



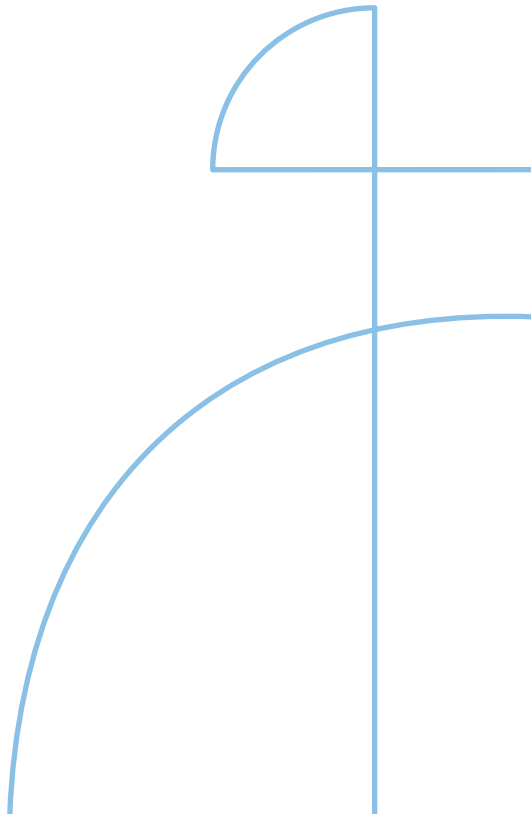
Licentiate Thesis in Machine Learning for Tandem Mass Spectrometry

Learning Representations for Tandem Mass Spectra

Self-Supervised Methods and Inductive Biases

ALFRED NILSSON

KTH ROYAL INSTITUTE OF TECHNOLOGY



Learning Representations for Tandem Mass Spectra

Self-Supervised Methods and Inductive Biases

ALFRED NILSSON

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Licentiate of Engineering on Friday the 17th of April 2026, at 2:00 p.m. in Gamma-6 SciLifeLab, Stockholm.

Licentiate Thesis in Machine Learning for Tandem Mass Spectrometry
KTH Royal Institute of Technology
Stockholm, Sweden 2026

© Alfred Nilsson
© Joel Lapin
© Mathias Wilhelm
© Lukas Käll

TRITA-CBH-FOU-2026:21
ISBN 978-91-8106-586-2

Printed by: Universitetservice US-AB, Sweden 2026

Abstract

Mass spectrometry (MS) is central to modern proteomics, enabling analysis of proteins and peptides based on their mass-to-charge ratio. Tandem mass spectrometry (MS²) encodes peptide fragmentation patterns and forms the basis for sequence identification. While database search has long dominated this process, deep learning has opened new paths for the direct interpretation of spectra. This thesis investigates how neural networks can learn representations of MS² spectra. Two complementary research directions are explored.

First, selected self-supervised pretraining strategies are evaluated through controlled downstream experiments using encoders pretrained on unlabeled MS² corpora. Self-distillation yields global embeddings that implicitly encode aspects of peptide chemical properties, and masked autoencoding provides modest improvements in de novo optimization and accuracy. However, the resulting improvements fall short of state-of-the-art supervised de novo sequencing performance.

Second, we introduce *Pairwise Attention*, a transformer architecture that incorporates a domain-aligned relational inductive bias by conditioning attention on pairwise mass differences between peaks. This yields consistent performance improvements on standard de novo sequencing benchmarks and strong generalization across datasets.

Overall, the results show that self-supervised learning can recover meaningful structure from raw MS² data, while architectural inductive biases currently offer the most robust and reliable gains for de novo peptide sequencing.

Keywords

proteomics, mass spectrometry, de novo sequencing, transformers, representation learning

Sammanfattning

Masspektrometri (MS) är central inom modern proteomik och möjliggör analys av proteiner och peptider baserat på deras massa. Tandem-masspektrometri (MS²) kodar fragmenteringsmönster för peptider och utgör grunden för sekvensidentifiering. Även om databassökning länge har dominerat denna process har djupinlärning öppnat nya möjligheter för direkt tolkning av spektra.

Denna avhandling undersöker hur neurala nätverk kan lära sig representationer av MS²-spektra. Två kompletterande forskningsinriktningar studeras.

Först utvärderas utvalda självövervakade förträningsstrategier genom kontrollerade experiment med encoders som förträns på oetiketterade MS²-korpor. Självdistillation ger globala inbäddningar som implicit kodar aspekter av peptiders kemiska egenskaper, och masked autoencoding ger måttliga förbättringar i de novo-precision. De resulterande förbättringarna når dock inte upp till prestandan hos dagens state-of-the-art-metoder för övervakad de novo-sekvensering.

Sedan introduceras *Pairwise Attention*, en transformerarkitektur som inkorporerar en domänanpassad induktiv bias genom att villkora *Attention* på parvisa masskillnader mellan toppar. Detta ger prestandaförbättringar på etablerade de novo-sekvenseringsbenchmarkar samt stark generalisering över dataset.

Sammantaget visar resultaten att självövervakad inlärning kan återvinna meningsfull struktur ur råa MS²-data, medan induktiva biaser för närvarande erbjuder de mest robusta förbättringarna för de novo-peptidsekvensering.

Nyckelord

proteomik, masspektrometri, de novo-sekvensering, transformers, representation learning

Acknowledgment

I would like to express my sincere gratitude to my advisor, Lukas Käll, for his invaluable guidance, support, and insightful feedback throughout this research.

Sincerely,

A handwritten signature in black ink, appearing to read 'Alfred Nilsson', with a stylized, flowing script.

Alfred Nilsson
Stockholm, March 27, 2026

List of included papers

Contributions This licentiate thesis consists of two parts. The first part provides a general overview of the research area and summarizes the author's contributions. The second part consists of the following peer-reviewed publication:

Paper A **Pairwise Attention: Leveraging Mass Differences to Enhance De Novo Sequencing of Mass Spectra**, Joel Lapin, Alfred Nilsson, Mathias Wilhelm, and Lukas Käll. In *Journal of Proteome Research*, 2025

Paper B **Self-Supervised Learning for Tandem Mass Spectra: Methods, Dynamics, and Downstream Effects**, Alfred Nilsson and Lukas Käll. 2025

Other contributions The author has contributed to the following work, which is not included in this thesis:

1 **Indirectly Parameterized Concrete Autoencoders**, Alfred Nilsson, Klas Wijk, Sai Bharath Chandra Gutha, Erik Englesson, Alexandra Hotti, Carlo Saccardi, Oskar Kviman, Jens Lagergren, Ricardo Vinuesa Motilva, and Hossein Azizpour. In *Proceedings of the 41st International Conference on Machine Learning*, 2024

2 **Regularizing and Interpreting Vision Transformer by Patch Selection on Echocardiography Data**, Alfred Nilsson and Hossein Azizpour. In *Proceedings of the fifth Conference on Health, Inference, and Learning*, 2024

3 **Better Inputs, Better Learning**, Luke Squires, Jose Humberto Giraldez Chavez, Alfred Nilsson, Lukas Käll, and Samuel H Payne. In *Journal of Proteome Research*, 2026

Contents

Abstract	iii
Sammanfattning	v
Acknowledgment	vii
List of included papers	ix
Contents	xi
List of Figures	xiii
List of Tables	xv
I Research Overview	1
1 Introduction	3
1.1 Peptide identification in proteomics	3
1.2 De novo sequencing	4
1.3 Representation learning for MS ²	4
1.4 Incorporating domain knowledge	5
1.5 Thesis scope	5
2 Background	7
2.1 Interpreting Tandem Mass Spectra	7
2.2 From Database Search to De Novo Sequencing	8
2.3 Supervised Deep Learning for Spectral Translation	8
2.4 Representation Learning from MS ² Data	10
2.5 Relational Inductive Biases and Attention	11
2.6 Summary	12
3 Materials and Methods	13
3.1 Transformer Architecture and Peak Representations	13
3.2 De Novo Peptide Sequencing with Transformers	14
3.3 Self-Supervised Pretraining for MS ²	15
3.4 Pairwise Attention for De Novo Sequencing	19
4 Datasets	23

CONTENTS

5 Experiments and Results	25
5.1 Self-Supervised Encoder Pretraining Experiments	25
5.2 Pairwise Attention for De Novo Sequencing	29
6 Conclusions and Future Directions	35
Bibliography	41
II Appended papers	45

List of Figures

1.0.1 Example raw MS ² spectrum. The plot shows a typical set of (<i>m/z</i> , intensity) peaks as provided directly to machine-learning models.	4
2.1.1 Annotated MS ² spectrum. Matched b-ions (blue) and y-ions (red) illustrate how inter-residue mass differences correspond to characteristic fragment peaks.	8
3.3.1 Masked spectrum autoencoder architecture for MS ² pretraining. The encoder processes visible peaks, and the decoder reconstructs the Fourier-encoded representations of masked peaks. Reproduced from [2].	16
3.3.2 View construction for DINO pretraining on MS ² spectra. Global views retain most peaks, while local views are formed by peak subsampling. Reproduced from [2].	17
3.3.3 DINO-style self-distillation applied to MS ² spectra. A student and teacher encoder process different views of the same spectrum, and the student is trained to match the teacher's centered and sharpened output distribution. Reproduced from [2].	18
3.4.4 Schematic of the Pairwise Attention encoder architecture. Pairwise mass difference features are added as biases to self-attention logits in each encoder layer.	20
5.2.1 Training loss dynamics for Base and Pairwise Attention (PA) encoders across species. Curves show de novo sequencing training loss as a function of epoch for the Base transformer and the Pairwise Attention variant. Across all species, PA exhibits faster initial loss reduction and converges to a slightly lower training loss than the Base model, indicating improved optimization behavior under identical training conditions.	30
5.2.2 Precision–coverage curves for the Base transformer, Pairwise Attention (PA), and Casanovo on NineSpecies V1. Only the best of the three PA training seeds is shown.	33
5.2.3 Precision–coverage curves for the Base transformer, Pairwise Attention (PA), and Casanovo on NineSpecies V2. Only the best of the three PA training seeds is shown.	34

List of Tables

5.1.1	Minimum validation loss over 30 fine-tuning epochs.	26
5.1.2	De novo test-set accuracy.	27
5.1.3	Retention-time prediction from frozen spectrum embeddings learned though DINO.	28
5.1.4	Spectral-quality assessment (SQA).	29
5.2.5	Peptide precision at 100% coverage on NineSpecies V1.	31
5.2.6	Peptide precision at 100% coverage on NineSpecies V2.	32
5.2.7	Peptide precision at 100% coverage on NineSpecies V2 after supervised pretraining on MassIVE-KB.	32
5.2.8	Peptide precision at 100% coverage on the external bacterial test set after MassIVE-KB pretraining.	34

Research Overview

1 Introduction

Mass spectrometry (MS) is a central analytical technique in modern proteomics, enabling large-scale identification and quantification of peptides and proteins. In a typical workflow, peptides are ionized—meaning they are given an electrical charge so they can be manipulated by electromagnetic fields—and these charged molecules are referred to as *ions*. The mass spectrometer first isolates a specific set of ions corresponding to a single peptide, known as the *precursor ion*. In tandem mass spectrometry (MS²), these precursor ions are then deliberately broken into smaller *fragment ions* along the peptide backbone. The instrument records the masses and intensities of these fragments as a *spectrum*: a collection of m/z –intensity pairs encoding where the backbone has cleaved.

An example raw MS² spectrum is shown in Figure 1.0.1. Such spectra are typically sparse, feature irregular peak patterns, and span a wide dynamic range, reflecting the fact that different peptide bonds break with very different frequencies and intensities. Interpreting these patterns, meaning inferring which observed peaks originate from which physical peptide fragments, underlies most computational proteomics workflows.

1.1 Peptide identification in proteomics

Traditionally, peptide identification relies on *database search*. Experimental MS² spectra are compared against theoretical spectra generated from candidate peptide sequences, and statistical procedures control the false discovery rate. This approach is highly effective when the underlying proteome is known, but performs poorly in settings such as immunopeptidomics, antibody sequencing, or metaproteomics, where the peptide space is unknown or highly variable.

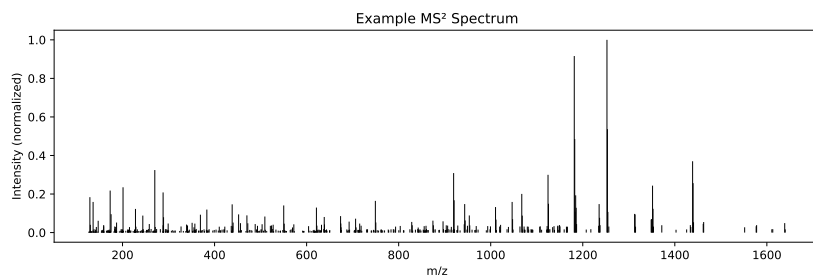


Figure 1.0.1: Example raw MS^2 spectrum. The plot shows a typical set of $(m/z, \text{intensity})$ peaks as provided directly to machine-learning models.

1.2 De novo sequencing

De novo sequencing addresses these limitations by reconstructing peptide sequences directly from MS^2 spectra without external databases. Early methods relied on dynamic programming or graph-based enumeration. Recent work has instead formulated de novo sequencing as a sequence-to-sequence learning problem: a transformer encoder represents the spectrum, and an autoregressive decoder predicts the amino-acid sequence.

Models such as DeepNovo [6], PointNovo [7], and Casanovo [8] demonstrate that deep architectures capture substantial statistical structure in fragmentation spectra and achieve strong performance on curated benchmarks. Despite these advances, current systems depend on large collections of paired spectra and peptide sequences. Although many labeled datasets are available, they inherit biases from the search engines that produced the identifications, and expanding supervised training by simply adding more such data does not eliminate these constraints. This creates motivation to explore methods that learn from unlabeled spectra or that incorporate domain knowledge directly into the model design.

1.3 Representation learning for MS^2

Self-supervised and unsupervised learning have become powerful tools in domains such as vision, language, and genomics, where models learn useful representations from raw data. Applying similar ideas to MS^2 is appealing: a model that captures the latent structure of fragmentation patterns without peptide supervision could, in principle, generalize across instruments, organisms, or acquisition protocols.

A detailed investigation of such self-supervised objectives for MS^2 is presented in a separate manuscript [2]. That work analyses the behavior, optimization dynamics, and downstream effects of large-scale pretraining using millions of

unlabeled spectra. In this thesis, those findings are used only to contextualize the challenges of learning spectral representations, as opposed to imposing structure through architectural design. The full methodological and empirical treatment of self-supervised pretraining is provided in the companion manuscript.

1.4 Incorporating domain knowledge

When generic representation learning proves insufficient for capturing the structure most relevant to peptide reconstruction, an alternative is to embed domain knowledge directly into the model architecture. Manual spectrum interpretation often relies on comparing pairs of peaks whose m/z differences match the masses of amino acids or characteristic fragment ions. This observation motivates the *Pairwise Attention* model [1], which augments the transformer encoder with pairwise mass-difference information. By introducing this inductive bias into the attention mechanism, the model is guided toward fragmentation-consistent relationships that are difficult for standard architectures to infer from raw peak tuples alone. This yields substantial improvements in peptide-level precision on established benchmarks.

1.5 Thesis scope

This thesis examines two complementary strategies for improving MS² representations:

1. **Representation learning** from unlabeled spectra, included here to motivate the challenges of learning representations directly from MS² data. These self-supervised objectives provide a weak but detectable learning signal, and their detailed analysis is presented in a separate companion manuscript [2].
2. **Domain-informed architectural design**, in the form of Pairwise Attention [1], which incorporates mass-difference structure directly into the encoder.

This thesis focuses primarily on supervised de novo sequencing and on the benefits of introducing explicit inductive biases into the architecture. The discussion of self-supervision is used to contextualize the challenges of learning spectral representations in the MS² setting, as opposed to imposing structure through architectural design.

2 Background

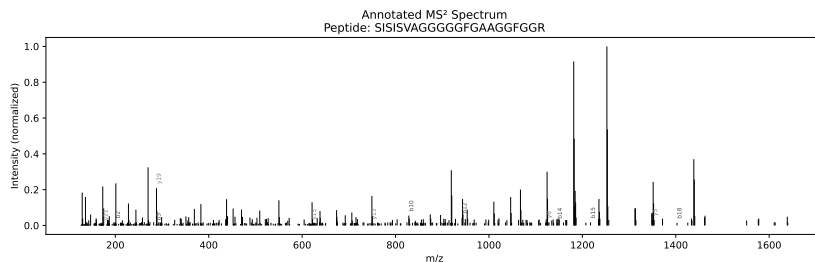
2.1 Interpreting Tandem Mass Spectra

Tandem mass spectrometry (MS^2) is the primary experimental technique used for peptide sequencing in proteomics. In a typical workflow, proteins are first enzymatically digested into shorter peptides prior to mass spectrometric analysis. Proteases such as trypsin are commonly used, as they cleave proteins at well-defined amino acid residues, producing peptides with predictable termini and lengths that are amenable to sequencing.

These peptides are subsequently ionized, isolated according to their precursor mass-to-charge ratio, and fragmented inside the mass spectrometer. The resulting MS^2 spectrum is represented as a set of peaks, each defined by a fragment ion mass-to-charge ratio and an associated intensity. Figure 2.1.1 illustrates how a subset of these peaks can be annotated as b- and y-ions, linking observed mass differences to specific cleavages between adjacent amino acids within the peptide.

Fragmentation predominantly occurs at peptide bonds along the amino acid chain, producing fragment ions that correspond to contiguous prefixes or suffixes of the original peptide sequence. In idealized cases, these fragment ions form b- and y-ion series whose mass differences reflect the underlying amino acid composition. In practice, only a subset of such fragments is observed, and the correspondence between peaks and physical peptide fragments is incomplete.

As a result, spectra are typically sparse and irregular. Many theoretical fragments are not observed, while additional peaks may arise from noise, neutral losses, or chimeric spectra caused by the co-isolation of multiple peptides. Peak



PointNovo [7] extended deep learning-based de novo sequencing by revisiting both the model architecture and the spectrum representation. Rather than discretizing spectra into fixed-length intensity vectors, PointNovo represents each MS² spectrum as an unordered set of $(m/z, \text{intensity})$ peaks. To the best of our knowledge, this was the first de novo sequencing model to operate directly on peak sets, avoiding the binning and discretization schemes used by earlier machine-learning and traditional approaches.

At each decoding step, PointNovo compares observed peaks against theoretical fragment ion masses derived from candidate amino acids and predefined ion types. This comparison yields a mass-difference feature tensor that scores the compatibility between individual observed peaks and expected fragment locations. These features are then aggregated across peaks using a PointNet-style, order-invariant network [13], producing a summary representation used to predict the next amino acid. In this formulation, mass differences serve to align observed peaks with hypothesized fragment ions, rather than to model relationships among observed peaks themselves. This design allows the model to reason explicitly over fragment-level evidence while remaining invariant to peak ordering, and leads to improved de novo sequencing accuracy.

Most recently, Casanovo [8] reframed de novo peptide sequencing as a sequence-to-sequence translation problem using a transformer architecture [14]. Building on the peak-set representation introduced by PointNovo, Casanovo treats each MS² spectrum as a variable-length sequence of $(m/z, \text{intensity})$ peaks and processes them using a standard encoder–decoder transformer.

To encode spectral inputs, Casanovo maps each peak’s m/z value to a high-dimensional sinusoidal embedding, analogous to positional encodings in natural language models, while intensities are embedded through a learned linear projection. These embeddings are summed and passed to the transformer encoder, where self-attention allows the model to learn contextual relationships among peaks across the entire spectrum. The decoder then autoregressively predicts the peptide sequence, conditioned on the encoded spectrum and precursor mass information, using beam search to enforce mass consistency.

By combining a peak-based spectrum representation with a general-purpose transformer architecture, Casanovo demonstrated that self-attention models are well suited to capturing the global structure of fragmentation spectra. When trained on large-scale spectral libraries, Casanovo achieved state-of-the-art de novo sequencing accuracy across multiple species and experimental conditions, and has since become a widely used baseline for deep learning-based de novo sequencing.

Together, these models represent the modern foundation of deep learning-based de novo sequencing and substantially outperform earlier heuristic and graph-based approaches. Much of this progress has been driven by the adoption of increasingly expressive neural architectures, including convolutional

encoders, order-invariant set models, and, most recently, transformer-based sequence-to-sequence models. In each case, performance improvements largely follow from applying general-purpose deep learning architectures to spectral data and scaling them with larger and higher-quality training sets. An important complementary direction is to investigate whether additional gains can be achieved by tailoring representations and model components more closely to the structure of MS² data, either by introducing explicit architectural inductive biases or by designing representation learning objectives that target characteristic spectral regularities.

2.4 Representation Learning from MS² Data

The vast majority of tandem mass spectra in public repositories remain unlabeled, motivating interest in learning spectral representations that generalize beyond specific peptide annotations. In other domains, this problem has been successfully addressed using contrastive and self-supervised representation learning methods, such as SimCLR [15] and MoCo [16], which rely on carefully designed data augmentations to define invariances in the input space. For MS² data, however, defining meaningful, label-free augmentations that preserve peptide identity is nontrivial, complicating direct application of these approaches.

GLEAMS [17] learns spectral embeddings using a weakly supervised contrastive objective. Spectra assigned to the same peptide by a database search engine are treated as positive pairs, while spectra assigned to different peptides form negative pairs. Thus, although GLEAMS does not perform explicit peptide prediction, it still relies on peptide labels obtained from search results to construct its training signal. In contrast to augmentation-based self-supervised methods such as SimCLR [15] or MoCo [16], GLEAMS avoids synthetic data augmentations, reflecting the difficulty of defining label-preserving transformations for MS² spectra. The learned embeddings enable large-scale spectral clustering and fast spectral library search.

yHydra [18] adopts a related but distinct contrastive strategy inspired by CLIP [19]. A transformer-based encoder maps spectra into a shared embedding space, while a separate encoder maps peptide sequences into the same space. The model is trained to align matched spectrum–peptide pairs, enabling rapid approximate search and peptide–spectrum alignment. As with GLEAMS, the construction of positive pairs relies on known peptide annotations, even though the learned representations can later be applied to unlabeled spectra.

A complementary line of work examines representations extracted from the encoder of Casanovo, as explored in a recent preprint framing Casanovo as a *foundation model* for MS² data [20]. In that work, the Casanovo encoder—pretrained for supervised de novo peptide sequencing—is reused as a fixed feature extractor for a range of spectrum-level downstream tasks, including spec-

trum quality assessment, chimericity detection, and prediction of phosphorylation and glycosylation status. While the resulting performance improvements over task-specific baselines are generally modest, the pretrained representations consistently outperform simple spectral baselines and show clear advantages in low-data regimes, where they can match or exceed end-to-end models trained from scratch.

Despite this progress, most existing representation learning approaches for MS² still rely on peptide annotations to define training signals, reflecting the inherent difficulty of formulating self-supervised objectives for fragmentation spectra. A broader investigation of genuinely self-supervised pretraining strategies for MS² data, including masked modeling, self-distillation, and fragment-reconstruction objectives that do not rely on peptide annotations, is presented in a separate companion manuscript [2]. That work demonstrates that meaningful structure can be learned directly from raw spectra at scale, while also characterizing the remaining challenges of such objectives. In this thesis, those results are referenced to contextualize the role of fully self-supervised objectives within the broader landscape of MS² representation learning, and to motivate the complementary use of domain-informed inductive biases in supervised de novo sequencing models.

2.5 Relational Inductive Biases and Attention

Most deep learning architectures operate on collections of elements using token-wise representations, with relationships between elements inferred implicitly through learned interactions. Fully connected layers allow unrestricted interactions between all inputs, but impose only weak structural assumptions. Other common building blocks, such as convolutional or recurrent layers, introduce stronger inductive biases through locality, weight sharing, or temporal structure, which can substantially improve learning efficiency when aligned with the underlying data-generating process.

Battaglia et al. [21] formalize this perspective in terms of *relational inductive biases*, which describe architectural assumptions about how entities interact and how these interactions should be composed. In their framework, many real-world systems lie between two extremes: models that ignore relations entirely, and models that assume all-to-all interactions. Effective architectures often occupy an intermediate regime, where interactions are structured according to domain-specific regularities rather than being absent or unrestricted.

Self-attention mechanisms provide a flexible way to model interactions among elements by allowing each token to attend to all others. In their standard form, however, attention mechanisms do not encode any prior preference for which interactions are likely to be informative. As a result, relational structure must be inferred entirely from data, which can be inefficient when strong and well-characterized interaction patterns are present.

AlphaFold [22] provides an example of incorporating explicit pairwise structure into an attention-based architecture. In that work, the Evoformer maintains a dedicated pairwise representation for residue pairs, initialized from multiple sequence alignment-derived coevolutionary signals and refined through iterative updates. These pairwise features are injected as biases into attention computations over per-residue representations, allowing the model to condition attention weights on learned residue–residue relationships. Their inclusion was shown empirically to improve structure prediction accuracy.

The relevance of this example lies not in the specific biological setting, but in the architectural principle it illustrates. Rather than relying solely on generic all-to-all attention, the model exposes learned pairwise information that reflects known structure in the domain and allows this information to influence how attention is allocated.

In the context of tandem mass spectrometry, fragmentation patterns are governed by well-defined physical constraints, and relationships between pairs of peaks often carry more direct information than individual peaks considered in isolation. Introducing relational inductive biases that reflect these constraints offers a principled way to guide attention toward physically meaningful interactions, while retaining the flexibility and expressiveness of transformer-based models.

2.6 Summary

Deep learning has substantially advanced MS² interpretation, particularly through modern de novo sequencing models based on increasingly expressive neural architectures. Much of this progress has been driven by adopting general-purpose deep learning components and scaling them with large, curated training datasets. At the same time, the structure of tandem mass spectra presents challenges that are not always naturally captured by generic architectures alone.

This thesis examines two complementary responses to this tension. First, it investigates how incorporating domain-informed inductive biases into model architectures can improve supervised de novo sequencing performance by aligning learning dynamics with known physical structure in fragmentation spectra. Second, it situates these contributions within the broader context of representation learning for MS², drawing on a separate companion manuscript to illustrate both the potential and the limitations of large-scale self-supervised learning from raw spectra. Together, these perspectives highlight the trade-off between architectural design and data-driven learning in modeling tandem mass spectrometry data.

3 Materials and Methods

3.1 Transformer Architecture and Peak Representations

All models in this work share a common base transformer encoder backbone that operates directly on the raw MS² spectra. Each input spectrum is treated as a variable-length set of N peaks,

$$S = \{(m_i, I_i)\}_{i=1}^N,$$

where m_i is the mass-to-charge ratio and I_i is the intensity of the i -th peak. We follow a fully set-based formulation and avoid binning or fixed-length vectorization.

Each m_i and I_i is expanded independently into a Fourier positional encoding. Specifically, the scalar $x_i \in \{m_i, I_i\}$ is mapped to a r -dimensional feature vector via sinusoidal basis functions:

$$\phi_{i,p}^{\sin} = \sin\left(\frac{x_i}{\lambda_{\min}/2\pi} \cdot \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{p/r}\right) \quad \text{for } p \leq r/2 \quad (3.1.1)$$

$$\phi_{i,p}^{\cos} = \cos\left(\frac{x_i}{\lambda_{\min}/2\pi} \cdot \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{p/r}\right) \quad \text{for } p > r/2 \quad (3.1.2)$$

where p indexes the frequency component, and λ_{\min} , λ_{\max} are hyperparameters controlling the minimum and maximum wavelengths. These encodings provide smooth, periodic features at multiple frequency scales.

Let r_m and r_I denote the Fourier feature dimensions for m_i and I_i , respectively. The final input token for each peak is constructed by concatenating the encoded

mass and intensity features:

$$t_i = \text{Linear}(\text{Concat}(\phi_i^m, \phi_i^I)),$$

where $t_i \in \mathbb{R}^d$ and the projection maps into the model dimension of the transformer encoder. This yields a sequence $\{t_1, \dots, t_N\} \in \mathbb{R}^{N \times d}$ passed into the standard transformer.

The transformer encoder consists of L blocks of multi-head self-attention and feed-forward layers. No additional position or order encoding is used beyond the Fourier features, as the model operates on spectra as permutation-invariant sets.

3.2 De Novo Peptide Sequencing with Transformers

Modern de novo peptide sequencing methods frame sequence reconstruction from MS/MS spectra as a sequence-to-sequence learning problem. A transformer encoder processes the input spectrum into a sequence of peak-level embeddings, and an autoregressive decoder predicts the amino-acid sequence conditioned on these embeddings.

We implement an encoder-decoder transformer model for autoregressive peptide generation from MS/MS spectra. The encoder maps the input peak tokens $\{t_1, \dots, t_N\}$ to a sequence of peak-level embeddings

$$H = (h_1, \dots, h_N),$$

and the decoder generates a peptide token-by-token using causal self-attention and cross-attention over all encoder embeddings. This setup follows the architecture of Casanovo [8] and is also used as the baseline for our Pairwise Attention model.

Because the decoder attends directly to the full token sequence H rather than to a pooled representation z , performance depends on the quality of the token-level (dense) encoder representations.

Training uses teacher forcing with a standard cross-entropy objective:

$$\mathcal{L}_{\text{CE}} = - \sum_{t=1}^T \log p_{\theta}(y_t | y_{<t}, H),$$

where y_t is the ground-truth amino acid at step t .

Inference is performed using greedy decoding or beam search. Following prior work [6, 8], evaluation metrics include amino-acid precision, peptide-level precision, coverage, and precision–coverage curves computed from a confidence score derived from the decoder’s output probabilities:

- **Amino acid precision:** mass tolerance < 0.1 Da and prefix or suffix mass error < 0.5 Da.
- **Peptide precision:** $N_{\text{match}}/N_{\text{total}}$.
- **Coverage:** $N_{\text{total}}/N_{\text{spec}}$.
- **Precision–coverage curve:** confidence defined as the mean amino acid softmax probability, with spectra assigned confidence -1 if the precursor mass error exceeds 50 ppm.

This formulation serves as the reference de novo sequencing setup throughout the remainder of this thesis. Subsequent sections examine how different representation learning strategies and architectural modifications affect encoder representations within this framework.

3.3 Self-Supervised Pretraining for MS²

The majority of tandem mass spectra available in public repositories lack peptide annotations, motivating interest in self-supervised pretraining strategies that can learn directly from raw MS² data. Unlike domains such as vision or language, however, defining meaningful self-supervised objectives and label-preserving augmentations for fragmentation spectra is nontrivial. A detailed and systematic investigation of this problem is presented in *Self-Supervised Learning for Tandem Mass Spectra: Methods, Dynamics and Downstream Effects* [2], which studies multiple pretraining objectives and their optimization behavior at scale.

Here, we summarize the core objectives and architectures at a high level to provide context.

3.3.1 Masked Spectrum Autoencoding

Masked spectrum autoencoding adapts masked autoencoder[23] objectives to MS² spectra. A subset of peaks is removed from the input spectrum, and the encoder processes the remaining peaks into a latent representation. A lightweight decoder then predicts the Fourier-encoded representations of the missing peaks.

Rather than regressing raw m/z and intensity values, the model predicts their sinusoidal Fourier features as defined in Equations 3.1.1 and 3.1.2. Reconstruction is trained using a mean squared error loss over the masked peaks:

$$\mathcal{L}_{\text{AE}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\hat{\phi}_i - \phi_i\|^2,$$

where \mathcal{M} denotes the set of masked peaks.

This objective encourages the encoder to capture local and contextual relationships between peaks that are predictive of missing spectral content. The masked autoencoder architecture is illustrated in Figure 3.3.1.

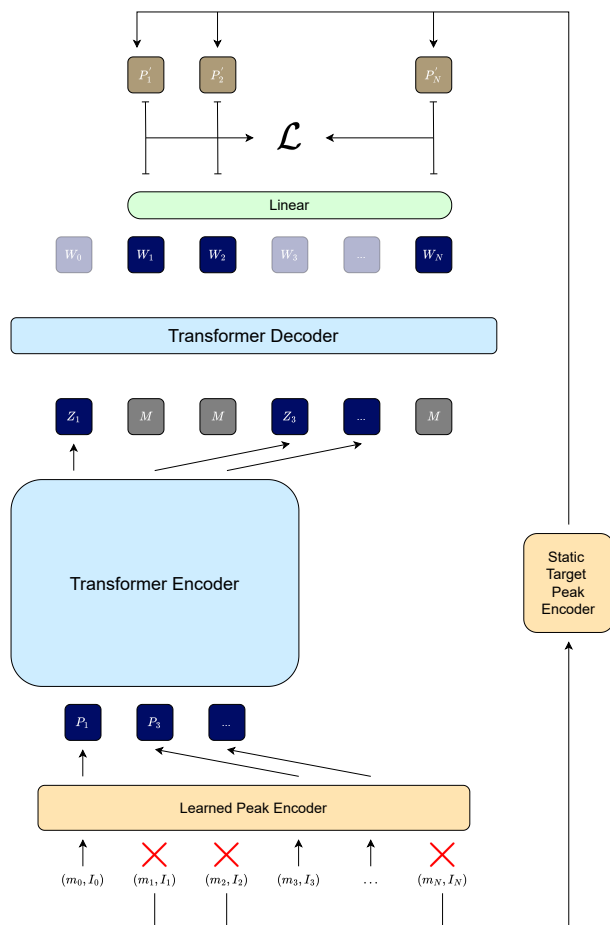


Figure 3.3.1: Masked spectrum autoencoder architecture for MS^2 pretraining. The encoder processes visible peaks, and the decoder reconstructs the Fourier-encoded representations of masked peaks. Reproduced from [2].

3.3.2 Self-Distillation via DINO

We also adapt the DINO self-distillation framework [24] to MS^2 spectra. In this setup, a student and teacher transformer encoder process different subsampled

views of the same spectrum. The student is trained to match the teacher’s output distribution, while the teacher parameters are updated as an exponential moving average of the student.

For each spectrum, multiple views are generated by randomly subsampling peaks. Two global views retain most peaks, while several local views contain smaller subsets. The teacher processes only global views, while the student processes both global and local views. A pooling head aggregates peak-level embeddings into a single spectrum-level representation, which is then passed through a projection head for the distillation loss. The view construction is shown in 3.3.2.

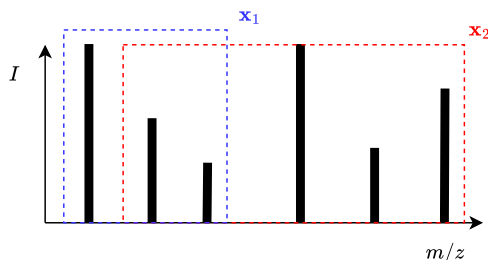


Figure 3.3.2: View construction for DINO pretraining on MS² spectra. Global views retain most peaks, while local views are formed by peak subsampling. Reproduced from [2].

The training objective minimizes the cross-entropy between the student and teacher output distributions after centering and temperature scaling:

$$\mathcal{L}_{\text{DINO}} = - \sum_{k=1}^K p_k^{(t)} \log p_k^{(s)}.$$

As in prior work, this objective is susceptible to several collapse modes, including entropy collapse and convergence to constant representations. In practice, these failure modes can be mitigated through the use of teacher momentum updates, output centering, temperature scheduling, and sufficiently distinct input views. A detailed analysis of these dynamics and stabilization mechanisms is provided in the companion manuscript [2]. The DINO-style architecture is shown in Figure 3.3.3.

Context within this thesis. These self-supervised objectives demonstrate that meaningful structure can be learned directly from raw MS² spectra at scale, while also exposing challenges that arise in the absence of peptide supervision.

In this thesis, self-supervised pretraining is included to contextualize the representation learning landscape for MS² data. The primary focus remains on supervised de novo sequencing and on the role of domain-informed architectural inductive biases, which provide a complementary and more targeted mechanism for improving peptide reconstruction performance.

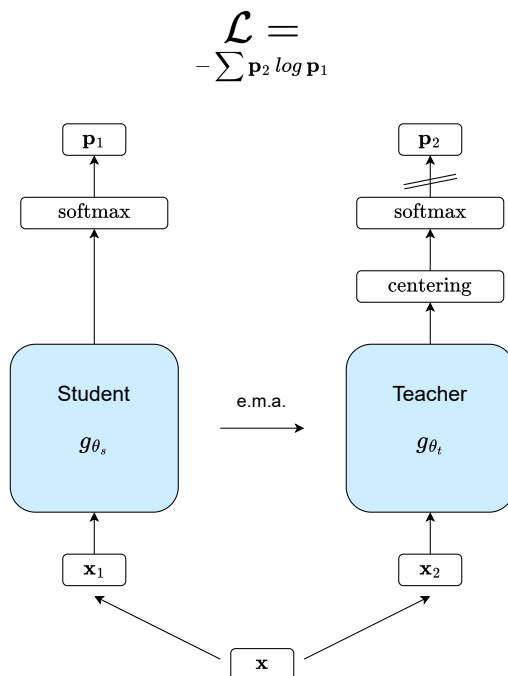


Figure 3.3.3: DINO-style self-distillation applied to MS² spectra. A student and teacher encoder process different views of the same spectrum, and the student is trained to match the teacher’s centered and sharpened output distribution. Reproduced from [2].

3.3.3 Downstream Probing Tasks

To assess the utility of the learned spectral embeddings, we consider a small set of downstream tasks. These tasks serve to characterize what information is captured by self-supervised representations, and to illustrate how different objectives emphasize token-level versus pooled spectrum-level features. They are not a primary focus of this thesis, but provide useful context for interpreting the behavior of the pretraining objectives.

De Novo Peptide Sequencing

De novo peptide sequencing provides a demanding downstream task that depends directly on the quality of token-level spectral representations. We evaluate pretrained encoders by initializing the encoder of a standard encoder-decoder Transformer model for autoregressive peptide generation from MS/MS spectra. The decoder attends to the full sequence of encoder outputs and predicts amino acids sequentially, following the setup used in DeepNovo and Casanovo [6, 8].

Because the decoder operates on the unpooled encoder outputs, performance on this task is particularly sensitive to the quality of dense, peak-level embeddings produced by the encoder.

Spectral Quality Assessment

Spectral quality prediction is framed as a binary classification task, with labels derived from a database-search pipeline. A linear classifier operates on the pooled encoder representation z , probing whether global spectral summaries encode information related to fragment completeness and noise. This task aligns naturally with global self-supervised objectives such as DINO, which supervise pooled representations directly.

Retention Time Prediction

Retention time prediction is formulated as a regression task using the pooled encoder embedding z . Although retention time is observed experimentally, this task probes whether learned embeddings capture physicochemical properties correlated with peptide behavior in chromatography. As with spectral quality prediction, this task depends only on global representations.

3.4 Pairwise Attention for De Novo Sequencing

We enhance the transformer encoder with Pairwise Attention (PA), introduced in *Pairwise Attention: Leveraging Mass Differences to Enhance De Novo Sequencing of Mass Spectra* [1]. PA introduces a domain-specific relational inductive bias into the self-attention operation. Standard self-attention treats all pairwise interactions between peaks as equally plausible a priori, requiring the model to infer relevant relationships implicitly from data. In tandem mass spectrometry, however, fragmentation patterns are governed by structured physical constraints, and mass differences between peaks often correspond directly to amino-acid residues or characteristic fragment transitions.

Pairwise Attention incorporates this structure by making pairwise m/z differences between observed peaks available to the attention mechanism as a learned bias. Rather than imposing a fixed neighborhood or hard-coded interaction pattern, the pairwise mass differences are embedded and transformed through

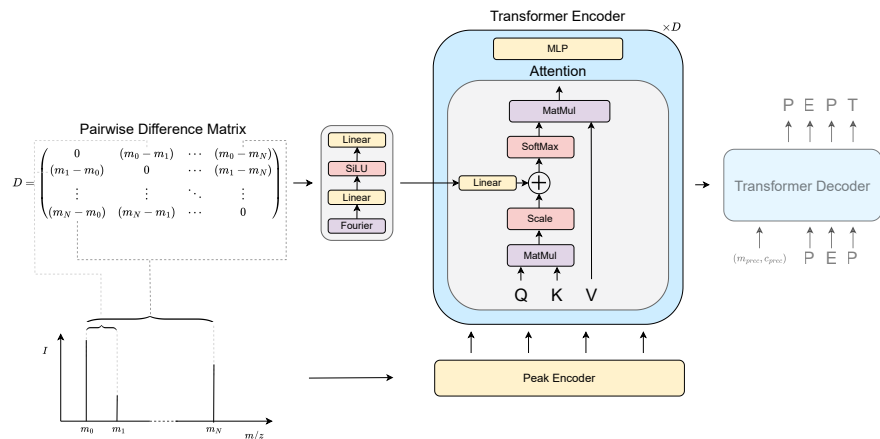


Figure 3.4.4: Schematic of the Pairwise Attention encoder architecture. Pairwise mass difference features are added as biases to self-attention logits in each encoder layer.

learnable layers before being added to the attention logits. This allows the model to learn which mass-difference relationships are informative, while guiding attention toward physically meaningful interactions aligned with fragmentation behavior. The result is a transformer encoder that remains fully expressive, but is biased toward fragmentation-consistent relationships for downstream *de novo* peptide sequencing.

Let a spectrum be represented as a set of N peaks, $S = \{(m_i, I_i)\}_{i=1}^N$. Following the same input processing as in our standard transformer encoder, each (m_i, I_i) is encoded via sinusoidal Fourier features and linearly projected into an embedding space $\mathbf{X} \in \mathbb{R}^{N \times d}$.

Pairwise Mass Difference Features

We define the pairwise mass difference matrix:

$$\Delta m_{ij} = m_i - m_j \in \mathbb{R}^{N \times N}.$$

Each Δm_{ij} is embedded using Fourier positional encoding:

$$\Phi_{i,j,p}^{\sin} = \sin \left(\Delta m_{ij} \cdot \frac{2\pi}{\lambda_{\min}} \cdot \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right)^{p/r'} \right), \quad p \leq r'/2, \quad (3.4.3)$$

$$\Phi_{i,j,p}^{\cos} = \cos \left(\Delta m_{ij} \cdot \frac{2\pi}{\lambda_{\min}} \cdot \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right)^{p/r'} \right), \quad p > r'/2, \quad (3.4.4)$$

where r' is the pairwise Fourier feature dimension, and λ_{\min} , λ_{\max} define the wavelength range. The concatenated result $\Phi_{i,j} \in \mathbb{R}^{r'}$ forms the input to the bias module.

Bias Injection into Self-Attention

The encoder uses standard scaled dot-product attention with queries \mathbf{Q} , keys \mathbf{K} , and values \mathbf{V} computed as:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V.$$

We define a learned MLP $f_{\text{pw}}: \mathbb{R}^{r'} \rightarrow \mathbb{R}^r$ for processing pairwise features:

$$f_{\text{pw}}(\Phi_{i,j}) = \mathbf{W}_2 \cdot \text{SiLU}(\mathbf{W}_1 \Phi_{i,j} + \mathbf{b}_1) + \mathbf{b}_2.$$

This is followed by a linear transformation $g^{(l)}$ for each encoder layer l :

$$g^{(l)}(z_{i,j}) = \mathbf{U}^{(l)} z_{i,j} + \mathbf{c}^{(l)}.$$

The pairwise attention bias $b_{ij}^{(l)}$ is then added to the attention logits:

$$A_{ij}^{(l)} = \frac{\mathbf{Q}_i \cdot \mathbf{K}_j^\top}{\sqrt{d}} + b_{ij}^{(l)},$$

$$\text{where } b_{ij}^{(l)} = g^{(l)}(f_{\text{pw}}(\Phi_{i,j})).$$

This bias is applied across all heads and layers, and added before the softmax in the attention mechanism. The decoder remains unmodified and autoregressively generates amino acid sequences via cross-attention.

This augmentation introduces a small number of additional parameters and modifies only the attention logits, leaving the overall transformer architecture and decoder unchanged.

Relation to Pairwise Biases in AlphaFold

Pairwise Attention is conceptually related to the use of pairwise representations in AlphaFold [22]. In AlphaFold, explicit residue–residue features are maintained alongside per-residue embeddings within the Evoformer architecture. These pair representations are initialized from multiple sequence alignments, updated through dedicated pair-update blocks, and linearly projected to produce biases that are added to attention logits, allowing residue–residue relationships to influence attention patterns during structure prediction.

In Pairwise Attention, pairwise information is constructed directly from the MS² input. For each pair of peaks, the mass difference Δm_{ij} is encoded using fixed Fourier features and transformed by the MLP and layer-specific linear projections described above to produce attention biases. These biases are injected into the self-attention logits at each encoder layer, without maintaining a separate pairwise state. The pairwise signal therefore operates directly on observed spectra and reflects physically grounded relationships between input tokens, as mass differences are tightly linked to peptide fragmentation.

In this sense, Pairwise Attention occupies an intermediate regime between generic self-attention and fully specified relational models. It preserves the expressiveness of transformer architectures while injecting domain-aligned information into the attention mechanism.

4 Datasets

We utilize a diverse set of mass spectrometry datasets spanning supervised benchmarks and large-scale unsupervised corpora.

NineSpecies V1 and NineSpecies V2. We follow the nine-species benchmark design introduced by Casanovo. NineSpecies V1 contains 1.5 million spectra and is distributed via Zenodo [25], while NineSpecies V2 is a more recent version curated for improved PSM quality, containing 2.8 million spectra. The V2 version is publicly hosted at MassIVE and contains raw and peak-picked data across nine diverse species. For each version, we parsed all modified peptides to generate amino acid vocabularies.

NineSpecies InstaUpdated. For some preliminary experiments we used the NineSpecies InstaUpdated subset hosted within the InstaNovo benchmark suite [26]. This version provides the nine-species data in a standardized, easy-to-load format and contains approximately 2.8 million MS² spectra. Although its exact relationship to the curated V1/V2 benchmarks is not formally documented, its scale and composition are consistent with a V2-sized corpus. This dataset was not used in any published research.

MassIVE-KB. We used the full MassIVE-KB corpus as a supervised training set. The data comprise approximately 30 million peptide-spectrum matches (PSMs), derived from the MassIVE-KB v1 spectral library, and represent one of the largest publicly available proteomics corpora [27]. PSMs were collected by selecting up to the top 100 hits for each peptidofrom-charge combination across the spectral library. Spectra were parsed using the official metadata catalog and downloaded in annotated MGF format. We reserved 98.75% of the

data for training, 1% for validation, and 0.25% for testing. All downstream evaluation was performed exclusively on external benchmarks to ensure generalization.

BactTrain (PXDO10000) and BactTest (PXDO10613). We used the BactTrain (PXDO10000) corpus, containing approximately 9.3 million spectra from nearly one million unique peptides, as an unsupervised pretraining source. Curated and published by Payne et al. [28], the dataset spans 51 organisms and includes high-quality spectra processed through standardized pipelines: trypsin digestion, LC-MS/MS acquisition, and MSGF+ search via PNNL’s DMS infrastructure. Its taxonomic breadth and scale make it a strong candidate for learning general-purpose spectral representations.

As an evaluation benchmark, we used BactTest (PXDO10613), a bacterial-only dataset introduced in a metaproteomics study of soil microbiomes [29]. It includes only oxidized methionine as a variable modification and excludes fixed modifications, simplifying the fragmentation landscape. With minimal overlap with MassIVE-KB and other training corpora, it provides a rigorous test of generalization and was previously used as a held-out benchmark in the Pairwise Attention study.

5 Experiments and Results

This chapter presents an empirical evaluation of the modeling choices introduced in this thesis. We first study the effect of self-supervised encoder pretraining on downstream de novo peptide sequencing, isolating initialization effects under controlled fine-tuning conditions. We additionally probe the global embeddings learned through DINO using spectral quality assessment and retention-time prediction. We then evaluate the impact of Pairwise Attention as an architectural inductive bias across multiple supervised regimes, including the standard nine-species V1 and V2 benchmarks, large-scale supervised pretraining on MassIVE-KB, and evaluation on an independent bacterial test set with minimal peptide overlap with the training data. Together, these experiments examine complementary attempts to improve de novo sequencing performance, one focusing on learning spectral representations through self-supervised pretraining, and the other introducing a relational inductive bias suited to peptide MS² spectra at the architectural level.

5.1 Self-Supervised Encoder Pretraining Experiments

This section evaluates how different self-supervised encoder pretraining strategies affect downstream transformer-based de novo peptide sequencing. All comparisons are performed under a controlled setting: identical encoder and decoder architectures, identical fine-tuning hyperparameters, identical peak preprocessing, and identical training, validation, and test splits. The only factor varied across experiments is the encoder initialization, allowing the effects of pretraining to be interpreted as differences in initialization quality rather than architectural or optimization changes.

All fine-tuning experiments in this section use a deliberately simple and largely untuned optimization setup. Hyperparameters were not exhaustively optimized.

Instead, we selected a stable configuration that produced reasonable convergence for a randomly initialized encoder (“scratch”) while prioritizing training throughput, which biases the comparison slightly against encoder pretraining. In particular, a relatively large batch size was used to accelerate experimentation, despite being known to be suboptimal.

All self-supervised encoder pretraining experiments reported in this chapter use the BactTrain (PXD010000) corpus as the unlabeled training source. This dataset contains approximately 9.3 million MS² spectra spanning 51 bacterial organisms and was used *without* peptide annotations during pretraining. Only raw peak lists and precursor mass and charge were provided to the encoder, ensuring that all learning signals arose solely from spectral structure rather than identification metadata.

5.1.1 Effect on De Novo Sequencing Optimization

All de novo fine-tuning experiments in this section are conducted on the Nine-Species InstaUpdated dataset. The encoder is initialized from a self-supervised checkpoint trained on BactTrain (PXD010000), and the full encoder–decoder model is then fine-tuned in a supervised manner. Following common practice in the nine-species benchmark, spectra from *yeast* are held out for evaluation: half are used as a validation set for model selection, and the remaining half form the test set. All other species are used for supervised training.

For each self-supervised objective, encoder checkpoints were selected based on the most appropriate pretraining validation signal for that method. For masked objectives, the checkpoint with the lowest validation loss was used. For DINO, where the training loss is not a reliable indicator of representation quality, checkpoint selection followed the procedure described in [2]. All downstream results reported here use these selected checkpoints; full details of pretraining dynamics and selection criteria are provided in the pretraining manuscript.

We first examine fine-tuning behavior on the de novo sequencing task by comparing validation loss trajectories under identical optimization schedules. Table 5.1.1 reports the minimum validation loss achieved over 30 fine-tuning epochs.

Table 5.1.1: Minimum validation loss over 30 fine-tuning epochs.

Initialization	Min. Val. Loss
Scratch	1.776
MAE	0.981
DINO	1.666

For MAE, the selected encoder corresponds to the checkpoint achieving the lowest validation reconstruction loss during pretraining. For DINO, the selected

Table 5.1.2: De novo test-set accuracy.

Initialization	AA Acc.	Pep. Prec.@100%
Scratch	0.293	0.035
MAE	0.581	0.338
DINO	0.319	0.047

encoder is the checkpoint with highest linear-probe performance, as described in [2]. For the scratch baseline, the encoder is randomly initialized.

Across both data regimes, MAE-pretrained encoders converge faster and reach substantially lower validation losses than scratch and DINO initializations. DINO exhibits high sensitivity to checkpoint choice, with some pretrained states offering mild benefit and others degrading optimization.

5.1.2 Final De Novo Sequencing Accuracy

We next evaluate held-out test-set performance using amino-acid accuracy and peptide-level precision at 100% coverage. Results are shown in Table 5.1.2.

These results mirror the validation-loss analysis. MAE improves both amino-acid accuracy and peptide-level precision. The improvement is more pronounced at the peptide level, which aggregates errors across the entire autoregressive sequence and therefore exhibits threshold-like behavior. Once amino-acid accuracy enters a favorable regime, small additional gains lead to large increases in fully correct peptide reconstructions. Amino-acid accuracy provides a more faithful view of relative model performance.

DINO offers small improvements, reflecting a mismatch between its pretraining objective and the dense, token-level requirements of de novo decoding.

All results above are obtained under a fixed and deliberately simple fine-tuning configuration. More aggressive hyperparameter optimization can substantially improve scratch baselines, thereby reducing the relative gains from encoder pretraining. Therefore, these experiments should be interpreted as controlled ablations on encoder initialization quality rather than as claims about absolute model performance.

In contrast, the Pairwise Attention results presented in section 5.2 demonstrate improvements at competitive, state-of-the-art performance levels across multiple datasets and training regimes, highlighting the impact of introducing domain-aligned architectural inductive bias.

5.1.3 Downstream Evaluation of Global Spectrum Embeddings

In addition to de novo sequencing, we evaluate whether self-supervised pretraining yields useful *global* spectrum representations for downstream tasks that operate on a single pooled embedding. These experiments are intended as diagnostic probes of representation quality rather than as primary application results.

A key distinction among the pretraining objectives is that DINO explicitly supervises a pooled spectrum-level embedding, whereas MAE primarily shapes token-level representations. Accordingly, the experiments in this section focus on DINO-pretrained encoders.

Unless otherwise noted, the encoder is kept frozen and a small MLP head is trained on top of the pooled embedding, isolating the quality of the learned representation.

Retention-Time Prediction

Retention time (RT) correlates with peptide physicochemical properties such as hydrophobicity and length. We therefore use RT prediction as a probe for whether a pooled spectrum embedding captures broad, chemically meaningful structure.

We attach a small MLP regressor to *frozen* embeddings and report the coefficient of determination (R^2) on a held-out test set. Binned-spectrum and random baselines are included for reference. All transformer embeddings in Table 5.1.3 are obtained from *DINO-pretrained* encoders; the comparison isolates the effect of increasing encoder capacity and input peak budget under the same self-supervised pretraining setup.

Table 5.1.3: Retention-time prediction from frozen spectrum embeddings learned though DINO.

Embedding	Params	Dim	Peaks	RT R^2
Binned (512d)	–	512	–	0.693
Random (512d)	–	512	–	–0.282
Small Transformer	19M	512	150	0.805
Large Transformer	127M	1024	150	0.832
Large Transformer	127M	1024	300	0.842
Large Transformer	127M	1024	1000	0.852

DINO-pretrained embeddings substantially outperform the random and binned-spectrum baselines. Within the DINO-pretrained family, increasing model ca-

capacity and allowing more peaks both improve RT prediction, indicating that this self-supervised setup scales with encoder size and input resolution.

Spectral-Quality Assessment

Spectral-quality assessment (SQA) is framed as a binary classification task distinguishing high-confidence spectra from low-confidence spectra using labels derived from database search. As with RT prediction, we evaluate frozen embeddings with a small MLP classifier. For reference, we also report fully fine-tuned models where the encoder and head are optimized jointly.

Table 5.1.4: Spectral-quality assessment (SQA).

Model	AUC	Accuracy
Finetune Scratch	0.777	0.707
Finetune DINO	0.801	0.724
Frozen DINO	0.776	0.705
Binned Spectrum	0.770	0.700

As shown in Table 5.1.4, frozen DINO embeddings modestly outperform the binned-spectrum baseline and perform similarly to fully fine-tuned scratch models. While the absolute gains are small, they provide evidence that self-supervised pretraining induces a nontrivial global signal related to spectral quality. The result is not a strong application outcome, but rather as confirmation that some meaningful structure is captured in the pooled representation.

Summary. Across both retention-time prediction and spectral-quality assessment, DINO-pretrained embeddings provide global structure without task-specific supervision. These results support the use of DINO for learning pooled spectrum representations, while reinforcing that dense downstream tasks such as de novo sequencing benefit more strongly from objectives and architectures that explicitly supervise or encode token-level structure.

5.2 Pairwise Attention for De Novo Sequencing

Pairwise Attention (PA) was introduced in our earlier work as a lightweight architectural bias for transformer-based de novo sequencing. The mechanism augments standard self-attention with pairwise *mass-difference* features, allowing the model to attend not only to fragment intensities but also to the physical relationships between peaks. This bias injects domain knowledge that cannot be easily learned from raw tuples alone, and empirically produces higher peptide-level accuracy at negligible parameter cost.

While the encoder used in the PA architecture is structurally compatible with the pretraining methods evaluated earlier, we restrict the analysis to the conventional supervised training protocol from previous work[6, 8].

5.2.1 Optimization Dynamics

To isolate the effect of Pairwise Attention on optimization, we perform an architectural ablation comparing a baseline transformer model without pairwise bias (denoted **Base**) and the PA model under identical training conditions. Figure 5.2.1 shows the de novo sequencing training loss across species.

Across all cases, Pairwise Attention reduces training loss more rapidly during early optimization and converges to a lower final value. Because the models differ only in the presence of the pairwise attention bias, this behavior indicates that PA improves the conditioning of the learning problem.

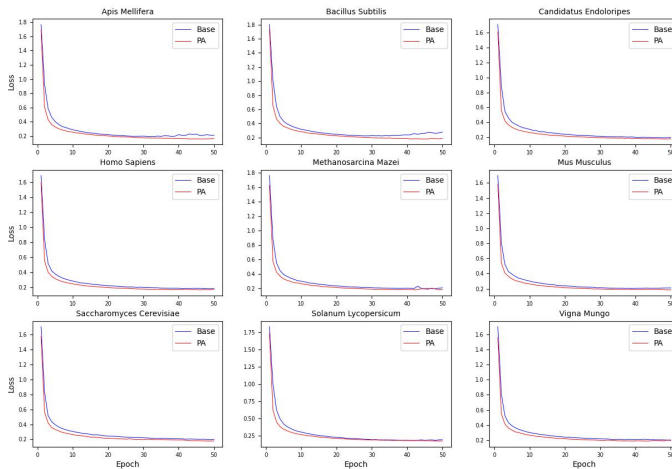


Figure 5.2.1: Training loss dynamics for Base and Pairwise Attention (PA) encoders across species. Curves show de novo sequencing training loss as a function of epoch for the Base transformer and the Pairwise Attention variant. Across all species, PA exhibits faster initial loss reduction and converges to a slightly lower training loss than the Base model, indicating improved optimization behavior under identical training conditions.

5.2.2 Nine-species Evaluation

We first evaluate PA and Base on NineSpecies V1 and NineSpecies V2. All models are trained from scratch on the corresponding training data for each version. For comparison we include the published performance of Casanovo[8] (without beam search).

Figures 5.2.2 and 5.2.3 show the average precision–coverage curves across species. Tables 5.2.5 and 5.2.6 report peptide precision at 100% coverage.

V1. On NineSpecies V1, PA improves upon the Base model by an average of 5.9 percentage points (11.3% relative). These gains occur consistently across all nine species. Compared to Casanovo, PA yields a 3.6 percentage-point improvement on average. Given that PA adds only 29.5k parameters to a 19M-parameter encoder, the magnitude and consistency of the improvement underscores the utility of incorporating mass-difference structure directly into the attention mechanism.

Table 5.2.5: Peptide precision at 100% coverage on NineSpecies V1.

Species	Base	PA	Casanovo
<i>Apis Mellifera</i>	0.390	0.463 (0.004)	0.433
<i>Bacillus Subtilis</i>	0.536	0.612 (0.009)	0.573
<i>Candidatus Endoloripes</i>	0.357	0.409 (0.002)	0.390
<i>Homo Sapiens</i>	0.340	0.391 (0.003)	0.383
<i>Methanosarcina Mazei</i>	0.503	0.554 (0.002)	0.515
<i>Mus Musculus</i>	0.433	0.472 (0.003)	0.431
<i>Solanum Lycopersicum</i>	0.509	0.590 (0.008)	0.522
<i>Saccharomyces Cerevisiae</i>	0.537	0.612 (0.009)	0.580
<i>Vigna Mungo</i>	0.570	0.625 (0.002)	0.552
Average	0.464	0.523 (0.004)	0.487

V2. On the higher-confidence NineSpecies V2dataset, PA again improves upon Base, now by 6.7 percentage points on average (14.2% relative). The comparison to Casanovo is more heterogeneous: PA outperforms Casanovo for several species but trails for others. This variability likely reflects differences in preprocessing, scoring conventions, and updated data quality, as also noted in earlier analyses and detailed in our manuscript [1]. Importantly, the *Base vs. PA* comparison remains stable and consistent across V1 and V2.

Table 5.2.6: Peptide precision at 100% coverage on NineSpecies V2.

Species	Base	PA	Casanovo
<i>Apis Mellifera</i>	0.390	0.446 (0.009)	0.456
<i>Bacillus Subtilis</i>	0.494	0.583 (0.006)	0.538
<i>Candidatus Endoloripes</i>	0.356	0.425 (0.006)	0.468
<i>Homo Sapiens</i>	0.451	0.521 (0.004)	0.533
<i>Methanosarcina Mazei</i>	0.509	0.579 (0.012)	0.529
<i>Mus Musculus</i>	0.410	0.435 (0.004)	0.395
<i>Solanum Lycopersicum</i>	0.543	0.623 (0.007)	0.608
<i>Saccharomyces Cerevisiae</i>	0.558	0.631 (0.010)	0.561
<i>Vigna Mungo</i>	0.525	0.598 (0.022)	0.428
Average	0.471	0.538 (0.009)	0.502

5.2.3 Supervised Pretraining on MassIVE-KB

Table 5.2.7: Peptide precision at 100% coverage on NineSpecies V2 after supervised pre-training on MassIVE-KB.

Test Set	Base	PA	Casanovo
<i>Apis Mellifera</i>	0.618	0.640	0.662
<i>Bacillus Subtilis</i>	0.706	0.732	0.778
<i>Candidatus Endoloripes</i>	0.525	0.549	0.656
<i>Homo Sapiens</i>	0.737	0.746	0.740
<i>Methanosarcina Mazei</i>	0.700	0.720	0.710
<i>Mus Musculus</i>	0.563	0.579	0.552
<i>Solanum Lycopersicum</i>	0.735	0.751	0.799
<i>Saccharomyces Cerevisiae</i>	0.765	0.784	0.840
<i>Vigna Mungo</i>	0.753	0.783	0.762
Average	0.678	0.698	0.722

To investigate performance in a large-data regime, we next train the Base and PA models on MassIVE-KB, containing approximately 30 million PSMs. We then evaluate the resulting models on NineSpecies V2.

Table 5.2.7 shows that PA again improves upon Base, but the gap narrows (from +6.7 points to +2.0 points), as expected when the supervised dataset becomes extremely large. Nevertheless, a nontrivial improvement remains, indicating that pairwise attention captures inductive structure that complements brute-force scaling of the training set.

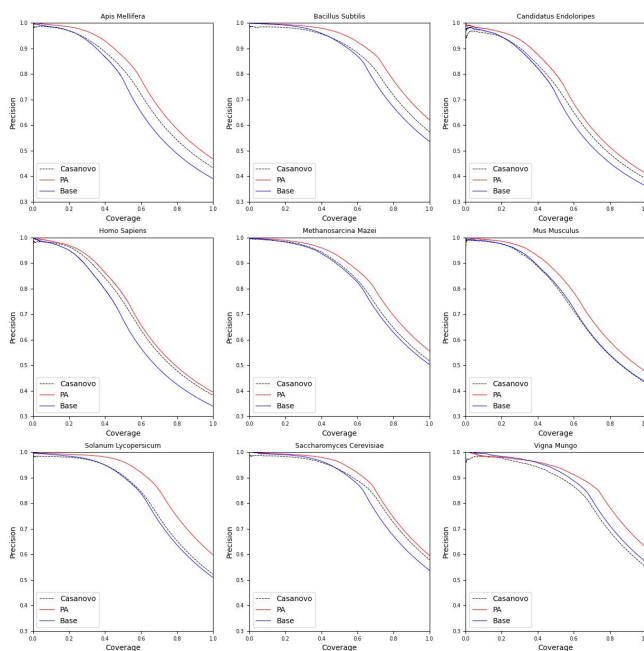


Figure 5.2: Precision–coverage curves for the Base transformer, Pairwise Attention (PA), and Casanovo on NineSpecies V1. Only the best of the three PA training seeds is shown.

5.2.4 Evaluation on an External Bacterial Dataset

To assess generalization beyond the benchmark species, we evaluate the MassIVE-KB-pretrained models on BactTest (PXD010613). This dataset differs substantially from MassIVE-KB in species composition and contains only oxidized methionine as a variable modification, reducing concerns about training–test leakage.

Table 5.2.8 shows that PA improves over Base by +2.5 percentage points, consistent with the NineSpecies V2 result. Strikingly, both Base and PA substantially outperform Casanovo’s published MassIVE-KB model, with PA achieving an 8.5-point (24% relative) improvement.

Summary. Across all settings, training from scratch on NineSpecies V1/NineSpecies V2, large-scale supervised pretraining on MassIVE-KB, and evaluation on an external bacterial dataset, Pairwise Attention consistently improves peptide-level

CHAPTER 5. EXPERIMENTS AND RESULTS

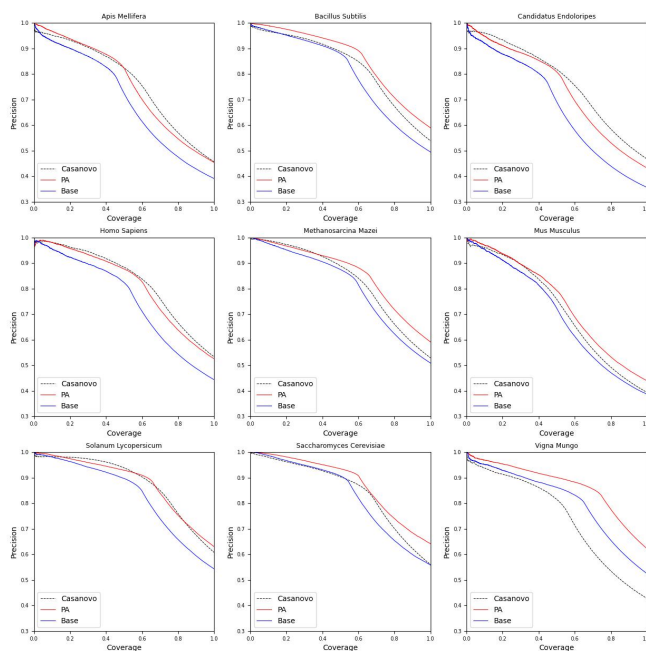


Figure 5.2.3: Precision–coverage curves for the Base transformer, Pairwise Attention (PA), and Casanovo on NineSpecies V2. Only the best of the three PA training seeds is shown.

Table 5.2.8: Peptide precision at 100% coverage on the external bacterial test set after MassIVE-KB pretraining.

Model	Precision
Base	0.408
PA	0.433
Casanovo	0.348

precision over a baseline transformer. The magnitude of improvement varies with dataset size but remains robust across domains, indicating that explicit modeling of mass-difference structure provides a consistent inductive advantage.

6 Conclusions and Future Directions

This thesis examined how representation learning and architectural design choices affect transformer-based de novo peptide sequencing from tandem mass spectra. Two complementary strategies were studied: (i) self-supervised encoder pretraining on large unlabeled MS² corpora, and (ii) the introduction of a domain-aligned relational inductive bias through Pairwise Attention. Together, these investigations clarify what can currently be gained from self-supervised learning in proteomics, where its limitations lie, and highlight architectural inductive bias as a particularly reliable mechanism for improving de novo sequencing performance.

Self-Supervised Pretraining for MS²: Progress with Clear Limits

Self-supervised encoder pretraining improves downstream de novo sequencing relative to random initialization, but the magnitude and nature of the gains depend on the alignment between the pretraining objective and the downstream task. Masked autoencoding provides improvements in de novo optimization and accuracy under the controlled, intentionally untuned fine-tuning setup reported in this thesis, indicating that reconstructive, token-level objectives can shape encoder representations in a way that is beneficial for autoregressive peptide decoding.

DINO behaves differently. It learns non-degenerate global spectrum embeddings and captures meaningful structure, as evidenced by downstream retention-time prediction and spectral-quality assessment. However, likely because DINO supervises only a pooled spectrum-level representation, its impact on de novo sequencing is limited. The decoder operates on token-level embeddings, which are only weakly constrained by a purely global objective. As a result, DINO

improves de novo performance relative to scratch initialization, but does not match the gains achieved by dense objectives such as masked autoencoding.

This behavior reflects a general pattern also observed in other domains: objectives that supervise global representations transfer naturally to global tasks, while dense prediction problems require supervision that directly targets token-level structure. The results here indicate that MS² follows the same principle.

At the same time, the fact that DINO learns useful global embeddings at all is notable. Despite being developed for images and applied here to sparse, unordered, set-structured MS² spectra, it recovers physicochemical and fragmentation-related information from raw spectra without peptide annotations. This demonstrates that high-level spectral structure is accessible through fully self-supervised learning, but also that the amount of structure captured so far is limited. Strong effects are observed in diagnostic probing tasks such as retention-time prediction, while gains on more practically relevant tasks, including spectral-quality assessment and de novo sequencing, remain modest.

Global Embedding Evaluations Confirm Nontrivial Structure Learning

To disentangle representation quality from de novo decoding performance, we evaluated self-supervised encoders on downstream tasks that operate directly on pooled spectrum embeddings. These tasks serve as diagnostic probes of whether self-supervised pretraining extracts meaningful global structure from MS² spectra, independent of sequence-level reconstruction.

Frozen DINO embeddings yield strong retention-time prediction performance ($R^2 \approx 0.80$ – 0.85 , depending on model size and peak budget) and modest but consistent improvements in spectral-quality assessment relative to binned-spectrum baselines. Because neither retention time nor spectral quality is observed during pretraining, these results demonstrate that the learned representations encode nontrivial information correlated with peptide physicochemical properties and fragmentation characteristics.

The magnitude of these improvements should be interpreted conservatively. Retention-time prediction from spectra is not a practically useful task, since retention time is already observed whenever an MS² spectrum is acquired. Its role here is purely diagnostic, probing whether global physicochemical structure is encoded in the embedding. Spectral-quality assessment shows only small gains over simple baselines.

In this sense, the global embedding evaluations provide supporting evidence for the conclusions drawn from de novo sequencing: self-supervised learning on MS² spectra uncovers real signal, but current objectives extract only limited structure. Bridging the gap between this weak but measurable signal and representations that materially improve dense prediction tasks remains an open

challenge.

Pairwise Attention: Relational Inductive Bias as a Robust Alternative

The Pairwise Attention experiments clarify an orthogonal but critical point. By modifying only the encoder to incorporate explicit pairwise mass-difference information, consistent and robust improvements are obtained across all evaluated regimes: training from scratch on NineSpecies V1 and NineSpecies V2, large-scale supervised pretraining on MassIVE-KB, and evaluation on an external bacterial dataset with minimal peptide overlap.

The external bacterial benchmark is particularly informative. Both the Base and Pairwise Attention models substantially outperform Casanovo’s published results, despite the bacterial test set containing peptides that are largely disjoint from both MassIVE-KB and the nine-species benchmarks. Pairwise Attention further improves upon the Base model in this setting, indicating strong generalization across species and peptide distributions. This suggests that the induced bias captures fragmentation structure that transfers beyond the training domain, approaching state-of-the-art generalization under realistic distribution shift.

From a conceptual standpoint, Pairwise Attention exemplifies the role of *relational inductive biases* as formalized by Battaglia et al. Standard self-attention assumes unrestricted all-to-all interactions and leaves the discovery of meaningful relations entirely to data. Pairwise Attention instead injects a structured but flexible relational signal, conditioning attention on pairwise mass differences without imposing fixed neighborhoods or sparsity patterns. The relevance of each relation is learned, not prescribed.

Relational Inductive Biases and the Bitter Lesson

The results of this thesis also speak to a broader discussion about the role of inductive bias in modern machine learning. Sutton’s *Bitter Lesson* [30] argues that methods relying on hand-designed structure tend to be outperformed, in the long run, by approaches that scale computation and data while minimizing domain-specific assumptions. From this perspective, architectural interventions such as Pairwise Attention might be viewed as temporary advantages that would eventually be subsumed by sufficiently large models trained on sufficiently large datasets.

However, Battaglia et al. [21] articulate a contrasting view that is particularly relevant for the present setting. They argue that many real-world problems are fundamentally *relational* and that learning in such domains benefits from architectures that encode appropriate relational inductive biases. Crucially, this view does not reject end-to-end learning. Rather, it rejects the false choice between hand-engineering and data-driven learning, emphasizing that architec-

tural structure can guide learning toward the interactions that matter, while still allowing flexibility and scalability.

The empirical results with Pairwise Attention strongly align with this perspective. By injecting a lightweight, physically grounded relational bias based on pairwise mass differences, the encoder consistently improves optimization behavior, data efficiency, and generalization across datasets. These gains persist not only in small supervised regimes, but also under large-scale supervised pre-training and on an external bacterial test set with minimal peptide overlap. This indicates that the improvement is not merely a form of regularization or short-cut learning, but reflects alignment with stable, domain-level structure in MS² fragmentation.

Importantly, Pairwise Attention does not impose a rigid relational graph or fixed neighborhood. Instead, it biases attention weights on pairwise mass differences while allowing the relevance of each interaction to be learned. In this sense, it occupies the intermediate regime emphasized by Battaglia et al.: neither ignoring relational structure nor hard-coding it, but exposing it to the model in a form that supports efficient learning.

From this viewpoint, the success of Pairwise Attention does not contradict the Bitter Lesson so much as it highlights its limits in domains where appropriate learning signals are scarce or misaligned with the task. In proteomics, labeled data are expensive, incomplete, and biased by upstream search engines, while unlabeled data are abundant but difficult to exploit for dense prediction. In such settings, relational inductive bias provides a practical and principled means of improving representation learning without sacrificing end-to-end optimization.

Together with the self-supervised learning results presented earlier, these findings suggest that progress in MS² modeling is likely to come from a combination of scale and structure: large unlabeled datasets, objectives that better match downstream tasks, and architectures that reflect the relational nature of fragmentation spectra.

Toward Foundation Models for MS²

Taken together, the results indicate that self-supervised learning for MS² is feasible and informative, but that the learning capacity of the objectives explored here appears to be the main limiting factor. They recover measurable structure from raw spectra, yet the amount of structure they extract remains insufficient to push state-of-the-art performance in demanding downstream tasks such as *de novo* sequencing.

A key implication is that the bottleneck is not primarily dataset size, but objective capability. If future objectives or training setups cross a qualitative threshold in how much fragmentation-relevant structure they can absorb from unlabeled

beled spectra, then scale becomes decisive: the combination of a sufficiently strong self-supervised signal with massive unlabeled corpora could make pre-training genuinely worthwhile and broadly transferable. In other words, large-scale unlabeled data are likely to matter most *after* the self-supervised learning problem is formulated well enough to exploit them.

From this perspective, the appeal of MS^2 foundation models is real but conditional. Public repositories contain orders of magnitude more unlabeled spectra than can be annotated reliably. However, turning that raw scale into usable representation learning requires objectives that can extract substantially richer structure than the ones tested here.

Reaching that regime is not only a modeling challenge but also a data-engineering one. Constructing large, diverse, and usable unlabeled corpora requires substantial effort in data collection, filtering, and standardization.

Future Research Directions

Domain-aligned self-supervised objectives for MS^2 . The results of this thesis indicate that generic self-supervised objectives developed for images or text do not transfer directly to MS^2 spectra. While masked autoencoding and self-distillation recover measurable signal, the amount of structure learned remains limited, suggesting a mismatch between these objectives and the spectral data-generating process.

Rather than abandoning self-supervised learning for MS^2 , these findings motivate the development of objectives and augmentations that better reflect domain structure. In particular, contrastive or self-distillation frameworks that exploit precursor-mass constraints, spectral additivity, or spectrum composition and decomposition may be more effective. For example, DINO-style objectives could be adapted to operate on structured spectral views formed by combining or selectively excluding peaks under shared precursor constraints. Such formulations could probe whether models can disentangle mixed spectral inputs without explicit supervision. Identifying effective domain-aligned objectives remains an open problem.

Hybrid supervised–self-supervised training. Another promising direction is to combine supervised and self-supervised learning signals within a single training framework. Rather than relying exclusively on self-supervised pretraining, models could be trained jointly on labeled and unlabeled spectra, using supervised sequence-level losses alongside auxiliary representation-level objectives.

Related ideas have been explored in recent work such as ContraNovo [31], which combines a supervised de novo sequencing loss with a contrastive objective applied to encoder representations. However, in that setting the contrastive loss

is applied only to spectra that already have peptide annotations, and no additional unlabeled data are introduced. As a result, such approaches do not directly leverage large-scale unlabeled MS² corpora.

Extending this paradigm to truly hybrid data regimes, where labeled spectra provide task-aligned supervision while large unlabeled collections shape representations through self-supervised objectives, may offer a more effective trade-off between structure and scale.

Hybrid dense–global objectives. The experiments reveal a separation between objectives that supervise token-level structure and those that supervise pooled spectrum-level representations. Dense objectives benefit autoregressive decoding, while global objectives yield transferable embeddings. Combining these signals within a single self-supervised objective, jointly supervising token-level and pooled representations, may bridge this gap, analogous to dense contrastive methods in vision.

Alternative decoding paradigms for de novo sequencing. All de novo sequencing experiments in this thesis rely on autoregressive decoding, which accumulates errors sequentially. An alternative direction is to explore non-autoregressive or iterative decoding schemes, such as diffusion-based sequence models. By refining predictions iteratively rather than committing to a left-to-right trajectory, such approaches may reduce exposure bias and improve robustness. Whether diffusion-based decoders can be effectively applied to peptide generation conditioned on MS² spectra remains an open but well-motivated question.

Closing Remarks

This thesis shows that the quality of learned spectral representations plays an important and often underexplored role. Self-supervised learning and relational inductive bias offer complementary routes toward improving encoder representations. While self-supervised pretraining reveals that some meaningful structure can be recovered from raw MS² spectra, the resulting gains remain limited with current objectives. In contrast, Pairwise Attention demonstrates that introducing a domain-aligned architectural inductive bias can deliver immediate, robust, and generalizable improvements across datasets and distribution shifts.

Together, these results suggest that further progress will likely require a combination of higher-quality labeled data, large-scale unlabeled corpora, objectives aligned with spectral structure, and architectures that explicitly encode relational information, rather than relying on any single factor in isolation.

Bibliography

- [1] Joel Lapin, Alfred Nilsson, Mathias Wilhelm, et al. “Pairwise Attention: Leveraging Mass Differences to Enhance De Novo Sequencing of Mass Spectra”. In: *Journal of Proteome Research* 24.7 (2025). Open Access under CC BY 4.0. doi: 10.1021/acs.jproteome.5c00063. url: <https://pubs.acs.org/doi/10.1021/acs.jproteome.5c00063>.
- [2] Alfred Nilsson and Lukas Käll. *Self-Supervised Learning for Tandem Mass Spectra: Methods, Dynamics, and Downstream Effects*. Manuscript in preparation. 2025.
- [3] Alfred Nilsson, Klas Wijk, Sai Bharath Chandra Gutha, et al. “Indirectly Parameterized Concrete Autoencoders”. In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, et al. Vol. 235. Proceedings of Machine Learning Research. PMLR, 21–27 Jul 2024, pp. 38237–38252. url: <https://proceedings.mlr.press/v235/nilsson24b.html>.
- [4] Alfred Nilsson and Hossein Azizpour. “Regularizing and Interpreting Vision Transformer by Patch Selection on Echocardiography Data”. In: *Proceedings of the fifth Conference on Health, Inference, and Learning*. Ed. by Tom Polard, Edward Choi, Pankhuri Singhal, et al. Vol. 248. Proceedings of Machine Learning Research. PMLR, 27–28 Jun 2024, pp. 155–168. url: <https://proceedings.mlr.press/v248/nilsson24a.html>.
- [5] Luke Squires, Jose Humberto Giraldez Chavez, Alfred Nilsson, et al. “Better Inputs, Better Learning: A Peptide Embedding Tutorial for Proteomic Mass Spectrometry”. In: *Journal of Proteome Research* 25.2 (Feb. 2026), pp. 1160–1165. issn: 1535-3893. doi: 10.1021/acs.jproteome.5c00563. url: <https://doi.org/10.1021/acs.jproteome.5c00563> (visited on 03/27/2026).
- [6] Ngoc Hieu Tran, Xianglilan Zhang, Lei Xin, et al. “De novo peptide sequencing by deep learning”. In: *Proceedings of the National Academy of Sciences* 114.31 (2017), pp. 8247–8252. doi: 10.1073/pnas.1705691114. eprint: <https://doi.org/10.1073/pnas.1705691114>.

BIBLIOGRAPHY

- [//www.pnas.org/doi/pdf/10.1073/pnas.1705691114](https://www.pnas.org/doi/pdf/10.1073/pnas.1705691114). url: <https://www.pnas.org/doi/abs/10.1073/pnas.1705691114>.
- [7] Rui Qiao, Ngoc Hieu Tran, Lei Xin, et al. “Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices”. In: *Nature Machine Intelligence* 3.5 (May 2021), pp. 420–425. issn: 2522-5839. doi: 10.1038/s42256-021-00304-3. url: <https://doi.org/10.1038/s42256-021-00304-3>.
- [8] Melih Yilmaz, William E. Fondrie, Wout Bittremieux, et al. “Sequence-to-sequence translation from mass spectra to peptides with a transformer model”. In: *Nature Communications* 15.1 (July 2024), p. 6427. issn: 2041-1723. doi: 10.1038/s41467-024-49731-x. url: <https://doi.org/10.1038/s41467-024-49731-x>.
- [9] Bin Ma, Kaizhong Zhang, Christopher Hendrie, et al. “PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry”. In: *Rapid Communications in Mass Spectrometry* 17.20 (2003), pp. 2337–2342. doi: <https://doi.org/10.1002/rcm.1196>. eprint: <https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/rcm.1196>. url: <https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/rcm.1196>.
- [10] J. Alex Taylor and Richard S. Johnson. “Sequence database searches via de novo peptide sequencing by tandem mass spectrometry”. In: *Rapid Communications in Mass Spectrometry* 11.9 (1997), pp. 1067–1075. doi: [https://doi.org/10.1002/\(SICI\)1097-0231\(19970615\)11:9<1067::AID-RCM953>3.0.CO;2-L](https://doi.org/10.1002/(SICI)1097-0231(19970615)11:9<1067::AID-RCM953>3.0.CO;2-L). eprint: <https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291097-0231%2819970615%2911%3A9%3C1067%3A%3AAID-RCM953%3E3.0.CO%3B2-L>. url: <https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0231%2819970615%2911%3A9%3C1067%3A%3AAID-RCM953%3E3.0.CO%3B2-L>.
- [11] Bin Ma. “Novor: Real-Time Peptide de Novo Sequencing Software”. In: *Journal of the American Society for Mass Spectrometry* 26.11 (Nov. 2015). Publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved., pp. 1885–1894. doi: 10.1007/s13361-015-1204-0. url: <https://doi.org/10.1007/s13361-015-1204-0>.
- [12] Ari Frank and Pavel Pevzner. “PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling”. In: *Analytical Chemistry* 77.4 (Feb. 2005). Publisher: American Chemical Society, pp. 964–973. issn: 0003-2700. doi: 10.1021/ac048788h. url: <https://doi.org/10.1021/ac048788h>.
- [13] R. Qi Charles, Hao Su, Mo Kaichun, et al. “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 77–85. doi: 10.1109/CVPR.2017.16.
- [14] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, et al. Vol. 30. Curran Associates, Inc., 2017. url:

- https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- [15] Ting Chen, Simon Kornblith, Mohammad Norouzi, et al. “A Simple Framework for Contrastive Learning of Visual Representations”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 1597–1607. url: <https://proceedings.mlr.press/v119/chen20j.html>.
- [16] Kaiming He, Haoqi Fan, Yuxin Wu, et al. “Momentum Contrast for Unsupervised Visual Representation Learning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020.
- [17] Wout Bittremieux, Damon H. May, Jeffrey Bilmes, et al. “A learned embedding for efficient joint analysis of millions of mass spectra”. en. In: *Nature Methods* 19.6 (June 2022), pp. 675–678. issn: 1548-7091, 1548-7105. doi: 10.1038/s41592-022-01496-1. url: <https://www.nature.com/articles/s41592-022-01496-1> (visited on 11/03/2023).
- [18] Tom Altenburg, Thilo Muth, and Bernhard Y. Renard. *yHydra: Deep Learning enables an Ultra Fast Open Search by Jointly Embedding MS/MS Spectra and Peptides of Mass Spectrometry-based Proteomics*. en. preprint. Bioinformatics, Dec. 2021. doi: 10.1101/2021.12.01.470818. url: <http://biorxiv.org/lookup/doi/10.1101/2021.12.01.470818> (visited on 11/03/2023).
- [19] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 8748–8763. url: <https://proceedings.mlr.press/v139/radford21a.html>.
- [20] Justin Sanders, Melih Yilmaz, Jacob H. Russell, et al. *Foundation model for mass spectrometry proteomics*. 2025. arXiv: 2505.10848 [cs.LG]. url: <https://arxiv.org/abs/2505.10848>.
- [21] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, et al. *Relational inductive biases, deep learning, and graph networks*. 2018. arXiv: 1806.01261 [cs.LG]. url: <https://arxiv.org/abs/1806.01261>.
- [22] John Jumper, Richard Evans, Alexander Pritzel, et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (Aug. 2021), pp. 583–589. issn: 1476-4687. doi: 10.1038/s41586-021-03819-2. url: <https://doi.org/10.1038/s41586-021-03819-2>.
- [23] Kaiming He, Xinlei Chen, Saining Xie, et al. “Masked Autoencoders Are Scalable Vision Learners”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 16000–16009.
- [24] Mathilde Caron, Hugo Touvron, Ishan Misra, et al. “Emerging Properties in Self-Supervised Vision Transformers”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 9650–9660.
- [25] Merve Yilmaz, William Fondrie, Wout Bittremieux, et al. *Nine Species Benchmark Dataset for De Novo Peptide Sequencing*. Zenodo, 2022. doi: 10.5281/zenodo.5976002. url: <https://doi.org/10.5281/zenodo.5976002>.

BIBLIOGRAPHY

- [26] InstaDeepAI. *Nine-Species Benchmark (Updated Version)*. https://huggingface.co/datasets/InstaDeepAI/ms_ninespecies_benchmark. 2024.
- [27] Wout Bittremieux. *MassIVE-KB v1: 30 Million Peptide-Spectrum Matches*. Zenodo, 2025. doi: 10.5281/zenodo.14973855. url: <https://doi.org/10.5281/zenodo.14973855>.
- [28] *DeNovo Peptide Identification Deep Learning Dataset PXD010000*. <https://www.ebi.ac.uk/pride/archive/projects/PXD010000>.
- [29] Joon-Yong Lee, Hugh D. Mitchell, Meagan C. Burnet, et al. “Uncovering Hidden Members and Functions of the Soil Microbiome Using De Novo Metaproteomics”. In: *Journal of Proteome Research* 21.8 (Aug. 2022), pp. 2023–2035. issn: 1535-3893. doi: 10.1021/acs.jproteome.2c00334. url: <https://doi.org/10.1021/acs.jproteome.2c00334> (visited on 03/27/2026).
- [30] Richard S. Sutton. *The Bitter Lesson*. Online essay. Accessed: 2025-08-XX. 2019. url: <http://www.incompleteideas.net/IncIdeas/BitterLesson.html>.
- [31] Zhi Jin, Sheng Xu, Xiang Zhang, et al. “ContraNovo: A Contrastive Learning Approach to Enhance De Novo Peptide Sequencing”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 38 (Mar. 2024), pp. 144–152. doi: 10.1609/aaai.v38i1.27765.

Appended papers

