UNIVERSITY OF CALIFORNIA SAN DIEGO

Metabolite Identification in ^{1}H-^{13}C HSQC Spectra is an Image Tagging Problem

A Thesis submitted in partial satisfaction of the
requirements for the degree Master of Science

in

Computer Science

by

Jerry Zhang

Committee in charge:

Professor Garrison Cottrell, Chair
Professor William Gerwick
Professor Ndapandula Nakashole

2023

The Thesis of Jerry Zhang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

TABLE OF CONTENTS

Thesis Approval Page iii

Table of Contents iv

List of Figures v

List of Tables vii

Acknowledgements viii

Abstract of the Thesis ix

Chapter 1 Introduction 1
1.1 Nuclear magnetic resonance (NMR) spectroscopy and metabolomics 1
1.2 Previous approaches 5
1.3 Convolutional neural networks 6

Chapter 2 Methods 10
2.1 Dataset 10
2.2 Models 12
2.3 Training 14

Chapter 3 Results and Discussion 17
3.1 Training performance 17
3.2 Test set performance 19
3.3 Experimental dataset performance 22

Chapter 4 Conclusion and Future Work 27

Bibliography 29

LIST OF FIGURES

Figure 1.1. An illustration of the NMR data collection process. (A) The emission of EM waves. (B) The capturing of waves in the time domain as free induction decay (FIDs). (C) The application of the Fourier transformation to obtain the NMR spectrum in the frequency domain. Figure taken from [3]. 3

Figure 1.2. An example of NMR spectra from yeast metabolite extract. (A) is a 1D ^{1}H spectrum and (B) is a 2D HSQC spectrum of the same sample. Each of the axes represent resonant frequency, also referred to as chemical shift: ^{1}H is given on the x axis and ^{13}C on the y axis. Figure taken from [3]. 4

Figure 1.3. AlexNet, a CNN which was trained on natural images from the ImageNet database. The convolution operation is often depicted as a volume, since the images retain their 2-dimensional shape and the feature maps from each kernel can conceptualized as stacked planes. Figure from [13]. 6

Figure 1.4. The SMART-Miner model, which is a U-Net model with a softmax prediction layer in the middle of the model. Figure from [12]. 8

Figure 1.5. The ResNet34 model, which is the variant of ResNet with 34 layers. Skip connections, which cause the model to learn residuals between layers, are shown to connect inputs and outputs across convolutional layers. Figure from [10]. 9

Figure 2.1. An example of the dataset generation process. Top row: the process of combining several metabolites into a mixture. Bottom row: Performing augmentation by removing actual peaks and adding noisy ones. 11

Figure 2.2. An example of inputs and outputs to SMART-Miner. In our model, we remove the need for the second channel from the input, and the label becomes a one-hot vector which encodes all present metabolites. Figure from [12]. 13

Figure 2.3. An example of inputs and outputs to our U-Net approach. In the output image, delete peaks have been recovered and added peaks removed. The label is a list of metabolites rather than a single prediction. 15

Figure 3.1. Results of grid-search over hyper-parameters for the ResNet18 model ("lr"=learning rate and "weight_decay"=L2 penalty). 18

Figure 3.2. Results of grid-search over hyper-parameters for the U-Net model. 19

Figure 3.3. Loss on the hold-out dataset of the ResNet18 and U-Net models during training at each noise level. Each plot is the average over three runs. 20

Figure 3.4. F1 performance of the ResNet18 and U-Net models during training at each noise level. Each plot is the average over three runs. 21

Figure 3.5. The three experimental spectra gathered in [21] and processed using DeepPicker [14]. The top row shows the experimental data by themselves, and the bottom row shows the experimental data with a synthetic reconstruction overlaid in red. 24

LIST OF TABLES

Table 2.1. The parameters used to generate each of the four datasets by noise level. . . 12

Table 3.1. Performance (F1 score) of each model on the testing datasets. “No shift” means a dataset in which the peaks were not shifted. For our models, the number in the parentheses indicates which noise level the model was trained on. 23

Table 3.2. Performance of each method on the N925 experimental spectrum. Data from [12]. 25

Table 3.3. Performance of each method on the N987 experimental spectrum. Data from [12]. 25

Table 3.4. Performance of each method on the N988 experimental spectrum. Data from [12]. 26

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Garrison Cottrell for his support as the chair of my committee. Through his mentorship throughout my university career, his guidance has proved to be invaluable.

I would also like to acknowledge the members of Gary's Unbelievable Research Unit (GURU), whose tips and support have helped me in an immeasurable way.

Finally, I would like to dedicate this to St. Josemaría Escrivá, a model teacher and scholar; without his intercession and admonitions to study and work, I surely would not have been able to complete this work.

ABSTRACT OF THE THESIS

Metabolite Identification in ^{1}H-^{13}C HSQC Spectra is an Image Tagging Problem

by

Jerry Zhang

Master of Science in Computer Science

University of California San Diego, 2023

Professor Garrison Cottrell, Chair

Metabolomics, or the study of compounds essential for cellular function, is a field with increasing application in the research of organisms and organic systems; one component of this is the automatic identification of metabolites from the spectral analysis of samples. This can be challenging, however, due to chemical shifts on account of laboratory conditions, as well as noise arising from the experiment itself. In this work, we develop an approach to this task using a convolutional neural network (CNN) on ^{1}H-^{13}C nuclear magnetic resonance (NMR) spectra. With very limited experimental data, we synthesized our own set of metabolic mixture data and trained the neural network to achieve good performance compared to other automatic metabolite identification methods predicting from NMR data.

Chapter 1

Introduction

Deep learning is rapidly becoming the most widely used method for automating tedious tasks in other fields; metabolomics is not an exception [5]. Within metabolomics, several tasks are ripe for deep learning applications, such as identifying molecular structure [22], classification of compounds and samples [9], and medical diagnoses [23]. The numerous and varied downstream applications of metabolomics research makes advancement in automated methods all the more important, especially for tasks more basic and general such as the identification of compounds from NMR spectra. This work therefore seeks to take one step in the improvement of this area.

The following introduction is laid out as follows: a brief explanation of the relevant parts of NMR spectroscopy is given, then we discuss previous methods applied to similar tasks, and finally we explain the mechanics of convolutional neural networks, which constitute the approach of this work.

1.1 Nuclear magnetic resonance (NMR) spectroscopy and metabolomics

"Metabolomics" refers to the study of small molecular compounds responsible for essential cellular functions, including energy production and storage: these compounds are appropriately named "metabolites" [11]. Metabolomics is a rapidly expanding field with increasing applications. The extraction, processing, and analysis of the "metabolome" of an organism or

system can provide insight into its metabolic health, as well as for the development of therapies including for cancer and neuro-degenerative diseases [6]. To this end several methods have been developed for representing and analyzing these molecular compounds, and among these the most commonly used are liquid chromatography with mass spectrometry (LC-MS), gas chromatography with mass spectrometry (GC-MS), and nuclear magnetic resonance spectroscopy (NMR) [7].

In NMR spectroscopy, experiments reveal how the nuclei of a specific element are distributed in the molecules of a sample [3]. This is data is collected by exploiting magnetic properties in the nuclei of these biological compounds using radio-frequency (RF) waves. During the experiment a sample is irradiated with RF waves, which may be grouped into "pulse sequences," causing the nuclei to flip spin states. When the RF waves are disabled, the nuclei emit electromagnetic waves in the time domain as they return to equilibrium. These emitted signals are then processed via Fourier transformation and are ultimately what are represented in the NMR spectra (this is why the axes are measured in terms of resonant frequency). The whole process may be likened to moving the needle of a compass, and observing its returning to face north [3]. This method of spectroscopy has seen an increasing interest over other experiments due to its relatively easy preparation process, its quantification of metabolite levels, and its nondestructive nature; however, among its disadvantages is its lower sensitivity when compared to other experiments [7].

Within NMR spectroscopy exist several more types of experiments. For example, one-dimensional ^{1}H spectra are one of the more commonly analyzed for metabolomics [3]. However, this method suffers from overlapping resonances that obscure peak identification. Another method of NMR spectroscopy involves heteronuclear pulse sequences which detect bonds between ^{1}H nuclei and ^{13}C nuclei: this makes the the process of identifying compounds faster and easier. Experiments of this kind includes heteronuclear single quantum coherence (HSQC), multiple quantum coherence (HMQC), and multiple bond coherence (HMBC) [1]. Examples of these spectra are shown in figure 1.2. In this work we deal with exclusively HSQC spectra.

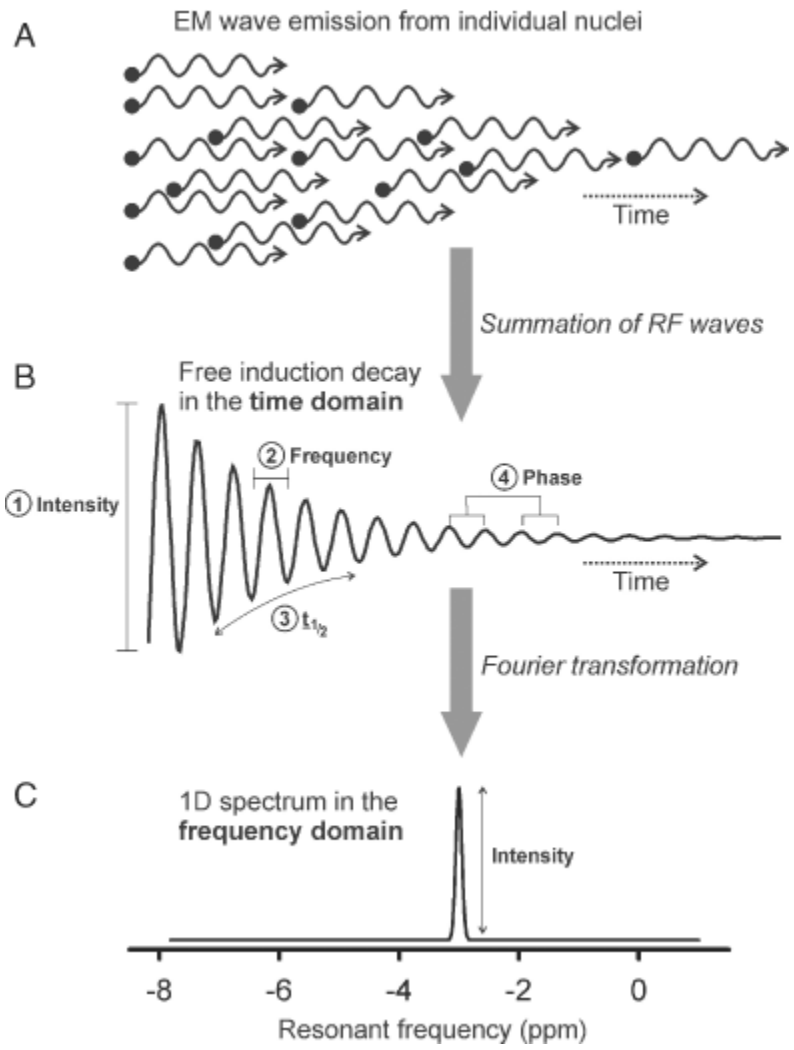


Figure 1.1. An illustration of the NMR data collection process. (A) The emission of EM waves. (B) The capturing of waves in the time domain as free induction decay (FIDs). (C) The application of the Fourier transformation to obtain the NMR spectrum in the frequency domain. Figure taken from [3].

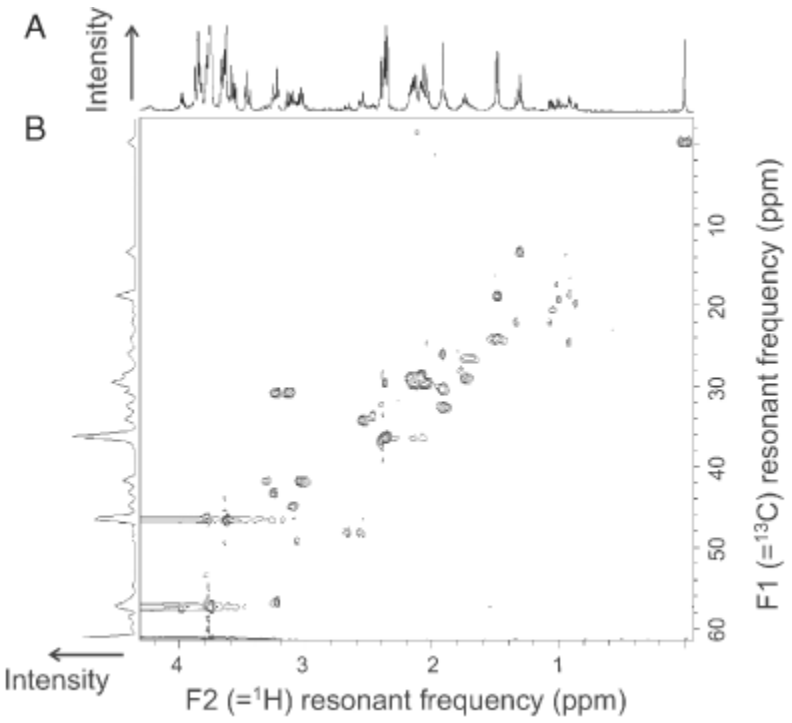


Figure 1.2. An example of NMR spectra from yeast metabolite extract. (A) is a 1D ^{1}H spectrum and (B) is a 2D HSQC spectrum of the same sample. Each of the axes represent resonant frequency, also referred to as chemical shift: ^{1}H is given on the x axis and ^{13}C on the y axis. Figure taken from [3].

1.2 Previous approaches

Many approaches to this problem use a comparative approach to databases which contain spectral samples of several hundred individual metabolites under different experiments, for example: the Human Metabolome Database (HMDB) [20] and the Biological Magnetic Resonance Data Bank (BMRB) [15], which are both also used in our approach. MetaboMiner [21] is one such approach which algorithmically assigns "uniqueness" values to peaks based on a library of spectra to compare to spectra gathered for inference. These are used to query peaks to identify them according to the reference library, conditioned under a set of authenticity checks—rules which help reduce false positives. The algorithm first performs a reverse-search using these values, where library peaks are matched against query peaks. Then, for unmatched peaks, a forward search is performed against the reference. This approach was performed both for total correlation spectroscopy (TOCSY) and HSQC spectra. COLMAR is a similar approach which uses a matching ratio to compare experimental spectra to reference spectra, and they also define a "uniqueness" parameter, though different from MetaboMiner and mostly used to reduce false positives. In addition, they consider spectra under different isomeric states to further refine their analysis [2]. SpinAssign is another approach which uses statistical indexing and a reference chemical shift database [4].

Some approaches use learning algorithms, including deep neural networks, for similar tasks in metabolomics. Decision trees, discriminants, support vector machines (SVM), and k-nearest neighbors (kNN) have been used to classify spectral data gathered from food [9]. A deep neural network has also been used by [19] on ^{1}H spectra to predict microbes. MetFID is a deep convolutional neural network (CNN) to predict molecular fingerprints and rank metabolite identification, but on mass spectroscopy data [8].

SMART-Miner uses a CNN to identify metabolites from HSQC spectra [12]. Specifically, they use a U-Net [17], adding a classification output in the middle of the model. The addition of this classification output is necessary because by design U-Net is a semantic segmentation

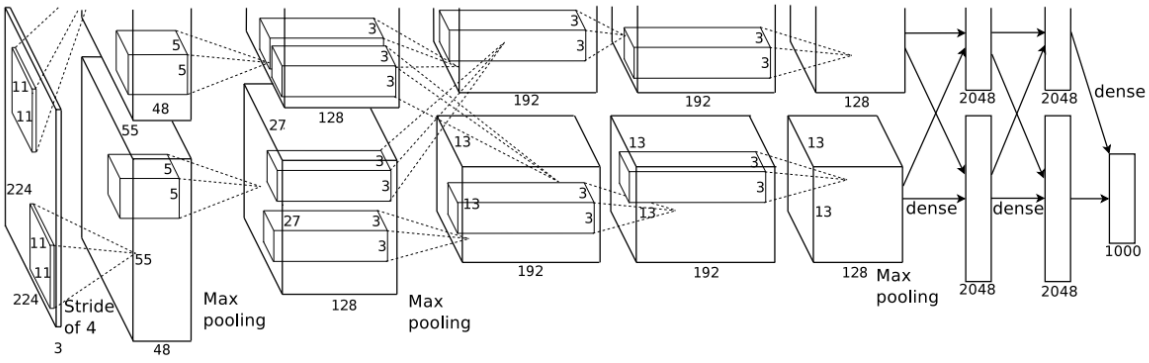


Figure 1.3. AlexNet, a CNN which was trained on natural images from the ImageNet database. The convolution operation is often depicted as a volume, since the images retain their 2-dimensional shape and the feature maps from each kernel can conceptualized as stacked planes. Figure from [13].

model. That is, the model predicts outputs over an array of values—effectively a map over the input—rather than a single vector of labels. With this classification output, they treat the task as a kind of image categorization problem: given a set of peaks and a query location, they classify the image into one of several hundred potential compounds. Simultaneously, at the output of the U-Net, they identify all the other peaks in the mixture that correspond to the predicted compound (see figure 2.2). Our work takes this approach as a baseline and seeks to show how reformulating the task as an image tagging problem can make the model simpler, while also avoiding the problem of overlapping peaks which is present in SMART-Miner.

1.3 Convolutional neural networks

Convolutional neural networks are deep neural networks that make use of convolutions to compute the same features across an image, reducing the number of parameters in the model necessary to extract features from the image [13]. Specifically, instead of each input being connected to every output by a weight in a given layer of the neural network, several kernels (which are the parameters of the model) are cross-correlated across the input, and the output can be considered a "feature map" of the kernels (see figure 1.3). This significantly reduces the complexity of the model since the parameters consist only of these kernels, which each can

be tens or hundreds of parameters, instead of connections across every input and output. At the same time, this operation exploits properties of images, e.g. the correlation of pixels in a local region and the multiple-occurrence of features across and image. The sequencing of many layers—as in the case of fully-connected neural networks—enables the model to learn more complex features for the task.

ResNet [10] is a further specification of a family of CNNs which make use of residual connections—connections across layers that cause the model to learn residual functions instead of unreferenced functions. These are shown to be easier to optimize than standard convolutional networks, and have an increased gain from depth. This family of models was trained and evaluated on the ImageNet [18] dataset, and like many deep CNNs is built up by a repeating set of similar blocks which terminate finally with a fully-connected linear layer to generate a prediction over a distribution of classes (see figure 1.5). Because of its design as a classification model, ResNet seems an appropriate choice to approach the task of tagging metabolites given images of HSQC spectra.

U-Net [17], on the other hand, is a model primarily designed for semantic segmentation and as a consequence has an entirely different structure (see figure 1.4). It consists of an encoding portion, where the input trades in fine-grained detail for a deeper feature representation, as well as a decoding portion, where the fine-grained information is reproduced from this feature space. To make this reproduction easier, copy connections are allowed at each level of the U-Net. Most importantly, the output of this model is the same shape as the input; thus in order to obtain predictions over classes, an output prediction has to be obtained from somewhere else in the model. Following SMART-Miner, it makes sense to collect this prediction at the "bottom" of the U-Net, where the representation is the most highly encoded in latent space. The ordinary output of the U-Net might then be used for some auxiliary task to improve the model's performance on the main task; such is the case with SMART-Miner's prediction of all peaks corresponding to the metabolite which it predicts.

In addition to the above applications of CNNs, CNNs have also been used for other

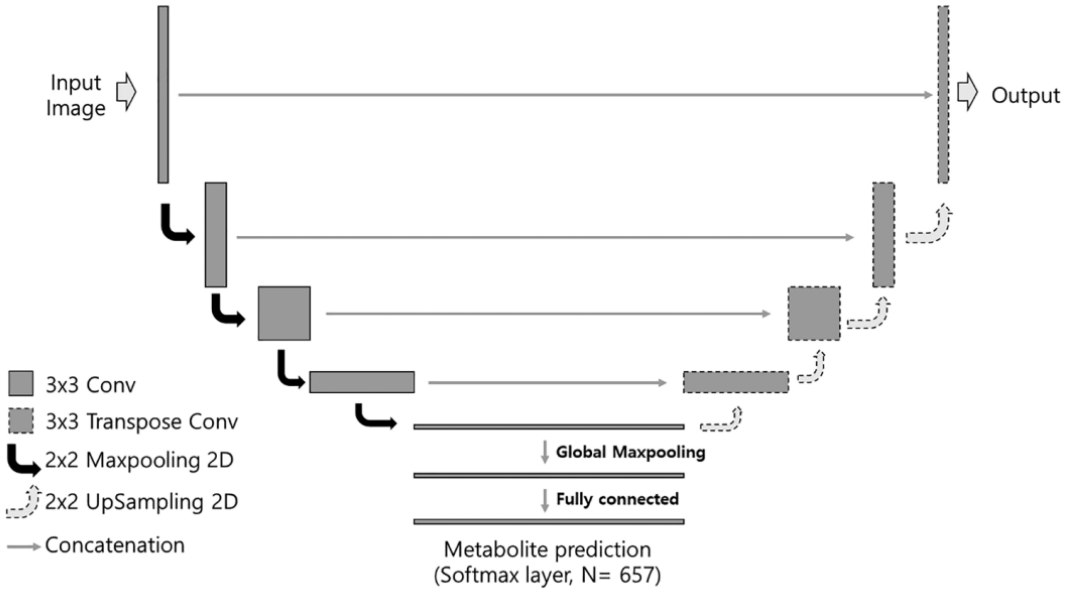


Figure 1.4. The SMART-Miner model, which is a U-Net model with a softmax prediction layer in the middle of the model. Figure from [12].

tasks within NMR metabolomics, such as predicting molecular structure [16] and spectral peak deconvolution [14]. These examples show the viability of CNNs as an avenue for processing and analyzing these types of data. In our formulation of the task as a tagging problem, we use ResNet models as well as demonstrate how SMART-Miner's use of the U-Net model can be adapted from classification.

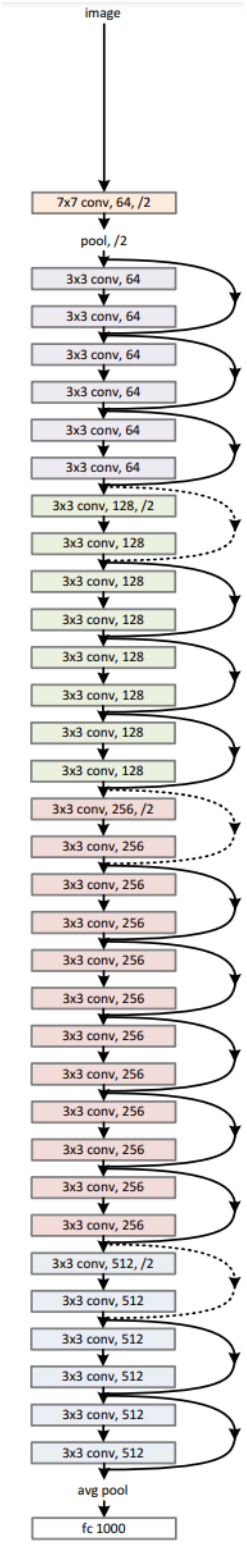


Figure 1.5. The ResNet34 model, which is the variant of ResNet with 34 layers. Skip connections, which cause the model to learn residuals between layers, are shown to connect inputs and outputs across convolutional layers. Figure from [10].

Chapter 2

Methods

Our work for the task is constituted by three main efforts: first, we specify and collect the data (or more specifically synthesize the data), then we develop and implement our models for the problem, and finally we conduct experiments in order to achieve the best-performing model for the data.

2.1 Dataset

One of the primary obstacles in automated metabolomics, particularly within the subset of HSQC spectroscopy, is a lack of available training data for which a list of known metabolites can be provided for an example spectrum. In fact, the few examples that have been studied were the result of synthesizing pure compounds in a laboratory [21]. In automated methods, they are most often reserved for use at evaluation time [12].

Therefore, following [12] an artificial dataset of HSQC spectra was collected; these examples were constructed by combining the known spectra of individual metabolites from the HMDB and BMRB databases, which in total represent 657 metabolites and 1006 HSQC spectra (some metabolites have multiple spectra corresponding to various laboratory conditions). From these data, which are provided as a list of peaks for each metabolite, a mixture can be simulated by overlaying several spectra on top of each other. It should be noted that this process ignores interactions between metabolites that might occur in an actual experiment. In addition, the data

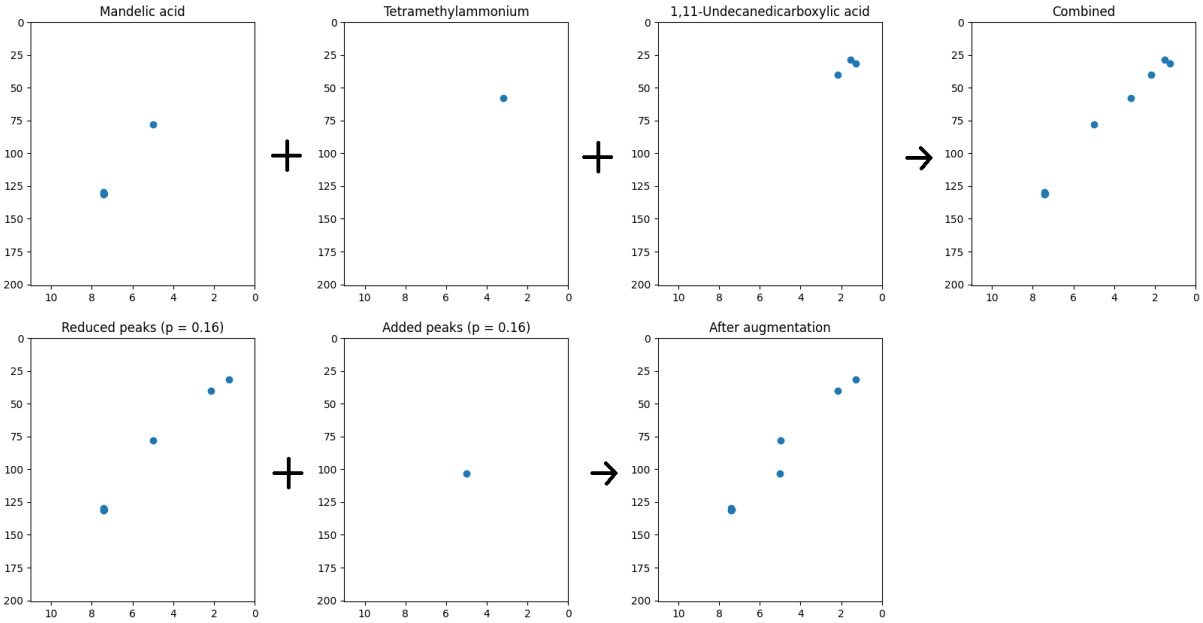


Figure 2.1. An example of the dataset generation process. Top row: the process of combining several metabolites into a mixture. Bottom row: Performing augmentation by removing actual peaks and adding noisy ones.

is provided in bitmap format, which means that rather than displaying peaks as concentric rings, a single bit is placed at location of highest intensity of each peak, and the rest of the intensity information is discarded (see figure 2.1).

Since under experimental conditions some peaks might fall below the detection range of the device, and spurious peaks may be present (due to water artifacts, for example) [21], we created four datasets at various noise levels (table 2.1). To simulate the above conditions, we apply random augmentation parameterized by deletion probability and additive proportion: deletion probability specifies the probability for each peak to be randomly deleted from the dataset, and additive proportion is multiplied with the number of peaks already in the mixture to specify how many additional peaks to add, randomly distributed through the spectrum.

Each dataset is constructed by randomly selecting 10 to 40 metabolites from the database and overlaying their spectra to synthesize a mixture. Then, in addition to the noise augmentation procedure describe above, we also randomly shift the spectrum by 0.05 in the proton dimension and 0.5 in the carbon dimension. This shift is also performed to increase the robustness of the

Table 2.1. The parameters used to generate each of the four datasets by noise level.

Noise level	Deletion probability	Additive proportion
0	0.00	0.00
1	0.04	0.04
2	0.08	0.08
3	0.16	0.16

model to experimental conditions. These datasets each contain 500,000 training examples and 50,000 held-out examples.

2.2 Models

Following [12], we used deep convolutional networks to approach this task; however, while they use a query peak channel to obtain a probability distribution over metabolites for each individual peak, we use an output layer of logistic units, thus obtaining a probability for each metabolite over the whole spectrum.

One of the advantages of this approach is simplicity: the use of a query channel results in inputs with an extra channel of inputs, which doubles the number of input data points. Moreover, our approach is faster, since the metabolites for all peaks are predicted simultaneously in one pass of the model, while in the previous method each peak would require its own pass through the model (or a batch size which is equal to the number of peaks in the mixture, if batched input is allowed).

The other advantage of this approach is the resolution of overlapping peaks. In HSQC spectra, there is a possibility that peaks from multiple metabolites overlap and appear as one peak [12]. In the previous peak-by-peak approach, each metabolite competes for weight in the probability distribution predicted by the model, and it is therefore impossible to distinguish when the model has an uncertain prediction versus when there are multiple metabolites present under one peak. This is a result of formulating the problem as categorization so that a softmax is computed over the output of the model, which calculates a probability distribution (equation 2.1).

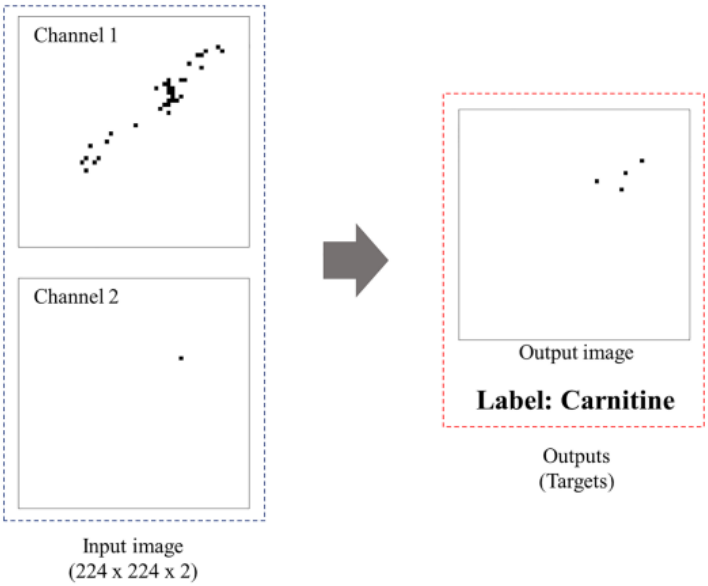


Figure 2.2. An example of inputs and outputs to SMART-Miner. In our model, we remove the need for the second channel from the input, and the label becomes a one-hot vector which encodes all present metabolites. Figure from [12].

This issue is resolved in the current approach, since the case of overlapping peaks is handled implicitly by the model. That is, in the formulation of the task as a tagging problem, at the output of the model the predicted probability of each metabolite is independent from each other metabolite (equation 2.2).

$$\text{softmax}(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{2.1}$$

$$\sigma(\mathbf{z})_i = \frac{1}{1+e^{z_i}} \tag{2.2}$$

The trade-off made with this approach is interpretability. Namely, by predicting metabolites peak-by-peak, a "map" of the predictions can be obtained from the model, while with the current approach the correspondence between peaks and metabolites is not so clear. This is, however, not a crippling trade-off, since by using the existing databases and basic pattern-matching it is not too difficult to recover a metabolite-to-peaks correspondence map.

In this work we use ResNet models, which are deep convolutional networks with residual connections [10]. In order to fit the models to our task, we add a final linear layer to the model which changes the number of output units from 1000 (the number of classes in ImageNet) to 657 (the number of metabolites in our database), as well as an initial convolutional layer to increase the number of input channels to 3, which is expected from the model since it was trained on RGB images. We also experiment with a U-Net model in a similar fashion to SMART-Miner; however, in order to adapt the model to an image tagging approach we continue to apply the logistic activation at the output computed at the bottom of the U-Net. At the ordinary output of the U-Net, we perform an auxiliary task of reproducing the input but without peak augmentation—that is, the spectrum without any additional peaks, and the deleted peaks recovered (see figure 2.3. The loss function for this model is then the sum of the classification loss and this auxiliary reproduction loss. This was designed with the aim of encouraging the model to learn the correspondence of peaks to metabolites, even in the presence of noise.

2.3 Training

A hyper-parameter grid-search was performed with the ResNet18 and U-Net models over the following: a learning rate of 1×10^{-3} and 1×10^{-4} and a L2 penalty of 1×10^{-4}, 1×10^{-6}, and 0. The ResNet18 model is trained for 80 epochs, and the U-Net model for 40 (the U-Net model converges more quickly, but takes more time per iteration). Each model is trained with the Adam optimizer on binary cross entropy loss. An initial test showed that training from pretrained weights (e.g. from ImageNet) and from scratch did not affect the final result of the model on the hold-out dataset. From these results, the best models were chosen by best F1 score on the hold-out dataset.

The training of these models was performed on a GPU cluster with GPUs such as NVIDIA GeForce GTX 1070, 1080, and 1080Ti, which range in video memory from 8GB to 12GB. With the above parameters, ResNet18 took about 14.5 hours to train, and U-Net about 36

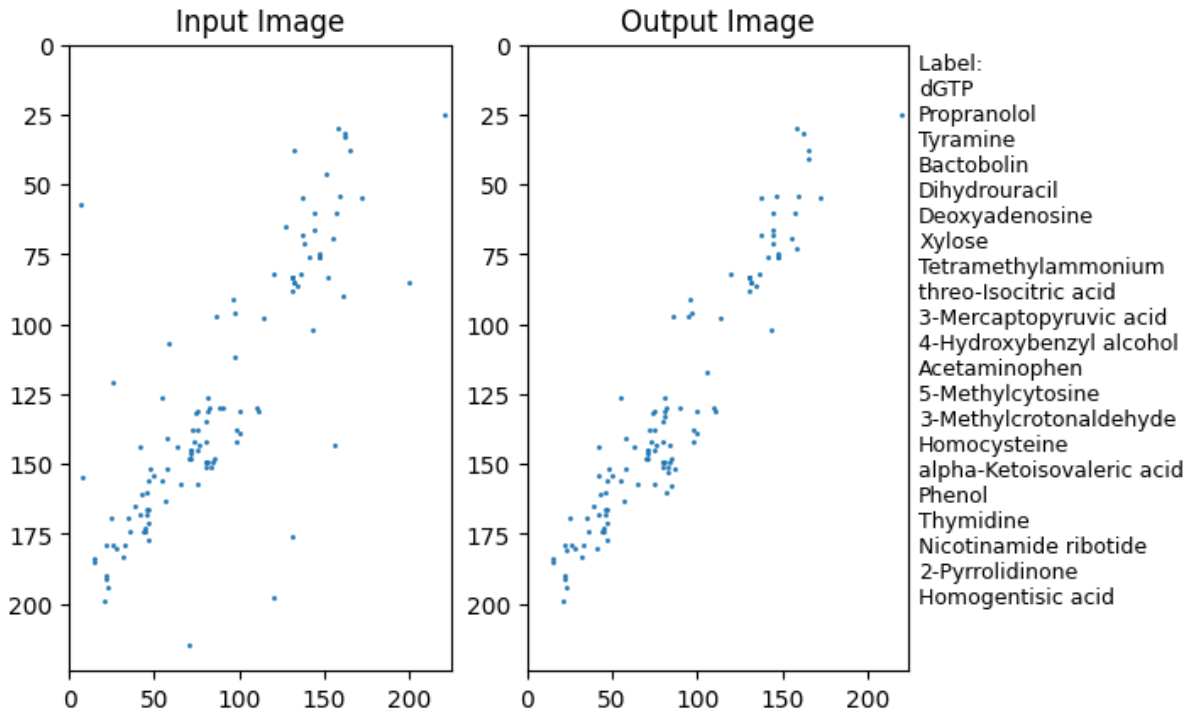


Figure 2.3. An example of inputs and outputs to our U-Net approach. In the output image, delete peaks have been recovered and added peaks removed. The label is a list of metabolites rather than a single prediction.

hours. The models with all training and evaluation pipelines were implemented in Python with the PyTorch framework.

Chapter 3

Results and Discussion

After training the model, we evaluated its performance on both a generated test set as well as a set of experimental data used by other methods for evaluation [12]. In the following section, we show and explain our results on the training process, our generated dataset, and this experimental dataset.

3.1 Training performance

Results for the hyper-parameter grid search are shown in figures 3.1 and 3.2. From these data it is clear that a higher learning rate yields better performance on the ResNet model, but the reverse is true for the U-Net model. For both models, we observed that precision increased as the L2 weight penalty increased, but at the expense of recall. Optimizing for F1 score, we selected the middle-ground on this parameter for both models.

The models' performance (loss and F1 score) on the hold-out dataset over training iterations is shown in figures 3.3 and 3.4. In all these plots it is universally the case that as the noise level increases, the model converges to a lower F1 score and to a higher loss. This suggests that as the noise level increases, the model is not able fully recover its ability to encode the metabolites' spectral data. However, the robustness that the noise augmentation lends the model is still helpful in test cases, as discussed below. A direction for future work might include developing more complex models capable of distinguishing signal and noise, or training a model

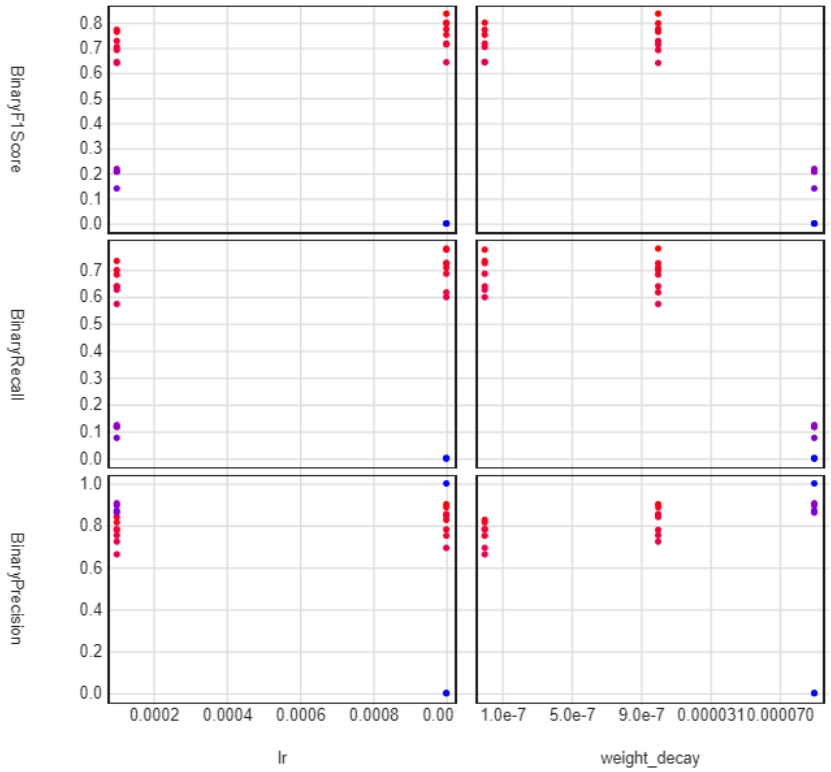


Figure 3.1. Results of grid-search over hyper-parameters for the ResNet18 model ("lr"=learning rate and "weight_decay"=L2 penalty).

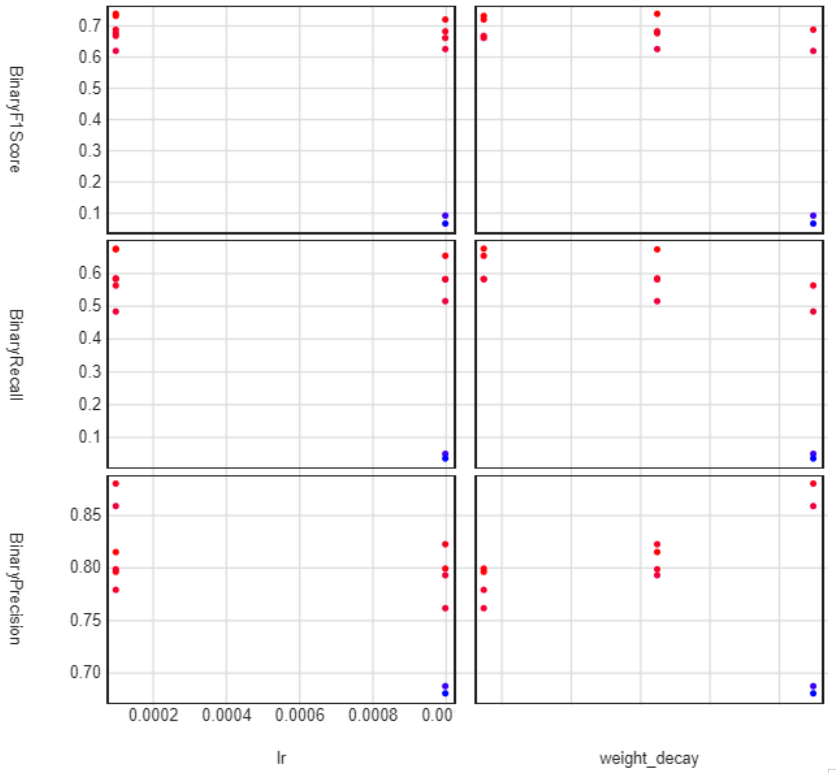


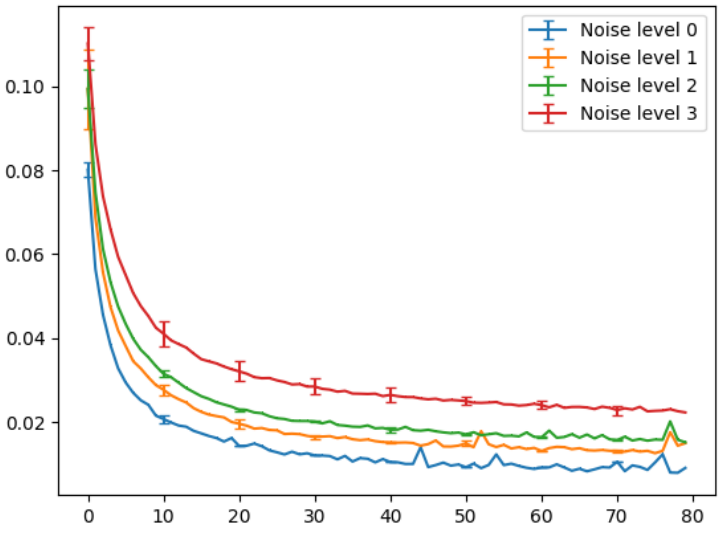
Figure 3.2. Results of grid-search over hyper-parameters for the U-Net model.

using curriculum-based learning to gradually make the model more robust to noise.

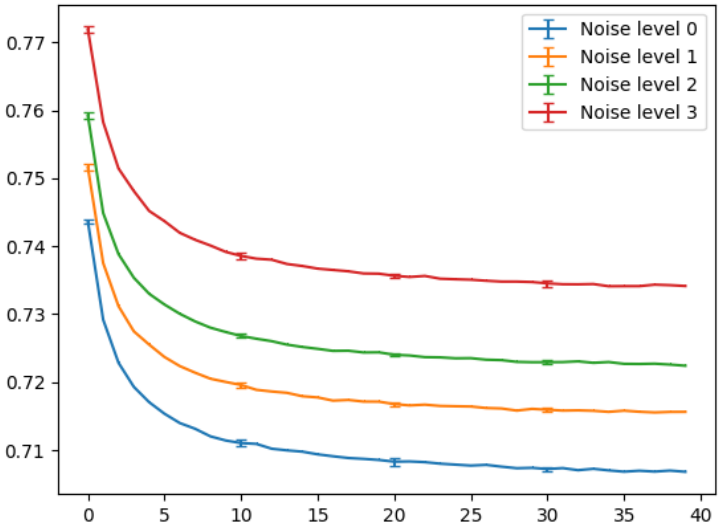
Also apparent in the plots is U-Net's degradation in performance as noise level increases, especially when compared to ResNet18 (U-Net also has a higher loss overall, but this can be attributed to the additional loss of the auxiliary task of regenerating the peaks at its output). This finding agrees with the test results shown below, and indicates that the model might not be complex enough for the more difficult noise-augmented task.

3.2 Test set performance

The model was evaluated on a test set of data generated by the same process outlined above, but never seen by the model or used for tuning (table 3.1). The performance shows that it is possible for a model to capture the relationship between HSQC spectra and metabolites, without the need for hand-crafted rules or algorithms. From these test results we observe that the

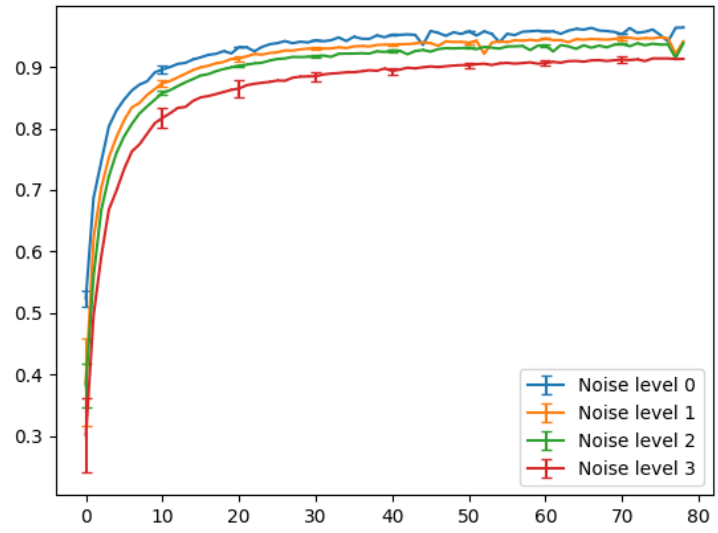


(a) ResNet18

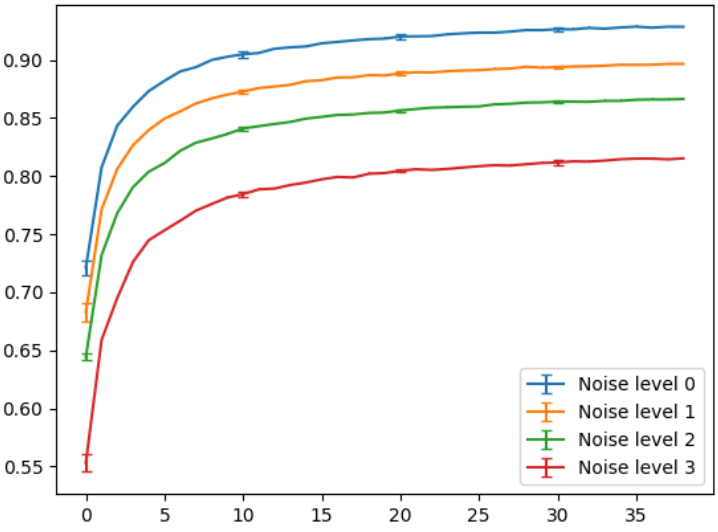


(b) U-Net

Figure 3.3. Loss on the hold-out dataset of the ResNet18 and U-Net models during training at each noise level. Each plot is the average over three runs.



(a) ResNet18



(b) U-Net

Figure 3.4. F1 performance of the ResNet18 and U-Net models during training at each noise level. Each plot is the average over three runs.

noise augmentation process improves the performance of the ResNet18 model across all datasets, while on the U-Net model this is only the case for the noisier datasets; in fact, the U-Net model actually suffers a performance decrease on less noisy data when trained on noiser data. This seems to indicate that, at least for the ResNet model, increasing the difficulty of the task allows the model to generalize HSQC data better under various conditions. This may not be observed for the U-Net model due to its smaller number of parameters and thus reduced complexity when compared to ResNet18. This is supported by the model's generally lower scores as well as its greater performance degradation as the noise level of the test data increases.

We also evaluated SMART-Miner on our test datasets. Each example was reformatted to a peak-by-peak approach; that is, each example becomes several examples—equal to the number of peaks in the spectrum—which are then evaluated batch-wise on the model. Despite SMART-Miner being trained on a similar synthetic dataset, we observe that its performance across all datasets is markedly worse than ours, which suggests that there is a difference between their synthetic dataset and ours (beyond the changes which are a consequence of the reformulation of the task). Interestingly, the model still performs well on data without any shifts, which indicates that our data encodes the information necessary to predict metabolite outputs. However, this performance drops steeply once shift is added, despite SMART-Miner being trained on the same shifting augmentation. Performance continues to drop when noise augmentation is applied, which is expected since they obscure metabolite information, and SMART-Miner was not trained with such augmentations.

3.3 Experimental dataset performance

The model was also evaluated on spectra obtained from actual NMR experiments, though these mixtures were obtained synthetically so that their composition is known [21]. These are the N925, N987, and N988 mixtures, which contain 27, 21, and 24 compounds, respectively (see figure 3.5). The data was processed as in [12], i.e. DeepPicker was applied for noise removal

Table 3.1. Performance (F1 score) of each model on the testing datasets. "No shift" means a dataset in which the peaks were not shifted. For our models, the number in the parentheses indicates which noise level the model was trained on.

	Dataset (noise level)				
Model	level 0 & no shift	level 0	level 1	level 2	level 3
SMART-Miner	0.801	0.388	0.402	0.373	0.285
ResNet18 (0)	0.959	0.956	0.910	0.877	0.778
ResNet18 (1)	0.963	0.964	0.942	0.928	0.864
ResNet18 (2)	0.962	**0.969**	**0.951**	0.937	0.905
ResNet18 (3)	**0.965**	0.965	0.950	**0.945**	**0.918**
U-Net (0)	0.938	0.936	0.878	0.834	0.713
U-Net (1)	0.931	0.928	0.893	0.859	0.781
U-Net (2)	0.925	0.923	0.895	0.879	0.793
U-Net (3)	0.911	0.919	0.872	0.866	0.823

and peak picking.

Also shown in figure 3.5 is a synthetic reconstruction of the experimental spectra—that is, artificial spectra generated by the same method as the training data using the ground truth labels provided with the spectra. From these overlays it is clear that the data remains very noisy even after being processed. In particular there are many extra peaks in the streaks found in N925 and N988 (the result of water being present in the mixture during the experiment), as well as the missing and spurious peaks across the whole spectra.

The results of our models, as well as a comparison to other metabolite identification methods, are shown in tables 3.2, 3.3, and 3.4. From these results, we see that this deep learning approach outperforms database look-up methods (HMDB and BMRB) as well as the SpinAssign algorithmic method. Further, the results varying over noise level show some indication that the noise augmentation method was effective in increasing the robustness of the model to noise. In general our models show high precision scores; this is especially the case for our U-Net model, which outperforms SMART-Miner in terms of precision on all spectra. This means that out of the model's predictions, the metabolites were more likely to be actually present in the mixture. However, our models' recall performance is not as good, and this results in a lower overall F1 score; even still, this F1 score still comes close to SMART-Miner's and is competitive with other

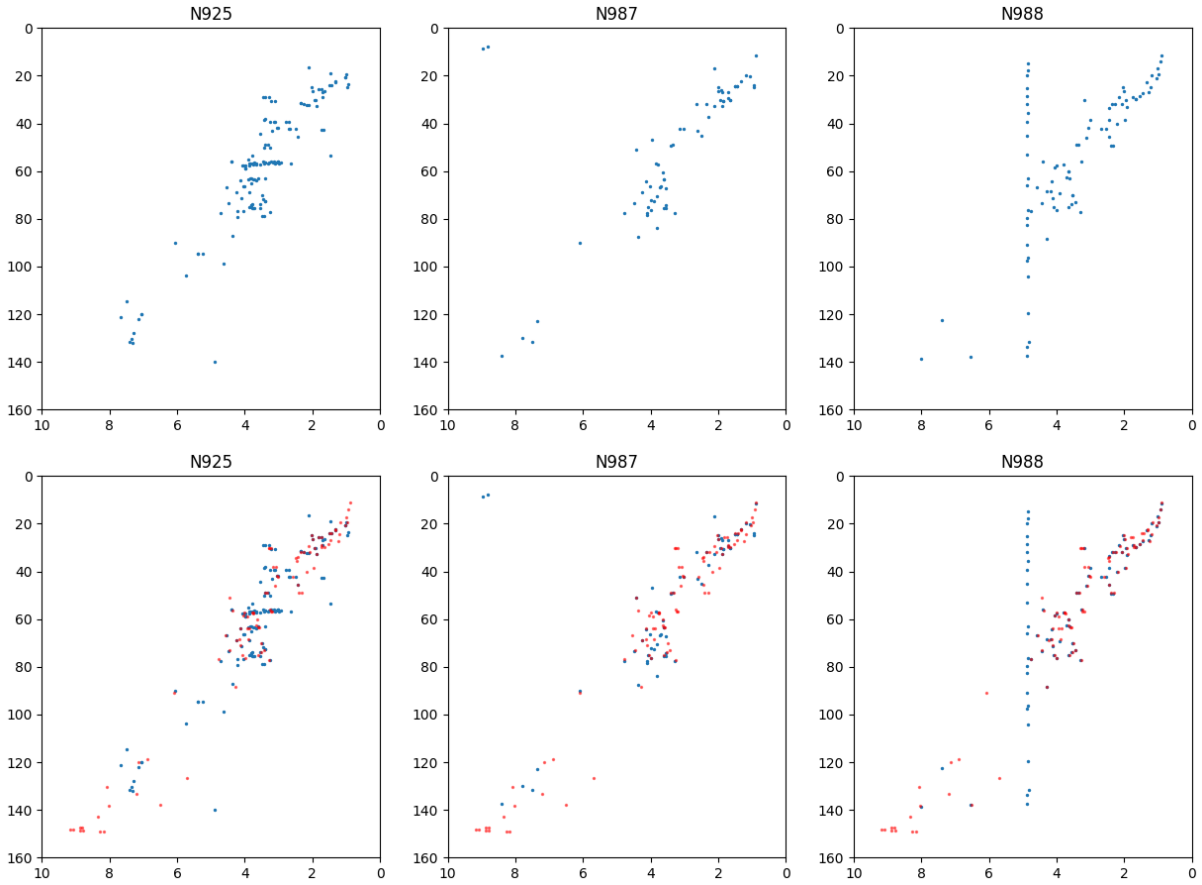


Figure 3.5. The three experimental spectra gathered in [21] and processed using DeepPicker [14]. The top row shows the experimental data by themselves, and the bottom row shows the experimental data with a synthetic reconstruction overlaid in red.

methods. On the N925 spectra, the ResNet18 model outperforms Metabominer; and on N988, the ResNet18 model outperforms COLMAR. Our model obtains the worst performance on the N987 spectrum, and it seems like generally the other methods also suffer worse performance on this example. From the overlay shown in figure 3.5, this is likely due to the high number of deleted peaks in this particular example.

Table 3.2. Performance of each method on the N925 experimental spectrum. Data from [12].

Method	Precision	Recall	F1 Score
SMART-Miner	0.70	**0.70**	0.70
COLMAR-HQSC	**0.86**	0.67	**0.75**
SpinAssign	0.19	0.65	0.29
Metabominer	0.65	0.56	0.60
HMDB	0.30	0.30	0.30
BMRB	0.22	0.22	0.22
ResNet18 (0)	0.50	0.41	0.45
ResNet18 (1)	0.36	0.30	0.33
ResNet18 (2)	0.56	0.52	0.54
ResNet18 (3)	0.67	0.67	0.67
U-Net (0)	0.63	0.52	0.57
U-Net (1)	0.67	0.59	0.63
U-Net (2)	0.62	0.48	0.54
U-Net (3)	0.75	0.56	0.64

Table 3.3. Performance of each method on the N987 experimental spectrum. Data from [12].

Method	Precision	Recall	F1 Score
SMART-Miner	0.74	**0.67**	**0.70**
COLMAR-HQSC	**0.90**	0.43	0.58
SpinAssign	0.19	0.65	0.29
Metabominer	1.00	0.43	0.60
HMDB	0.05	0.05	0.05
BMRB	0.05	0.05	0.05
ResNet18 (0)	0.73	0.38	0.50
ResNet18 (1)	0.67	0.48	0.56
ResNet18 (2)	0.5	0.38	0.43
ResNet18 (3)	0.62	0.38	0.47
U-Net (0)	0.38	0.24	0.29
U-Net (1)	0.64	0.33	0.44
U-Net (2)	0.72	0.38	0.50
U-Net (3)	0.75	0.29	0.41

Table 3.4. Performance of each method on the N988 experimental spectrum. Data from [12].

Method	Precision	Recall	F1 Score
SMART-Miner	0.68	0.63	0.65
COLMAR-HQSC	**1.00**	0.42	0.59
SpinAssign	0.24	0.38	0.30
Metabominer	0.84	**0.67**	**0.74**
HMDB	0.38	0.38	0.38
BMRB	0.13	0.13	0.13
ResNet18 (0)	0.67	0.25	0.36
ResNet18 (1)	0.67	0.33	0.44
ResNet18 (2)	0.75	0.375	0.50
ResNet18 (3)	0.92	0.46	0.61
U-Net (0)	0.62	0.33	0.43
U-Net (1)	0.67	0.33	0.44
U-Net (2)	0.70	0.29	0.41
U-Net (3)	0.70	0.29	0.41

Chapter 4

Conclusion and Future Work

We proposed an automated system for identifying metabolites from HSQC spectra of a mixture of compounds. We found that using by a logistic "tagging" model, we could achieve competitive results by F1 score, without needing to use hand-crafted rule sets or using a peak-by-peak query method. This results in a simpler model, with the added advantage that overlapping peaks do not result in conflicting predictions. We used synthesized HSQC mixtures to show that the model could learn the correspondence between peaks and metabolites, and evaluated on experimental spectra to show competitive performance when compared to other methods.

A direction for future work is to recover the mapping from the peaks in a spectrum to its corresponding metabolite, a mapping which was originally generated automatically in the peak-by-peak approach. This should be relatively simple, because it is easier to predict which peaks correspond to a given metabolite than to identify a set of metabolites from several hundred candidates.

In addition, more research should be done to refine the model architecture. CNNs are powerful vision models; however, in this case one of the core assumptions of CNNs (i.e. the multiple-occurrence of features across an image) seems to be violated. Thus it seems a more efficient model could be designed to exploit the specific properties of HSQC spectra.

Finally, one of the most challenging aspects of the task was the very limited amount of available experimental mixture data. Ideally, we would want to train the model not on artificially

synthesized mixtures, but on data collected from actual NMR experimental mixtures. This would enable the model to capture the actual interactions between metabolites in NMR experiments, as well as other side-effects of experimental data such as the presence of noisy peaks. If not enough data could be collected for training, even the use of more data as a held-out dataset would greatly help to tune the model for the down-stream experimental task. Thus more work should be done to collect more known samples of experimental HSQC data.

Bibliography

[1] Stefan Berger, Siegmar Braun, and Hans-Otto Kalinowski. Nmr spectroscopy of the non-metallic elements. *Ed. John Wiley &*, 1997.

[2] Kerem Bingol, Da-Wei Li, Lei Bruschweiler-Li, Oscar A Cabrera, Timothy Megraw, Fengli Zhang, and Rafael Bruschweiler. Unified and isomer-specific nmr metabolomics database for the accurate analysis of 13c–1h hsqc spectra. *ACS chemical biology*, 10(2):452–459, 2015.

[3] John HF Bothwell and Julian L Griffin. An introduction to biological nuclear magnetic resonance spectroscopy. *Biological Reviews*, 86(2):493–510, 2011.

[4] Eisuke Chikayama, Yasuyo Sekiyama, Mami Okamoto, Yumiko Nakanishi, Yuuri Tsuboi, Kenji Akiyama, Kazuki Saito, Kazuo Shinozaki, and Jun Kikuchi. Statistical indices for simultaneous large-scale metabolite detections for a single nmr spectrum. *Analytical Chemistry*, 82(5):1653–1658, 2010.

[5] Carmelo Corsaro, Sebastiano Vasi, Fortunato Neri, Angela Maria Mezzasalma, Giulia Neri, and Enza Fazio. Nmr in metabolomics: From conventional statistics to machine learning and neural network approaches. *Applied Sciences*, 12(6):2824, 2022.

[6] Chiara Damiani, Daniela Gaglio, Elena Sacco, Lilia Alberghina, and Marco Vanoni. Systems metabolomics: From metabolomic snapshots to design principles. *Current opinion in biotechnology*, 63:190–199, 2020.

[7] Abdul-Hamid Emwas, Raja Roy, Ryan T McKay, Leonardo Tenori, Edoardo Saccenti, GA Nagana Gowda, Daniel Raftery, Fatimah Alahmari, Lukasz Jaremko, Mariusz Jaremko, and David S Wishart. Nmr spectroscopy for metabolomics research. *Metabolites*, 9(7):123, 2019.

[8] Shijinqiu Gao, Hoi Yan Katharine Chau, Kuijun Wang, Hongyu Ao, Rency S Varghese, and Habtom W Ressom. Convolutional neural network-based compound fingerprint prediction for metabolite annotation. *Metabolites*, 12(7):605, 2022.

[9] Mason Greer, Cheng Chen, and Soumyajit Mandal. Automated classification of food products using 2d low-field nmr. *Journal of Magnetic Resonance*, 294:44–58, 2018.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[11] Caroline H Johnson, Julijana Ivanisevic, and Gary Siuzdak. Metabolomics: beyond biomarkers and towards mechanisms. *Nature reviews Molecular cell biology*, 17(7):451–459, 2016.

[12] Hyun Woo Kim, Chen Zhang, Garrison W Cottrell, and William H Gerwick. Smart-miner: A convolutional neural network-based metabolite identification from 1h-13c hsqc spectra. *Magnetic Resonance in Chemistry*, 60(11):1070–1075, 2022.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[14] Da-Wei Li, Alexandar L Hansen, Chunhua Yuan, Lei Bruschweiler-Li, and Rafael Brüschweiler. Deep picker is a deep neural network for accurate deconvolution of complex two-dimensional nmr spectra. *Nature communications*, 12(1):5229, 2021.

[15] John L Markley, Eldon L Ulrich, Helen M Berman, Kim Henrick, Haruki Nakamura, and Hideo Akutsu. Biomagresbank (bmrb) as a partner in the worldwide protein data bank (wwpdb): new policies affecting biomolecular nmr depositions. *Journal of biomolecular NMR*, 40:153–155, 2008.

[16] Raphael Reher, Hyun Woo Kim, Chen Zhang, Huanru Henry Mao, Mingxun Wang, Louis-Félix Nothias, Andres Mauricio Caraballo-Rodriguez, Evgenia Glukhov, Bahar Teke, Tiago Leao, Kelsey L Alexander, Brendan M Duggan, Ezra L Van Everbroeck, Pieter C Dorrestein, Garrison W Cottrell, and William H Gerwick. A convolutional neural network-based approach for the rapid annotation of molecularly diverse natural products. *Journal of the American Chemical Society*, 142(9):4114–4120, 2020.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[19] Danhui Wang, Peyton Greenwood, and Matthias S Klein. Deep learning for rapid identification of microbes using metabolomics profiles. *Metabolites*, 11(12):863, 2021.

[20] David S Wishart, Yannick Djoumbou Feunang, Ana Marcu, An Chi Guo, Kevin Liang, Rosa Vázquez-Fresno, Tanvir Sajed, Daniel Johnson, Carin Li, Naama Karu, Zinat Sayeeda, Elvis Lo, Nazanin Assempour, Mark Berjanskii, Sandeep Singhal, David Arndt, Yonjie Liang, Hasan Badran, Jason Grant, Arnau Serra-Cayuela, Yifeng Liu, Rupa Mandal, Vanessa Neveu, Allison Pon, Craig Knox, Michael Wilson, Claudine Manach, and Augustin Scalbert. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research*, 46(D1):D608–D617, 11 2017.

[21] Jianguo Xia, Trent C Bjorndahl, Peter Tang, and David S Wishart. Metabominer–semi-automated identification of metabolites from 2d nmr spectra of complex biofluids. *BMC bioinformatics*, 9(1):1–16, 2008.

[22] Chen Zhang, Yerlan Idelbayev, Nicholas Roberts, Yiwen Tao, Yashwanth Nannapaneni, Brendan M Duggan, Jie Min, Eugene C Lin, Erik C Gerwick, Garrison W Cottrell, and William H Gerwick. Small molecule accurate recognition technology (smart) to enhance natural products research. *Scientific reports*, 7(1):14243, 2017.

[23] Hong Zheng, Jiansong Ji, Liangcai Zhao, Minjiang Chen, An Shi, Linlin Pan, Yiran Huang, Huajie Zhang, Baijun Dong, and Hongchang Gao. Prediction and diagnosis of renal cell carcinoma using nuclear magnetic resonance-based serum metabolomics and self-organizing maps. *Oncotarget*, 7(37):59189, 2016.