Geração de materiais NLO para computação quântica usando aprendizado de máquina

André Lopes, Carlos Eduardo e Rosiane de Freitas

***Resumo*— Neste trabalho é apresentado um pipeline baseado em aprendizado de máquina para a geração de vetores candidatos a materiais ópticos não lineares (NLO), voltados a aplicações em computação quântica, com bandgap direcionado. Utilizando dados reais do Benchmark *Materials Project*, featurização automatizada via matminer e predição com Random Forest, foi treinado um modelo CVAE-χ^2 para gerar vetores condicionados. O modelo alcançou 65,1% de candidatos com bandgap previsto na faixa de 2–4 eV. Os resultados demonstram o potencial da abordagem para triagem eficiente de materiais NLO em contextos de fotônica e tecnologias quânticas, com futura aplicação em validações estruturais e simulações físicas.**

***Palavras-Chave*— aprendizado de máquina, bandgap, computação quântica, materiais ópticos não lineares, Random Forest.**

***Abstract*— This work proposes a machine learning-based pipeline for generating candidate vectors for non-linear optical (NLO) materials, aimed at applications in quantum communication, with targeted bandgap. Using real data from the Materials Project, automated feature extraction via matminer, and prediction with Random Forest, a CVAE-χ^2 model was trained to generate conditioned vectors. The model achieved 65.1% of candidates with predicted bandgap in the range of 2–4 eV. The results demonstrate the potential of the approach for efficient screening of NLO materials in the contexts of photonics and quantum technologies, with future applications in structural validations and physical simulations.**

***Keywords*— bandgap, quantum computing, machine learning, nonlinear optical materials, Random Forest.**

I. Introdução

A óptica não linear (NLO) é fundamental em fotônica, computação e comunicação quântica, viabilizando fenômenos como geração de segunda harmônica e efeito Pockels [2]. Contudo, a descoberta de novos materiais NLO é limitada pelo alto custo computacional de métodos tradicionais, como a Teoria do Funcional da Densidade (DFT). O aprendizado de máquina (ML) surge como alternativa promissora para acelerar essa triagem, permitindo prever propriedades eletrônicas, como o bandgap, a partir de descritores químicos e estruturais. Modelos supervisionados, como *Random Forest* (RF) e *Graph Neural Networks* (GNNs), já demonstraram eficácia nessa tarefa [1][5], enquanto abordagens generativas, como os *Autoencoders* Variacionais Condicionais (CVAE), permitem direcionar a criação de candidatos com propriedades específicas [4]. Neste trabalho, é proposto um pipeline que combina dados do Benchmark *Materials Project* [3][6], featurização com matminer, predição via RF e geração de vetores condicionados com CVAE-χ^2, visando identificar materiais NLO na faixa de bandgap entre 2–4 eV, crítica para aplicações fotônicas e quânticas.

[1]André Lopes, [1]Carlos Eduardo, [1]Rosiane de Freitas. [1]Instituto de Computação, Universidade Federal do Amazonas, Manaus- AM, e-mails: {andre.teixeira, carlos.santos, rosiane}@icomp.ufam.edu.br.

II. Protocolo Experimental

O pipeline desenvolvido integra dados reais, featurização, aprendizado supervisionado e geração condicional para obtenção de candidatos a materiais NLO. Os dados foram coletados da base *Materials Project* via API, aplicando filtros de estabilidade e simetria não centrossimétrica, e organizados em DataFrame com informações cristalográficas, energéticas e eletrônicas.

A featurização foi realizada com o matminer, empregando descritores de composição (ElementProperty, Stoichiometry e ValenceOrbital), normalizados com StandardScaler. A base foi dividida em treino, validação e teste, aplicando oversampling apenas no treino para equilibrar as faixas de bandgap.

Na etapa supervisionada, utilizou-se o RF Regressor, avaliado por MAE, RMSE e R^2, além de validação cruzada 5-fold. Já na etapa generativa, foi treinado um CVAE-χ^2 em PyTorch, condicionado ao bandgap, com função de perda combinando reconstrução e divergência KL. Após o treinamento, foram gerados 1000 vetores sintéticos na faixa de 2–4 eV, dos quais 651 (65,1%) foram considerados válidos, confirmando a viabilidade do pipeline para triagem de materiais NLO.

III. Desempenho do modelo Random Forest

O RF foi empregado como baseline para a predição de bandgap. No conjunto original, sem balanceamento, o modelo apresentou MAE = 0,3964, RMSE = 0,5840 e R^2 = 0,8245, confirmando desempenho consistente frente ao desbalanceamento de faixas. Após aplicação de oversampling no conjunto de treino, o desempenho no conjunto de teste apresentou ligeira queda (MAE = 0,4071; RMSE = 0,6383; R^2 = 0,7903). Entretanto, a validação cruzada 5-fold revelou ganhos expressivos, alcançando MAE = 0,0911 ± 0,0053; RMSE = 0,2797 ± 0,0198; R^2 = 0,9891 ± 0,0015, indicando robustez do modelo quando avaliado em múltiplas partições.

TABELA I
Desempenho do Random Forest sem balanceamento

Modelo	MAE	RMSE	R^2
RF	0.3964	0.5840	0.8245

TABELA II

DESEMPENHO DO RANDOM FOREST COM BALANCEAMENTO

Modelo	MAE	RMSE	R^2
RF	0.4071	0.6383	0.7903

TABELA III

RESULTADOS DA VALIDAÇÃO CRUZADA

Modelo	MAE	RMSE	R^2
RF	0.0911 ± 0.0053	0.2797 ± 0.0198	0.9891 ± 0.0015

IV. TREINAMENTO DO CVAE-χ^2

O CVAE-χ^2 apresentou convergência estável até a 25ª época, com queda consistente nas perdas de treino e validação. Após esse ponto, observou-se oscilação na perda de validação, sinalizando início de overfitting. Esse comportamento reforçou a importância do uso de early stopping para preservar o melhor desempenho.

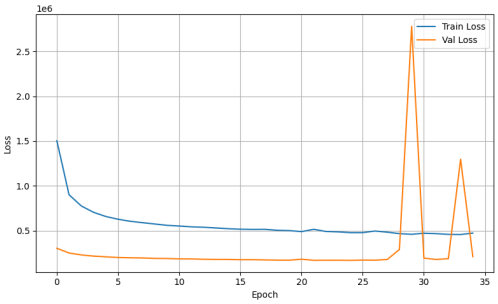

Fig. 1. Evolução da perda do treino e validação do CVAE-χ^2.

V. GERAÇÃO DE VETORES E AVALIAÇÃO PREDITIVA

Foram gerados 1000 vetores sintéticos, condicionados a bandgaps entre 2–4 eV. Destes, 651 vetores (65,1%) apresentaram valores previstos na faixa-alvo, confirmando a eficácia parcial do condicionamento. A distribuição resultante mostrou maior concentração entre 2,0 e 3,0 eV, alinhando-se à região de interesse, mas indicando necessidade de ajustes para ampliar a proporção de candidatos válidos.

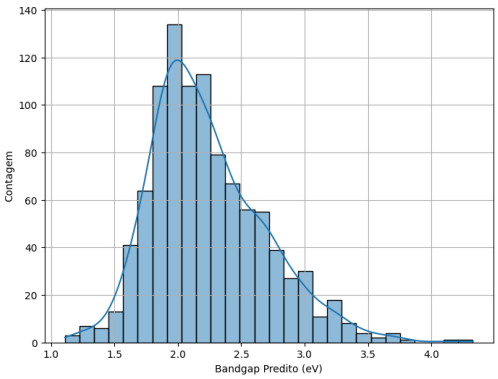

Fig. 2. Distribuição dos bandgaps preditos para vetores gerados.

VI. ANÁLISE GRÁFICA

A comparação entre distribuições reais e geradas evidenciou sobreposição significativa na faixa de 2–4 eV. Enquanto os dados reais estavam concentrados fora da janela ideal, os vetores gerados ocuparam de forma mais equilibrada a região de interesse, demonstrando que o CVAE-χ^2 internalizou relações relevantes entre atributos químicos e bandgap, ainda que com eficiência moderada.

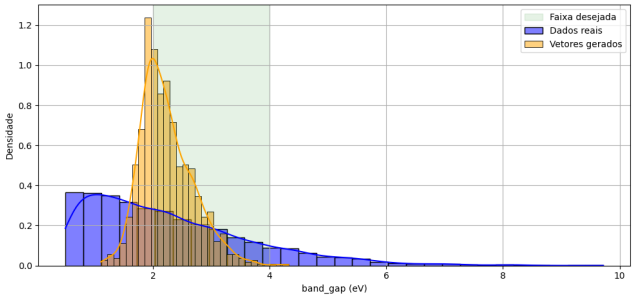

Fig. 3. Distribuição comparativa entre bandgaps reais e gerados, destacando a faixa de interesse.

VII. CONCLUSÃO

O pipeline de aprendizado supervisionado proposto para geração de materiais NLO com bandgap direcionado, combinando regressão RF e modelo generativo CVAE-χ^2, apresentou resultados promissores. O RF obteve desempenho consistente (R^2 acima de 0,82 e próximo de 0,99 na validação cruzada), enquanto o CVAE-χ^2 gerou 65,1% de candidatos válidos na faixa de 2–4 eV. A análise gráfica confirmou que os vetores sintéticos ocuparam de forma mais equilibrada toda região de interesse, evidenciando o potencial da abordagem para aplicações em óptica quântica integrada e na fotônica quântica, importantes na computação e comunicação quântica. Como trabalhos futuros, destaca-se a validação com simulações DFT e integração com bases experimentais, visando consolidar o pipeline como ferramenta estratégica na descoberta de novos materiais.

REFERÊNCIAS

[1] Chen, C.; et al. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Nature Communications, v. 12, p. 3135* (2021). DOI: 10.48550/arXiv.1812.05055.

[2] Haq, S.; et al. Design and evaluation of indacenothienothiophene based functional materials for second and third order nonlinear optics properties via DFT approach. *Scientific Reports, v. 15, p. 13262* (2025). DOI: 10.1038/s41598-025-96902-x.

[3] Horton, M. K.; et al. Accelerated data-driven materials science with the Materials Project. *Nature Materials* (2025). DOI: 10.1038/s41563-025-02272-0.

[4] Matsunoshita, K.; et al. Optimization of force-field potential parameters using conditional variational autoencoder. *Science and Technology of Advanced Materials: Methods, v. 3, n. 1* (2023). DOI: 10.1080/27660400.2023.2253713.

[5] Ward, L.; et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science, v. 152, p. 60–69* (2018). DOI: 10.1016/j.commatsci.2018.05.018.

[6] Xie, C.; et al. A prediction-driven database to enable rapid discovery of nonlinear optical materials. *npj Computational Materials, v. 9, p. 116* (2023). DOI: 10.1007/s40843-023-2592-x.