Artificial Intelligence in Spectroscopy: Advancing Chemistry from Prediction to Generation and Beyond

Kehan Guo[1], Yili Shen[1], Gisela Abigail Gonzalez-Montiel[2], Yue Huang[1], Yujun Zhou[1], Mihir Surve[2], Zhichun Guo[3], Payel Das[4], Nitesh V. Chawla[1], Olaf Wiest[2], Xiangliang Zhang[1]

[1]Department of Computer Science and Engineering, University of Notre Dame, IN, USA
[2]Department of Chemistry and Biochemistry, University of Notre Dame, IN, USA
[3]Institute for Protein Design, University of Washington, WA, USA
[4]Trusted AI Department, IBM Thomas J. Watson Research Center, NY, USA

Abstract

The rapid advent of machine learning (ML) and artificial intelligence (AI) has catalyzed major transformations in chemistry, yet the application of these methods to spectroscopic and spectrometric data—termed Spectroscopy Machine Learning (SpectraML)—remains relatively underexplored. Modern spectroscopic techniques (MS, NMR, IR, Raman, UV-Vis) generate an ever-growing volume of high-dimensional data, creating a pressing need for automated and intelligent analysis beyond traditional expert-based workflows. In this survey, we provide a unified review of SpectraML, systematically examining state-of-the-art approaches for both forward tasks (molecule-to-spectrum prediction) and inverse tasks (spectrum-to-molecule inference). We trace the historical evolution of ML in spectroscopy—from early pattern recognition to the latest foundation models capable of advanced reasoning—and offer a taxonomy of representative neural architectures, including graph-based and transformer-based methods. Addressing key challenges such as data quality, multimodal integration, and computational scalability, we highlight emerging directions like synthetic data generation, large-scale pretraining, and few- or zero-shot learning. To foster reproducible research, we release an open-source repository containing curated datasets and code implementations. Our survey serves as a roadmap for researchers, guiding advancements at the intersection of spectroscopy and AI.

1 Introduction

The rapid advancements in Artificial Intelligence (AI) and Machine Learning (ML) are reshaping scientific disciplines of chemistry, by streamlining tasks such as molecular property prediction [Guo *et al.*, 2021] and reaction modeling [Coley *et al.*, 2019]. Despite these breakthroughs, the application of ML to spectroscopy—hereafter referred to as **Spectroscopy Machine Learning (SpectraML)** [Elias *et al.*, 2004; Ralbovsky and Lednev, 2020]—remains relatively underexplored. Spectroscopic and spectrometric techniques, which provide high-sensitivity insights into molecular structure, dynamics, and properties, are now generating large volumes of data due to advances in high-throughput experiments and automated acquisition. Consequently, traditional manual analysis methods, reliant on expert interpretation and reference libraries [Alberts *et al.*, 2024b; Zhu and Jonas, 2023], are increasingly inadequate for handling the scale and complexity of modern spectral datasets.

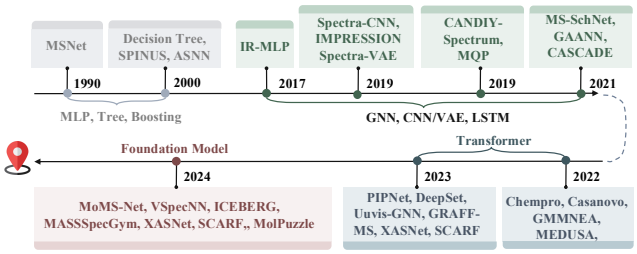

Figure 1: Timeline of ML progression and its application to spectroscopic studies.

The growing interest in this field is reflected in the increasing number of research papers that expand the scope of tasks addressed by spectrum-based ML models [Gabriel *et al.*, 2024; Alberts *et al.*, 2024b; Guo *et al.*, 2024; Bushuiev *et al.*, 2024]. However, existing overviews often focus on a single modality (e.g., UV alone or MS) [Beck *et al.*, 2024] or lack a clear framework for distinguishing forward (molecule-to-spectrum) from inverse (spectrum-to-molecule) tasks [Sridharan *et al.*, 2022a]. By contrast, our survey unifies five major spectroscopic techniques—MS, NMR, IR, Raman, and UV-Vis—within a single methodological framework. Moreover, we highlight the rapid progression of spectroscopic analysis driven by ML advances in generative modeling, few- or zero-shot learning, and large-scale pretraining, and we provide an open-source repository of datasets and code. By bridging computational and experimental viewpoints, our work fills a key gap in the literature and highlights new avenues for interdisciplinary collaboration in SpectraML.

The rapid advancements in ML and AI have been transforming workflow automation in spectral analysis, as illustrated by the timeline in Fig. 1. Deep learning models, such as convolutional neural networks (CNNs) [O'Shea, 2015]

and recurrent neural networks (RNNs) [Schuster and Paliwal, 1997], have proven effective in tasks like peak detection, deconvolution, and reaction monitoring. Additionally, transfer learning and pre-trained models [Guo *et al.*, 2024] enable these algorithms to generalize across diverse spectra, thereby reducing the need for extensive retraining. Emerging foundation models [Bommasani *et al.*, 2021] further extend the capabilities of SpectraML by offering advanced reasoning and planning for complex tasks such as molecular structure elucidation and reaction pathway prediction [Guo *et al.*, 2024]. As AI techniques continue to evolve, there is a critical need for a structured discussion on positioning the different capabilities of AI models across various spectroscopy tasks, as well as underscoring key challenges, limitations, and future directions.

This survey addresses these needs with the following contributions:

1. We offer a comprehensive overview of current SpectraML techniques across five major spectroscopic modalities—MS, NMR, IR, Raman, and UV-Vis—highlighting both methodological innovations and practical applications. Unlike existing surveys that focus on a single modality or overlook the distinction between forward (molecule-to-spectrum) and inverse (spectrum-to-molecule) tasks [Beck *et al.*, 2024; Sridharan *et al.*, 2022a], our work provides a unified perspective and frames these tasks within AI's problem-solving role.
2. We present a **unified roadmap** that traces the evolution of ML in spectroscopy, from early pattern recognition and predictive analytics to advanced generative and reasoning frameworks, thus situating current progress within a broader historical context. It helps researchers understand how foundational techniques have shaped modern approaches and guides them in innovating future methodologies in SpectraML.
3. We identify key challenges (e.g., data quality, multimodal integration, and computational scalability) and emerging opportunities (e.g., foundation models, synthetic data generation, few- or zero-shot learning, and large-scale pretraining) in SpectraML. To facilitate further research, we provide, and will maintain, an open-source GitHub repository containing datasets and code. This work thus serves as a valuable resource for researchers and practitioners in this interdisciplinary field.

The remainder of this paper is organized as follows. Section 2 introduces spectral data representations and the definition of fundamental **Forward and Inverse Problems** in spectral analysis. Section 3 categorizes and summarize SpectraML approaches in solving forward and inverse problems. In Section 4, we discuss major challenges and highlight emerging directions such as **foundation models**, and **synthetic data generation**. Section 5 concludes the work.

2 Background

2.1 Applications of Spectroscopy in Chemistry

Spectroscopy, the study of the interaction between matter and electromagnetic radiation, produces data that resembles **audio signals** in its representation—peaks, shifts, and patterns that encode molecular information [Elias *et al.*, 2004]. Spectrometry, on the other hand, focuses on measuring chemical interaction to gain insight into molecular structures and properties [Ralbovsky and Lednev, 2020]. Common spectroscopic techniques include mass spectrometry (MS), infrared (IR), Raman, ultraviolet/visible (UV-Vis), and nuclear magnetic resonance (NMR). Each of these techniques is akin to a "lens" providing a different perspective of the molecular world, and when combined, they reveal a fuller picture of molecular structures.

- **Mass Spectrometry (MS)** allows for the determination of the molecular mass and formula of a compound, as well as some of its structural features by identifying the fragments produced when the molecule breaks apart.
- **Infrared (IR)** and **Raman** spectra data allow the identification of the types of functional groups in a compound.
- **UV-Vis** spectra data provides information about compounds that have conjugated double bonds.
- **NMR** spectra data provide information about atomic nuclei (e.g., the carbon-hydrogen framework of a compound). Advanced techniques, 2D and 3D-NMR, further enable the characterization of complex molecules such as natural products, proteins, and nucleic acids.

The obtained spectra data are widely used across chemistry, biology, and related fields, akin to a "molecular microscope" that enables researchers to explore the unseen. These spectral data are often presented as plots or graphs that visually represent the relationship between intensity and a specific variable, such as wavelength, wavenumber, or mass-to-charge ratio (*m/z*), as demonstrated on the left part of Figure 2. Studies involving these data are generally divided into two main categories:

- ***Forward Problem***: *predicting a spectrum based on molecular structure information*. While spectroscopy devices can generate spectra from molecular samples, solving the **Forward Problem** (structure-to-spectrum problem) using AI models is highly valuable and offers several key advantages. First, it reduces the need for costly and time-consuming experimental measurements by enabling rapid spectral predictions. Second, it enhances the understanding of fundamental relationships between molecular structures and their spectral signatures. Such structure-to-spectrum correlation is crucial for scientists to know what molecule(s) are present for drug discovery, biomarker research, natural product synthesis, and other research areas [Mandelare *et al.*, 2018]. Lastly, it expands applications beyond experimental limits. Some molecules are difficult to analyze using standard spectroscopy due to low concentrations, unstable intermediates, or extreme environmental conditions. AI solutions enables insights into such challenging cases where direct measurement is impractical.
- ***Inverse (Backward) Problem****: deducing the molecular structure based on experimentally obtained spectra,* also known as molecule elucidation, is a crucial task in life

sciences, chemical industries, and other fields [Sridharan *et al.*, 2022a; Yao *et al.*, 2023]. Resolving this problem enables researchers to identify unknown compounds, verify chemical compositions, and gain deeper insights into molecular behavior, ultimately advancing scientific discovery and industrial applications. However, molecular elucidation remains a time-consuming and complex process that heavily relies on human expertise. Identifying spectrum-to-structure correlation is particularly challenging, requiring analysts to distinguish real peaks and accurately deduce their chemical meaning. Manual interpretation is labor-intensive, has limited scalability, and is also prone to misinterpretation due to overlapping signals, sample impurities, and isomerization issues. This is where AI can play a transformative role, automating spectral interpretation, and accelerating the resolution of inverse problems.

Note that the above definition of the forward/inverse problem is in accordance with what is commonly referred to in the community [Lu *et al.*, 2024]. However, the opposite definition exists in some contexts, e.g., in [Beck *et al.*, 2024], where the inverse problem focuses on predicting spectra, while the forward problem refers to molecular deduction from given spectra. This difference in terminology highlights the slightly varying perspectives across disciplines and underscores the need for clear definitions when discussing these concepts in the context of spectroscopy and ML applications.

2.2 Roadmap of SpectraML

ML has revolutionized the way spectroscopic data is analyzed, offering new pathways to extract deeper insights, accelerate workflows, and uncover patterns beyond human capability. Historically, the use of computational techniques in spectroscopy was limited to basic pattern recognition and property prediction tasks [Elias *et al.*, 2004]. This changed with the advent of deep learning and advanced ML frameworks that have enabled transformative capabilities across the entire spectrum analysis pipeline. For instance, CNNs excel in tasks such as peak detection [O'Shea, 2015] and deconvolution [Hu *et al.*, 2024], akin to identifying features in an image, while RNNs and transformers [Schuster and Paliwal, 1997] handle sequential spectral data, similar to interpreting audio signals, making them suitable for reaction monitoring and dynamic studies. For example, CASCADE [Guan *et al.*, 2021] accelerates the prediction of chemical shift in NMR spectra by *6000 times* comparing to the fastest DFT method, enabling real-time NMR chemical shift predictions from simple molecular representations.

As spectroscopic datasets have grown in size and complexity, ML has demonstrated exceptional scalability and adaptability. The shift from early predictive models to modern **generative** and **reasoning** frameworks, such as attention-based transformers and foundation models, has redefined the scope of spectral analysis. Generative models enable the simulation of spectra based on molecular structures [Goldman *et al.*, 2023], addressing the **forward problem**, while reasoning-driven models tackle the **inverse problem**, predicting molecular structures with enhanced accuracy [Alberts *et al.*, 2024a; Alberts *et al.*, 2023]. More discussion regarding these two types of probelms is presented in next section. These developments have brought unprecedented precision and speed to applications ranging from molecular characterization to reaction pathway prediction. For example, IMPRESSION [Gerrard *et al.*, 2020] predicts NMR parameters with near-quantum chemical accuracy while accelerating computational time from days to seconds.

3 SpectraML Methodologies Summary

In this section, we present a detailed discussion of the machine learning methodologies that address the twin challenges: the **forward (molecule-to-spectrum) and inverse (spectrum-to-molecule) problems**. Our discussion is organized around four core components. We begin by examining the data representations and preprocessing strategies that serve as the foundation for effective spectral modeling. We then focus on the forward problem of predicting spectral signatures from molecular structures, followed by a discussion of the inverse problem of inferring molecular structures from spectral data. Finally, we describe emerging unified frameworks and cross-modal integration approaches that promise to address both challenges simultaneously. A summary of the discussed work is presented in Table 1.

3.1 Data Representations and Preprocessing

The quality of spectral analysis is fundamentally determined by how both molecular and spectral data are represented and preprocessed. In SpectraML, spectral data may be expressed as vectors, sequences, or images, while molecular structures are encoded using vector-based descriptors, simplified molecular input line entry system (SMILES) strings, 2D graphs, or 3D coordinates. Such diverse representations are essential for capturing the intricate details of molecular interactions. However, the high-dimensional and heterogeneous nature of spectral data, combined with challenges such as noise, baseline drift, and instrument variability, demands robust preprocessing pipelines. Early work demonstrated that conventional normalization and alignment techniques were insufficient for fully preserving the chemical information embedded in these datasets. Recent studies, including those by [Gastegger *et al.*, 2017; Gerrard *et al.*, 2020], have underscored the importance of integrating domain-specific knowledge, such as physics-informed normalization and tailored feature extraction into the preprocessing stage. More recent work, such as [Alberts *et al.*, 2024a; Alberts *et al.*, 2024b], is building a large-scale spectral dataset and harnessing the power of transformer-based models to map the latent representations of spectral data, thereby paving the way for robust and generalizable spectral analysis frameworks. These advances ensure that the learned representations are both resilient to experimental artifacts and chemically meaningful, thereby establishing a strong foundation for addressing both forward and inverse tasks.

3.2 Forward Problem: Molecule-to-Spectrum

The forward problem in SpectraML aims at predicting spectral information directly from known molecular structures,

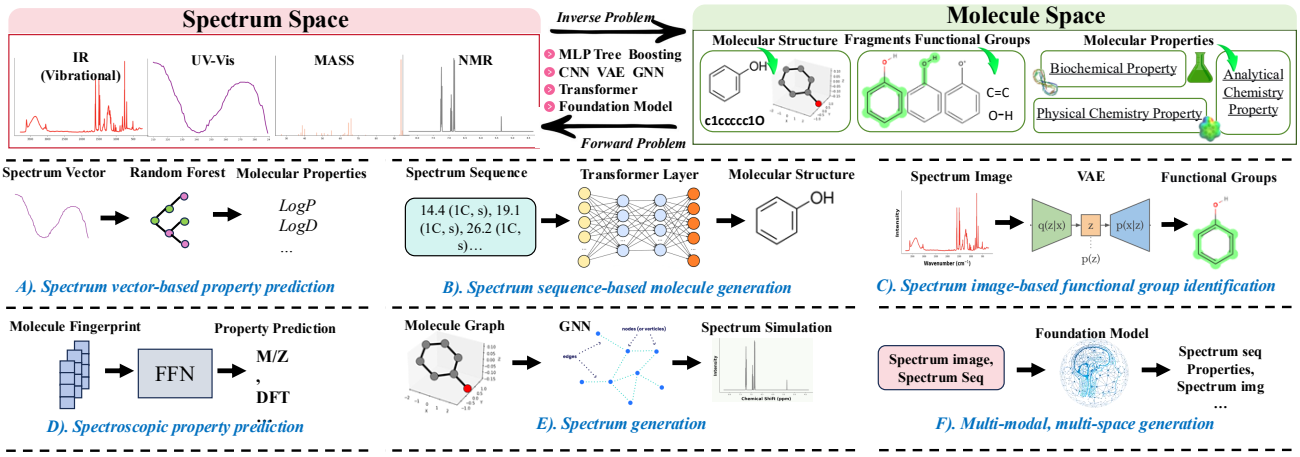

Figure 2: (Top) Overview of SpectraML, translating between **Spectrum Space** and **Molecule Space**. (Middle and Bottom) Illustration of key tasks in SpectraML, including their **inputs**, **outputs**, and the **machine learning models** used for mapping them, such as Random Forest, Feed Forward Networks (FFN), Variational Autoencoders (VAE), Transformers, Graph Neural Networks (GNN), and Foundation Models.

serving as an efficient alternative to computationally expensive quantum-chemical simulations and laborious experimental measurements. Forward-problem can also be extended to extract critical spectral features and related chemical properties. Therefore, the **input** of these ML-empowered solutions consists of molecules represented in different forms, such as SMILES strings, molecular graphs, or three-dimensional coordinates. The **output** can be either full spectra across different modalities (MS, NMR, IR, Raman, UV-Vis) or specific spectral features and chemical properties relevant to the target application. These **ML approaches** typically adopt an **encoding–prediction** framework, which predicts spectral features or related chemical properties in forms of regression or classification. In such architectures, the encoder transforms the molecular structure into a latent feature space that captures its essential chemical characteristics. The subsequent prediction stage then leverages this representation to predict partial spectra or specific spectral properties, depending on the target modality. While encoding and prediction are often implemented and trained end-to-end within a single model, without a strict separation (as demonstrated in example tasks D and E in Fig. 2), we structure the following discussion of related work based on the various forms of input and output involved in these problems, as they directly influence the selection and design of applicable machine learning models.

Input Encoding. As summarized in Table 1, the input to the forward problem is often in the form of vector-based molecular features/descriptors, 2D, or 3D molecular graphs. This determines the choice of encoding, which is typically implemented as an MLP for vector-based molecular features/descriptors and a GNN for 2D and 3D molecular graphs. While vector representations are straightforward to handle, molecules represented as graphs require more sophisticated processing. Message-passing layers within GNNs effectively capture the structural and relational information between atoms and bonds. These graph-based encoders are typically paired with regression or classification modules to predict continuous properties, such as ^{1}H and ^{13}C chemical shifts [Guan *et al.*, 2021; Kwon *et al.*, 2020; Kang *et al.*, 2020; Jonas and Kuhn, 2019], or to learn spectral features like excitation energies and spectral line shapes [McNaughton *et al.*, 2023; Chen *et al.*, 2022; Singh *et al.*, 2022]. Alternatively, encoders may utilize direct coordinate-based features. For example, physics-informed neural networks extract vibrational properties directly from atomic coordinates—integrating experimental observations with quantum chemical insights—to predict key quantities such as dipole moment derivatives and polarizability tensors [Schütt *et al.*, 2021; Gastegger *et al.*, 2021; Chen *et al.*, 2024; Sowa and Rossky, 2024].

Output Prediction. The "Task Type" and "Output" column in Table 1 for the forward problem indicate that the output prediction is mostly in the form of regression. For example in MS prediction, molecular substructures are mapped to fragment *m/z* values and intensities. In NMR spectroscopy, three-dimensional molecular graphs serve as inputs to predict continuous ^{1}H and ^{13}C chemical shifts, which in turn enables the accurate prediction of coupling constants and supports MAS-based spectral reconstructions [Cordova *et al.*, 2023]. For IR, Raman, and UV spectroscopy, the prediction stage typically employs regression or classification layers to generate vibrational properties. In particular, key quantities—such as dipole moment derivatives and polarizability tensors—are predicted to capture the essential physical characteristics of the molecules [Schütt *et al.*, 2021; Gastegger *et al.*, 2021; Chen *et al.*, 2024; Sowa and Rossky, 2024]. For ultraviolet and electronic spectra, excitation energies and spectral line shapes are predicted [McNaughton *et al.*, 2023; Chen *et al.*, 2022; Singh *et al.*, 2022]. The prediction could also involve validating subformulas and predicting discrete spectral features [Goldman *et al.*, 2023; Zhu *et al.*, 2020; Young *et al.*, 2024; Park *et al.*, 2024; Zhu and Jonas, 2023; Murphy *et al.*, 2023; Goldman *et al.*, 2024], with some approaches further extending the framework to perform joint property prediction [Voronov *et al.*, 2022]. The output can also take the form of a sequence of spectral tokens [Wei *et al.*, 2019]. In this case, a generative model is employed to map a

SMILES string to the full spectrum, enabling sequence-based spectral prediction.

Extension of Forward Problems. Beyond simulating spectral profiles, these forward-modeling approaches also support property-focused tasks: classification models can reveal reaction behaviors in MS-based reactivity analyses [Fine *et al.*, 2020a], while hybrid ML-first-principles methods utilize IR data to infer adsorption energies and bond strengths [Du *et al.*, 2023]. Additional efforts have extended these frameworks to predict other physicochemical parameters, such as logD values at various pH levels [Leniak *et al.*, 2024], thereby supporting broader applications in drug discovery, catalyst design, and materials optimization.

3.3 Inverse Problem: Spectrum-to-Molecule

The inverse problem in spectral analysis aims at inferring a molecular structure directly from its measured spectrum, providing a complementary approach to traditional structure elucidation methods. In this task, the **input** consists of spectral measurements that can vary widelyfrom one-dimensional NMR signals and high-dimensional spectral vectors to image-like two-dimensional matrices (e.g., from NMR or IR) and sequential data from mass spectrometry (MS). The **output** is the predicted molecular structure, commonly represented as a molecular graph or an SMILES string. **ML approaches** to the inverse problem typically adopt either an **encoding–decoding** scheme, where the spectral data is transformed into a latent representation and then decoded into a molecular structure (e.g., task B in Fig. 2), or an **encoding–prediction** framework, which directly predicts molecular substructures or functional groups from the spectral features (e.g., task C in Fig. 2). In such architectures, the encoder processes the input spectra to capture the critical information necessary for structure elucidation, and the subsequent decoder or classifier reconstructs the corresponding molecular representation.

Input Encoding. As presented in the bottom part of Table 1, the input to inverse modeling typically consists of one-dimensional ^{1}H or ^{13}C NMR spectra, which are signals along a single frequency axis, representing the chemical shift of the nuclei being observed, and often represented as high-dimensional vectors. For example, [Hu *et al.*, 2024] employs a multitask, transformer-based model to encode 1D NMR signals into a latent space, facilitating the reconstruction of full molecular structures and substructure arrays. Similarly, [Huang *et al.*, 2021] integrates convolutional neural networks with beam search to process spectral inputs, predicting substructure probabilities and iteratively assembling complete molecular graphs. In another approach, [Yao *et al.*, 2023] leverages a bidirectional, auto-regressive transformer (BART) [Lewis, 2019] that is pre-trained on large-scale molecular data and fine-tuned with ^{13}C NMR constraints. Additional methods, such as that of [Alberts *et al.*, 2023], tokenize NMR spectral features into sequences for encoding, while [Sridharan *et al.*, 2022b] combines Monte Carlo Tree Search with graph convolutional networks to iteratively build molecular graphs guided by spectral cues. These encoder designs are crucial for capturing both local and global spectral patterns that underpin accurate molecular reconstruction.

Output Decoding and Prediction. When the output of the inverse problem is a molecular structure, the task becomes a **generative** problem, where the decoder functions as a generator to reconstruct the molecular structure from spectral data, either as a sequence of tokens representing SMILES strings or by progressively constructing molecular graphs. For instance, [Alberts *et al.*, 2023] utilizes a transformer decoder to convert tokenized NMR or IR spectra into SMILES strings, treating each spectral absorption value as a sequence element in a translation-like process. Similarly, [Jonas, 2019] frames the molecule reconstruction task as a Markov decision process (MDP) and incrementally reconstructs the molecule through a relation network. Moreover, additional constraints can be incorporated to refine the generative process; for example, [Sun *et al.*, 2024] couples generative models with contrastive retrieval to enhance candidate matching accuracy, [Zheng *et al.*, 2024] focuses on classifying seized substances from ^{1}H and ^{13}C NMR data, and [Tian *et al.*, 2024] verifies proposed structures through joint analysis of image-like spectral data and graph-based molecular features.

Alternatively, an **encoding–prediction** framework maps the spectral representation to discrete structural elements, such as molecular substructures or functional groups, without generating an entire molecular structure. In this paradigm, the model deduces how atoms and functional groups are arranged to produce the observed spectral features, a capability that is particularly valuable for applications ranging from natural product identification to forensic analysis. For example, in MS context, MEDUSA [Boiko *et al.*, 2022] and CANOPUS [Dührkop *et al.*, 2020] incorporate classification and ranking layers to discriminate between candidate metabolites based on MS and MS/MS features, while CandyCrunch [Urban *et al.*, 2024] predicts glycan topologies by analyzing tandem MS data. In IR spectroscopy, CANDIY-spectrum [Fine *et al.*, 2020b] and CNN-based methods [Enders *et al.*, 2021] focus on identifying diagnostic functional groups from characteristic absorption patterns. Similarly, transformer-based networks for NMR leverage ^{1}H and ^{13}C spectra to predict both key substructures and complete molecular formulas for robust classification and reconstruction [Huang *et al.*, 2021; Hu *et al.*, 2024; Alberts *et al.*, 2023]. By directly extracting these critical features, the encoding–prediction paradigm offers an interpretable and efficient alternative to generative approaches for structure elucidation.

Extension of Inverse Problems. Beyond full structure elucidation, inverse SpectraML can be extended to recover detailed substructural information and functional group classifications that are essential for rapid compound identification and downstream analysis. For instance, sequence-to-sequence models applied to MS/MS data—such as those demonstrated in Casanovo [Yilmaz *et al.*, 2022]—and hybrid systems that combine substructure detection with full structure generation [Kim *et al.*, 2023; Yao *et al.*, 2023] further enhance compound identification when integrated with spectral databases, as seen in CFLS [Sun *et al.*, 2024]. These extended approaches broaden the applicability of inverse SpectraML to diverse fields, from natural product discovery and metabolite screening to forensic investigations, thereby significantly reducing the reliance on time-intensive manual verification.

Paper	Forward Problem: Molecule-to-Spectrum Prediction				
	Task Type	**Input**	**Output**	**Model**	**Dataset**
Vector-Based Molecular Representations					
[Binev *et al.*, 2007]	REG	Molecular features	Chem. shift, Coupling const.	ASNN	Custom
[Gastegger *et al.*, 2017]	REG	3D coordinates	Simulated IR spectrum	MLP	Custom
[Gerrard *et al.*, 2020]	REG	Coulomb matrix	Chem. shift, Coupling const.	KRR	CSD subset
[Ye *et al.*, 2020]	REG	Coulomb matrix	IR properties	MLP	Custom
[Chen *et al.*, 2022]	REG	Bispectrum components	Vertical excitation energy	LASSO	Custom
[Lin *et al.*, 2022]	REG	Geometric descriptors	Chemical shift	MLP	Custom
[Sowa and Rossky, 2024]	REG	Geometric descriptors	Polarizability tensor	KRR	Custom
[Ho Manh *et al.*, 2024]	REG	Molecular features	Vacuum UV spectrum	Random Forest	Custom
2D Graph-Based Molecular Representations					
[Jonas and Kuhn, 2019]	REG	2D graph	Chemical shift	GNN	NMRshiftdb2
[Kwon *et al.*, 2020]	REG	2D graph	Chemical shift	GNN	NMRshiftdb2
[Kang *et al.*, 2020]	REG	2D graph	Chemical shift	GNN	NMRshiftdb2
[Zhu *et al.*, 2020]	REG	2D graph	MS peaks vector	GNN	NIST17
[Young *et al.*, 2024]	REG	2D graph	MS peaks vector	Transformer	NIST20
[Goldman *et al.*, 2023]	CLS, REG	2D graph	Subformula classification	GNN	NIST20, NPLIB1
[Zhu and Jonas, 2023]	REG	2D graph	MS peaks vector	GNN	NIST17
[Murphy *et al.*, 2023]	CLS, REG	2D graph	Subformula classification	GNN	NIST20
[Park *et al.*, 2024]	REG	2D graph	MS peaks vector	GNN	NIST20
[Goldman *et al.*, 2024]	REG	2D graph	MS peaks vector	GNN	NIST20
3D Molecular Representations					
[Gastegger *et al.*, 2021]	REG	3D graph	Multiple spectral properties	GNN	MD17
[Schütt *et al.*, 2021]	CLS	3D graph	Peptide-spectrum matches	GNN	MD17, QM9
[Guan *et al.*, 2021]	REG	3D graph	Chemical shift	GNN	NMRshiftdb2
[Singh *et al.*, 2022]	REG	3D graph	Excitation spectrum	GNN	QM9
[Chen *et al.*, 2024]	REG	3D graph	Energy, forces, dipole moments	GNN	Custom
SMILES Representations					
[Wei *et al.*, 2019]	GEN	SMILES Seq	(EI-MS) prediction	MLP	NIST17
Paper	**Inverse Problem: Spectrum-to-Molecule Prediction**				
	Task Type	**Input**	**Output**	**Model**	**Dataset**
NMR Spectral Representations					
[Jonas, 2019]	GEN	Formula + NMR vector	Molecule Graph	GNN	NMRshiftdb
[Sridharan *et al.*, 2022b]	GEN	NMR Vector	Molecule Graph	GCN	NMRshiftdb2
[Yilmaz *et al.*, 2022]	GEN	MS Seq	SMILES Seq	Transformer	DeepNovo
[Kim *et al.*, 2023]	CLS	NMR image	Molecule structure classification	CNN	Custom
[Yao *et al.*, 2023]	GEN	NMR Seq	SMILES Seq	Transformer	CReSS
[Alberts *et al.*, 2023]	GEN	NMR sequence	SMILES sequence	Transformer	Pistachio
[Hu *et al.*, 2024]	GEN	NMR vector	SMILES sequence	Transformer	SpectraBase
[Leniak *et al.*, 2024]	REG	NMR vector	LogD value	SVR	SpecFAI
[Yan *et al.*, 2024]	GEN	Low-resolution NMR image	High-resolution image	GAN	Custom
[Guo *et al.*, 2024]	GEN,REA	IR, NMR, MS image	SMILES sequence	MLLM	MolPuzzle
[Su *et al.*, 2024]	GEN,REA	BitMap image	Molecule Graph	LLM	Custom
Other Spectral Representation					
[Wei *et al.*, 2019]	REG	MS vector	Intensity values	MLP	NIST2017
[Fine *et al.*, 2020b]	CLS	IR/MS vector	Functional group classification	MLP	CANDIY
[Fine *et al.*, 2020a]	CLS	MS vector	Reaction classification	Decision Tree	MoP
[Enders *et al.*, 2021]	CLS	IR image	Functional group classification	CNN	FTIRML
[Alberts *et al.*, 2024a]	GEN, REA	IR Seq	SMILES Seq	Transformer	NIST2010

Table 1: Summary of ML approaches in spectral analysis categorized into **Forward Problems** (molecule-to-spectrum) and **Inverse Problems** (spectrum-to-molecule). Studies are grouped by input representation. Task types are annotated regarding output: CLS (Classification), REG (Regression), and GEN (Generation) REA (Reasoning). Papers in each section are ordered by year of publication.

3.4 Unified Frameworks and Cross-Modal Integration

Recognizing that forward and inverse problems share common underlying chemical principles, recent research has begun to develop unified frameworks capable of addressing both tasks concurrently. **Foundation models** pre-trained on large, heterogeneous spectral datasets are at the forefront of this endeavor. These models leverage cross-domain learning to capture shared features across diverse modalities, such as IR, NMR, MS, and Raman—thereby enabling few-shot and zero-shot learning capabilities [Bommasani *et al.*, 2021]. Concurrently, physics-informed generative models, including diffusion models and GAN-based super-resolution techniques, have been introduced to synthesize high-fidelity spectra while respecting known chemical constraints [Cordova *et al.*, 2023]. Hybrid architectures that combine the relational modeling strength of GNNs with the sequence modeling capabilities of transformers offer a particularly promising route toward integrated spectrum analysis [Young *et al.*, 2024]. Moreover, foundation models are steering the **advancement of reasoning-driven spectrum analysis**, particularly in complex inference tasks such as spectral deconvolution, peak assignment, and spectral consistency validation [Guo *et al.*, 2024; Su *et al.*, 2024; Alberts *et al.*, 2024b]. These models can reason about ambiguous spectra by lever-

aging prior chemical knowledge to infer plausible molecular structures, resolve overlapping spectral features, and predict missing spectral regions. By unifying forward and inverse tasks within a single framework, these emerging approaches not only alleviate issues of data scarcity through synthetic data generation but also enhance model robustness and interpretability. This integrated perspective is poised to accelerate discovery in diverse domains such as drug development, materials science, and environmental monitoring.

4 Challenges and Opportunities

4.1 Data Quality, Scarcity, and Complexity

SpectraML faces several interrelated challenges arising from the inherent nature of experimental spectral data and the limitations of current ML approaches. First, the variability in data quality is a significant obstacle. Experimental spectra are often compromised by noise, baseline drifts, and instrument-to-instrument discrepancies, leading to inconsistencies in spectral resolution and intensity. Such variability complicates model training and can severely degrade the predictive performance of ML algorithms—especially when preprocessing pipelines are insufficiently robust. Moreover, the scarcity and imbalance of high-quality, annotated spectral datasets, particularly for rare or complex compounds, further exacerbate the issue. The limited availability of training data not only hinders the generalization of models across diverse chemical spaces but also increases the risk of overfitting, necessitating strategies such as data augmentation and transfer learning.

In addition to data quality challenges, the intrinsic complexity of spectral data presents a formidable hurdle. Spectral measurements typically exhibit high dimensionality and overlapping peaks, making feature extraction a nontrivial task. Current ML models often struggle to capture the nuanced, high-dimensional patterns inherent in such data, which leads to suboptimal performance in tasks like peak detection and feature discrimination. Furthermore, many existing architectures are not designed to fully leverage the domain knowledge embedded in spectroscopic data, thereby limiting their ability to exploit underlying chemical and physical principles.

A further challenge arises from the need to integrate data from multiple spectroscopic techniques (e.g., IR, MS, and NMR), each characterized by distinct scales, formats, and noise properties. Developing effective fusion strategies to reconcile these differences into a unified model is nontrivial. Most current ML architectures are optimized for single-modality inputs and often fail to capture the critical cross-domain relationships needed for accurate spectral analysis. Additionally, achieving model interpretability—so as to provide meaningful insights into the underlying chemical phenomena—requires a careful balance between model complexity and transparency. Addressing these challenges is crucial for advancing the state of SpectraML and ensuring that ML-driven approaches can fully harness the rich information contained in spectral data.

4.2 Opportunities and Emerging Paradigms

Synthetic Data Generation and Physics-Informed Methods. To address challenges stemming from scarce and variable spectral data, AI-based generative models—such as large language models (LLMs) and diffusion models—have emerged as effective tools for advanced data augmentation [Luo *et al.*, 2023; Bushuiev *et al.*, 2024]. These models can learn complex, high-dimensional distributions from experimental spectra and subsequently generate synthetic data that mirrors key structural patterns and nonlinear dependencies [Alberts *et al.*, 2024a; Alberts *et al.*, 2024b]. Such synthetic spectra can supplement limited training sets, improving model robustness and generalization.

Incorporating physics-informed constraints—such as conservation laws, known intensity ratios, and chemical shift rules—into the generative process further enhances the realism and interpretability of the outputs. These priors guide the model toward producing chemically valid spectra, offering a compelling alternative to traditional simulation techniques like DFT or molecular dynamics. This hybrid strategy not only mitigates data scarcity but also reduces computational cost, facilitating faster iteration in applications ranging from materials science to pharmaceutical research.

Foundation Models: A New Paradigm for SpectraML. Foundation models [Bommasani *et al.*, 2021; Moor *et al.*, 2023], trained on large-scale, multimodal spectral datasets (e.g., IR, NMR, MS, Raman), offer a unified framework for spectral analysis. By leveraging cross-domain learning, they capture both global chemical patterns and local spectral signatures, supporting few-shot and zero-shot adaptation for diverse tasks such as peak prediction, spectral reconstruction, and structure inference.

These models integrate domain-specific priors directly into their architectures, enabling both forward and inverse reasoning. Their advanced inference capabilities allow them to resolve ambiguous or overlapping spectral signals through multi-step reasoning and modality fusion. Built-in mechanisms for uncertainty quantification and error detection enhance model reliability and interpretability. Overall, foundation models provide a flexible, data-efficient, and explainable approach to tackling both conventional and novel challenges in spectral learning.

5 Conclusion

The SpectraML establishes a crucial intersection between machine learning and spectroscopy. In this work, we provide a comprehensive overview of SpectraML and present a unified roadmap that traces methodologies across multiple spectroscopic techniques and categorizes key advancements in forward and inverse problems. To support future research, we highlight emerging trends such as generative modeling and foundation models and release an open-source repository. This survey serves as a valuable resource for researchers in both chemistry and AI, fostering interdisciplinary collaboration and driving innovation in spectral analysis.

Acknowledgments

This work was supported by the National Science Foundation under the NSF Center for Computer Assisted Synthesis (C-CAS), grant number CHE-2202693.

References

[Alberts *et al.*, 2023] Marvin Alberts, Federico Zipoli, and Alain C Vaucher. Learning the language of NMR: Structure elucidation from NMR spectra using transformer models. *ChemRxiv*, 2023.

[Alberts *et al.*, 2024a] Marvin Alberts, Teodoro Laino, and Alain C Vaucher. Leveraging infrared spectroscopy for automated structure elucidation. *Communications Chemistry*, 7(1):268, 2024.

[Alberts *et al.*, 2024b] Marvin Alberts, Oliver Schilter, Federico Zipoli, Nina Hartrampf, and Teodoro Laino. Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry. *Advances in Neural Information Processing Systems*, 37:125780–125808, 2024.

[Beck *et al.*, 2024] Armen G. Beck, Matthew Muhoberac, Caitlin E. Randolph, Connor H. Beveridge, Prageeth R. Wijewardhane, Hilkka I. Kenttämaa, and Gaurav Chopra. Recent developments in machine learning for mass spectrometry. *ACS Measurement Science Au*, 4(3):233–246, 2024.

[Binev *et al.*, 2007] Yuri Binev, Maria M. B. Marques, and João Aires-de Sousa. Prediction of 1H NMR coupling constants with associative neural networks trained for chemical shifts. *Journal of Chemical Information and Modeling*, 47(6):2089–2097, 2007.

[Boiko *et al.*, 2022] Daniil A Boiko, Konstantin S Kozlov, Julia V Burykina, Valentina V Ilyushenkova, and Valentine P Ananikov. Fully automated unconstrained analysis of high-resolution mass spectrometry data with machine learning. *Journal of the American Chemical Society*, 144(32):14590–14606, 2022.

[Bommasani *et al.*, 2021] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[Bushuiev *et al.*, 2024] Roman Bushuiev, Anton Bushuiev, Niek F de Jonge, Adamo Young, Fleming Kretschmer, Raman Samusevich, Janne Heirman, Fei Wang, Luke Zhang, Kai Dührkop, Marcus Ludwig, Nils A Haupt, Apurva Kalia, Corinna Brungs, Robin Schmid, Russell Greiner, Bo Wang, David S Wishart, Li-Ping Liu, Juho Rousu, Wout Bittremieux, Hannes Rost, Tytus D Mak, Soha Hassoun, Florian Huber, Justin JJ van der Hooft, Michael A Stravs, Sebastian Böcker, Josef Sivic, and Tomáš Pluskal. MassSpecGym: A benchmark for the discovery and identification of molecules. *Advances in Neural Information Processing Systems*, 37:110010–110027, 2024.

[Chen *et al.*, 2022] Zekun Chen, Fernanda C. Bononi, Charles A. Sievers, Wang-Yeuk Kong, and Davide Donadio. UV–Visible absorption spectra of solvated molecules by quantum chemical machine learning. *Journal of Chemical Theory and Computation*, 18(8):4891–4902, 2022.

[Chen *et al.*, 2024] Yuzhuo Chen, Sebastian V Pios, Maxim F Gelin, and Lipeng Chen. Accelerating molecular vibrational spectra simulations with a physically informed deep learning model. *Journal of Chemical Theory and Computation*, 20(11):4703–4710, 2024.

[Coley *et al.*, 2019] Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical Science*, 10(2):370–377, 2019.

[Cordova *et al.*, 2023] Manuel Cordova, Pinelopi Moutzouri, Bruno Simões de Almeida, Daria Torodii, and Lyndon Emsley. Pure isotropic proton NMR spectra in solids using deep learning. *Angewandte Chemie International Edition*, 62(8):e202216607, 2023.

[Du *et al.*, 2023] Wenjie Du, Fenfen Ma, Baicheng Zhang, Jiahui Zhang, Di Wu, Edward Sharman, Jun Jiang, and Yang Wang. Spectroscopy-guided deep learning predicts solid–liquid surface adsorbate properties in unseen solvents. *Journal of the American Chemical Society*, 146(1):811–823, 2023.

[Dührkop *et al.*, 2020] Kai Dührkop, Louis-Félix Nothias, Markus Fleischauer, Raphael Reher, Marcus Ludwig, Martin A. Hoffmann, William H. Petras, Daniel ans Gerwick, Juho Rousu, Pieter C. Dorrestein, and Sebastian Böcker. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology*, 39(4):462–471, 2020.

[Elias *et al.*, 2004] Joshua E Elias, Francis D Gibbons, Oliver D King, Frederick P Roth, and Steven P Gygi. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature biotechnology*, 22(2):214–219, 2004.

[Enders *et al.*, 2021] Abigail A Enders, Nicole M North, Chase M Fensore, Juan Velez-Alvarez, and Heather C Allen. Functional group identification for ftir spectra using image-based machine learning models. *Analytical Chemistry*, 93(28):9711–9718, 2021.

[Fine *et al.*, 2020a] Jonathan Fine, Judy Kuan-Yu Liu, Armen Beck, Kawthar Z Alzarieni, Xin Ma, Victoria M Boulos, Hilkka I Kenttämaa, and Gaurav Chopra. Graph-based machine learning interprets and predicts diagnostic isomer-selective ion–molecule reactions in tandem mass spectrometry. *Chemical Science*, 11(43):11849–11858, 2020.

[Fine *et al.*, 2020b] Jonathan A Fine, Anand A Rajasekar, Krupal P Jethava, and Gaurav Chopra. Spectral deep learning for prediction and prospective validation of functional groups. *Chemical Science*, 11(18):4618–4630, 2020.

[Gabriel *et al.*, 2024] Wassim Gabriel, Omar Shouman, Eva Ayla Schröder, Florian Bößl, and Mathias Wilhelm. Prospect ptms: Rich labeled tandem mass spectrometry dataset of modified peptides for machine learning in proteomics. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[Gastegger *et al.*, 2017] Michael Gastegger, Jörg Behler, and Philipp Marquetand. Machine learning molecular dynamics for the simulation of infrared spectra. *Chemical Science*, 8:6924–6935, 2017.

[Gastegger *et al.*, 2021] Michael Gastegger, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of solvent effects on molecular spectra and reactions. *Chemical Science*, 12(34):11473–11483, 2021.

[Gerrard *et al.*, 2020] Will Gerrard, Lars A. Bratholm, Martin J. Packer, Adrian J. Mulholland, David R. Glowacki,

and Craig P. Butts. Impression – prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chemical Science*, 11:508–515, 2020.

[Goldman *et al.*, 2023] Samuel Goldman, John Bradshaw, Jiayi Xin, and Connor Coley. Prefix-tree decoding for predicting mass spectra from molecules. *Advances in Neural Information Processing Systems*, 36:48548–48572, 2023.

[Goldman *et al.*, 2024] Samuel Goldman, Janet Li, and Connor W Coley. Generating molecular fragmentation graphs with autoregressive neural networks. *Analytical Chemistry*, 96(8):3419–3428, 2024.

[Guan *et al.*, 2021] Yanfei Guan, SV Shree Sowndarya, Liliana C Gallegos, Peter C St John, and Robert S Paton. Real-time prediction of 1 H and 13 C chemical shifts with DFT accuracy using a 3D graph neural network. *Chemical Science*, 12(36):12012–12026, 2021.

[Guo *et al.*, 2021] Zhichun Guo, Chuxu Zhang, Wenhao Yu, John Herr, Olaf Wiest, Meng Jiang, and Nitesh V Chawla. Few-shot graph learning for molecular property prediction. In *Proceedings of the web conference 2021*, pages 2559–2567, 2021.

[Guo *et al.*, 2024] Kehan Guo, Bozhao Nan, Yujun Zhou, Taicheng Guo, Zhichun Guo, Mihir Surve, Zhenwen Liang, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Can LLMs solve molecule puzzles? A multimodal benchmark for molecular structure elucidation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[Ho Manh *et al.*, 2024] Linh Ho Manh, Victoria C. P. Chen, Jay Rosenberger, Shouyi Wang, Yujing Yang, and Kevin A. Schug. Prediction of vacuum ultraviolet/ultraviolet gas-phase absorption spectra using molecular feature representations and machine learning. *Journal of Chemical Information and Modeling*, 64(14):5547–5556, 2024.

[Hu *et al.*, 2024] Frank Hu, Michael S Chen, Grant M Rotskoff, Matthew W Kanan, and Thomas E Markland. Accurate and efficient structure elucidation from routine one-dimensional NMR spectra using multitask machine learning. *ACS Central Science*, 10(11):2162–2170, 2024.

[Huang *et al.*, 2021] Zhaorui Huang, Michael S Chen, Cristian P Woroch, Thomas E Markland, and Matthew W Kanan. A framework for automated structure elucidation from routine NMR spectra. *Chemical Science*, 12(46):15329–15338, 2021.

[Jonas and Kuhn, 2019] Eric Jonas and Stefan Kuhn. Rapid prediction of NMR spectral properties with quantified uncertainty. *Journal of cheminformatics*, 11:1–7, 2019.

[Jonas, 2019] Eric Jonas. Deep imitation learning for molecular inverse problems. *Advances in neural information processing systems*, 32, 2019.

[Kang *et al.*, 2020] Seokho Kang, Youngchun Kwon, Dongseon Lee, and Youn-Suk Choi. Predictive modeling of NMR chemical shifts without using atomic-level annotations. *Journal of Chemical Information and Modeling*, 60(8):3765–3769, 2020.

[Kim *et al.*, 2023] Hyun Woo Kim, Chen Zhang, Raphael Reher, Mingxun Wang, Kelsey L Alexander, Louis-Félix Nothias, Yoo Kyong Han, Hyeji Shin, Ki Yong Lee, Kyu Hyeong Lee, et al. DeepSAT: learning molecular structures from nuclear magnetic resonance data. *Journal of Cheminformatics*, 15(1):71, 2023.

[Kwon *et al.*, 2020] Youngchun Kwon, Dongseon Lee, Youn-Suk Choi, Myeonginn Kang, and Seokho Kang. Neural message passing for NMR chemical shift prediction. *Journal of chemical information and modeling*, 60(4):2024–2030, 2020.

[Leniak *et al.*, 2024] Arkadiusz Leniak, Wojciech Pietrus, and Rafał Kurczab. From NMR to AI: designing a novel chemical representation to enhance machine learning predictions of physicochemical properties. *Journal of Chemical Information and Modeling*, 64(8):3302–3321, 2024.

[Lewis, 2019] Mike Lewis. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[Lin *et al.*, 2022] Min Lin, Jingfang Xiong, Mintao Su, Feng Wang, Xiangsi Liu, Yifan Hou, Riqiang Fu, Yong Yang, and Jun Cheng. A machine learning protocol for revealing ion transport mechanisms from dynamic NMR shifts in paramagnetic battery materials. *Chemical Science*, 13:7863–7872, 2022.

[Lu *et al.*, 2024] Xin-Yu Lu, Hao-Ping Wu, Hao Ma, Hui Li, Jia Li, Yan-Ti Liu, Zheng-Yan Pan, Yi Xie, Lei Wang, Bin Ren, and Guo-Kun Liu. Deep learning-assisted spectrum-structure correlation: State-of-the-art and perspectives. *Analytical Chemistry*, 96(20):7959–7975, 2024.

[Luo *et al.*, 2023] Tianze Luo, Zhanfeng Mo, and Sinno Jialin Pan. Fast graph generation via spectral diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3496–3508, 2023.

[Mandelare *et al.*, 2018] P.E. Mandelare, D.A. Adpressa, E.N. Kaweesa, L.N. Zakharov, and S. Loesgen. Coculture of two developmental stages of a marine-derived aspergillus alliaceus results in the production of the cytotoxic bianthrone allianthrone A. *Journal of Natural Products*, 81(4):1014–1022, 2018.

[McNaughton *et al.*, 2023] Andrew D McNaughton, Rajendra P Joshi, Carter R Knutson, Anubhav Fnu, Kevin J Luebke, Jeremiah P Malerich, Peter B Madrid, and Neeraj Kumar. Machine learning models for predicting molecular UV–Vis spectra with quantum mechanical properties. *Journal of Chemical Information and Modeling*, 63(5):1462–1471, 2023.

[Moor *et al.*, 2023] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[Murphy *et al.*, 2023] Michael Murphy, Stefanie Jegelka, Ernest Fraenkel, Tobias Kind, David Healey, and Thomas Butler. Efficiently predicting high resolution mass spectra with graph neural networks. In *International Conference on Machine Learning*, pages 25549–25562. PMLR, 2023.

[O'Shea, 2015] K O'Shea. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[Park *et al.*, 2024] Jiwon Park, Jeonghee Jo, and Sungroh Yoon. Mass spectra prediction with structural motif-based graph neural networks. *Scientific Reports*, 14(1):1400, 2024.

[Ralbovsky and Lednev, 2020] Nicole M Ralbovsky and Igor K Lednev. Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning. *Chemical Society Reviews*, 49(20):7428–7453, 2020.

[Schuster and Paliwal, 1997] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[Schütt *et al.*, 2021] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR, 2021.

[Singh *et al.*, 2022] Kanishka Singh, Jannes Munchmeyer, Leon Weber, Ulf Leser, and Annika Bande. Graph neural networks for learning molecular excitation spectra. *Journal of Chemical Theory and Computation*, 18(7):4408–4417, 2022.

[Sowa and Rossky, 2024] Jakub K. Sowa and Peter J. Rossky. A bond-based machine learning model for molecular polarizabilities and a priori raman spectra. *Journal of Chemical Theory and Computation*, 20(22):10071–10079, 2024.

[Sridharan *et al.*, 2022a] Bhuvanesh Sridharan, Manan Goel, and U. Deva Priyakumar. Modern machine learning for trackling inverse problems in chemistry: molecular desgin to realization. *Chemical Communication*, 58(35):5316–5331, 2022.

[Sridharan *et al.*, 2022b] Bhuvanesh Sridharan, Sarvesh Mehta, Yashaswi Pathak, and U Deva Priyakumar. Deep reinforcement learning for molecular inverse problem of nuclear magnetic resonance spectra to molecular structure. *The Journal of Physical Chemistry Letters*, 13(22):4924–4933, 2022.

[Su *et al.*, 2024] Yuming Su, Xue Wang, Yuanxiang Ye, Yibo Xie, Yujing Xu, Yibin Jiang, and Cheng Wang. Automation and machine learning augmented by large language models in a catalysis study. *Chemical Science*, 15(31):12200–12233, 2024.

[Sun *et al.*, 2024] Hanyu Sun, Xi Xue, Xue Liu, Hai-Yu Hu, Yafeng Deng, and Xiaojian Wang. Cross-modal retrieval between 13C NMR spectra and structures based on focused libraries. *Analytical Chemistry*, 96(15):5763–5770, 2024.

[Tian *et al.*, 2024] ZiJing Tian, Yan Dai, Feng Hu, ZiHao Shen, HongLing Xu, HongWen Zhang, JinHang Xu, YuTing Hu, YanYan Diao, and HongLin Li. Enhancing chemical reaction monitoring with a deep learning model for NMR spectra image matching to target compounds. *Journal of Chemical Information and Modeling*, 64(14):5624–5633, 2024.

[Urban *et al.*, 2024] James Urban, Chunsheng Jin, Kristina A Thomsson, Niclas G Karlsson, Callum M Ives, Elisa Fadda, and Daniel Bojar. Predicting glycan structure from tandem mass spectrometry via deep learning. *Nature Methods*, 21(7):1206–1215, 2024.

[Voronov *et al.*, 2022] Gennady Voronov, Rose Lightheart, Joe Davison, Christoph A Krettler, David Healey, and Thomas Butler. Multi-scale sinusoidal embeddings enable learning on high resolution mass spectrometry data. *arXiv preprint arXiv:2207.02980*, 2022.

[Wei *et al.*, 2019] Jennifer N Wei, David Belanger, Ryan P Adams, and D Sculley. Rapid prediction of electron–ionization mass spectrometry using neural networks. *ACS central science*, 5(4):700–708, 2019.

[Yan *et al.*, 2024] Yan Yan, Michael T Judge, Toby Athersuch, Yuchen Xiang, Zhaolu Liu, Beatriz Jiménez, and Timothy MD Ebbels. Resolution enhancement of metabolomic J-Res NMR spectra using deep learning. *Analytical Chemistry*, 96(29):11707–11715, 2024.

[Yao *et al.*, 2023] Lin Yao, Minjian Yang, Jianfei Song, Zhuo Yang, Hanyu Sun, Hui Shi, Xue Liu, Xiangyang Ji, Yafeng Deng, and Xiaojian Wang. Conditional molecular generation net enables automated structure elucidation based on 13C NMR spectra and prior knowledge. *Analytical chemistry*, 95(12):5393–5401, 2023.

[Ye *et al.*, 2020] Sheng Ye, Kai Zhong, Jinxiao Zhang, Wei Hu, Jonathan D. Hirst, Guozhen Zhang, Shaul Mukamel, and Jun Jiang. A machine learning protocol for predicting protein infrared spectra. *Journal of the American Chemical Society*, 142(45):19071–19077, 2020.

[Yilmaz *et al.*, 2022] M Yilmaz, WE Fondrie, W Bittremieux, S Oh, and WS Noble. De novo mass spectrometry peptide sequencing with a transformer model. biorxiv, 2022.02. 07.479481, 2022.

[Young *et al.*, 2024] Adamo Young, Bo Wang, and Hannes Röst. Tandem mass spectrum prediction for small molecules using graph transformers. *Nature Machine Intelligence*, 6(4):404–416, 2024.

[Zheng *et al.*, 2024] Xiaoshan Zheng, Boyi Tang, Peng Xu, Youmei Wang, Bin Di, Zhendong Hua, Mengxiang Su, and Jun Liao. Cbmaff-net: An intelligent NMR-based nontargeted screening method for new psychoactive substances. *Analytical Chemistry*, 96(47):18672–18680, 2024.

[Zhu and Jonas, 2023] Richard Licheng Zhu and Eric Jonas. Rapid approximate subset-based spectra prediction for electron ionization–mass spectrometry. *Analytical chemistry*, 95(5):2653–2663, 2023.

[Zhu *et al.*, 2020] Hao Zhu, Liping Liu, and Soha Hassoun. Using graph neural networks for mass spectrometry prediction. *arXiv preprint arXiv:2010.04661*, 2020.