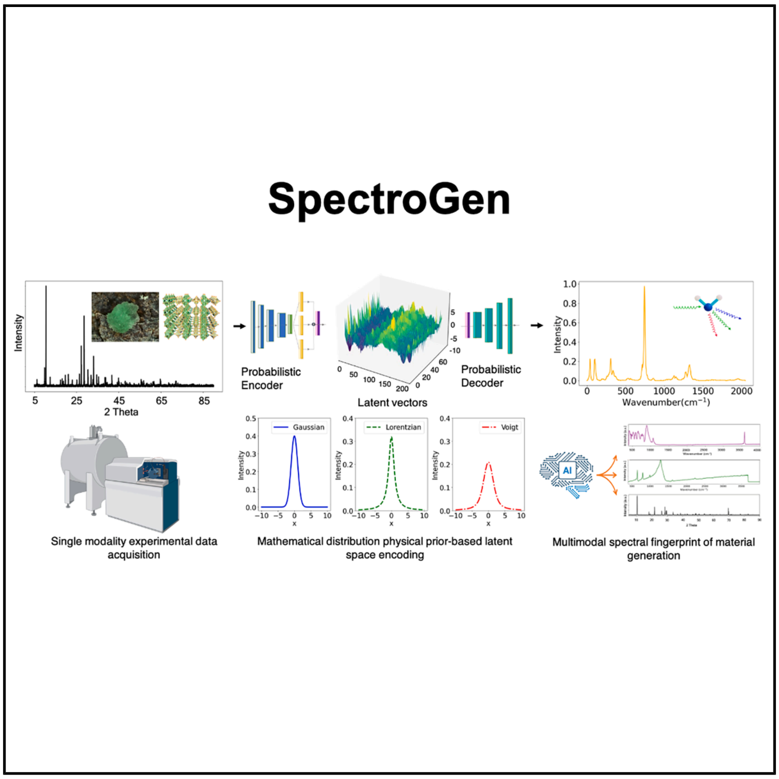
Matter

Article

SpectroGen: A physically informed generative artificial intelligence for accelerated cross-modality spectroscopic materials characterization

Graphical abstract



Highlights

- Physics-inspired models and autoencoders unify for fast "virtual" spectra
- Realistic spectra enable rapid screening in materials discovery
- Universal platform accommodates any spectroscopic technique
- Accelerates AI-driven discovery and verification for vital societal breakthroughs

Authors

Yanmin Zhu, Loza F. Tadesse

Correspondence

lozat@mit.edu

In brief

SpectroGen seamlessly couples physics-driven distribution models with a variable autoencoder to generate synthetic spectra indistinguishable from real data. By speeding up high-throughput screening, it closes the gap between AI-based materials discovery and experimental confirmation. Its flexible architecture accommodates diverse spectroscopic techniques, extending its utility across multiple scientific domains. The synergy of rapid AI-driven design and swift AI-enabled characterization expedites validation of innovative materials, bridging lab-based discovery and industry-ready applications to address urgent societal needs.

2
Benchmark
First qualification/assessment of material properties and/or performance

Zhu & Tadesse, 2026, Matter 9, 102434
January 7, 2026 © 2025 Published by Elsevier Inc.
https://doi.org/10.1016/j.matt.2025.102434

CellPress

Matter

CellPress

Article

SpectroGen: A physically informed generative artificial intelligence for accelerated cross-modality spectroscopic materials characterization

Yanmin Zhu[1] and Loza F. Tadesse[1,2,3,4,*]

[1]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

[2]Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA

[3]Jameel Clinic for AI & Healthcare, Massachusetts Institute of Technology, Cambridge, MA, USA

[4]Lead contact

*Correspondence: lozat@mit.edu

https://doi.org/10.1016/j.matt.2025.102434

PROGRESS AND POTENTIAL Recent advances in artificial intelligence (AI) have propelled materials discovery by identifying unique composition pathways at unprecedented speed. However, experimental characterization—the step where new materials are actually tested—still lags behind. Traditional characterization requires specialized instruments that measure electromagnetic responses in a painstaking, expert-driven process. SpectroGen offers a transformative solution. By coupling physics-inspired distribution models (e.g., Gaussians and Lorentzians) with a robust variable autoencoder framework, SpectroGen rapidly generates “virtual” spectra that correlate almost perfectly with actual measurements. This approach effectively bridges the gap between AI-driven materials discovery and real-world verification. SpectroGen’s universal compatibility also makes it flexible: any spectroscopy technique that can be represented by analytic functions may be harnessed within its platform.

The potential impact is substantial. High-throughput screening—vital for developing next-generation catalysts, batteries, superconductors, and pharmaceuticals—can now be accelerated without sacrificing accuracy. Researchers stand to gain significant time and resource savings, as they can prioritize the most promising candidate materials for detailed follow-up. This synergy of fast AI-driven discovery and swift AI-enabled characterization could catalyze breakthroughs vital to society, from clean energy solutions to advanced medical treatments. Beyond accelerating fundamental research, SpectroGen’s capacity for rapid prototyping and validation is poised to reshape how we innovate, ultimately translating into critically needed technologies that better serve humanity.

SUMMARY

Artificial intelligence (AI)-driven materials discovery offers rapid design of novel material compositions, yet synthesis and characterization lag behind. Characterization, in particular, remains bottlenecked by labor-intensive experiments using expert-operated instruments that typically rely on electromagnetic spectroscopy. We introduce SpectroGen, a generative AI model for transmodality spectral generation, designed to accelerate materials characterization. SpectroGen generates high-resolution, high-signal-to-noise ratio spectra with 99% correlation to ground truth and a root-mean-square error of 0.01 a.u. Its performance is driven by two key innovations: (1) a novel distribution-based physical prior and (2) a variational autoencoder (VAE) architecture. The prior simplifies complex structural inputs into interpretable Gaussian or Lorentzian distributions, while the VAE maps them into a physically grounded latent space for accurate spectral transformation. SpectroGen generalizes across spectral domains and promises rapid, accurate spectral predictions, potentially transforming high-throughput discovery in domains such as battery materials, catalysts, superconductors, and pharmaceuticals.

INTRODUCTION

High-throughput materials discovery aims to accelerate the identification and optimization of novel materials with fit-for-purpose exceptional properties.[1] However, the rate of synthesis and characterization of computer-generated material candidates remains a significant challenge to the need for speed. Spectroscopy-based modalities are the core of materials characterization, as they enable molecular- and atomic-level analysis.[2–5] For example, infrared (IR) absorption reveals functional groups within molecules; Raman scattering provides insights into molecular vibrations, symmetry, and crystal structures;[6–9] X-ray scattering determines the elemental composition; and X-ray diffraction (XRD) visualizes crystal structures.[10,11] However, it is a time-consuming and expensive endeavor, costing up to half a million USD per equipment and domain expertise for data interpretation. Furthermore, samples are often scarce, fragile, and/or hazardous, limiting experimental interrogation.[12] These limitations present significant challenges to achieving high-throughput characterization—an essential element for keeping pace with the speed of computational material generation. Thus, there is a need for a paradigm shift in how materials are characterized.

Conventional deep learning techniques, which focus primarily on classification and regression tasks, have been employed for performance prediction,[13] the analysis and preprocessing of microscopic images and spectroscopic data,[14–17] and the design of novel materials.[18] Furthermore, physics-informed neural networks, which embed physical principles, e.g., conservation laws, differential equations, and boundary conditions into the network architecture or loss function,[19] are reducing the "black-box" nature of such models and enhancing interpretability. Recently, generative artificial intelligence (AI) algorithms, such as variational autoencoder (VAE),[20] which learns a probabilistic representation of the data by encoding the input as a distribution in the latent space, thus capturing inherent uncertainty in the data and enabling more accurate generation, are being introduced. VAEs are ideal as compared to traditional deterministic methods (such as autoencoders or regression models) that directly map inputs to outputs: they provide a means to learn the uncertainty involved in the spectral transformation process, and unlike other generative networks (e.g., generative adversarial networks [GANs]), VAEs learn the latent representation of the data, making them particularly suitable for applications where understanding the underlying data structure and generating diverse yet plausible outputs are critical. VAEs have been successfully used in gene editing,[21] protein design,[22] drug discovery,[23] and inverse design of solid-state materials.[24] However, almost all the output data from these VAE models are new structures for discovery applications, and implementation in transmodality transfer applications is yet to be demonstrated. Furthermore, the generated data fidelity of existing VAEs is questionable, whereas materials characterization, such as spectral data generation, requires high data fidelity, as the generated spectra need to match otherwise experimentally collected fingerprints. Thus, there is a need for a high-fidelity custom generative AI model that can address the following two critical challenges: (1) a computationally efficient representation of material structure-to-characterization output pairs and (2) a simple interpretation of the said characterization output for the network to understand and train on. A notable example of a successful generative AI implementation task that addressed these challenges is the protein structure generation tool AlphaFold,[25] which received the 2024 Nobel Prize. Here, instead of the daunting, nearly impossible task of computing every molecule representation and intermolecular force at play, the task of generation was creatively simplified to a triangle inequality problem, which the algorithm superbly optimized, leading to highly accurate predictions. A similar creative approach is needed for transmodality generation for applications in materials characterization.

Here, we introduce SpectroGen, a custom generative AI model that can computationally generate high-resolution spectra from multiple types of spectroscopic techniques using only a single spectroscopy modality experimental input (Figure 1), enabling high-throughput materials characterization. We demonstrate successful transmodality spectral generation with 99% correlation to experimental results by (1) creatively representing spectral data as a mathematical distribution curve, such as a Lorentzian, Gaussian, or Voigt distribution, to represent Raman, IR, and XRD spectra instead of computationally dense molecular and crystal structure inputs and (2) building a physical-prior-informed custom variable autoencoder-based generative algorithm. Our model outputs spectral transformations that are both physically meaningful and computationally accurate, sensitively accounting for characteristics like line broadening, superposition, and wavenumber shifts (Figure 2). Our mathematical distribution-based physical priors act as fundamental constraints, successfully capturing the inherent complexity of the fingerprint, enhancing the interpretability of the model, and reducing its black-box nature. We demonstrate SpectroGen on the RRUFF dataset,[26] comprising 6,006 International Mineralogical Association (IMA)-approved standard mineral samples (Figure 1B), from which 319 IR-Raman and 371 XRD-Raman data pairs were examined (Data S1–S4; Figure 3). We computed wavenumber shifts, peak heights, and the full width at half maximum (FWHM) of peaks to evaluate the accuracy of spectrum generation (Figure 4). Furthermore, we conducted a material source classification task to compare the classification accuracy of the generated spectra with that of experimentally collected spectra, thereby assessing the informational efficacy of the generated spectra (Figure 5; Notes S1 and S2; Data S5 and S6). SpectroGen exhibited 99% correlation on peak characteristics, a root-mean-square error (RMSE) of 0.01 on intensity (in arbitrary units [a.u.]), and a peak signal-to-noise ratio (PSNR) of 43 ± 4 dB normalized by the peak heights, compared to experimentally acquired ground-truth spectra. Moreover, it achieves a mean classification accuracy of 90.476% on the classification test, surpassing the 69.879% accuracy obtained from experimentally acquired Raman spectra. Overall, SpectroGen eliminates the need for multiple time- and resource-intensive spectroscopic instruments, revolutionizing materials characterization throughput. Furthermore, by abstracting spectra as "data" independent of physical material properties, we can decouple the generation of spectra from the constraints imposed by specific material behaviors or interactions. This "data as a link" concept positions the spectral

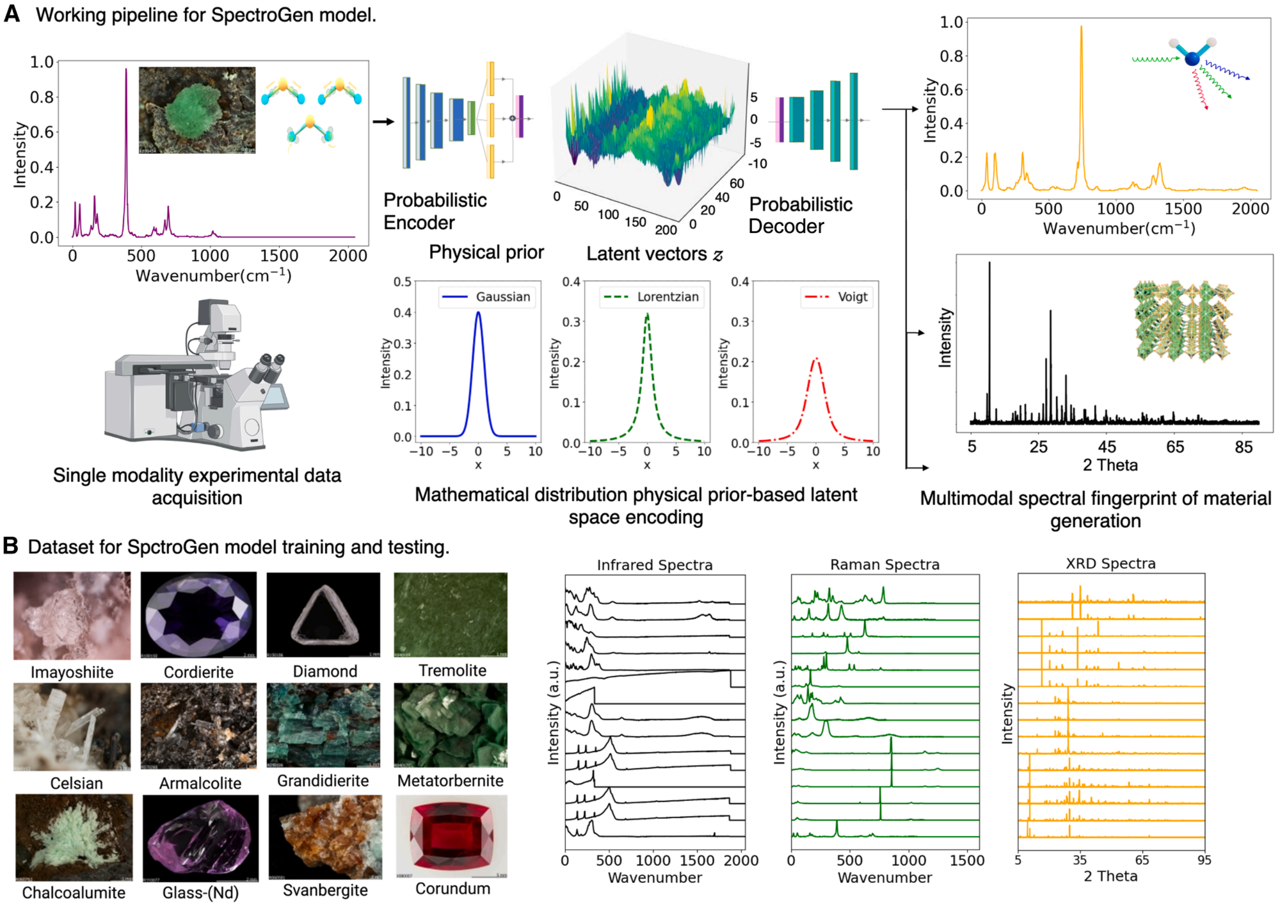


Figure 1. SpectroGen workflow

(A) Schematic showing the flow of spectroscopic modality transfer with SpectroGen. Experimental data acquired from one spectroscopic modality are input into the physical-prior-informed variational autoencoder model. Mathematical distributions of the spectroscopic curve (e.g., Gaussian distribution, Lorentzian distribution, and Voigt distribution) are used as a physical prior to represent the input and output spectra. Spectra output from another modality is then generated.

(B) The dataset used for SpectroGen model training and testing: images of example single-crystalline materials (left) and example infrared, Raman, and XRD spectra data pairs for material samples (right).

fingerprint as a bridge between the physical and computational domains, enabling SpectroGen to generate spectra with greater flexibility and efficiency through understanding spectra in a mathematical approach rather than deconstructing molecular structures and representations from spectra. We believe that SpectroGen could be transformational in bringing the much-needed throughput to match advances in material, accelerating the real-life translation of efficient materials and life-saving pharmaceuticals.

RESULTS

SpectroGen is as accurate as experimental collection

We demonstrate SpectroGen on the RRUFF dataset,[26] a publicly available IMA-approved standard mineral samples dataset, from which 319 IR-Raman and 371 XRD-Raman data pairs were examined (Data S1–S4). For the transformation from IR to Raman with a Gaussian distribution prior, SpectroGen precisely captures and reconstructs the information of 8 peaks in the barrerite Raman spectrum (Figure 3B) from their respective IR spectra (Figure 3A). Notably, the generated barrerite Raman spectrum matches the respective broadening and peak height and exhibits a smoother waveform. Similarly, for the actinolite sample (Figure 3C), the generated Raman spectrum closely follows the peak heights, peak number, and wavenumber of the experimentally collected data, with low residual values (Figure 3D). We evaluated the accuracy of peak positions by measuring their absolute peak position deviation (APPD), which all demonstrate low values smaller than 0.2 cm^{-1} (Figures S1 and S15). As for the XRD-Raman transformation test, the sample spectra of clinohumite show that the generated Raman spectrum precisely matches the peak positions and heights with lower noise compared to the experimentally collected Raman spectra, where the SNRs for the generated and experimentally collected Raman spectra are 11.10 and 3.11, respectively. For the demonstration of XRD-to-Raman transformation for cordierite, the generated Raman spectrum aligns with the peak heights and wavenumber shifts, showing low residual values. Here, we

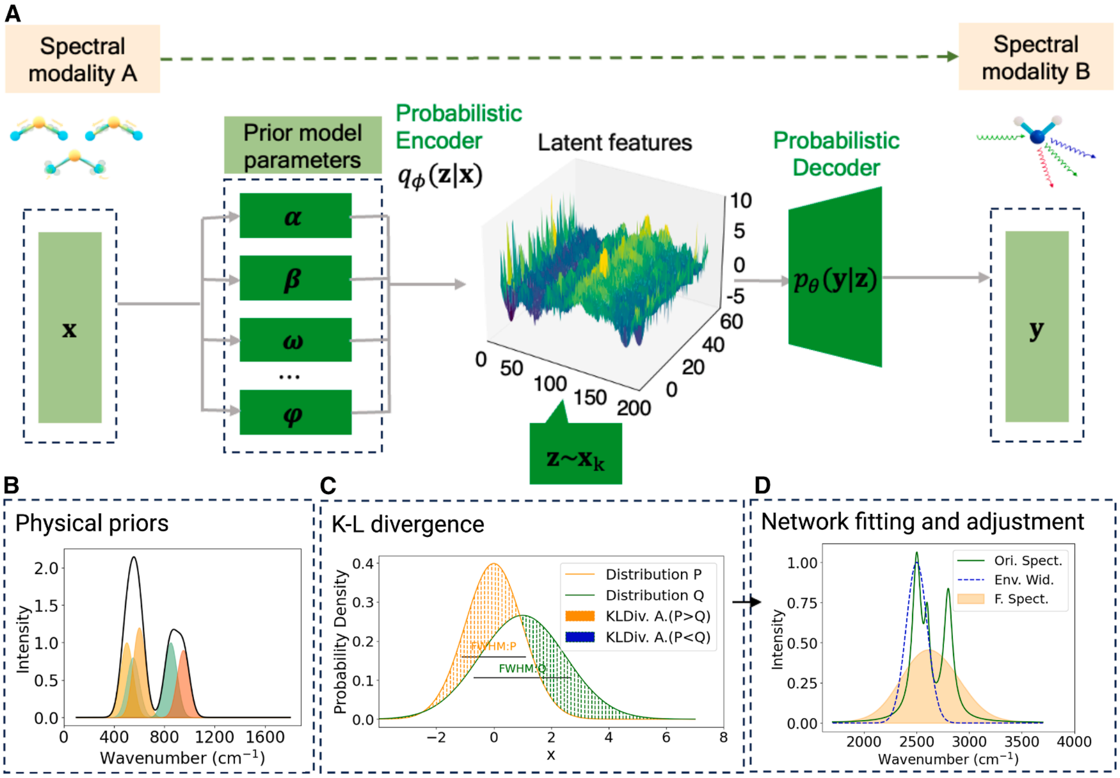


Figure 2. Modeling strategy

(A) SpectroGen employs an experimentally derived spectrum from modality A.

(B) Physical priors are employed to deconstruct the spectral distribution and map it to latent features $\boldsymbol{z}$ via a frequency encoder $\boldsymbol{q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ (Methods S3). SpectroGen designs a probabilistic decoder $\boldsymbol{p}_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z})$ to reconstruct the second (generated) spectral distribution that would have been experimentally collected from modality B.

(C) SpectroGen computes Kullback-Leibler (K-L) divergence to perform spectral fitting (Methods S5).

(D) The network's fitting capabilities are fully exploited to accommodate the irregular environmental broadening present in the spectra (Methods S4).

sampled four sets of results: two using a Gaussian distribution prior for the input of IR (Figures 3A–3D) with their corresponding residual value plots (supplemental methods, Equation S3-1) and two using a Voigt distribution prior for XRD spectra with their residual value plots (Figures 3E–3H; supplemental methods, Equation S3-3), and the complete generated spectral dataset is available in Data S1 and S20.

The quality of the generated spectrum was evaluated using spectrum-based and image-based evaluation metrics to assess its performance, as illustrated in Figure 4. We quantified the average peak height, FWHM, and SNR for all 97 and 110 test pairs of SpectroGen-generated and experimentally obtained Raman spectra in the IR-Raman and XRD-Raman datasets, respectively. We performed a Jensen-Shannon (JS) divergence test, which typically shows a relatively small value (e.g., <0.1), indicating strong alignment between distributions.[27] Structural similarity index (SSIM), RMSE, PSNR, and correlation assessments were performed as part of the image-based evaluation, which compares graphical structure and image content. SSIM is a dimensionless metric ranging from 0 to 1 and can evaluate pixel intensity, image structure, and context similarity, which, in our research task, refers to the evaluation of peak height, wavenumber shifts, and noise level. The value of SSIM approaches 1 as the similarity between the two images increases.[28] RMSE is calculated based on the spectra intensity (a.u.), and PSNR is normalized by the maximum intensity on the spectrum data. A lower RMSE value signals a smaller difference between the two spectra, while a PSNR exceeding 40 dB demonstrates a high degree of similarity of the generated spectrum to the reference spectrum.

As shown in Figures 4A and 4D and Table S1, the average peak height and FWHM of the generated spectra exhibit a similar distribution to that of the collected spectra, with JS divergences of 0.11 and 0.09 for IR-to-Raman and 0.05 and 0.06 for XRD-to-Raman tasks, respectively. Notably, the generated spectra, on average, possess a higher SNR compared to the experimentally collected spectra, which is consistent with the spectral examples provided in Figures 4B and 4E. In the residual plots, the generated Raman spectra exhibit small residual values, indicating minimal differences from the experimentally collected spectra. As depicted in Figure 4B and Table S2, for the IR-to-Raman task, the SpectroGen-generated spectrum has a mean SSIM of 0.96 ± 0.03, RMSE of 0.010 ± 0.006, correlation of 0.99 ± 0.01, and PSNR of 39 ± 4 dB. The XRD-to-Raman transformation task also shows similar outstanding performance. As shown in Figure 4E, SpectroGen displays a 0.97 ± 0.04 mean

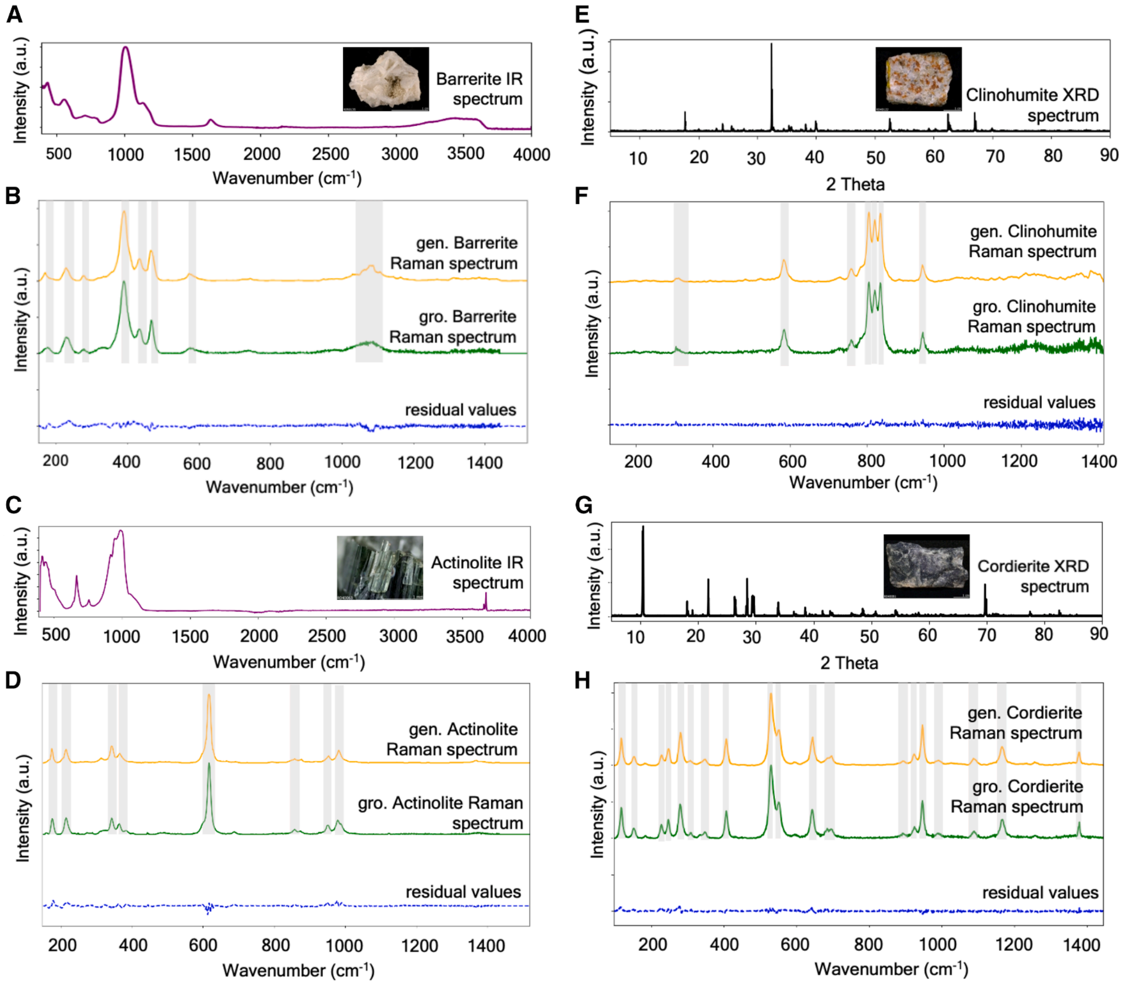


Figure 3. SpectroGen accurately generates spectra across different modalities

(A and B) Demonstration of IR-to-Raman transfer test with (A) barrerite IR spectrum and its material sample photo and (B) generated (yellow) and experimentally collected (green, termed as ground truth) barrerite Raman spectrum, showing alignment in peak positions and lower noise. The residual plot between the generated and experimentally collected Raman spectra shows low residual values across the wavelength range.

(C) Actinolite IR spectrum and its material sample photo.

(D) Generated actinolite Raman spectrum and ground-truth actinolite Raman spectrum. Peaks are correctly reconstructed with low residual values.

(E–H) Demonstration of XRD-to-Raman transfer tests with (E) clinohumite-input XRD spectrum and its material sample photo; (F) the generated and experimentally collected clinohumite Raman spectra, along with the residual plot between the two spectra; (G) cordierite XRD spectrum and its material sample photo; and (H) the generated and ground-truth cordierite Raman spectra with their residual values. SpectroGen accurately predicts the Raman spectrum from XRD inputs for both clinohumite and cordierite samples using a Voigt distribution prior, correctly identifying their peak locations with reduced noise levels, which correspond to low residual values across the wavelength range.

SSIM, a 0.010 ± 0.009 RMSE, a 43 ± 4 dB PSNR, and a 0.98 ± 0.08 correlation. In the area under the curve (AUC) test (see Figures 4C and 4F), for both the IR-to-Raman task and XRD-to-Raman task, the generated spectra are well aligned with experimentally collected spectra (Data S5–S9). These results demonstrate the exceptional similarity of the visually observed trajectory of the generated and experimentally obtained spectra.

SpectroGen is able to make accurate cross-modality spectral generation, primarily because of two key aspects. Firstly, the accurate physical priors were input to represent respective spectra from the modalities of interest, which removes the original model formulation constraints of the decoder. Secondly, the VAE backbone architecture employed was the best suited for matching curves. These combined strengths enable SpectroGen to achieve robust and precise spectral generation across diverse modalities.

Assessment of information transfer effectiveness via classification performance

To evaluate SpectroGen's information transfer effectiveness, we compared its performance on a material-type classification task using its generated spectra versus using experimentally obtained spectra. The diagonal values in a confusion matrix represent the number of correctly classified instances for each class. Each value

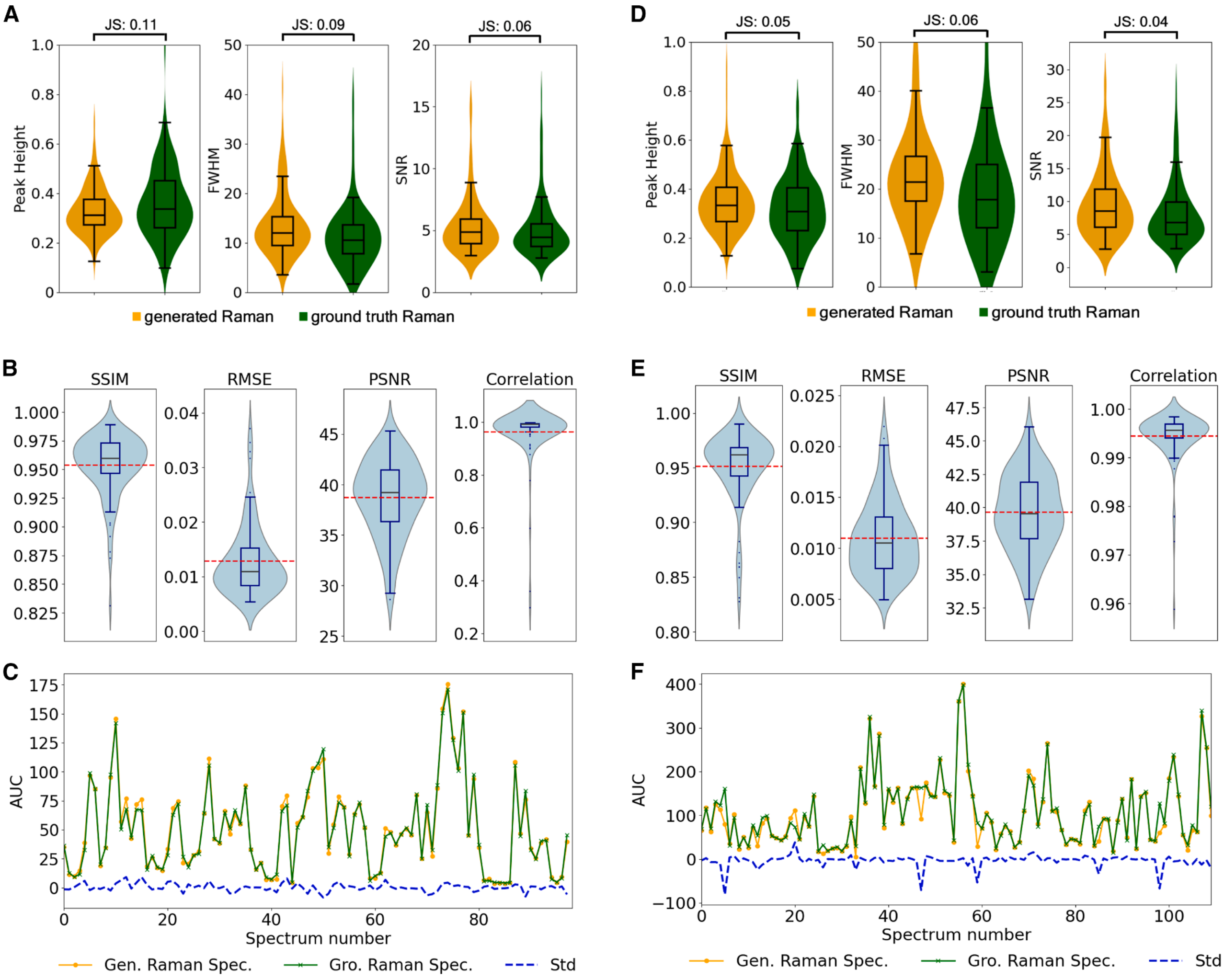


Figure 4. Spectral characteristic assessments were conducted for the entire generated datasets of IR-to-Raman and XRD-to-Raman tasks

(A) Average peak height, FWHM, and SNR of the generated Raman spectra (yellow) were compared to the experimentally collected Raman spectra (green). The Jensen-Shannon (JS) divergence was calculated to assess the similarity between the generated and experimentally collected spectra, showing values smaller than 0.11.

(B) SSIM, RMSE, PSNR, and correlation tests were performed between the generated Raman spectra and the experimentally collected Raman spectra, demonstrating that SpectroGen achieves high similarity in image-based assessments.

(C) AUC tests of the generated (Gen.) and experimentally collected (Gro.) Raman spectra, along with their standard deviations, revealed near-zero deviations.

(D) For the XRD-to-Raman transformation experiments, average peak height, FWHM, and SNR assessments showed strong alignment between the generated (yellow) and experimentally collected (green) Raman spectra datasets. JS divergences close to zero confirmed this alignment.

(E) SSIM, RMSE, PSNR, and correlation tests between the generated Raman spectra and the experimentally collected Raman spectra demonstrated that SpectroGen achieved excellent prediction performance on the XRD-to-Raman transformation task.

(F) AUC tests of the generated and ground-truth spectra also revealed a strong alignment.

on the diagonal corresponds to the count of samples where the predicted class matches the true class. As shown in Figures 5A and 5B, the confusion matrix reveals that the classification performance of generated spectra and experimentally collected spectra is similar for individual categories on a randomly selected training round. The correctly predicted samples of each class have similar values for the generated and experimentally collected spectra. (Detailed data from the 10 rounds of repetitive classification tests are available in Data S39–S68. The confusion matrices for the full dataset are shown in Figures S23 and S24.) As shown in Figures 5C and 5D, generated spectra achieved a mean accuracy of 90.476% across 26 categories of mineral materials (test set accuracy: 50.100%) for 10 rounds of repetitive classification tasks. Under identical network parameter conditions, the experimentally collected spectra had a mean classification accuracy of 69.879% (test set accuracy: 61.644%). Detailed data from the 10 rounds of repetitive classification tests are available in Data S39–S68. Even though it is beyond the scope of our current study, we generally observe poor classification performance due to the limited number of samples in the dataset; the majority of categories have

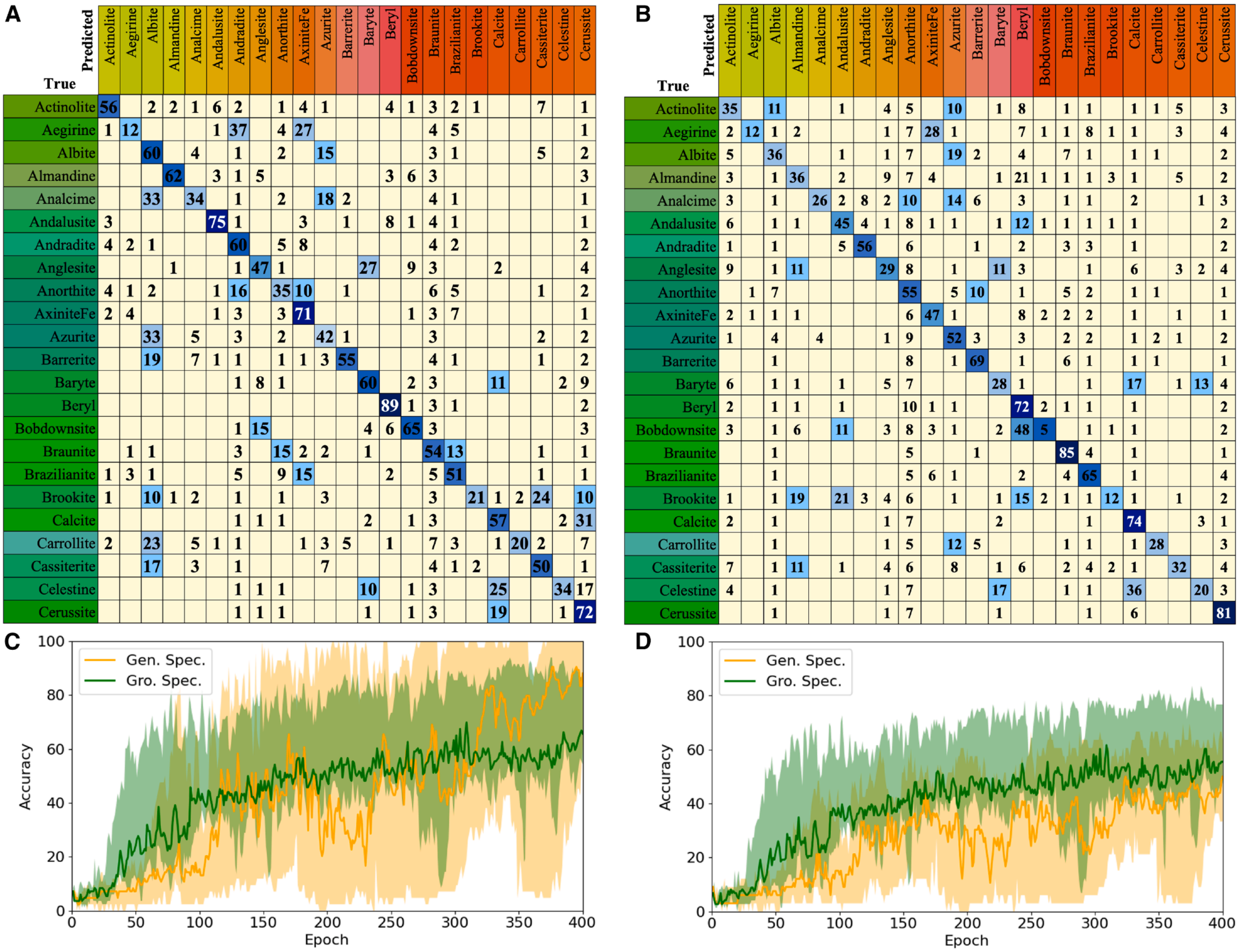


Figure 5. SpectroGen precisely transfers information, where generated spectra outperform experimental spectra

(A and B) Confusion matrix of classification test using (A) generated Raman spectra and (B) experimentally collected Raman spectra (confusion matrix for the full dataset is shown in Figures S23 and S24). The generated spectra deliver similar classification results to the experimentally collected spectra.

(C) Accuracy results based on the train set. Green line: ground-truth spectra; orange line: generated spectra. Both the generated and experimentally collected spectra improve their classification accuracy during the training process. Generated Raman spectra achieve competitive accuracy results compared to ground-truth spectra while reaching a higher final accuracy.

(D) Accuracy based on the test set. Both the generated and experimentally collected spectra show an increasing trend as the number of epochs increases. The spectrum generated with SpectroGen provides competitive information effectiveness compared to experimentally collected data. Both the generated and experimentally collected spectra show slightly lower accuracy than their training set.

fewer than five samples. We believe that the lower accuracy observed in the test set of the generated spectra, compared to the experimentally collected spectra, may be attributed to the instability in classification performance resulting from the small dataset size of much less than 10 spectra per material type. We expect this to improve significantly with a more substantial dataset. Overall, despite these constraints, this result effectively demonstrates SpectroGen's ability to transfer the fingerprint information that depicts molecular vibration.

SpectroGen interpretability test via physical prior distribution analysis

We validated the importance of the physical prior in the network by intentionally misrepresenting the respective spectra and their distribution, as shown in Figure 6. To this end, when incorrectly using a Lorentzian distribution as the physical prior for IR, we obtained an average peak height of 0.59, an average FWHM of 134.54, and an SNR of 47.69 for the generated Raman spectra, compared to an average peak height of 0.39, an average FWHM of 14.75, and an SNR of 5.22 for the experimentally collected Raman spectra. When XRD is incorrectly represented using a Gaussian distribution, the performance of SpectroGen on the generated Raman spectra similarly declines, yielding an average peak height of 0.27, an FWHM of 26.17, and an SNR of 12.87 compared to an average peak height of 0.24, an FWHM of 20.21, and an SNR of 7.88 for the experimentally collected Raman spectra. Similar drops also appear in image-based assessments (see Figures S4, S5, S12, and S13). These results

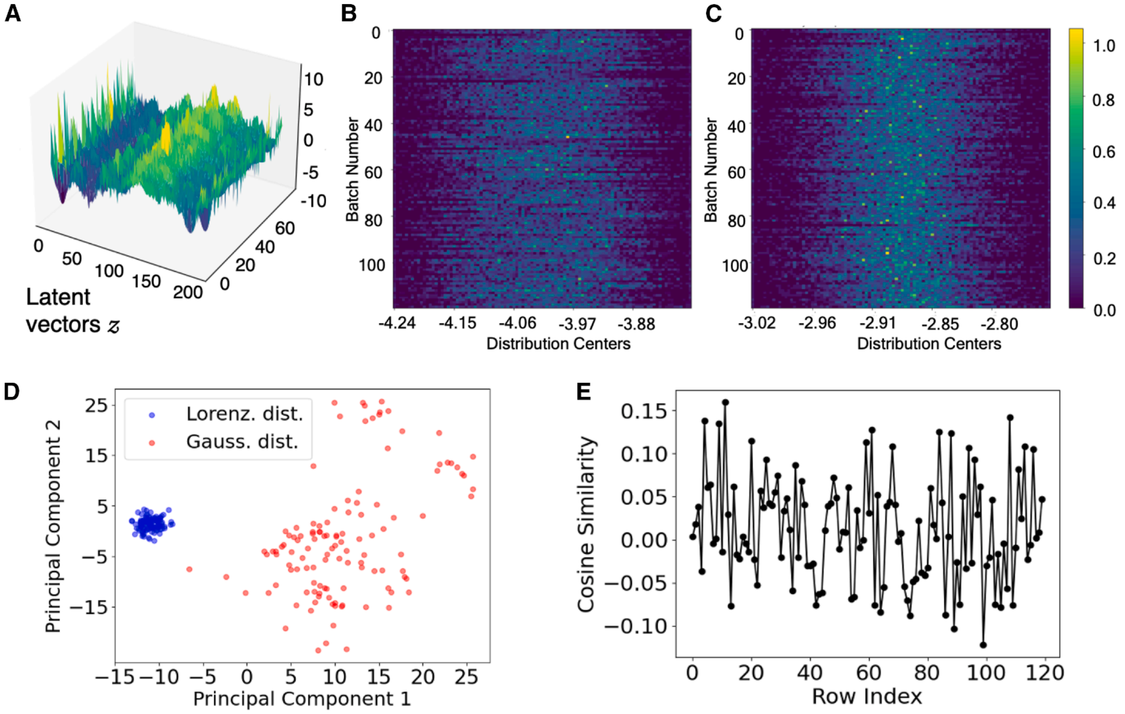


Figure 6. SpectroGen provides physical-prior-informed spectrum deconvolution and generation

(A) The probabilistic encoder represents the input spectra with latent vectors z with the guidance of physical priors. The probabilistic decoder learns the distribution from z to map the generated spectrum. Even with the same input, different physical priors will result in discrepancies in latent vectors (A) (the detailed network parameter is provided in Methods S6).

(B and C) Visualization of latent vectors of IR-Raman transfer with (B) Gaussian distribution prior and (C) Lorentzian distribution prior. Latent vectors show differences in width, distribution centers, and values for each training epoch between Gaussian and Lorentzian distribution-prior-guided experiments.

(D) Principal-component analysis for latent vector between Gaussian distribution prior and Lorentzian distribution prior. The Gaussian prior led to a dispersed distribution in principal components 1 and 2 for the IR-to-Raman transformation, compared to a concentrated distribution with the same input data for the Lorentzian prior-guided SpectroGen.

(E) Cosine similarity with Gaussian and Lorentzian distribution priors. Both positive and negative cosine similarities indicate significant differences in the latent features in terms of their magnitudes and directions.

underscore the critical role of physical prior models in the interpretability of the network, leading to precise generation, in contrast to a purely black-box approach, which relies solely on the network without incorporating physical priors.

To further elucidate the guidance and impact of the physical prior model on the network, we visualized the latent vector values (Figure 6) under the two previously tested conditions (IR-to-Raman transformation with Gaussian and Lorentzian priors). The results after 120 training epochs are shown in Figures 6B and 6C. We observed that different physical priors influenced the latent space in terms of their spatial distributions and values. We further performed principal-component analysis (PCA) and calculated the cosine similarity between two distributions, as presented in Figures 6D and 6E. As shown, we observed significant differences in the spatial distribution of latent features when using Gaussian versus Lorentzian distribution priors for the IR-to-Raman transformation task. The Gaussian prior led to a dispersed distribution in principal components 1 and 2, whereas the Lorentzian prior resulted in a concentrated distribution while using the same input data. This substitution not only changed the magnitude of latent feature values but also their distribution, affecting the generated results. Cosine similarity calculations revealed differences ranging from −0.15 to 0.20, highlighting significant variations in both the magnitude and direction of the latent features. A value of zero would indicate complete overlap between features.

This stress test demonstrates that physical priors are essential for maintaining the fidelity and accuracy of spectral transformations, as they guide the network to produce results that align with experimental data. A mismatch in physical priors could lead to deviations in the latent space and generated outputs, degrading performance and reducing interpretability. Our PCA of the latent space vectors obtained from IR-Raman spectral transformation using Gaussian and Lorentzian priors reveals that the choice of prior influences the distribution characteristics of the latent space representation for the same spectral data. Specifically, latent vectors derived from the Lorentzian prior exhibit a more concentrated distribution compared to those generated with the Gaussian prior. This suggests that the distribution of latent space vectors can serve as a reference for assessing the alignment of physical priors with the underlying spectral data.

DISCUSSION

It has long been recognized that the analysis of interactions between matter and electromagnetic waves illuminates structure, property, and function across broad fields such as biology, chemistry, and materials sciences. Using the latest advancements in generative deep learning, we demonstrate effective cross-modality spectral transfer with precise multi-dimensional molecular, structural, and other material property representations. SpectroGen, our physical-prior-informed deep generative model, achieves state-of-the-art performance in cross-domain spectral transformation, with generated spectra showing 99% average correlation and a 0.01 RMSE in intensity (in a.u.) compared to experimentally obtained spectra. Furthermore, transformation testing of multiple spectral modalities provides compelling evidence of SpectroGen's strong generalization capabilities. Experimental results indicate that, under the premise of objective, spectrum-based physical priors, we can accurately generate spectral data from another completely different spectroscopy modality. Our results demonstrate comparable peak ratio, FWHM, and AUC. Notably, we have statistically significant improvement in SNR compared to experimentally generated spectra, which led to a competitive classification test accuracy and 16% higher training performance.

We evaluate extrapolation performance by including spectra data excluded from the training set and adopt a random separation strategy for the training and test sets, with the test set potentially containing both seen and unseen crystal structures. To reduce overfitting and improve generalization, we employ normalization strategies during training and implement several safeguards: incorporating low-SNR spectra to simulate real-world noise, using a Gaussian distribution as the default prior for non-strictly matching spectra and spectra with an unknown prior, and applying regularization techniques.[29] Additionally, we explore zero-shot learning to transfer knowledge of unseen crystal structures from trained crystal categories.[30,31]

This first-of-its-kind demonstration promises spectroscopy implementation without the need for physical instrumentation, which is key to matching the pace of AI-enabled materials discovery efforts. In addition, our approach is key for research work where sample-specific experimental challenges, such as active specimens or *in vivo* biological samples, impose considerable limitations on spectral acquisition. By treating spectral data as an abstract mathematical distribution representation, our model enables the generation of spectra independent of the canonical physical representation of bonds and crystal structures, traditionally tied to specific materials. This abstraction allows SpectroGen to bridge the gap between physical experimentation and computational analysis, expanding the versatility of spectral generation across various domains. The success of AlphaFold similarly underscores the importance of physical priors, as it incorporates biochemical and physical constraints to bridge the gap between raw data and the complex rules governing protein folding. Without such priors, both SpectroGen and AlphaFold would lack the necessary physical grounding, leading to reduced fidelity, interpretability, and generalization. This highlights the indispensable role of physical priors in advancing computational approaches to complex systems, including vibrational spectroscopy and molecular biology.

SpectroGen effectively revolutionizes the reach of spectroscopy-based analysis across disciplines and research areas. Our study demonstrates that computational technologies can be integrated with the principles of spectroscopy to provide potential solutions for addressing challenges in materials characterization. This approach holds significant potential to advance the applications of spectroscopy in areas such as molecular structure analysis, material performance prediction, and biomolecular dynamic monitoring. By enhancing the interpretive capabilities of spectroscopic data, improving cross-modal applications, and driving the intelligent evolution of traditional high-precision methodologies, it paves the way for technological innovation. In addition, the spectral modalities are governed by shared principles of light-matter interactions,[2,4,5,8,9] and their complementary strengths can be harnessed through AI to provide a more holistic understanding of materials and molecules. Further study using surface-sensitive spectral information such as surface-enhanced Raman spectroscopy,[10,32] X-ray photoelectron spectroscopy,[33] and others will provide insight into surface property representations and enrich the latent space representation. We believe that our approach not only enhances existing technologies but could also assist in pioneering novel spectroscopic methods, revealing previously uncharacterized material properties and generating characteristics of materials that are challenging to probe experimentally.

Overall, SpectroGen can redefine the future of materials science and spectroscopy by enabling spectral transformations across modalities with minimal experimental input, eliminating the need for costly, time-intensive, and limited-access instrumentation. This could democratize advanced materials characterization, allowing researchers worldwide to access high-quality spectral data without expensive facilities. It could accelerate the discovery of next-generation materials, such as high-efficiency batteries, superconductors, and catalysts, by providing rapid, multi-modal insights into material properties. In pharmaceuticals, it could revolutionize drug development by streamlining molecular profiling and quality control processes. Its ability to synthesize high-fidelity spectra might pave the way for real-time diagnostics in healthcare, where portable devices equipped with SpectroGen could instantly identify biomarkers or pathogens. On a larger scale, SpectroGen could serve as a foundation for automated, AI-driven research ecosystems, enabling breakthroughs at a speed and scale previously unimaginable, leading to new technologies addressing climate change, novel therapies, and sustainable development.

METHODS

Cross-domain spectral transfer via generation

SpectroGen is an algorithm that incorporates mathematical distribution-based physical prior representation of spectra coupled with a deep generative model that specializes in tracking curves. It is implemented by first establishing a probabilistic encoder $q_\phi(z|x)$ that learns the physical prior probability distribution of experimentally derived input spectrum A, for example, an XRD spectrum with a Voigt distribution prior, capturing the physical constraints inherent in the spectral transformation process, such as the complex dependencies of line broadening, superposition, and wavenumber shifts. The intermediate extracted features from the encoder are captured in the latent, low-dimensional vectors z. A probabilistic decoder $p_\theta(z)$ then up-samples and reconstructs the probability distribution of the generated spectrum B (e.g., a Raman spectrum with Lorentzian distribution). The algorithm training entails multiple waveform distribution analyses to deconstruct single-frequency peaks, wavenumber shifts, and broadening (Figure 2A). The stability and performance of the spectral generation are verified through physical prior spectral deconstruction (Figure 2B), such as a Gaussian distribution prior, a Voigt distribution prior, and a Lorentzian distribution prior, and model fitting (Figures 2C and 2D).

The physical priors in SpectroGen describe and represent the fundamental backbone structure of spectroscopic curves, as validated by established findings in the scientific literature. For example, it is widely acknowledged that the IR spectra of solid mineral materials follow a Gaussian distribution prior,[34] and the peaks in X-ray spectra follow a Voigt distribution.[33] The integral width of the intrinsic XRD profile is determined by factors such as the average crystallite size and lattice strain. The observed XRD profile for a powder reflection is obtained by convolving the intrinsic profile with instrumental broadening effects, which can be approximated as the convolution of a Gaussian function and a Lorentzian function,[34] resulting in a Voigt distribution. Based on this, we employed a Gaussian prior for IR-to-Raman transformation and a Voigt prior for XRD-to-Raman transformation. Notably, the experimentally acquired spectra, aside from conforming to the physical prior, are also influenced by various broadening mechanisms. A key feature of our method is its ability to fit the difference between physical priors and actual spectra through the network's automatic fitting capabilities to address the non-uniform environmental broadening of the spectrum (Figure 2D), e.g., collision broadening,[35] Doppler broadening,[36] transit-time broadening,[37,38] and instrumental influences (see Methods S4). Notably, this approach mitigates the limitations of fitting based solely on physical priors with the flexibility of generative learning-based curve matching to support the precise model transformations, allowing an accurate fitting of peak overlap and broadening. Moreover, Kullback-Leibler (KL) divergence loss (Figure 2C; Methods S5) between the generated and input spectra is computed and iteratively minimized during the training phase, increasing generation accuracy.[39,40]

Further details regarding the methods can be found in Methods S1–S9.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Loza F. Tadesse (lozat@mit.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Data used in this research are all from open-source dataset RRUFF.[26]
- The code of this research is available at https://github.com/ymzhu19eee/Raman-generation.

ACKNOWLEDGMENTS

This research was funded by the Massachusetts Institute of Technology, specifically through Prof. Loza F. Tadesse's startup package (1578169) from the Department of Mechanical Engineering. We thank Sujan Manna, Dr. Jeong-Hee (Jenn) Kim, and Dr. Jia Dong for their valuable suggestions and comments during the manuscript writing process.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.Z. and L.F.T.; methodology, Y.Z.; investigation, Y.Z., and L.F.T.; writing – original draft, Y.Z.; writing – review and editing, Y.Z. and L.F.T.; funding acquisition, L.F.T.; resources, L.F.T.; supervision, L.F.T.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.matt.2025.102434.

Received: October 3, 2024
Revised: May 6, 2025
Accepted: August 25, 2025
Published: October 14, 2025

REFERENCES

1. Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., and Walsh, A. (2018). Machine learning for molecular and materials science. Nature *559*, 547–555.
2. Morton, R. (1962). Spectroscopy as a Biochemical Tool. Nature *193*, 314–318.
3. Smith, B.C. (2018). Infrared Spectral Interpretation: A Systematic Approach (CRC Press).
4. Aruldhas, G. (2007). Molecular Structure and Spectroscopy (PHI Learning Pvt. Ltd).
5. Thompson, H.W. (1945). Application of Infra-Red Spectroscopy to Chemical Problems. Nature *155*, 191–193.
6. Gibson, A. (1968). Infrared Spectroscopy. Nature *218*, 49.
7. (1933). Raman Spectra and Chemistry. Nature *131*, 263–265.
8. Wang, Y.H., Zheng, S., Yang, W.M., Zhou, R.Y., He, Q.F., Radjenovic, P., Dong, J.C., Li, S., Zheng, J., Yang, Z.L., et al. (2021). In situ Raman spectroscopy reveals the structure and dissociation of interfacial water. Nature *600*, 81–85.
9. Tadesse, L.F., Ho, C.S., Chen, D.H., Arami, H., Banaei, N., Gambhir, S.S., Jeffrey, S.S., Saleh, A.A.E., and Dionne, J. (2020). Plasmonic and electrostatic interactions enable uniformly enhanced liquid bacterial surface-enhanced Raman scattering (SERS). Nano Lett. *20*, 7655–7661.
10. Hanke, R., Fuchs, T., and Uhlmann, N. (2008). X-ray based methods for non-destructive testing and material characterization. Nucl. Instrum. Methods Phys. Res. *591*, 14–18.
11. Morcrette, M., Chabre, Y., Vaughan, G., Amatucci, G., Leriche, J.B., Patoux, S., Masquelier, C., and Tarascon, J.M. (2002). In situ X-ray diffraction techniques as a powerful tool to study battery electrode materials. Electrochim. Acta *47*, 3137–3149.
12. Zheng, X., Bommier, C., Luo, W., Jiang, L., Hao, Y., and Huang, Y. (2019). Sodium metal anodes for room-temperature sodium-ion

batteries: Applications, challenges and solutions. Energy Storage Mater. *16*, 6–23.

13. Huang, L., Sun, H., Sun, L., Shi, K., Chen, Y., Ren, X., Ge, Y., Jiang, D., Liu, X., Knoll, W., et al. (2023). Rapid, label-free histopathological diagnosis of liver cancer based on Raman spectroscopy and deep learning. Nat. Commun. *14*, 48.
14. Karunanithy, G., Shukla, V.K., and Hansen, D.F. (2024). Solution-state methyl NMR spectroscopy of large non-deuterated proteins enabled by deep neural networks. Nat. Commun. *15*, 5073.
15. Webel, H., Niu, L., Nielsen, A.B., Locard-Paulet, M., Mann, M., Jensen, L.J., and Rasmussen, S. (2024). Imputation of label-free quantitative mass spectrometry-based proteomics data using self-supervised deep learning. Nat. Commun. *15*, 5405.
16. Ziatdinov, M., Ghosh, A., Wong, C.Y., and Kalinin, S.V. (2022). AtomAI framework for deep learning analysis of image and spectroscopy data in electron and scanning probe microscopy. Nat. Mach. Intell. *4*, 1101–1112.
17. Karunanithy, G., Shukla, V.K., and Hansen, D.F. (2024). Solution-state methyl NMR spectroscopy of large non-deuterated proteins enabled by deep neural networks. Nat. Commun. *15*, 5073.
18. Tong, X., Qu, N., Kong, X., Ni, S., Zhou, J., Wang, K., Zhang, L., Wen, Y., Shi, J., Zhang, S., et al. (2024). Deep representation learning of chemical-induced transcriptional profile for phenotype-based drug discovery. Nat. Commun. *15*, 5378.
19. Cuomo, S., Di Cola, V.S., Giampaolo, F., Rozza, G., Raissi, M., and Piccialli, F. (2022). Scientific machine learning through physics–informed neural networks: Where we are and what's next. J. Sci. Comput. *92*, 88.
20. Doersch, C. (2016). Tutorial on variational autoencoders. Preprint at arXiv. https://doi.org/10.48550/arXiv.1606.05908.
21. Xiang, W., Chen, D., Cui, Y., and Peng, S. (2021). H-VAE: A Hybrid Variational AutoEncoder with Data Augmentation in Predicting CRISPR/Cas9 Off-target. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE), pp. 550–555.
22. Lyu, S., Sowlati-Hashjin, S., and Garton, M. (2024). Variational autoencoder for design of synthetic viral vector serotypes. Nat. Mach. Intell. *6*, 147–160.
23. Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent. Sci. *4*, 268–276.
24. Noh, J., Gu, G.H., Kim, S., and Jung, Y. (2020). Machine-enabled inverse design of inorganic solid materials: promises and challenges. Chem. Sci. *11*, 4871–4881.
25. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.
26. Lafuente, B., Downs, R.T., Yang, H., and Stone, N. (2015). The power of databases: the RRUFF project. Highlights Mineral. Crystallogr. *1*, 1–30.
27. Menéndez, M.L., Pardo, J.A., Pardo, L., and Pardo, M.C. (1997). The jensen-shannon divergence. J. Franklin Inst. *334*, 307–318.
28. Hore, A., and Ziou, D. (2010). Image quality metrics: PSNR vs. SSIM. In 2010 20th International Conference on Pattern Recognition (IEEE), pp. 2366–2369.
29. Torrey, L., and Shavlik, J. (2010). Transfer learning. In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques (IGI global), pp. 242–264.
30. Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C.P., Wang, X.Z., and Wu, Q.M.J. (2022). A review of generalized zero-shot learning methods. IEEE Trans. Pattern Anal. Mach. Intell. *45*. 4051-20.
31. Zhu, Y., Lo, H.K.A., Yeung, C.H., and Lam, E.Y. (2022). Microplastic pollution assessment with digital holography and zero-shot learning. APL Photon *7*, 076102.
32. Lee, J., McDonald, M., Mhlanga, N., Kang, J.W., Karnik, R., and Tadesse, L.F. (2023). More than magnetic isolation: Dynabeads as strong Raman reporters toward simultaneous capture and identification of targets. J. Raman Spectrosc. *54*, 905–916.
33. Suortti, P., Maija, A., and Lars, U. (1979). Voigt function fit of X-ray and neutron powder diffraction profiles. J. Appl. Cryst. *12*, 365–369.
34. Bradley, M. (2007). Curve Fitting in Raman and IR Spectroscopy: Basic Theory of Line Shapes and Applications (Thermo Fisher Scientific). Application Note 50733.
35. Van Vleck, J.H., and Weisskopf, V.F. (1945). On the shape of collision-broadened lines. Rev. Mod. Phys. *17*, 227–236.
36. Rautian, S.G., and Igor, I.S. (1967). The Effect of Collisions on the Doppler Broadening of Spectral Lines. Sov. Phy. Usp. *9*, 701.
37. Xiao, Y. (2009). Spectral line narrowing in electromagnetically induced transparency. Mode. Phys. Lett. B *23*, 661–680.
38. Bruvelis, M., Ulmanis, J., Bezuglov, N.N., Miculis, K., Andreeva, C., Mahrov, B., Tretyakov, D., and Ekers, A. (2012). Analytical model of transit time broadening for two-photon excitation in a three-level ladder and its experimental validation. Phys. Rev. A *86*, 012501.
39. Hall, P. (1987). On Kullback-Leibler loss and density estimation. Ann. Stat. *15*, 1491–1519.
40. Ji, S., Zhang, Z., Ying, S., Wang, L., Zhao, X., and Gao, Y. (2022). Kullback–Leibler divergence metric learning. IEEE T. Cybern. *52*, 2047–2058.