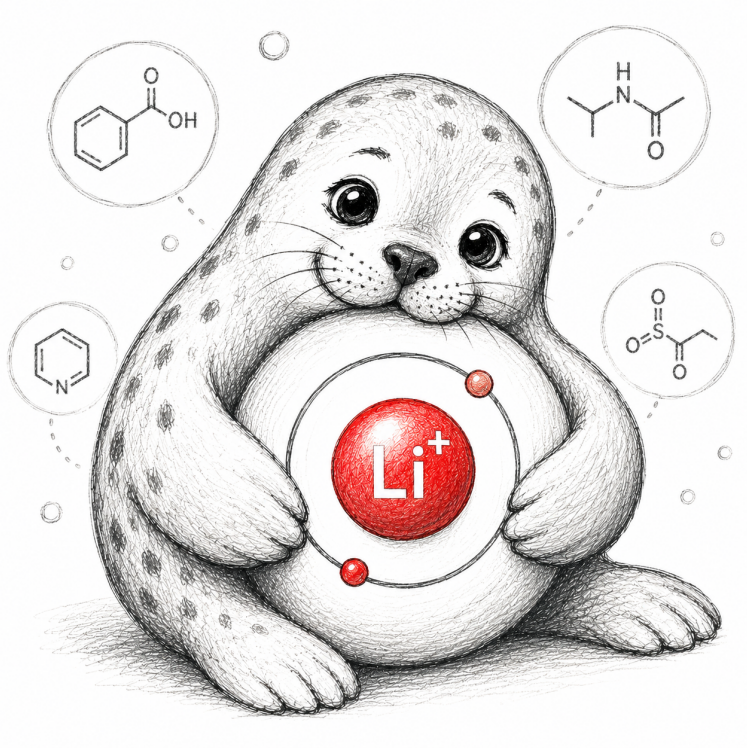
CHALMERS
UNIVERSITY OF TECHNOLOGY




# SÄLEN and the Art of Metal Binding

## An Explainable Graph Neural Network for Fragment-Wise Prediction of Metal-Adduction from Mass Spectrometry Data

Master's thesis in Complex Adaptive Systems

JONATAN LARSEN & FILIP NYMAN

DEPARTMENT OF PHYSICS AND ASTRONOMY

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2026

www.chalmers.se

Master's thesis 2026

# SÄLEN and the Art of Metal Binding

An Explainable Graph Neural Network for Fragment-Wise Prediction of Metal-Adduction from Mass Spectrometry Data

JONATAN LARSEN & FILIP NYMAN


CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Physics and Astronomy
Chalmers University of Technology
Gothenburg, Sweden 2026

SÄLEN and the Art of Metal Binding
An Explainable Graph Neural Network for Fragment-Wise Prediction of Metal-Adduction from Mass Spectrometry Data
JONATAN LARSEN & FILIP NYMAN



Supervisors: Gustaf Hulte and Mats Josefson, Astra Zeneca
Advisor: Felix Bölle, Astra Zeneca
Examiner: Jan Gerken, Department of Mathematical Sciences

Master's Thesis 2026
Department of Physics and Astronomy
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Cartoon of a seal contemplating the binding of a metal ion with various molecules, symbolising the SÄLEN model. Generated with OpenAI's ChatGPT image-generation tools.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

SÄLEN and the Art of Metal Binding
An Explainable Graph Neural Network for Fragment-Wise Prediction of Metal-Adduction from Mass Spectrometry Data
JONATAN LARSEN & FILIP NYMAN
Department of Physics and Astronomy
Chalmers University of Technology

# Abstract

When searching for and analyzing existing pharmaceutical molecules, mass spectrometry is often used. One method utilizes metal-adduct formation, adding metals to the sample to see which compound its rate and location of binding match. Detecting and interpreting metal-adduct formation in mass spectrometry data is challenging, especially when the goal is not only to predict whether a signal will appear, but also to understand which parts of a molecule contribute to that prediction. In this thesis, we develop the Substructure Aware Ligand Explanation Network (SÄLEN), an ion-conditioned, fragment-wise graph neural network for predicting Mass Spectrometry (MS) derived metal-adduct-like signals from molecule–ion pairs.
SÄLEN was trained on experimental data from AstraZeneca and designed to provide intrinsic fragment-level explanations by decomposing each prediction into contributions from molecular substructures. The model achieved competitive regression performance compared with established molecular models, with a test Mean Absolute Error (MAE) of 0.81 in log-intensity space. In the multitask setting, SÄLEN reached a positive classification precision of 92% and recall of 70%, making it suitable for high-confidence prioritization of molecule–ion pairs.
The fragment-wise explanations were evaluated through counterfactual atom edits, fragment swaps, inter-model comparison, and comparison with Quantum Mechanical (QM) descriptors. These analyses suggest that SÄLEN often highlights chemically plausible regions; although the explanations should be interpreted as model sensitivities rather than proof of physical binding mechanisms.
Overall, SÄLEN provides a useful framework for interpretable prediction of metal-adduct-like MS signals and may support hypothesis generation and prioritization in experimental workflows.

Keywords: ms-ms, machine learning, GNN, experimental training data, XAI, DFT, metal-adduct

# Acknowledgements

Firstly, we would like to express our gratitude to our supervisors, Gustaf Hulte and Mats Josefson, for creating an environment of curiosity and learning, and for generously giving us their time during our many spontaneous meetings.
We would also like to thank our examiner, Jan Gerken, for providing timely and constructive feedback throughout the project.
We are grateful to Felix Bölle for introducing us to the business side of neural networks, as well as to the office and its people.
Finally, we would like to thank our fellow thesis writer, André Maio, for sharing his hard-earned simulation data with us.

Jonatan Larsen & Filip Nyman, Gothenburg, June 2026

# List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

| | |
|---|---|
| AZ | AstraZeneca |
| BRICS | Breaking of Retrosynthetically Interesting Chemical Substructures |
| CHELPG | Charges from Electrostatic Potentials using a Grid-based method |
| DFT | Density Functional Theory |
| DMSO | Dimethyl sulfoxide, a solvent. |
| ESP | Electrostatic Potential |
| FiLM | Feature-wise Linear Modulation |
| GNN | Graph Neural Network |
| MAE | Mean Absolute Error |
| MLP | Multi Layer Perceptron |
| MPNN | Message Passing Neural Network |
| MSE | Mean Square Error |
| QM | Quantum Mechanical |
| RMSE | Root Mean Square Error |
| SÄLEN | Substructure Aware Ligand Explanation Network |
| SEAL | Substructure Explanation via Attribution Learning |
| SMILES | Simplified Molecular Input Line Entry System |
| TIMS | Trapped Ion Mobility Spectrometry |
| TOF | Time Of Flight |

# Contents

# List of Figures

# List of Tables

# 1
# Introduction

The identification and analysis of new pharmaceutical compounds at AstraZeneca (AZ) is a labour-intensive process that relies on both chemical intuition and advanced analytical instrumentation. The chemical space relevant to pharmaceutical development is vast, making it difficult not only to identify promising compounds from experimental measurements, but also to design molecules with a desired set of properties[1]. In this search mass spectrometry is extensively used, providing vast amounts of data of several dimensions to sift through[2]. These dimensions can be as simple things as pH of the compound with liquid chromatography (LC), or more intricate properties such as the rate of metal-adduct formation. The latter can be used to identify or narrow down the search for chemicals by adding metal ions to a sample and see what compound its rate and location of binding matches. There still exist no classical computer models that can compete with a human mind in this analysis.

To support this workflow, AZ is exploring machine learning methods for predicting and interpreting molecular properties from experimental data. This would give analysts a tool to limit the chemical space to a few possible candidates in a given search, freeing up both time and lab-resources.

Our contribution to this effort is SÄLEN, a graph neural network (GNN) designed to predict calibrated MS-derived metal-adduct signals for molecule–ion pairs. In addition to prediction, SÄLEN provides fragment-wise explanations, allowing the user to inspect which parts of the molecular structure contributed to the predicted signal.

## 1.1 Aim

The aim of this thesis is to develop an interpretable machine-learning model for predicting MS-derived metal-adduct-like signals from molecule–ion pairs. The model should fit into the existing AstraZeneca data-processing workflow and provide predictions that can be inspected at the level of molecular fragments.

A central goal is therefore not only predictive performance, but also explainability. The model should help identify molecule–ion pairs worth further inspection and suggest which molecular substructures contribute to the predicted signal.

**Table 1.1:** Mean absolute error, R2, root mean square and their standard deviation of log1p, described in Equation 3.1, of predicted values from 5 runs of different models

| Model | MAE ± std | R2 ± std | RMSE ± std |
|---|---|---|---|
| SÄLEN | 0.8085 ± 0.0029 | 0.7562 ± 0.0031 | 1.1172 ± 0.0070 |
| Chemprop | 0.810 ± 0.003 | 0.748 ± 0.004 | 1.125 ± 0.008 |
| Attentive-fp | 0.794 ± 0.003 | 0.759 ± 0.003 | 1.111 ± 0.007 |
| XG-boost | 0.864 ± 0.001 | 0.737 ± 0.0004 | 1.1602 ± 0.0009 |

## 1.2 Summarized Results

The result of this is SÄLEN, a molecular GNN wherein the molecule is not evaluated as a whole, rather, it is split up into chemically relevant fragments which in turn pool their individual predictions. These predictions come in the form of whether a binding is expected and to what degree it is then predicted to bind. The predictions can be looked at as the sum to give information about the molecule or as the individual contributions to see which part of the molecule is associated with binding.

SÄLEN is trained with a experimentally derived dataset of 92K samples with a split of 80%, 10% and 10% for training, validation and test-set. Each sample included the identity of the molecule, the calibrated intensity from the spectrometer and the ion it was tested with. The calibration was performed to get the intensity of calibration compounds to roughly the same level for all experiments. This adjustment of the measurements was key in gaining usable results from the models.

SÄLEN was compared to the following baseline models. Chemprop, a chemical GNN that evaluates the molecule using the bindings to carry the information, lacking SÄLENs division into larger pieces. Attentive-fp, another chemical GNN that utilizes a "supernode" to allow some communication between all nodes. XGBoost, a tree-ensemble method without usage of neural networks.

The SÄLEN model in its final form performed comparably to the stock version of Chemprop, outperforming XGBoost, but was outperformed by the model Attentive-fp, while maintaining intrinsic fragment explainability. Relevant performance metrics are displayed below in Table , 1.1.

At the chosen threshold, 92% of SÄLENs predicted positive molecule–ion pairs were positive according to the constructed detection labels, with a recall of 70%.

The model can provide explainability in part of its prediction, what structures are associated with binding in the molecule, but not in the chemical reasoning of said predictions. These explanations were, in a small number, analysed by an expert and deemed accurate in some instances and less so i others. The very nature of this analysis made it impossible to gather a statistically meaningful sample size. A small number of molecules, 700, of quantum mechanical simulations were performed to compare to the predictions of SÄLEN and gave a weak but noticeable correlation.

## 1.3 Limitations

The work conducted and conclusions drawn is limited by a number of factors. The chief part of these are mentioned below.
The number of metals tested and the number of compounds tested for each metal is limited by the what was feasible for the lab to produce. Some metals interact badly with the machinery and others are simply inaccessible. Silver, for example, gave nice strong signals from the measurements but precipitated in the spectrometer causing fouling and reduced sensitivity in following runs. This metal was therefore only included in one run, reducing the dataset and thereby the reliability concerning this specific metal. Other metals simply never showed up with high enough intensity to be deemed reliable to use in the dataset.
We have no real control group or ground truth to compare our explanation related results to. A more extensive list of better simulations could have given us answers to the question if the bind actually happens where the model predicts and if the effects of the fragments can explain physical phenomena. The quantum mechanical simulations performed were done at a simulated 0K in temperature and with no ion present to distort the electron clouds or the molecule itself. A comprehensive simulation where the physical placement of the ion and the induced 3D structure of the molecule is iterated would far exceed our computational resources.
Detections by the measuring equipment can only confirm the existence of complexes while there is no way to confirm the absence of a complex. This reduces the available data considerably as most combination of molecule and ion yielded zero intensity and were thereby discarded. Even the zero intensities used from the repeated measurements had a risk of being misclassification which may have skewed the model to produce more false negatives.

# 2
# Background

This thesis combines experimental mass spectrometry data with graph-based machine learning. This chapter introduces the chemical, analytical, and machine-learning concepts needed to understand the dataset, the models, and the interpretation methods used in the thesis.

## 2.1 Chemistry

To understand the methodology of this project it is important to get acquainted with a few chemical concepts. Here we will talk about how molecules and ions bind together, how the complex structure of a molecule can be easily written down and different ways computers can generate chemical information.

### 2.1.1 Ligand Binding

Atoms and molecules can interact through several types of bonding. The more well known covalent bonds are strong and often form stable molecules, while coordination interactions between a ligand and a metal ion are usually weaker and more reversible. It is this weak type of binding that this thesis will focus on. A ligand–metal complex may form and dissociate repeatedly, making the measured signal dependent on the experimental conditions.
Metal coordination is often driven by local molecular features [3]. Donor atoms, such as oxygen, nitrogen, sulphur or phosphorus, can act as coordination sites for metal ions. If several such sites are arranged so that the molecule can fold around the ion, the molecule may chelate the metal. Chelation can strongly increase complex stability, which makes both local donor groups and the structure connecting them important for predicting metal-adduct behaviour.

### 2.1.2 Coordination-Relevant Substructures in Organic Ligands

Metal ions are commonly positively charged in solution and can coordinate to electron-rich regions of organic ligands. The strength and selectivity of this interaction depend on several factors, including donor atoms, local charge distribution, coordination number, ligand geometry, and the Lewis acid character of the metal ion [3]. Hard–soft acid–base theory is a useful simplified framework for this: hard metal ions tend to interact more strongly with hard donor atoms such as oxygen, while

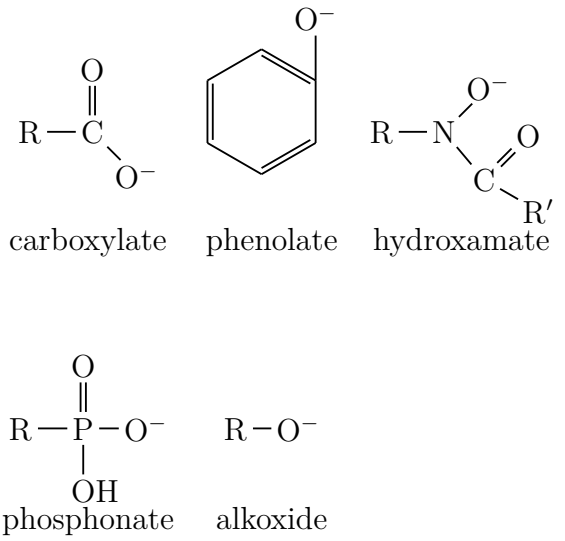



**Figure 2.1:** Examples of oxygen-containing donor motifs relevant for metal coordination.

softer metal ions may interact more favourably with softer donors such as sulphur or certain nitrogen environments [4].

The local environment around a donor atom is often as important as the donor atom itself. Functional groups can stabilize negative charge, delocalize electron density, or position several donor atoms close to each other. If several donor atoms can bind the same metal ion, the ligand may form a chelate complex. Such multidentate binding can increase complex stability compared with a single isolated donor atom, making both the donor atoms and the molecular scaffold connecting them important [5].

Common coordination-relevant substructures include carboxylates, phenolates, hydroxamates, phosphonates, alkoxides, amines, pyridine-like nitrogens, and sulphur-containing groups. Oxygen-containing motifs are especially common in metal coordination and several examples are shown in Figure 2.1. Similar motifs are also central in biological metal-binding systems such as siderophores, where hydroxamate, catecholate, and carboxylate groups are frequently used to coordinate metal ions [6].

Carboxylates are common coordination motifs because their negatively charged oxygens can interact strongly with metal ions. Phenolates and other aromatic systems can also be relevant because charge and electron density can be delocalized through the ring system. Hydroxamates and phosphonate-like groups are important because they combine strongly donating atoms with local geometries that can support metal coordination.

However, the presence of a heteroatom alone does not guarantee coordination. A nitrogen, oxygen, sulphur, or phosphorus atom may be a poor donor if its electron density is unavailable due to bonding, resonance, protonation state, steric shielding, or local geometry. For this reason, coordination-relevant substructures are better understood as local chemical environments rather than isolated atoms. This motivates using fragment-level representations in SÄLEN: fragments can preserve donor atoms together with the surrounding structure that determines whether they are

likely to contribute to metal-adduct-like signal.

### 2.1.3 Simplified Molecular Input Line Entry System (SMILES)

The Simplified Molecular Input Line Entry System (SMILES) represents molecular structures as strings of ASCII characters [7]. A SMILES string defines a molecular graph, although the same compound can often be written using several different valid SMILES depending on the starting atom and traversal order. This makes SMILES useful for storing molecular structures in a form that can be read by both cheminformatics software and machine-learning pipelines.

### 2.1.4 Density Functional Theory (DFT)

Density Functional Theory (DFT) is a quantum-chemical method used to estimate molecular structure and electronic properties [8]. Instead of explicitly solving the full many-electron wave-function which computationally scales exponentially with electrons, DFT describes the system through its electron density which instead scales cubically to the number of electrons. This makes it more computationally feasible than many wave-function-based methods, while still providing chemically useful information. Even so, DFT simulations quickly exceed modern computational power when molecules in the scale of pharmaceuticals are modelled.

In this thesis, DFT-derived descriptors are used as a plausibility check for the model explanations. Relevant quantities include electrostatic potential (ESP), atomic polarizability, and charges from electrostatic potentials using a grid-based method (CHELPG). These descriptors can indicate electron-rich or polarizable regions of a molecule, although they remain approximations and depend on the simulation setup.

### 2.1.5 RDKit

RDKit is an open-source cheminformatics toolkit used to parse molecules, compute descriptors, and manipulate molecular graphs [9]. The descriptors are less in depth than quantum-chemical calculations, but they are fast to compute and useful for large datasets. In this thesis, RDKit is used both for molecular processing and for generating atom-, bond-, molecule-, and fragment-level features.

## 2.2 Mass Spectrometry

Most of the data used in this thesis were collected using mass spectrometry. Mass spectrometry measures ions according to their mass-to-charge ratio and produces spectra from which molecular signals can be identified. In this project, the relevant output is the integrated intensity of detected molecule–ion combinations after processing [10].

### 2.2.1 timsTOF HT with Chromatography System

The experimental setup used chromatography followed by a Bruker timsTOF HT instrument [11]. Liquid chromatography separates compounds before mass-spectrometric analysis, in this case partly according to retention behaviour during a solvent gradient. Compounds that interact differently with the chromatograph elute at different times.
Trapped Ion Mobility Spectrometry (TIMS) provides additional separation based on ion mobility. Ion mobility can be interpreted as how easily an ion moves through gas under an electric field and is dependent on its charge, size, and shape. This helps separate ions that may have similar mass-to-charge ratios.
Time-of-flight (TOF) mass spectrometry then estimates the mass-to-charge ratio by accelerating ions and measuring their flight time [12]. Lighter ions, or ions with higher charge, travel faster under the same accelerating field. Together, chromatography, TIMS, and TOF provide retention time, ion mobility, and mass information for each measurement, of which thousands occur every second. An example of the data collected can be seen in Figure 2.2

### 2.2.2 MetaboScape

Raw mass spectrometry data require extensive processing before they can be used for modelling. MetaboScape was used to search the spectra for predefined molecule–ion complexes [13]. The software compares expected molecular properties, such as mass and isotope pattern, with observed peaks in the data.
This approach requires that the possible compounds and adducts are specified beforehand. A molecule–ion combination that is not searched for will not appear in the processed output, even if it is present in the raw data. As more combinations are searched for, the risk of ambiguous or overlapping assignments also increases. The processed output used in this thesis is a list of detected SMILES–ion combinations and their integrated peak intensities, referred to as intensities.

## 2.3 Graph Neural Networks

Artificial neural networks approximate complex functions by combining many simple computational units. A basic neuron takes an input vector $\mathbf{x} \in \mathbb{R}^n$, applies a weight vector $\mathbf{w} \in \mathbb{R}^n$ and bias $b \in \mathbb{R}$, and passes the result through an activation function $\phi : \mathbb{R} \to \mathbb{R}$:

$$y = \phi\left(\mathbf{w}^\top \mathbf{x} + b\right), \tag{2.1}$$

where $y \in \mathbb{R}$ is the neuron output.
A common activation function is the rectified linear unit, ReLU:

$$\mathrm{ReLU}(z) = \max(0, z) = \begin{cases} z, & z > 0 \\ 0, & z \leq 0. \end{cases} \tag{2.2}$$

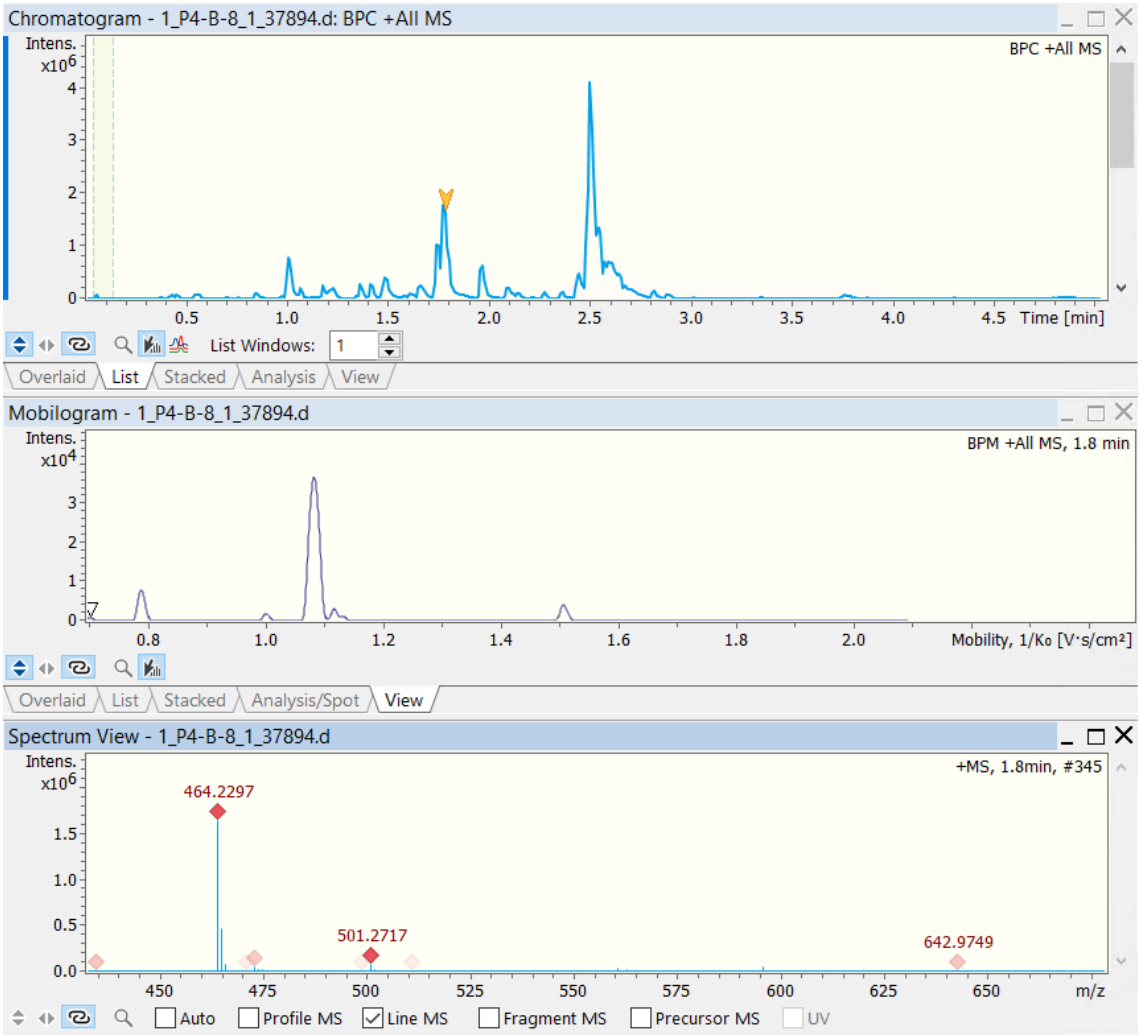



**Figure 2.2:** Example of processed MS output for one signal. The top panel shows the chromatogram, where intensity is plotted against retention time. Selecting a retention-time interval gives the middle panel, the mobilogram, where ions are separated by ion mobility. Selecting a mobility interval gives the bottom panel, the mass spectrum, where intensity is plotted against mass-to-charge ratio ($m/z$). Together, the three panels illustrate how a detected signal is localized by retention time, ion mobility, and $m/z$.

Here, ReLU is a scalar function $\mathrm{ReLU} : \mathbb{R} \to \mathbb{R}$. When used on vectors, it is applied element-wise.
Neurons are arranged in layers, where each layer transforms a representation $\mathbf{h}^{(l-1)} \in \mathbb{R}^{d_{l-1}}$ into a new representation $\mathbf{h}^{(l)} \in \mathbb{R}^{d_l}$:

$$\mathbf{h}^{(l)} = \phi\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right), \tag{2.3}$$

where $\mathbf{W}^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$, $\mathbf{b}^{(l)} \in \mathbb{R}^{d_l}$, and $\phi : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l}$ is applied element-wise.
A multilayer perceptron (MLP) is a basic example of a neural network, where layers of neurons are stacked to produce some output.
Graph neural networks (GNNs) extend this idea to graph-structured data [14, 15]. For molecular data, one common representation is to define atoms as nodes and chemical bonds as edges, which is the representation used in this thesis. Other graph definitions are also possible; for example, if bond information is unavailable or if 3D structure is used, edges can instead be defined by a distance cut-off between atoms.
A broad way to describe many GNNs is message passing, where node representations are updated by aggregating information from neighbouring nodes. Equation 2.4 shows a simplified special case of such an update, where each node representation $\mathbf{h}_v^{(l)} \in \mathbb{R}^{d_l}$ is updated using its own representation and the sum of transformed neighbouring representations:

$$\mathbf{h}_v^{(l+1)} = \phi\left(\mathbf{W}_{\text{self}}^{(l)}\mathbf{h}_v^{(l)} + \sum_{u \in \mathcal{N}(v)} \mathbf{W}_{\text{neigh}}^{(l)}\mathbf{h}_u^{(l)} + \mathbf{b}^{(l)}\right), \tag{2.4}$$

where $\mathcal{N}(v)$ is the neighbourhood of node $v$, $\mathbf{h}_u^{(l)} \in \mathbb{R}^{d_l}$ is the representation of a neighbouring node, $\mathbf{W}_{\text{self}}^{(l)}, \mathbf{W}_{\text{neigh}}^{(l)} \in \mathbb{R}^{d_{l+1} \times d_l}$, $\mathbf{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$, and $\phi : \mathbb{R}^{d_{l+1}} \to \mathbb{R}^{d_{l+1}}$ is applied element-wise.
More general message-passing layers may use separate message and update functions, edge features, attention mechanisms, or learned aggregation schemes. In the simplified form above, each layer transfers information one step through the graph. Therefore, after several layers, a node can contain information from increasingly distant parts of the molecule. After the graph layers, the node representations are usually pooled into a graph-level representation and passed through a multilayer perceptron (MLP) to produce the prediction $\hat{y}$.
Each message-passing layer transfers information one step through the graph. Therefore, after several layers, a node can contain information from increasingly distant parts of the molecule. After message passing, the node representations are sent into a readout function, usually summed, giving a graph-level representation and passed through a MLP to produce the prediction $\hat{y}$.
GNNs are trained by comparing $\hat{y}$ with the target $y$ using a loss function. For a dataset with $N$ samples, where $\hat{y}_i \in \mathbb{R}$ is the prediction and $y_i \in \mathbb{R}$ is the target for sample $i$, two common regression losses are mean squared error (MSE),

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2, \tag{2.5}$$

and Smooth L1 loss,

$$\mathcal{L}_{\text{SmoothL1}} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}\frac{1}{2}\left(\hat{y}_i - y_i\right)^2, & |\hat{y}_i - y_i| < 1\\ |\hat{y}_i - y_i| - \frac{1}{2}, & |\hat{y}_i - y_i| \geq 1.\end{cases} \tag{2.6}$$

Both losses map from prediction-target pairs to a scalar loss value, $\mathcal{L} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$. Smooth L1 behaves quadratically for small errors and linearly for larger errors, making it less sensitive to outliers than MSE.
Model parameters are updated using backpropagation. In gradient descent form, the weight and bias updates are

$$\mathbf{W}_{t+1}^{(l)} = \mathbf{W}_t^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(l)}}, \tag{2.7}$$

$$\mathbf{b}_{t+1}^{(l)} = \mathbf{b}_t^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(l)}}, \tag{2.8}$$

where $\eta \in \mathbb{R}_{>0}$ is the learning rate, $\mathbf{W}_t^{(l)}$ and $\mathbf{b}_t^{(l)}$ are the parameters at optimization step $t$, and the gradients have the same dimensions as the corresponding parameters.
A key property of molecular GNNs is permutation invariance at the graph level: the final molecular prediction should not depend on the order in which atoms are listed. This is essential for molecular modelling, since atom ordering in a SMILES-derived graph is arbitrary.

### 2.3.1 Feature-Wise Linear Modulation

Feature-wise Linear Modulation (FiLM) is a conditioning method for neural networks [16]. It uses an external conditioning vector to produce a feature-wise scaling vector and shift vector, which are then used to modulate hidden representations.
Given a hidden representation $\mathbf{h}^{(l)} \in \mathbb{R}^{d_l}$ at layer $l$ and a conditioning vector $\mathbf{c} \in \mathbb{R}^{d_c}$, FiLM computes

$$\boldsymbol{\gamma}^{(l)}(\mathbf{c}), \boldsymbol{\beta}^{(l)}(\mathbf{c}) = f_{\text{FiLM}}^{(l)}(\mathbf{c}), \tag{2.9}$$

where $f_{\text{FiLM}}^{(l)} : \mathbb{R}^{d_c} \to \mathbb{R}^{2d_l}$ is a small neural network. The hidden representation is then modulated as

$$\tilde{\mathbf{h}}^{(l)} = \boldsymbol{\gamma}^{(l)}(\mathbf{c}) \odot \mathbf{h}^{(l)} + \boldsymbol{\beta}^{(l)}(\mathbf{c}), \tag{2.10}$$

where $\boldsymbol{\gamma}^{(l)}, \boldsymbol{\beta}^{(l)} \in \mathbb{R}^{d_l}$ and $\odot$ denotes element-wise multiplication.
In this thesis, the conditioning vector $\mathbf{c}$ represents the metal ion. Instead of providing ion information only at the input, FiLM allows the model to reintroduce ion-specific context at several graph layers. The FiLM modules are initialized so that they initially behave close to the identity transformation, reducing the risk of adding unnecessary noise at the start of training.

### 2.3.2 Edge-Conditioned Message Passing

In a basic GNN, messages are passed along graph edges without necessarily accounting for the type of edge. In molecular graphs, however, bonds contain important

chemical information. Edge-conditioned message passing allows bond features to influence how messages are transferred [15].
Let $\mathbf{h}_u^{(l)} \in \mathbb{R}^{d_l}$ be the representation of a neighbouring node $u$ at layer $l$, and let $\mathbf{e}_{uv} \in \mathbb{R}^{d_e}$ be the edge-feature vector for the bond between nodes $u$ and $v$. In this work, the edge features are passed through a small neural network followed by a sigmoid activation to produce a gate,

$$\mathbf{g}_{uv}^{(l)} = \sigma\left(f_{\text{edge}}^{(l)}(\mathbf{e}_{uv})\right), \tag{2.11}$$

where $f_{\text{edge}}^{(l)} : \mathbb{R}^{d_e} \to \mathbb{R}^{d_l}$ and $\mathbf{g}_{uv}^{(l)} \in \mathbb{R}^{d_l}$. The message from node $u$ to node $v$ is then scaled element-wise by this gate,

$$\mathbf{m}_{u\to v}^{(l)} = \mathbf{g}_{uv}^{(l)} \odot \mathbf{h}_u^{(l)}, \tag{2.12}$$

where $\odot$ denotes element-wise multiplication. The gate allows the model to learn which bond features should promote or suppress different parts of the message passed between atoms.

### 2.3.3 Substructure Explanation via Attribution Learning (SEAL)

SEAL is an explainable molecular GNN designed to produce fragment-wise attributions [17]. Before message passing, each molecule is partitioned into fragments, with each atom belonging to exactly one fragment. In the original SEAL model, this partitioning is based on an expanded version of the Breaking of Retrosynthetically Interesting Chemical Substructures (BRICS) algorithm [18]. BRICS decomposes molecules by cutting chemically meaningful bonds, originally motivated by retrosynthetic fragmentation. This produces substructures that often correspond to recognizable chemical building blocks.
SEAL separates messages depending on whether they are passed within a fragment or between fragments. Let $\mathcal{F}(v)$ denote the fragment containing node $v$. The intra-fragment neighbourhood is defined as

$$\mathcal{N}_{\text{in}}(v) = \{u \in \mathcal{N}(v) : \mathcal{F}(u) = \mathcal{F}(v)\}, \tag{2.13}$$

while the inter-fragment neighbourhood is

$$\mathcal{N}_{\text{out}}(v) = \{u \in \mathcal{N}(v) : \mathcal{F}(u) \neq \mathcal{F}(v)\}. \tag{2.14}$$

The corresponding mean neighbour representations are

$$\bar{\mathbf{h}}_{\text{in},v}^{(l)} = \frac{1}{|\mathcal{N}_{\text{in}}(v)|} \sum_{u\in\mathcal{N}_{\text{in}}(v)} \mathbf{h}_u^{(l)}, \tag{2.15}$$

and

$$\bar{\mathbf{h}}_{\text{out},v}^{(l)} = \frac{1}{|\mathcal{N}_{\text{out}}(v)|} \sum_{u\in\mathcal{N}_{\text{out}}(v)} \mathbf{h}_u^{(l)}. \tag{2.16}$$

If one of these neighbourhoods is empty, the corresponding term is omitted. The SEAL node update can then be written as

**Table 2.1:** Standard atomic properties used by Chemprop.

| Property | Explanation |
|---|---|
| Degree | Number of bonds to the atom. |
| Formal charge | Difference between the assigned valence electrons in the molecule and the neutral free-atom state. |
| Chiral tag | Stereochemical configuration around the atom. |
| Number of hydrogens | Number of hydrogens attached to the atom. Hydrogens are often implicit in SMILES. |
| Hybridization | Hybridization state of the atom's valence orbitals. |

$$\mathbf{h}_v^{(l+1)} = \phi\left(\mathbf{W}_{\text{self}}^{(l)}\mathbf{h}_v^{(l)} + \mathbf{W}_{\text{intra}}^{(l)}\bar{\mathbf{h}}_{\text{in},v}^{(l)} + \mathbf{W}_{\text{inter}}^{(l)}\bar{\mathbf{h}}_{\text{out},v}^{(l)} + \mathbf{b}^{(l)}\right). \tag{2.17}$$

Here, $\mathbf{W}_{\text{intra}}^{(l)}$ and $\mathbf{W}_{\text{inter}}^{(l)}$ are separate weight matrices for messages within and between fragments. This allows the model to treat information exchange inside a fragment differently from information exchange across fragment boundaries.
After the graph layers, atom representations are pooled within each fragment. Each fragment representation is passed through an MLP to produce a fragment contribution, and the final prediction is obtained by summing these contributions. This creates an intrinsic explanation, since the model output is directly decomposed into fragment-level terms.
SEAL can also regularize inter-fragment message passing, encouraging information to remain localized within fragments. This supports interpretability, since fragment contributions are easier to interpret when they are not dominated by information from distant parts of the molecule. In this thesis, SÄLEN builds on this idea but adapts the architecture and fragmentation to metal-adduct-like MS prediction.

### 2.3.4 Chemprop

Chemprop is a molecular GNN based on directed message passing [19]. It encodes atoms and bonds as feature vectors and passes messages along directed bonds to learn molecular representations. The direction of a bond is not a chemical direction; instead, each undirected molecular bond is represented by two directed edges, one in each direction. This allows Chemprop to update messages along directed paths in the molecular graph.
Standard atom and bond features include the properties listed in Tables 2.1 and 2.2 [20, 21]. The learned molecular representation, optionally combined with user-defined descriptors, is then passed through an MLP to predict molecular properties. Both atom and bond feature sets can be extended by the user.

### 2.3.5 Attentive FP

Attentive FP, where FP stands for fingerprint, is an attention-based molecular GNN designed for molecular property prediction [22]. Instead of treating all neighbouring messages equally, it uses attention mechanisms to weight information according

**Table 2.2:** Standard bond properties used by Chemprop.

| Property | Explanation |
|---|---|
| Bond | Whether a bond is present between the two atoms. |
| Bond type | Type of bond, such as single, double, triple, or aromatic. |
| Conjugated | Whether the bond is part of a conjugated system. |
| In ring | Whether the bond is part of a ring. |
| Stereochemistry | Stereochemical configuration associated with the bond. |

to learned relevance. It also forms a graph-level molecular representation through attentive readout.
The attention weights can provide some insight into which atoms or neighbourhoods influence the learned molecular representation. However, attention weights should not automatically be interpreted as faithful explanations of the final prediction, since previous work has shown that attention can be an unreliable explanation signal [23, 24]. Attentive FP was therefore treated as a predictive baseline rather than as an explainability method in this thesis.
This makes Attentive FP a strong baseline for tasks where distant parts of the molecule may jointly influence the target. Such long-range effects are relevant for metal coordination, where donor atoms separated in the graph may become spatially close through molecular conformation.

## 2.4 XGBoost

Extreme Gradient Boosting, or XGBoost, is a tree-ensemble method based on gradient boosting [25]. It models the prediction $\hat{y}_i$ as an additive ensemble of $K$ regression trees:

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i). \tag{2.18}$$

Each new tree is fitted to improve the predictions of the existing ensemble. Regularization and pruning are used to reduce overfitting, and hyperparameters control model complexity and sensitivity to individual samples. In this thesis, XGBoost was useful both as a descriptor-based baseline and as a fast method for estimating the importance of RDKit descriptors.

# 3
# Data Gathering & Pre-Processing

This project covered the workflow from experimental planning and data processing to model design and evaluation. The experimental setup, including the choice of metals and molecule–ion combinations to search for, was planned in collaboration with AstraZeneca. The MS experiments were performed by our supervisor Gustaf Hulte, and the raw spectral data were processed in MetaboScape by our advisor Felix Bölle according to the selected search targets. The work in this thesis then focused on calibration, dataset construction, featurization, model development, and evaluation.
This chapter describes how the experimental MS data were generated, how the processed MetaboScape output was further calibrated, and how it was converted into model-ready targets and features.

## 3.1 Experimental Setup

Plates with $16 \times 24$ wells were prepared with 100 compounds per well, corresponding to 38400 compounds in total. Each compound was present at a concentration of 3 $\mu$mol $\mathrm{L}^{-1}$ in dimethyl sulfoxide (DMSO). Two types of metal solutions were prepared, using either water or DMSO as solvent. These solutions contained different metals at 5 mmol $\mathrm{L}^{-1}$, together with four calibration compounds, hereafter referred to as calibrants. The calibrants were added to all solutions, while only a subset of metals was included in each run to avoid overwhelming the analysis software with too many possible molecule–ion combinations. The solutions were also acidified with 0.1% formic acid by volume to improve metal solubility.
Some plates were tested with all metal solutions, while others were tested with only a subset, since certain metals proved impractical to continue with. To assess measurement accuracy and repeatability, selected plates were measured several times with the same metal solution. The plates were numbered 1 to 4, and a fifth plate that was named Mölndal-homepool. The full set of plate and metal combinations is shown in Table A.2.1.
The instrument was tuned to detect the majority of molecular masses in the sample set. However, the lowest masses, around 100 Da, were less reliably detected.
The plates were analysed using a chromatography system followed by TIMS-TOF HT. At the start of each injection, the solvent consisted of the water-based metal solution. The concentration of the DMSO-based metal solution was then gradually increased to 100%, separating compounds partly by retention behaviour before they entered the TIMS-TOF HT. For each well, this produced spectra containing

retention time, ion mobility, and mass information.

## 3.2 Data Processing Pipeline

Before the spectral data could be used for modelling, it required further processing, calibration, target construction, and featurization.

### 3.2.1 Spectral Analysis using MetaboScape

A set of possible molecule–ion combinations was defined based on chemical plausibility and the expected outcomes of the metal mixtures. These ranged from simple adducts such as [M+K]+ to more complex combinations such as [2M+Cr-H+NH4]+ and [M+Cr-2H+CH3(SO)CH3]+, where M denotes a molecule from the plate. The full setup can be found in Table A.6.
The spectral data were processed in MetaboScape by Felix Bölle using queries corresponding to the selected metal mixtures and SMILES structures present in each plate. These query definitions were developed as part of the project to target chemically plausible molecule–ion combinations. The resulting MetaboScape output associated each detected molecule–ion combination with an integrated intensity value and was used as the starting point for calibration and dataset construction.

### 3.2.2 Calibration

The raw MS intensities varied substantially due to technical effects, making calibration necessary before model training. Of the four calibrants, three were detected in a sufficient fraction of wells, above 40%, and were therefore used for calibration. Plate regions where no calibrants were detected were excluded both from the calibration procedure and from the final dataset. The calibrant intensities were log-transformed to reduce the influence of extreme values.
A shared least-squares calibration model was fitted across all plates. The model included calibrant identity, plate effect, and smooth run-order drift. Calibrant identity captured differences between the calibration compounds, while plate effect and run-order drift represented technical variation. The estimated technical bias was then transformed back from log space and applied as a multiplicative correction factor to the raw intensities, excluding regions where no calibrant signal was available.

### 3.2.3 Target Definition

Raw MS ion-count intensities are inherently noisy and vary by several orders of magnitude. The regression target was therefore defined by log-transforming the raw intensity $I$ according to Equation 3.1.

$$y = \log(1 + I) \tag{3.1}$$

This transformation reduces the influence of very large intensity values and improves numerical stability during training. The addition of 1 makes the expression well-defined for zero intensity values.

The log-transformed targets were then normalized using the mean and standard deviation of the training set, as shown in Equation 3.2.

$$\tilde{y} = \frac{y - \text{mean}(y_{\text{train}})}{\text{std}(y_{\text{train}})} \tag{3.2}$$

For the classification task, molecule–ion pairs were assigned positive or negative labels corresponding to detected and non-detected metal-adduct-like behaviour. Since non-detection does not necessarily imply absence of adduct formation, negative labels required additional selection criteria, described in Section 3.3.

### 3.2.4 Featurization

Chemical features were added to the molecular graph representation to support learning from the available data. Atom features were assigned to graph nodes, while bond features were assigned to graph edges. These features were selected based on their expected relevance to metal-adduct formation and molecular representation.
Molecular and fragment descriptors were selected using XGBoost. First, an XGBoost model was trained using the full set of available RDKit descriptors on the training set. The descriptors were then evaluated using a rolling mask procedure, where the effect of removing descriptor groups was assessed on the validation set. The selection criterion was the mean absolute error (MAE) of the log-intensity prediction.
Because different molecules produce different numbers of fragments, the fragments were ordered by the number of donor-like atoms before being stored in a fixed representation. Fragment descriptors were computed after fragmentation by treating each fragment as an individual RDKit molecule. These descriptors were then saved as fragment-level feature vectors. XGBoost selected similar descriptor groups for whole molecules and fragments, suggesting that related chemical properties were informative at both levels. The selected feature groups are listed in Tables A.2–A.4 in the appendix.

### 3.2.5 Statistical Analysis

During calibration analysis, the calibrants were found to have lower detection reliability than expected. This raised concern about interpreting undetected molecule–ion combinations as true negatives. To estimate detection reliability, repeated measurements of selected plates were analysed. Plate 4 was measured five times, while Mölndal-homepool was measured three times.
In a perfect measurement setup, every real and detectable molecule–ion complex would be found in every repeated run. Non-detection could then be interpreted as absence. In practice, some real signals may be missed, and the repeated measurements were therefore used to estimate the probability of missing a real signal in all runs.
The number of detections across repeated runs, $K$, was modelled as a binomial random variable:

$$K \sim \text{Binomial}(n, p), \tag{3.3}$$

where $n$ is the number of repeated runs and $p$ is the probability of detecting a given sample in one run. Since only samples with $K > 0$ were observed, a zero-truncated expectation was used:

$$\mathbb{E}[K \mid K > 0] = \frac{\mathbb{E}[K]}{P(K > 0)}. \tag{3.4}$$

For a binomial random variable,

$$\mathbb{E}[K] = np, \tag{3.5}$$

and

$$P(K > 0) = 1 - P(K = 0) = 1 - (1 - p)^n. \tag{3.6}$$

Therefore,

$$\mathbb{E}[K \mid K > 0] = \frac{np}{1 - (1 - p)^n}. \tag{3.7}$$

The detection probability $p$ was estimated by matching this expression to the observed mean number of detections among observed samples:

$$\frac{np}{1 - (1 - p)^n} = \bar{K}_{\text{obs}}. \tag{3.8}$$

The probability that a real sample was missed in all repeated runs was then

$$P(K = 0) = (1 - p)^n. \tag{3.9}$$

This gave an estimated all-run miss probability of 7.9% for plate 4 and 5.5% for Mölndal-homepool. These estimates provide an approximate upper bound on the achievable reliability of negative labels in this dataset. However, as discussed in Section 3.4, the observed detection counts deviate from a simple independent binomial model, meaning that these values should be interpreted as rough estimates rather than exact probabilities.

## 3.3 Synthetic Negatives

A central limitation of the data-generation procedure is the lack of definite negative samples, meaning molecule–ion pairs experimentally confirmed not to form metal-adduct-like signals. A non-detection can occur for several reasons: the molecule may not react, may ionize poorly, may fall below the detection limit, may form an unsearched adduct, or may be missed by the processing workflow. Therefore, assigning all undetected molecule–ion pairs as confident negatives would introduce substantial label noise.

The term synthetic negative is used because these negative labels were not directly measured as true non-binding cases. Instead, they were constructed from repeated non-detections under strict filtering criteria. A molecule–ion pair was selected as a synthetic negative only if three criteria were satisfied. First, the molecule was never detected with the specific metal ion. Second, the molecule was detected in

protonated form at least twice during repeated measurements, with intensity above 30000. Third, any detected molecule–ion pairings involving the same molecule and other metal ions had intensities above 10000.
These criteria were intended to ensure that the molecule itself was present and detectable, while the absence of the specific metal-adduct-like signal was consistent across repeated observations.
All other non-detected molecule–ion combinations were excluded rather than treated as negatives. Thus, the classification dataset consisted of detected positives together with the filtered synthetic negatives. The synthetic negatives were added before model training and then assigned to train, validation, and test according to the SMILES-level split.
Using these criteria on the repeated measurements of plate 4 and Mölndal-homepool yielded 20000 synthetic negatives. These samples were used only for the classification component of the multitask SÄLEN model and were not used for regression-loss updates. The regression-only models, including SÄLEN regression, Chemprop, AttentiveFP, and XGBoost, were trained and evaluated only on detected molecule–ion pairs with measured intensities.

## 3.4 Effect of Data Pre-Processing on the Dataset

Calibration had a limited effect on the remaining within-plate pattern, shown in Figure 3.1, but a large effect on variation between runs, shown in Figure 3.2. Both diagnostics are shown in log-intensity units. Here, “residual” and “offset” refer to differences from the average log calibrant intensity, not to the logarithm of a positive residual. Positive values indicate higher-than-average calibrant signal, while negative values indicate lower-than-average calibrant signal. Figure 3.1 shows the mean log-intensity residual by well position, while Figure 3.2 shows estimated plate-level log-intensity offsets.
In an ideal case, the calibrated plate surface would be flat, with no systematic dependence on well position. A stronger spatial correction could have been applied, but a smoother and more conservative calibration was preferred because the relationship between calibrant signal and sample signal was not fully known. A well could, by chance, have weak calibrant signal while still giving strong signals for other compounds. An overly aggressive correction would then risk amplifying those compound intensities too much.
Calibration diagnostics were evaluated in log-intensity space. The terms residual and offset therefore refer to differences between log-transformed calibrant intensities and fitted calibration terms, not to the logarithm of a positive residual. Figure 3.1 shows the mean calibrant log residual by well position, while Figure 3.2 shows fitted run-level log offsets before and after correction.
Figure 3.2 shows that the measurements differed substantially between runs. This makes calibration necessary before regression training, since technical variation of this magnitude could obscure the chemical signal the model is intended to learn.
Some plates were measured only in runs yielding low-intensity and less reliable results. However, the model performed better when these data were included rather

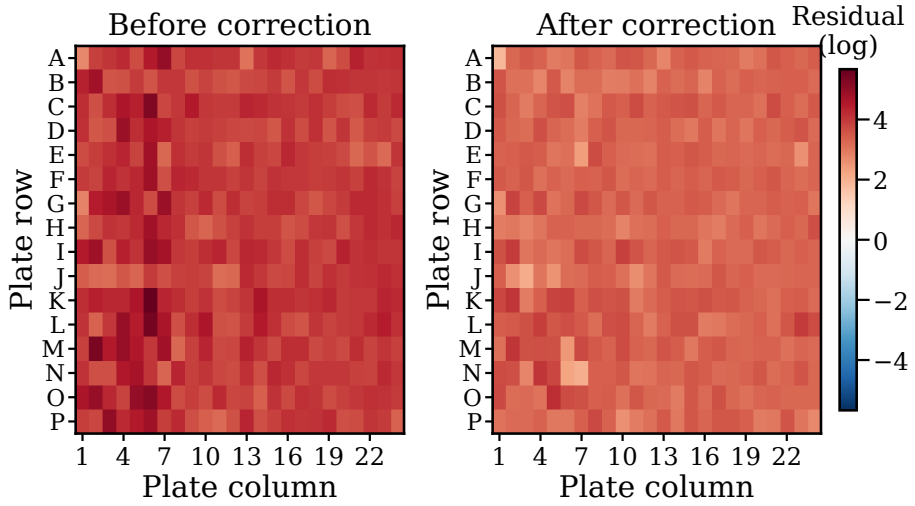



**Figure 3.1:** Mean calibrant log-intensity residual by well position before and after correction. Rows and columns correspond to the plate layout. The residual is computed in log-intensity space after subtracting the calibrant-specific baseline; it is not the logarithm of a residual. Positive values indicate higher-than-expected calibrant signal and negative values indicate lower-than-expected calibrant signal. The weak spatial pattern indicates that within-plate effects were small compared with run-to-run variation.

**Table 3.1:** Number of molecules detected in one or more repeated runs for plate 4 and Mölndal-homepool. The table counts whether a molecule was detected in any form, independent of the specific ion.

| Times found | Plate 4 | Mölndal-homepool |
|---|---|---|
| 1 | 3010 | 141 |
| 2 | 4181 | 292 |
| 3 | 5247 | 1760 |
| 4 | 7673 | – |
| 5 | 14192 | – |

than excluded. This suggests that the increased chemical coverage was more beneficial than the reduction in measurement noise.

The repeated measurements also showed that molecules were detected more consistently than specific molecule–ion complexes. Plate 4 was measured five times, while Mölndal-homepool was measured three times.

As shown in Table 3.1, many molecules were detected repeatedly, and for plate 4 most molecules were detected in all five runs. The same pattern was not observed for specific molecule–ion complexes.

Table 3.2 supports two conclusions. First, the instrument and processing workflow can miss molecule–ion complexes, since many complexes were detected in only a subset of repeated runs. Second, the detections do not appear to be independent and identically distributed. Under a simple binomial model with a fixed detection probability, the number of detections should follow a unimodal binomial-like distri-

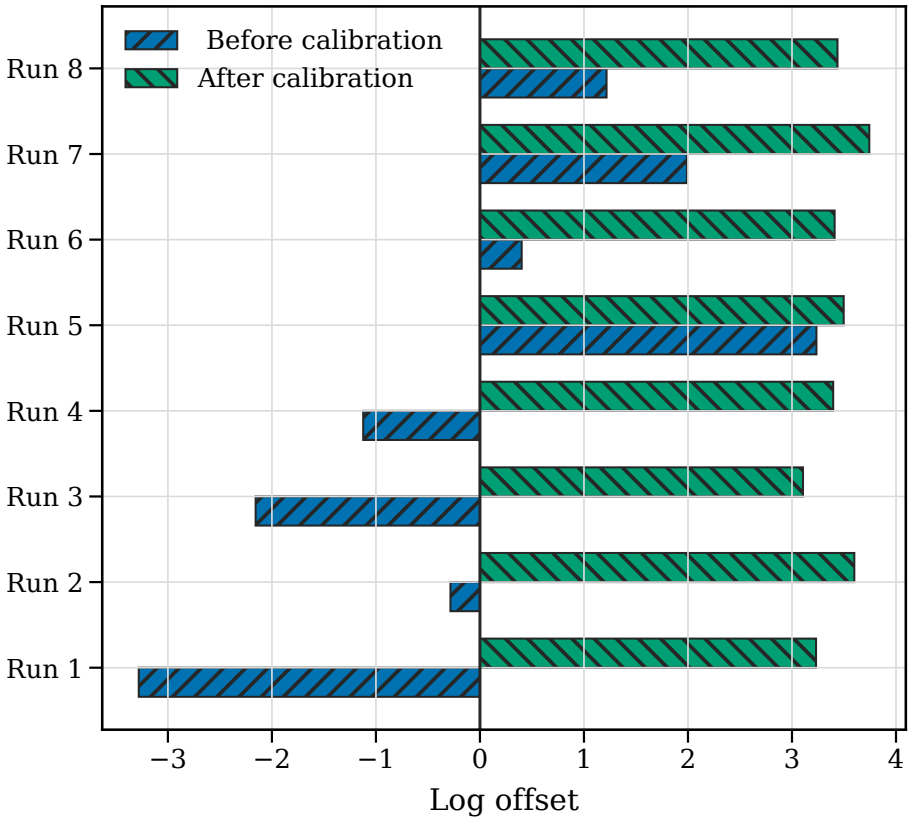



**Figure 3.2:** Estimated run-level calibrant log offsets before and after calibration. Each row corresponds to one run or plate. The before-correction values are fitted plate effects, while the after-correction values are median corrected calibrant residuals in log-intensity space. Before correction, the fitted run-level offsets spanned about 6.5 log units, corresponding to roughly a 670-fold difference in calibrant signal between runs.

**Table 3.2:** Number of molecule–ion complexes detected in one or more repeated runs for plate 4 and Mölndal-homepool. Complex-level detections were less repeatable than molecule-level detections.

| Times found | Plate 4 | Mölndal-homepool |
|---|---|---|
| 1 | 45620 | 2267 |
| 2 | 21478 | 1414 |
| 3 | 10134 | 2093 |
| 4 | 7542 | – |
| 5 | 12017 | – |

**Table 3.3:** Observed complex-detection counts compared with expected counts under the fitted zero-truncated binomial model. Expected counts were computed using the detection probabilities corresponding to all-run miss rates of 7.9% for plate 4 and 5.5% for Mölndal-homepool.

| Times found | Plate 4 obs. | Plate 4 exp. | Mölndal obs. | Mölndal exp. |
|---|---|---|---|---|
| 1 | 45620 | 27441 | 2267 | 1635 |
| 2 | 21478 | 36316 | 1414 | 2677 |
| 3 | 10134 | 24031 | 2093 | 1461 |
| 4 | 7542 | 7951 | – | – |
| 5 | 12017 | 1052 | – | – |

bution, and under perfect measurement, all detections would occur 5 and 3 times for plate 3 and Mölndal-homepool respectively. Instead, plate 4 contains many complexes detected only once and many detected in all five runs, suggesting that some complexes are intrinsically much easier to detect than others.
The deviation between observed and expected counts in Table 3.3 shows that the binomial model is only an approximation. In particular, the large number of complexes detected in all five plate 4 runs indicates that detection probability varies substantially between complexes. Nevertheless, the calculation is still useful as a simple estimate of false-negative risk. Using the method described in Section 3.2.5, the estimated probability of missing a real signal in all repeated runs was 7.9% for plate 4 and 5.5% for Mölndal-homepool.

# 4

# Model Architecture & Performance

Four models were used for the regression task: AttentiveFP, Chemprop, XGBoost, and our own SÄLEN model, based on the SEAL architecture. The baseline models were kept as close as possible to their standard versions, with the addition of relevant molecular and ion features to make the comparison fair. Only SÄLEN was then fine-tuned with a multitask regression and classification head.

## 4.1 SÄLEN Architecture

SÄLEN builds on the SEAL architecture by adding ion conditioning, edge-feature gates, fragment descriptors, and a multitask prediction head. These additions are summarized below.

Ion information is introduced in two ways. First, the ion feature vector $\mathbf{u} \in \mathbb{R}^{d_u}$ is concatenated to each initial atom feature vector $\mathbf{x}_v \in \mathbb{R}^{d_x}$,

$$\tilde{\mathbf{x}}_v = [\mathbf{x}_v \,\|\, \mathbf{u}], \tag{4.1}$$

where $\|$ denotes concatenation. This gives every atom direct access to the ion identity and ion descriptors at the input layer.

Second, ion information is reintroduced at each graph layer using Feature-wise Linear Modulation (FiLM) [16]. For layer $l$, a single layer perceptron maps the ion vector to a scaling vector and a shift vector,

$$\boldsymbol{\gamma}^{(l)}(\mathbf{u}), \boldsymbol{\beta}^{(l)}(\mathbf{u}) = f^{(l)}_{\text{FiLM}}(\mathbf{u}), \tag{4.2}$$

where $\boldsymbol{\gamma}^{(l)}, \boldsymbol{\beta}^{(l)} \in \mathbb{R}^{d_l}$. The node representation is then modulated as

$$\tilde{\mathbf{h}}^{(l)}_v = \boldsymbol{\gamma}^{(l)}(\mathbf{u}) \odot \mathbf{h}^{(l)}_v + \boldsymbol{\beta}^{(l)}(\mathbf{u}), \tag{4.3}$$

where $\odot$ denotes element-wise multiplication. This allows the model to condition the message-passing layers on the ion, rather than relying only on the input representation.

Edge-feature-informed gates were also added to account for how bonds affect the messages being transferred. For an edge between nodes $u$ and $v$, with edge feature vector $\mathbf{e}_{uv}$, a gate is computed as

$$\mathbf{g}^{(l)}_{uv} = \sigma\left(f^{(l)}_{\text{edge}}(\mathbf{e}_{uv})\right), \tag{4.4}$$

where $\mathbf{g}_{uv}^{(l)} \in \mathbb{R}^{d_l}$ and $\sigma$ is the sigmoid function. The message from node $u$ to node $v$ is then scaled by this gate,

$$\mathbf{m}_{u\to v}^{(l)} = \mathbf{g}_{uv}^{(l)} \odot \mathbf{h}_u^{(l)}. \tag{4.5}$$

This lets the model learn which bond features should promote or suppress different parts of the message.

Whole-molecule descriptors improved the performance of Chemprop and AttentiveFP, but using them directly in SÄLEN would weaken the fragment-wise explanation. Part of the prediction would then come from global molecular information rather than from individual fragments. To keep the information localized, descriptors were instead calculated for each fragment separately, treating each fragment as an individual molecule. For fragment $F_k$, the pooled fragment representation $\mathbf{r}_k$ was concatenated with a fragment descriptor vector $\mathbf{d}_k$,

$$\tilde{\mathbf{r}}_k = [\mathbf{r}_k \,\|\, \mathbf{d}_k]\,. \tag{4.6}$$

The resulting vector was then passed to the fragment-level prediction heads.

To stabilize training, all non-binary and non-one-hot features were normalized using statistics from the training set. For a feature value $x$, normalization was performed as

$$\tilde{x} = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}}}, \tag{4.7}$$

where $\mu_{\text{train}}$ and $\sigma_{\text{train}}$ are the mean and standard deviation of that feature in the training set.

Finally, the model was modified by adding a classification head in parallel with the regression head. This version is hereafter referred to as the multitask version of the model, or SÄLEN multitask, to distinguish it from the purely regression-based version.

For the regression head, each fragment $F_k$ predicts a scalar contribution $c_k^{\text{reg}}$. The final regression prediction is obtained by summing the fragment contributions,

$$\hat{y}_{\text{reg}} = \sum_{k=1}^{K} c_k^{\text{reg}}, \tag{4.8}$$

where $K$ is the number of fragments in the molecule.

For the classification head, each fragment instead predicts a logit contribution $c_k^{\text{cls}}$. These contributions are summed to produce the molecule-level binding logit,

$$z_{\text{bind}} = \sum_{k=1}^{K} c_k^{\text{cls}}. \tag{4.9}$$

The predicted probability of binding is then obtained using the sigmoid function,

$$\hat{p}_{\text{bind}} = \sigma(z_{\text{bind}}) = \frac{1}{1 + e^{-z_{\text{bind}}}}. \tag{4.10}$$

The classification head was trained using binary cross-entropy loss,

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i^{\mathrm{cls}}\log(\hat{p}_i) + \left(1-y_i^{\mathrm{cls}}\right)\log\left(1-\hat{p}_i\right)\right], \tag{4.11}$$

where $y_i^{\mathrm{cls}} \in \{0,1\}$ is the binary binding label and $\hat{p}_i$ is the predicted binding probability.

## 4.2 Molecular Data Input & Fragmentation

The target is only one part of the data; the corresponding input representation is also needed. In all models, the molecular input was given as a SMILES string. From this string, RDKit was used to construct a molecular graph, where atoms were represented as nodes and bonds as edges. The associated metal ion was also recorded and used as an additional input.

For SEAL and the derived SÄLEN model, each molecule is split into fragments before prediction. The original SEAL model uses an expanded BRICS-based fragmentation algorithm. In this project, that approach was found to sometimes produce fragments with limited chemical relevance for metal coordination, including both very small fragments and large fragments containing several distinct donor environments. Therefore, a separate rule-based fragmentation algorithm was developed for SÄLEN. The algorithm was inspired by the motivation behind BRICS, namely to split molecules into chemically meaningful substructures, but it was implemented as a custom donor-protected fragmentation procedure rather than as a direct modification of the BRICS output.

The purpose of the algorithm was to partition molecules into fragments that are chemically meaningful for the metal-adduct task while preserving likely coordination motifs. The rules were chosen from general coordination-chemistry reasoning, inspection of failure cases from the original SEAL/BRICS fragmentation, and practical requirements for fragment-wise interpretability. Metal coordination is strongly influenced by donor atoms, local charge distribution, denticity, chelate-ring formation, and hard–soft acid–base compatibility [3, 4, 5]. Motifs such as carboxylates, phenolates/catecholates, hydroxamates, and phosphonate-like groups are common coordination-relevant substructures and were therefore protected from fragmentation where possible [6].

The algorithm operates on the RDKit molecular graph. Candidate bonds for breaking are first generated from the bond list, beginning with non-ring single bonds. The surrounding local structure is then inspected using manually defined protected motifs and heuristic donor-atom rules. Thus, the algorithm does not first identify fixed building blocks and then apply rules to them; instead, the final fragments are the connected components that remain after candidate cuts have been accepted or vetoed.

The proposed cuts are vetoed based on several criteria: whether the bond is part of or adjacent to a protected coordination motif, whether the cut would separate a likely donor atom from its local chemical environment, whether it would fragment a ring system, and whether it would create very small fragments. In a second round of cuts, fragments already within the desired size range are skipped, while some veto

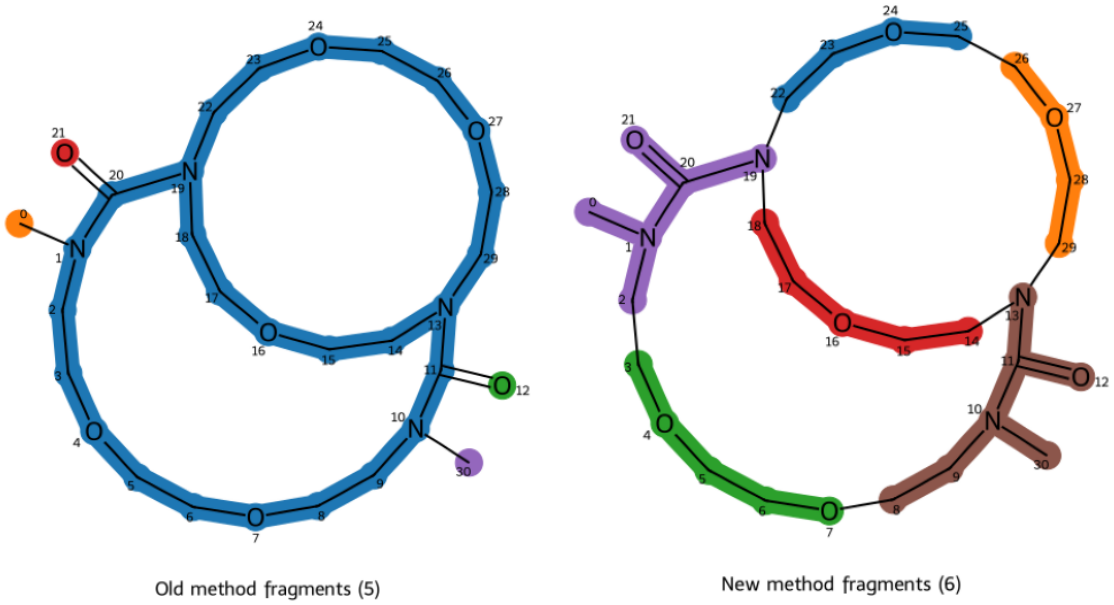



**Figure 4.1:** Comparison between the original BRICS-based fragmentation used in SEAL and the chemically informed fragmentation developed for SÄLEN. The new algorithm better preserves coordination-relevant motifs and avoids many chemically uninformative fragments.

requirements are relaxed for oversized fragments. Finally, adjacent small fragments are merged if the resulting fragment would not become too large.
The main rule categories are summarized in Table 4.1.
The result is a fragmentation algorithm better suited to the metal-adduct task. A comparison between the BRICS-based algorithm used in SEAL and the chemically informed fragmentation developed for SÄLEN is shown in Figure 4.1. Further examples are shown in Appendix A.3. The new algorithm generally produces larger and more chemically coherent fragments, although some artifacts remain. In some cases, fragments can still become too large for clear interpretation or too small to be chemically meaningful.

## 4.3 SÄLEN Training

The data were split by SMILES using an 8:1:1 train, validation, and test ratio. The multitask dataset contained 73589 training samples, 9177 validation samples, and 9242 test samples. Each sample consisted of a SMILES string, an ion, a regression target, and a binary adduct-detection label. Of these samples, 20000 were synthetic negatives, distributed according to the same split. Synthetic negatives were used only for the classification loss and were masked out of the regression loss.
Although predicting whether a metal-adduct-like signal is detected for a given molecule–ion pair is the main classification goal, it is a difficult first modelling task. The signal is noisy, as described in Section 3.2.2, and the log-transformed

**Table 4.1:** Main rule categories used in the donor-protected fragmentation algorithm.

| **Rule category** | **Examples** | **Purpose** |
|---|---|---|
| Protected coordination motifs | Carboxylate/carboxylic acid, hydroxamate-like groups, phosphate-like groups, $\beta$-dicarbonyl-like motifs, picolinate-like motifs | Preserve groups that may act as metal-binding or chelating motifs. |
| Ring protection | Ring bonds, bonds within the same ring system, donor-containing ring systems | Avoid fragmenting chemically coherent cyclic structures and heteroaromatic donor environments. |
| Donor protection | Acidic or anionic oxygens, pyridine-like nitrogens, neutral non-amide nitrogens, reduced sulfurs, anionic phosphorus atoms | Prevent likely coordination atoms from being separated from their local chemical environment. |
| Cut veto rules | Conjugated donor/carbonyl contexts, ring-attached exocyclic donors, acidic carbonyl contexts, neutral linker heteroatoms | Reject cuts that would split chemically coupled donor or carbonyl environments. |
| Size control | Minimum fragment size, preferred minimum fragment size, maximum side size, maximum clique diameter | Avoid single-atom or overly small fragments while preventing fragments from becoming too large for interpretation. |
| Macrocycle | Optional paired cuts in large non-aromatic ring systems | Allow oversized macrocyclic fragments to be split when ordinary non-ring cuts are insufficient. |
| Cleanup | Reattachment of isolated neutral linker heteroatoms, merging of small non-donor fragments | Reduce chemically uninformative fragments and improve interpretability of the final decomposition. |

regression target provides a smoother training signal than the binary classification label. The final model was therefore first pre-trained on the measured regression data using only the regression head. The best regression checkpoint, selected by validation MAE, was then used as the starting point for multitask fine-tuning. In this stage, the classification head was added and the synthetic negatives described in Section 3.3 were included.
For a batch of $N$ samples, the multitask loss was defined as

$$\mathcal{L} = w_{\text{reg}}\mathcal{L}_{\text{reg}} + w_{\text{cls}}\mathcal{L}_{\text{cls}}, \tag{4.12}$$

where $w_{\text{reg}}$ and $w_{\text{cls}}$ control the relative importance of the regression and classification tasks. The regression loss was computed only for samples with measured positive intensities:

$$\mathcal{L}_{\text{reg}} = \frac{1}{\sum_{i=1}^{N} m_i} \sum_{i=1}^{N} m_i \,\text{SmoothL1}\left(\hat{y}_i, \tilde{y}_i\right), \tag{4.13}$$

where $\hat{y}_i$ is the predicted normalized log-intensity, $\tilde{y}_i$ is the normalized regression target, and $m_i = 1$ for measured positive samples and $m_i = 0$ for synthetic negatives. The classification loss was a weighted binary cross-entropy loss,

$$\mathcal{L}_{\text{cls}} = \frac{1}{\sum_{i=1}^{N} \alpha_i} \sum_{i=1}^{N} \alpha_i \,\text{BCE}\left(\hat{p}_i, y_i^{\text{cls}}\right), \tag{4.14}$$

where $\hat{p}_i$ is the predicted probability of a detected metal-adduct-like signal, $y_i^{\text{cls}} \in \{0, 1\}$ is the binary classification label, and $\alpha_i$ is a confidence weight used to account for differences in label reliability. Synthetic negatives contributed to this classification loss but not to the regression loss.
For the first 5 epochs of fine-tuning, only the classification head was trained. After this warm-up stage, both the regression and classification heads were trained jointly according to Equation 4.12, with the regression term applied only to measured positive samples.

### 4.3.1 Hyperparameter Optimization

The regression pre-training model was optimized separately from the fine-tuned multitask model. This was faster, and the best starting point for multitask fine-tuning was consistently the best regression model as evaluated by MAE.
Hyperparameter sweeps were performed using Weights & Biases Sweeps with the Bayesian search strategy. In contrast to grid or random search, Bayesian search uses a probabilistic surrogate model to propose new hyperparameter values based on the results of previous trials. This makes the search more informed, although it is mainly suited to smaller numbers of continuous parameters [26, 27]. Validation MAE was used as the optimization target, and a fixed random seed was used during the sweeps.
A smaller and faster Bayesian sweep was then performed for the multitask model. The results are shown in Section 4.6, while the final hyperparameters are listed in Appendix A.4.

### 4.3.2 Model Evaluation

The regression models were evaluated using mean absolute error (MAE), root mean squared error (RMSE), and $R^2$. MAE was used as the primary metric during model comparison, since it is directly interpretable in log-intensity space and is less sensitive to large outliers than RMSE. RMSE was included to show whether individual large errors remained, while $R^2$ was used to measure how much of the variance in the target was explained by the model.

For the classification task, evaluation was less straightforward because the labels were imbalanced and derived from non-detection rules rather than direct ground truth. The most important trade-off was therefore between precision and recall. In the intended use case, false positive binding predictions are more problematic than false negatives, since the model should preferably suggest molecule–ion pairs that are likely to be real positives. Some pessimism was therefore considered acceptable if it increased the reliability of positive predictions.

Classification metrics were calculated using a probability threshold of 0.5. This threshold gave a reasonable compromise between precision and recall on the validation set. Since the negative labels were generated from incomplete experimental evidence, overly permissive negative selection introduced label noise and degraded performance. This also increased class imbalance, which was handled during training using weighted classification loss.

## 4.4 Chemprop

To provide a basis for comparison, previously available neural-network models were also tested. One of these was Chemprop. In addition to the molecular features generated by Chemprop, the same RDKit descriptors selected through XGBoost were added as molecular features. The ion features listed in Table A.3 were concatenated to the molecular feature vector according to the ion associated with each sample.

This gave the GNN an opportunity to learn molecular structure from the graph, while the final MLP could also use additional molecular descriptors and ion-specific information.

The network was trained using mean squared error (MSE) as the loss function. Bayesian hyperparameter optimization was performed over learning rate, number of MPNN layers, number of MLP layers, dropout, and hidden dimension of the MLP. The final values are listed in Table A.9.

## 4.5 AttentiveFP

AttentiveFP was also tested, since it was considered well suited for the long-range interactions that can occur during chelation between a ligand and a metal ion. Ion features and molecular features were provided to the model in much the same way as for Chemprop. A hyperparameter sweep was performed to give the model a fair comparison with the others. The final setup is shown in Table A.10.

**Table 4.2:** Regression performance on the test set, reported as mean ± standard deviation across 5 random seeds. Model parameters are approximate. Parameter counts are reported for neural-network models only; XGBoost is a tree ensemble and is therefore not directly comparable by neural-network parameter count.

| Model | Parameters | MAE | $R^2$ | RMSE |
|---|---|---|---|---|
| SÄLEN | 900K | $0.8085 \pm 0.0029$ | $0.7562 \pm 0.0031$ | $1.1172 \pm 0.0070$ |
| Chemprop | 1M | $0.810 \pm 0.003$ | $0.748 \pm 0.004$ | $1.125 \pm 0.008$ |
| AttentiveFP | 12M | $0.794 \pm 0.003$ | $0.759 \pm 0.003$ | $1.111 \pm 0.007$ |
| XGBoost | – | $0.864 \pm 0.001$ | $0.737 \pm 0.0004$ | $1.1602 \pm 0.0009$ |

**Table 4.3:** Classification performance of the final multitask SÄLEN model on the test set.

| **Metric** | **Test value** |
|---|---|
| Precision | 0.9198 |
| Recall | 0.6989 |
| True positives | 5078 |
| True negatives | 1533 |
| False positives | 443 |
| False negatives | 2188 |

## 4.6 Model Performance

Predictive performance is essential for the usefulness of any model. In this project, the goal was not only high performance, but also to retain explainability. The mean scores and standard deviations across 5 seeds are shown in Table 4.2.

Table 4.2 shows similar but distinct performance for all neural-network models, with AttentiveFP performing best and SÄLEN close behind. This is reasonable, since AttentiveFP receives a more complete global representation of the molecule before making its prediction, which may better allow distant atoms to interact. Nevertheless, the difference is small, especially considering that AttentiveFP had close to 12M parameters compared with approximately 900K for SÄLEN. Chemprop performed slightly worse, possibly due to its graph-level pooling before prediction, while SÄLEN pools and weighs fragment-level information. XGBoost performed worst, but still achieved reasonable performance using only molecular descriptors.

Table 4.3 shows the classification performance of the multitask SÄLEN model on the test set. The model reached a precision of 92% and a recall of 70% with constructed negative labels. This means that 92% of the molecule–ion pairs predicted as binding were true positives, while 70% of all true binding pairs were detected by the model. This trade-off was considered suitable for the intended use case, where high confidence in predicted binders was prioritized over recovering every possible positive.

Figure 4.2 shows the performance of SÄLEN over different fractions of the training set, averaged over 4 seeds. The error bars show one standard deviation.

As shown in Figure 4.2, model performance improves markedly with more training

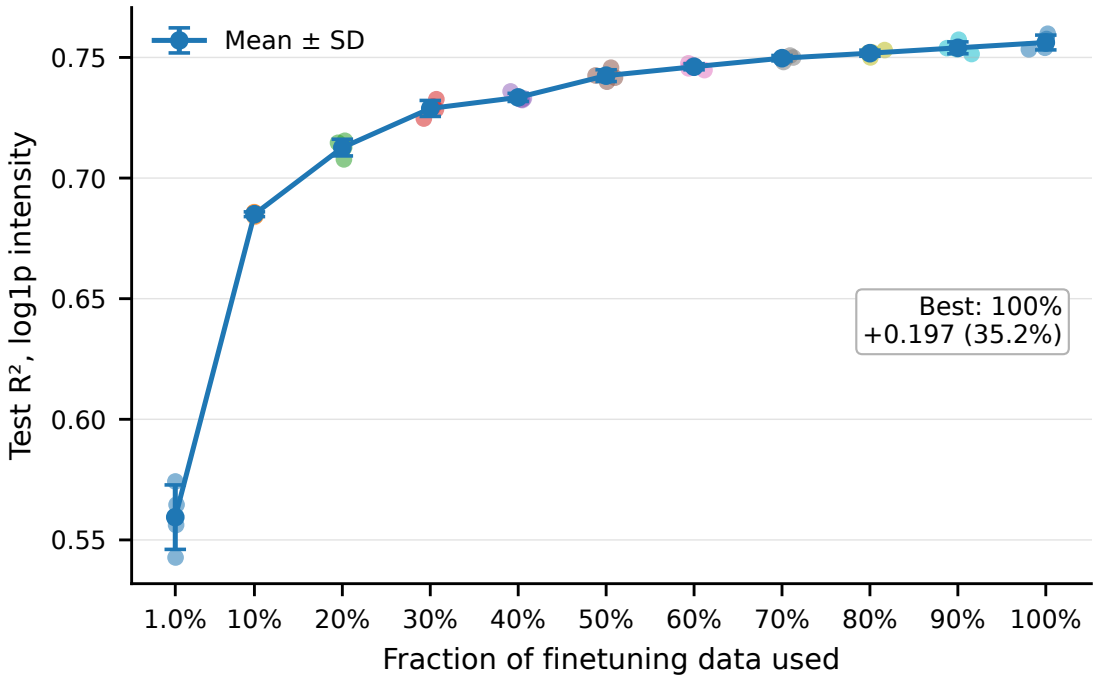



**Figure 4.2:** SÄLEN regression performance as a function of the fraction of the training set used. Results are averaged across 4 random seeds, and error bars show one standard deviation. The best performance was achieved with the full dataset, giving a 35.2% relative increase in test $R^2$ compared with the lowest training-data fraction.

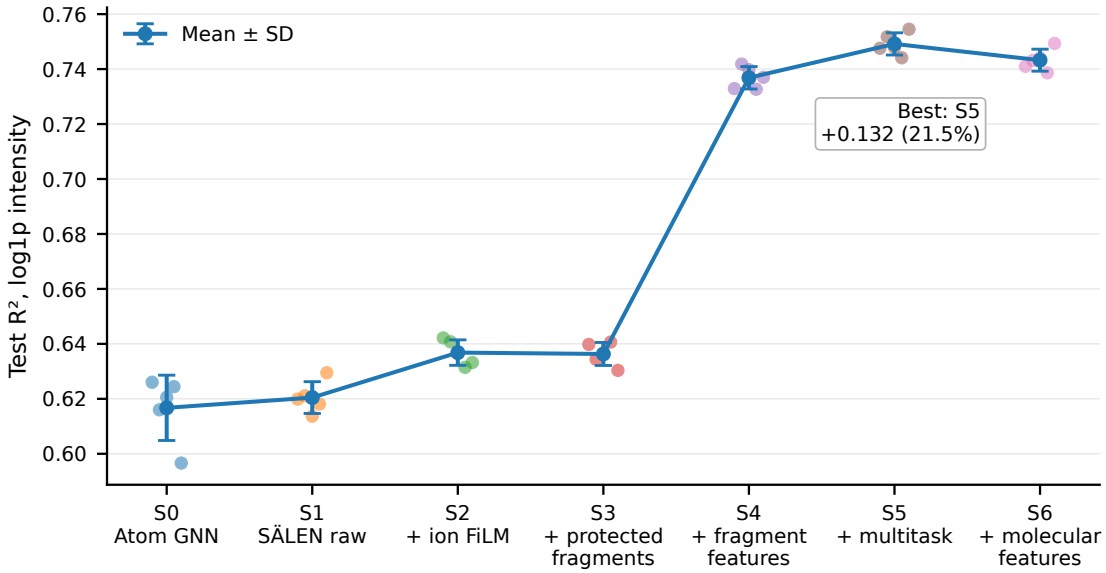



**Figure 4.3:** Regression performance of SÄLEN on the test set after consecutive additions to the architecture. Results are averaged across 5 random seeds, and error bars show one standard deviation. The largest gain came from adding fragment-level features.

data, and a clear performance plateau does not appear to have been reached.
Much of the thesis work focused on making SÄLEN competitive while maintaining intrinsic explainability. The performance of the model through different stages of development is shown in Figure 4.3.
The stages in Figure 4.3 are discussed here. During all stages, ion features were concatenated to each node to make the comparison fair and allow the model to differentiate between ions. All models also used the same hyperparameter settings and atom features. S0, the atom GNN, did not use the SEAL architecture and was included mainly as a comparison. S1, SÄLEN raw, was essentially the SEAL model before any modifications. With the addition of the ion FiLM submodule in S2, the model better retained ion information across layers, slightly improving performance. The chemically informed fragmentation algorithm did not by itself improve performance much, but it enabled the addition of fragment features in S4, which gave the largest performance gain. The multitask fine-tuning in S5 slightly improved performance, likely because the model was exposed to more molecules, although not all were used for regression. The final stage, S6, showed that adding molecule descriptors hurt performance, possibly due to over-saturation of repeated information, since the fragment features together with the learned embeddings already provided a strong signal.

# 5
# Explainability

In this chapter, we examine whether the intrinsic fragment-wise explanations produced by SÄLEN are chemically plausible. This is done by comparing them with model-independent counterfactual explanations, inter-model agreement, and DFT-derived QM descriptors. The methods used to compare the models to each other and to simulated features can be found below.

## 5.1 Methods of Explanation

The explanation methods were divided into intrinsic and model-independent methods. The intrinsic explanation was specific to SÄLEN, since the model produces fragment-wise contribution scores that are summed to form the final prediction. These scores were used to visualize which fragments supported or opposed the predicted metal-adduct signal.

Two model-independent counterfactual methods were also used. First, atom counterfactuals were generated by individually replacing oxygen and nitrogen atoms with carbon. This was done for every oxygen and nitrogen atom in every molecule in the test set. A counterfactual was kept only if RDKit accepted the edited molecule after sanitization. Since oxygen and nitrogen are common donor atoms in metal-ligand interactions, this tested how sensitive each model was to local donor-like atoms.

Second, fragment counterfactuals were generated from the SÄLEN fragmentation. Only test-set molecules were modified. For each molecule, SÄLEN fragments with a single attachment point were considered for replacement. Replacement fragments were taken from the training-set fragment library and were also required to have a compatible single attachment point. A proposed swap was kept only if the resulting molecule was accepted by RDKit. For SÄLEN, the expected effect of a swap could be estimated from the difference between the original and replacement fragment contribution scores. The edited molecules were then evaluated with SÄLEN, Chemprop, and AttentiveFP to compare how the models responded to the same structural change.

For each atom replacement or fragment swap, the prediction change was defined as

$$\Delta\hat{y} = \hat{y}_{\text{original}} - \hat{y}_{\text{edited}}. \tag{5.1}$$

A positive value therefore means that the original atom or fragment increased the predicted signal relative to the edited molecule, while a negative value means that the edited molecule received a higher prediction.

Model agreement was measured in three ways. Strong sign agreement measured whether two models agreed on the direction of a perturbation effect, considering only examples where the absolute prediction change exceeded 0.25 log1p-intensity units. Spearman rank correlation measured whether the models ranked the perturbations similarly. Pearson magnitude correlation was computed on the absolute prediction changes and measured whether the models agreed on effect size. Together, these metrics compare agreement in direction, ranking, and magnitude.

These counterfactual explanations should be interpreted as model-sensitivity tests rather than direct chemical validation, since atom and fragment replacements may create unrealistic or out-of-distribution molecules.

DFT simulations were available for a subset of 740 molecules through the related thesis work of André Maio [28]. This allowed the intrinsic SÄLEN explanations to be compared with independent simulated molecular descriptors. These descriptors should not be considered a true ground truth, since the simulations were performed at 0 K in vacuum and did not include the metal ion interacting with the molecule. However, they can still be used as a plausibility check, either supporting or challenging the explanations produced by SÄLEN.

We reasoned that the most relevant QM features for metal binding would be related to charge distribution and polarizability. Therefore, three atom-level descriptors were selected: CHELPG partial charge, ESP electron contribution, and atomic polarizability. The CHELPG partial charge is an atom-wise charge fitted to reproduce the molecular electrostatic potential; more negative values indicate more electron-rich atoms, which are plausible interaction sites for positively charged metal ions. In the fragment-level statistics, the minimum CHELPG charge in each fragment was used, corresponding to the most negative atom in the fragment.

The ESP electron contribution describes the electron-density contribution to the electrostatic potential. More negative values indicate regions where the electron cloud contributes more strongly to attracting a positive probe. For the fragment-level comparison, the mean ESP electron contribution was used, as this gave a clearer fragment-level signal than the total ESP, which appeared to be dominated by a positive nuclear contribution.

Atomic polarizability describes how easily the electron cloud around an atom can be distorted. Higher polarizability may be relevant for metal binding, especially for larger or softer ions, since a nearby metal ion can induce polarization in the ligand. For this descriptor, the maximum atomic polarizability in each fragment was used. However, in the final comparisons, CHELPG charge and ESP electron contribution appeared to align more clearly with the SÄLEN binding attributions than polarizability.

## 5.2 Results of Explanations

Figure 5.1 shows two examples of intrinsic SÄLEN fragment-wise explanations: one true positive and one false negative. Red fragments increase the predicted binding probability, while blue fragments decrease it. The numbers shown on the fragments are binding-logit contributions, which are summed and passed through the sigmoid

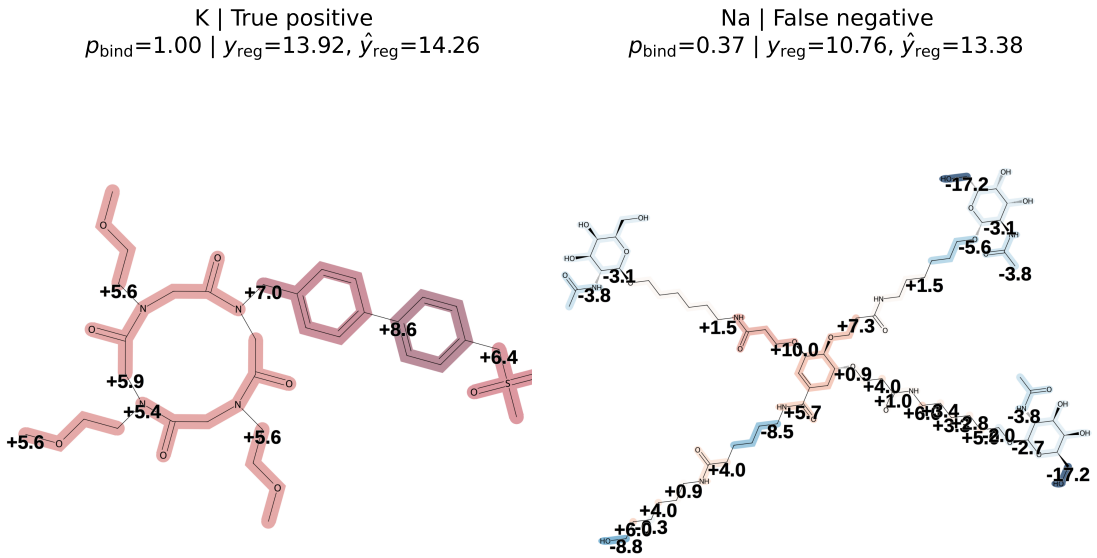



**Figure 5.1:** Examples of intrinsic SÄLEN fragment-wise explanations. Fragment colour indicates the binding-logit contribution, with red fragments increasing and blue fragments decreasing the predicted binding probability. The left example is a true positive and the right example is a false negative. The predicted binding probability, predicted log-intensity, target log-intensity, and ion are shown above each molecule.

function in Equation 4.10. The classification threshold was set to 0.5.

In the true positive example, the explanation does not point to a single isolated binding site. Instead, several fragments receive positive contributions. This may indicate that the model uses a distributed binding pattern, where several donor-containing regions or flexible arms contribute to the prediction. This is chemically plausible if multiple parts of the molecule can fold toward the ion, but it also shows a limitation of the explanation: when many fragments receive high scores, the attribution becomes less specific and does not clearly identify one binding motif.

The false negative example illustrates a different limitation. The molecule is large and partly over-fragmented, especially in the lower arms. This may make it difficult for SÄLEN to represent long-range interactions between distant parts of the molecule. The model may therefore miss cases where one region binds the ion while other parts of the molecule are less relevant. The regression prediction is nevertheless close to the target, suggesting that the model has captured part of the signal even though the classification is incorrect.

The next Figure 5.2 shows how SÄLENs intrinsic explanation changes for different ions. As the target is different for each ion it has hopefully learned what makes a fragment attractive or repelling to each metal ion.

Figure 5.2 shows how the SÄLEN explanation changes for the same molecule under different ion conditions. The model correctly classifies all four cases, but the predicted probabilities are relatively close to the classification threshold for Li, Na, and K. The fragment scores also change with ion identity, indicating that the model uses the ion features when assigning fragment contributions.

For Li, the explanation highlights the right-hand region around the oxygen-containing groove. This is chemically plausible, since Li is small and charge-dense and may fit into a compact donor environment. For Na, the same region is still relevant, but the prediction is weaker, which may reflect the larger ion size and a less favourable local geometry. For K and Cs, the predicted binding probabilities are lower, consistent with the larger ions being less able to access the same compact site. In these cases, the model assigns weaker or more distributed contributions, suggesting that no single fragment provides a strong binding-like signal.

These examples show both the usefulness and the limitation of the fragment-wise explanation. The model can reveal ion-dependent changes in which molecular regions support the prediction, but the explanation should not be interpreted as a precise binding geometry. Especially for flexible or larger molecules, binding may depend on long-range conformational effects that are only indirectly represented by the fragment-level model.

### 5.2.1 Inter-Model Comparison

To compare whether different model architectures relied on similar molecular regions, the same atom and fragment counterfactuals were evaluated with SÄLEN, Chemprop, and AttentiveFP. For the atom counterfactuals, oxygen and nitrogen atoms were individually replaced with carbon, and the prediction change was defined as

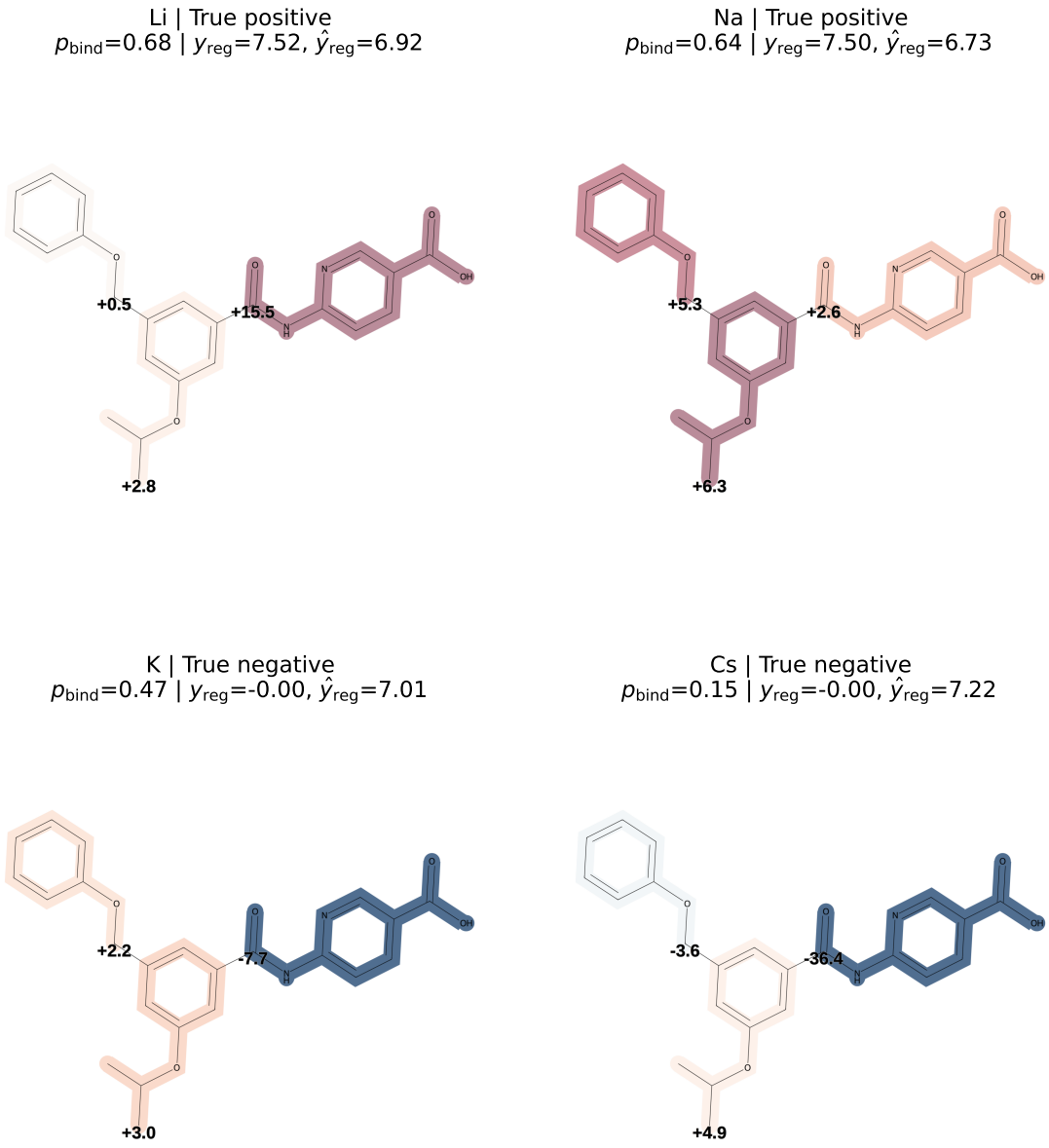



**Figure 5.2:** Ion-conditioned SÄLEN explanations for the same molecule under four metal-ion conditions. Labels show the classification outcome and predicted/target log-intensity.

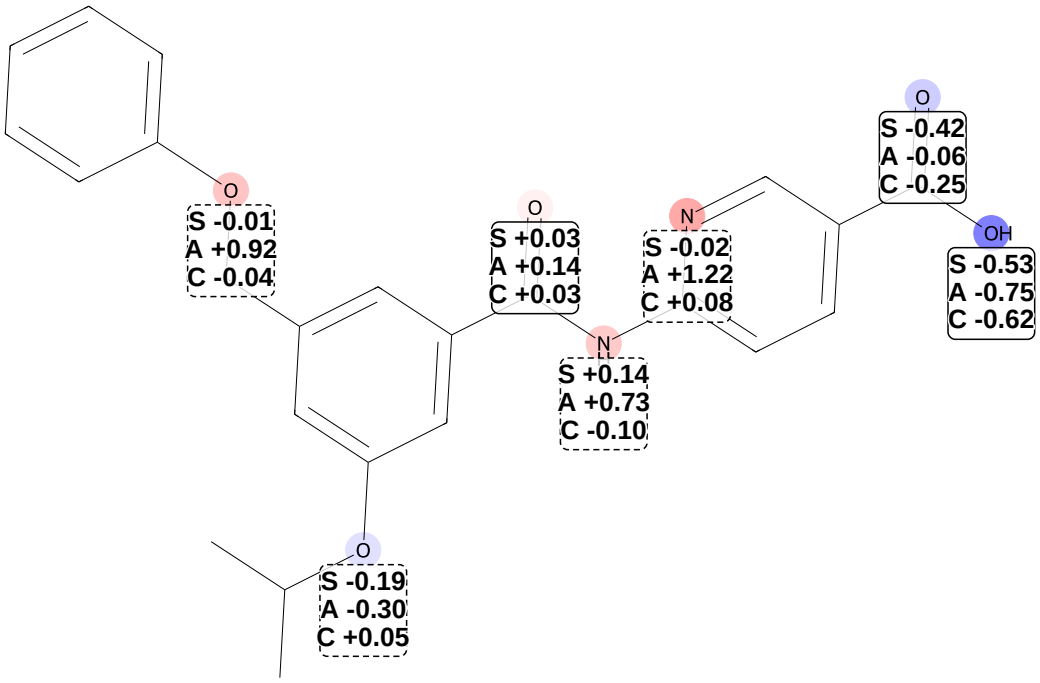



**Figure 5.3:** Inter-model comparison for atom counterfactuals for a molecule with Li ion. Oxygen and nitrogen atoms were individually replaced with carbon, and the displayed values show $\Delta\hat{y} = \hat{y}_{\text{original}} - \hat{y}_{\text{edited}}$. Positive values indicate that the original atom increased the predicted signal relative to carbon, while negative values indicate that the prediction increased after replacement. S, C, and A denote SÄLEN, Chemprop, and AttentiveFP.

$$\Delta\hat{y} = \hat{y}_{\text{original}} - \hat{y}_{\text{edited}}. \tag{5.2}$$

A positive value means that the original atom increased the predicted signal relative to carbon, while a negative value means that replacing the atom with carbon increased the prediction. The magnitude, $|\Delta\hat{y}|$, describes the strength of the effect. In total, atom counterfactuals were evaluated for 7266 molecules and 52185 atom edits. Figure 5.3 shows one example for a molecule evaluated with lithium. The target log1p intensity for this molecule was 7.52. The original predictions were 8.49 for AttentiveFP, 6.99 for Chemprop, and 6.92 for SÄLEN. Thus, all three models predicted a signal in the same approximate range, although AttentiveFP overestimated the intensity while Chemprop and SÄLEN underestimated it.
In this example, the models mainly agree for the two oxygens on the far right, where all three assign negative scores. This means that replacing these oxygens with carbon increased the predicted signal in all three models. One possible interpretation is that this carboxyl-containing region can stabilize negative charge and may attract the lithium ion away from other regions that are more favourable for the measured signal [29]. However, this remains a model-based hypothesis rather than direct evidence of the physical lithium-binding geometry.

**Table 5.1:** Test-set-wide model agreement over O/N→C atom edits. Agreement was calculated over 52185 atom edits from 7266 molecules. Spearman denotes Spearman rank correlation, Pearson denotes Pearson magnitude correlation, and strong sign denotes sign agreement for prediction changes larger than 0.25 log1p-intensity units.

| Models | Spearman | Pearson | Strong sign |
|---|---|---|---|
| SÄLEN–Chemprop | 0.57 | 0.58 | 76% |
| AttentiveFP–Chemprop | 0.58 | 0.59 | 78% |
| SÄLEN–AttentiveFP | 0.53 | 0.54 | 73% |

**Table 5.2:** Test set wide model agreement over fragment swaps. Agreement was calculated over 59638 swaps from 7266 molecules.

| Models compared | Spearman correlation | Strong sign agreement |
|---|---|---|
| SÄLEN – Chemprop | 0.67 | 78% |
| AttentiveFP – Chemprop | 0.70 | 80% |
| SÄLEN – AttentiveFP | 0.61 | 74% |

The atom-level effects are generally small, and the dataset-wide statistics in Table 5.1 show only moderate agreement between models. Strong sign agreement ranges from 73–78%, using a threshold of 0.25 log1p-intensity units, while the Spearman and Pearson correlations are around 0.5–0.6. Taken together with the intrinsic SÄLEN explanation in Figure 5.2, the cross-model comparison suggests that atoms near the central carbonyl-containing region are relevant to the predicted lithium-associated signal, but does not identify a unique binding geometry.

The fragment swaps showed stronger agreement between models than the atom edits. In total 59638 fragment swaps were evaluated across 7266 molecules. The higher correlations in Table 5.2 suggest that larger substructural changes produce more consistent model responses than individual atom replacements. This supports the use of fragment-level explanations, although part of the stronger agreement may also come from fragment swaps causing larger prediction changes.

Figure 5.4 shows one fragment-swap example. In this case, a fragment that SÄLEN considered supportive of lithium binding was replaced by a fragment with a lower predicted contribution. The predicted binding probability dropped below the classification threshold after the swap, consistent with the expected direction of the counterfactual. However, the edited molecule also illustrates a limitation of the method: after the swap, the fragmentation algorithm produced a large fragment that was not split adequately. This shows that fragment counterfactuals evaluate both the model response and the stability of the fragmentation procedure.

### 5.2.2 SÄLEN Compared to DFT-Derived QM Descriptors

The comparison between SÄLEN and DFT-derived quantum mechanical (QM) descriptors was limited in scope, and the results should therefore be interpreted as a plausibility check rather than a validation of the explanations. The DFT calcula-

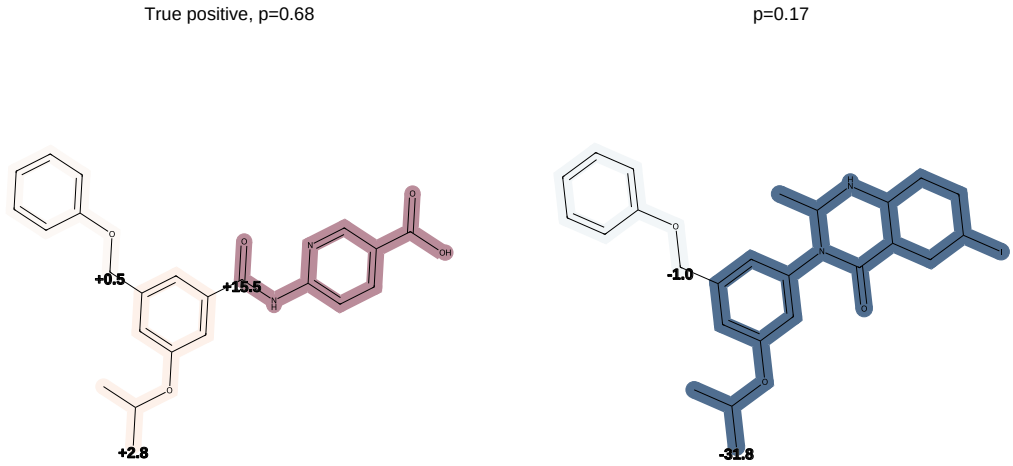



**Figure 5.4:** Example of a fragment counterfactual. The original molecule with intrinsic SÄLEN fragment contributions is shown on the left, and the molecule after replacing an important fragment is shown on the right. The swap decreases the predicted binding probability, but also exposes a fragmentation failure in the edited molecule, where a large fragment remains unsplit.

tions were performed on isolated molecules at 0 K in vacuum and did not include the interacting metal ion. They therefore do not represent the full experimental system measured by MS. However, the descriptors still provide independent information about charge distribution and electronic structure, which can support or challenge the intrinsic SÄLEN explanations.

The QM descriptors were available at atom level, while SÄLEN produces fragment-level binding contributions. To compare them, the atom-level QM descriptors were aggregated over each SÄLEN fragment. Three descriptors were selected. For CHELPG partial charge, the minimum value in each fragment was used, corresponding to the most negative atom in that fragment. For ESP electron contribution, the mean value over the fragment was used. For atomic polarizability, the maximum value in the fragment was used. These were compared with the SÄLEN fragment contribution scores for binding.

Figure 5.5 shows an example where the SÄLEN explanation is supported by the QM descriptors. The highly contributing fragment highlighted by SÄLEN contains atoms with descriptor values consistent with interaction with a metal cation. In particular, the ESP electron contribution indicates a region where the electron density contributes strongly to attracting a positive probe. This agrees with SÄLEN assigning a positive binding contribution to that region. The polarizability map also highlights parts of the same region, although polarizability was generally a weaker explanatory signal than charge and ESP electron contribution in the broader comparison.

Not all molecules showed clear agreement between SÄLEN and the QM descriptors.

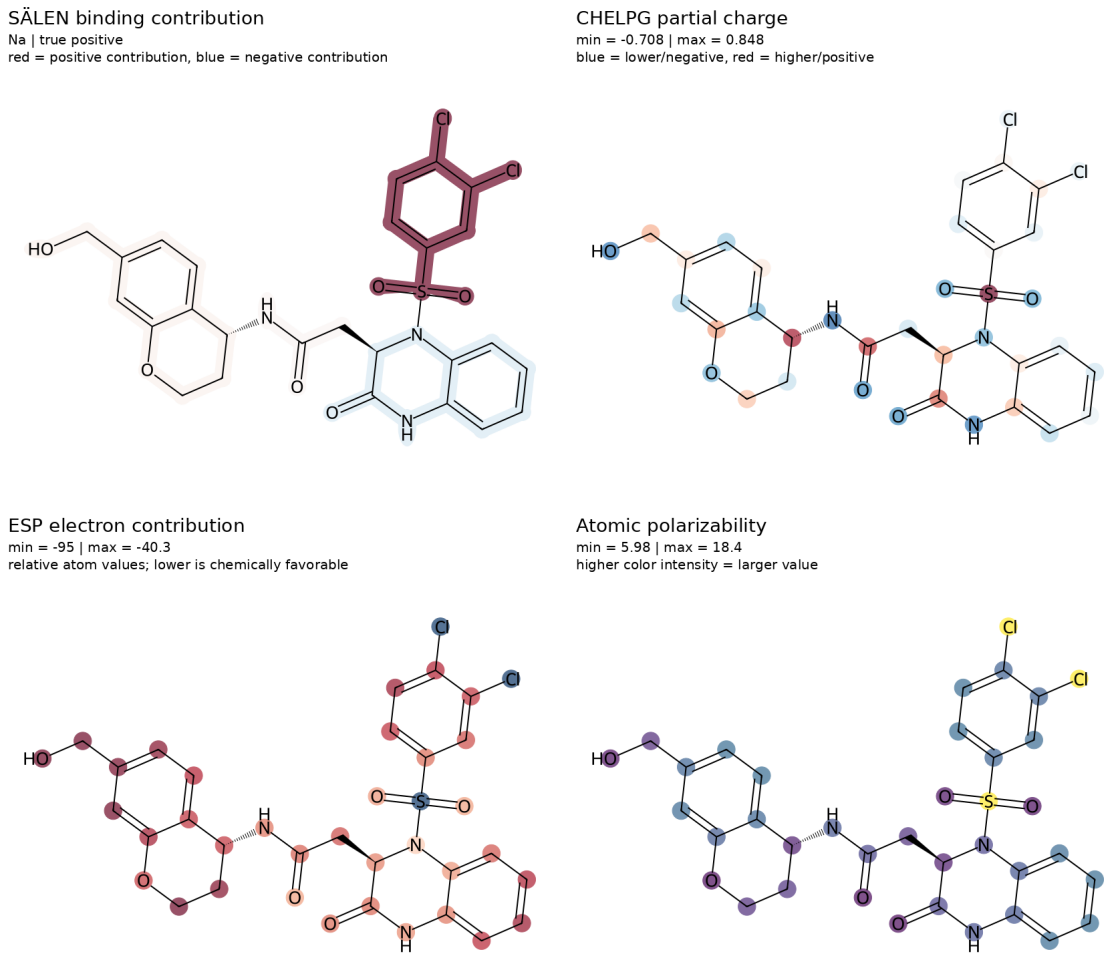



**Figure 5.5:** Comparison between atom-level QM descriptors and SÄLEN binding attribution for one molecule. The maps show CHELPG partial charge, ESP electron contribution, atomic polarizability, and the SÄLEN fragment-wise binding contribution.

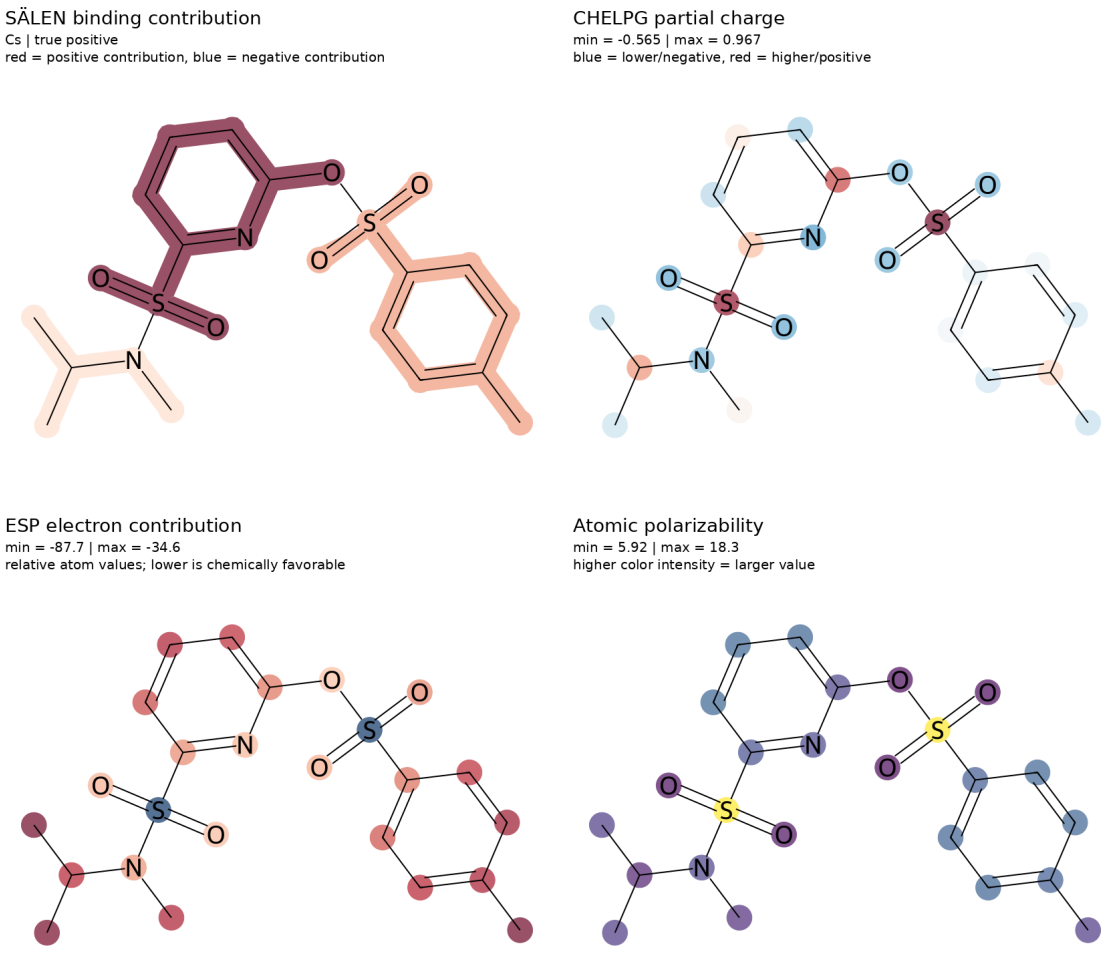



**Figure 5.6:** Example where the correspondence between QM descriptors and SÄLEN binding attribution is less clear. The maps show CHELPG partial charge, ESP electron contribution, atomic polarizability, and the SÄLEN fragment-wise binding contribution.

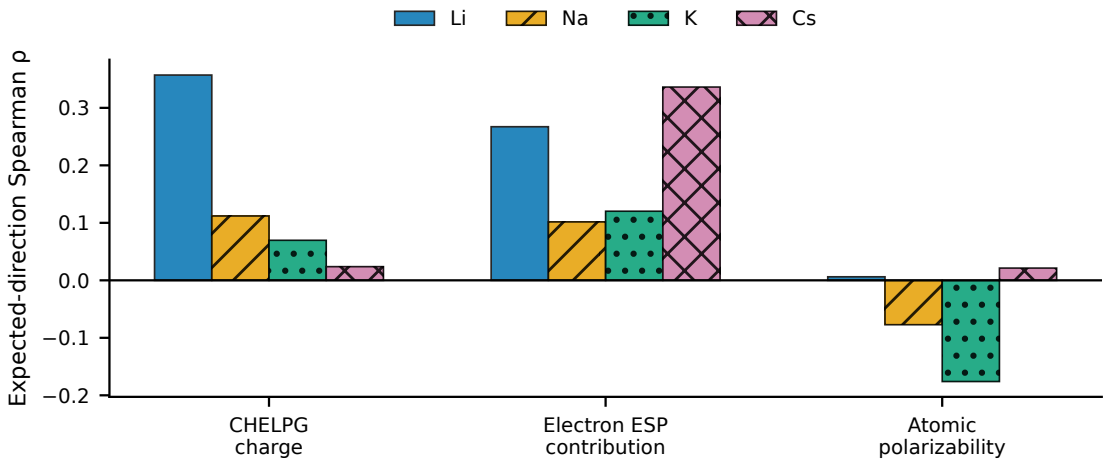



**Figure 5.7:** Spearman correlation between SÄLEN fragment binding contributions and fragment-aggregated QM descriptors for true binders. Descriptor directions were transformed so that positive values indicate agreement with the expected binding direction.

Figure 5.6 shows an example where the fragment ranking assigned by SÄLEN is harder to support from the selected QM features. This does not necessarily mean that the SÄLEN explanation is wrong. The QM descriptors describe the isolated molecule, while the experimental target depends on the molecule, the metal ion, the solvent conditions, ionization, and detection in the MS pipeline. However, such cases highlight that the explanation quality depends both on the model and on whether the fragmentation captures chemically meaningful regions. Figure 5.7 summarizes the dataset-level comparison between SÄLEN binding contributions and the selected QM descriptors. The descriptor directions were transformed so that a positive Spearman correlation means agreement with the expected binding direction. For example, more negative CHELPG charge and more negative ESP electron contribution were treated as more supportive of metal binding. With this convention, CHELPG charge and ESP electron contribution showed positive correlations for all ions, although the strength varied. This supports the interpretation that SÄLEN often assigns higher binding contributions to fragments with more favourable electrostatic properties. The agreement with polarizability was weaker and generally less consistent. This suggests that, for the ions and molecules studied here, localized charge and electron-density descriptors aligned more clearly with SÄLEN binding attributions than polarizability, though with weak Spearman correlation of 0.3. The decrease in charge-based agreement from lithium toward larger alkali ions is also chemically plausible, since larger ions have lower charge density and may depend less strongly on localized negative charge. This trend should nevertheless be interpreted cautiously, as the DFT descriptors do not include the metal ion or the experimental environment.

# 6
# Conclusion

SÄLEN was developed as an ion-conditioned, fragment-wise GNN for predicting MS-derived metal-adduct-like signals. The main goal was not only predictive performance, but to make the model's predictions interpretable at the level of chemically meaningful fragments.

## 6.1 Main Findings

SÄLEN achieved regression performance close to strong molecular GNN baselines, although AttentiveFP performed slightly better. This suggests that intrinsic fragment-wise interpretability could be introduced without a large loss in predictive accuracy. The stronger performance of AttentiveFP may partly reflect its ability to relate graphically distant parts of a molecule, which can be important for chelation and other effects involving separated donor groups.
The multitask version of SÄLEN also showed useful classification performance. Its high precision makes it suitable for identifying molecule–ion pairs worth further inspection, while its lower recall means that negative predictions should be treated cautiously. In practice, SÄLEN is therefore better suited for prioritization than for ruling out possible metal-adduct formation.
The main value of SÄLEN is the combination of competitive prediction and intrinsic fragment-wise explanations. Instead of applying a separate post-hoc explanation method, the final prediction is built from summed fragment contributions. This makes it possible to inspect both the predicted signal and the molecular regions that drive it.

## 6.2 Reliability and Plausibility of the Explanations

Fragment-level explanations are chemically useful because metal-adduct formation is often reasoned about through donor groups, acidic motifs, aromatic systems, and heteroatom-rich regions rather than isolated atoms. The counterfactual results support this level of analysis. Atom edits showed only moderate agreement between models, while fragment swaps gave stronger agreement between SÄLEN, Chemprop, and AttentiveFP, with Spearman correlations of 0.61-0.70 and strong sign agreement of 74-80%. This suggests that fragment perturbations capture more robust model sensitivities, although they also represent larger chemical changes.

The comparison with QM descriptors provides an additional plausibility check. CHELPG charge and ESP electron contribution aligned better with SÄLEN attributions than polarizability, suggesting that the model often emphasizes electrostatically favourable fragments. However, these descriptors were computed without the metal ion or the experimental MS environment. They therefore support chemical plausibility, but do not validate the true binding mechanism.
Overall, the explanations should be interpreted as model sensitivities. They indicate which fragments SÄLEN uses for its prediction, not necessarily which fragments physically coordinate the metal ion.

## 6.3 Boundaries of Interpretation

SÄLEN should not be interpreted as a direct model of thermodynamic metal binding. The target is a calibrated MS signal, which reflects metal-adduct-like detectability under the experimental and data-processing conditions used in this work.
A non-detected molecule–ion pair is not necessarily a true negative. It may be absent, but it may also be below the detection limit, poorly ionized, form an adduct that was not searched for, or be missed due to instrumental or processing limitations. The synthetic negatives were selected to reduce this problem, but they do not remove it. The classification labels should therefore be viewed as operational labels rather than independent chemical ground truth.
The explanations are also limited by both the fragmentation and the explanation procedure itself. If a molecule is split into chemically poor or overly large fragments, the attribution can only be interpreted at that same resolution. In addition, the counterfactual atom edits and fragment swaps mainly test local, first-order sensitivity: how the prediction changes when one atom or one fragment is modified. They may therefore miss higher-order effects, where the model responds to combinations of several fragments or features present at the same time.
Counterfactual edits can also move molecules outside the training distribution. In such cases, the model may be evaluated on structures unlike those seen during training, making the resulting prediction changes less reliable. The counterfactual explanations should therefore be interpreted as approximate model-sensitivity tests, not as complete or guaranteed-faithful descriptions of the model.
For these reasons, SÄLEN explanations identify which fragments the model appears to use for its prediction, not necessarily which fragments physically coordinate the metal ion or all interactions the model relies on.

## 6.4 Limitations

Several sources of error may have affected both the dataset and the model evaluation. Some metals were difficult to dissolve in DMSO, making it uncertain whether they remained available during the later stages of the chromatography gradient. This could reduce the chance of detecting complexes with more lipophilic molecules and introduce systematic bias.

The MetaboScape analysis only searched for predefined molecule–ion combinations. Unexpected adducts, fragments, or complexes may therefore have been missed, further limiting the interpretation of non-detections.
The dataset was updated several times as new measurements became available, and the train, validation, and test splits were regenerated rather than extended from a fixed split. This may have allowed model development decisions to be indirectly influenced by previous test sets, making the final performance estimate somewhat optimistic.
The compounds also originated from a chemical library where some samples had been stored for long periods of time. Degradation, precipitation, evaporation, or incorrect compound identity could mean that some intended compounds were not present during measurement. This is especially important for negative labels, where non-detection may reflect absence of the molecule rather than absence of metal-adduct formation.
Finally, MS intensity is not a direct measurement of binding affinity. Ionization efficiency, adduct stability, co-elution, ion suppression, and instrumental variation all influence the measured signal. Calibration reduces some technical variation, but relies on the assumption that the calibrants represent the behaviour of the broader dataset.

## 6.5 Future Work

Several possible improvements were identified during the project. Some measurements showed signs of spectrometer recalibration during a run. These were excluded, but future work could investigate whether parts of such runs can be cropped and retained. Repeated plates could also be combined to increase coverage, although this would require a principled method for merging repeated intensity values.
The larger improvement would likely come from more targeted data collection. More repeated measurements would improve estimates of detection reliability, support better negative labels, and make it possible to define a more realistic upper bound for classification performance. Independent negative controls or orthogonal assays would also help distinguish true absence of metal-adduct formation from non-detection in the MS workflow.
The fragmentation algorithm is another important direction. Since SÄLEN's explanations depend directly on the molecular split, future work should benchmark the donor-protected fragmentation more systematically against BRICS and other chemically informed schemes. Important goals would be to reduce overly large fragments, better handle macrocycles and donor-rich ring systems, and preserve coordination-relevant motifs.
The model could also be extended with features that improve chemical realism without removing fragment-level attribution. Possible directions include metal-aware QM descriptors, conformer or 3D features, uncertainty estimates, and controlled global molecular context. A hybrid architecture combining the predictive strengths of AttentiveFP with the interpretability of SÄLEN could also be explored.
More realistic quantum-chemical calculations would also strengthen the evaluation of the explanations. The DFT-derived descriptors used in this thesis were computed

for isolated molecules and did not include the metal ion or experimental environment. Future work could instead model molecule–metal complexes directly, include relevant conformers, and account for solvent or ionization effects where feasible. Such calculations would provide a more chemically faithful reference for testing whether highlighted fragments correspond to plausible coordination sites.
Finally, the explanations should be evaluated more directly. Chemist review, prospective testing of predicted important motifs, or experimental comparison of molecules before and after targeted fragment changes would help determine whether SÄLEN highlights chemically meaningful drivers of metal-adduct-like signal, rather than only model-relevant patterns. Together, these directions could make interpretable modelling of MS-derived metal-adduct signals both more accurate and more chemically actionable.

# Bibliography


[1] J.-L. Reymond, "The chemical space project," *Accounts of Chemical Research*, vol. 48, no. 3, pp. 722–730, 2015.

[2] M. Khalikova, J. Jireš, O. Horáček, M. Douša, R. Kučera, and L. Nováková, "What is the role of current mass spectrometry in pharmaceutical analysis?," *Mass Spectrometry Reviews*, vol. 43, no. 3, pp. 560–609, 2024.

[3] K. L. Haas and K. J. Franz, "Application of metal coordination chemistry to explore and manipulate cell biology," *Chemical Reviews*, vol. 109, no. 10, pp. 4921–4960, 2009.

[4] R. G. Pearson, "Hard and soft acids and bases," *Journal of the American Chemical Society*, vol. 85, no. 22, pp. 3533–3539, 1963.

[5] R. D. Hancock and A. E. Martell, "Ligand design for selective complexation of metal ions in aqueous solution," *Chemical Reviews*, vol. 89, no. 8, pp. 1875–1914, 1989.

[6] R. C. Hider and X. Kong, "Chemistry and biology of siderophores," *Natural Product Reports*, vol. 27, no. 5, pp. 637–657, 2010.

[7] D. Weininger, "SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[8] K. Burke, "Perspective on density functional theory," *The Journal of Chemical Physics*, vol. 136, no. 15, p. 150901, 2012.

[9] G. Landrum *et al.*, "RDKit: Open-source cheminformatics." `https://www.rdkit.org/`. Accessed: 2026-06-03.

[10] J. M. Hollas, *Modern Spectroscopy*. John Wiley & Sons, 4 ed., 2004.

[11] Bruker, "timsTOF mass spectrometry platform." `https://www.bruker.com/en/products-and-solutions/mass-spectrometry/timstof.html`. Accessed: 2026-06-03.

[12] SepSolve Analytical, "What is TOF-MS?." `https://www.sepsolve.com/techniques/what-is-tof-ms.aspx`. Accessed: 2026-06-03.

[13] Bruker, "MetaboScape: Software for metabolomics and lipidomics." `https://www.bruker.com/en/products-and-solutions/mass-spectrometry/ms-software/metaboscape.html`. Accessed: 2026-06-03.

[14] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.

[15] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proceedings of the 34th Inter-*

*national Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272, 2017.

[16] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[17] S. Musiał, B. Zieliński, and T. Danel, "Fragment-wise interpretability in graph neural networks via molecule decomposition and contribution analysis," 2025.

[18] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, and M. Rarey, "On the art of compiling and using 'Drug-Like' chemical fragment spaces," *ChemMedChem*, vol. 3, no. 10, pp. 1503–1507, 2008.

[19] E. Heid, K. P. Greenman, Y. Chung, S.-C. Li, D. E. Graff, F. H. Vermeire, H. Wu, W. H. Green, and C. J. McGill, "Chemprop: A machine learning package for chemical property prediction," *Journal of Chemical Information and Modeling*, vol. 64, no. 1, pp. 9–17, 2024.

[20] Chemprop Developers, "Chemprop atom featurizers documentation." `https://chemprop.readthedocs.io/en/main/tutorial/python/featurizers/atom_featurizers.html`. Accessed: 2026-06-03.

[21] Chemprop Developers, "Chemprop bond featurizers documentation." `https://chemprop.readthedocs.io/en/main/tutorial/python/featurizers/bond_featurizers.html`. Accessed: 2026-06-03.

[22] Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang, and M. Zheng, "Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism," *Journal of Medicinal Chemistry*, vol. 63, no. 16, pp. 8749–8760, 2020.

[23] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 3543–3556, 2019.

[24] S. Serrano and N. A. Smith, "Is attention interpretable?," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951, 2019.

[25] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[26] Weights & Biases, "Sweep configuration options." `https://docs.wandb.ai/models/sweeps/sweep-config-keys`. Accessed: 2026-06-03.

[27] SigOpt, "Bayesian optimization primer." `https://web.archive.org/web/20240209053347/https://static.sigopt.com/b/20a144d208ef255d3b981ce419667ec25d8412e2/static/pdf/SigOpt_Bayesian_Optimization_Primer.pdf`. Accessed: 2026-06-03.

[28] A. Maio, "Data-driven approaches to predict mass spectrometry response factors for high-throughput experimentation." Master's thesis, University of Strasbourg, to be published, 2026.

[29] Chemistry LibreTexts, "20.4: Substituent effects on acidity." `https://chem.libretexts.org/Courses/Athabasca_University/Chemistry_360%3A_Organic_Chemistry_II/Chapter_20%3A_Carboxylic_Acids_and_Nitriles/20.04_Substituent_Effects_on_Acidity`. Accessed: 2026-06-03.

# A

# Appendix 1

## A.1 Feature Sets Used by the Models

**Table A.1:** Atom-level features used as initial node features in the molecular graph.

| Feature | Encoding | Description |
|---|---|---|
| Element identity | One-hot vector | Encodes the atomic element. Elements outside the predefined set are assigned to an unknown category. |
| Atom degree | Scalar | Number of directly bonded neighboring atoms. Captures the local connectivity of the atom. |
| Aromaticity | Binary scalar | Indicates whether the atom is part of an aromatic system. |
| Formal charge | Scalar | Formal atomic charge assigned by RDKit. Provides information about charged donor or acceptor sites. |
| Number of bonded hydrogens | Scalar | Total number of hydrogens attached to the atom. Helps distinguish, for example, protonated and non-protonated heteroatoms. |
| Hybridization | One-hot vector | Encodes the RDKit hybridization state of the atom, with an additional unknown category. Gives a simple proxy for local bonding geometry. |

**Table A.2:** Bond-level features used as edge features in the molecular graph.

| **Feature** | **Encoding** | **Description** |
|---|---|---|
| Bond type | One-hot vector | Encodes whether the bond is single, double, triple, or aromatic. |
| Aromaticity | Binary scalar | Indicates whether the bond is aromatic according to RDKit. |
| Conjugation | Binary scalar | Indicates whether the bond is part of a conjugated system. This helps represent electronically connected regions of the molecule. |
| Ring membership | Binary scalar | Indicates whether the bond is part of a ring. This captures cyclic structural constraints in the molecular graph. |

**Table A.3:** Ion-level features used to describe the metal ion or adduct-forming ion.

| **Feature** | **Encoding** | **Description** |
|---|---|---|
| Atomic number | Scalar | Identifies the element and provides periodic information about the ion. |
| Atomic mass | Scalar, Da | Mass of the ion in daltons. This may influence the measured mass-to-charge ratio and distinguish different adducts. |
| Charge | Scalar | Formal charge of the ion used in the experiment. Determines the charge state of the resulting complex or adduct. |
| Ionic radius | Scalar, Å | Approximate ionic radius of the bare ion. Gives information about ion size and possible coordination geometry. |
| Number of d-electrons | Scalar | Number of d-electrons in the relevant oxidation state. This helps distinguish transition-metal ions with different electronic configurations. |
| HSAB class | Ordinal scalar | Heuristic hard–soft acid–base category, where hard, borderline, and soft ions are encoded numerically. |
| Hydrated radius | Scalar, Å | Approximate radius of the hydrated ion in aqueous solution. This describes the effective size of the ion including its hydration shell. |
| Coordination number | Scalar | Typical coordination number of the hydrated ion or common coordination environment. |
| Electronegativity | Scalar | Pauling electronegativity of the element, used as a simple descriptor of electron-attracting tendency. |

**Table A.4:** Whole-molecule descriptor groups used to provide global chemical context.

| **Feature category** | **Examples** | **Description** |
|---|---|---|
| Molecular size and mass | MolWt, ExactMolWt, HeavyAtomCount | Describe the overall size and mass of the molecule, which may influence ionization, adduct formation, and signal intensity. |
| Polarity and hydrogen bonding | TPSA, NumHDonors, NumHAcceptors, NHOHCount | Capture polar surface area and hydrogen-bonding capacity. |
| Lipophilicity and surface distribution | MolLogP, SlogP_VSA, SMR_VSA | Describe hydrophobicity and surface-area contributions distributed by lipophilicity or molar refractivity. |
| Flexibility and saturation | NumRotatableBonds, FractionCSP3, SPS | Capture molecular flexibility, saturation, and simple 3D-like structural character derived from the 2D graph. |
| Ring and aromatic structure | NumRings, aromatic/aliphatic/saturated ring descriptors | Describe cyclic and aromatic character of the molecule. |
| Electronic and charge-related descriptors | PEOE_VSA, EState_VSA, VSA_EState, BCUT2D | Approximate electronic, partial-charge, and electrotopological properties. |
| Functional-group counts | fr_C_O, fr_C_O_noCOO, fr_NH0, fr_Ar_NH, NumAmideBonds | Count selected functional groups and chemically relevant substructures. |
| Structural complexity | BertzCT, FpDensityMorgan3 | Describe molecular complexity and fingerprint density. |

**Table A.5:** Fragment-level descriptor groups computed for each molecular fragment.

| Feature category | Examples | Description |
|---|---|---|
| Relative fragment size | frac_parent_heavy | Fraction of the parent molecule's heavy atoms contained in the fragment. |
| Fragment mass and atom count | MolWt, ExactMolWt, HeavyAtomMolWt, HeavyAtomCount | Describe the size and mass of the individual fragment. |
| Fragment polarity | TPSA, NumHDonors, NumHAcceptors | Capture the polar surface area and hydrogen-bonding capacity of the fragment. |
| Fragment lipophilicity and refractivity | LogP, MR | Describe hydrophobicity and molar-refractivity-related properties of the fragment. |
| Donor and heteroatom content | donorish_atoms, n_hetero | Approximate the number of heteroatoms and donor-like atoms that may contribute to metal interaction or polar coordination. |
| Ring structure | n_rings, NumAromaticRings, NumAliphaticRings, NumSaturatedRings | Describe whether the fragment contains cyclic, aromatic, aliphatic, or saturated ring systems. |
| Flexibility and saturation | NumRotatableBonds, FractionCSP3 | Capture local flexibility and saturation within the fragment. |
| Complexity and electronic size | BertzCT, NumValenceElectrons | Describe fragment complexity and approximate electronic size. |

## A.2 All Molecule (M) and Ion Combinations Searched For in MetaboScape

### A.2.1 Experimental Configurations Performed

Plate-1 containing 38400 compounds was run with Fe, Ag, Cr, Ni, Zn, Co, Li Na K Cs.
Plate-2 containing 38400 compounds was run with Na, K Fe, Ag, Cr, Ni, Zn, Co.
Plate-3 containing 38400 compounds was run 5 times with Li, Na, K and Cs.
Plate-4 containing 38400 compounds was run with Li Na K Cs.
Mölndal-homepool containing 3000 compounds was run three times with Li Na K

**Table A.6:** Molecule–ion combinations searched for in MetaboScape. Here, M denotes a molecule from the plate.

| Ion | Searched molecule–ion combinations |
|---|---|
| K | [M+K]+, [M+K+H]2+, [M+K+H+NH4]+, [M+K+H-H2O]2+, [M+K+CH3(SO)CH3]+, [M+K+H-CO2]2+, [2M+K]+ |
| Li | [M+Li]+, [M+Li+H]2+, [M+Li+H+NH4]+, [M+Li+H-H2O]2+, [M+Li+CH3(SO)CH3]+, [M+Li+H-CO2]2+, [2M+Li]+ |
| Na | [M+Na]+, [M+Na+H]2+, [M+Na+H+NH4]+, [M+Na+H-H2O]2+, [M+Na+CH3(SO)CH3]+, [M+Na+H-CO2]2+, [2M+Na]+ |
| Cs | [M+Cs]+, [M+Cs+H]2+, [M+Cs+H+NH4]+, [M+Cs+H-H2O]2+, [M+Cs+CH3(SO)CH3]+, [M+Cs+H-CO2]2+, [2M+Cs]+ |
| Ag | [M+Ag]+, [M+Ag+H]2+, [M+Ag+H+NH4]+, [M+Ag+H-H2O]2+, [M+Ag+CH3(SO)CH3]+, [M+Ag+H-CO2]2+, [2M+Ag]+ |
| Fe | [M+Fe]2+, [M+Fe-H]+, [M+Fe+H+NH4]2+, [M+Fe+NH4]+, [M+Fe+CH3(SO)CH3]2+, [2M+Fe]2+, [2M+Fe-H]+, [M+Fe]+ |
| Ni | [M+Ni]2+, [M+Ni-H]+, [M+Ni+H+NH4]2+, [M+Ni+NH4]+, [M+Ni+CH3(SO)CH3]2+, [2M+Ni]2+, [2M+Ni-H]+, [M+Ni]+ |
| Cu | [M+Cu]2+, [M+Cu-H]+, [M+Cu+H+NH4]2+, [M+Cu+NH4]+, [M+Cu+CH3(SO)CH3]2+, [2M+Cu]2+, [2M+Cu-H]+, [M+Cu]+ |
| Co | [M+Co]2+, [M+Co-H]+, [M+Co+H+NH4]2+, [M+Co+NH4]+, [M+Co+CH3(SO)CH3]2+, [2M+Co]2+, [2M+Co-H]+, [M+Co]+ |
| Mn | [M+Mn]2+, [M+Mn-H]+, [M+Mn+H+NH4]2+, [M+Mn+NH4]+, [M+Mn+CH3(SO)CH3]2+, [2M+Mn]2+, [2M+Mn-H]+, [M+Mn]+ |
| Sc | [M+Sc-H]2+, [M+Sc-2H]+, [M+Sc+NH4]2+, [M+Sc-H+NH4]+, [M+Sc-H+CH3(SO)CH3]2+, [M+Sc-2H+CH3(SO)CH3]+, [2M+Sc-H]2+, [2M+Sc-H+NH4]+, [M+Sc-H-H2O]2+ |
| Cr | [M+Cr-H]2+, [M+Cr-2H]+, [M+Cr+NH4]2+, [M+Cr-H+NH4]+, [M+Cr-H+CH3(SO)CH3]2+, [M+Cr-2H+CH3(SO)CH3]+, [2M+Cr-H]2+, [2M+Cr-H+NH4]+, [M+Cr-H-H2O]2+ |

Cs.

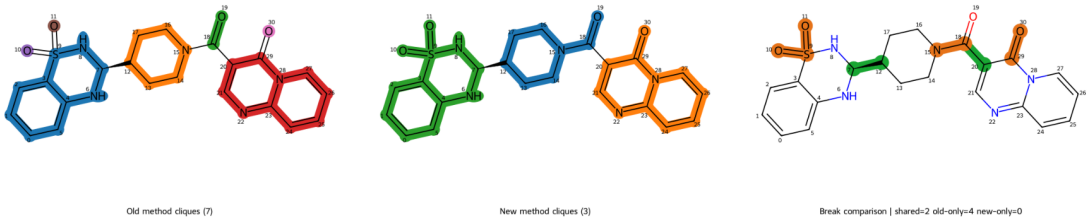



**Figure A.1:** A comparison between the old and new fragmentation algorithm, as well as the overlap between the two

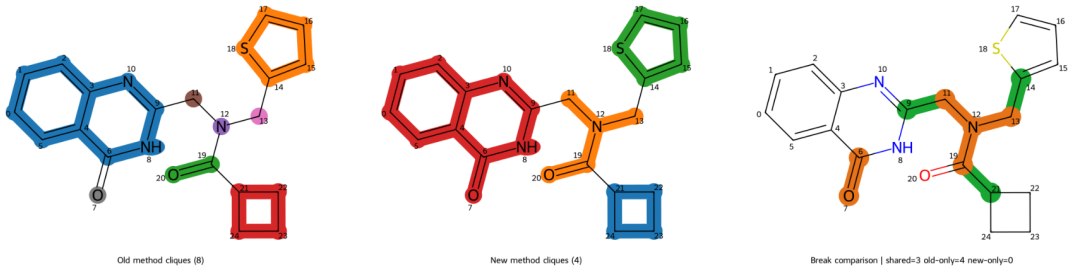



**Figure A.2:** A comparison between the old and new fragmentation algorithm, as well as the overlap between the two

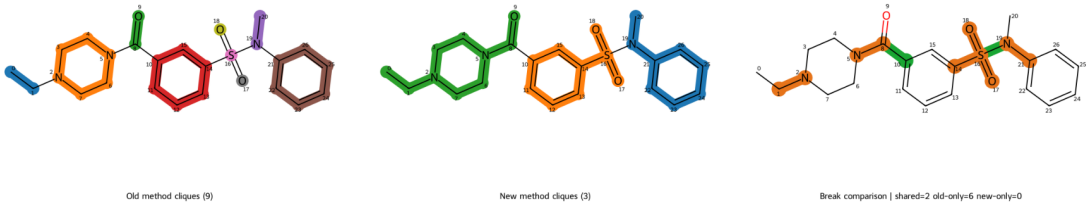



**Figure A.3:** A comparison between the old and new fragmentation algorithm, as well as the overlap between the two

## A.3 SEAL Fragmentation vs SÄLEN

## A.4 Hyperparameters

**Table A.7:** Final hyperparameters used for regression pretraining of SÄLEN.

| Setting | Value |
|---|---|
| cross_cluster_l1_weight | 0 |
| cluster_sparsity_weight | 0.1 |
| dropout | 0.14118576320924647 |
| learning_rate | 0.001248136367613311 |
| weight_decay | 0.000031899280517308704 |
| epochs | 140 |
| batch_size | 256 |
| prefetch_factor | 128 |
| num_workers | 16 |
| num_layers | 4 |
| patience | 30 |
| hidden_features | 256 |
| min_improvement | $1 \times 10^{-4}$ |
| cluster_head | mlp |
| cluster_head_hidden | 256 |

**Table A.8:** Final hyperparameters used for multitask finetuning of SÄLEN. Settings not listed here were inherited from the regression pretraining setup in Table A.7.

| Setting | Value |
|---|---|
| transfer_epochs | 60 |
| head_only_epochs | 5 |
| transfer_head_lr | 0.0001515649286298793 |
| transfer_encoder_lr | 0.00009683715127554987 |
| transfer_weight_decay | 0.000037064690203515 54 |
| regression loss weight | 1.0 |
| binding loss weight | 0.8067913472450075 |
| high-confidence sample weight | 1.0 |
| low-confidence sample weight | 3.963621822271748 |
| cls_threshold | 0.5 |
| finetune_selection_metric | balanced_accuracy |
| bind_head_dropout | 0.7051254524296289 |

**Table A.9:** Final hyperparameters used for regression training of Chemprop.

| **Setting** | **Value** |
|---|---|
| Max epochs | 70 |
| Learning rate | 0.0011277072012324102 |
| MP layers | 4 |
| MP dropout | 0.4888541982056695 |
| FFN layers | 5 |
| FFN dropout | 0.03269100267373665 |
| FFN hidden dim | 400 |
| Patience | 20 |

**Table A.10:** Final hyperparameters used for the AttentiveFP baseline model.

| **Setting** | **Value** |
|---|---|
| num_timesteps | 4 |
| use_global_features | True |
| epochs | 300 |
| patience | 20 |
| batch_size | 256 |
| num_layers | 5 |
| hidden_features | 512 |
| dropout | 0.10412510875331604 |
| learning_rate | 0.000137470823596769 |
| weight_decay | $4.4158657649666434 \times 10^{-8}$ |

**DEPARTMENT OF PHYSICS AND ASTRONOMY**
**CHALMERS UNIVERSITY OF TECHNOLOGY**
Gothenburg, Sweden
www.chalmers.se


CHALMERS
UNIVERSITY OF TECHNOLOGY