and the set of three features used in regression experiments previously. 10-fold cross-validation was performed in both cases, where splits were made according to groups representing specific N_{def} and r_{def} combinations. Results are shown in Table 20.

Table 20: Classification results of random/clustered cases using different features.

Features	Accuracy	Precision	Recall	F1
Voronoi σ	0.95	0.98	0.84	0.90
Spectral features	0.90	0.85	0.78	0.82

2.4.3. Defect set parameter prediction

As previously shown, the clustering of defects affects the modeled EIS spectrum not just in terms of its minimum point coordinates ($\log f_{min}$ and $\arg Y_{min}$), but also the overall shape of the curve. This prompts us to re-evaluate the regression models for predicting membrane parameters from EIS spectral features, as presented in subsection 1.5.4. For the reference, linear regression models for predicting v_1 and v_2 coefficients (equations 36 and 37, chapter 1) trained on the random EIS dataset were applied for the clustered dataset to predict the two coefficients. Additionally, the same models were also cross-validated using the clustered EIS dataset only (Table 21). As the modeling of all clustered defect distribution cases was performed with $\rho_{sub} = 10^5$, the same value was used to reconstruct N_{def} and r_{def} based on v_1 and v_2 values predicted by the models.

As Table 21 shows, the prediction accuracy of N_{def} is worse (when the models are trained on the random EIS dataset), but comparable to results obtained with random defect set spectra only (Table 12, Chapter 1), while the r_{def} predictions are extremely inaccurate. This resonates with an earlier observation where Voronoi σ was shown to have a major impact on the position of

the minimum point in the EIS spectrum, especially in terms of $\arg Y_{min}$ (Figure 36). Same models cross-validated on the clustered EIS dataset show even higher errors in predicting N_{def}, although r_{def} predictions are relatively better.

Training dataset	N_{def}		r_{def}	
	MAE	MAPE	MAE	MAPE
Random	12.318	46.823	314.526	14896.688
Clustered	22.195	69.728	12.425	442.098

Table 21: Comparison of the membrane parameter prediction accuracy using the linear regression models for predicting v_1 and v_2 coefficients, trained using random or clustered EIS datasets.

2.5. Clustering model parameter estimation

Although all three presented defect clustering algorithms enable the generation of synthetic defect sets with varying degrees of clustering (adjusted via the model parameters), an important aspect of evaluating the model applicability to experimental AFM data is the comparison of EIS spectra modeled using both generated and real defect sets. However, reconstructing the parameter values of the selected clustering model from a given defect set is not possible analytically (with an exception of the minimum contrast estimation method applicable to the point process clustering model). Although some dependencies between the numerical properties of the generated defect sets and the clustering algorithm parameters are evident (Figures 26, 28, 29), unambiguously determining the parameters from i.e. Voronoi σ values alone is not possible. This poses the question of whether the parameter values could be estimated with an acceptable accuracy using machine learning approaches and certain numeric features of defect sets.

Normalized Voronoi sector areas of all defect sets generated with parameters listed in Table 15 were selected as the initial data source from which numerical features for machine learning models could be derived. The four previously discussed statistical measures (standard deviation, MAD, skewness, kurtosis) were first used as predictor variables for linear regression models predicting parameter values of each clustering model. 10-fold cross-validation was performed to evaluate the regression models, where the training and validation splits were made according to sample groups representing different

clustering parameter value combinations.

Table 22 shows the performance of the regression models, presented in terms of R^2 and MAE scores. The prediction accuracy of the models is rather poor in all cases, with a relative exception of the LCN clustering algorithm.

Metric	Attraction			LCN		Point process		
	N_{def}	r_{def}	d_T	P	S	κ	r	α
R^2	0.150	0.089	0.105	0.254	0.401	−0.428	0.374	0.225
MAE	33.931	8.651	6.871	0.036	0.367	4.456	0.026	3.859

Table 22: Cross-validation scores of linear regression models predicting clustering algorithm parameters based on statistical properties of Voronoi sector areas.

Another approach was based on using the histograms of Voronoi sector areas as the input features for the regression models. The histograms were computed for each instance of generated clustered defect sets and consisted of 40 bins in a range from 0 to 4, normalized for the values to sum up to 1. PCR models with a varying number of principal components were cross validated following the same principle as described earlier. Despite the more informative feature set, no significant performance gains for any clustering algorithm parameter were achieved (Table 23) in comparison to simpler linear regression models.

Model	Attraction			LCN		Point process		
	N_{def}	r_{def}	d_T	P	S	κ	r	α
PCR(2)	0.165	0.084	0.044	0.310	0.504	−0.875	0.548	0.023
PCR(5)	0.264	0.099	0.113	0.309	0.513	−0.829	0.542	0.049
PCR(8)	0.260	0.096	0.116	0.310	0.518	−0.832	0.541	0.072

Table 23: Cross-validation scores (R^2) of PCR models predicting clustering algorithm parameters based on histograms of Voronoi sector areas.

Lastly, the linear models were replaced with a non-linear alternative, the k-nearest neighbors regression. KNN models were fitted by using the EMD metric for computing the distances between points (represented as histograms, computed in the same way as before) and predictions were computed using uniform weights assigned to neighborhood points. Table 24 shows the cross-validation scores of all clustering parameter predictions, where the number of

neighbors for the KNN algorithm was varied. Still, the use of a completely different regression approach did not result in a significantly improved prediction accuracy, while the scores for some parameters got even lower.

Model	Attraction			LCN		Point process		
	N_{def}	r_{def}	d_T	P	S	κ	r	α
KNN(2)	−0.228	−0.094	−0.227	−0.048	0.199	−0.502	0.180	0.247
KNN(5)	−0.089	0.081	−0.010	0.144	0.360	−0.342	0.307	0.387
KNN(10)	−0.025	0.140	0.070	0.205	0.412	−0.296	0.352	0.413
KNN(20)	−0.003	0.183	0.106	0.240	0.442	−0.285	0.385	0.424
KNN(50)	0.021	0.202	0.124	0.259	0.457	−0.295	0.408	0.420

Table 24: Cross-validation scores (R^2) of KNN regresion models predicting clustering algorithm parameters based on the histograms of Voronoi sector areas.

One explanation for the poor performance of all tested regression models is related to the intrinsic properties of the data used in experiments. As for each clustering algorithm and a specific combination of its parameters (Table 15) a total of 100 defect sets were generated independently, each such group displays certain variability due to the stochastic nature of all three clustering algorithms. By treating each group of 100 such cases as a cluster (not to be confused with defect clusters) of multivariate data points, where each point is a histogram of the Voronoi sector areas of the defect set, one can define the centroid of the data cluster as an averaged histogram of all defect sets belonging to that group. Then, an average distance (in terms of EMD metric) between each histogram of an individual defect set and its corresponding averaged histogram of the group can be defined as follows:

$$\bar{d}_w = \frac{1}{GM} \sum_{i=1}^{G} \sum_{k=1}^{M} d(h_{i,k}, \bar{h}_i). \tag{52}$$

Here G denotes the total number of groups (unique clustering parameter value combinations), M is the number of generated cases in each group ($M = 100$), $h_{i,k}$ is the k-th histogram from i-th group and $\bar{h}_i$ is the averaged histogram of i-th group. Similarly, an average distance between an averaged histogram $\bar{h}_i$ of i-th group and averaged histograms of all other groups can be defined:

$$\bar{d}_b = \frac{1}{G(G-1)} \sum_{i=1}^{G} \sum_{j=1, i\neq j}^{G} d(\bar{h}_i, \bar{h}_j). \tag{53}$$

The defined quantities are loosely related to the concepts of within-cluster sum of squares (WCSS) and between-cluster sum of squares (BCSS), commonly used in evaluating the results of clustering algorithms, such as K-Means [18]. As shown in Table 25, the values of $\bar{d}_w$ and $\bar{d}_b$ are in the same order of magnitude for all defect clustering models, which suggests that defect set groups generated with different parameter combinations can still overlap significantly. The same is reflected by the silhouette scores, where negative or zero values indicate poor separation of the groups.

Table 25: Group separation evaluated for different defect clustering models.

Clustering model	Silhouette score	$\bar{d}_w$	$\bar{d}_b$
Attraction	0.021	2.42×10^{-3}	5.22×10^{-3}
LCN	-0.057	3.95×10^{-3}	5.42×10^{-3}
Point process	-0.209	7.08×10^{-3}	8.48×10^{-3}

2.6. Methodology validation with experimental AFM data

2.6.1. AFM dataset description

To validate the proposed methodology against real-world data, three AFM images of actual tBLMs were obtained. Each tBLM sample was formed in a separate vial (following the experimental procedure described in earlier work [77]) and injected with vaginolysin (VLY) solution to induce membrane defects. Samples were then imaged with an AFM microscope (Dimension Icon AFM, Bruker) by scanning a single $2\,\mu\text{m} \times 2\,\mu\text{m}$ surface at a time. Raw images were pre-processed by performing 3rd-degree polynomial flattening, using NanoScope Analysis software. The experiment and data acquisition were conducted by Dr. Marija Jankunec, researcher at the Department of Bioelectrochemistry and Biospectroscopy, Life Sciences Center, Vilnius University.

Each image was stitched from a grid of 9 AFM image fragments taken at 512×512 resolution and covering $2\,\mu\text{m} \times 2\,\mu\text{m}$ membrane surface area, thus the final image represents the area of $6\,\mu\text{m} \times 6\,\mu\text{m}$ at 1536×1536 resolution. Figure 40 illustrates one such example with Voronoi diagram overlaid on top of the AFM image where several defect clusters, characterized by a large number of small sectors, can be observed. Coordinates of the defects present in each image were annotated manually by a domain expert and Voronoi diagrams with corresponding statistical properties were computed for all cases (Table

26). Results show that experimentally registered defect distributions are significantly different from baseline random cases in terms of standard deviation and MAD, although this does not apply for skewness and kurtosis. Finding the most closely matching probability distribution (a similar approach used for the random defect distribution model, described in section 2.3) was not attempted for this dataset, as its sample size was considered to be too small for this task.

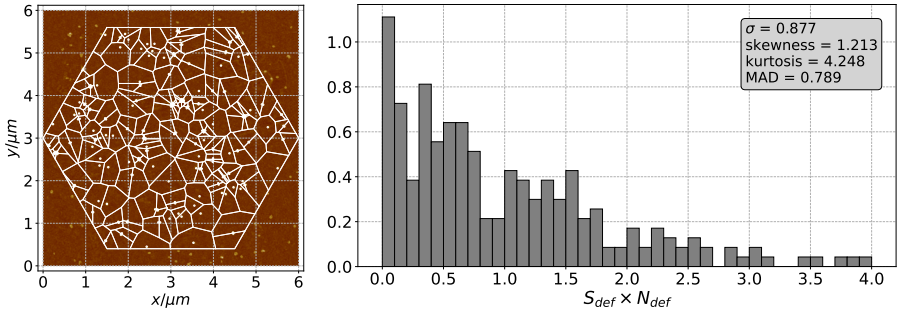


Figure 40: Left: combined AFM image, scan size is 36 μm^2. Each highlighted dot represents a membrane defect. A total of 235 defects are present in the hexagonal domain denoted in the image, overlaid with the Voronoi diagram. Right: histogram of Voronoi sector areas for the given defect set.

Table 26: Statistical properties of Voronoi sector areas of experimentally registered defect distributions in comparison with random defect distribution properties.

Domain	**Image**	N	N_{def}	**Stdev**	**Skewness**	**Kurtosis**	**MAD**
Hexagonal	1	234	10.01	1.22	2.21	10.00	0.83
	2	148	6.33	1.12	1.80	6.52	0.70
	3	235	10.05	0.88	1.21	4.25	0.79
Rectangular	1	374	10.39	1.17	1.78	6.40	0.83
	2	235	6.53	1.06	2.02	8.63	0.72
	3	328	9.11	0.81	1.06	3.78	0.78
-	Random	-	-	0.54	1.18	5.38	0.49

2.6.2. Clustering model evaluation

In order to evaluate how applicable the proposed defect clustering models are to the real defect sets (measured by AFM), the modeled EIS spectra obtained

from the generated defect sets were compared with the corresponding spectra of AFM defect sets (described in subsection 2.6.1). To perform that, the parameters for each clustering algorithm had to be selected so that the resulting generated defect sets would be as similar as possible (in terms of clustering effect) to the real ones. As shown earlier, Voronoi σ cannot be used alone to unambiguously select parameters of any clustering algorithm, while the regression models (subsection 2.5) exhibited poor performance of parameter prediction. For these reasons, the algorithm parameters were instead selected by comparing (using the EMD metric) the histograms of AFM defect sets to the averaged histograms of generated defect sets (Table 15) and choosing the closest match (lowest EMD value). All calculations were performed with three defect size r_{def} options of 0.5 nm, 13 nm and 25.5 nm.

Table 27 shows the selected attraction clustering algorithm parameters for each AFM surface as well as the EMD distances to the closest matching averaged histogram of the generated defect sets. Due to the algorithm's dependency on N_{def} and r_{def} parameters, d_T was selected for each AFM surface and r_{def} value, assuming the fixed defect density of $N_{def} = 10$.

AFM surface	r_{def}	d_T	EMD
1	0.5	10	1.24×10^{-2}
	13.0	30	8.27×10^{-3}
	25.5	25	8.15×10^{-3}
2	0.5	10	7.62×10^{-3}
	13.0	30	7.22×10^{-3}
	25.5	20	6.55×10^{-3}
3	0.5	10	5.69×10^{-3}
	13.0	15	4.54×10^{-3}
	25.5	10	4.69×10^{-3}

Table 27: Modeling parameters of the attraction algorithm selected for comparison with AFM data.

Parameters for the LCN algorithm were selected in a similar way (Table 28), except no distinction between different defect densities or sizes was made. EMD distances to the best matching averaged histograms are somewhat lower among all AFM surface cases, suggesting a relatively better fit between this

model and the experimental data.

AFM surface	P	S	EMD
1	0.00	0.75	7.94×10^{-3}
2	0.06	1.00	4.95×10^{-3}
3	0.06	0.75	2.62×10^{-3}

Table 28: Modeling parameters of LCN algorithm selected for comparison with AFM data.

Parameters for the point process model were selected in two different ways. In the first approach, the parameter values (Table 29) were selected by histogram comparison, in the same way as for attraction and LCN models. The other approach involved fitting the point process model to AFM data using the minimum contrast method (described in subsection 2.2.3). Table 30 shows this alternative set of parameters with the addition of EMD distances, computed by generating 100 defect set instances with each parameter combination and comparing their averaged histograms with corresponding histograms of AFM defect sets.

AFM surface	κ	r	α	EMD
1	5	0.03	5	2.29×10^{-3}
2	10	0.06	5	3.95×10^{-3}
3	10	0.12	5	3.82×10^{-3}

Table 29: Modeling parameters of point process algorithm selected for comparison with AFM data (minimum EMD distances from histogram comparison).

AFM surface	κ	r	α	EMD
1	61.14	0.006	6.12	6.04×10^{-3}
2	80.71	0.005	2.91	2.59×10^{-3}
3	159.64	0.005	2.05	2.77×10^{-3}

Table 30: Modeling parameters of point process algorithm selected for comparison with AFM data (minimum contrast estimation).

To perform further modeling of EIS spectra, for each one of the three AFM surfaces and each r_{def} value a total of 10 defect set cases were generated, using the parameter values listed in the previous tables. In the case of attraction and LCN models both defect density N_{def} and count N were set equal to the corresponding N_{def} and N values of the AFM defect sets (Table 26), whereas the defect sets produced by the point process model also had fixed N_{def}, but varying N values, ranging from 7 to 78 and from 174 to 511 for histogram comparison and minimum contrast approaches correspondingly. The comparison of EIS spectra obtained from AFM data and the generated clustered cases is visualized in Figures S.15–S.16, Appendix 1.

Table 31 shows the means and standard deviations of Voronoi σ values of defect sets generated with each AFM surface and defect clustering model (using the selected parameters).

AFM surface	Attraction	LCN	PP (EMD)	PP (MC)
1	0.78 (0.19)	1.03 (0.10)	1.26 (0.37)	1.43 (0.11)
2	0.69 (0.12)	0.98 (0.11)	1.12 (0.32)	1.08 (0.08)
3	0.60 (0.05)	0.87 (0.08)	0.83 (0.18)	0.94 (0.06)

Table 31: Mean Voronoi σ values (standard deviations in parentheses) of clustered defect sets generated using the selected parameters. PP denotes the point process model, where EMD and MC stand for histogram comparison and minimum contrast methods, respectively.

Due to the moderate clustering effect evident in AFM data (Table 26) and the corresponding EIS spectra containing no anomalous features (such as the double minima), the EIS spectra of both experimental and generated clustered defect sets were compared in terms of the minimum point coordinates, expressed as $\log f_{min}$ and $\arg Y_{min}$. The following notation is used to represent the differences between the generated and AFM cases on both axes:

$$D_f = \log f_{min}^{(model)} - \log f_{min}^{(afm)}, \tag{54}$$

$$D_Y = \arg Y_{min}^{(model)} - \arg Y_{min}^{(afm)}. \tag{55}$$

For reference purposes, the random defect distribution model was initially used to generate defect sets with N and N_{def} properties equivalent to AFM data. Table 32 shows the comparison of EIS spectra of both datasets. In all cases

a systemic shift of both $\log f_{min}$ and $\arg Y_{min}$ is evident with the magnitude depending on r_{def}. This is consistent with earlier observations discussed in subsection 1.5.3. However, the shifts can be noticed to decrease as the Voronoi σ gets lower in AFM defect sets.

AFM	D_f			D_Y		
surf.	$r_{def} = 0.5$	$r_{def} = 13.0$	$r_{def} = 25.5$	$r_{def} = 0.5$	$r_{def} = 13.0$	$r_{def} = 25.5$
1	0.35 (0.01)	0.51 (0.01)	0.55 (0.01)	-3.73 (0.58)	-1.01 (0.38)	1.07 (0.36)
2	0.27 (0.01)	0.39 (0.02)	0.41 (0.02)	-3.24 (0.52)	-0.57 (0.66)	0.62 (0.63)
3	0.21 (0.01)	0.31 (0.03)	0.38 (0.02)	-2.55 (0.45)	-1.27 (0.45)	-0.21 (0.38)

Table 32: Comparison of the minimum point coordinates of EIS spectra obtained from AFM data and generated by using the random defect distribution model.

A comparison of EIS spectra of attraction clustering model cases (Table 33) still indicate dependency on r_{def}, but D_f tends to decrease when r_{def} increases, in contrast to the random model. Although D_Y values are comparable to the ones of random cases, the trend in their decrease (when r_{def} increases) is less identifiable as D_Y is highest at $r_{def} = 13$ for all AFM cases. Standard deviations of both D_f and D_Y are similar to the random model, indicating relatively low variation among the generated defect set instances of the attraction model. Despite the lower Voronoi σ values (Table 31) of the generated attraction model cases relative to AFM defect sets, this clustering model shows a slightly better match to AFM data in terms of D_f and D_Y, as compared to the random defect distribution model.

AFM	D_f			D_Y		
surf.	$r_{def} = 0.5$	$r_{def} = 13.0$	$r_{def} = 25.5$	$r_{def} = 0.5$	$r_{def} = 13.0$	$r_{def} = 25.5$
1	0.36 (0.01)	0.14 (0.02)	-0.04 (0.03)	-3.48 (0.44)	-3.52 (0.23)	1.64 (0.37)
2	0.27 (0.01)	0.14 (0.02)	0.05 (0.01)	-3.09 (0.54)	-3.43 (0.43)	-0.62 (0.45)
3	0.21 (0.02)	0.24 (0.03)	0.24 (0.02)	-2.16 (0.35)	-3.04 (0.56)	-1.32 (0.55)

Table 33: Comparison of the minimum point coordinates of EIS spectra obtained from AFM data and the attraction model instances.

The comparison of LCN clustering model data (Table 34) against AFM data reveals somewhat different trends. Mean D_f values show little dependency on r_{def} or specific AFM cases and are, in general, lower than corre-

sponding $\log f_{min}$ shifts produced by the random model. However, D_Y values are significantly higher in most cases, indicating a consistent upward shift of $\arg Y_{min}$. This is observed despite the fact that the Voronoi σ values of LCN defect sets are closer to the AFM cases than the corresponding attraction model instances. Standard deviations of both D_f and D_Y are also much higher than in the case of the attraction model, indicating more significant variations among defect sets generated with the same clustering algorithm parameters.

AFM	D_f			D_Y		
surf.	$r_{def} = 0.5$	$r_{def} = 13.0$	$r_{def} = 25.5$	$r_{def} = 0.5$	$r_{def} = 13.0$	$r_{def} = 25.5$
1	0.22 (0.04)	0.35 (0.10)	0.31 (0.16)	3.41 (1.18)	6.17 (0.80)	7.23 (1.29)
2	0.16 (0.04)	0.11 (0.13)	0.18 (0.05)	2.32 (0.98)	6.24 (1.53)	6.20 (0.91)
3	0.15 (0.04)	0.21 (0.04)	0.24 (0.04)	1.60 (0.75)	2.92 (1.07)	3.77 (0.98)

Table 34: Comparison of the minimum point coordinates of EIS spectra obtained from AFM data and the LCN model instances.

The point process model showed qualitatively different results depending on the method by which the algorithm parameters for generating defect set instances were selected - histogram comparison (Table 35) or minimum contrast method (Table 36). Results of the first approach indicate mean D_f and D_Y values comparable to LCN results in most cases. However, the standard deviations of both D_f and D_Y are the highest among all tested defect clustering models (often much exceeding the standard deviations of $\log f_{min}$ and $\arg Y_{min}$ of comparable random defect distribution cases, presented in Table 4), reflecting the influence of varying and relatively low defect counts N on the features of EIS spectra.

The other approach by which the point process model parameters were selected for modeling (minimum contrast method) produced the overall best match to AFM data, compared to other tested defect clustering algorithms. As Table 35 shows, the magnitudes of D_f values for AFM surfaces 2 and 3 are much lower than any previous scores of other models, with relatively worse results of surface 1 still being better than the reference cases of the random defect distribution model. An improvement over the said approach and other defect clustering models is evident in D_Y values as well. Standard deviations of D_f and D_Y are significantly lower than in the alternative point process model parameter selection approach and are comparable to the attraction clustering model.

AFM	D_f			D_Y		
surf.	$r_{def} = 0.5$	$r_{def} = 13.0$	$r_{def} = 25.5$	$r_{def} = 0.5$	$r_{def} = 13.0$	$r_{def} = 25.5$
1	-0.04 (0.13)	-0.30 (0.19)	-0.15 (0.14)	3.20 (3.07)	4.17 (2.77)	3.62 (1.02)
2	0.11 (0.10)	0.18 (0.21)	0.15 (0.21)	5.37 (4.01)	7.29 (2.70)	7.08 (2.50)
3	0.20 (0.04)	0.28 (0.04)	0.35 (0.04)	2.64 (2.70)	2.46 (1.28)	3.20 (1.80)

Table 35: Comparison of the minimum point coordinates of EIS spectra obtained from AFM data and the LCN model instances (parameters fitted by histogram comparison).

AFM	D_f			D_Y		
surf.	$r_{def} = 0.5$	$r_{def} = 13.0$	$r_{def} = 25.5$	$r_{def} = 0.5$	$r_{def} = 13.0$	$r_{def} = 25.5$
1	-0.13 (0.06)	-0.24 (0.05)	-0.28 (0.09)	3.23 (0.84)	2.71 (1.06)	1.87 (0.93)
2	-0.03 (0.03)	-0.07 (0.05)	-0.08 (0.03)	0.74 (1.03)	1.47 (0.88)	1.26 (0.80)
3	-0.03 (0.02)	-0.04 (0.02)	-0.04 (0.02)	1.14 (0.77)	0.93 (0.47)	0.68 (0.79)

Table 36: Comparison of the minimum point coordinates of EIS spectra obtained from AFM data and the LCN model instances (parameters fitted with the minimum contrast method).

2.7. Conclusions

- Three different algorithms (attraction, LCN, point process) for generating synthetic clustered defect sets have been presented and compared against the random defect distribution model. All algorithms are capable of producing defect sets distinct from random cases and exhibiting a varying degree of defect clustering.

- The standard deviation of Voronoi diagram sector areas (Voronoi σ) computed for a defect set enables the quantification of the clustering effect by a simple and interpretable metric. The mean value of 0.54 was determined for the random defect distribution model which can be used for the differentiation of random and clustered defect sets.

- Modeled EIS spectra of clustered defect sets exhibit qualitative differences compared to similar EIS data of random defect sets. The shapes of EIS curves show dependency on Voronoi σ, where higher amounts of defect clustering cause distortions (such as double minima) of EIS

curves relative to the canonical cases of random or regular defect distributions.

- Linear regression models for predicting Voronoi σ based on engineered spectral features (describing the properties of EIS spectral peak) have been shown to perform relatively poorly ($R^2 < 0.45$) on the full clustered EIS dataset. However, the analogous models performed better ($R^2 > 0.7$) on most of the subsets of the dataset corresponding to specific combinations of N_{def} and r_{def}. This indicates that the current approach cannot fully decouple the variations in EIS spectra caused by clustering effects from the variations related to membrane parameter values (such as N_{def} and r_{def}). Such an issue is also reflected by significantly worse accuracy of the membrane parameter predictions obtained with the methodology presented in Chapter 1.

- The presented defect clustering models were compared with the random defect distribution model in the context of experimentally-measured AFM defect sets. Analysis of modeled EIS data showed that the point process model with its parameters estimated by the minimum contrast method displayed the best match.

3. Automated defect detection in AFM images

Automated object detection in digital images has been a key area of interest in computer vision research during the last few decades [90, 86]. A multitude of traditional image processing methods based on handcrafted feature extraction have, for many tasks, recently been outperformed by deep learning approaches, based on various architectures of convolutional neural networks (CNN) and enabled by the availability of large amounts of training data and high-performance specialized computing hardware. The automated detection of membrane defects in AFM images also falls into the described problem category and is relevant in the context of computational membrane models. EIS modeling of real defect distributions obtained from AFM images (discussed in Chapters 1 and 2) was so far performed using the defect coordinate sets manually annotated by a domain expert. Automating this process would be beneficial in various aspects of EIS modeling and quantifying membrane damage from AFM data.

The task of defect detection in AFM images can be predefined by several assumptions:

- Membrane defects present in the grayscale height-channel AFM images are visible as light objects on a dark background and can be characterized as either complete or incomplete ring-like structures with an approximately constant diameter.
- Defects of varying completeness can be clustered together.
- Membrane images are affected by noise and distortions inherent in the AFM image acquisition process.
- The amount of available image data is often limited due to expensive imaging equipment or materials and complex measurement process.

To the best of our knowledge, no generic methods have yet been proposed for automated defect detection in AFM images of tBLM membranes, taking into account the listed properties of the problem. Our literature review suggests that in many cases researchers still resort to the manual work of annotating and quantifying objects of interest in AFM images of lipid membranes [68, 84, 53].

Despite the absence of available solutions for this specific defect detection task, substantial progress has been recently made in developing practical methods for other similar problems. Meng et al. [82] presented an algorithm

based on local adaptive Canny edge detection and circular Hough transform which is suitable for recognizing particles in scanning electron microscope (SEM) or transmission electron microscope (TEM) images. Another study conducted by Venkataraman et al. [25] showed that rotavirus particles in AFM images can be detected by applying a series of image pre-processing, segmentation and morphological operations. Marsh et al. [81] proposed the Hessian blob algorithm for detecting biomolecules in AFM images and showed its superiority against the threshold and watershed image segmentation algorithms. Other recent studies also showed that deep learning techniques can be successfully applied to detect complex-shaped objects in microscopy images. Sotres et al. [98] used the YOLOv3 object detection model and a Siamese neural network to determine the locations of DNA molecules in AFM images and identify the same molecule in different images. Okunev et al. [93] applied a Cascade Mask-RCNN neural network to detect metal nanoparticles in scanning tunneling microscopy (STM) images. An open-source software tool for the automated biomolecule tracing in AFM data based was also recently developed and presented by Beton et al. [96]

3.1. Object detection algorithms

3.1.1. Hough transform

Hough transform (HT) is one of the classic image processing algorithms, first developed for the detection of straight lines in noisy images [2]. The original algorithm operates on a binary image and is based on a voting procedure, in which each pixel in the image coordinate space is mapped to all possible object locations in its parameter space. In the case of straight-line detection, points (x and y coordinates) belonging to a single line are defined by a parametric equation in polar coordinates (line parameterized by θ and ρ):

$$\rho = x\cos\theta + y\sin\theta. \tag{56}$$

HT for a given binary image is performed by initializing a 2D array (referred to as the accumulator array) representing all possible values of θ and ρ at a desired level of discretization. For each non-zero pixel in the image, the accumulator array values corresponding to all lines possibly passing through that point are incremented. Afterwards, the coordinates of the maximum values in the accumulator array are selected (by a specified threshold or some non-maximum suppression technique) as the detected lines in the image.

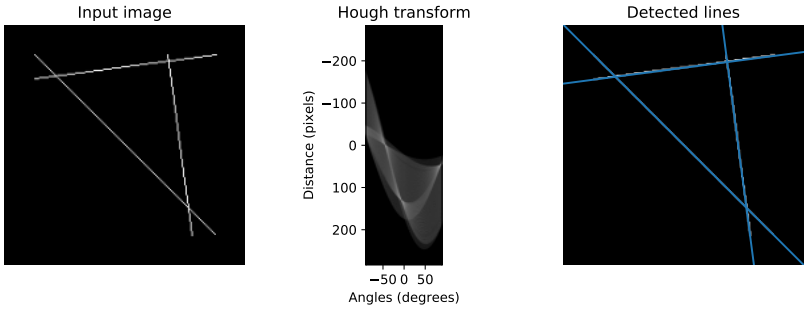


Figure 41: Example of the original image and the corresponding Hough accumulator array [55].

Circular Hough transform (CHT) is a modification of standard HT used to detect circular patterns in images, where a circle is characterized by its center coordinates a and b and the radius r:

$$(x-a)^2+(y-b)^2=r^2. \tag{57}$$

CHT is implemented in a similar way as the standard HT, where the accumulator array is three-dimensional (representing possible circle parameters a, b and r).

3.1.2. Convolutional neural network

Convolutional neural networks (CNN) were first introduced by LeCun et al. [11] as a novel method for handwritten character recognition. The main advantages of CNNs over the typical feed-forward artificial neural networks (ANN) are the reduced number of trainable weights and sensitivity to spatial features of an input image. Such models typically consist of the following layers [65]:

- Convolutional layer. This layer represents a number of convolutional image filters of selected dimensions (i.e. 3×3 pixels) and strides that are applied on the input matrix, resulting in feature maps which are then passed to the subsequent layer of the network.
- Pooling layer. This special type of non-trainable layer (also refered to as the subsampling layer) reduces the dimensions of the feature maps (outputs of the preceding convolutional layer) by computing the average or maximum values.

- Fully connected (FC) layer. Analogous to feed-forward neural networks, FC layers consist of fully connected neurons with some activation functions (i.e. sigmoid), use the flattened output of previous convolutional and pooling layers as its input and perform the actual classification task on the image.

Figure 42 illustrates an example of CNN architecture, consisting of multiple convolution, pooling (subsampling) and fully-connected layers.

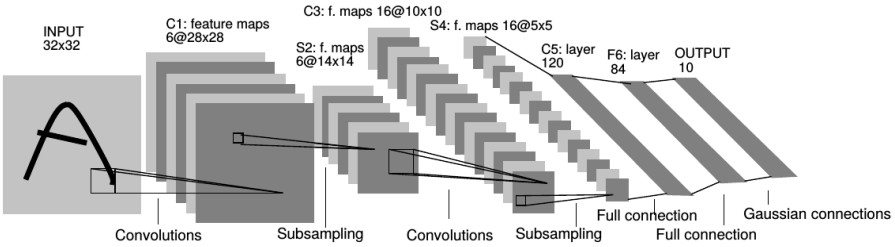


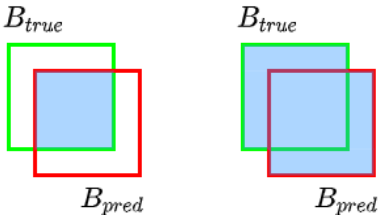
Figure 42: Example of LeNet-5 CNN architecture [11].

Although the original CNN models were most often applied for image classification tasks, various architectures were recently developed for object detection as well. Region-based CNN (R-CNN) [47] was one of the initial solutions based on a two-stage object detection approach, where candidate regions (likely to contain the objects of interest) are first selected in the original image and then subsequently classified using a CNN model. The original R-CNN was later modified to achieve better performance (Fast R-CNN [58]) and perform image segmentation tasks (Mask R-CNN [74]).

Another family of object detection models, referred to as one-stage detectors was recently introduced with the YOLO model [69]. In this approach, a single deep neural network is applied for the full image, subdivided into candidate regions, and probabilities of object presence are predicted simultaneously for all possible locations. In comparison to two-stage detectors, such implementation greatly improved detection performance, albeit at the expense of lower object localization accuracy. SSD (single-shot multibox detector) [67] is another notable example of one-stage detectors, which operate on multiple-scale feature maps.

3.1.3. Detection accuracy evaluation

Although membrane defects are primarily characterized by their center coordinates and defect radius, these attributes can be used to express the defect position in the image as its bounding rectangle. By comparing two sets of bounding rectangles, corresponding to true and predicted defect positions, defect detection accuracy can be quantitatively evaluated.



(a) Intersection (b) Union

Figure 43: Bounding rectangle overlap of true and predicted defect positions.

$$IoU = \frac{B_{true} \cap B_{pred}}{B_{true} \cup B_{pred}}. \tag{58}$$

To count the number of correct detections, the bounding rectangle of each true defect position (B_{true}) is matched with its closest prediction (B_{pred}). The overlap between each such pair of true and predicted bounding rectangles is evaluated by the intersection over union (IoU) metric (58) (also known as the Jaccard index), which is expressed as the ratio of bounding rectangle intersection and union areas (Fig. 43). Higher IoU values correspond to a better match between both bounding rectangles. If IoU value is above the chosen threshold (i.e. 0.5), the detection is assumed to be a true positive (TP). Otherwise, if no matching prediction exists for a given true position, such detection is counted as a false negative (FN). In the opposite case, when no true bounding rectangle can be matched for a given prediction, a false positive (FP) is assumed. By counting all such cases of correct and incorrect detections, overall defect detection accuracy is summarized by precision and recall metrics [62]:

$$Precision = \frac{TP}{TP+FP}. \tag{59}$$

$$Recall = \frac{TP}{TP+FN}. \tag{60}$$

Both precision and recall can also be expressed by the $F1$ metric:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{61}$$

Although the described metrics are most commonly used to measure classification model performance, they have also been applied in evaluating object detection accuracy in many works [26, 64, 57].

3.2. Defect detection experiments

3.2.1. AFM image dataset

The AFM dataset described earlier (Table 26, chapter 2) was reused for defect detection experiments further described in this section. Image fragment sets of each cell were partitioned into training and test subsets by assigning 5 fragments for training and 4 for testing. Test fragments were selected to represent a cohesive $4\,\mu m \times 4\,\mu m$ surface patch at the lower right corner of the fully stitched image. Table 37 shows the total number of annotated defects (N) and average defect density (N_{def}) for each AFM image cell and training/test subset.

Table 37: AFM image sets used for the defect detection model training and testing.

AFM surface	Subset	Image fragments	N	N_{def}	Voronoi σ
1	Training	5	202	10.10	1.18
2	Training	5	138	6.90	1.12
3	Training	5	170	8.50	0.77
1	Test	4	172	10.75	1.20
2	Test	4	97	6.06	1.02
3	Test	4	158	9.88	0.91

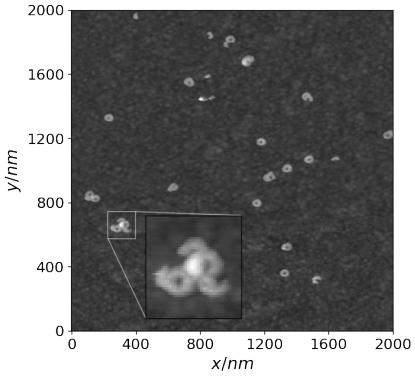


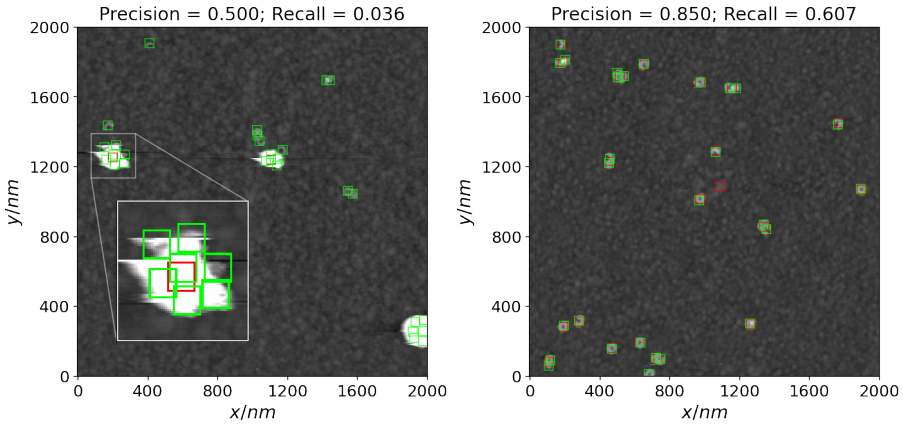
Figure 44: Example of an AFM image fragment with an instance of defect cluster zoomed in.

3.2.2. TopoStats

The first defect detection experiment was performed using an open-source software tool TopoStats, recently developed for biomolecule detection in AFM images [96]. Although the tool is not specifically designed for defect detection in tBLM membrane images, it does support several types of molecules including the so-called membrane attack complexes (MAC) which are ring-like protein structures somewhat resembling the tBLM membrane defects. The default parameter settings specified for MAC pore detection were adjusted by setting the minimum area to 2×10^{-7} and the maximum and minimum deviation from the median pixel size to 5.0 and 0.5 respectively, to adapt the tool for the specific resolution of AFM images. Each detected grain was treated as an individual defect with its coordinates derived from the center of the grain area. Precision, recall and F1 values were computed as described in subsection 3.1.3. Table 38 summarizes the obtained detection results.

Table 38: Defect detection (with TopoStats) accuracy in test AFM images.

AFM surface	N_{true}	N_{pred}	Precision	Recall	F1
1	172	53	0.754	0.233	0.355
2	97	31	0.742	0.237	0.359
3	158	79	0.886	0.443	0.591



(a) Illustrative image fragment containing defect clusters (surface 2).

(b) Illustrative image fragment without defect clusters (surface 3).

Figure 45: Examples of true defect positions (green rectangles) and detected ones by using TopoStats (red rectangles). An instance of a defect cluster and the corresponding true and predicted defect positions is zoomed in on the left image.

Despite the relatively high precision, the recall is low in all cases, indicating a large number of false negatives. The visual inspection of the detection results revealed that the tool poorly resolved defect clusters where the majority of such instances were treated as a single defect (example shown in Figure 45). This can be explained by the lack of the tool's sensitivity to the sizes of detected objects, which is relevant for the detection task.

3.2.3. Area measurement method

To address the issue of defect cluster separation observed in the previous experiment, a simple method based on basic image processing operations was implemented and tested. The main assumption of the approach was that a single defect takes up a certain number of pixels in the image and their clusters can be resolved by dividing them into equally-sized parts, ignoring the fine details of image data in those regions. The algorithm (further referred to as the area measurement method) consists of the following steps:

1. Image thresholding. The initial grayscale AFM image is converted into a binary image using a fixed threshold value T.

2. Morphological processing. Thresholded regions are denoised by applying the binary closing operation and removing small objects (lesser than 5 pixels) [55].

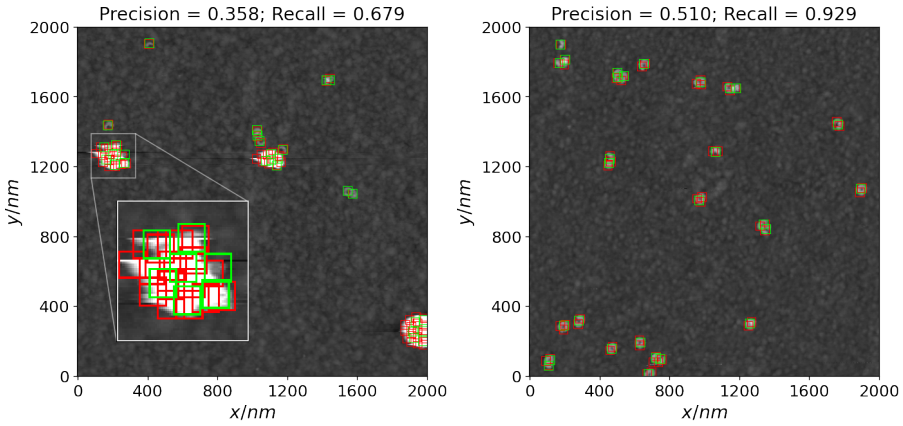
3. Defect count determination. The number of defects making up each region is determined by dividing the region area (in pixels) by a predefined amount S and rounding up the resulting ratio.

4. Defect coordinate assignment. The exact center coordinates of the defects are determined by performing K-means clustering of the pixel coordinates of each region (using the cluster count from the previous step) and using the centroids of each pixel cluster.

To perform the experiment on the test AFM images, the parameter values of $T = 100$ and $S = 130$ were selected by running the algorithm on the training set AFM images with varying S and T values and selecting the ones yielding the highest average F1 value. Table 39 shows the test results. Despite the simplistic approach, the algorithm performed significantly better than TopoStats in terms of recall, although precision was reduced for test images 2 and 3 (detection examples presented in Fig. 46). Overall F1 scores were higher for all test cases.

AFM surface	N_{true}	N_{pred}	Precision	Recall	F1
1	172	114	0.860	0.570	0.685
2	97	114	0.553	0.649	0.597
3	158	194	0.613	0.753	0.676

Table 39: Defect detection accuracy in test AFM images of the area measurement method.



(a) Illustrative image fragment containing defect clusters (surface 2).

(b) Illustrative image fragment without defect clusters (surface 3).

Figure 46: Examples of true defect positions (green rectangles) and detected ones by using the area measurement method (red rectangles).

3.2.4. Hough transform

The next algorithm, based on circular Hough transform (CHT), is defined with an assumption that membrane defects in AFM images are visible as light circle-like structures on dark background, having an approximately constant diameter. The implemented algorithm consists of the following steps:

1. Image thresholding. The initial grayscale AFM image is converted into a binary image, where ones represent the objects (defects) and zeros represent the background (membrane). The exact threshold is obtained using the minimum method [1].

2. Morphological processing. Binary image areas corresponding to defects are converted into single-pixel width contours by applying the morphological thinning operation [6].

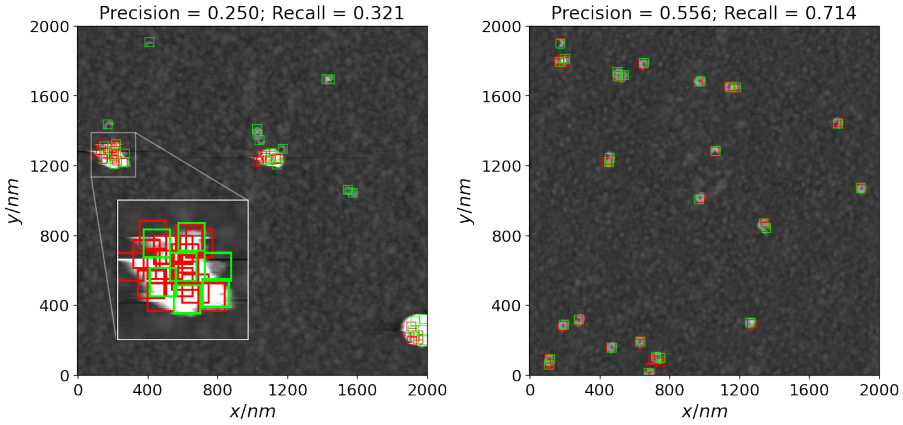
3. Circular Hough transform. The final defect detection is performed by applying CHT on the thresholded and thinned binary image, using some pre-selected Hough threshold and a varying circular radius.

The algorithm was applied to the test image set using the Hough threshold value of 0.28 and circle radius varied from 3 to 7 pixels. These specific parameter values were selected from a set of different combinations by first running this algorithm on the training image set, evaluating defect detection accuracy

in terms of F1 and selecting the parameter combination resulting in the highest average F1 value (0.468 on the training image set). Test results (Table 40) are comparable to the previously described area measurement method, although the average precision is lower due to overall higher detection counts, corresponding to more false positives. Figure 47 illustrates the detection results with two different AFM image fragments.

AFM surface	N_{true}	N_{pred}	Precision	Recall	F1
1	172	201	0.577	0.674	0.622
2	97	156	0.436	0.701	0.538
3	158	140	0.614	0.544	0.577

Table 40: Defect detection accuracy in test AFM images of the CHT-based algorithm.



(a) Illustrative image fragment containing defect clusters (surface 2).

(b) Illustrative image fragment without defect clusters (surface 3).

Figure 47: Examples of true defect positions (green rectangles) and detected ones by using the CHT-based method (red rectangles).

3.2.5. Convolutional neural network

To perform defect detection experiments using convolutional neural networks, we used an SSD FPN architecture object detector. This model consists of multiple components. ResNet-50 is used as the backbone for deep feature extraction. Feature Pyramid Network (FPN) uses ResNet to construct a multi-scale

feature pyramid from a single input image. Finally, a variation of a single-shot multibox detector (SSD) with focal loss (RetinaNet) is used to perform the actual object detection. Network architectures are described in more detail in publications [59, 61, 76].

The initial model [97] was pre-trained with the COCO image dataset [54] to detect objects of 90 different types. In order to adapt it for defect detection in AFM images, the model was re-trained to detect a single type of object (membrane defect) using 15 AFM images (training fragments) described in Table 37 and containing a total of 510 annotated defect instances. Data augmentation techniques were not applied for the initial experiments with the described CNN model, considering the defect instances to be relatively simple objects, weakly affected by scale or orientation. Each training image fragment with 512×512 resolution was scaled to match the model input of 640×640 color (RGB) images. Tensorflow 2.0 framework was used to train and evaluate the model and the training was performed using Nvidia GTX 1080 GPU hardware. The model was trained for 10^4 epochs and total training loss (expressed as the weighted sum of localization and classification loss) is visualized in Figure 48.

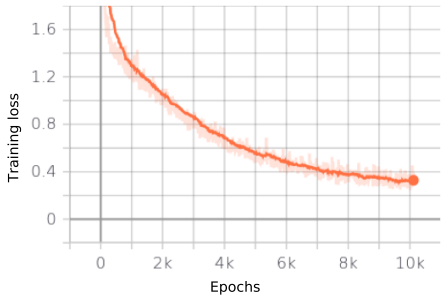
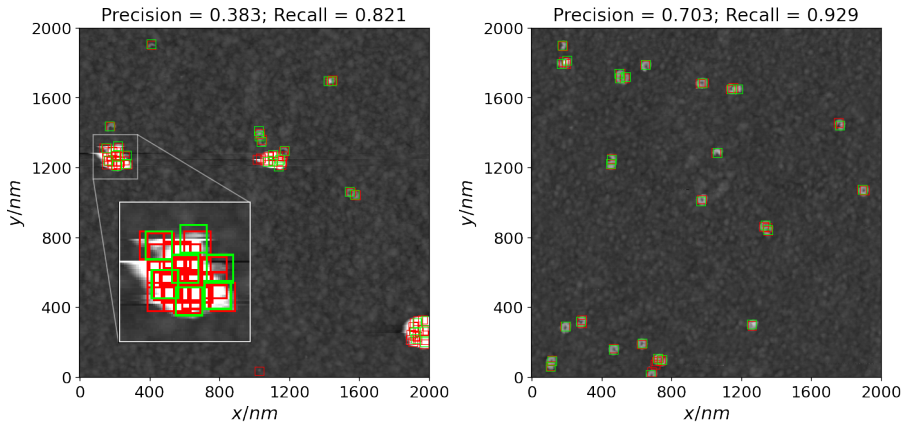


Figure 48: Total training loss of the model.

The trained model was evaluated with each of 12 test image fragments (Table 37) and the detection results were aggregated to match the layout of 4 stitched fragments per each AFM surface. Bounding boxes of all detected defect instances were equalized to match the width and height of 50 nm, corresponding to defects with a circular radius of 25 nm. Defect instances predicted by the model were compared with the true defect positions and the overall model accuracy was evaluated using the precision, recall and F1 metrics for each AFM surface (Table 41).

Precision and recall scores are comparable to previous experiments con-

ducted with simpler algorithms, although the average F1 score among all test images is slightly higher. Defect clusters (Figure 49, left) still proved to be relatively difficult to resolve due to poorly visible surface features inside the clusters. However, the model performed fairly well for certain image fragments with no defect clusters present (Figure 49, right). This is also illustrated by the fact that the test image of AFM surface 3 which indicates the lowest amount of defect clustering in terms of σ (Table 37) also have the highest overall F1 score among all tested algorithms.



(a) Illustrative image fragment containing defect clusters (surface 2)

(b) Illustrative image fragment without defect clusters (surface 3)

Figure 49: Examples of the true defect positions (green rectangles) and the predicted (red rectangles) ones by using the convolutional neural network. An instance of a defect cluster and the corresponding true and predicted defect positions is zoomed in on the left image.

Table 41: Defect detection accuracy of the test AFM images and the resulting differences between EIS spectra of true and predicted defect sets.

AFM surface	N_{true}	N_{pred}	Precision	Recall	F1
1	172	129	0.775	0.581	0.664
2	97	119	0.555	0.680	0.611
3	158	152	0.757	0.728	0.742

3.3. Defect detection accuracy effect on EIS spectra

While the investigated defect detection algorithms indicated adequate performance with real AFM data, it remains unclear how much the detection inaccuracies impact modeled EIS spectra, relative to true known defect coordinates. In order to quantitatively assess the relationship between the defect detection accuracy and the corresponding variations in EIS spectra, a substantial number of defect detection result sets is required. Such detection results should exhibit different precision and recall values distributed in a certain range. However, such specific detection results can be difficult to acquire by applying object detection models trained using real AFM images and annotated true defect positions. Another issue is the limited amount of available AFM image data, which prevents from collecting a sufficiently large collection of true and predicted defect set pairs for EIS modeling.

To work around these issues, an alternative approach was chosen by which defect sets are synthetically generated to emulate imperfect defect detection results at varying accuracy levels. Each synthetic case is generated by starting with the initial set of annotated true defect coordinates and applying certain modifications (defect addition, removal, coordinate shifting) to acquire a new defect set equivalent to the defects being detected by some model with imperfect accuracy. Such a method is different from other defect set generation algorithms presented in this work in the way that it aims to modify an existing (AFM-registered) defect set, instead of generating a completely new defect set with selected N, N_{def} or other properties (such as the clustering model parameters).

3.3.1. EIS modeling

To quantify the discrepancy between the EIS spectra modeled for any given pair of true (annotated by the domain expert) and predicted (by an object detection algorithm) defect sets we used the positions of the minima points of the curves (example in Figure 50) along both frequency and admittance phase axes (assuming that the clustering effect on the EIS spectral shapes is negligible):

$$\Delta \mathrm{f}_{\log} = \log_{10}(f_{min}^{(true)}) - \log_{10}(f_{min}^{(pred)}), \tag{62}$$

$$\Delta \arg Y = \arg Y_{min}^{(true)} - \arg Y_{min}^{(pred)}. \tag{63}$$

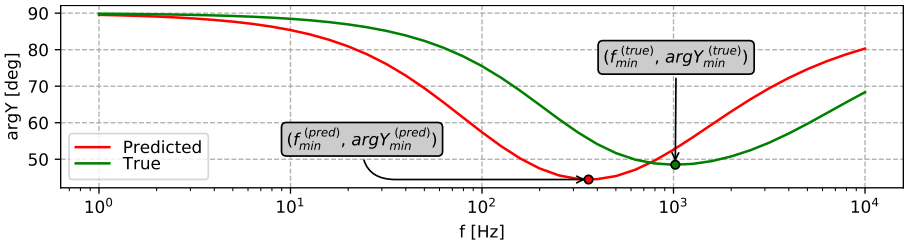


Figure 50: Spectral features of modeled EIS spectra used in quantifying the difference between true and predicted defect set cases.

To characterize the relationship between the defect detection accuracy and deviations in the resulting EIS spectra, using the F1 metric alone is not enough due to the observation that EIS spectral features are more strongly influenced by the defect size and density than by the specific positions of the defects (as discussed in chapter 1). For this reason, a predicted defect set might poorly match the true one and thus exhibit a low F1 value, although their corresponding EIS spectra might closely match, as long as the overall properties of defect count and size are similar. To take this effect into account we also use an additional Q_N metric which represents the ratio of defect densities (number of defects per square micrometer) from predicted and true defect sets:

$$\mathrm{Q_N} = N_{def}^{(pred)} / N_{def}^{(true)}. \tag{64}$$

3.3.2. Synthetic non-clustered defect set generation

The initial approach of producing defect sets imitating non-perfect defect detection results was based on the assumption that the original defect set contains randomly distributed defects and no clustering is in effect. Such defect sets would imitate the results of a defect detection algorithm that poorly resolves defect clusters present in AFM images (such as the TopoStats tool, described in subsection 3.2.2). A single defect set is generated by the following procedure:

1. True coordinates ($x^{(true)}$ and $y^{(true)}$) of each existing defect (in the original defect set) are shifted by adding normally-distributed random values:

$$x^{(pred)} = x^{(true)} + \delta; \qquad y^{(pred)} = y^{(true)} + \delta; \qquad \delta \sim \mathcal{N}(\mu, \sigma^2).$$

   This results in realistically imperfect matches between true and predicted bounding rectangles of the defects.

2. False negatives are introduced by removing n_{remove} number of randomly selected defects from the original defect sets.

3. False positives are introduced by adding n_{add} number of defects with coordinates sampled from the uniform distribution, matching the image dimensions.

For each of the 3 test AFM surfaces, a total of 100 defect sets were generated by independently adjusting n_{add} and n_{remove} from 0 to $N/2$ (half of the original defect count). Due to the varying numbers of artificially introduced false positive and false negative detections, the defect sets exhibited precision and recall values ranging from approximately 0.5 to 1 (Table 42).

Table 42: Summary of generated non-clustered defect sets for each AFM test image. Precision, recall and F1 values were computed against true defect sets annotated in the given AFM image.

AFM surface	Cases	Precision	Recall	F1
1	100	0.52 - 1.00	0.52 - 0.99	0.52 - 0.98
2	100	0.53 - 1.00	0.53 - 1.00	0.53 - 1.00
3	100	0.55 - 1.00	0.53 - 1.00	0.54 - 0.99

In addition to the aforementioned detection accuracy metrics, Voronoi σ values were also computed for all generated defect sets and compared with the corresponding values of the original AFM test surfaces. Voronoi σ histograms (Figure 51) indicate that most of the generated defect sets exhibit a lower degree of clustering as compared to the original AFM defect sets.

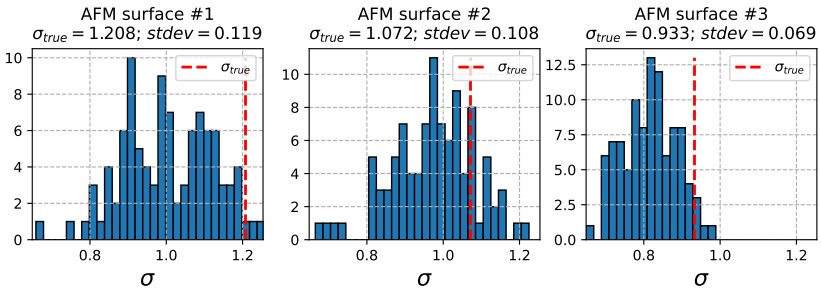


Figure 51: Voronoi σ histograms of synthetically generated non-clustered defect sets. Red dashed lines are the Voronoi σ of original AFM test defect sets.

EIS spectra were modeled for all of the generated defect sets and compared against the initial EIS spectra of AFM defect sets in terms of Δf_{log} and $\Delta \arg Y$. To summarize the effect of defect detection accuracy on the resulting EIS spectra, generated defect sets of each AFM surface were partitioned into several subsets corresponding to fixed-width intervals of F1 and Q_N values. Mean Δf_{log} and $\Delta \arg Y$ values were then computed for each subset.

The results presented in Figure 52 indicate that mean Δf_{log} values across all F1 and Q_N intervals significantly differ among the three AFM test surfaces and tend to get lower in magnitude as the Voronoi σ of original AFM defect sets decreases (the lowest absolute Δf_{log} values can be observed for AFM surface 3). However, mean $\Delta \arg Y$ values do not indicate clear dependency on Voronoi σ of original AFM defect sets, although on average they still approach 0 as the simulated defect detection accuracy increases. As could be expected from the simplistic approach of non-clustered defect set generation, the effect of decreasing Voronoi σ is evident, for example, in mean Δf_{log} versus Q_N intervals where the average deviations are significantly below 0 for $[0.9; 1.1]$ interval (which corresponds to a relatively small difference in defect count between original and generated defect sets). In total, Δf_{log} and $\Delta \arg Y$ range from -0.47 to 0.23 and from -2.38 to 2.79 respectively (full datasets are visualized in Fig. 62, App. 2), which could be considered as relatively low deviations assuming the widely varying F1 and Q_N values and systemic decrease of Voronoi σ.

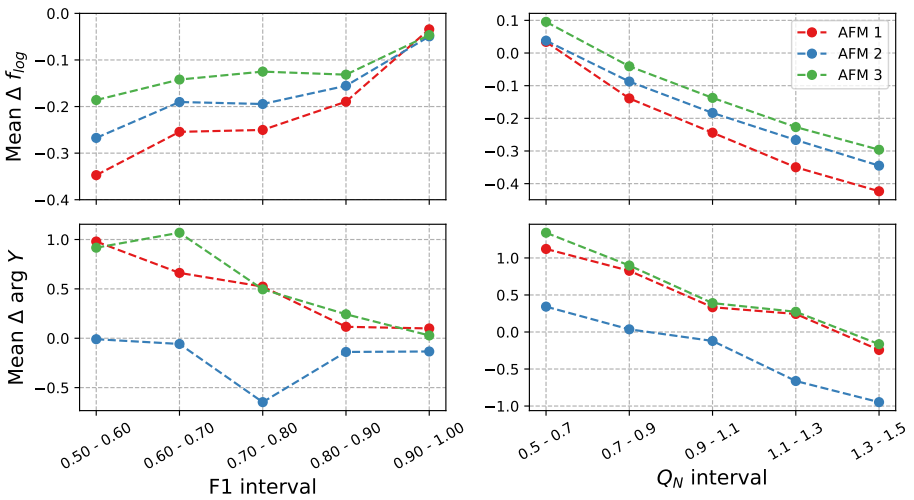


Figure 52: Mean Δf_{log} and $\Delta \arg Y$ values computed for subsets of generated non-clustered defect sets corresponding to different intervals of F1 and Q_N.

3.3.3. Synthetic clustered defect set generation

An alternative method for generating synthetic defect sets is based on the assumption that the defect detection algorithm can resolve defect clusters reasonably well and the overall clustering effect in the predicted set is retained (relative to the true set) despite some incorrect detections. To generate such defect sets, the previously described method was modified to sample the coordinates of FP or FN cases using the method of kernel density estimation (KDE) [34] instead of the uniform distribution. The KDE model represents an empirical probability distribution function of the X and Y coordinates of the defects. Figure 53 shows an example of a clustered defect set and its corresponding KDE model, where warmer colors correspond to the higher values of its probability density function.

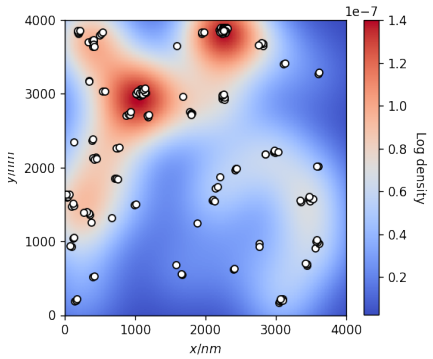


Figure 53: KDE model of true defect coordinates (white dots) annotated for AFM test image #1. The background color represents the log probability density of the fitted KDE distribution.

The modified procedure for generating a series of clustered defect sets consists of the following steps:

1. KDE model is fitted to the coordinates of the original defect set.
2. For each synthetic defect set case:
   (a) True coordinates ($x^{(true)}$ and $y^{(true)}$) of each existing defect are modified by adding normally-distributed random values:

$$x^{(pred)} = x^{(true)} + \delta; \qquad y^{(pred)} = y^{(true)} + \delta; \qquad \delta \sim \mathcal{N}(\mu, \sigma^2).$$

(b) A number n_{remove} of defect coordinate pairs are sampled from the KDE model. True defects closest to the sampled coordinates are selected and removed from the initial defect set. This introduces false negatives (FN) into the generated defect set and reduces recall accordingly.

(c) A number n_{add} of new coordinate pairs are sampled from the KDE model and defects with these coordinates are added to the generated defect set. This represents false positives (FP) and corresponds to lowered precision values.

The described algorithm was used to generate the synthetic cases for each of the three AFM test images independently. KDE instances were fitted using the Gaussian kernel and bandwidth parameter set to 400. Parameters n_{remove} and n_{add} were initially set to 0 and then incremented throughout the generation process by a step quantity corresponding to 3% of true defect count N until the maximum value of $N/2$ was reached. Table 43 shows the properties of the synthetic defect sets generated by the described procedure. Although some variability of the clustering effect is still present in the defect sets (Figure 54) a qualitative difference relative to the previously described method (Figure 51) is evident as the Voronoi σ of generated defect sets are approximately normally distributed around the initial Voronoi σ values.

Table 43: Summary of generated defect sets for each AFM test image. Precision, recall and F1 values were computed against true defect sets annotated in the given AFM image.

AFM surface	Cases	Precision	Recall	F1
1	324	0.49 – 1.00	0.48 – 0.99	0.49 – 0.98
2	256	0.54 – 1.00	0.51 – 1.00	0.53 – 1.00
3	256	0.53 – 1.00	0.51 – 1.00	0.53 – 0.99

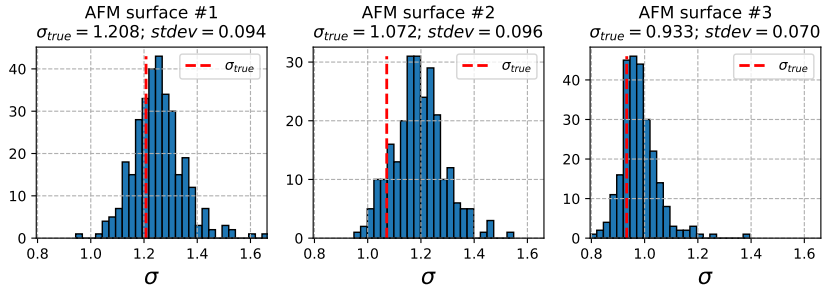


Figure 54: Voronoi σ histograms of synthetically generated clustered defect sets. Red dashed lines are the Voronoi σ values of the original AFM defect sets.

The EIS spectra of the generated cases were compared to the initial data and summarized in the same way as previously, with results presented in Figure 55. Despite the modification in the defect set generation process, allowing to better retain the clustering effect of the original defect set, discrepancies in the modeled spectra in terms of Δf_{log} decreased by no more than 40%, while the mean $\Delta \arg Y$ values show a significant fewfold increase. The full dataset from which the mean values were derived is visualized in Figure 63, Appendix 2.

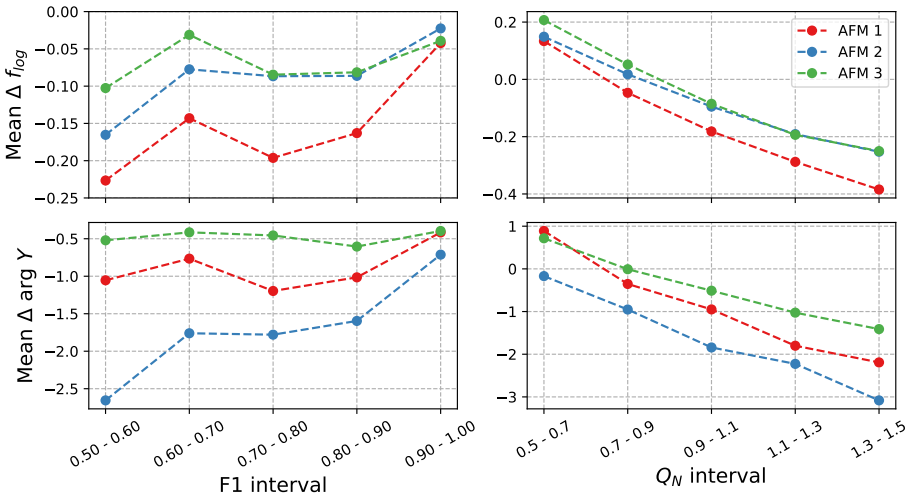


Figure 55: Mean Δf_{log} and $\Delta \arg Y$ values computed for subsets of generated clustered defect sets corresponding to different intervals of F1 and Q_N.

3.4. Comparison of modeled and experimental EIS spectra

The effect of imperfect defect detections on EIS spectra was also evaluated using the actual detection results (further referred to as the predicted defect set) obtained with a convolutional neural network (subsection 3.2.5) as well as the experimental EIS data measured for the three AFM surfaces (Table 26, Chapter 2). Experimental data was obtained by the similar procedures as described in sections 1.5.6 and 2.6.1. For the initial comparison, EIS spectra were modeled using both true and predicted defect sets. Table 44 shows the discrepancy in terms of Δf_{log} and $\Delta \arg Y$.

Table 44: Discrepancies in EIS spectra of true and predicted (with CNN model) defect sets.

AFM surface	F1	Q_N	Δf_{log}	$\Delta \arg Y$
1	0.664	0.750	0.009	0.735
2	0.611	1.227	-0.013	0.681
3	0.742	0.962	-0.027	0.864

To evaluate how this mismatch would translate to membrane parameter (ρ_{sub} and r_{def}) predictions from EIS spectra, a series of FEA modeling tasks were performed with each pair of true and predicted defect sets for all three AFM surfaces (test data) separately. Two parameters were varied in each scenario: defect radius r_{def} was adjusted from 1 nm to 13 nm with increments of 2 nm, while the specific conductivity of the submembrane layer ρ_{sub} was adjusted in logarithmic scale from 10^4 to 10^5 Ω·cm with power increments of 0.1, resulting in a total of 77 parameter combinations. Modeled curves of both true and predicted defect sets were matched against the experimental EIS data by minimizing the L1 norm of minimum point coordinates (frequency and admittance phase axes) between a pair of curves.

Figure 56 shows the modeled and experimental curves of each surface as well as the specific r_{def} and ρ_{sub} values of the corresponding modeled cases. As the experimental EIS measurements were conducted at a narrower frequency range (from 10^0 Hz to 10^4 Hz), the modeled curves were truncated accordingly. The best-matched modeled curves from both true and predicted defect sets display a minor difference of 0.1 in terms of $\log \rho_{sub}$ in all cases, while r_{def} differences range from 0 nm to 4 nm.

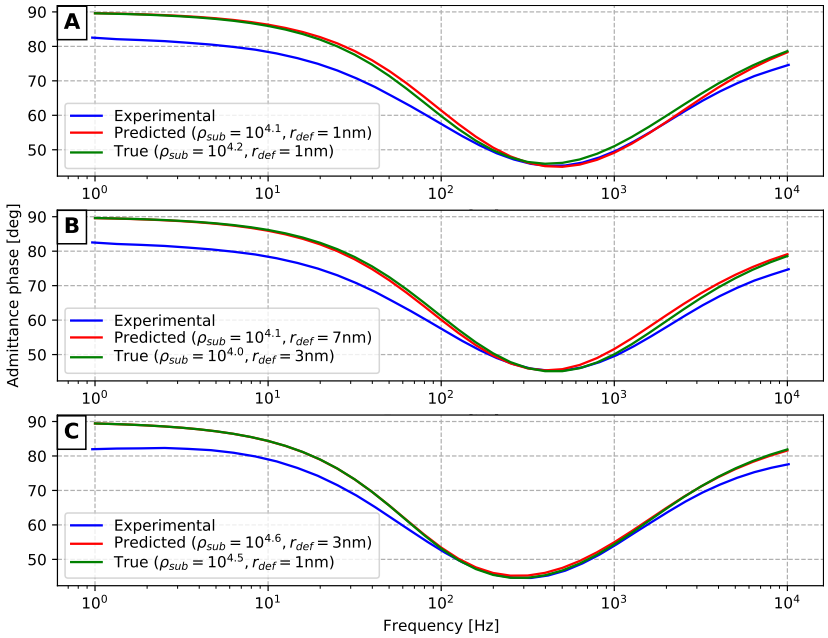


Figure 56: Admittance phase data of experimental EIS measurements (blue curves) versus modeled cases (green and red curves corresponding to manually annotated defect coordinates and CNN model predictions, respectively). The A, B and C panes correspond to AFM surfaces 1, 2 and 3, respectively.

For reference, the regression equations (36) and (37) were also used to estimate ρ_{sub} and r_{def} values from experimental EIS curves by using the defect densities N_{def} determined from AFM data (Table 26). Table 45 shows predicted parameter values for each membrane sample. The data shows that regression models over-estimate both ρ_{sub} and r_{def} in comparison to the previous set of results (Figure 56), although the values from both sets still show correlation, where $\log \rho_{sub}$ difference ranges approximately from 0.3 to 0.5 in all test cases. Such systemic shifts can be explained by the fact that the currently used regression models do not take into account the clustering effect, which is evident from AFM data although being relatively low (in terms of Voronoi σ).

Table 45: Membrane parameters predicted from experimental EIS data by using linear regression models and N_{def} values determined from AFM data.

AFM surface	r_{def}	$\log_{10}\rho_{sub}$
1	4.289	4.717
2	5.435	4.516
3	3.534	4.810

3.5. Conclusions

- Defect detection experiments performed with a small amount of AFM image data produced F1 scores ranging from 0.538 to 0.742 (excluding the results of the TopoStats tool). The convolutional neural network applied for the task showed only marginally better results compared to the simpler methods based on Hough transform or basic image processing operations.

- With the goal of estimating the effect of defect detection errors on the modeled EIS spectra (given that only a small amount of annotated AFM image data is available), two methods of generating synthetic defect detection sets with specific F1 and Q_N values were presented. The comparison of EIS spectra obtained from original and generated defect sets defines the approximate error margins (in terms of Δf_{log} and $\Delta \arg Y$) for defect detection algorithms with known accuracy (expressed in F1 metric). However, both defect set generation methods produce significant variations of the clustering properties (in terms of Voronoi σ). This leads to systemic deviations in the modeled EIS spectra which cannot yet be reliably decoupled from deviations caused by varying F1 and Q_N values.

- The methodology of membrane parameter prediction was applied in conjunction with a defect detection algorithm (convolutional neural network) to predict membrane properties by using both EIS and AFM data, measured for the same tBLM samples. This presented a proof-of-concept case of estimating membrane parameters (specific resistance of submembrane layer ρ_{sub}) which are inaccessible to EIS or AFM measurement techniques in laboratory conditions.

General conclusions

- The three-dimensional membrane model (implemented with FEA technique) is capable of simulating valid EIS responses with arbitrary membrane defect distributions. The methodology of EIS data analysis based on machine learning techniques was demonstrated by predicting quantitative membrane parameters of defect density, size and submembrane specific resistance (all of which are not directly accessible from EIS spectra), using EIS data modeled by using random defect distributions as well as experimental EIS data.

- Three presented defect clustering models can be used to generate realistic defect sets exhibiting a varying degree of clustering and parameterize real AFM-registered defect sets. The standard deviation of Voronoi diagram sector areas computed for clustered defect sets has been proposed as a simple metric suitable for quantifying clustering effect and differentiating clustered and randomly-distributed defect sets. The defect clustering effect is reflected by changes in resulting EIS spectra which cannot yet be fully decoupled (with investigated methods) from the influence of other membrane parameters (such as defect density and size).

- Tests of automated defect detection algorithms on AFM membrane images indicated F1 scores ranging from 0.538 to 0.742, with the convolutional neural network performing marginally better than simpler methods based on Hough transform and basic image processing operations. Due to the limited amount of AFM image data, two methods were presented for generating synthetic defect sets corresponding to specific detection accuracy levels and estimating error margins in resulting EIS spectra.

References

[1] Judith M. S. Prewitt and Mortimer L. Mendelsohn. The analysis of cell images. *Annals of the New York Academy of Sciences*, 128.3 (1966), pp. 1035–1053.

[2] Richard O. Duda and Peter E. Hart. Use of the Hough Transformation to Detect Lines and Curves in Pictures. *Commun. ACM*, 15.1 (Jan. 1972), pp. 11–15.

[3] Peter J. Diggle and Richard J. Gratton. Monte Carlo Methods of Inference for Implicit Statistical Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46.2 (1984), pp. 193–227.

[4] Paul Geladi and Bruce R. Kowalski. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185 (1986), pp. 1–17.

[5] Youcef Saad and Martin H Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on scientific and statistical computing*, 7.3 (1986), pp. 856–869.

[6] L. Lam, S.-W. Lee, and C.Y. Suen. Thinning methodologies-a comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14.9 (1992), pp. 869–885.

[7] J. Fleig and J. Maier. Finite element calculations of impedance effects at point contacts. *Electrochimica Acta*, 41.7 (1996). Electrochemical Impendance Spectroscopy, pp. 1003–1009.

[8] Claudia Steinem, Andreas Janshoff, Wolf-Peter Ulrich, Manfred Sieber, and Hans-Joachim Galla. Impedance analysis of supported lipid bilayer membranes: a scrutiny of different preparation techniques. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1279.2 (1996), pp. 169–180.

[9] B. A. Cornell, V. L. B. Braach-Maksvytis, L. G. King, P. D. J. Osman, B. Raguse, L. Wieczorek, and R. J. Pace. A biosensor that uses ion-channel switches. *Nature*, 387 (1997), pp. 580–583.

[10] E. Vigneau, M. F. Devaux, E. M. Qannari, and P. Robert. Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration. *Journal of Chemometrics*, 11.3 (1997), pp. 239–249.

[11] Yann Lecun, Leon Bottou, Y. Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86 (Dec. 1998), pp. 2278–2324.

[12] J.W. Thomas. *Numerical Partial Differential Equations: Finite Difference Methods*. Texts in Applied Mathematics. Springer New York, 1998.

[13] Roy De Maesschalck et al. The development of calibration models for spectroscopic data using principal component regression. *Internet J. Chem.*, 2 (July 1999).

[14] R Ferrigno and H.H Girault. Finite element simulation of electrochemical ac diffusional impedance. Application to recessed microdiscs. *Journal of Electroanalytical Chemistry*, 492.1 (2000), pp. 1–6.

[15] Atsu Okabe, Barry Boots, Kokichi Sugihara, and Sung Chiu. Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. *Wiley Series in Probability and Mathematical Statistics, Chichester, New York: Wiley, 1992*, 43 (Jan. 2000).

[16] Yossi Rubner, Carlo Tomasi, and Leonidas Guibas. The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, 40 (Jan. 2000), pp. 99–121.

[17] Patrick R. Amestoy, Iain S. Duff, Jacko Koster, and Jean-Yves L'Excellent. A Fully Asynchronous Multifrontal Solver Using Distributed Dynamic Scheduling. *SIAM Journal on Matrix Analysis and Applications*, 23 (2001), pp. 15–41.

[18] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, 17 (Oct. 2001).

[19] Britta Lindholm-Sethson, Paul Geladi, and Andrew Nelson. Interaction with a phospholipid monolayer on a mercury electrode: Multivariate analysis of impedance data. *Analytica Chimica Acta*, 446.1 (2001). 7th International Conference on Chemometrics and Analytical Chemistry Antwerp, Belgium, 16-20 October 2000, pp. 121–130.

[20] David Ebert, F.K. Musgrave, D. Peachey, Ken Perlin, Steve Worley, W.R. Mark, and John Hart. *Texturing and Modeling: A Procedural Approach: Third Edition*. Dec. 2002, pp. 1–687.

[21] Jesper Møller and Rasmus Waagepetersen. Statistical Inference and Simulation for Spatial Point Process. 100 (Jan. 2004).

[22] Sarah Tilley, E.V. Orlova, Robert Gilbert, Peter Andrew, and Helen Saibil. Structural Basis of Pore Formation by the Bacterial Toxin Pneumolysin. *Cell*, 121 (May 2005), pp. 247–256.

[23] Philip Dixon. Ripley's K Function. Vol. 3. Sept. 2006.

[24] Yongtao Guan. A Composite Likelihood Approach in Fitting Spatial Point Process Models. *Journal of the American Statistical Association*, 101.476 (2006), pp. 1502–1512.

[25] S. Venkataraman, D.P. Allison, H. Qi, J.L. Morrell-Falvey, N.L. Kallewaard, J.E. Crowe, and M.J. Doktycz. Automated image analysis of atomic force microscopy images of rotavirus particles. *Ultramicroscopy*, 106.8 (2006). Proceedings of the Seventh International Conference on Scanning Probe Microscopy, Sensors and Nanostructures, pp. 829–837.

[26] Christian Wolf and Jean-Michel Jolion. Object count/area graphs for the evaluation of object detection and segmentation algorithms. *International Journal of Document Analysis and Recognition (IJDAR)*, 8.4 (2006), pp. 280–296.

[27] Duncan J. McGillivray et al. Molecular-scale structural and functional characterization of sparsely tethered bilayer lipid membranes. *Biointerphases*, 2 (2007), pp. 21–33.

[28] Jose Palacios Santander, Laura Cubillana-Aguilera, I. Naranjo-Rodríguez, and J.L. Hidalgo-Hidalgo-de-Cisneros. A chemometric strategy based on peak parameters to resolve overlapped electrochemical signals. *Chemometrics and Intelligent Laboratory Systems - CHEMOMETR INTELL LAB SYST*, 85 (Jan. 2007), pp. 131–139.

[29] Ignacio Castillo, Frank J. Kampas, and János D. Pintér. Solving circle packing problems by global optimization: Numerical results and industrial applications. *European Journal of Operational Research*, 191.3 (2008), pp. 786–802.

[30] Fred Lisdat and D. Schäfer. The use of electrochemical impedance spectroscopie for biosensing. *Analytical and Bioanalytical Chemistry*, 391 (Jan. 2008), p. 1555.

[31] Tao Sun, Nicolas G. Green, and Hywel Morgan. Analytical and numerical modeling methods for impedance analysis of single cells on-chip. *Nano*, 03.01 (2008), pp. 55–63.

[32] Sandro Cattarin, Marco Musiani, Bernard Tribollet, and Vincent Vivier. Impedance of passive oxide films with graded thickness: Influence of the electrode and cell geometry. *Electrochim. Acta*, 54 (2009), pp. 6963–6970.

[33] Patrick Frederix, Patrick Bosshart, and Andreas Engel. Atomic Force Microscopy of Biological Membranes. *Biophysical Journal*, 96 (Feb. 2009), pp. 329–338.

[34] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2009.

[35] Hai-Chau Chang and Lih-Chung Wang. A Simple Proof of Thue's Theorem on Circle Packing (Sept. 2010).

[36] Peter Eaton and Paul West. *Atomic Force Microscopy*. Mar. 2010.

[37] Kwang Joo Kwak et al. Formation and finite element analysis of tethered bilayer lipid structures. *Langmuir*, 26 (2010), pp. 18199–18208.

[38] Thet Naing Tun and A. Toby A. Jenkins. An electrochemical impedance study of the effect of pathogenic bacterial toxins on tethered bilayer lipid membrane. *Electrochemistry Communications*, 12.10 (2010), pp. 1411–1415.

[39] Wei Zhan et al. Photocurrent Generation from Porphyrin/Fullerene Complexes Assembled in a Tethered Lipid Bilayer. *Langmuir : the ACS journal of surfaces and colloids*, 26 (Oct. 2010), pp. 15671–9.

[40] William H. Hayt and John A. Buck. *Engineering Electromagnetics, 8th ed*. McGraw-Hill, 2012.

[41] Sarah Kißler, Sebastien Pierrat, Tom Zimmermann, Holger Vogt, Hoc-Khiem Trieu, and Ingo Köper. CMOS based capacitive biosensor with integrated tethered bilayer lipid membrane for real-time measurements. *Biomedical Engineering / Biomedizinische Technik*, 57.SI-1-Track-E (2012), pp. 1010–1013.

[42] R. Montoya, F.R. García-Galván, A. Jiménez-Morales, and J.C. Galván. Effect of conductivity and frequency on detection of heterogeneities in solid/liquid interfaces using local electrochemical impedance: Theoretical and experimental study. *Electrochemistry Communications*, 15.1 (2012), pp. 5–9.

[43] Andrew Mugler, Aimee Bailey, Koichi Takahashi, and Pieter Wolde. Membrane Clustering and the Role of Rebinding in Biochemical Signaling. *Biophysical journal*, 102 (Mar. 2012), pp. 1069–78.

[44] Christian Ulrich, Louthander Dan, Per Mårtensson, André Kluftinger, Michael Gawronski, and Fredrik Björefors. Evaluation of industrial cutting fluids using electrochemical impedance spectroscopy and multivariate data analysis. *Talanta*, 97 (2012), pp. 468–472.

[45] Gintaras Valincius, Tadas Meškauskas, and Feliksas Ivanauskas. Electrochemical Impedance Spectroscopy of Tethered Bilayer Membranes. *Langmuir*, 28 (2012), pp. 977–990.

[46] Rima Budvytyte, Milda Pleckaityte, Aurelija Zvirbliene, David J. Vanderah, and Gintaras Valincius. Reconstitution of Cholesterol-Dependent Vaginolysin into Tethered Phospholipid Bilayers: Implications for Bioanalysis. *PLOS ONE*, 8.12 (Dec. 2013), null.

[47] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Nov. 2013).

[48] Mats Larson and Fredrik Bengzon. *The Finite Element Method: Theory, Implementation, and Applications*. Vol. 10. Jan. 2013.

[49] Olek C. Zienkiewicz, Robert L. Taylor, and J. Z. Zhu. *The Finite Element Method: Its Basis and Fundamentals, 7th ed*. Butterworth-Heinemann, 2013.

[50] William F Ames. *Numerical methods for partial differential equations*. Academic press, 2014.

[51] Å. Björck. *Numerical Methods in Matrix Computations*. Texts in Applied Mathematics. Springer International Publishing, 2014.

[52] Chao Hu, Gaurav Jain, Puqiang Zhang, Craig Schmidt, Parthasarathy Gomadam, and Tom Gorka. Data-driven method based on particle swarm optimization and k-nearest neighbor regression for estimating capacity of lithium-ion battery. *Applied Energy*, 129 (2014), pp. 49–55.

[53] Carl Leung et al. Stepwise visualization of membrane pore formation by suilysin, a bacterial cholesterol-dependent cytolysin. *eLife*, 3 (Dec. 2014). Ed. by Volker Dötsch, e04247.

[54] Tsung-Yi Lin et al. Microsoft COCO: Common Objects in Context. *Computer Vision – ECCV 2014*. Ed. by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars. Cham: Springer International Publishing, 2014, pp. 740–755.

[55] Stéfan van der Walt et al. scikit-image: image processing in Python. *PeerJ*, 2 (June 2014), e453.

[56] Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial Point Patterns: Methodology and Applications with R*. London: Chapman and Hall/CRC Press, 2015.

[57] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE transactions on image processing*, 24.12 (2015), pp. 5706–5722.

[58] Ross Girshick. Fast r-cnn (Apr. 2015).

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *CoRR*, abs/1512.03385 (2015).

[60] Frédéric Lavancier, Jesper Møller, and Ege Rubak. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77.4 (2015), pp. 853–877.

[61] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. *CoRR*, abs/1512.02325 (2015).

[62] David Scott. *Multivariate density estimation: Theory, practice, and visualization: Second edition*. John Wiley & Sons, Mar. 2015, pp. 1–360.

[63] Gintaras Valincius and Mindaugas Mickevicius. "Tethered Phospholipid Bilayer Membranes: An Interpretation of the Electrochemical Impedance Response". *Advances in Planar Lipid Bilayers and Liposomes*. Ed. by Aleš Iglič, Chandrashekhar V. Kulkarni, and Michael Rappolt. Vol. 21. Academic Press, 2015. Chap. 2, pp. 27–61.

[64] Gong Cheng and Junwei Han. A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117 (2016), pp. 11–28.

[65] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016.

[66] H. Igel. *Computational Seismology: A Practical Introduction*. OUP Oxford, 2016.

[67] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander Berg. SSD: Single Shot MultiBox Detector. Vol. 9905. Oct. 2016, pp. 21–37.

[68] Giulio Preta, Marija Jankunec, Frank Heinrich, Sholeem Griffin, Iain Martin Sheldon, and Gintaras Valincius. Tethered bilayer membranes as a complementary tool for functional and structural studies: The pyolysin case. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1858.9 (2016), pp. 2070–2080.

[69] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. June 2016, pp. 779–788.

[70] Solomon Tesfalidet, Paul Geladi, Kenichi Shimizu, and Britta Lindholm-Sethson. Detection of methotrexate in a flow system using electrochemical impedance spectroscopy and multivariate data analysis. *Analytica Chimica Acta*, 914 (2016), pp. 1–6.

[71] Gintaras Valincius, Mindaugas Mickevicius, Tadas Penkauskas, and Marija Jankunec. Electrochemical Impedance Spectroscopy of Tethered Bilayer Membranes: An Effect of Heterogeneous Distribution of Defects in Membranes. *Electrochim. Acta*, 222 (2016), pp. 904–913.

[72] Matthieu Clavaud, Yves Roggo, Klara Dégardin, Pierre-Yves Sacré, Philippe Hubert, and Eric Ziemons. Global regression model for moisture content determination using near-infrared spectroscopy. *European Journal of Pharmaceutics and Biopharmaceutics*, 119 (2017), pp. 343–352.

[73] Chirantan Das, Subhadip Chakraborty, Krishnendu Acharya, Nirmal Kumar Bera, Dipankar Chattopadhyay, Anupam Karmakar, and Sanatan Chattopadhyay. FT-MIR supported Electrical Impedance Spectroscopy based study of sugar adulterated honeys from different floral origin. *Talanta*, 171 (2017), pp. 327–334.

[74] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. Oct. 2017, pp. 2980–2988.

[75] Muhammad Shuja Khan, Noura Sayed Dosoky, Darayas Patel, Jeffrey Weimer, and John Dalton Williams. Lipid Bilayer Membrane in a Silicon Based Micron Sized Cavity Accessed by Atomic Force Microscopy and Electrochemical Impedance Spectroscopy. *Biosensors*, 7.3 (2017).

[76] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2999–3007.

[77] Tadas Ragaliauskas, Mindaugas Mickevicius, Bozena Rakovska, Tadas Penkauskas, David J. Vanderah, Frank Heinrich, and Gintaras Valincius. Fast formation of low-defect-density tethered bilayers by fusion of multilamellar vesicles. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1859.5 (2017), pp. 669–678.

[78] C. Siontorou, Georgia-Paraskevi Nikoleli, D. Nikolelis, and Stefanos K Karapetis. Artificial Lipid Membranes: Past, Present, and Future. *Membranes*, 7 (2017).

[79] Varun Jain, Mark C. Biesinger, and Matthew R. Linford. The Gaussian-Lorentzian Sum, Product, and Convolution (Voigt) functions in the context of peak fitting X-ray photoelectron spectroscopy (XPS) narrow scans. *Applied Surface Science*, 447 (2018), pp. 548–553.

[80] T.N. Krishnamupti and Lahouari Bounoua. *An Introduction to Numerical Weather Prediction Techniques*. May 2018.

[81] Brendan Marsh, Nagaraju Chada, Raghavendar Gari, Krishna Sigdel, and Gavin King. The Hessian Blob Algorithm: Precise Particle Detection in Atomic Force Microscopy Imagery. *Scientific Reports*, 8 (Jan. 2018).

[82] Yingchao Meng, Zhongping Zhang, Huaqiang Yin, and Tao Ma. Automatic detection of particle size distribution by image analysis based on local adaptive canny edge detection and modified circular Hough transform. *Micron*, 106 (2018), pp. 34–41.

[83] Melissa Piontek and Wouter Roos. Atomic Force Microscopy: An Introduction. Vol. 1665. Jan. 2018, pp. 243–258.

[84] M. Ewald et al. High speed atomic force microscopy to investigate the interactions between toxic Aβ1-42 peptides and model membranes in real time: impact of the membrane composition. *Nanoscale*, 11 (15 2019), pp. 7229–7238.

[85] Derek M. Hall, Timothy Duffy, Margaret Ziomek-Moroz, and Serguei N. Lvov. Electrochemical impedance spectroscopy and finite element analysis modeling of a 4-electrode humidity sensor for natural gas transportation pipelines. *Review of Scientific Instruments*, 90.1 (2019), p. 015005.

[86] Licheng Jiao, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng, and Rong Qu. A Survey of Deep Learning-Based Object Detection. *IEEE Access*, 7 (2019), pp. 128837–128868.

[87] Georgia-Paraskevi Nikoleli, Christina G. Siontorou, Marianna-Thalia Nikolelis, Spyridoula Bratakou, and Dimitrios K. Bendos. Recent Lipid Membrane-Based Biosensing Platforms. *Applied Sciences*, 9.9 (2019).

[88] Tadas Penkauskas and Giulio Preta. Biological applications of tethered bilayer lipid membranes. *Biochimie*, 157 (2019), pp. 131–141.

[89] Dayvison Ribeiro Rodrigues, Alejandro César Olivieri, Wallace Duarte Fragoso, and Sherlan Guimarães Lemos. Complex numbers-partial least-squares applied to the treatment of electrochemical impedance spectroscopy data. *Analytica Chimica Acta*, 1080 (2019), pp. 1–11.

[90] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object Detection in 20 Years: A Survey. *CoRR*, abs/1905.05055 (2019).

[91] Julia Alvarez-Malmagro, Gabriel García-Molina, and Antonio López De Lacey. Electrochemical Biosensors Based on Membrane-Bound Enzymes in Biomimetic Configurations. *Sensors*, 20.12 (2020).

[92] Xiaoyu Gong, Chaofang Dong, Jiajin Xu, Li Wang, and Xiaogang Li. Machine learning assistance for electrochemical curve simulation of corrosion and its application. *Materials and Corrosion*, 71.3 (2020), pp. 474–484.

[93] Alexey G. Okunev, Mikhail Yu. Mashukov, Anna V. Nartova, and Andrey V. Matveev. Nanoparticle Recognition on Scanning Probe Microscopy Images Using Computer Vision and Deep Learning. *Nanomaterials*, 10.7 (2020).

[94] Azin Sadat and Iris J. Joye. Peak Fitting Applied to Fourier Transform Infrared and Raman Spectroscopic Analysis of Proteins. *Applied Sciences*, 10.17 (2020).

[95] Shinya Tanaka, Kaiken Kimura, Ko-ichiro Miyamoto, Yuhki Yanase, and Shigeyasu Uno. Simulation and Experiment for Electrode Coverage Evaluation by Electrochemical Impedance Spectroscopy using Parallel Facing Electrodes. *Analytical Sciences*, advpub (2020).

[96] Joseph G. Beton et al. TopoStats – A program for automated tracing of biomolecules from AFM images. *Methods* (2021).

[97] OpenVINO. *SSD ResNet50 V1 FPN COCO*. `https://docs.openvinotoolkit.org/latest/omz_models_model_ssd_resnet50_v1_fpn_coco.html`. [Online; accessed 7-July-2021]. 2021.

[98] Javier Sotres, Hannah Boyd, and Juan F. Gonzalez-Martinez. Enabling autonomous scanning probe microscopy imaging of single molecules with deep learning. *Nanoscale*, 13 (20 2021), pp. 9193–9203.

Appendix 1

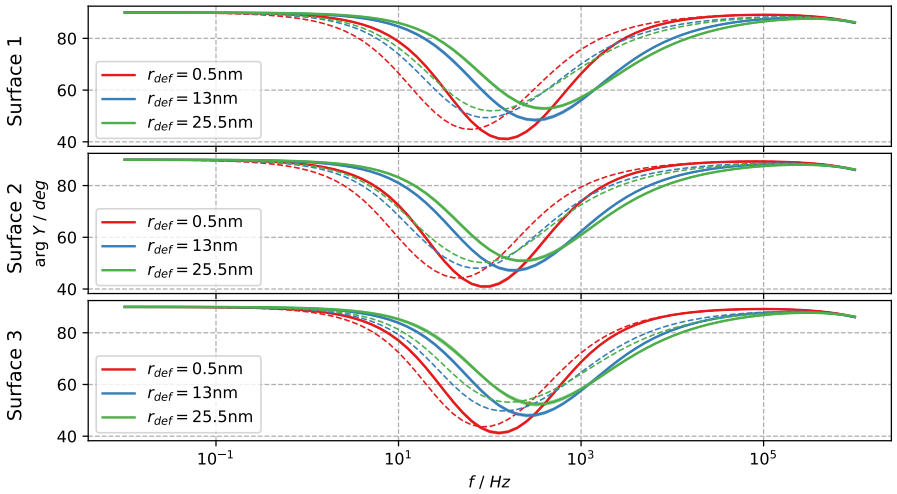


Figure 57: Comparison of EIS spectra modeled using AFM-measured defect sets (dashed curves) and instances of random defect distribution model (solid bands).

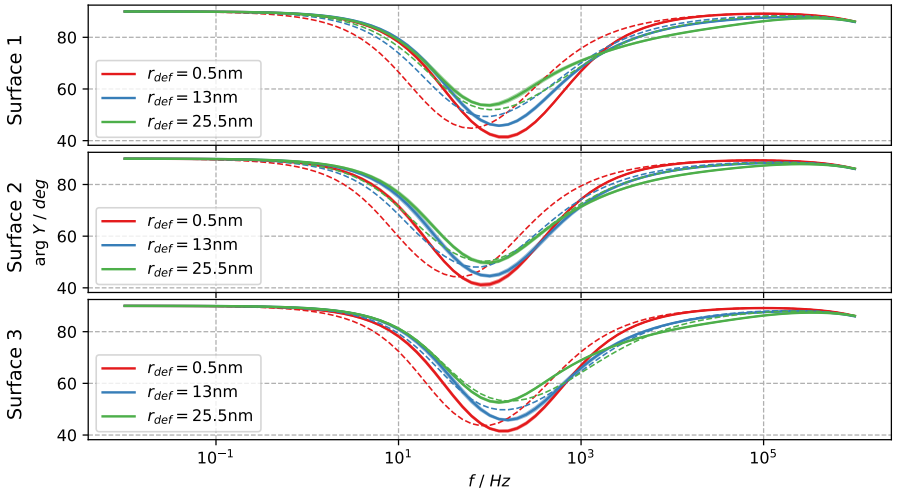


Figure 58: Comparison of EIS spectra modeled using AFM-measured defect sets (dashed curves) and instances of attraction clustering model (solid bands).

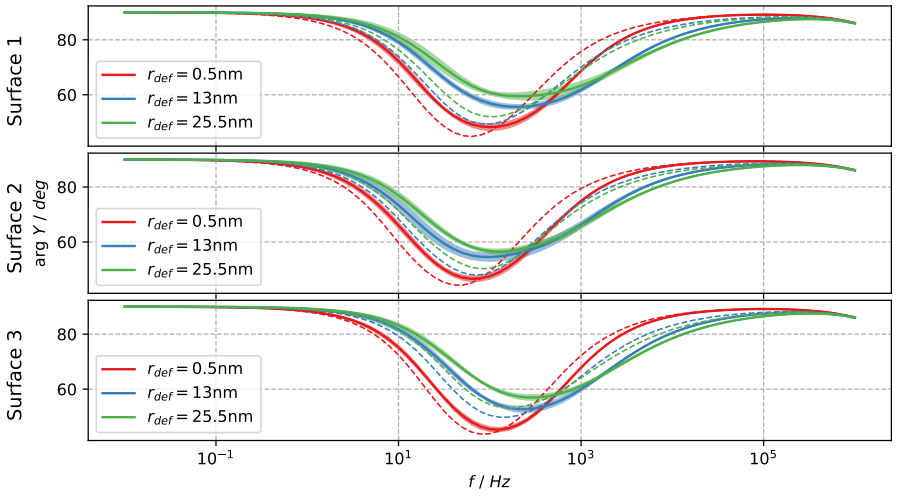


Figure 59: Comparison of EIS spectra modeled using AFM-measured defect sets (dashed curves) and instances of LCN clustering model (solid bands).

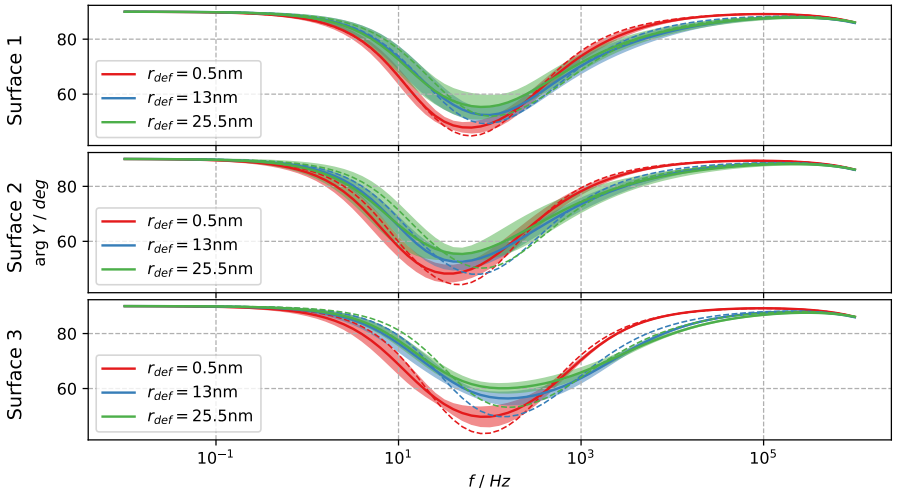


Figure 60: Comparison of EIS spectra modeled using AFM-measured defect sets (dashed curves) and instances of point process clustering model (solid bands), with parameters determined by histogram comparison.

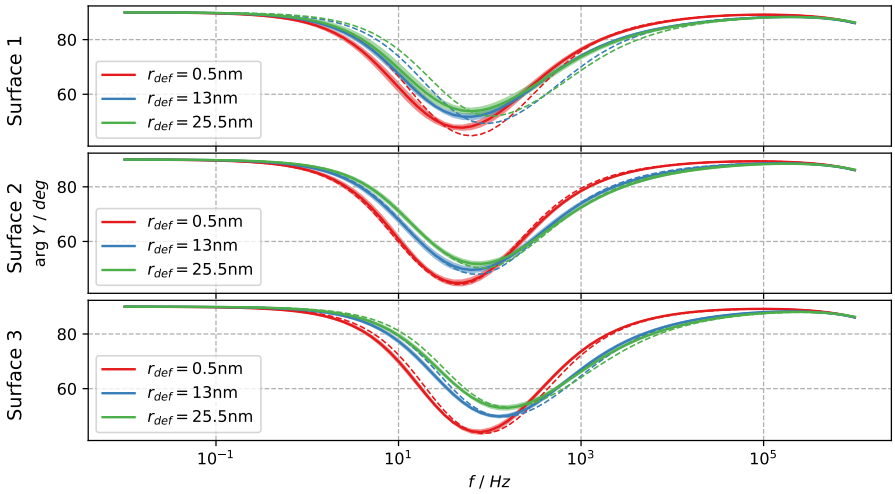


Figure 61: Comparison of EIS spectra modeled using AFM-measured defect sets (dashed curves) and instances of point process clustering model (solid bands), with parameters determined by minimum contrast method.

Appendix 2

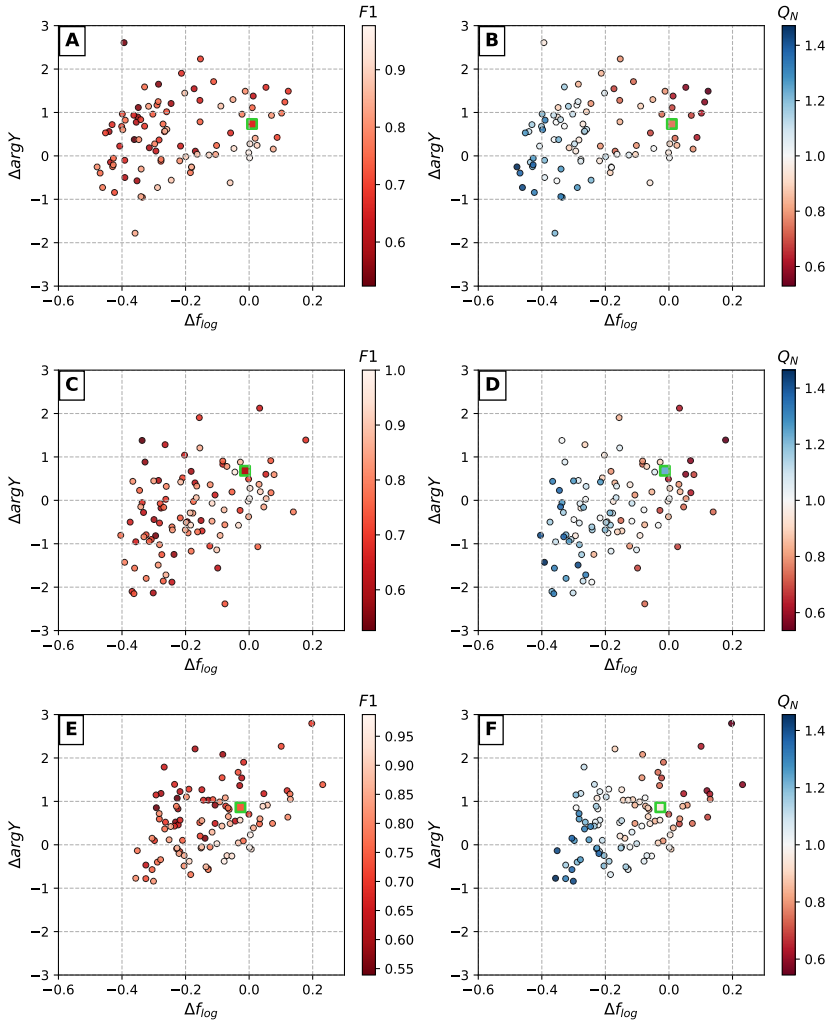


Figure 62: Dependencies between defect detection accuracy (expressed in terms of $F1$ and Q_N) and deviations in corresponding EIS spectra. Coloured dots represent synthetically generated non-clustered defect sets at varying detection accuracy levels, squares with green borders indicate real detection results obtained with CNN model (Table 41). Scatter plot pairs A/B, C/D and E/F represent AFM surfaces 1, 2 and 3 respectively.

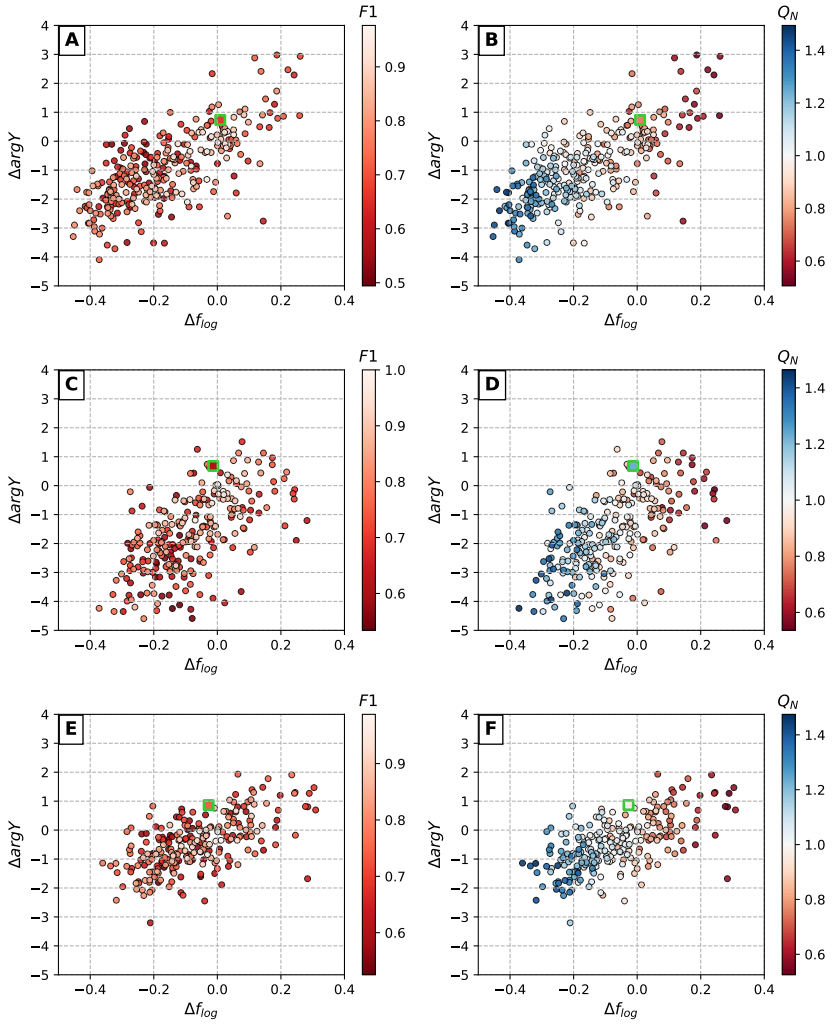


Figure 63: Dependencies between defect detection accuracy (expressed in terms of $F1$ and Q_N) and deviations in corresponding EIS spectra. Coloured dots represent synthetically generated clustered defect sets at varying detection accuracy levels, squares with green borders indicate real detection results obtained with CNN model (Table 41). Scatter plot pairs A/B, C/D and E/F represent AFM surfaces 1, 2 and 3 respectively.

Santrauka (Summary in Lithuanian)

Tyrimų sritis

Disertacijoje pristatomo tarpdisciplininio tyrimo sritis - prikabintų dvisluoksnių fosfolipidinių membranų (tBLM) pažaidos įvertinimas, taikant kompiuterinio modeliavimo metodus. Laboratorinėmis sąlygomis šio tipo membranos tiriamos taikant tokius eksperimentinius metodus, kaip atominės jėgos mikroskopija (AJM) bei elektrocheminio impedanso spektroskopija (EIS). Gaunamų šių tipų duomenų analizėje yra svarbūs informatikos mokslo metodai, leidžiantys tiksliai ir efektyviai interpretuoti duomenis bei kiekybiškai arba kokybiškai įvertinti membranų savybes. Šiame darbe nagrinėjami tBLM membranų EIS ir AJM duomenų analizės ir modeliavimo metodai, paremti baigtinių elementų metodu, mašininio mokymosi metodais bei kitais skaitiniais algoritmais.

Tyrimo tikslas ir uždaviniai

Pagrindinis tyrimo tikslas – sukurti metodiką, leidžiančią modeliuoti trimačių tBLM membranų su įvairiai išsidėsčiusiais defektais elektrocheminį atsaką, ir interpretuoti EIS duomenis naudojant mašininio mokymosi metodus, siekiant įvertinti kokybines ir kiekybines membranos pažeidimo savybes. Sprendžiami šie uždaviniai:

1. Sukurti trimatį skaitinį modelį, leidžiantį simuliuoti tBLM membranų EIS spektrus, esant bet kokiam defektų išsidėstymui.
2. Naudojant mašininio mokymosi metodus sudaryti prognozavimo modelius, skirtus kiekybinių tBLM membranų charakteristikoms įvertinti pagal jų EIS spektrus.
3. Sukurti defektų išsidėstymo modelius, leidžiančius kompiuteriu generuoti tikroviškus defektų rinkinius, apibrėžti jų palyginimo metrikas ir ištirti modelių ryšį su modeliuotais EIS duomenimis.
4. Ištirti AJM vaizdo duomenų automatizuoto defektų aptikimo ir jų tikslumo įvertinimo metodus, įvertinti veikimo įtaką modeliuojamiems EIS spektrams.
5. Patvirtinti siūlomus metodus ir sintetinius modeliavimo duomenis, palyginant juos su eksperimentiniu būdu gautais EIS ir AJM matavimų duomenimis.

Tyrimo metodai ir priemonės

Trimatis tBLM membranos modelis įgyvendintas naudojant baigtinių elementų metodą ir COMSOL Multiphysics paketą (5.3–5.4 versijos), specialus modelių rengimo įrankis (naudojantis COMSOL API) realizuotas Java kalba. Daugumą duomenų analizei naudotų programų parašytos naudojant Python (3.7 versija), pagrindines mokslines bibliotekas (Numpy, SciPy, Pandas, Matplotlib), mašininio mokymosi bibliotekas (scikit-learn, Tensorflow) ir Jupyter Notebook aplinką.

Mokslinis naujumas ir praktinė reikšmė

Darbe aprašoma metodika yra nauja savo galimybėmis, pritaikomomis skirtingomis savybėmis pasižyminčių tBLM membranų modelių analizei ir jų elektrocheminiam atsakui interpretuoti, siekiant įvertinti įvairias kokybines ir kiekybines membranų charakteristikas. Mokslinį naujumą pagrindžia šie esminiai rezultatai:

1. Taikant baigtinių elementų metodą pirmą kartą įgyvendintas trimatis tBLM membranos modelis, leidžiantis modeliuoti sistemos EIS spektrą esant bet kokiam defektų išsidėstymui.

2. Sukurti nauji defektų klasterizacijos modeliai ir parodyta, kad jie gali generuoti realistiškus defektų rinkinius įvairiais klasterizacijos lygiais. Taip pat apibrėžtos metrikos, skirtos klasterizacijos efektui įvertinti iš defektų rinkinių.

3. Sukurti ir palyginti įvairūs automatinio defektų aptikimo AFM vaizduose algoritmai ir pirmą kartą ištirtas ryšys tarp jų tikslumo ir modeliuojamų EIS spektrų.

Praktiniu aspektu metodika galėtų būti taikoma greitam kiekybiniam membranos pažaidos įvertinimui tBLM pagrindu veikiančiuose impedanso biojutikliuose ar kitose panašiose sistemose. Ji taip pat leidžia įvertinti tam tikras membranos savybes (pvz., pomembraninio sluoksnio savitąjį laidumą arba membranos defektų klasterizaciją), kurių negalima išmatuoti tiesiogiai EIS arba AJM metodais. Automatinio defektų aptikimo AJM vaizduose metodai galėtų būti naudingi tyrėjams, dirbantiems šioje srityje, ir padaryti šių duomenų analizės procesą greitesnį ir tikslesnį.

Ginamieji teiginiai

1. Fosfolipidinių membranų su defektais kiekybines savybes galima įvertinti pagal jų elektrocheminio impedanso spektrus, naudojant baigtinių elementų modeliavimo ir mašininio mokymosi metodus.

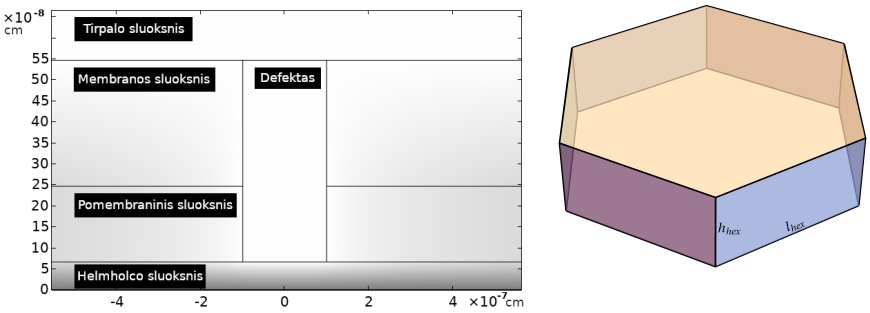
2. Defektų klasterizacijos reiškinius fosfolipidinėse membranose galima aprašyti skaičiavimo modeliais, kurie leidžia kiekybiškai įvertinti klasterizacijos efektą iš atominės jėgos mikroskopijos vaizdų arba elektrocheminio impedanso spektrų.

3. Kompiuterinės regos metodai gali būti taikomi automatizuotam membranos defektų aptikimui modeliavimo tikslams pakankamu tikslumu atominės jėgos mikroskopijos vaizduose.

S.1. Pažeistų fosfolipidinių membranų electrocheminio atsako modeliavimas

S.1.1. Trimatis membranos modelis

Šiame darbe nagrinėjamo trimačio membranos modelio paskirtis – simuliuoti kintamosios elektros srovės tekėjimą per membraną, turinčią tam tikru būdu išsidėsčiusių defektų, ir įvertinti elektrocheminį sistemos atsaką srovės dažnio atžvilgiu. Modelis susideda iš keturių sluoksnių, atitinkančių (iš viršaus į apačią) tirpalą, membraną, pomembraninį sluoksnį bei Helmholco sluoksnį, kartu su pasirinktu membranos defektų kiekiu (S.1 pav., kairėje). Tirpalo, pomembraniniai sluoksniai ir defektai yra laidūs elektros srovei, o membranos ir Helmholco sluoksniai – nelaidūs (dielektrikai). Panašios struktūros tBLM membranos modelis buvo nagrinėjamas ankstesniame darbe [37].



S.1 pav.: Trimačio tBLM membranos modelio schema. Kairėje: modelio skerspjūvis defekto aplinkoje. Dešinėje: šešiakampė prizmė, atitinkanti modeliavimo sritį.

Modeliavimo sritis, susidedanti iš aprašytų sluoksnių, atitinka šešiakampę prizmę (S.1 pav., dešinėje) su briaunos ilgiu l_{hex}. Prizmės aukštis h_{hex} atitinka Helmholco, pomembraninio sluoksnio, membranos ir tirpalo sluoksnių aukščių sumą:

$$h_{hex} = d_H + d_{sub} + d_m + d_{sol}. \tag{S.1}$$

Prizmės viršutinės arba apatinės plokštumos (šešiakampio) plotas S_{hex} apibrėžiamas:

$$S_{hex} = \frac{3\sqrt{3}(l_{hex})^2}{2}. \tag{S.2}$$

Konkrečiu atveju modelis gali turėti N membranos defektų, išsidėsčiusių norima tvarka. Kiekvienas defektas apibrėžiamas kaip cilindras su spinduliu r_{def}, kertantis membranos bei pomembraninį sluoksnius ir turintis fiksuotą aukštį $d_m + d_{sub}$. Defektų tankis N_{def} atitinka defektų skaičių viename kvadratiniame mikrometre:

$$N_{def} = \frac{N}{S_{hex}}. \tag{S.3}$$

Kintamosios elektros srovės tekėjimas per membraną yra apibrėžiamas Laplaso lygtimi, kur lygties sprendinys yra kompleksinė įtampa Φ, atitinkanti trimatėje modelio srityje apibrėžtą funkciją:

$$\nabla \cdot (\tilde{\sigma}(x,y,z)\nabla\Phi(x,y,z)) = 0, \tag{S.4}$$

Čia $\tilde{\sigma}$ žymi kompleksinį laidumą tam tikrame srities taške:

$$\tilde{\sigma}(x,y,z) = \sigma(x,y,z) + j\,\omega\varepsilon(x,y,z). \tag{S.5}$$

Realioji ir menamoji dalys (menamasis vienetas žymimas j) atitinka elektrinį laidumą bei dielektrinę skvarbą skirtingose sistemos dalyse. Elektrinio laidumo konstanta σ galioja laidiems modelio sluoksniams (tirpalas, pomembraninis sluoksnis ir defektai), o dielektrinė skvarba ε apibrėžia dielektrinių sluoksnių (membranos ir Helmholco) savybes. $\omega = 2\pi f$ žymi kampinį kintamosios elektros srovės dažnį, kur f – dažnis hercais (Hz).

Laikoma, kad šešiakampės modelio prizmės viršuje yra fiksuotas 1 V elektrinis potencialas, o po Helmholco sluoksniu potencialas lygus 0 V. Šios prielaidos modelyje išreiškiamos kaip Dirichlė kraštinės sąlygos:

$$\Phi(x,y,h_{hex}) = 1, \tag{S.6}$$

$$\Phi(x,y,0) = 0. \tag{S.7}$$

Taip pat laikoma, kad šešiakampės prizmės šonai yra nelaidūs srovei, apibrėžiant atitinkamą kraštinės sąlygą, kur n – prizmės šoninės sienos normalinis vektorius:

$$n \cdot \nabla\Phi(x,y,z) = 0. \tag{S.8}$$

Išsprendus (S.4) lygtį kintamosios srovės Φ atžvilgiu, modeliavimo srityje apskaičiuojamas srovės tankis J, taikant Omo dėsnį [40]:

$$J(x,y,z) = -\tilde{\sigma}(x,y,z)\,\nabla\Phi(x,y,z). \tag{S.9}$$

Pagal srovės tankio reikšmes modeliavimo srities viršutinėje plokštumoje skaičiuojamas admitansas Y. Tai – atvirkštinis impedansui kompleksinis dydis, nusakantis, kaip lengvai sistema praleidžia elektros srovę:

$$Y = \frac{\iint\limits_{(x,y)\in\Gamma_{hex}} -n \cdot J(x,y,h_{hex})\,dx\,dy}{S_{hex}} \times \frac{1}{\Phi(x,y,h_{hex})}. \tag{S.10}$$

Sprendžiant aprašytą modeliavimo uždavinį su diskrečiomis srovės dažnio reikšmėmis pasirinktame intervale, sumodeliuojamas išsamus sistemos elektrocheminis atsakas. Visiems šiame darbe aprašomiems modeliavimo atvejams naudotas dažnių intervalas nuo 10^{-2} Hz iki 10^{6} Hz, reikšmes išdėstant logaritmiškai po 10 taškų dekadai (iš viso 81 dažnio reikšmė).

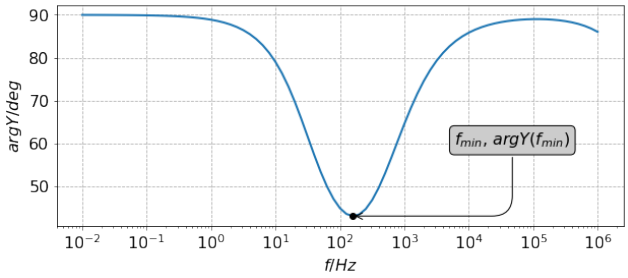
Sumodeliuotas elektrocheminis atsakas išreiškiamas kaip admitanso fazės (laipsniais) priklausomybė nuo dažnio (S.2 pav.). Tokių reikšmių rinkinys toliau įvardijimas kaip elektrocheminio impedanso arba EIS spektras. Šiame ir

ankstesniuose susijusiuose darbuose [71, 46] nagrinėjami EIS spektrai pasižymi charakteringa kreivės forma, turinčią vieną minimumo tašką. Ankstesniame tyrime įrodyta, kad šio taško koordinatės yra susijusios su įvairiais kiekybiniais pažeistos membranos parametrais, tokiais kaip defektų tankis ar dydis [45]. Šiems spektriniams požymiams naudojami toliau nurodyti žymėjimai:

$$f_{min} \quad \text{– dažnis } f\text{, kuriame } \arg Y(f) \text{ įgyja mažiausią reikšmę,} \tag{S.11}$$

$$\arg Y_{min} \text{ – admitanso fazės reikšmė taške } f_{min}. \tag{S.12}$$

Aprašytas trimatis membranos modelis realizuotas taikant baigtinių elementų metodą (BEM) ir naudojant COMSOL Multiphysics programinį paketą. Modeliavimo sritis diskretizuojama naudojant tetraedrų arba prizmių elementus, sudaryta tiesinių lygčių sistema sprendžiama tiesioginiu algoritmu (angl. direct solver) MUMPS [17]. Skaičiavimai lygiagretinami pagal srovės dažnio f parametrą.

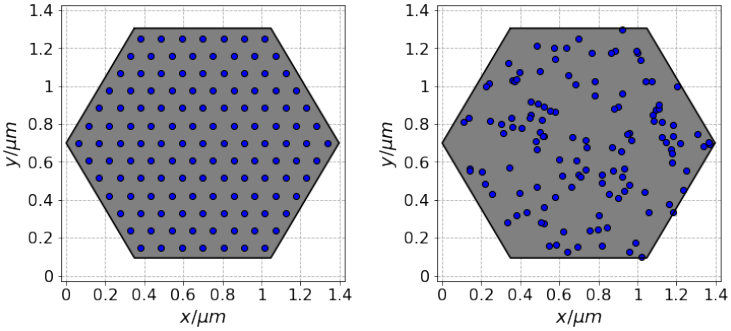


S.2 pav.: Sumodeliuotas EIS spektras ir jo požymiai.

Membranos modelyje defektai gali būti išsidėstę bet kokia norima tvarka. Konkretų defektų rinkinį sugeneruoja tam tikras algoritmas, parenkantis defektų centrų koordinates pagal nustatytas defektų kiekio N ir defektų tankio N_{def} reikšmes. Šiame skyriuje nagrinėjami du defektų rinkinių modeliai (algoritmai):

- Tolygus defektų išsidėstymas (S.3 pav., kairėje). Šiame modelyje defektai yra išdėstomi vienodais tarpusavio atstumais, taip sudarant reguliarų tinklelį. Membranų modeliai su tokiu defektų išdėstymu yra analogiški nagrinėtiems ankstesniame darbe, kuriame membranos elektrocheminis atsakas buvo modeliuojamas analitiškai [45].

- Atsitiktinis defektų išsidėstymas (S.3 pav., dešinėje). Laikant, kad defektų tarpusavio sąveikos nėra, kiekvieno atskiro defekto koordinatės parenkamos atsitiktinai iš tolygiojo skirstinio, kurio intervalas atitinka modeliuojamos srities išmatavimus.



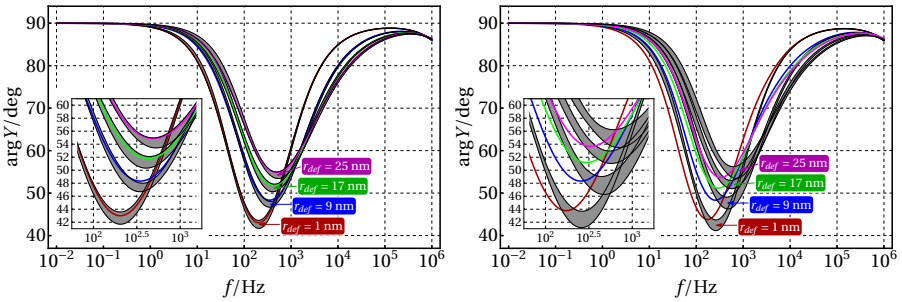
S.3 pav.: Defektų išsidėstymo membranos modelyje pavyzdžiai ($N = 127$). Kairėje: tolygus defektų išsidėstymas. Dešinėje: atsitiktinis defektų išsidėstymas.

S.1.2. Palyginimas su eksperimentiniais AJM duomenimis

Norint įvertinti, kaip atsitiktinis defektų išsidėstymo modelis atitinka faktinėse tBLM membranose pastebėtų defektų padėtis, buvo lyginami kompiuteriu sugeneruotų ir eksperimentiškai išmatuotų defektų rinkinių EIS spektrai. Membranos defektus rankiniu būdu anotavo srities ekspertas, o jų koordinatės (šešiakampėje modelio srityje) buvo naudojamos EIS spektrams modeliuoti. Pirmasis modelis (be klasterių) turėjo $N = 74$ defektų esant $N_{def} = 12,66$ tankiui, o antrasis modelis (su klasteriais) turėjo $N = 41$ defektus esant $N_{def} = 15,78$ tankiui. Be to, sugeneruoti atsitiktinių defektų rinkiniai (10 atvejų kiekvienam parametrų deriniui), naudojant tuos pačius dviejų vaizdų defektų kiekius N ir tankius N_{def}, kad būtų gauti papildomi modeliai, besiskiriantys tik tiksliomis defektų padėtimis. Visų atvejų modeliavimas atliktas naudojant keturis skirtingus defektų spindulius $r_{def} = 1,9,17,25$ nm, atitinkančius labiausiai tikėtinus defektų dydžius, remiantis ankstesniais tyrimais [22].

EIS spektrų, gautų iš eksperimentinių ir kompiuteriu sugeneruotų atsitiktinių defektų rinkinių, palyginimas pateiktas S.4 pav. Neklasterizuoto defektų pasiskirstymo spektrai rodo gerą sutapimą su atitinkamais atsitiktiniais atvejais visuose r_{def} lygiuose. Tačiau akivaizdus neatitikimas tarp faktinių AFM

registruotų defektų koordinačių ir atsitiktinio defektų pasiskirstymo modelio matomas defektų klasterizacijos atveju, kur $\log f_{min}$ poslinkis yra nuo $0,17$ iki $0,30$, o $\arg Y_{min}$ skirtumas svyruoja nuo $1,35$ iki $-1,21$, didėjant r_{def}.



S.4 pav.: Sumodeliuoti kompiuteriu sugeneruotų atsitiktinių defektų pasiskirstymų (pilkos juostos) ir eksperimentiškai registruotų defektų rinkinių (spalvotos kreivės), EIS spektrai. Kairėje: neklasterizuotas defektų pasiskirstymas. Dešinėje: klasterizuotas defektų pasiskirstymas.

S.1.3. Membranos parametrų įvertinimas pagal EIS spektrus

Duomenų rinkinys membranos parametrų prognozavimo eksperimentams sudarytas modeliuojant EIS spektrus su įvairiomis N_{def}, r_{def} ir ρ_{sub} parametrų kombinacijomis. Kiekvienu atveju fiksuoto kiekio defektų ($N = 200$) koordinatės sugeneruotos pagal atsitiktinio defektų pasiskirstymo modelį. Iš viso sugeneruota po 10 tokių atvejų kiekvienam iš 546 unikalių parametrų derinių. Tada kiekvienam atvejui atliktas baigtinių elementų modeliavimas ir apskaičiuoti EIS spektrai. Šis duomenų rinkinys buvo naudojamas skirtingiems regresijos modeliams (tiesinė, Lasso, PCR, PLS, artimiausių kaimynų regresija) palyginti ir jų prognozavimo tikslumui įvertinti.

Kiekvienas regresijos modelis buvo validuojamas atliekant 10 dalių kryžminę patikrą (angl. 10-fold cross-validation). Laikant kiekvieną iš 546 unikalių parametrų derinių ir atitinkamų 10 modelių pavyzdžių vientisa pavyzdžių grupe, kiekvieno validavimo etapo mokymo ir validavimo rinkiniai sudaryti taip, kad bet kurios konkrečios grupės pavyzdžiai nebūtų padalyti tarp abiejų rinkinių. Kiekviename eksperimente buvo vertinami du atskiri regresijos modeliai numatant v_1 ir v_2 koeficientus (plačiau aprašyti disertacijos tekste). Modelių tikslumas buvo vertinamas pagal R^2 koeficientą, apskaičiuojant validavimo reikšmių vidurkį ir standartinį nuokrypį. S.1 lentelėje parodytas tiesinės regresijos modelių, apmokytų pagal du spektrinius požymius ($\log f_{min}$ ir $\arg Y_{min}$) ir

iš jų išvestus papildomus polinominius požymius (iki 3 laipsnio), tikslumas.

S.1 lentelė: Tiesinės regresijos modelių su polinominiais požymiais kryžminės patikros rezultatai. SN žymimas standartinis nuokrypis.

Poly. degree	Features	v_1		v_2	
		Mean	Stdev	Mean	Stdev
1	2	0.999	0.000	0.564	0.040
2	5	0.999	0.000	0.829	0.019
3	9	0.999	0.000	0.880	0.013

Naudojant iš EIS spektrų prognozuojamas koeficientų v_1 ir v_2 reikšmes papildomai įvertintas membranos parametrų N_{def}, r_{def} ir ρ_{sub} prognozavimo tikslumas. S.2 lentelėje parodytos MAE ir MAPE reikšmės kiekvienam membranos parametrui, kur vienas iš likusių dviejų parametrų laikomas žinomu. N_{def} ir $\log\rho_{sub}$ prognozavimo tikslumas priklauso nuo to, kuris iš kitų dviejų parametrų yra fiksuotas – pasirinkus r_{def} abiem atvejais gaunami žymiai mažiau tikslūs santykinės paklaidos įverčiai, o pagal ρ_{sub} prognozuojant N_{def} (arba atvirkščiai) gaunami geriausi rezultatai.

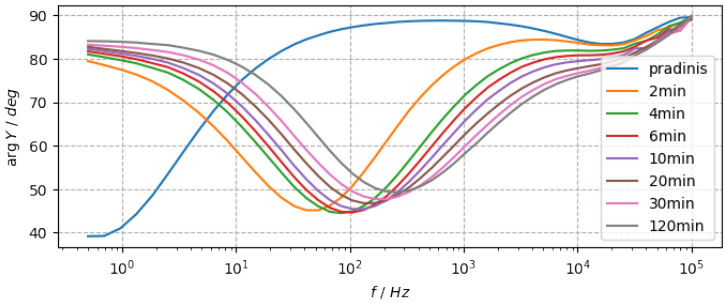
S.2 lentelė: Membranos parametrų nustatymo tikslumas naudojant tiesinės regresijos modelius.

Fixed parameter	N_{def}		r_{def}		$\log\rho_{sub}$	
	MAE	MAPE	MAE	MAPE	MAE	MAPE
N_{def}	-	-	2.607	28.320	0.034	0.762
r_{def}	6.028	77.580	-	-	0.246	5.507
ρ_{sub}	1.349	7.842	2.965	32.184	-	-

S.1.4. Membranos parametrų numatymas iš eksperimentinių EIS spektrų

Aprašyta kiekybinių membranų savybių prognozavimo metodika patvirtinta naudojant eksperimentinius EIS duomenis. Duomenų rinkinys (S.5 pav.) gautas eksperimentinėmis sąlygomis, kai surinktos tBLM buvo veikiamos tirpalu, kuriame yra poras formuojančio toksino vaginolizino (VLY) [46], ir atlikti atskiri EIS matavimai skirtingais laikotarpiais, praėjusiais po ekspozicijos. Tik-

rosios N_{def}, r_{def} arba ρ_{sub} reikšmės nebuvo žinomos dėl EIS matavimo metodo pobūdžio bei specifinių eksperimentinių sąlygų, kai nebuvo taikomi atskiri metodai šioms savybėms nepriklausomai išmatuoti.



S.5 pav.: Eksperimentiniai EIS duomenys, išmatuoti tBLM mėginiui skirtingu laiku (nurodyta legendoje), kai jis paveiktas poras formuojančiu toksinu.

Kiekybinėms membranų savybėms prognozuoti iš eksperimentinių EIS spektrų regresijos modeliai buvo mokomi naudojant modelių duomenų rinkinį, aprašytą S.1.3 skyrelyje, o konkretūs modeliai parinkti pagal anksčiau pateiktą prognozavimo tikslumą. Numatytos v_1 ir v_2 reikšmės panaudotos N_{def} ir r_{def} įverčiams apskaičiuoti, laikant ρ_{sub} reikšmę fiksuotą ir lygią $10^5\ \Omega \cdot cm$. S.3 lentelėje pateikiamos apskaičiuotos parametrų reikšmės kiekvienam eksperimentiniam EIS spektrui (išskyrus pradinį matavimą).

S.3 lentelė: Membranos parametrų įverčiai, numatyti iš eksperimentinių EIS spektrų, naudojant tiesinės regresijos modelius.

Time (min.)	N_{def}	r_{def}
2	2.019	10.795
4	3.696	6.300
6	4.482	6.036
10	4.990	7.750
20	5.458	11.956
30	6.601	14.225
120	7.863	19.305

N_{def} įvertinimai rodo monotonišką padidėjimą, atitinkantį eksperimentines membranos pažeidimo, besikaupiančio laikui bėgant (dėl ilgalaikio kontakto su poras formuojančiu toksinu), sąlygas. N_{def} prognozių diapazonas yra tokio

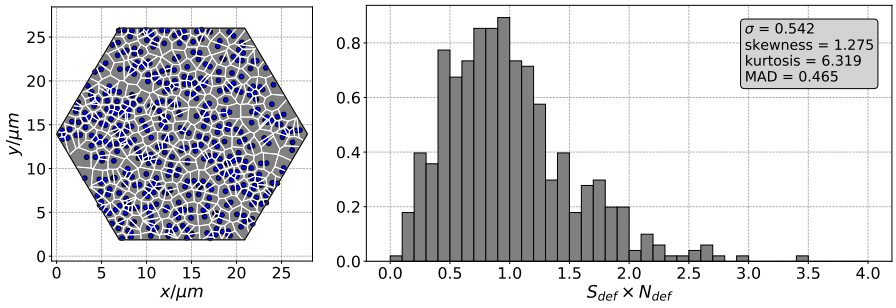
paties dydžio kaip ir panašiuose tyrimuose pateikti įverčiai [71] ir vertės, nustatytos iš eksperimentinių AFM duomenų, kaip aprašyta S.1.2. r_{def} reikšmės rodo pradinio sumažėjimo, po kurio matyti padidėjimas, tendenciją – tai gali būti siejama su sudėtingu defektų susidarymo procesu ir skirtingu metu membranos paviršiuje esančiu visų ir nevisų porų kiekiu [53]. Didžiausia numatoma r_{def} vertė $19,3$ taip pat atitinka apytikslį didžiausią porų dydį 25 nm (sukelta toksino, panašaus į VLY), kaip aprašyta kitame tyrime [22].

Išsamesni eksperimentų aprašymai ir skyriaus išvados pateikiamos disertacijos tekste ir publikacijose [A1, A4].

S.2. Defektų klasterizacijos modeliai

S.2.1. Klasterizacijos įvertinimo metodai

Norint įvertinti ir palyginti skirtingų defektų rinkinių, galinčių turėti skirtingą defektų kiekį, klasterizacijos stiprumą, naudojamos Voronojaus diagramos ir jų sektorių sričių histogramos (S.6 pav., dešinėje). Jos apskaičiuojamos naudojant fiksuotą vienodo dydžio intervalų kiekį iš sektorių plotų, normalizuotų pagal defektų tankį N_{def}.



S.6 pav.: Kompiuteriu sugeneruoto atsitiktinio defektų pasiskirstymo pavyzdys. Kairėje: šešiakampėje modeliavimo srityje išsidėsčiusių defektų Voronojaus diagrama. Dešinėje: Voronojaus diagramos sektorių plotų histograma.

Siekiant supaprastinti histogramų lyginimą ir nustatyti svarbiausias jų statistines savybes, naudojamos keturios apibendrinančios statistikos: standartinis nuokrypis, vidutinis absoliutusis nuokrypis (angl. mean absolute deviation - MAD), asimetrijos koeficientas (angl. skewness) ir ekscesas (angl. kurtosis). Taip pat dviem histogramoms lyginti naudojama EMD metrika (angl. earth mover's distance) [16], kuri apibrėžia minimalias sąnaudas, reikalingas

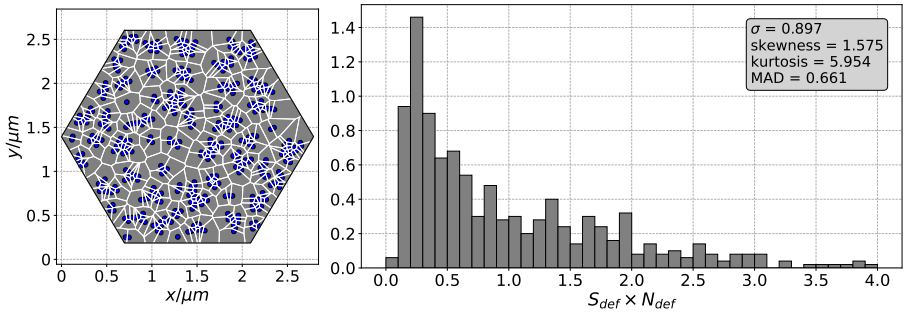
vienam duomenų skirstiniui transformuoti į kitą.

S.2.2. Klasterizacijos modeliai

Traukos modelis

Pirmasis klasterizuotų defektų rinkinių generavimo metodas pagrįstas prielaida, kad defektai natūraliai traukia vienas kitą ir todėl linkę telktis į grupes. Tokio tipo objektų sąveika yra fundamentali ir paplitusi gamtoje (t. y. gravitacinės ir elektromagnetinės jėgos), taip pat taikoma biologinių membranų modeliuose [43]. Šiame modelyje trauka veikia, jei atstumas tarp dviejų defektų yra mažesnis už iš anksto nustatytą slenkstį d_T, kurį galima išreikšti vienu iš dviejų būdų:

- Defekto spindulių skaičius (pritraukiančio defekto).
- Fiksuotas atstumas nanometrais.



S.7 pav.: Sintetinis defektų rinkinys, sugeneruotas pagal traukos modelį, kur $d_T = 15$ (išreikšta defektų spinduliais), $N = 500$, $N_{def} = 100$ ir $r_{def} = 13$.

Darant prielaidą, kad d_T išreiškiamas defekto spinduliais, klasterizacijos modelis turi tris parametrus, turinčius tiesioginę įtaką klasterizacijos efektui (neįskaitant defektų skaičiaus): defektų tankis N_{def}, defektų spindulys r_{def} ir traukos slenkstis d_T. S.7 pav. parodytas vienas sugeneruoto defektų rinkinio pavyzdys, gautas pagal aprašytą modelį.

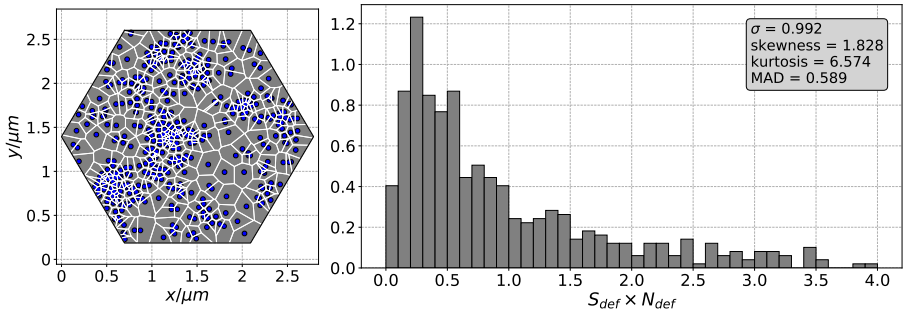
LCN modelis

Šis modelis paremtas idėja, kad membranos defektų sankaupos linkusios sudaryti sudėtingas įvairaus dydžio ir formos struktūras, vizualiai primenančias

debesis. Ši koncepcija aktuali kompiuterinėje grafikoje, kur debesų ar dūmų tekstūroms procedūriškai generuoti naudojami įvairūs algoritmai. Modeliui įgyvendinti pasirinktas *lattice convolutional noise* (LCN) algoritmas [20], išplėstas įvedant du papildomus parametrus, kuriais koreguojami klasterizacijos efektai:

- Vidutinis santykinis klasterio dydis: S
- Minimali defekto atsiradimo tikimybė: P

Defektai atsitiktinai parenkami pagal LCN algoritmu sugeneruotą rastrinį vaizdą, laikomą tikimybių lauku, kuriame vaizdo taškų reikšmės patenka į intervalą $[P, 1]$. Parametras S yra teigiamas realusis skaičius, kuris koreguoja LCN sugeneruoto pradinio vaizdo mastelį – mažesnės reikšmės atitinka didesnį mažų defektų grupių skaičių. Šis algoritmas sugeneruoja klasterizuotus defektų rinkinius (S.8 pav.), kurie vizualiai skiriasi nuo tų, kurie gaunami taikant traukos modelį (S.7 pav.). Klasteriai pasižymi skirtingu dydžiu ir įvairiais nelygumais, kuriuos atspindi ir Voronojaus sektorių plotų statistinės savybės, kur didelį mažų sektorių kiekį atsveria gana nedaug didelių sektorių.



S.8 pav.: Sintetinis defektų rinkinys, sugeneruotas pagal LCN modelį, kur $S = 1$ ir $P = 0,1$.

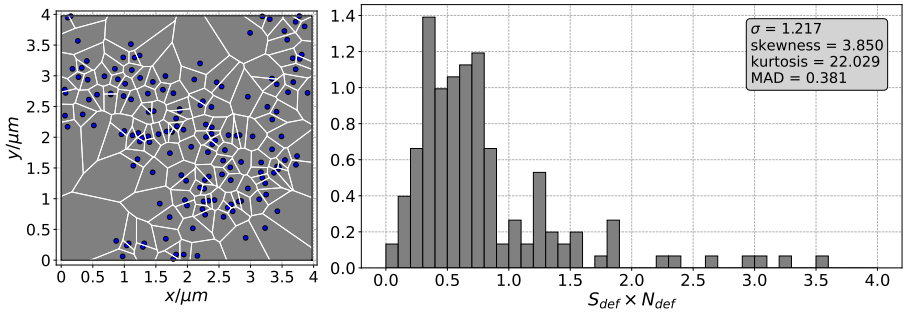
Taškinio proceso modelis

Dar vienas būdas modeliuoti klasterinį defektų išdėstymą membranos paviršiuje yra pagrįstas erdvinių taškų procesų teorija. Konkretus pasirinktas modelis yra Thomas klasterių taškinis procesas (toliau vadinamas taškiniu procesu), kuris generuoja atsitiktinį pirminių taškų (klasterių centrų) skaičių, o kiekvienam iš jų priskiriamas atsitiktinis palikuonių taškų (klasterio narių), atsitiktinai

išstumtų iš centro vektoriaus, paimto iš izotropinio Gauso skirstinio (su ta pačia skale kiekvienoje ašyje), skaičius. Procesas valdomas trimis parametrais:

- Pirminių taškų vidutinis kiekis κ
- Klasterių mastelis r
- Palikuonių taškų vidutinis kiekis α

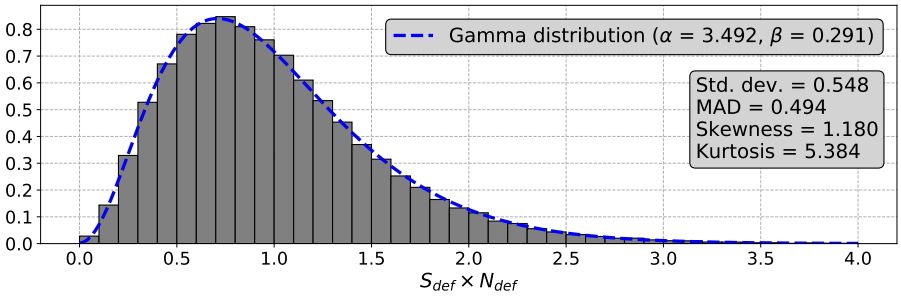
Kitaip nei anksčiau aprašyti traukos ir LCN modeliai, taškinio proceso modelis negeneruoja defektų rinkinių su tiksliu defektų skaičiumi N, nors šiam kiekiui įtakos turi parametrai κ ir α. S.9 pav. parodytas defektų rinkinio, sugeneruoto naudojant šį modelį, pavyzdys. Taškinio proceso modelio pranašumas, palyginti su traukos ir LCN modeliais, yra galimybė tiesiogiai iš duomenų nustatyti modelio parametrus. Šiame darbe tam naudojamas mažiausio kontrasto metodas [3].



S.9 pav.: Sintetinis defektų rinkinys, sugeneruotas pagal taškinio proceso modelį, kur $\kappa = 10$, $\sigma = 0.1$ ir $\alpha = 10$.

Klasterizacijos modelių palyginimas

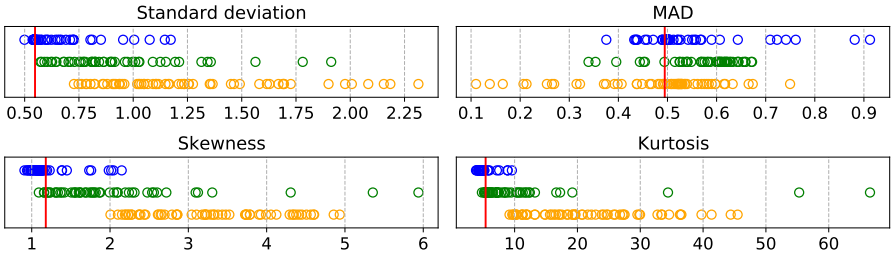
Aprašyti defektų klasterizacijos modeliai pirmiausia įvertinti atsižvelgiant į atsitiktinio defektų pasiskirstymo modelį. Šiuo tikslu nepriklausomai sugeneruota 100 atsitiktinių defektų pasiskirstymo atvejų, iš kurių kiekvienas susideda iš $N = 500$ defektų, kai defektų tankis yra $N_{def} = 10$. S.10 pav. parodyta visų normalizuotų Voronojaus sektorių plotų histograma ir statistinės jų pasiskirstymo savybės.



S.10 pav.: Normalizuotų Voronojaus diagramos sektorių plotų histograma iš 100 nepriklausomai sugeneruotų atsitiktinių defektų rinkinių.

Norint palyginti atsitiktinius ir klasterizuotus defektų pasiskirstymus, sugeneruoti sintetiniai defektų rinkiniai, taikant aprašytus klasterizacijos modelius su skirtingais parametrų deriniais. Buvo nagrinėjama 54, 48 ir 60 parametrų kombinacijų atitinkamai traukos, LCN ir taškinio proceso modeliams. Kiekvienam parametrų deriniui sugeneruota po 100 defektų rinkinių. Traukos ir LCN modelių atveju kiekvieną defektų rinkinį sudarė 500 defektų.

S.11 pav. pateiktos statistinės klasterizuotų defektų rinkinių savybės (apskaičiuotos kiekvienai unikaliai parametrų kombinacijai) ir palygintos su atitinkamomis atsitiktinio defektų pasiskirstymo modelio reikšmėmis. Standartinis nuokrypis geriausiai atskleidžia klasterizacijos efektą, kur beveik visi atvejai (išskyrus kelis traukos modelio atvejus) pasižymi didesnėmis kaip 0,54 (atsitiktinio defektų paskirstymo modelis) reikšmėmis. Toliau šiame darbe įvardinta metrika bus vadinama Voronojaus σ.

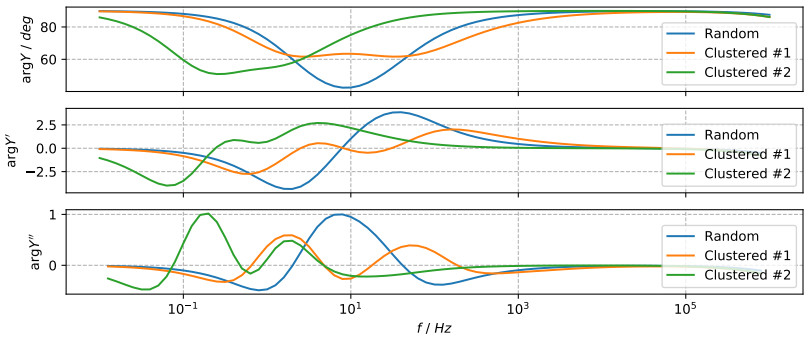


S.11 pav.: Klasterizuotų defektų rinkinių statistinės savybės (spalvoti apskritimai), palygintos su atsitiktinio defektų pasiskirstymo modeliu (vertikalios raudonos linijos). Traukos, LCN ir taškinio proceso modelių atvejai pavaizduoti atitinkamai mėlynais, žaliais ir geltonais apskritimais.

S.2.3. Klasterizacijos efekto įvertinimas pagal EIS spektrus

Siekiant įvertinti defektų klasterizacijos efektą EIS spektrams, sudarytas EIS duomenų rinkinys, modeliuojant įvairius klasterizuotų defektų rinkinių atvejus. Taškinio proceso klasterizacijos modelis buvo naudojamas generuoti membranos modelio egzempliorius, keičiant klasterizacijos modelio parametrus bei defektų tankį ir dydį (viso 216 kombinacijų). Kiekvienam deriniui atskirai sugeneruota 10 modelių egzempliorių, todėl iš viso buvo modeliuojama 2160 unikalių atvejų.

Pirminė sumodeliuotų klasterizuotų defektų rinkinių EIS spektrų peržiūra atskleidė keletą svarbių kokybinių skirtumų, palyginti su atsitiktinių defektų pasiskirstymo EIS duomenimis, išnagrinėtais S.1 skyriuje. Visais atsitiktinio defektų išsidėstymo modelio atvejais EIS spektrai turėjo vieną minimumo tašką ir nulį arba vieną maksimumo tašką. Šių EIS spektrų pirmoms išvestinėms taip pat būdingas vienas minimumo ir vienas maksimumo taškas, o antrosios išvestinės turėjo nuo vieno iki dviejų minimumų ir nuo vieno iki dviejų maksimumų. Tačiau šios savybės negaliojo reikšmingai daliai klasterizuotų atvejų. Maždaug 1 % iš jų turėjo aiškiai išskiriamus dvigubus minimumus, o papildomi 30 % turėjo kitų neįprastų spektrinių požymių, kuriuos atspindėjo didesnis ekstremumų skaičius jų pirmoje ir antroje išvestinėse, palyginti su atsitiktiniais atvejais. S.12 pav. iliustruoja šį reiškinį, kai du klasterizuoti atvejai pasižymi aprašytais skirtumais, palyginti su atsitiktinio atvejo pavyzdžiu.



S.12 pav.: Sumodeliuotų EIS spektrų pavyzdžiai, gauti naudojant atsitiktinius ir klasterizuotus defektų rinkinius. Viršutinė diagrama rodo pradinius EIS spektrus, vidurinė ir apatinė diagramos rodo atitinkamai pirmą ir antrą spektro išvestinę.

Priklausomai nuo konkrečios klasterizacijos modelio parametrų kombina-

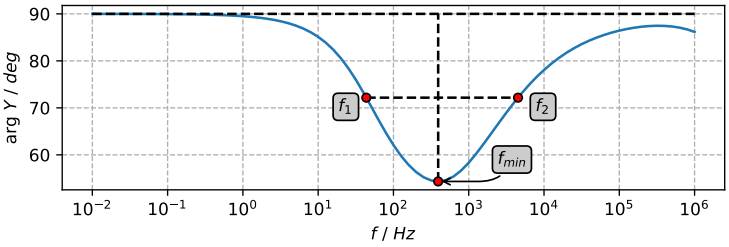
cijos, spektrų su neįprastais kokybiniais požymiais dalis svyruoja nuo 3 % iki 70 %. Didžiausią įtaką turi mažos r ir didelės α reikšmės, lemiančios nedidelį kiekį klasterių su dideliu glaudžiai išsibarsčiusių defektų skaičiumi. Atitinkamai pastebimas šio efekto ryšys su Voronojaus σ reikšmėmis, kurių augimas lemia didesnę tikimybę atsirasti anomalijoms spektruose.

Kadangi defektų klasterizacija lemia pokyčius EIS spektruose, šis efektas, išreikštas Voronojaus σ reikšmėmis, galėtų būti kiekybiškai įvertinamas pagal tam tikrus spektrinius požymius. Šie požymiai, užuot siedamiesi su minimumo taško padėtimi, turėtų apibrėžti spektro kreivės formą, apibūdinamą f_{min} ir $\arg Y_{min}$ reikšmėmis. Viena iš tokių charakteristikų, apibūdinančių kreivės formą ir dažnai naudojamų spektroskopijos duomenų analizėje, yra pusaukščio plotis (FWHM) [94, 73]:

$$FWHM = \log_{10} f_2 - \log_{10} f_1. \tag{S.13}$$

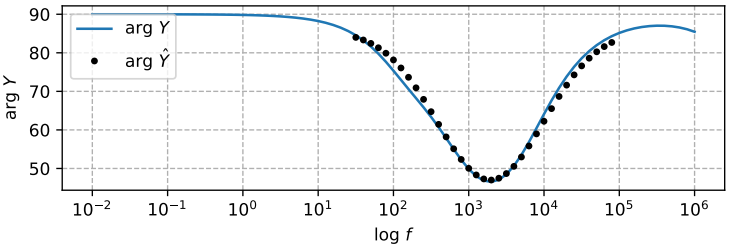
Čia dažniai f_1 ir f_2 reiškia du taškus (pavaizduoti S.13 pav.), kur:

$$\arg Y = 90 - \frac{90 - \arg Y_{min}}{2}. \tag{S.14}$$



S.13 pav.: EIS spektro taškai, naudojami apskaičiuoti pusaukščio plotį.

Darant prielaidą, kad tam tikra EIS spektro sritis gali būti aproksimuojama pasirinkta analitine funkcija, spektro neatitikimas jai (pvz., dėl dvigubų minimumų) galėtų reikšti klasterizacijos efektą. Tokiu atveju atliekamas Gauso kreivės pritaikymas spektrinėje srityje aplink minimumo tašką (kur $\arg Y$ vertės neviršija 87 deg), kaip parodyta S.14 pav.



S.14 pav.: Gauso kreivės, pritaikytos EIS spektrui minimumo taško aplinkoje, pavyzdys.

k faktinių ($\arg Y$) ir pritaikytų ($\arg \hat{Y}$) spektro reikšmių (indeksai nuo t_1 iki t_2) neatitikimas tuomet apibendrinamas keliais dydžiais:

- Vidutinis absoliutus skirtumas:

$$p_{mean} = \frac{1}{k} \sum_{i=t_1}^{t_2} |\arg Y_i - \arg \hat{Y}_i|. \tag{S.15}$$

- Absoliučių skirtumų standartinis nuokrypis:

$$p_{std} = \sqrt{\frac{1}{k} \sum_{i=t_1}^{t_2} (\arg \hat{Y}_i - p_{mean})^2}. \tag{S.16}$$

- Didžiausias absoliutus skirtumas:

$$p_{max} = max(|\arg Y_i - \arg \hat{Y}_i|), \quad i = t_1, ..., t_2. \tag{S.17}$$

Taip pat naudojami spektriniai požymiai, apibūdinantys ekstremumų kiekius spektre ar jo išvestinėse. Visas regresijos užduočiai pasirinktas požymių rinkinys apėmė FWHM, p_{mean}, p_{std}, p_{max} ir 6 ekstremumų kiekių požymius (viso 10 reikšmių). Tiesinės regresijos modelis su L1 reguliarizacija (Lasso) buvo naudojamas 10 išvardytų požymių ir Voronojaus σ ryšiui išreikšti bei informatyviausiems požymiams išskirti. Norint ištirti, kaip modelis apibūdina spektrinius pokyčius, susijusius tik su defektų klasterizacija, o ne su kintančiomis N_{def} ir r_{def} reikšmėmis, atlikta 8 dalių kryžminė patikra, kur kiekvienoje iteracijoje modelis buvo validuojamas naudojant visus konkretaus N_{def} ir r_{def} derinio EIS spektrus. Reguliarizacijos parametras λ buvo keičiamas nuo 10^{-3} iki 10^{-1}.

Priimtinas validavimo tikslumas buvo stebimas ties $\lambda = 10^{-1,5}$ su 3 požymiais: FWHM, p_{max} ir m_2. Šio konkretaus modelio MAE ir MAPE paklaidos

buvo atitinkamai lygios 0,22 ir 24 %. Gana žemas bendras tikslumas rodo, kad parinkti spektriniai požymiai ir tiesinis modelis negali visiškai atspindėti kintančio Voronojaus σ poveikio spektrui, tuo pačiu atmetant panašius pokyčius, kuriuos lemia N_{def} ir r_{def} parametrai.

S.2.4. Metodikos patvirtinimas naudojant AJM duomenis

Norint patikrinti siūlomą metodiką su realiais duomenimis, eksperimentiškai gauti trys tBLM membranų, paveiktų poras formuojančiu toksinu vaginolizinu (VLY), AJM vaizdai. Kiekvienas vaizdas sujungtas iš 9 AJM vaizdo fragmentų, fiksuotų 512×512 skiriamąja geba ir apimančių $2\,\mu m \times 2\,\mu m$ membranos paviršiaus plotą, todėl galutinis vaizdas sudaro $6\,\mu m \times 6\,\mu m$ esant 1536×1536 skiriamajai gebai. Kiekviename vaizde esančių defektų koordinates dalykinės srities ekspertas anotavo rankiniu būdu (visame vaizde arba pagal šešiakampę modeliavimo sritį). S.4 lentelėje pateiktos pagrindinės šių defektų rinkinių savybės.

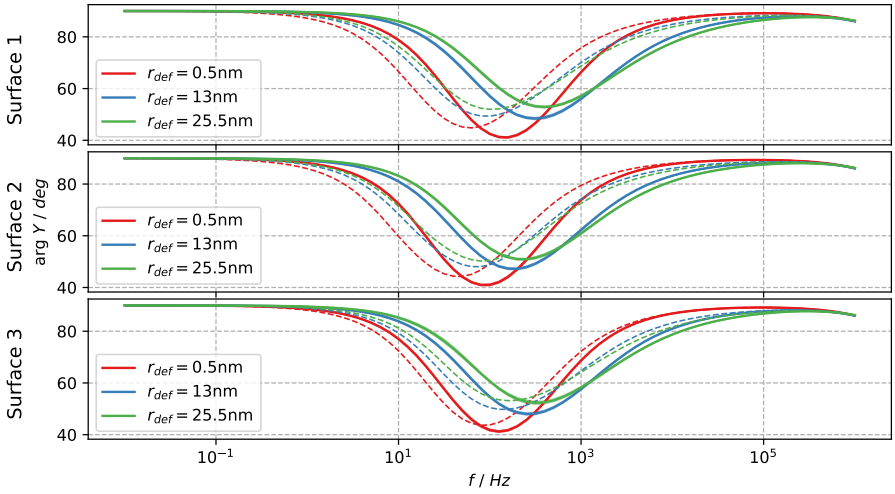
S.4 lentelė: AJM vaizduose sužymėtų defektų rinkinių savybės.

Modeliavimo sritis	**AJM vaizdo Nr.**	N	N_{def}	**Voronojaus σ**
Šešiakampė	1	234	10,01	1,22
	2	148	6,33	1,12
	3	235	10,05	0,88
Keturkampė	1	374	10,39	1,17
	2	235	6,53	1,06
	3	328	9,11	0,81

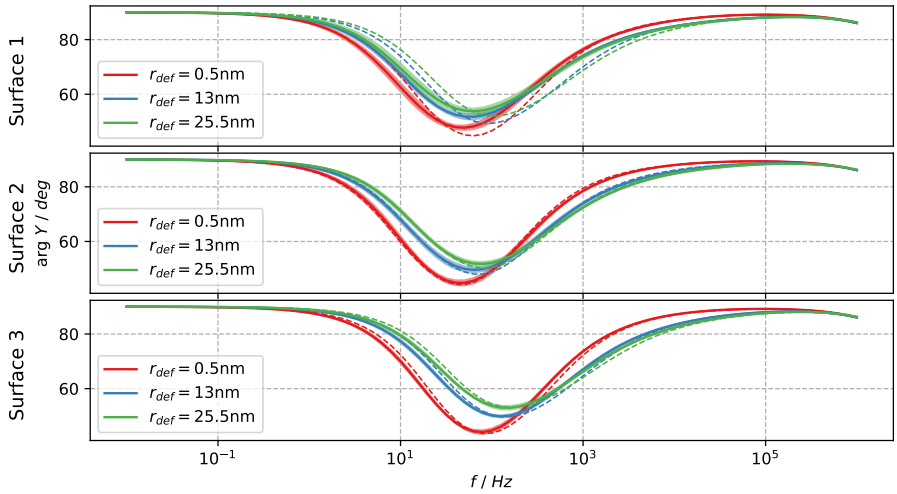
Siekiant įvertinti, ar aprašyti defektų klasterizacijos modeliai yra pritaikomi realiems defektų rinkiniams (išmatuotiems AJM), sugeneruotų klasterizuotų defektų rinkinių EIS spektrai buvo lyginami su atitinkamais AJM defektų rinkinių spektrais. Skirtingų klasterizacijos algoritmų parametrai modeliuoti parinkti naudojant S.2.2 skyrelyje nagrinėtus atvejus, lyginant sintetinių klasterizuotų defektų rinkinių histogramas su atitinkamomis realių AJM duomenų rinkinių histogramomis (naudojant EMD metriką). Taškinio proceso modelio parametrams parinkti papildomai buvo taikomas mažiausio kontrasto metodas, pagal kurį parametrai įvertinti tiesiogiai iš AJM duomenų. Visi skaičiavimai atlikti naudojant tris defekto dydžio r_{def} parinktis: 0.5 nm, 13 nm ir 25.5 nm.

Kiekvienai klasterizacijos algoritmo, AJM defektų rinkinio ir r_{def} kombinacijai buvo nepriklausomai generuojama 10 klasterizuotų defektų rinkinių, su kuriais atliktas EIS modeliavimas.

Palyginimui analogiški skaičiavimai atlikti naudojant ir atsitiktinį defektų išsidėstymo modelį (S.15 pav.). EIS kreivėse matomi sintetinių defektų rinkinių neatitikimai AJM atvejams, taip pat pastebėti ankstesnio eksperimento metu (S.1.2 punktas). Iš nagrinėtų klasterizacijos modelių geriausiai AJM duomenis atitinka taškinio proceso modelis ir jo parametrai, parinkti naudojant mažiausio kontrasto metodą (S.16 pav.). Šio modelio atvejo skirtumai tarp minimumo taškų $\log f$ ir $\arg Y$ ašyse svyruoja atitinkamai nuo -0,28 iki -0,03 ir nuo 0,74 iki 3,32.



S.15 pav.: EIS spektrų, modeliuotų naudojant AJM vaizdų defektų rinkinius (punktyrinės kreivės) ir atsitiktinį defektų išsidėstymo modelį (vientisos juostos), lyginimas.



S.16 pav.: EIS spektrų, modeliuotų naudojant AJM vaizdų defektų rinkinius (punktyrinės kreivės) ir taškinio proceso klasterizacijos modelį (vientisos juostos), lyginimas.

Išsamūs klasterizacijos modelių ir eksperimentų aprašymai bei skyriaus išvados yra pateiktos disertacijos tekste ir publikacijose [A2, A3].

S.3. Automatinis defektų aptikimas AJM vaizduose

S.3.1. Defektų aptikimo eksperimentai

Anksčiau aprašytas AJM duomenų rinkinys (S.4 lentelė, S.2 skyrius) buvo pakartotinai naudojamas defektų aptikimo eksperimentams, aprašytiems toliau šiame skyriuje. Kiekvieno tBLM membranos mėginio vaizdo fragmentų rinkiniai suskirstyti į mokymo ir testavimo poaibius, priskiriant 5 fragmentus mokymui ir 4 testavimui. Pastarieji fragmentai parinkti taip, kad atitiktų vientisą $4\,\mu\text{m} \times 4\,\mu\text{m}$ paviršiaus plotą apatiniame dešiniajame visiškai susiūto vaizdo kampe. S.5 lentelėje rodomas bendras anotuotų defektų skaičius (N) ir vidutinis defektų tankis (N_{def}) kiekviename AJM vaizde ir mokymo bei testavimo fragmentų poaibyje.

S.5 lentelė: AJM vaizdų rinkiniai, naudojami defektų aptikimo modeliams mokyti ir testuoti.

AJM vaizdas	Poaibis	Fragmentų sk.	N	N_{def}	Voronojaus σ
1	Mokymo	5	202	10,10	1,18
2	Mokymo	5	138	6,90	1,12
3	Mokymo	5	170	8,50	0,77
1	Testavimo	4	172	10,75	1,20
2	Testavimo	4	97	6,06	1,02
3	Testavimo	4	158	9,88	0,91

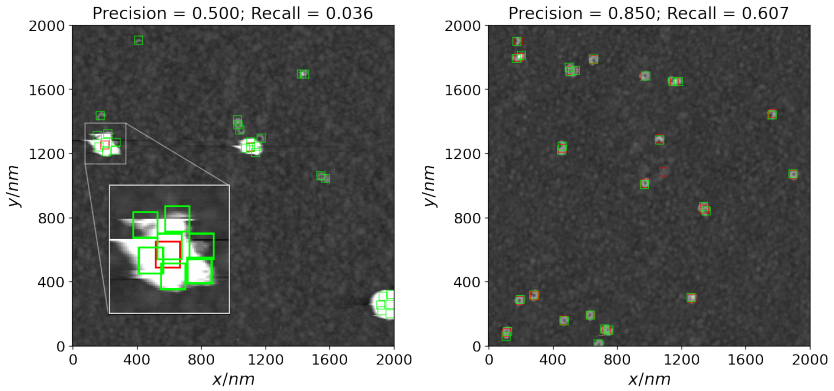
TopoStats

Pirmasis defektų aptikimo eksperimentas atliktas naudojant atvirojo kodo programinės įrangos įrankį TopoStats, sukurtą biomolekulėms aptikti AJM vaizduose [96]. Defektų aptikimo tikslumas buvo vertinamas tikslumo (angl. *precision*), jautrumo (angl. *recall*) bei F1 metrikomis (S.6 lentelė).

S.6 lentelė: Defektų aptikimo tikslumas testiniuose AJM vaizdų fragmentuose, naudojant TopoStats įrankį.

AJM vaizdas	N_{true}	N_{pred}	Tikslumas	Jautrumas	F1
1	172	53	0,754	0,233	0,355
2	97	31	0,742	0,237	0,359
3	158	79	0,886	0,443	0,591

Nepaisant gana aukšto tikslumo, visais atvejais jautrumas yra žemas, o dėl to gaunama daug klaidingų neigiamų rezultatų. Vizualiai patikrinus aptikimo rezultatus paaiškėjo, kad įrankis prastai aptiko defektų grupes, kuriose dauguma tokių atvejų buvo traktuojami kaip vienas defektas (pavyzdys parodytas S.17 pav.).



(a) Iliustratyvus AJM vaizdo fragmentas su defektų klasteriais.

(b) Iliustratyvus AJM vaizdo fragmentas be defektų klasterių.

S.17 pav.: Tikrųjų defektų padėčių (žali stačiakampiai) ir aptiktų naudojant TopoStats įrankį pavyzdžiai (raudoni stačiakampiai). Defektų klasterio pavyzdys ir atitinkamos tikrosios ir numatomos defektų vietos yra priartintos kairiajame vaizde.

Ploto matavimo metodas

Siekiant išspręsti ankstesniame eksperimente pastebėtą defektų klasterių atskyrimo problemą, išbandytas paprastas metodas, pagrįstas tipinėmis skaitmeninių vaizdų apdorojimo operacijomis. Algoritmas (toliau vadinamas ploto matavimo metodu) susideda iš šių žingsnių:

1. Objektų ir fono atskyrimas. Pradinis pilkos skalės AJM vaizdas konvertuojamas į dvejetainį vaizdą naudojant fiksuotą slenkstinę vertę T.

2. Morfologinis apdorojimas. Objektus atitinkantys regionai apdorojami taikant dvejetainę uždarymo operaciją (angl. *binary closing*) ir pašalinant mažus objektus (mažesnius kaip 5 pikseliai) [55].

3. Defektų skaičiaus nustatymas. Defektų, sudarančių kiekvieną regioną, skaičius nustatomas padalijus regiono plotą (pikseliais) iš nustatytos reikšmės S (vidutinis vieno defekto užimamas plotas) ir suapvalinant gautą santykį.

4. Defekto koordinačių priskyrimas. Tikslios defektų centro koordinatės nustatomos atliekant kiekvienos srities pikselių koordinačių K-means klasterizavimą (naudojant ankstesniame žingsnyje gautą klasterių skaičių) ir naudojant kiekvieno pikselių klasterio centrus.

Norint atlikti eksperimentą su testiniais AJM vaizdų fragmentais, pasirinktos parametrų reikšmės $T = 100$ ir $S = 130$, paleidus algoritmą apmokymo rinkinio AJM vaizdo fragmentams su kintančiomis S ir T reikšmėmis ir pasirinkus tuos, kurie duoda didžiausią vidutinę F1 vertę. S.7 lentelėje rodomi bandymo rezultatai.

S.7 lentelė: Defektų aptikimo tikslumas testiniuose AJM vaizdų fragmentuose, naudojant ploto matavimo metodą.

AJM vaizdas	N_{true}	N_{pred}	Tikslumas	Jautrumas	F1
1	172	114	0,860	0,570	0,685
2	97	114	0,553	0,649	0,597
3	158	194	0,613	0,753	0,676

Hough transformacija

Alternatyvus defektų aptikimo algoritmas realizuotas naudojant apskritiminę Hough transformaciją, remiantis prielaida, kad membranos defektai matomi kaip apytiksliai vienodo spindulio žiedinės struktūros. Algoritmas susideda iš šių žingsnių:

1. Objektų ir fono atskyrimas. Pradinis pilkos skalės AJM vaizdas konvertuojamas į dvejetainį vaizdą naudojant minimumo slenksčio metodą [1].

2. Morfologinis apdorojimas. Defektus atitinkančios vaizdo sritys paverčiamos vieno pikselio pločio kontūrais, taikant morfologinio retinimo operaciją (angl. *thinning*) [6].

3. Apskritiminė Hough transformacija (CHT). Galutinis defekto aptikimas atliekamas taikant CHT apdorotam dvejetainiam vaizdui, naudojant pasirinktą Hough slenkstį ir kintantį apskritimo spindulį.

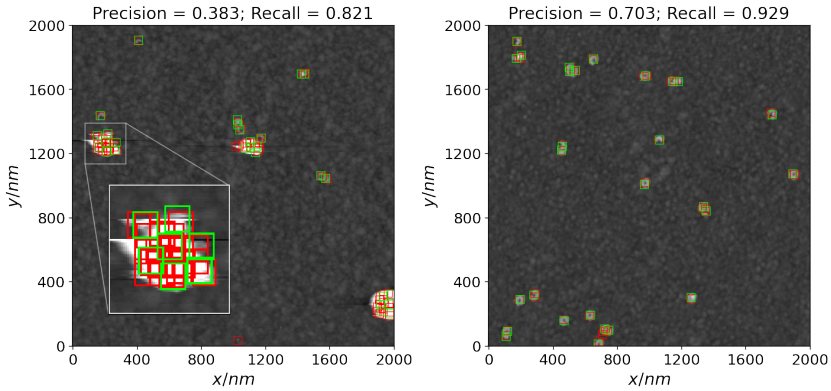
Aprašytas algoritmas pritaikytas testiniams vaizdų fragmentams, naudojant Hough slenksčio reikšmę, lygią 0,28, o apskritimo spindulys svyravo nuo 3 iki 7 taškų. Bandymo rezultatai (S.8 lentelė) palyginami su anksčiau aprašytu ploto matavimo metodu, nors vidutinis tikslumas yra mažesnis dėl bendro didesnio aptikimo skaičiaus, atitinkančio daugiau klaidingai teigiamų rezultatų.

AJM vaizdas	N_{true}	N_{pred}	Tikslumas	Jautrumas	F1
1	172	201	0.577	0.674	0.622
2	97	156	0.436	0.701	0.538
3	158	140	0.614	0.544	0.577

S.8 lentelė: Defektų aptikimo tikslumas testiniuose AJM vaizdų fragmentuose, naudojant apskritiminę Hough transformaciją.

Konvoliucinis neuroninis tinklas

Defektams aptikti su konvoliuciniu neuroniniu tinklu panaudotas SSD FPN architektūros objektų aptikimo modelis [76], pritaikytas defektams aptikti AJM vaizduose pagal apmokymo poaibio fragmentus. Nustatyti tikslumo ir jautrumo balai (S.9 lentelė) yra panašūs į ankstesnius eksperimentus, atliktus naudojant paprastesnius algoritmus, nors vidutinis F1 balas tarp visų bandomųjų vaizdų yra šiek tiek didesnis. Defektų klasteriai (S.18 pav., kairėje) vis dar buvo sunkiai išskiriami dėl prastai matomų paviršiaus savybių grupių viduje. Tačiau modelis gana gerai veikė su tam tikrais vaizdų fragmentais, kuriuose nebuvo defektų klasterių (S.18 pav., dešinėje). Tai taip pat iliustruoja faktas, kad testinis AJM 3-ojo paviršiaus vaizdas, rodantis mažiausią defektų klasterizaciją pagal Voronojaus σ (S.5 lentelė), taip pat turi aukščiausią bendrą F1 balą iš visų patikrintų algoritmų.



(a) Iliustratyvus AJM vaizdo fragmentas su defektų klasteriais.

(b) Iliustratyvus AJM vaizdo fragmentas be defektų klasterių.

S.18 pav.: Tikrųjų defektų padėčių (žali stačiakampiai) ir aptiktų naudojant konvoliucinį neuroninį tinklą pavyzdžiai (raudoni stačiakampiai). Defektų klasterio pavyzdys ir atitinkamos tikrosios ir numatomos defektų vietos priartintos kairiajame vaizde.

S.9 lentelė: Defektų aptikimo tikslumas testiniuose AJM vaizdų fragmentuose, naudojant konvoliucinį neuroninį tinklą.

AJM vaizdas	N_{true}	N_{pred}	Tikslumas	Jautrumas	F1
1	172	129	0,775	0,581	0,664
2	97	119	0,555	0,680	0,611
3	158	152	0,757	0,728	0,742

S.3.2. Defektų aptikimo tikslumo įtaka EIS spektrams

Norint įvertinti ryšį tarp defektų aptikimo tikslumo ir atitinkamų pokyčių EIS spektruose, reikalingas didelis defektų aptikimo rezultatų rinkinių skaičius. Tokie aptikimo rezultatai turėtų pasižymėti skirtingomis tikslumo ir jautrumo reikšmėmis, tačiau tokius specifinius aptikimo rezultatus gali būti sunku gauti taikant objektų aptikimo modelius, sudarytus naudojant tikrus AJM vaizdus ir anotuotas tikras defektų vietas. Kita problema – ribotas turimų AJM vaizdo duomenų kiekis. Dėl šių priežasčių nagrinėjamas alternatyvus būdas, kuriuo generuojami sintetiniai defektų rinkiniai, imituojantys defektų aptikimo rezultatus skirtinguose tikslumo lygiuose. Kiekvienas sintetinis atvejis generuojamas iš pradinio tikrųjų defektų koordinačių rinkinio, pritaikant tam tik-

ras modifikacijas (defektų pridėjimą, pašalinimą, koordinačių perkėlimą), kad būtų gautas naujas defektų rinkinys analogiškiems rezultatams, gaunamiems taikant kurį nors defektų aptikimo algoritmą.

Siekiant kiekybiškai įvertinti neatitikimą tarp EIS spektrų, sumodeliuotų pateiktai porai tikrųjų (sužymėtų eksperto) ir automatiškai aptiktų (objekto aptikimo algoritmu) defektų rinkinių, naudojama kreivių minimumo taškų padėtis $\log f$ ir $\arg Y$ ašyse (darant prielaidą, kad defektų klasterizacijos poveikis EIS spektrų formoms yra nereikšmingas):

$$\Delta \mathrm{f}_{\log} = \log_{10}(f_{min}^{(true)}) - \log_{10}(f_{min}^{(pred)}). \tag{S.18}$$

$$\Delta \arg Y = \arg Y_{min}^{(true)} - \arg Y_{min}^{(pred)}. \tag{S.19}$$

Taip pat naudojama papildoma Q_N metrika, kuri parodo defektų tankio santykį (defektų skaičių kvadratiniame mikrometre) iš numatytų ir tikrų defektų rinkinių:

$$\mathrm{Q_N} = N_{def}^{(pred)} / N_{def}^{(true)}. \tag{S.20}$$

Algoritmas sintetinių defektų rinkiniams (analogiškiems defektų aptikimo rezultatams) generuoti sudarytas remiantis prielaida, kad defektų klasteriai yra išskiriami gana gerai, o klasterizacijos lygis pradinio (tikrojo) defektų rinkinio atžvilgiu yra išlaikomas. Jį sudaro šie žingsniai:

1. Tikrosioms defektų koordinatėms atliekamas dvimatis branduolio tankio įvertinimas (angl. *kernel density estimation* – KDE) [34].

2. Kiekvienam generuojamam defektų rinkiniui:

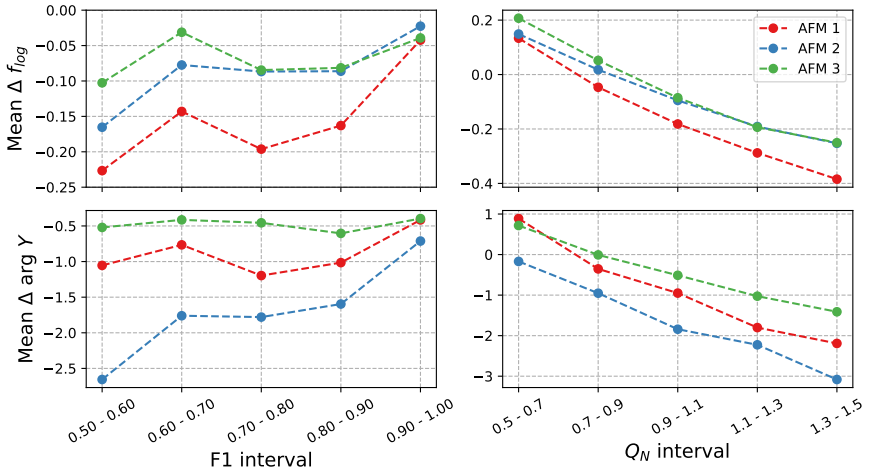
   (a) Tikrosios koordinatės ($x^{(true)}$ ir $y^{(true)}$) kiekvienam defektui yra modifikuojamos pridedant atsitiktinę reikšmę, parinktą iš normaliojo skirstinio:

   $$x^{(pred)} = x^{(true)} + \delta; \qquad y^{(pred)} = y^{(true)} + \delta; \qquad \delta \sim \mathcal{N}(\mu, \sigma^2)$$

   (b) Pagal KDE modelį parenkamas n_{remove} defektų koordinačių porų skaičius. Tikrieji defektai, esantys arčiausiai atrinktų koordinačių, atrenkami ir pašalinami iš pradinio defektų rinkinio. Tai įveda klaidingai neigiamus atvejus į generuojamą defektų rinkinį ir atitinkamai sumažina aptikimo jautrumą.

(c) Pagal KDE modelį parenkamas n_{add} naujų koordinačių porų skaičius, o defektai su šiomis koordinatėmis įtraukiami į sugeneruotą defektų rinkinį. Tai atitinka klaidingai teigiamus atvejus ir sumažina aptikimo tikslumą.

Aprašytas algoritmas buvo naudojamas sintetiniams atvejams generuoti kiekvienam iš trijų AJM bandomųjų vaizdų atskirai. KDE modeliai buvo pritaikyti naudojant Gauso branduolį, o pralaidumo (angl. *bandwidth*) parametras nustatytas į 400. Parametrų n_{remove} ir n_{add} reikšmės buvo keičiamos nuo 0 iki $N/2$, su žingsniu, atitinkančiu 3 % bendro defektų skaičiaus N.



S.19 pav.: Vidutinės Δf_{log} ir $\Delta \arg Y$ vertės, apskaičiuotos sugeneruotų defektų rinkinių poaibiams, atitinkantiems skirtingus F1 ir Q_N intervalus.

Apibendrinti rezultatai, pateikti S.19 pav., rodo vidutinius minimumo taško koordinačių nuokrypius skirtinguose F1 ir Q_N intervaluose. Vidutinės absoliučios Δf_{log} ir $\Delta \arg Y$ vertės reikšmingai skiriasi tarp trijų AJM vaizdų, kur mažiausiais nuokrypiais pasižymi AJM defektų rinkinys su mažiausia Voronojaus σ reikšme.

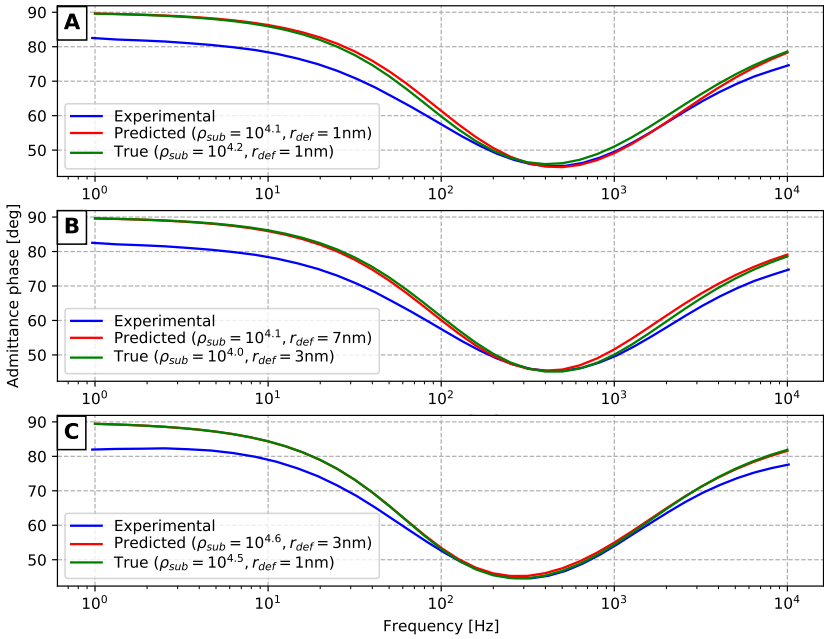
S.3.3. Modeliuotų ir eksperimentinių EIS spektrų palyginimas

Siekiant įvertinti, kaip automatizuoto defektų aptikimo netikslumas realiuose AJM vaizduose paveiktų parametrų ρ_{sub} ir r_{def} prognozavimą iš atitinkamų EIS spektrų, atlikta modeliavimo užduočių serija su kiekviena tikrų ir prognozuotų defektų rinkinių (gautų naudojant konvoliucinį neuroninį tinklą) pora,

naudojant skirtingas r_{def} ir ρ_{sub} reikšmes. Sumodeliuotos tikrų ir prognozuotų defektų rinkinių kreivės sulygintos su eksperimentiniais EIS duomenimis, gautais atlikus matavimus kiekvienam iš trijų tBLM membranų mėginių. Visų trijų tipų EIS kreivės sugretintos pagal minimumo taško koordinates ($\log f_{min}$ ir $\arg Y_{min}$).

S.20 pav. parodytos modeliuojamos ir eksperimentinės kiekvieno paviršiaus kreivės, taip pat konkrečios atitinkamų modeliuotų atvejų r_{def} ir ρ_{sub} reikšmės. Geriausiai suderintos modeliuotos kreivės iš tikrųjų ir prognozuotų defektų rinkinių visais atvejais rodo nedidelį 0,1 skirtumą pagal $\log \rho_{sub}$, o r_{def} skirtumai svyruoja nuo 0 nm iki 4 nm.

Išsamūs defektų aptikimo metodų ir eksperimentų aprašymai bei skyriaus išvados pateikiamos disertacijos tekste ir publikacijoje [A5].



S.20 pav.: Eksperimentinių EIS matavimų duomenys (mėlynos kreivės), palyginti su modeliuotais atvejais (žalios ir raudonos kreivės, atitinkančios atitinkamai rankiniu būdu anotuotas defektų koordinates ir CNN modelio prognozes). A, B ir C dalys atitinka 1, 2 ir 3 AJM paviršius.

Bendrosios išvados

- Aprašytas trimatis membranos modelis (įgyvendintas naudojant baigtinių elementų metodą) leidžia imituoti realistiškus EIS atsakus su įvairaus

pobūdžio membranos defektų išsidėstymais. Mašininio mokymosi metodais pagrįstos EIS duomenų analizės metodika pademonstruota prognozuojant kiekybinius membranos parametrus, tiesiogiai nepasiekiamus iš EIS spektro – defektų tankį, dydį ir pomembraninio sluoksnio savitąją varžą.

- Trys pateikti defektų klasterizacijos modeliai gali būti naudojami generuojant tikroviškus defektų rinkinius, pasižyminčius skirtingu klasterizacijos laipsniu, ir parametrizuojant realius defektų rinkinius, gaunamus iš AJM duomenų. Voronojaus diagramos sektorių plotų standartinis nuokrypis, apskaičiuotas klasterizuotiems defektų rinkiniams, pasiūlytas kaip paprasta metrika, tinkanti kiekybiškai įvertinti klasterizacijos efektą ir atskirti klasterizuotus ir atsitiktinai paskirstytus defektų rinkinius. Defektų klasterizacijos efektą atspindi EIS spektrų pokyčiai, kurie dar negali būti visiškai atsieti (ištirtais metodais) nuo kitų membranos parametrų (tokių kaip defekto tankis ir dydis) įtakos.

- Automatizuoto defektų aptikimo algoritmų bandymai su realių membranų AJM vaizdais parodė F1 reikšmes nuo 0,538 iki 0,742. Konvoliucinis neuroninis tinklas veikė nežymiai tiksliau už paprastesnius metodus, pagrįstus Hough transformacija, ir paprastomis vaizdų apdorojimo operacijomis. Dėl riboto AJM vaizdo duomenų kiekio įgyvendintas metodas, leidžiantis generuoti sintetinius defektų rinkinius, atitinkančius įvairius defektų aptikimo tikslumo lygius ir įvertinti gautų EIS spektrų paklaidų ribas.

Publications by the author

[A1] Tomas Raila, Tadas Penkauskas, Marija Jankunec, Gintaras Dreižas, Tadas Meškauskas, and Gintaras Valinčius. Electrochemical impedance of randomly distributed defects in tethered phospholipid bilayers: Finite element analysis. *Electrochimica Acta*, 299 (2019), pp. 863–874.

[A2] Tomas Raila, Filipas Ambrulevičius, Tadas Penkauskas, Marija Jankunec, Tadas Meškauskas, David J. Vanderah, and Gintaras Valincius. Clusters of protein pores in phospholipid bilayer membranes can be identified and characterized by electrochemical impedance spectroscopy. *Electrochimica Acta*, 364 (2020), p. 137179.

[A3] Tomas Raila, Marija Jankunec, Tadas Meškauskas, and Gintaras Valinčius. Computational Models of Defect Clustering for Tethered Bilayer Membranes. *Computational Science and Its Applications – ICCSA 2020*. Ed. by Osvaldo Gervasi et al. Cham: Springer International Publishing, 2020, pp. 496–504.

[A4] Tomas Raila, Tadas Meškauskas, Gintaras Valinčius, Marija Jankunec, and Tadas Penkauskas. Computer Modeling of Electrochemical Impedance Spectra for Defected Phospholipid Membranes: Finite Element Analysis. *Numerical Computations: Theory and Algorithms*. Ed. by Yaroslav D. Sergeyev and Dmitri E. Kvasov. Cham: Springer International Publishing, 2020, pp. 462–469.

[A5] Tomas Raila, Tadas Penkauskas, Filipas Ambrulevičius, Marija Jankunec, Tadas Meškauskas, and Gintaras Valinčius. AI-based atomic force microscopy image analysis allows to predict electrochemical impedance spectra of defects in tethered bilayer membranes. *Scientific Reports*, 12.1 (2022), p. 1127.

About the author

Tomas Raila was born in Pakruojis, Lithuania in 1989. In 2007 he graduated from "Atžalynas" high school in Biržai. He obtained his BSc and MSc degrees of informatics in 2012 and 2015 at Vilnius University, Faculty of Mathematics and Informatics. In 2017-2021 he studied in PhD study program of informatics in Vilnius University. Since 2011 Tomas has been working as a software engineer in various private companies, since 2019 he also teaches at Vilnius University, Faculty of Mathematics and Informatics.

Trumpos žinios apie doktorantą

Tomas Raila gimė 1989 m. gegužės 16 d. Pakruojyje. 2007 m. baigė Biržų „Atžalyno" vidurinę mokyklą. 2007-2012 m. studijavo programų sistemas Vilniaus universiteto Matematikos ir informatikos fakultete ir įgijo informatikos bakalauro laipsnį. 2013-2015 m. ten pat studijavo kompiuterinį modeliavimą, studijas baigė su Cum Laude diplomu ir įgijo informatikos magistro laipsnį. 2017-2021 m. studijavo informatikos doktorantūroje Vilniaus universitete. Nuo 2011 m. dirbo įvairiose privačiose įmonėse programuotoju, nuo 2019 m. dėsto Vilniaus universiteto Matematikos ir informatikos fakultete.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my scientific advisor Prof. Dr. Tadas Meškauskas for the continuous guidance, support and motivation. His scientific and personal advice made this thesis possible and I could not have imagined having a better mentor for my doctoral studies.

I am very thankful to the biochemistry researchers at Life Sciences Center: Prof. Dr. Gintaras Valinčius and his team, Dr. Marija Jankunec, Dr. Tadas Penkauskas and Filipas Ambrulevičius. The prolific scientific collaboration which resulted in the research papers was a truly inspiring experience and I am grateful for the opportunity to take part.

My thanks also go to fellow doctoral students and university colleagues for valuable advice, ideas and discussions. I am also grateful to my colleague Dr. Valdas Rapševičius for inspiration and encouragement in pursuing the doctorate degree.

Finally, I am thankful to my family and friends for their support and understanding throughout this challenging undertaking.

NOTES

Tomas Raila

Computer modeling methods for phospholipid membrane damage assessment

Doctoral Dissertation
Natural Sciences
Informatics (N 009)
Thesis Editor: Zuzana Šiušaitė

Kompiuterinio modeliavimo metodai fosfolipidinių membranų pažaidos įvertinimui

Daktaro disertacija
Gamtos mokslai
Informatika (N 009)
Santraukos redaktorė: Jorūnė Rimeisytė-Nekrašienė

Vilnius University Press
9 Saulėtekio Ave., Building III, LT-10222 Vilnius
Email: info@leidykla.vu.lt, www.leidykla.vu.lt
bookshop.vu.lt, journals.vu.lt
Print run of 20 copies