Uma abordagem com autoencoders variacionais condicionais para geração de materiais NLO com bandgap alvo

André Lopes[1], **Carlos Eduardo de Souza Santos**[1], **Rosiane de Freitas**[1]

[1]Instituto de Computação – Universidade Federal Amazonas (IComp/UFAM)
Manaus – AM, Brasil

{andre.teixeira, carlos.santos, rosiane}@icomp.ufam.edu.br

Abstract. *This work presents a machine learning pipeline to generate candidate vectors for nonlinear optical materials (NLOs) with bandgaps in the 2–4 eV range, essential for applications in quantum computing, such as entangled photon sources. Using data from the Materials Project database, automated featurization, and a Random Forest model for bandgap prediction ($R^2 > 0.99$), a Conditional Variational Autoencoder (CVAE-χ^2) was trained to generate synthetic vectors conditioned to the target range. Of the generated vectors, 65.1% showed predicted bandgaps between 2 and 4 eV, demonstrating the potential of the approach to accelerate the discovery of NLO materials for quantum technologies.*

Resumo. *Neste trabalho, é apresentado um pipeline de aprendizado de máquina para gerar vetores candidatos à materiais ópticos não lineares (NLO) com bandgap na faixa de 2–4 eV, essenciais para aplicações em computação quântica, como fontes de fótons emaranhados. Utilizando dados da base Materials Project, featurização automatizada e um modelo Random Forest para predição de bandgap ($R^2 > 0,99$), foi treinado um Autoencoder Variacional Condicional (CVAE-χ^2) para gerar vetores sintéticos condicionados à faixa alvo. Dos vetores gerados, 65,1% apresentaram bandgap previsto entre 2 e 4 eV, demonstrando o potencial da abordagem para acelerar a descoberta de materiais NLO voltados a tecnologias quânticas.*

1. Introdução

A óptica não linear (NLO) é um campo que estuda a interação de materiais com luz laser intensa. Materiais com propriedades NLO desempenham um papel importante em dispositivos fotônicos, telecomunicações, armazenamento óptico de dados, processamento de sinais e imagem holográfica. Aplicações específicas incluem a geração de segunda harmônica (SHG) e o efeito Pockels eletro-óptico (EOPE) [Haq et al. 2025]. Para investigar as propriedades de novos materiais, são empregados cálculos de química quântica, utilizando a Teoria do Funcional da Densidade (DFT). Em tais estudos teóricos, o custo computacional é uma consideração, e estratégias como a substituição de grupos moleculares volumosos por outros menores são utilizadas para reduzir as despesas computacionais [Haq et al. 2025].

Nos últimos anos, o uso de técnicas de aprendizado de máquina (ML) tem emergido como uma abordagem promissora para acelerar a triagem de materiais com propriedades desejadas. Estudos recentes demonstraram que modelos supervisionados baseados

em redes neurais gráficas, regressão e florestas aleatórias podem prever com precisão propriedades eletrônicas, como o bandgap, a partir de representações estruturais e químicas extraídas de bases de dados computacionais [Chen et al. 2019]. Complementando os métodos supervisionados, o uso de modelos generativos, como os Autoencoders Variacionais Condicionais (CVAE), tem ganhado destaque na geração de vetores sintéticos que representam materiais com propriedades-alvo definidas. Essa estratégia tem sido aplicada com sucesso à criação de materiais fotovoltaicos, ligas metálicas e, mais recentemente, ao design inverso de materiais NLO com bandgap direcionado [Matsunoshita et al. 2023].

Neste contexto, este trabalho propõe um pipeline completo que integra a extração de dados reais do Materials Project, a featurização (vetorização) automatizada com o matminer, a predição do bandgap por meio de modelos de regressão supervisionados e a geração de vetores sintéticos utilizando um modelo CVAE-χ^2 condicionado a faixas específicas de bandgap. O objetivo é possibilitar a geração de vetores candidatos a materiais NLO com maior eficiência, alinhando-se às práticas de ponta da descoberta de materiais guiada por dados.

2. Fundamentação Teórica

A descoberta de novos materiais com propriedades ópticas não lineares (NLO) é um desafio central para o avanço de tecnologias fotônicas e quânticas, uma vez que o comportamento óptico de segunda e terceira ordens está diretamente relacionado à estrutura eletrônica do material, em especial ao seu *bandgap* eletrônico [Haq et al. 2025]. Simulações baseadas na Teoria do Funcional da Densidade (*Density Functional Theory* – DFT) permitem a predição precisa dessas propriedades, porém apresentam elevado custo computacional, restringindo a exploração sistemática do vasto espaço químico [Xie et al. 2023]. Nesse contexto, métodos de aprendizado de máquina (ML) têm emergido como alternativas eficazes para acelerar a triagem de materiais. Modelos supervisionados, como *Random Forest* e redes neurais gráficas, têm sido empregados com sucesso na predição de propriedades eletrônicas a partir de descritores composicionais e estruturais extraídos de bases de dados computacionais como o *Materials Project* [Ward et al. 2018, Chen et al. 2019].

Paralelamente, modelos generativos condicionais, como os Autoencoders Variacionais Condicionais (CVAE), têm demonstrado grande potencial no *design* inverso de materiais, permitindo a geração de novos candidatos com propriedades-alvo predefinidas [Matsunoshita et al. 2023]. Diferentemente de abordagens baseadas exclusivamente em busca ou otimização, o CVAE aprende uma representação latente contínua dos dados, condicionada a variáveis de interesse — como o *bandgap* — e é capaz de amostrar novos vetores que respeitam as características desejadas.

Essa arquitetura combina um codificador que mapeia os dados de entrada para uma distribuição latente, uma etapa de reparametrização que permite o treinamento estocástico, e um decodificador que reconstrói os vetores originais a partir das amostras latentes e da variável condicional [Zuo et al. 2021]. A integração entre modelos supervisionados de regressão e modelos generativos condicionais constitui, portanto, uma estratégia promissora para a descoberta orientada de materiais NLO com *bandgap* direcionado, alinhando-se às demandas de aplicações em computação e comunicação quântica.

3. Geração Automática de Materiais por Aprendizado de Máquina

A descoberta de materiais com propriedades ópticas não lineares (NLO) adequadas para aplicações em computação quântica e fotônica representa um desafio multidisciplinar de grande complexidade. Tais aplicações, incluindo fontes de fótons emaranhados, moduladores eletro-ópticos e dispositivos de processamento de sinais quânticos, exigem materiais com *bandgap* eletrônico na faixa de 2 a 4 eV, que concilia transparência óptica na região do visível com resposta não linear suficiente [Haq et al. 2025]. A busca por esses materiais é dificultada pela vastidão do espaço químico — estimado em dezenas de milhares de compostos cristalinos estáveis — e pela natureza computacionalmente dispendiosa das simulações baseadas em primeiros princípios, como a Teoria do Funcional da Densidade (DFT), que inviabiliza a exploração exaustiva de candidatos potenciais por métodos exclusivamente teóricos [Xie et al. 2023].

Nesse contexto, surge a necessidade de estratégias baseadas em dados e aprendizado de máquina que permitam explorar de forma inteligente e eficiente o espaço de materiais, reduzindo o custo computacional associado à triagem inicial de candidatos. Especificamente, coloca-se a seguinte questão de pesquisa: *como gerar vetores candidatos a materiais NLO que apresentem bandgap dentro de uma faixa predefinida (2–4 eV), a partir de dados reais de materiais estáveis, utilizando técnicas de aprendizado de máquina supervisionado e generativo?*

A resposta a essa questão envolve a integração de múltiplas etapas — coleta e filtragem de dados, extração automatizada de descritores químicos, treinamento de modelos preditivos de *bandgap* e desenvolvimento de modelos generativos condicionais — de modo a viabilizar o *design* inverso de materiais com propriedades direcionadas, alinhando-se às demandas de dispositivos fotônicos e tecnologias quânticas emergentes [Matsunoshita et al. 2023, Zuo et al. 2021].

Para isto, está sendo proposto o desenvolvimento de um pipeline de aprendizado de máquina capaz de gerar vetores representativos de materiais NLO com bandgap direcionado, utilizando dados reais do Materials Project.

3.1. Trabalhos Relacionados

A aplicação de aprendizado de máquina na descoberta de materiais ópticos não lineares (NLO) tem avançado significativamente nos últimos anos, com destaque para abordagens que integram predição de propriedades e geração de candidatos. Uma abordagem relevante nesse campo é a de Xie et al. [2023], que propuseram um banco de dados orientado por predição para acelerar a triagem de materiais NLO com alto desempenho. O estudo combina aprendizado supervisionado com validação experimental e demonstra que estratégias preditivas podem reduzir drasticamente o custo computacional na descoberta de materiais NLO com bandgap desejado [Xie et al. 2023].

Essa linha de pesquisa está alinhada com iniciativas anteriores que utilizam bases de dados computacionais como o Materials Project e ferramentas de featurização como o matminer para treinar modelos de regressão capazes de prever propriedades como bandgap, energia de formação e simetria cristalina [Ward et al. 2018]. A extração automatizada de atributos estruturais e composicionais, como implementado em pipelines baseados em matminer, tem se mostrado essencial para a obtenção de vetores preditivos robustos.

Complementando os métodos supervisionados, o uso de modelos generativos condicionais, como os Autoencoders Variacionais Condicionais (CVAE), tem se tornado cada vez mais comum. Matsunoshita et al. [2023] demonstraram que é possível treinar um CVAE para gerar vetores sintéticos de materiais com bandgaps direcionados, obtendo alta taxa de consistência preditiva quando avaliados por modelos regressivos externos. Essa abordagem representa um avanço no design inverso de materiais, especialmente para NLO, onde a transparência óptica e a estabilidade dielétrica são condicionadas por janelas específicas de bandgap.

Finalmente, Zuo et al. [2021] realizaram uma avaliação sistemática do desempenho de diferentes algoritmos de aprendizado de máquina na predição de propriedades eletrônicas e estruturais. O estudo mostrou que, com uma featurização bem calibrada, modelos simples como Random Forest ainda são altamente competitivos, especialmente quando aplicados a bases de dados balanceadas e limpas.

O presente trabalho se diferencia ao combinar essas estratégias em um único pipeline: coleta de dados do Materials Project, featurização com matminer, predição de bandgap com modelos rasos e geração dirigida com CVAE-χ^2. Tal abordagem, além de estar em consonância com o estado da arte, busca demonstrar o potencial do uso conjunto de modelos supervisionados e generativos na descoberta orientada de materiais NLO.

4. Protocolo Experimental

A fim de avaliar a eficácia de modelos supervisionados e generativos na geração de candidatos a materiais ópticos não lineares com bandgap específico, foi desenvolvido um pipeline experimental baseado em dados reais, featurização automatizada, modelagem preditiva e geração condicionada. As etapas que compõem esse protocolo são descritas nas subseções a seguir.

4.1. Base de Dados

O conjunto de dados utilizado neste trabalho foi obtido a partir do Materials Project, uma das maiores plataformas computacionais de dados de materiais disponíveis publicamente. A coleta foi realizada via NextGen API, utilizando a biblioteca mp-api, e seguiu os seguintes critérios de filtragem: bandgap 0,5 eV; energy above hull 0,05 eV; número de sítios 30 e pertencentes a grupos espaciais não centrossimétricos (NCS).

Esses critérios visam garantir a estabilidade termodinâmica e a viabilidade óptica dos candidatos a materiais NLO, conforme sugerido por Xie et al. [2023]. O uso da nova infraestrutura aberta do Materials Project NextGen possibilitou a extração eficiente e reproduzível de dados cristalográficos e eletrônicos, com suporte robusto à integração com pipelines de aprendizado de máquina [Horton et al. 2025].

Após a coleta, os dados foram organizados em um DataFrame contendo informações como fórmula química, número de grupo espacial, volume da célula unitária, energia de formação por átomo, bandgap e estrutura cristalina em formato serializado.

4.2. Métodos Investigados

Dois conjuntos principais de métodos foram explorados neste trabalho:

1. Modelo supervisionado de regressão

Foi treinado o modelo Random Forest Regressor para prever o bandgap dos materiais. Os dados de entrada foram gerados com o pacote matminer, aplicando os seguintes featurizadores:

- ElementProperty (preset "magpie");
- Stoichiometry;
- ValenceOrbital.

Os atributos numéricos foram normalizados com StandardScaler, conforme prática recomendada em pipelines preditivos em ciência dos materiais [Ward et al. 2018].

2. Modelo generativo CVAE-χ^2

A etapa de geração foi conduzida por um Autoencoder Variacional Condicional (CVAE), baseado na arquitetura proposta por Matsunoshita et al. [2023]. O modelo foi treinado com os vetores featurizados como entrada e o bandgap como variável condicional, com o objetivo de aprender a distribuição dos dados e gerar novos vetores sintéticos com bandgap pré-definido.

A denominação CVAE-χ^2 adotada neste trabalho refere-se a uma variante do Autoencoder Variacional Condicional (CVAE) utilizada para geração condicionada de vetores de materiais com bandgap direcionado. O termo ² foi empregado para destacar o caráter probabilístico e estatisticamente regularizado do espaço latente aprendido pelo modelo, diferenciando a abordagem de um CVAE convencional. Essa nomenclatura está alinhada a estudos recentes que investigam estratégias de regularização e organização do espaço latente em Variational Autoencoders, visando melhorar estabilidade, generalização e capacidade generativa [Kingma and Welling 2022, Sinha and Dieng 2022].

4.3. Etapas dos Experimentos Desenvolvidos

As etapas executadas foram as seguintes:

1. Coleta dos dados via NextGen API do Materials Project, com critérios estruturais e energéticos específicos para materiais NLO.
2. Pré-processamento das estruturas cristalinas, convertendo descrições textuais em objetos Structure (pymatgen) e serializando em formato JSON.
3. Organização e exportação dos dados para arquivo CSV para reúso no pipeline.
4. Featurização automatizada com matminer, gerando vetores a partir da composição química (ElementProperty, Stoichiometry, ValenceOrbital).
5. Divisão do conjunto de dados em treino (70%), validação (15%) e teste (15%).
6. Aplicação de oversampling estratificado apenas no conjunto de treino, de modo a balancear as faixas de bandgap antes do treinamento.
7. Treinamento do modelo supervisionado Random Forest Regressor.
8. Avaliação do modelo preditivo com métricas R^2, MAE e RMSE.
9. Validação cruzada 5-fold do Random Forest para verificar robustez.
10. Treinamento do modelo CVAE-χ^2 com vetores e bandgap como variável condicional.
11. Geração de 1000 vetores sintéticos condicionados a valores de bandgap entre 2 e 4 eV.

12. Predição do bandgap dos vetores gerados utilizando o modelo Random Forest treinado.
13. Avaliação da proporção de candidatos válidos (2–4 eV).
14. Visualização gráfica da distribuição dos bandgaps reais e sintéticos, utilizando Matplotlib e Seaborn.

5. Análise dos Resultados

Nesta seção, são apresentados os principais resultados obtidos nas etapas de predição e geração de vetores candidatos a materiais não lineares ópticos (NLO) com bandgap direcionado. As análises são divididas em duas frentes: desempenho do modelo supervisionado de regressão e avaliação do modelo generativo CVAE-χ^2.

5.1. Desempenho do modelo de regressão Random Forest

Para avaliar a capacidade do modelo de aprendizado de máquina na predição do bandgap de materiais NLO, foi treinado o algoritmo Random Forest Regressor. O desempenho foi avaliado inicialmente com o dataset original e, em seguida, com o dataset balanceado por faixa de bandgap (oversampling aplicado apenas na base de treino). As métricas utilizadas foram: Erro Absoluto Médio (MAE), Erro Quadrático Médio (RMSE) e Coeficiente de Determinação (R^2).

O Random Forest Regressor foi empregado como baseline para a predição do bandgap. No conjunto original (sem balanceamento), o modelo obteve MAE = 0,3964, RMSE = 0,5840 e R^2 = 0,8245 (ver Tabela 1), refletindo desempenho consistente frente ao desbalanceamento de faixas de bandgap.

Modelo	MAE	RMSE	R^2
Random Forest	0.3964	0.5840	0.8245

Tabela 1. Desempenho do Random Forest antes do balanceamento do dataset

Em seguida, aplicou-se oversampling estratificado apenas no conjunto de treino e reavaliou-se o desempenho. No conjunto de teste, os resultados não apresentaram ganho: MAE = 0,4071, RMSE = 0,6383 e R^2 = 0,7903 (ver Tabela 2), indicando leve piora em relação ao cenário sem balanceamento.

Modelo	MAE	RMSE	R^2
Random Forest	0.4071	0.6383	0.7903

Tabela 2. Desempenho do Random Forest após o balanceamento do dataset (oversampling no treino)

A validação cruzada 5-fold conduzida após o balanceamento do treino indicou bom ajuste do modelo sob esse procedimento, com MAE = 0,0911 ± 0,0053, RMSE = 0,2797 ± 0,0198 e R^2 = 0,9891 ± 0,0015 (ver Tabela 3), além de baixa variabilidade entre os folds.

O Random Forest teve desempenho inicial razoável (R^2 = 0,8245), mas com erros relativamente altos. Após o balanceamento por oversampling no conjunto de treino, o resultado no teste particionado não melhorou — na verdade, piorou ligeiramente, com R^2 reduzido para 0,7903 e erros maiores (MAE e RMSE aumentaram).

Modelo	MAE	RMSE	R²
Random Forest	0.0911 ± 0.0053	0.2797 ± 0.0198	0.9891 ± 0.0015

Tabela 3. Resultados da validação cruzada (5-fold) do modelo Random Forest após balanceamento do dataset

Por outro lado, a validação cruzada 5-fold aplicada após o balanceamento mostrou um cenário muito mais favorável, com R² próximo de 0,99 e baixa variabilidade, indicando que o modelo consegue generalizar bem quando avaliado em múltiplas partições do conjunto balanceado. Esses resultados comprovam que o Random Forest é um modelo altamente adequado para a tarefa, conseguindo explicar quase toda a variância do band-gap e fornecendo uma base sólida para a etapa posterior de avaliação dos vetores gerados pelo CVAE-χ^2.

5.2. Treinamento do CVAE-χ^2

O modelo generativo foi implementado como um Autoencoder Variacional Condicional (CVAE), desenvolvido em PyTorch. A entrada consistiu nos vetores featurizados e normalizados, juntamente com o valor condicional do bandgap.

A arquitetura foi composta por três módulos:

- Encoder: concatenação do vetor de entrada com o bandgap, processada em camadas densas com ReLU, retornando μ e $\log(\sigma^2)$.
- Reparameterização: geração de amostras latentes $z \sim N(\mu, \sigma^2)$ para viabilizar o treinamento com variáveis estocásticas.
- Decoder: reconstrução do vetor original a partir de z concatenado ao valor condicional.

O modelo CVAE-χ^2 utilizado neste trabalho baseia-se na aprendizagem de uma distribuição latente probabilística, diferentemente de autoencoders tradicionais, que mapeiam cada entrada para um ponto fixo no espaço latente [Kingma and Welling 2022]. No codificador (encoder), os vetores featurizados concatenados ao valor condicional do bandgap são transformados em dois parâmetros fundamentais da distribuição latente: a média μ e a variância σ^2.

A média μ representa a posição central da distribuição latente e corresponde à representação mais provável do vetor de entrada condicionado ao bandgap. Já a variância σ^2 modela a dispersão em torno dessa média, permitindo representar a incerteza associada à codificação dos dados. Dessa forma, o encoder aprende uma distribuição normal multivariada $N(\mu, \sigma^2)$, a partir da qual novas amostras latentes podem ser geradas [Sohn et al. 2015].

Essa representação probabilística é essencial para garantir continuidade e organização no espaço latente, permitindo que pontos próximos produzam vetores com propriedades semelhantes após a decodificação. Consequentemente, o modelo torna-se capaz de realizar geração condicionada de novos candidatos a materiais com bandgap direcionado.

Durante o treinamento, a geração das variáveis latentes utiliza a técnica de reparameterização proposta por Kingma and Welling [2014], necessária para viabilizar o fluxo de gradientes no processo de otimização. Em vez de amostrar diretamente da distribuição aprendida, define-se a variável latente z como:

$$z = \mu + \sigma, N(0, 1)$$

onde corresponde a um ruído gaussiano aleatório. Essa formulação desloca a aleatoriedade para fora da rede neural, permitindo o treinamento do modelo via back-propagation. Além disso, a combinação entre o erro de reconstrução e a divergência KL força o espaço latente a assumir uma distribuição contínua e regularizada, favorecendo a geração de vetores sintéticos coerentes e diversificados.

A função de perda combinou erro de reconstrução (MSE) e divergência KL. O treinamento foi realizado por até 100 épocas, com early stopping (paciência de 10 épocas), otimizador Adam (lr = 0,001) e minibatches de 64.

Na Figura 1, observa-se que o modelo apresenta convergência estável até aproximadamente a 25ª época, com redução consistente das perdas de treino e validação. A partir desse ponto, a perda de validação sofre fortes oscilações, indicando instabilidade no processo de otimização e início de overfitting. Esse comportamento reforça a importância do uso de early stopping, interrompendo o treinamento antes dessas flutuações, de modo a preservar o melhor desempenho generalizável do CVAE-χ^2.

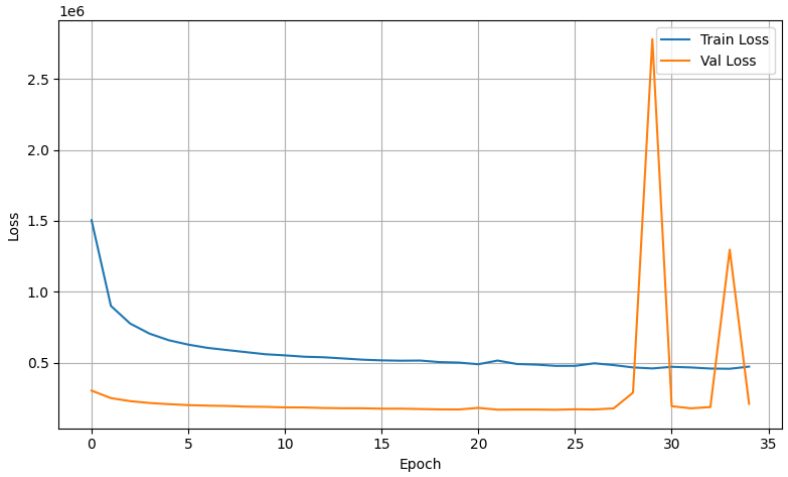

Figura 1. Evolução da perda de treino e validação do modelo CVAE-χ^2

O processo de otimização convergiu de forma estável até aproximadamente a 25ª época, confirmando a capacidade do CVAE-χ^2 em capturar as relações entre atributos químicos e bandgap, viabilizando a etapa de geração condicionada.

5.3. Geração de Vetores e Avaliação Preditiva

Após o treinamento, foram gerados 1000 vetores sintéticos condicionados a bandgaps entre 2 e 4 eV. Esses vetores foram avaliados pelo Random Forest Regressor previamente treinado. Dos vetores gerados, 651 apresentaram bandgap previsto na faixa de interesse (2–4 eV), correspondendo a 65,1% de candidatos válidos, conforme Figura 2.

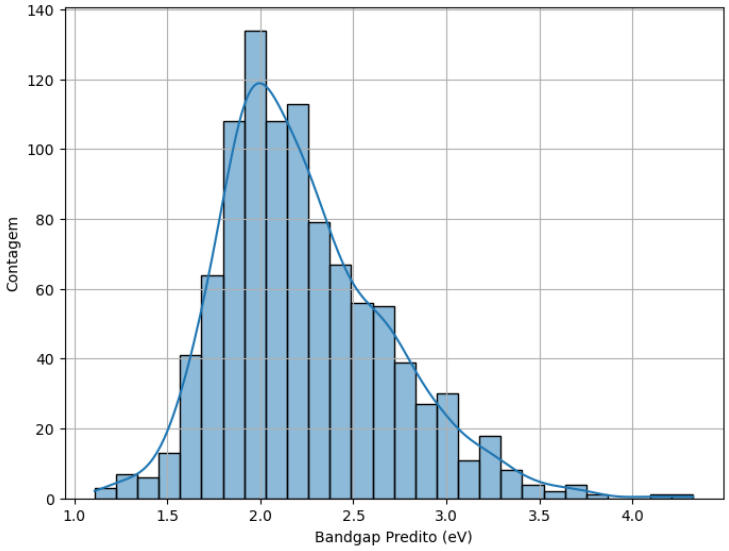

Figura 2. Distribuição dos bandgaps preditos para dados gerados pelo CVAE-χ^2.

Conforme evidenciado na Figura 1, esse resultado demonstra que o CVAE-χ^2 é capaz de guiar a geração de vetores em direção à janela desejada de bandgap, reduzindo o viés presente nos dados originais. A distribuição dos valores gerados concentrou-se majoritariamente entre 2,0 e 3,0 eV, confirmando a eficácia parcial do condicionamento no espaço latente.

5.4. Análise gráfica

A análise gráfica complementa as métricas numéricas ao permitir a comparação visual entre os bandgaps dos dados reais e dos vetores gerados, destacando a aderência do modelo à faixa de interesse para materiais NLO. A Figura 3 apresenta a comparação gráfica entre a distribuição de bandgap dos materiais reais do dataset original (em azul) e a dos vetores gerados pelo modelo CVAE-χ^2 (em amarelo), com a faixa-alvo de 2 a 4 eV destacada em verde-claro como região de interesse para aplicações NLO.

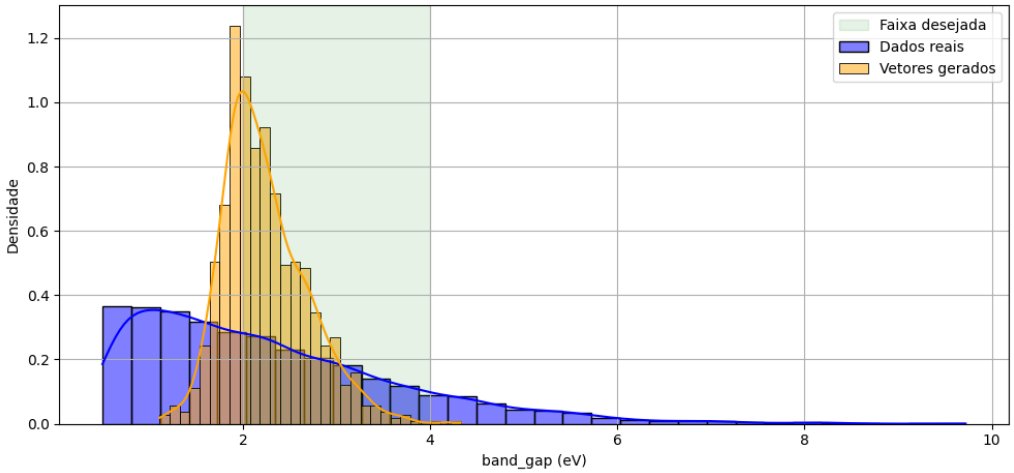

Figura 3. Distribuição dos bandgaps: reais vs preditos.

A Figura 3 mostra claramente a sobreposição entre as distribuições: enquanto os dados reais se concentram em valores sub-ótimos, os vetores gerados ocupam de forma mais equilibrada a janela de interesse. Essa sobreposição, especialmente na região destacada entre 2 e 4 eV, demonstra que o CVAE-χ^2 internalizou relações relevantes entre os atributos e o bandgap, embora com eficiência moderada.

Em síntese, os resultados gráficos reforçam a conclusão de que o modelo consegue gerar candidatos mais próximos da faixa ideal para materiais NLO, mas também evidenciam que ajustes adicionais no condicionamento são necessários para aumentar a proporção de vetores válidos dentro da janela alvo.

A Tabela 4 apresenta um quadro comparativo com os principais trabalhos relacionados da literatura, e com a última linha sendo a proposta apresentada neste artigo.

Tabela 4. Quadro comparativo entre os trabalhos relacionados e a proposta do presente estudo.

Referência	**Método Principal**	**Dataset Usado**	**Métricas**	**Contribuição / Diferencial**
Xie et al. (2023)	Predição supervisionada + validação experimental	Base proprietária + validação experimental	Precisão preditiva; validação experimental	Banco de dados orientado por predição; redução do custo computacional na busca por materiais com bandgap desejado
Ward et al. (2018)	Featurização automatizada (matminer)	Materials Project	Tempo de extração; escalabilidade	Ferramenta de código aberto para extração de descritores composicionais e estruturais
Matsunoshita et al. (2023)	Autoencoder Variacional Condicional (CVAE)	Dados sintéticos de materiais	Taxa de consistência preditiva; erro de reconstrução	Demonstração da eficácia do CVAE na geração de vetores sintéticos com alta consistência preditiva
Zuo et al. (2021)	Avaliação sistemática de algoritmos de ML	Materials Project + bases complementares	R^2; MAE; RMSE; acurácia	Comparação entre diferentes métodos; destaque para Random Forest em bases calibradas
Presente Trabalho	**CVAE-χ^2 + Random Forest + matminer**	**Materials Project (NextGen API)**	**65,1% candidatos válidos (2–4 eV); $R^2 > 0,99$ (validação cruzada)**	**Pipeline integrado: coleta de dados, featurização, predição supervisionada e geração condicional; foco específico em aplicações quânticas (2–4 eV)**

6. Considerações Finais

Este trabalho apresentou um pipeline de aprendizado de máquina para a geração de candidatos a materiais ópticos não lineares (NLO) com bandgap direcionado, combinando dados reais do Materials Project, featurização automatizada via matminer, regressão supervisionada com Random Forest e um modelo generativo CVAE-χ^2. O estudo reforça a relevância de integrar modelos preditivos e generativos na descoberta de materiais orientada por dados.

Do ponto de vista preditivo, o Random Forest demonstrou desempenho robusto, com R^2 superior a 0,82 em dados originais e chegando a quase 0,99 na validação cruzada após balanceamento. Esses resultados confirmam a adequação do modelo como baseline para avaliação de candidatos, mesmo diante de distribuições desbalanceadas de bandgap. Já o modelo generativo CVAE-χ^2 apresentou convergência estável e foi capaz de gerar 651 vetores sintéticos válidos (65,1%) na faixa-alvo de 2–4 eV. A análise gráfica mostrou que, embora os dados reais se concentrem fora da janela desejada, os vetores gerados ocuparam de forma mais equilibrada a região de interesse, evidenciando a eficácia parcial do condicionamento latente.

Esses achados demonstram que a combinação entre featurização química, regressão não paramétrica e geração condicional oferece uma estratégia promissora para reduzir o espaço de busca por candidatos a materiais NLO. Ainda que o percentual de vetores válidos indique espaço para melhorias — como ajustes no condicionamento ou no balanceamento dos dados —, os resultados já apontam para um avanço significativo na triagem inicial de candidatos com potencial tecnológico.

Como perspectivas futuras, destaca-se a necessidade de incorporar validações adicionais, tanto estruturais quanto baseadas em simulações de first principles (DFT), para confirmar a viabilidade física dos vetores gerados. Além disso, a integração com bancos de dados experimentais e estratégias de active learning poderá ampliar a eficiência do pipeline, aproximando ainda mais a descoberta de materiais NLO das aplicações práticas em fotônica, telecomunicações e tecnologias quânticas.

Agradecimentos

Os autores fazem parte do grupo de Computação Quântica - Algoritmos, Otimização e Complexidade (ALGOX) do Programa de Pós-Graduação em Informática da UFAM e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), com pesquisa realizada com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES-PROEX) - Código de Financiamento 001, bem como parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM – por meio do projeto POSGRAD 2024/2025.

Referências

Chen, C., Ye, W., Zuo, Y., Zheng, C., and Ong, S. P. (2019). Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572.

Haq, S., Khalid, M., Braga, A. A. C., Alhokbany, N., and Chen, K. (2025). Design and evaluation of indacenothienothiophene based functional materials for second and third order nonlinear optics properties via dft approach. *Scientific Reports*, 15(1):13262.

Horton, M. K., Huck, P., Yang, R. X., Munro, J. M., Dwaraknath, S., Ganose, A. M., Kingsbury, R. S., Wen, M., Shen, J. X., Mathis, T. S., et al. (2025). Accelerated data-driven materials science with the materials project. *Nature Materials*, 24(10):1522–1532.

Kingma, D. P. and Welling, M. (2022). Auto-encoding variational bayes.

Matsunoshita, K., Yamaguchi, Y., Hamaie, M., Horibe, M., Tanibata, N., Takeda, H., Nakayama, M., Karasuyama, M., and Kobayashi, R. (2023). Optimization of force-field potential parameters using conditional variational autoencoder. *Science and Technology of Advanced Materials: Methods*, 3(1):2253713.

Sinha, S. and Dieng, A. B. (2022). Consistency regularization for variational auto-encoders.

Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N. E., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M., et al. (2018). Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69.

Xie, C., Tikhonov, E., Chu, D., Wu, M., Kruglov, I., Pan, S., and Yang, Z. (2023). A prediction-driven database to enable rapid discovery of nonlinear optical materials. *Science China Materials*, 66(11):4473–4479.

Zuo, Y., Qin, M., Chen, C., Ye, W., Li, X., Luo, J., and Ong, S. P. (2021). Accelerating materials discovery with bayesian optimization and graph deep learning. *Materials Today*, 51:126–135.