Lucas Antonio Gigante

Revisão teórica da previsão de *band gap* em perovskitas por meio de modelos de *Machine Learning* e Teoria do Funcional da Densidade (DFT)

São Carlos

2025

Lucas Antonio Gigante

Revisão teórica da previsão de *band gap* em perovskitas por meio de modelos de *Machine Learning* e Teoria do Funcional da Densidade (DFT)

Trabalho de Conclusão de Curso para obtenção do título de Bacharel em Engenharia Física da Universidade Federal de São Carlos. Área de concentração: São Carlos

Orientação: Vivaldo Leiria Campo Junior

Universidade Federal de São Carlos

Centro de Ciências Exatas e Tecnológicas

Departamento de Física

São Carlos

2025

Gigante, Lucas Antonio

Revisão teórica da previsão de bandgap em perovskitas por meio de modelos de Machine Learning e Teoria do Funcional da Densidade (DFT) / Lucas Antonio Gigante -- 2025.
58f.

TCC (Graduação) - Universidade Federal de São Carlos, campus São Carlos, São Carlos
Orientador (a): Vivaldo Leiria Campo Junior
Banca Examinadora: Matheus Paes Lima, Leonardo Kleber Castelano
Bibliografia

1. Bandgaps em perovskitas. 2. Machine learning. 3. Teoria do funcional da densidade. I. Gigante, Lucas Antonio. II. Título.

Ficha catalográfica desenvolvida pela Secretaria Geral de Informática (SIn)

DADOS FORNECIDOS PELO AUTOR

Bibliotecário responsável: Arildo Martins - CRB/8 7180

FOLHA DE APROVAÇÃO

Lucas Antonio Gigante

Revisão teórica da previsão de *band gap* em perovskitas por meio de modelos de *Machine Learning* e Teoria do Funcional da Densidade (DFT)

Trabalho de Conclusão de Curso para obtenção do título de Bacharel em Engenharia Física da Universidade Federal de São Carlos. São Carlos, 04 de dezembro de 2025.

Orientador(a)

_____

Dr. Vivaldo Leiria Campo Junior

Universidade Federal de São Carlos

Examinador(a)

_____

Dr. Matheus Paes Lima

Universidade Federal de São Carlos

Examinador(a)

_____

Dr. Leonardo Kleber Castelano

Universidade Federal de São Carlos

FICHA DE AVALIAÇÃO DE TRABALHO FINAL DE CURSO

ENGENHARIA FÍSICA

Aluno(a): Lucas Antonio Gigante

Título: Revisão teórica da previsão de bandgap em perovskitas por meio de modelos de Machine Learning e Teoria do Funcional da Densidade (DFT)

Prof(a). Orientador(a): Vivaldo Leiria Campo Júnior

Prof(a). Examinador(a) 1: Matheus Paes Lima

Prof(a). Examinador(a) 2: Leonardo Kleber Castelano

Itens Avaliados	Orientador	Examinador 1	Examinador 2
Redação **(atribuir notas de 0 a 2)**	2,0	2,0	2,0
Apresentação oral **(atribuir notas de 0 a 2)**	2,0	1,5	1,6
Conteúdo desenvolvido no trabalho **(atribuir notas de 0 a 4)**	3,5	3,5	4,0
Arguição **(atribuir notas de 0 a 2)**	1,5	2,0	2,0
Total	9,0	9,0	9,6

Nota Final do Estudante: 9,2

Observações:

São Carlos, 04 de dezembro de 2025.

gov.br Documento assinado digitalmente
VIVALDO LEIRIA CAMPO JUNIOR
Data: 04/12/2025 15:58:19-0300
Verifique em https://validar.iti.gov.br

Prof. Vivaldo Leiria Campo Jr.

gov.br Documento assinado digitalmente
MATHEUS PAES LIMA
Data: 04/12/2025 16:08:51-0300
Verifique em https://validar.iti.gov.br

Prof. Matheus Paes Lima

gov.br Documento assinado digitalmente
LEONARDO KLEBER CASTELANO
Data: 04/12/2025 17:42:12-0300
Verifique em https://validar.iti.gov.br

Prof. Leonardo Kleber Castelano

Ao meu falecido pai.
Honrar teu nome é minha eterna missão.

AGRADECIMENTO

Primeiramente, agradeço minha família, em especial meu irmão Diego e minha mãe Tânia, que me apoiaram todos estes anos e são os principais responsáveis por eu chegar até aqui.

Em segundo lugar, agradeço a meu orientador, Prof. Dr. Vivaldo, que me ajudou muito durante este semestre com a preparação da monografia, sempre atencioso e paciente com minhas dúvidas.

Agradeço à República Voodoo, minha casa durante a graduação, e às inúmeras pessoas incríveis que, além de compartilharem moradia comigo, dividiram momentos únicos, sem vocês a graduação não teria sido tão especial. Agradeço também à Tia Gleice, que foi praticamente uma segunda mãe para mim nesse período.

Agradeço aos amigos de Orlândia e Ribeirão Preto, em especial Victor, Rafael, Eduardo, João Carlos, Julio, Marcelo, Davi e Leonardo. Mesmo distantes, vocês se mantiveram presentes em minha vida, nos bons e nos maus momentos.

Agradeço aos amigos que fiz na graduação, Victor, Hugo, Felipe, Leandro, Jucelio e Luiz. As várias madrugadas viradas estudando com vocês jamais serão esquecidas.

Enfim, agradeço ao Prof. Dr. Fábio, coordenador do curso, pela disposição em me ajudar e a esclarecer dúvidas pertinentes a respeito da graduação ao longo desse tempo. Agradeço também aos demais professores que me trouxeram uma enorme bagagem de conhecimento, a qual levarei para a vida.

RESUMO

Este trabalho apresenta uma revisão teórica acerca da previsão de *band gap* em perovskitas, com foco na integração entre métodos tradicionais de primeiros princípios, representados pela Teoria do Funcional da Densidade (DFT), e modelos de *Machine Learning* (ML) aplicados à ciência dos materiais. Inicialmente, são postulados os fundamentos teóricos de materiais sólidos, como estruturais cristalinas e fatores que influenciam o *band gap*. A continuidade dessa discussão é prosseguida com a descrição das particularidades das perovskitas, sua estrutura e comportamento optoeletrônico. Em seguida, explora-se a formulação da DFT, suas limitações práticas associadas às aproximações do funcional de troca e correlação e ao alto custo computacional em sistemas extensos. A partir disso, apresenta-se a relevância de modelos de *Machine Learning* como alternativa para mitigar tais limitações, com análise comparativa de diferentes algoritmos utilizados na predição do *band gap* de perovskitas, incluindo *Random Forest*, *Support Vector Regression* e métodos de *Boosting*. Os resultados reportados na literatura indicam que esses modelos alcançam métricas de desempenho competitivas quando treinados em bases de dados adequadas, reforçando o potencial da abordagem híbrida entre DFT e ML. Por fim, discute-se a importância da expansão contínua de bases de dados para melhor interpretabilidade dos modelos, da seleção criteriosa de descritores e da investigação de estratégias de engenharia composicional para a consolidação de vantagens das perovskitas como materiais promissores para tecnologias optoeletrônicas emergentes.

Palavras-chave: *band gap*; perovskitas; teoria do funcional da densidade; *machine learning*; ciência dos materiais; algoritmo.

ABSTRACT

This work presents a theoretical review of band gap prediction in perovskites, focusing on the integration between traditional first-principles methods, represented by Density Functional Theory (DFT), and Machine Learning (ML) models applied to materials science. Initially, the theoretical foundations of solid materials are postulated, such as crystalline structures and factors that influence the band gap. This discussion continues with a description of the particularities of perovskites, their structure, and optoelectronic behavior. Next, the formulation of DFT is explored, along with its practical limitations associated with exchange and correlation functional approximations and the high computational cost in extensive systems. Based on this, the relevance of Machine Learning models as an alternative to mitigate these limitations is presented, with a comparative analysis of different algorithms used in perovskite band gap prediction, including Random Forest, Support Vector Regression, and Boosting methods. The results reported in the literature indicate that these models achieve competitive performance metrics when trained on suitable datasets, reinforcing the potential of the hybrid approach between DFT and ML. Finally, it is discussed the importance of the continuous expansion of datasets for better interpretability of the models, the careful selection of descriptors, and the investigation of compositional engineering strategies for consolidating the advantages of perovskites as promising materials for emerging optoelectronic technologies.

Keywords: band gap; perovskites; density functional theory; machine learning; materials science; algorithm.

LISTA DE FIGURAS

Figura 1 – Fluxograma de atuação de diversos campos de estudo para descoberta de novas aplicações na ciência dos materiais..... 13

Figura 2 – Célula de perovskita em preparação para realização de ensaios..... 14

Figura 3 – Representação 2D de (a) SiO2 cristalino e (b) SiO2 não cristalino 16

Figura 4 – Diferentes tipos de redes de Bravais. (a), (b) e (c) são: cúbica simples (SC), cúbica de corpo centrado (BCC) e cúbica de face centrada (FCC); (d) e (e) são tetragonais; (f), (g), (h) e (i) sistemas ortorrômbicos; (j) e (k) sistemas rhomboédricos; (l) e (m) monoclínicos e (n) triclínico..... 17

Figura 5 – *Band gap* direto (a) versus *band gap* indireto (b)..... 19

Figura 6 – Disposição estrutural do arranjo ABX_3, em uma forma cúbica com cantos octaédricos ao longo do espaço tridimensional..... 20

Figura 7 – Representação esquemática da célula unitária de uma perovskita dupla..... 21

Figura 8 – Estruturas cristalinas de perovskitas para diferentes valores de fator de tolerância..... 22

Figura 9 – Viabilidade de aplicação fotovoltaica de uma perovskita de acordo com o fator de tolerância de Goldschmidt versus raios efetivos do cátion do sítio A..... 23

Figura 10 – Escada de Jacó de DFT, uma representação gráfica verticalizada das aproximações da DFT (à direita) e seus parâmetros (à esquerda) em relação a acurácia química..... 29

Figura 11 – Fluxo para utilização de algoritmos de acordo com o tipo de aprendizado em relação ao conjunto de dados usado..... 33

Figura 12 – Comportamento do tubo ε, em que as variáveis de folga assumem valores nulos dentro dele..... 37

Figura 13 – Esquema simples de uma árvore de regressão..... 38

Figura 14 – Esquema de funcionamento de uma rede neural simples..... 40

Figura 15 – Performance do modelo MLP durante o treinamento..... 53

LISTA DE TABELAS

Tabela 1 – Categorização de perovskitas de acordo com seu fator dimensional 24

Tabela 2 – Dados coletados de células solares de diferentes materiais.....45

Tabela 3 – Métricas de performance em perovskitas ABX_3.....47

Tabela 4 – Métricas de performance para predição de *band gap* direto.....48

Tabela 5 – Métricas de performance para predição de *band gap* indireto.....49

Tabela 6 – Métricas de performance em perovskitas duplas e 2D.....50

Tabela 7 – Métricas de performance em perovskitas de nitreto.....52

LISTA DE ABREVIATURAS E SIGLAS

CINE – Centro de Inovação em Novas Energias

CNPEM – Centro Nacional de Pesquisa em Energia e Materiais

DFT – *Density Functional Theory* (Teoria do Funcional da Densidade)

E_g – Energia de *band gap* (Energia da banda de energia proibida)

GBM – *Gradient Boosting Machine* (Máquina de Aumento de Gradiente)

HSE06 – Heyd–Scuseria–Ernzerhof 2006 (funcional híbrido para DFT)

MAE – *Mean Absolute Error* (Erro Absoluto Médio)

ML – *Machine Learning* (Aprendizado de Máquina)

MLP – *Multilayer Perceptron* (Perceptron de Multicamadas)

PBE – Perdew–Burke–Ernzerhof (funcional GGA)

PSC – *Perovskite Solar Cell* (Célula Solar de Perovskita)

RF – *Random Forest*

RMSE – *Root Mean Squared Error* (Raiz do Erro Quadrático Médio)

R^2 – Coeficiente de determinação

SVR – *Support Vector Regression* (Regressão de Vetor de Suporte)

UFSCar – Universidade Federal de São Carlos

UNESP – Universidade do Estado de São Paulo

UNICAMP – Universidate Estadual de Campinas

XGBoost – *Extreme Gradient Boosting* (Aumento de Gradiente Extremo)

SUMÁRIO

1 INTRODUÇÃO 13
2 FUNDAMENTOS TEÓRICOS 16
2.1 ESTRUTURA CRISTALINA DOS SÓLIDOS 16
2.2 PEROVSKITAS 19
2.2.1 Estrutura das Perovskitas 21
2.2.2 Band gap em Perovskitas 23
2.3 TEORIA DO FUNCIONAL DA DENSIDADE 25
2.3.1 O Problema de Muitos Corpos e a Proposta da DFT 25
2.3.2 Aproximações para o Funcional de Troca e Correlação 29
2.4 MACHINE LEARNING E ALGORITMOS 32
2.4.1 Regressões Lineares e Métodos de Regularização 34
2.4.2 Métodos Baseados em Kernel 35
2.4.3 Árvores de Decisão, Random Forests e Boosting 37
2.4.4 Redes Neurais Artificiais 39
2.4.5 Métricas de Avaliação em Modelos de Machine Learning 40
3 DISCUSSÃO 42
3.1 A IMPORTÂNCIA DA PREDIÇÃO DO BAND GAP EM MATERIAIS 43
3.2 POTENCIAL DE APLICAÇÃO OPTOELETRÔNICA DAS PEROVSKITAS 43
3.3 MACHINE LEARNING COMO COMPLEMENTO À DFT 45
3.3.1 Predição de Band gap em Perovskitas ABX3 47
3.3.2 Predição de Band gap em Perovskitas ABX3 com Discrição de Band gap 48
3.3.3 Predição de Band gap em Perovskitas Duplas e 2D 50
3.3.4 Predição de Band gap em Perovskitas de Nitreto 52
4 CONCLUSÃO 53

1 INTRODUÇÃO

A capacidade de prever propriedades estruturais, eletrônicas e ópticas de materiais com precisão e eficiência tem se tornado um dos pilares da ciência moderna dos materiais. Nas últimas décadas, métodos computacionais baseados em primeiros princípios, como a Teoria do Funcional da Densidade (DFT), transformaram a maneira como novos compostos são estudados, permitindo compreender fenômenos fundamentais sem a necessidade exclusiva de experimentação. Apesar desse avanço, o custo computacional associado à DFT, especialmente em cálculos envolvendo funcionais mais sofisticados com sistemas na ordem de milhares de átomos, traz custos elevados e restringe sua aplicação em larga escala. Esse panorama motivou o surgimento de alternativas complementares baseadas em *Machine Learning* (ML), capazes de acelerar previsões e auxiliar na descoberta de propriedades de novos materiais (SCHLEDER et al., 2019).

Nesse contexto, a integração entre DFT e ML representa uma estratégia capaz de combinar rigor teórico com eficiência computacional. Enquanto a DFT fornece um alicerce robusto para o cálculo de propriedades fundamentais, os algoritmos interpretam os dados e são treinados a prever padrões matemáticos presentes em bases de dados previamente calculadas, oferecendo predições rápidas para novos compostos com desempenho comparável a métodos tradicionais. Logo, a parceria entre essas duas vertentes otimiza o fluxo de pesquisa em materiais, ilustrado na Figura 1:

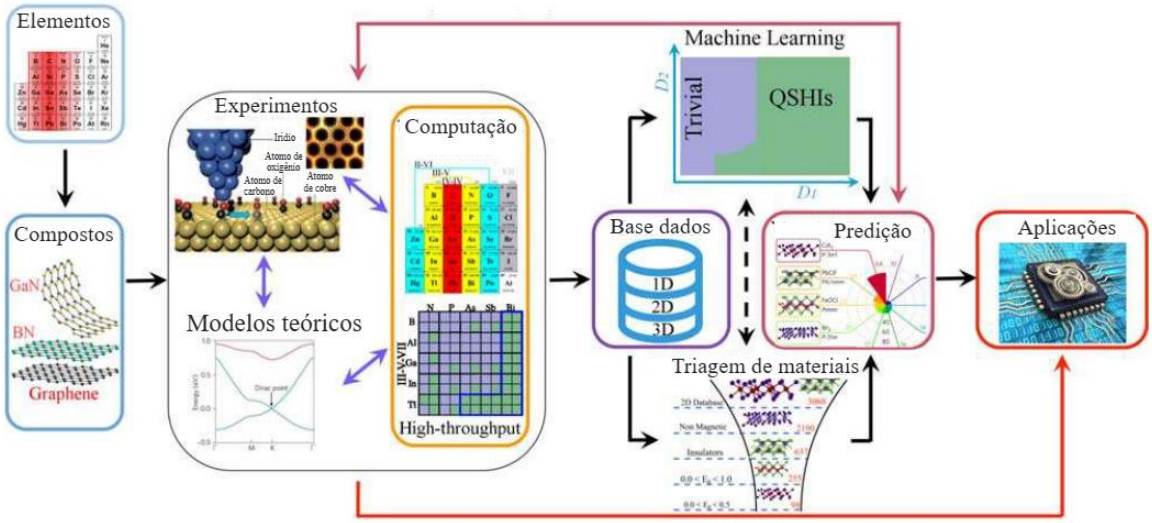


Figura 1: Fluxograma de atuação de diversos campos de estudo para descoberta de novas aplicações na ciência dos materiais. Fonte: Adaptado de (SCHLEDER et al, 2019).

Como é possível analisar pela Figura 1, existem diversas etapas anteriores de modelos teóricos, experimentos e outros tratamentos para se consolidar uma base de dados adequada a ser consumida por ML. A partir disso, a triagem de materiais é direcionada através dos resultados obtidos pelas predições, sintetizando os candidatos mais promissores e poupando recursos durante o processo (SCHLEDER et al., 2019).

Dentre esses materiais, destacam-se as perovskitas, compostos minerais que possuem propriedades que trazem valor às aplicações optoeletrônicas, como LEDs, lasers, emissores, capacitores, dispositivos piezoelétricos e, em especial, as células solares de perovskitas (PSCs) (ZHANG et al., 2023). Apesar de desafios com a estabilidade desses dispositivos e a presença de elementos tóxicos, como o chumbo, as PSCs possuem o potencial para se tornar uma das tecnologias fotovoltaicas mais baratas do mercado, graças ao uso de materiais de baixo custo e métodos de fabricação simples e escaláveis, alinhado com o seu comportamento eletrônico ajustável, viabilizando maneiras de se buscar maior eficiência e resultados (SONG et al., 2017).

No Brasil, a pesquisa em células solares de perovskita tem ganhado força como parte de uma competição tecnológica global, com laboratórios nacionais trabalhando para superar desafios centrais como a instabilidade do material. Diversos grupos de pesquisados de universidades como UNESP e UNICAMP, por meio do CINE, estão engajados com a iniciativa, usufruindo da infraestrutura avançada proveniente do CNPEM para entender as propriedades do material e seu comportamento em células fotovoltaicas (JONES, 2023).

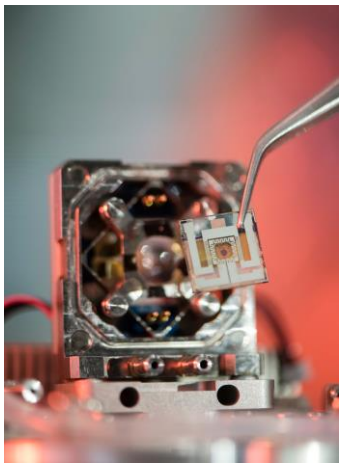


Figura 2: Célula de perovskita em preparação para realização de ensaios. Fonte: (JONES, 2023).

Portanto, o propósito deste trabalho é realizar a revisão teórica desta busca por resultados satisfatórios de predição do *band gap*, ou seja, a energia da banda proibida, por meio da sinergia entre DFT e *Machine Learning*. A metodologia empregada consiste na análise de diversos estudos e pesquisas indicando métricas de performance dos algoritmos envolvidos para

descrever perovskitas em diferentes condições. Dessa forma, é possível traçar uma conclusão sobre a importância desse método e entender perspectivas futuras de melhoria para conseguir proporcionar o entendimento das características das perovskitas e viabilizar sua aplicação.

2 FUNDAMENTOS TEÓRICOS

2.1 ESTRUTURA CRISTALINA DOS SÓLIDOS

Quando estudamos um material sólido, podemos caracterizá-lo, inicialmente, em duas categorias: amorfos e cristalinos. Em amorfos, como vidros, não existe necessariamente uma ordenação de longo alcance que rege a posição dos átomos dispersos ao longo da estrutura do sólido, o que torna suas propriedades físicas mais difíceis de serem descritas e, portanto, comumente isotrópicas (CALLISTER, 2016). Em sólidos cristalinos, os átomos estão arranjados de forma periódica no espaço, ocorrendo padrões que se repetem regularmente.

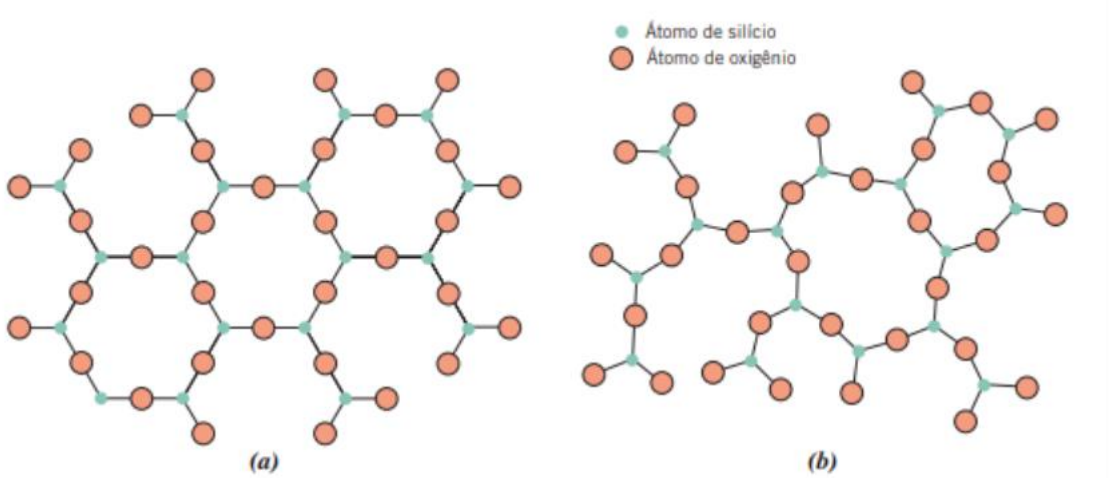


Figura 3: Representação 2D de (a) SiO2 cristalino e (b) SiO2 não cristalino. Fonte: (CALLISTER, 2016).

Pela Figura 3, é possível visualizar o conceito dessa ordenação nos átomos no espaço do sólido. A descrição matemática dessa periodicidade é feita a partir do conceito de rede de Bravais, que representa o conjunto de todos os pontos no espaço que possuem o mesmo ambiente atômico. Cada ponto R da rede pode ser expresso como uma combinação linear dos vetores primitivos a_1, a_2 e a_3 de acordo com a relação:

$$R = n_1 a_1 + n_2 a_2 + n_3 a_3 \quad (1)$$

em que n_1, n_2 e n_3 são números inteiros (ASHCROFT; MERMIN, 1976). Existem quatorze tipos distintos de redes de Bravais tridimensionais, classificadas conforme a simetria e os parâmetros de rede de cada sistema cristalino, apresentadas na Figura 4:

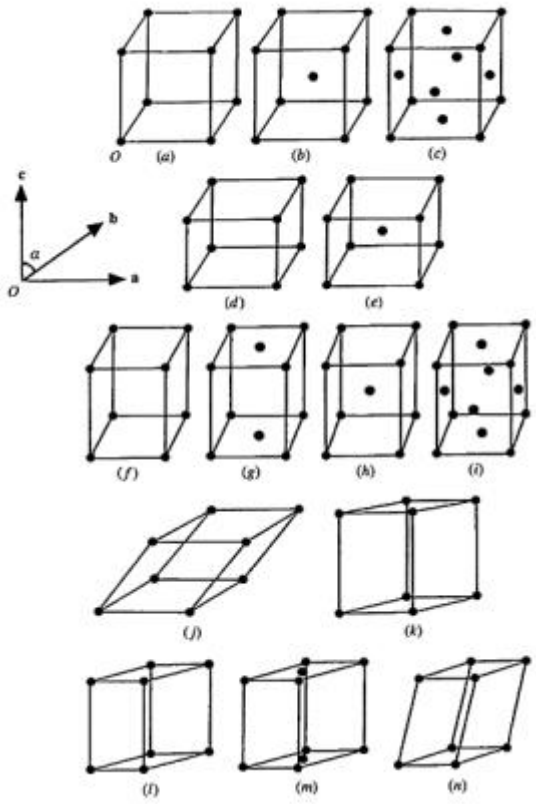


Figura 4: Diferentes tipos de redes de Bravais. (a), (b) e (c) são: cúbica simples (SC), cúbica de corpo centrado (BCC) e cúbica de face centrada (FCC); (d) e (e) são tetragonais; (f), (g), (h) e (i) sistemas ortorrômbicos; (j) e (k) sistemas rhomboédricos; (l) e (m) monoclínicos e (n) triclínico. Fonte: (MYERS, 2002).

Associada a cada rede de Bravais, encontra-se a célula unitária, definida como o menor volume do cristal capaz de, por repetições translacionais de acordo com os vetores primitivos, reconstruir toda a estrutura do sólido sem se sobrepor. Essa célula contém o conjunto de átomos que se repete periodicamente de maneira simétrica e que define a composição química e estrutural do material, constituindo a base para o estudo das propriedades físicas dos sólidos, uma vez que representa o elemento fundamental da estrutura periódica (ASHCROFT; MERMIN, 1976).

A disposição periódica dos átomos impõe um potencial periódico aos elétrons que se movem no interior do sólido. Esse potencial dá origem à rede recíproca, uma construção matemática derivada dos vetores primitivos da rede real. A rede recíproca é essencial para o entendimento da difração de elétrons e para a definição da zona de Brillouin, que representa o domínio fundamental no espaço dos vetores de onda onde se encontram os estados eletrônicos permitidos (ASHCROFT; MERMIN, 1976).

O comportamento dos elétrons sob esse potencial periódico é descrito pelo Teorema de Bloch, o qual estabelece que as funções de onda dos elétrons em um cristal podem ser expressas na forma:

$$\psi_k(r) = e^{ik \cdot r} u_k(r) \quad (2)$$

onde k é o vetor de onda e $u_k(r)$ é uma função com a mesma periodicidade da rede cristalina. Infere-se a partir do Teorema de Bloch que o movimento eletrônico em um sólido é quantizado por meio do vetor de onda cristalino k, que assume valores discretos dentro da zona de Brillouin (ASHCROFT; MERMIN, 1976).

A partir desse formalismo, observa-se que a energia dos elétrons em um cristal não é contínua como no vácuo, mas organizada em bandas de energia permitidas e regiões proibidas, denominadas *band gap*s. Essa estrutura de bandas surge devido à interferência entre ondas eletrônicas equivalentes, gerando intervalos de energia nos quais nenhum estado eletrônico pode existir. A periodicidade do potencial é a responsável pela abertura de lacunas de energia nas fronteiras da zona de Brillouin, caracterizando as diferentes naturezas de materiais metálicos, semicondutores e isolantes (ASHCROFT; MERMIN, 1976).

Nos metais, há o preenchimento parcial de elétrons nas bandas de energia, o que facilita a movimentação eletrônica dado que a lacuna de energia é inexistente, fenômeno ocasionado pela sobreposição de bandas (em geral para sólidos 2D e 3D). Já nos semicondutores e isolantes, existe uma lacuna de energia entre essas bandas. A diferença entre ambos está no tamanho do *band gap*: semicondutores possuem gaps relativamente pequenos (geralmente inferiores a 3 eV), enquanto isolantes apresentam gaps muito maiores (acima de 5 eV). Em materiais semicondutores, a magnitude do *band gap* determina a energia mínima necessária para a excitação eletrônica, sendo, portanto, um parâmetro central na determinação das propriedades ópticas e eletrônicas (KITTEL, 2005).

O *band gap* pode ainda ser classificado como direto ou indireto, conforme a posição relativa do topo da banda de valência e do fundo da banda de condução, como é possível visualizar na Figura 5 a seguir:

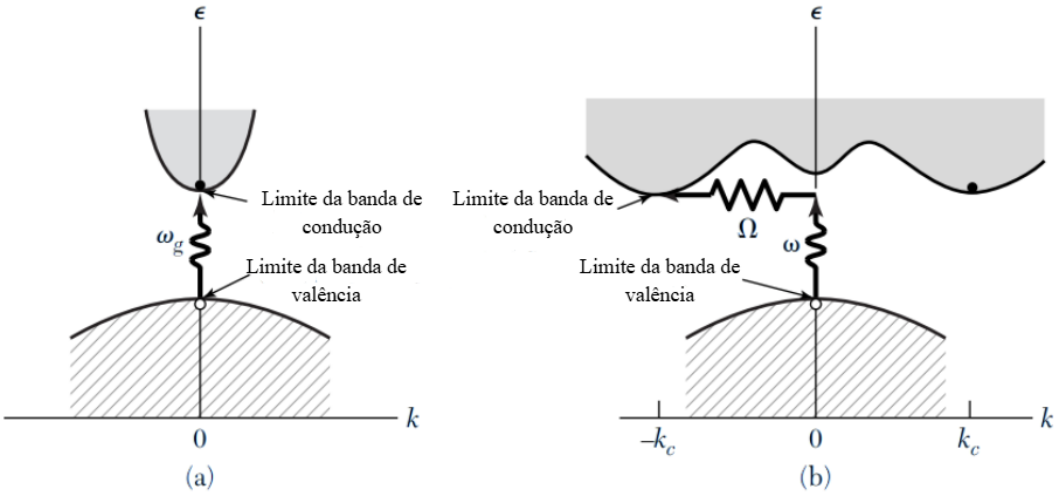


Figura 5: *Band gap* direto (a) versus *band gap* indireto (b). Fonte: Adaptado de (KITTEL, 2005).

A Figura 5 evidencia essas condições relativas através de processos de absorção de fótons. Para *band gap*s diretos, o fundo da banda de condução coincide, para um mesmo vetor de onda k, com o topo da banda de valência, fazendo o fóton ser absorvido em uma frequência ω_g, capaz de aferir o *band gap* como $E_g = \hbar\omega_g$. Para *band gap*s indiretos, caracterizados pela presença de elétrons e buracos, o máximo da banda de valência e o mínimo da banda de condução ocorrem em diferentes valores de k, assim, o *band gap* é definido como $E_g = \hbar\omega_g + \hbar\Omega$, em que Ω é a frequência do fônon emitido, tornando-se necessária tanto a absorção do fóton quanto do fônon para completar a transição eletrônica (KITTEL, 2005).

A estrutura cristalina e o tipo de ligação química influenciam diretamente a largura e a natureza do *band gap*. Estruturas compactas, com maior sobreposição entre orbitais atômicos, tendem a apresentar bandas mais largas e lacunas de energia menores. Por outro lado, redes mais abertas, com menor acoplamento eletrônico, resultam em gaps maiores (ASHCROFT; MERMIN, 1976). Outro aspecto importante é a massa efetiva dos portadores de carga, que pode ser obtida a partir da curvatura das bandas de energia próximas aos extremos de valência e condução. Bandas mais curvas correspondem a portadores mais leves, o que implica maior mobilidade eletrônica, enquanto bandas mais planas indicam portadores mais pesados e mobilidade reduzida. (KITTEL, 2005).

2.2 PEROVSKITAS

As perovskitas fazem parte de uma ampla classe de materiais cristalinos que despertam grande interesse na física do estado sólido e na ciência dos materiais devido à notável versatilidade estrutural, que permite uma série de aplicações industriais de grande valor. Sua origem está relacionada ao mineral titanato de cálcio ($CaTiO_3$), que foi descoberto em 1839 pelo alemão Gustav Rose e batizado em homenagem a Lev Perovski, um mineralogista russo. Atualmente, o termo perovskita designa qualquer composto similar cuja estrutura cristalina siga o arranjo ABX_3 (MIAH et al., 2024), como pode ser observado na Figura 6 abaixo:

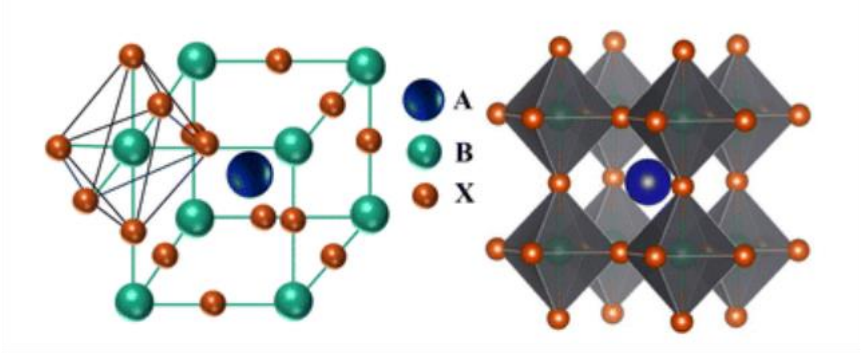
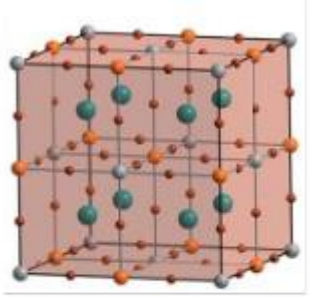


Figura 6: Disposição estrutural do arranjo ABX_3, em uma forma cúbica com cantos octaédricos ao longo do espaço tridimensional. Fonte: (MIAH et al., 2024).

De maneira geral, as perovskitas apresentam em sua estrutura o cátion A, de natureza orgânica-inorgânica, monovalente e de maior raio; o cátion B, metálico, bivalente e de menor raio e por fim, um ânion X relacionado aos halogênios. Em alguns casos, o elemento X pode ser substituído por outros elementos como oxigênio (como visto no mineral original), nitrogênio, carbono ou até mesmo boro, o que impacta na natureza dos cátions A e B, que, para essa particularidade, serão bivalentes e tetravalentes, respectivamente (MIAH et al., 2024).

Diferentes variações estruturais podem ocorrer nas perovskitas em decorrência das modificações de diversas variáveis, como será explicado na seção 2.2.2, mas um dos tipos mais interessantes deste material são as perovskitas duplas, que apresentam fórmula estrutural $A_2BB'X_6$, representando uma extensão do clássico composto ABX_3, em que dois cátions diferentes (B e B′) ocupam alternadamente a posição central da rede octaédrica, e o ânion X forma octaedros ao redor desses cátions, como podemos visualizar na Figura 7 abaixo:



Célula FCC
$Cs_2AgBiBr_6$

Figura 7: Representação esquemática da célula unitária de uma perovskita dupla. Fonte (SLAVNEY, 2019).

Um aspecto estrutural-chave é que a substituição sistemática dos cátions B e B′ altera diretamente os orbitais de fronteira (de valência e condução) que formam a banda de condução e o topo da banda de valência, o que impacta o tipo, direto ou indireto, e a amplitude do *band gap*. Com isso, a alta capacidade ajustável de seu *band gap* dão destaque para as perovskitas como uma via promissora para otimização de propriedades físicas como absorção de luz, mobilidade de portadores e estabilidade térmica-química, combinando a sua estrutura ordenada com a versatilidade dos pares iônicos nos sítios B/B′ (SLAVNEY, 2019).

2.2.1 Estrutura das Perovskitas

A estrutura cristalina das perovskitas é fundamental para determinar suas propriedades físicas. Retomando um pouco do que foi discutido na seção anterior, podemos definir um arranjo para estrutura cúbica FCC ideal das perovskitas, em que se apresente o cátion A nas posições de vértice da célula unitária cúbica, o cátion B no centro do cubo e o ânion X nas posições médias das arestas, formando octaedros BX_6 interconectados. Essa organização permite uma grande variedade de distorções estruturais, resultantes de variações no tamanho dos íons e nas condições de temperatura e pressão. (MIAH et al., 2024)

A estabilidade da estrutura é comumente avaliada pelo fator de tolerância de Goldschmidt (t), um parâmetro adimensional obtido através da análise de correlação entre os raios iônicos presentes na célula unitária:

$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)} \tag{3}$$

em que r_A, r_B e r_X são os raios iônicos associados aos cátions A e B e o ânion X, e o fator $\sqrt{2}$ é decorrente da geometria tipicamente cúbica das perovskitas. Na Figura 8 a seguir, é possível observar o comportamento da estrutura cristalina das perovskitas de acordo com o parâmetro do fator de tolerância de Goldschmidt:

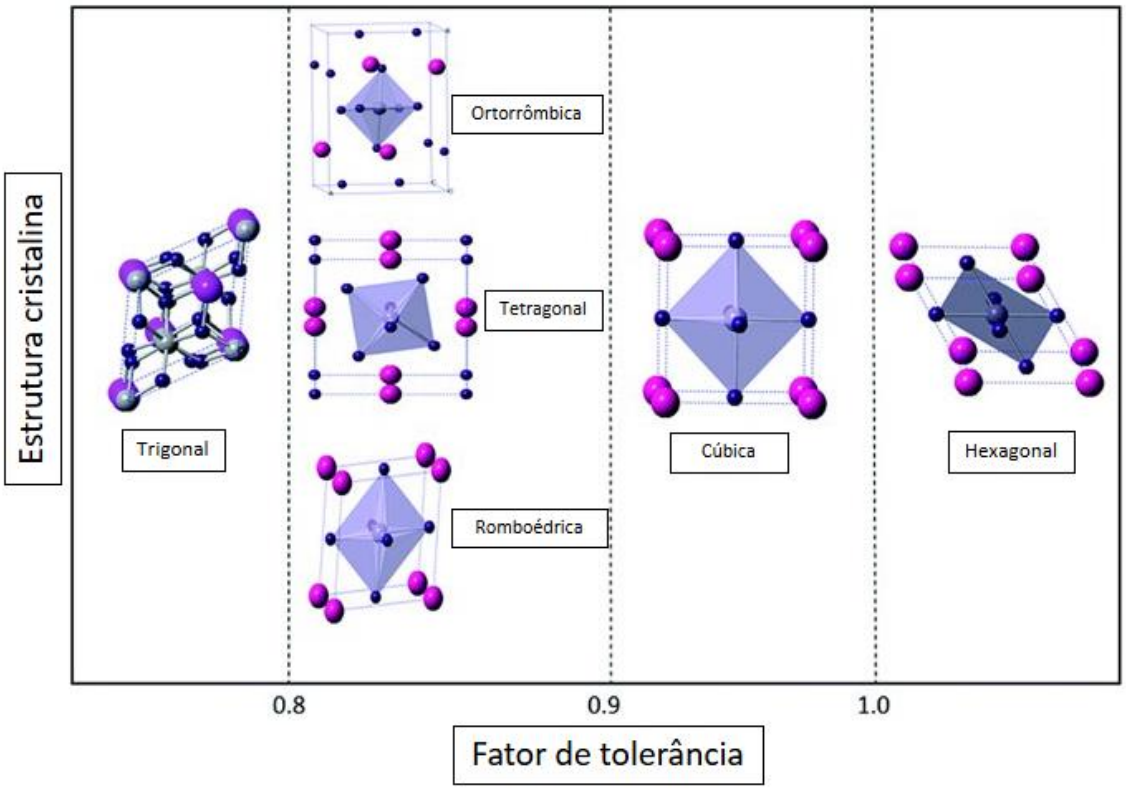


Figura 8: Estruturas cristalinas de perovskitas para diferentes valores de fator de tolerância. Fonte: Adaptado de (YI et al., 2019).

Para valores de fator de tolerância próximos de 1, tem-se a presença de uma estrutura cúbica estável, considerada ideal para as perovskitas. Conforme ocorre variação dos parâmetros da Equação 3, chega-se em resultados diferentes para t, em que pode se observar o início de distorções na estrutura da perovskita, levando a diferentes sistemas cristalinos, como hexagoinais, tetragonais, ortorrômbicos, romboédricos e trigonais. (YI et al., 2019). Essa faixa ideal também é replicável para analisar o potencial de uso de um material na indústria fotovoltaica:

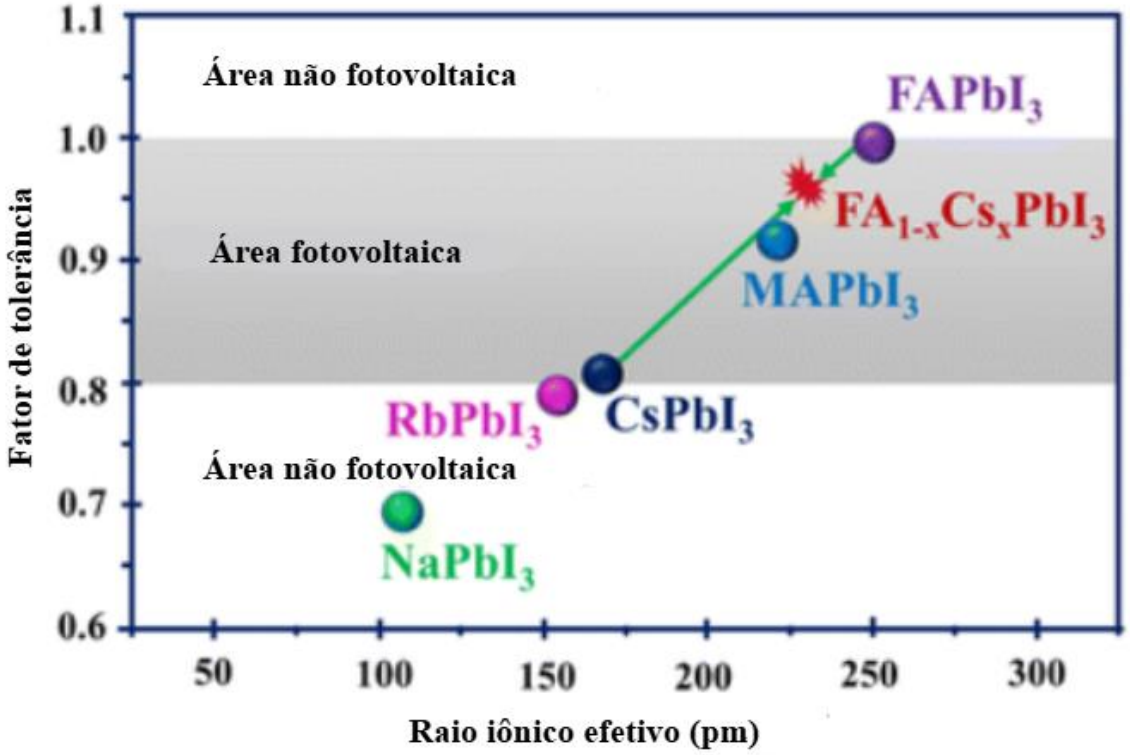


Figura 9: Viabilidade de aplicação fotovoltaica de uma perovskita de acordo com o fator de tolerância de Goldschmidt versus raios efetivos do cátion do sítio A. Fonte: Adaptado de (MIAH et al., 2024).

Portanto, pela Figura 9, que pelo fator de tolerância de Goldschmidt, também é possível de inferir se uma perovskita trará valor aplicável à indústria dada a avaliação da performance do material a partir de parâmetros estruturais, como simetria, estabilidade e formação de defeitos em decorrência de variações de temperatura e pressão (MIAH et al., 2024).

2.2.2 *Band gap* em Perovskitas

As propriedades físicas das perovskitas, em especial aquelas que impactam o *band gap*, derivam diretamente de sua estrutura cristalina. Nestes materiais, o topo da banda de valência é formado predominantemente pelos orbitais p do ânion X, enquanto a banda de condução resulta dos orbitais p e s do cátion metálico B. Assim, qualquer modificação estrutural ou composicional que altere a sobreposição desses orbitais resultará em variações no *band gap*, tornando esse tipo de material um grande alvo de estudos para aplicações que explorem o seu potencial optoeletrônico (MIAH et al., 2024).

A engenharia composicional é o método mais amplamente explorado para modular o *band gap* em perovskitas, baseando-se na substituição de cátions e ânions em sua estrutura cristalina. Fisicamente, a modificação do *band gap* decorre das mudanças nas ligações químicas

e nas interações orbitais entre o cátion metálico (geralmente Pb ou Sn) e o halogênio (I, Br ou Cl), que determinam as posições da banda de valência e da banda de condução. A substituição de íons no sítio B ou X altera o ângulo B-X-B do octaedro e a distância inter-atômica B-X, parâmetros que governam a sobreposição dos orbitais e, portanto, a largura do *band gap*. Substituições que aumentam a eletronegatividade do haleto ou reduzem o ângulo B-X-B tendem a aumentar o *band gap*, enquanto o uso de íons maiores ou menos eletronegativos (como Sn em substituição ao Pb) reduz essa energia. Essa engenharia permite adequar o *band gap* ao espectro solar, otimizando a absorção e a eficiência dos dispositivos fotovoltaicos, ao mesmo tempo em que ajusta parâmetros estruturais, como o fator de tolerância de Goldschmidt, para buscar preservar a estabilidade da rede cristalina (MIAH et al., 2024).

Outro método é o ajuste de *band gap* induzido por pressão, que se baseia na aplicação de uma pressão hidrostática uniforme sobre o cristal, capaz de alterar a distância inter-atômica e os ângulos de ligação dentro da rede perovskita. Do ponto de vista físico, o encurtamento das distâncias B-X sob pressão aumenta a sobreposição dos orbitais e, portanto, reduz o *band gap*. Por exemplo, no composto $FAPbI_3$, foi observada uma redução do *band gap* de 1,57 eV para 1,44 eV sob pressões de até 2,3 GPa, indicando que pequenas distorções estruturais podem alterar significativamente as propriedades eletrônicas. No entanto, a aplicação de pressões excessivas (acima de 5 GPa) pode induzir amorfização e perda da cristalinidade, comprometendo as propriedades fotônicas do material. Assim, o método é eficaz para estudos fundamentais e ajustes reversíveis, mas apresenta limitações práticas na incorporação direta em dispositivos fotovoltaicos devido à necessidade de manutenção da pressão (MIAH et al., 2024).

O fator estrutural de dimensionalidade das perovskitas também representa uma maneira de se ajustar o *band gap* a fim de obter um desempenho satisfatório. Na Tabela 1 a seguir, são apresentados os diferentes tipos de dimensões possíveis para perovskitas, englobando tridimensionais (3D) e as de baixa dimensão (LD):

Tabela 1: Categorização de perovskitas de acordo com seu fator dimensional.

Dimensão	**Formato estrutural**	**Aplicações na indústria**
0D	Pontos quânticos	LEDs, lasers
1D	Nanofios	Fotodetectores, dispositivos optoeletrônicos
2D	Camadas	Células solares, LEDs, detectores, lasers
3D	Formato livre	Dispositivos fotovoltaicos

Fonte: (GAO; HU; LIU, 2023).

À medida que a dimensionalidade diminui, ocorre um aumento do confinamento quântico dos portadores de carga, o que resulta no alargamento do *band gap* e em uma maior energia de ligação dos éxcitons. Nas perovskitas 3D, o forte acoplamento entre octaedros BX_6 produz bandas de condução e valência mais amplas, promovendo menor *band gap* e maior mobilidade eletrônica, características ideais para aplicações fotovoltaicas. Já nas estruturas 2D, a separação entre as camadas inorgânicas por ligantes orgânicos reduz o acoplamento orbital, ampliando o *band gap* e conferindo maior estabilidade química e ambiental. Em geometrias ainda mais confinadas, como 1D e 0D, o confinamento espacial intensifica o aumento do *band gap*, para 1D, e dos níveis discretos de energia, para 0D, possibilitando o controle preciso da emissão e absorção de luz. Assim, o fator de dimensionalidade permite modular o *band gap* das perovskitas, equilibrando eficiência eletrônica e estabilidade de acordo com a aplicação desejada (GAO; HU; LIU, 2023).

2.3 TEORIA DO FUNCIONAL DA DENSIDADE

A Teoria do Funcional da Densidade, do inglês *Density Functional Theory* (DFT), constitui um dos pilares da física do estado sólido e da química quântica moderna, oferecendo uma maneira eficiente de descrever a estrutura eletrônica da matéria sem recorrer explicitamente à função de onda de muitos corpos. Sua origem remonta aos trabalhos fundamentais de Hohenberg, Kohn e Sham nas décadas de 1960, mas sua relevância permanece crescente, impulsionada tanto pela consolidação de cálculos de primeiros princípios quanto pela integração com abordagens de aprendizado de máquina na ciência dos materiais (SCHLEDER et al., 2019).

2.3.1 O Problema de Muitos Corpos e a Proposta da DFT

O ponto de partida da DFT é o problema quântico de muitos elétrons, cuja descrição exata é dada pela equação de Schrödinger independente do tempo. Para um sistema de N elétrons interagentes sob a influência de um potencial externo v(r), o Hamiltoniano é escrito como:

$$H = \sum_{i}^{N} \left(-\frac{\hbar^2 \nabla_i^2}{2m} + v(r_i) \right) + \sum_{i<j} \frac{q^2}{|r_i - r_j|} \tag{4}$$

O primeiro termo da Equação 4 representa a energia cinética dos elétrons, associada ao momento quântico de cada partícula. O segundo termo expressa a energia potencial devido ao

campo externo e o terceiro termo descreve a interação entre elétrons. A função de onda $\Psi(r, r_2, ..., r_N)$ contém toda a informação do sistema, mas depende de 3N variáveis espaciais, o que torna o problema intratável para sistemas de muitos elétrons. (CAPELLE, 2006).

A grande ideia da DFT é reformular o problema de muitos corpos em termos da densidade eletrônica, n(r), uma função de apenas três variáveis espaciais que descreve a distribuição de carga no sistema. Essa densidade é definida a partir da função de onda de muitos corpos como:

$$n(r) = N \int |\Psi(r, r_2, \dots, r_N)|^2 \, d^3r_2 \dots d^3r_N \tag{5}$$

Fisicamente, n(r) representa a probabilidade de se encontrar um elétron na posição r. Ela é uma grandeza observável e experimentalmente acessível, diferentemente da função de onda, que é uma entidade puramente matemática. Além disso, a densidade eletrônica deve satisfazer a condição de normalização

$$\int n(r) \, d^3r = N \tag{6}$$

a qual assegura que a integração de n(r) sobre todo o espaço reproduza o número total de elétrons do sistema. Essa propriedade reforça o caráter físico da densidade como variável fundamental, que contém, de forma condensada, todas as informações relevantes sobre a distribuição eletrônica (CAPELLE, 2006).

Em 1964, Hohenberg e Kohn demonstraram que a densidade eletrônica do estado fundamental contém, em princípio, toda a informação necessária para determinar completamente um sistema quântico de muitos elétrons. O primeiro teorema de Hohenberg-Kohn afirma que existe uma correspondência entre a densidade eletrônica $n_0(r)$ e o potencial externo v(r), desconsiderando-se uma constante aditiva. Isso implica que duas densidades diferentes não podem corresponder ao mesmo potencial externo, o que significa que o conhecimento de $n_0(r)$ é suficiente para determinar, de forma única, a função de onda e todas as propriedades do sistema (KOHN, 1999).

O segundo teorema de Hohenberg-Kohn, de caráter variacional, estabelece que a energia total de um sistema pode ser expressa como um funcional da densidade:

$$E_v[n] = F[n] + \int n(r)\ v(r)\, d^3r \quad (7)$$

O primeiro termo da Equação 7, F[n], é o funcional universal de energia, que contém a energia cinética T[n] e a energia de interação entre elétrons U[n], sendo o mesmo para qualquer sistema eletrônico, já o segundo termo representa a energia potencial associada ao potencial externo. O teorema afirma que a energia total $E_v[n]$ assume seu valor mínimo para a densidade do estado fundamental $n_0(r)$, isto é, $E_v[n_0] \leq E_v[n']$ para qualquer densidade admissível n'(r) (CAPELLE, 2006). Dessa forma, essa formulação inaugura a base variacional da DFT, em que o estado fundamental pode ser obtido minimizando-se um funcional de energia dependente apenas da densidade.

Entretanto, a forma exata do funcional F[n] é desconhecida. Para contornar esse problema, Kohn e Sham, em1965, propuseram um modelo que introduz um sistema fictício de elétrons não interagentes que reproduz exatamente a mesma densidade do sistema interagente real. Essa técnica permite separar os diferentes componentes da energia total, que passa a ser escrita como:

$$E[n] = T_s[n] + U_H[n] + V_{ext}[n] + E_{xc}[n] \quad (8)$$

Na Equação 8, $T_s[n]$ é a energia cinética do sistema não interagente, $U_H[n]$ é a energia de Hartree, que descreve a repulsão clássica entre distribuições de carga eletrônica, $V_{ext}[n]$ corresponde à interação dos elétrons com o potencial externo, e $E_{xc}[n]$ é o funcional de troca e correlação, que contém todos os efeitos quânticos de correlação entre elétrons e as correções à energia cinética (CAPELLE, 2006). A forma funcional da Equação 8 indica uma importante informação: os três primeiros termos podem ser avaliados diretamente, enquanto $E_{xc}[n]$ é o termo que engloba toda a complexidade da interação eletrônica. Em princípio, se esse termo fosse conhecido exatamente, a DFT descreveria com exatidão o estado fundamental de qualquer sistema eletrônico (KOHN, 1999).

A partir dessa decomposição, Kohn e Sham derivaram as chamadas equações de Kohn-Sham, que governam o comportamento dos orbitais $\phi_j(r)$ do sistema:

$$\left[-\frac{1}{2}\nabla^2 + v_{eff}(r)\right]\phi_j(r) = \varepsilon_j \phi_j(r) \quad (9)$$

O primeiro termo da Equação 9 representa o operador energia cinética de uma partícula quântica livre, enquanto v_{eff} (r) é o potencial efetivo, responsável por englobar todas as interações relevantes. Esse potencial é dado por:

$$v_{eff}(r) = v_{ext}(r) + \int \frac{n(r')}{|r - r'|} d^3r' + v_{xc}(r) \quad (10)$$

em que v_{ext} é o potencial externo, o segundo termo é o potencial de Hartree v_H, que representa a repulsão eletrostática clássica entre elétrons, e o terceiro termo é o potencial de troca e correlação, definido como a derivada funcional do termo E_{xc}[n]:

$$v_{xc}(r) = \frac{\delta E_{xc}[n]}{\delta n(r)} \quad (11)$$

Esse termo é responsável por incorporar os efeitos quânticos de correlação eletrônica e de exclusão de Pauli, tornando o sistema fictício de Kohn-Sham equivalente, em densidade, ao sistema real (CAPELLE, 2006).

As equações de Kohn-Sham devem ser resolvidas de forma autoconsistente, pois o potencial efetivo depende da densidade eletrônica, e esta é obtida a partir dos próprios orbitais ϕ_j(r). A densidade é reconstruída pela soma das densidades associadas a cada orbital ocupado:

$$n(r) = \sum_{j=1}^{N} |\phi_j(r)|^2 \quad (12)$$

Como tanto o potencial v_H e v_{xc} dependem de n(r), que depende de ϕ_j(r),como demonstra a Equação 12, que por sua vez depende do potencial efetivo v_{eff}, o processo iterativo de resolução consiste em propor uma densidade inicial, calcular o potencial efetivo correspondente, resolver as equações de Kohn-Sham, atualizar a densidade e repetir o ciclo até a convergência. Esse procedimento fornece simultaneamente a densidade e a energia do estado fundamental do sistema (KOHN, 1999).

O formalismo de Kohn-Sham representa, portanto, um mapeamento exato entre um sistema de elétrons interagentes e um sistema de elétrons não interagentes que compartilham a mesma densidade. Esse mapeamento é o que confere à DFT sua força conceitual e aplicabilidade prática, pois reduz um problema quântico de muitos corpos a um conjunto de equações de partícula única, sem perda de rigor teórico. Embora a forma exata do funcional E_{xc}[n] permaneça desconhecida, o formalismo geral é, em princípio, exato, e constitui a base

de praticamente todas as abordagens computacionais modernas em estrutura eletrônica (SCHLEDER et al., 2019).

2.3.2 Aproximações para o Funcional de Troca e Correlação

Ao longo do desenvolvimento histórico da DFT, as aproximações práticas para E_{xc} foram evoluindo e sendo classificadas de diversas maneiras, que refletem o grau de sofisticação na descrição da densidade eletrônica. Essa classificação é comumente representada na bibliografia como a Escada de Jacó, apresentada por John Perdew (PERDEW, 2001), conforme a Figura 10 a seguir:

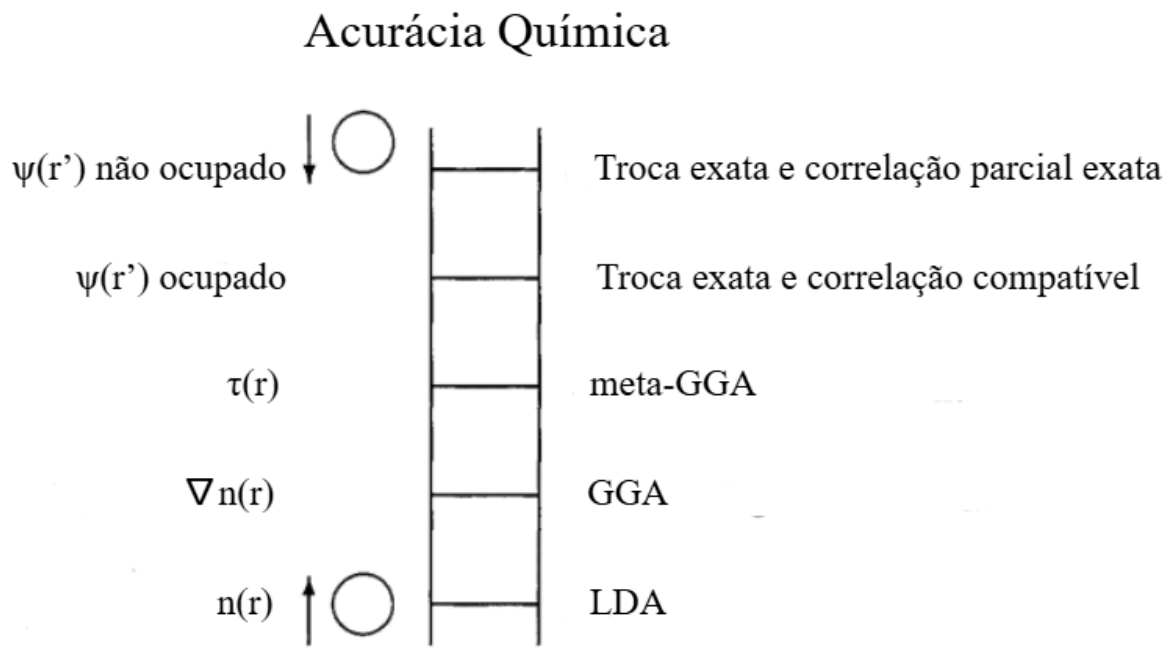


Figura 10: Escada de Jacó de DFT, uma representação gráfica verticalizada das aproximações da DFT (à direita) e seus parâmetros (à esquerda) em relação a acurácia química. Fonte: Adaptado de (PERDEW, 2001).

Iniciando pelo primeiro degrau da escada, a aproximação mais simples dentre elas é a Local Spin Density Approximation (LSDA), ou simplesmente Local Density Approximation (LDA), que assume que cada ponto do sistema pode ser tratado como se estivesse em um gás de elétrons homogêneo. Para este caso, a energia correspondente pode ser descrita como:

$$E_{xc}^{LDA}[\, n_{\uparrow}(r), n_{\downarrow}(r)] \;=\; \int dr\; n(r)\; \varepsilon_{xc}^{LDA}[\, n_{\uparrow}(r), n_{\downarrow}(r)] \qquad (13)$$

em que $n_{\uparrow}$ e $n_{\downarrow}$ são densidades de spin uniformes e ε_{xc} é a energia de correlação de troca por elétron do sistema. Assim, o funcional de troca-correlação é calculado localmente com base na

densidade eletrônica n(r) e na energia conhecida do gás homogêneo. Apesar de simples, a LDA é eficaz para descrever sistemas metálicos e sólidos cristalinos onde a densidade varia lentamente. Contudo, ela tende a trazer resultados pouco acurados devido a erros sistemáticos em relação ao comprimento das ligações químicas e energias de atomização, devido à ausência de dependência explícita dos gradientes da densidade. Além disso, a LDA pode falhar em sistemas que a interação elétron-elétron seja predominante (CAPELLE, 2006).

Para superar essa limitação, surgiram as aproximações generalizadas de gradiente, conhecidas como Generalized Gradient Approximations (GGA), cuja energia de correlação de troca pode ser expressa como:

$$E_{xc}^{GGA}[\, n_{\uparrow}(r), n_{\downarrow}(r)] \;=\; \int dr\; n(r)\; \varepsilon_{xc}^{GGA}[\, n_{\uparrow}(r),\, n_{\downarrow}(r),\, \nabla n_{\uparrow}(r),\, \nabla n_{\downarrow}(r)] \qquad (14)$$

que incorporam não apenas a densidade local n(r), mas também o seu gradiente ∇n(r) (CAPELLE, 2006). Essa dependência adicional permite que a GGA capture variações espaciais mais realistas, tornando-a mais adequada para moléculas e semicondutores. A GGA consegue corrigir os erros sistemáticos da LDA, ainda que continue subestimando os *band gap*s eletrônicos. Essa subestimação decorre do fato de que tanto LDA quanto GGA são funcionais localizados e semilocalizados, respectivamente, incapazes de reproduzir a descontinuidade no potencial de troca-correlação (PERDEW; BURKE; ERNZERHOF, 1996).

Uma terceira categoria de aproximações é composta pelas meta-GGA, com energia de correlação de troca descrita como:

$$E_{xc}^{MGGA}[\, n_{\uparrow}(r), n_{\downarrow}(r)] = \int drn(r)\; \varepsilon_{xc}^{MGGA}[n_{\uparrow}(r), n_{\downarrow}(r), \nabla n_{\uparrow}(r), \nabla n_{\downarrow}(r), \tau_{\uparrow}(r), \tau_{\downarrow}(r)] \qquad (15)$$

que acrescentam uma dependência explícita da densidade da energia cinética eletrônica local τ(r):

$$\tau(r) = \frac{\hbar}{2m}\sum_{i}^{n} |\nabla \phi_i(r)|^2 \qquad (16)$$

o que permite distinguir regiões de diferentes tipos de ligação (como ligações covalentes, iônicas, metálicas) (CAPELLE, 2006). Essas funções fornecem ganhos de precisão significativos em propriedades termodinâmicas e estruturais, mantendo o custo computacional

baixo. Na prática, Meta-GGAs representam um compromisso entre custo e precisão, oferecendo uma correção parcial à deficiência sistemática dos funcionais semilocalizados (SCHLEDER et al., 2019).

Uma evolução conceitual importante veio com os funcionais híbridos, que combinam o caráter semilocal dos funcionais GGA com uma fração de troca exata do termo de Hartree-Fock, cuja energia de correlação de troca agora é descrita pela relação:

$$E_{xc}^{hib} = (1-\alpha)E_{xc}^{GGA} + \alpha E_{x}^{HF} \tag{17}$$

em que $0 \leq \alpha \leq 1$ é um parâmetro de mistura, com valor normalmente atribuído na faixa de 0,15 a 0,25. Através da formulação de Hartree-Fock, introduz-se não localidade ao modelo pela inserção da energia de troca exata:

$$E_{x}^{HF} = \frac{1}{2}\sum_{i,j}^{n}\int dr \int dr' \frac{\phi_i^*(r)\phi_j^*(r')\phi_i(r')\phi_j(r)}{|r-r'|} \tag{18}$$

Essa mistura busca incorporar parte da natureza não local da interação de troca, ausente em LDA e GGA. Do ponto de vista teórico, esses funcionais melhoram a descrição da lacuna eletrônica e a estrutura de bandas, justamente por introduzir a correção da descontinuidade do potencial de troca-correlação. Em contrapartida, seu custo computacional, que envolve resolver a equação de Hartree-Fock, requer grandes esforços e recursos. (SCHLEDER et al., 2019).

A vantagem dos híbridos torna-se clara ao observar que eles aproximam o comportamento da energia de troca exata, corrigindo o erro de auto-interação presente nos funcionais locais e semi-locais. Enquanto LDA e GGA assumem que um elétron interage parcialmente consigo mesmo, o que desloca artificialmente níveis de energia e estreita o gap, os híbridos eliminam parte dessa inconsistência. Ainda assim, eles não são exatos, pois o grau de correção depende da fração de troca exata incluída e pode exigir calibração empírica para obter um parâmetro α mais ideal. (SCHLEDER et al., 2019).

Apesar de demonstrar resultados positivos, é essencial compreender que DFT é uma teoria conceitualmente exata, mas praticamente aproximada. O funcional exato E_{xc} para a energia de correlação de troca é desconhecido e, portanto, cada aproximação introduz erros sistemáticos. As LDA e GGA, por exemplo, tratam bem sistemas homogêneos, mas falham ao descrever corretamente a localização eletrônica e o *band gap*. Já os híbridos corrigem parte

dessas deficiências, mas a um custo computacional alto e sem uma regra universal para a fração ótima de troca exata. Além disso, o caráter de teoria de estado fundamental implica que as energias de Kohn-Sham não correspondem, formalmente, aos níveis energéticos observáveis pois, para obtê-los, é necessário recorrer a teorias que tragam respostas acerca dos estados de excitação como, por exemplo, a aproximação GW, a Equação de Bethe-Salpeter (BSE) e a TD-DFT (Time Dependent DFT, teoria do funcional da densidade dependente do tempo) que visam tratar essa deficiência existente na DFT e trazer a análise de sistemas de estados excitados. (SCHLEDER et al., 2019).

2.4 *MACHINE LEARNING* E ALGORITMOS

O aprendizado de máquina (*Machine Learning*, ou ML) consiste em um conjunto de métodos que buscam inferir padrões ou relações funcionais entre variáveis a partir de dados observacionais. Diferente dos métodos tradicionais de modelagem física, nos quais as leis são explicitamente formuladas, o ML busca construir uma função f(x) que aproxima uma relação subjacente entre descritores x e propriedades alvo y. No contexto científico, essa abordagem é particularmente útil para problemas de alta dimensionalidade ou cuja solução analítica é inviável. A incorporação de ML em ciências de materiais tem se mostrado fundamental para acelerar a descoberta e o entendimento de novos compostos, permitindo estabelecer relações entre estrutura e propriedade a partir de grandes volumes de dados teóricos ou experimentais (SCHLEDER et al., 2019).

Cada modelo pode ser construído a partir de uma configuração inicial de acordo com um conjunto de dados de domínio X que é desejável ser categorizado, como por exemplo o *band gap* e outras propriedades estruturais. Em seguida, define-se: o conjunto de rótulos descritores Y, que será responsável por englobar as categorias do algoritmo; um conjunto S de treinamento, representando uma sequência exemplo de elementos que corresponde a dados já categorizados adequadamente; o preditor, ou classificador, uma função baseada em uma regra de predição h: $X \rightarrow Y$ que é utilizada para prever a categoria de novos pontos de domínio; um modelo de geração de dados, os quais irão compor o conjunto S para treinamento, em que, nesse caso, assume-se uma função f: $X \rightarrow Y$ com a premissa de classificar corretamente os dados gerados para treinamento e, por fim, métricas de sucesso para definir o erro de classificação do modelo, ou seja, o erro de h, através da probabilidade de obter um dado aleatório de tal forma que, para um determinado valor x do conjunto X, h(x) seja diferente de f(x). Essencialmente,

os principais tipos de problemas envolvendo modelos de ML se dividirão em aprendizado supervisionado e aprendizado não supervisionado (SHALEV-SHWARTZ; BEN-DAVID, 2014), com o foco deste trabalho sendo os modelos supervisionados.

Em problemas de aprendizado supervisionado, a capacidade de generalização é o principal desafio teórico. Quando o modelo é excessivamente complexo, ele se ajusta aos ruídos do conjunto de treinamento, fenômeno conhecido como *overfitting* (sobreajuste). Por outro lado, modelos simples demais falham em capturar padrões relevantes, resultando em underfitting. Essa dualidade é explicada pelo trade-off entre viés e variância. A teoria da generalização introduz limites probabilísticos que quantificam o quão bem o desempenho no conjunto de treino se traduz no conjunto de teste (SHALEV-SHWARTZ; BEN-DAVID, 2014). Para evitar estas deficiências nos algoritmos, a Figura 11 a seguir ilustra os procedimentos para seleção ideal de modelos de *Machine Learning*:

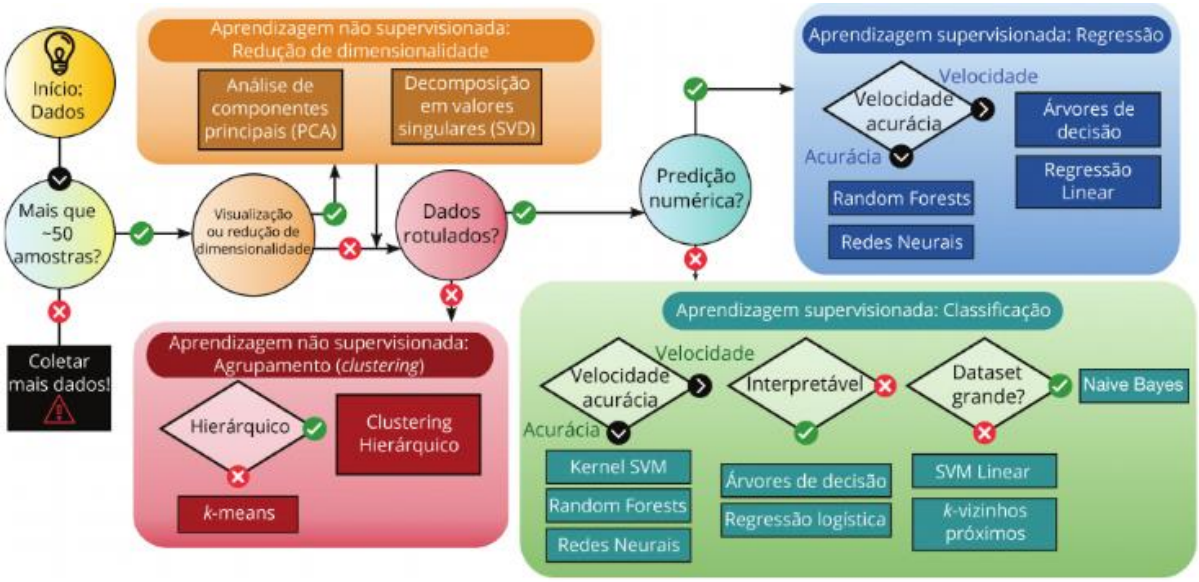


Figura 11: Fluxo para utilização de algoritmos de acordo com o tipo de aprendizado em relação ao conjunto de dados usado. Fonte: (SCHLEDER; FAZZIO, 2021).

De acordo com a Figura 11, existem diversas variáveis a se ponderar ao escolher um método de *Machine Learning*. Esse argumento é reforçado ainda mais pelo Teorema da Inexistência de Almoço Grátis, em que é provado que não existe um método de aprendizado de máquina universal, pois, para todo aprendizado, existe uma tarefa específica em que o método irá falhar, enquanto outro poderá triunfar (SHALEV-SHWARTZ; BEN-DAVID, 2014). Dessa forma, o entendimento do que deve ser analisado em relação aos dados disponíveis pode clarificar o caminho até a escolha de um algoritmo, agilizando este processo inicial existente na análise de materiais com DFT e trazendo resultados mais satisfatórios.

Um fluxo de trabalho de ML em ciência dos materiais compreende cinco etapas principais: geração e curadoria de dados em diferentes conjuntos: treinamento, validação e teste; escolha de descritores (features): propriedades do material que serão usadas para alimentar o modelo; seleção de modelo; treinamento e validação; e finalmente a avaliação preditiva. Essa estrutura é essencial para garantir reprodutibilidade e comparabilidade entre diferentes estudos. Em especial, a escolha dos descritores é um ponto crítico, pois representa a forma como o sistema físico é traduzido em um espaço matemático manipulável por algoritmos. Em ciência dos materiais, esses descritores podem incluir propriedades atômicas, como eletronegatividade e raio iônico e fatores geométricos, como parâmetros de rede (SCHLEDER et al., 2019). Nas seções a seguir, serão analisados diferentes modelos de ML empregados na predição de *band gap*, comumente regressão, e, no final, as métricas de avaliação de performance.

2.4.1 Regressões Lineares e Métodos de Regularização

Os modelos lineares constituem a base conceitual do aprendizado supervisionado. Nesses modelos, assume-se que a variável alvo y pode ser expressa como uma combinação linear das variáveis de entrada x. A forma mais simples dessa relação é dada por:

$$y = w^T x + b \tag{19}$$

onde $w^T x$ é o produto escalar entre o vetor de pesos w e o vetor de entrada x, e b é o termo de bias, ou viés (MURPHY, 2012). O treinamento busca determinar os parâmetros w e b que minimizem a função de perda, ou função de custo, comumente definida como a soma dos erros quadráticos:

$$E = \sum_i \quad (y_i - w^T x_i - b)^2 \tag{20}$$

Essa formulação é conhecida como método dos mínimos quadrados ordinários (OLS), cujo principal aspecto está na interpretabilidade, pois, cada coeficiente w_j indica a influência do atributo j sobre a propriedade alvo. Contudo, em contextos de alta dimensionalidade, como ocorre quando há muitos descritores correlacionados, o modelo OLS tende a apresentar colinearidade e *overfitting* (SHALEV-SHWARTZ; BEN-DAVID, 2014). Para mitigar esses efeitos, métodos de regularização foram desenvolvidos a fim de equilibrar o ajuste e a capacidade de generalização (SCHLEDER et al., 2020).

A regularização consiste em adicionar termos de penalização exclusivamente à função de custo, controlando a magnitude dos pesos, sem incluir o termo de viés. Os dois esquemas mais comuns são a regularização L2 (Regressão Ridge) e a regularização L1 (LASSO). No caso da Regressão Ridge, minimiza-se o erro:

$$E = \sum_i (y_i - w^T x_i)^2 + \lambda \sum_j w_j^2 \tag{21}$$

enquanto na LASSO a formulação assume a forma:

$$E = \sum_i (y_i - w^T x_i)^2 + \lambda \sum_j |w_j| \tag{22}$$

O hiperparâmetro λ, normalmente obtido na fase de validação do modelo, atua como fator de equilíbrio entre o erro de ajuste e a penalização. Para $\lambda = 0$, a função de perda retorna o resultado linear (Equação 20), enquanto para valores $\lambda >> 0$ torna os pesos pouco significativos, simplificando o modelo. A regularização L2 (Ridge) distribui o peso entre atributos correlacionados e reduz a variância do modelo, enquanto a L1 (LASSO) reduz os pesos, podendo forçá-los a zero, promovendo esparsidade e permitindo a seleção de atributos relevantes. Essa última característica é particularmente útil em problemas com muitos descritores físicos, possibilitando identificar quais propriedades atômicas ou estruturais têm maior relevância para o resultado. Ambas as formulações derivam de problemas convexos e possuem soluções analíticas ou numéricas eficientes (SHALEV-SHWARTZ; BEN-DAVID, 2014).

Do ponto de vista teórico, os modelos lineares regularizados possuem garantias formais de generalização, pois o espaço de hipóteses é limitado por restrições normativas que reduzem a complexidade do modelo. Na prática, a minimização regularizada atua como um mecanismo de controle de capacidade, impedindo que o modelo se ajuste aos ruídos do conjunto de treinamento, ou seja, para que o modelo consiga atuar sobre dados novos sem se prender ao viés dos dados de treinamento. Essa característica torna os modelos lineares regularizados uma referência fundamental para comparação com métodos não lineares mais sofisticados, discutidos nas seções seguintes (SHALEV-SHWARTZ; BEN-DAVID, 2014).

2.4.2 Métodos Baseados em *Kernel*

Os métodos de *kernel* (núcleo) expandem o conceito de regressão linear ao permitir a modelagem de relações não lineares entre as variáveis de entrada e saída. O princípio central é o mapeamento implícito dos dados de entrada x para um espaço de características de alta dimensão $\phi(x)$, no qual a relação entre as variáveis se torna aproximadamente linear. O modelo resultante é dado por:

$$f(x) = \sum_i \alpha_i \, \phi(x, \, x_i) \tag{23}$$

onde $\phi(x, x_i)$é a função *kernel*, que mede a similaridade entre amostras, e α_i é o vetor de coeficientes ajustáveis. Um exemplo de função *kernel* utilizada é a função de base radial:

$$\phi(x, \, x_i) \; = \; \exp\left(-\gamma \, ||\, x - x_i||^2\right) \tag{24}$$

Na Equação 24, γ é um parâmetro para controlar a influência de cada exemplo único do conjunto de dados do treinamento, a fim de controlar seu alcance e suavizar a modelagem das fronteiras de decisão do algoritmo (MOEINI; TEHRANI; NAEIMI-SADIGH, 2024). Essa formulação permite trabalhar com espaços de dimensão arbitrária sem computar explicitamente as transformações $\phi(x)$, utilizando a técnica do truque de *kernel*. Dessa forma, não há necessidade de conhecer todos os elementos dentro do espaço de descritores e, sim, apenas o produto interno deles, ou seja, a função *kernel*. (SHALEV-SHWARTZ; BEN-DAVID, 2014).

A *Kernel Ridge Regression* (KRR), ou Regressão de Ridge com *Kernel*, é uma das aplicações desse formalismo. Esse método é baseado tanto na regularização L2 quanto nos métodos dos mínimos quadrados ordinários. O objetivo é minimizar o erro:

$$E \; = \; ||y \; - \; \phi\alpha||^2 \; + \; \lambda\alpha^T \phi \, \alpha \tag{25}$$

O termo λ corresponde a regularização e controla a relação entre ajuste e suavidade de forma que valores pequenos permitem maior fidelidade aos dados de treino, mas aumentam o risco de *overfitting*, enquanto valores grandes promovem generalização (MURPHY, 2012).

Outro algoritmo amplamente utilizado é a Máquina de Vetores de Suporte (SVM – Support Vector Machine), que busca o hiperplano ($w^T x - b = 0$) que maximiza a margem entre classes (no caso de classificação) ou o intervalo de tolerância de erro (no caso de regressão),

potencializado pelo *kernel* para lidar com problemas não lineares. Para regressão, o algoritmo apresenta uma tolerância aos erros de previsão: o tubo ε, que, em caso de a previsão de erro estar dentro do intervalo desse tubo (erro menor que ε), o modelo não é penalizado e, caso contrário, é penalizado. O objetivo, portanto, da SVM é trazer a formulação que equilibra a minimização do erro e a regularização:

$$E = \frac{1}{2}||w||^2 + C\sum_i(\xi_i + \xi_i^*) \tag{26}$$

em que ξ_i e ξ_i^* representam as variáveis de folga e medem, respectivamente, o erro em previsões muito baixas e muitos altas. O termo C = 1/ λ define a penalização por erros acima de ε, enquanto $||w||^2$ atua como regularização (MURPHY, 2012). A Figura 12 a seguir ilustra como é o comportamento desse tubo no SVM:

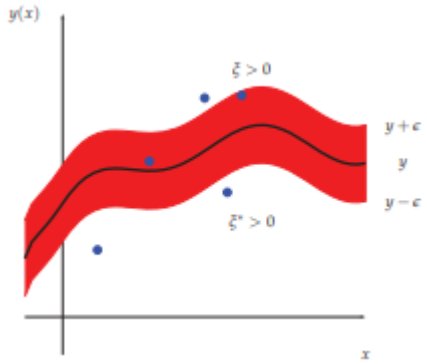


Figura 12: Comportamento do tubo ε, em que as variáveis de folga assumem valores nulos dentro dele. Fonte: (MURPHY, 2012).

Os métodos de *kernel* destacam-se por sua capacidade de capturar não linearidades complexas sem aumentar drasticamente o número de parâmetros. Entretanto, o custo computacional cresce com o número de amostras, já que é necessário calcular e armazenar uma matriz *kernel* n×n. Por isso, são mais adequados a bases de dados pequenas ou médias, nas quais oferecem excelente equilíbrio entre precisão e interpretabilidade (MURPHY, 2012).

2.4.3 Árvores de Decisão, *Random Forests* e *Boosting*

As árvores de decisão são modelos não paramétricos que particionam o espaço de atributos em regiões de decisão hierarquicamente organizadas. Cada nó representa um teste binário sobre um atributo e as divisões são escolhidas de modo a maximizar a pureza dos subconjuntos gerados, chamados de folhas, segundo métricas como o ganho de informação ou

o índice de Gini (SHALEV-SHWARTZ; BEN-DAVID, 2014). A Figura 13 exemplifica a diagramação comum de uma árvore de decisão:

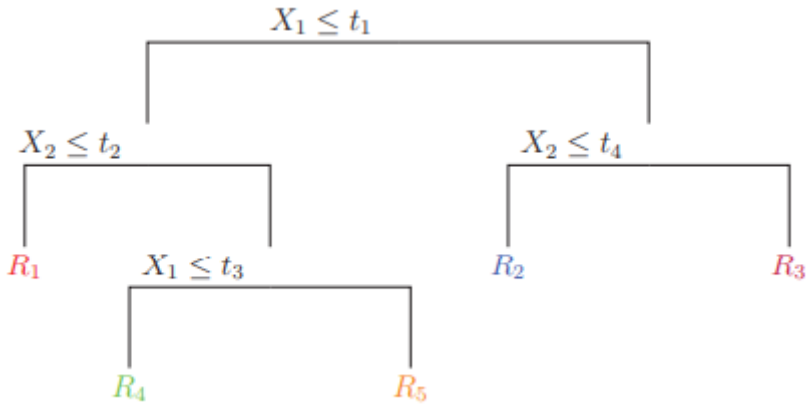


Figura 13: Esquema simples de uma árvore de regressão. Fonte: (MURPHY, 2012).

A construção é recursiva até que os dados em cada nó sejam homogêneos ou um critério de parada seja alcançado. Embora intuitivas e interpretáveis, as árvores são propensas ao *overfitting*, pois pequenas variações nos dados podem gerar estruturas significativamente diferentes. (SCHLEDER et al., 2019). Existem meios de se contornar esse problema gerado, os quais serão apresentados nos próximos parágrafos.

O método Random Forest (RF), ou Floresta Aleatória, mitiga essa limitação por meio da técnica de *bagging*, ou Agregação Bootstrap. Nesse esquema, um número M de árvores são treinadas em subconjuntos aleatórios do conjunto de dados e dos atributos, e a predição final é obtida pela média das saídas individuais, computadas como:

$$f(x) = \sum_{m=1}^{M} \frac{1}{M} f_m(x) \tag{27}$$

em que $f_m(x)$ é a saída da m-ésima árvore (MURPHY, 2012). Essa estratégia reduz a variância e melhora a estabilidade preditiva, ao custo de perda de interpretabilidade individual. O RF é amplamente utilizado na predição de propriedades de materiais, devido ao seu bom desempenho em pequenas bases de dados e robustez a correlações entre descritores (SCHLEDER et al., 2019).

Já o modelo *Gradient Boosting Machine* (GBM), traduzido como Máquina de Aumento de Gradiente, adota uma abordagem sequencial: em vez de árvores independentes, constrói-se

uma sequência de árvores fracas, em que cada nova árvore é ajustada aos resíduos do modelo anterior. O modelo final pode ser expresso como:

$$F_{m(x)} = F_{m-1(x)} + \gamma_m h(x, a_m) \tag{28}$$

onde h(x, a_m) é a nova árvore e γ_m a taxa de aprendizado (ISAYEV et al., 2017). O *boosting* minimiza uma função de perda geral L(y, f(x)) usando o gradiente descendente funcional, adaptando-se a diferentes tarefas de regressão e classificação (SHALEV-SHWARTZ; BEN-DAVID, 2014).

Tanto o RF quanto o GBM são fundamentados no princípio do ensemble learning, segundo o qual múltiplos estimadores combinados reduzem o erro de generalização. Essa combinação de modelos diversificados permite alcançar desempenho superior a qualquer modelo isolado, reduzindo a variância sem aumento de viés, tornando-se uma propriedade valiosa em bases de dados com ruído experimental ou incerteza computacional (SCHLEDER et al., 2019).

2.4.4 Redes Neurais Artificiais

As redes neurais artificiais, do inglês Artificial Neural Networks (ANNs), representam uma das classes mais flexíveis e poderosas de algoritmos supervisionados. São compostas por camadas de unidades de processamento (neurônios) que realizam transformações lineares seguidas por funções de ativação não lineares. O formalismo teórico envolve a presença de grafos G(V, E), o peso w sobre os limites E e a presença de nós dos grafos, análogos a neurônios. Cada neurônio disposto na rede neural consome a informação de neurônios anteriores de maneira linear, análogo ao que foi visto na seção 2.4.1 deste trabalho, por meio da soma ponderada das saídas com os pesos, e de maneira não linear, expressa pela relação:

$$a = \sigma(z) \tag{29}$$

em que σ é uma função de ativação, como a função sinal σ(a) = sign(a) ou a função sigmoide σ(a) = 1/(1+exp(-a)). Considerando a entrada $a_{t+1,j}$ para o vetor final de saída $v_{t+1,j}$, o comportamento da rede neural é matematicamente descrito como:

$$a_{t+1,j}(x) = \sum_{r:\,(v_{t,r}, v_{t+1,j}) \in R} w\left((v_{t,r}, v_{t+1,j})\right) o_{t,r}(x) \tag{30}$$

em que $v_{t,r}$ é o neurônio na camada t e $o_{t,r}(x)$o seu output de acordo com a entrada da camada inicial. Essencialmente, na camada V_0, estarão presentes os dados de entrada x do sistema. Subsequentemente, tem-se as camadas escondidas, em que ocorrerá o processamento em cada neurônio de acordo com a Equação 30, resultando na saídal final presente na camada V_t (SHALEV-SHWARTZ; BEN-DAVID, 2014). Essa esquematização é ilustra na Figura 14 a seguir:

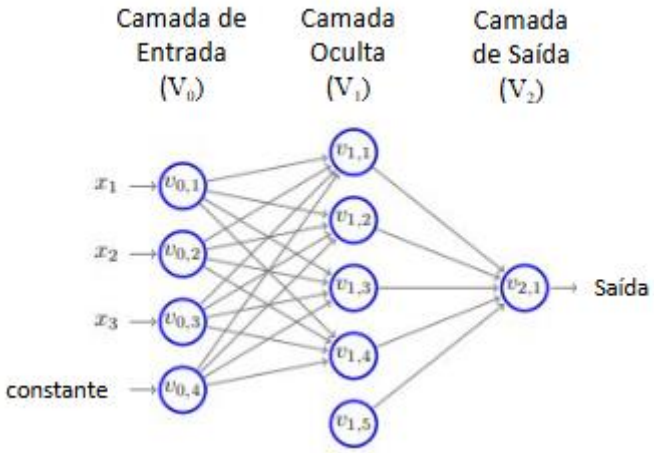


Figura 14: Esquema de funcionamento de uma rede neural simples. Fonte: Adaptado de (SHALEV-SHWARTZ; BEN-DAVID, 2014).

O Teorema da Aproximação Universal demonstra o poder de representação das redes neurais, garantindo que uma rede com apenas uma camada oculta pode aproximar qualquer função contínua em um domínio compacto, com erro arbitrariamente pequeno. Essa propriedade assegura o potencial teórico das redes para modelar relações extremamente complexas. Contudo, a capacidade de generalização depende do equilíbrio entre número de parâmetros e volume de dados disponíveis. Dessa forma, ela não garante que o treinamento será bem-sucedido com as escolhas corretas de pesos e vieses, além de também não explicitar a quantidade de neurônios necessária para realizar a aproximação, que, em caso de funções complexas, pode exigir uma grande alocação de recursos voltados a construção de muitos neurônios em uma só camada oculta (AUGUSTINE, 2024).

2.4.5 Métricas de Avaliação em Modelos de *Machine Learning*

A avaliação quantitativa de modelos de aprendizado de máquina é essencial para determinar não apenas a acurácia, mas também a robustez e a capacidade de generalização. Em problemas comuns de regressão, como a predição de *band gap*s, as métricas mais utilizadas são

o Erro Absoluto Médio (MAE), a Raiz do Erro Quadrático Médio (RMSE) e o Coeficiente de Determinação (R^2) (SCHLEDER et al., 2020), apresentadas em sequência:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{31}$$

No MAE, n representa o número total de amostras do conjunto de dados; y_i é o valor real (observado ou calculado por métodos de referência, como DFT), e $\hat{y}_i$ o valor previsto pelo modelo. O termo $|y_i - \hat{y}_i|$ indica o erro absoluto de cada previsão individual. Assim, o MAE expressa o erro médio absoluto cometido pelo modelo, fornecendo uma noção direta de quão distante, em média, as predições estão dos valores reais, sem ser fortemente prejudicado pela presença de outliers, ou seja, dados que desviam significativamente do padrão (SCHLEDER et al., 2020). Em seguida, discutimos o RMSE, obtido pela raiz quadrada do erro quadrático médio (MSE):

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{32}$$

O RMSE, por sua vez, enfatiza os erros de maior magnitude, pois eleva os desvios ao quadrado antes de calcular a média. Isso faz com que ele seja mais sensível a predições muito incorretas, funcionando como uma medida de penalização para erros grandes. Em termos físicos, o RMSE pode ser interpretado como a raiz do erro quadrático médio de energia, representando a dispersão das previsões em torno dos valores experimentais ou teóricos de referência. Em geral, um RMSE baixo indica que o modelo é consistente e pouco sujeito a flutuações aleatórias. Enfim, temos o coeficiente de determinação R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{33}$$

Na Equação 33, $\bar{y}$ remete à média dos valores reais presentes no conjunto de dados. O R^2 indica a performance do modelo em captar a adequação da predição em relação aos dados coletados. Quanto mais próximos os valores do coeficiente estão de 1, em teoria, mais o modelo demonstra a habilidade de descrever bem os resultados esperados, porém não garante que todos os desvios sistemáticos e outliers estejam devidamente representados. Além disso, o valor de R^2 depende da variabilidade dos dados reais, pois, se essa variância for reduzida, mesmo previsões com erro moderado podem resultar num R^2 aparente alto. Assim, R^2 deve ser

interpretado em conjunto com métricas como MAE e RMSE, bem como com análise de resíduos, para assegurar que a generalização do modelo é realmente robusta e fisicamente significativa (KHOSHVAGHT et al., 2025).

Em geral, essas três métricas são as mais populares para avaliação de modelos de *Machine Learning* e é ideal que sejam utilizadas em conjunto para construir uma melhor análise dos dados que traga a visão de ajuste entre modelo preditivo e dados calculados. O MAE reflete o erro médio absoluto e é útil para interpretação direta dos resultados, o RMSE identifica a presença de grandes desvios, e o R^2 avalia o grau de correlação estrutural entre predições e valores reais. Em conjunto, elas permitem caracterizar não apenas a precisão, mas também a estabilidade e confiabilidade estatística do modelo (KHOSHVAGHT et al., 2025).

3 DISCUSSÃO

A predição do *band gap* em perovskitas, fundamentada nos cálculos fornecidos pela DFT e potencializada por modelos preditivos baseados em *Machine Learning*, constitui um campo de investigação de grande valor para a ciência dos materiais. Existem diversos aspectos que contribuem, durante a síntese de um material, para alcançar as propriedades desejadas, como fatores estruturais e condições de ambiente exercidas, como temperatura e pressão. Dessa forma, a procura contínua pelos melhores métodos e materiais disponíveis é fomentada pela inovação do ramo industrial para se atingir novos patamares de eficiência (CALLISTER, 2016). Nesta seção, será discutida a importância de se prever o *band gap* de um material, o diferencial de perovskitas em potencial de aplicações de dispositivos optoeletrônicos e, por fim, a busca por utilização de *Machine Learning* como complemento à DFT para obter resultados mais satisfatórios.

3.1 A IMPORTÂNCIA DA PREDIÇÃO DO *BAND GAP* EM MATERIAIS

A predição do *band gap* de um material, além de nos dizer sua natureza, conforme discutido na seção 2.3.1, é uma etapa fundamental para descobrir seu comportamento em aplicações optoeletrônicas. Na área fotovoltaica, o valor ideal para células solares de junção única é de 1,34 eV, seja ele direto ou indireto. Em estudos ópticos, obter um *band gap* ideal contribui, com base no espectro de luz, à característica a qual se deseja obter em fenômenos de absorção (fotodetectores) ou emissão (lasers e LEDs) (LI; YANG, 2020). Dessa forma, a predição do *band gap* em um material se torna o direcionamento para sua posterior sintetização.

Além disso, conhecer o *band gap* constitui uma etapa fundamental do processo de triagem de materiais, principalmente em substâncias ainda não fabricadas. Prever o *band gap*, via cálculos de DFT e/ou modelos de ML, pode poupar esforços e oferecer um melhor direcionamento para a síntese de um composto (GAO et al., 2021). Caso a base de dados que alimenta o modelo de ML estiver bem alinhada com os atributos do material, o resultado obtido pode trazer uma maior confiança e, em alguns casos, ser tão eficiente quanto cálculos de funcionais modernos da DFT, a um custo e esforço relativamente menor (SCHLEDER et al., 2019). Entretanto, conforme o fluxo da Figura 1 da seção de Introdução ilustra, ainda é recomendável a realização de validações experimentais após síntese do material, pois há a possibilidade de ocorrer defeitos e impurezas na estrutura, por exemplo, das perovskitas durante o processo de fabricação, afetando sua performance (BERA et al., 2022).

3.2 POTENCIAL DE APLICAÇÃO OPTOELETRÔNICA DAS PEROVSKITAS

Um dos principais pontos positivos sobre o potencial de aplicação optoeletrônica de perovskitas é sua versatilidade de capabilidades dentro desse ramo, devido a um importante fator estratégico já discutido: a flexibilidade de seu *band gap*. Existem diversas maneiras de realizar esse ajuste, conforme a seção 2.2.2, abordando os métodos de engenharia composicional, dimensionalidade e pressão. Os dois primeiros casos vêm apresentando resultados promissores, principalmente a engenharia composicional, em que, pela simples manipulação de elementos presentes na estrutura da perovskita, consegue promover um caminho para desenvolver materiais especializados para as necessidades das PSCs. Entretanto, a modulação de *band gap* via pressão hidrostática, apesar de efetiva, deve ser analisada com cautela pois não demonstra viabilidade prática ao necessitar de um ambiente controlado para funcionamento do dispositivo, além de, para determinadas condições, ocorrer amorfização do material, resultando na perda de propriedades desejadas (MIAH et al., 2024). Ainda assim, existe um desafio em ajustar o *band gap* para um valor ideal abaixo de 1,48 eV, dentro dos limites de Shockley-Quessier para células solares, pois as substituições empregadas na engenharia composicional atual não são capazes de alcançar esse resultado. Apesar disso, a utilização de outros compostos, como o latão, no preparo do material conseguiu minimizar essa energia (AFROZ et al., 2025). No contexto tecnológico, conforme observado inicialmente pela Tabela 1, as perovskitas têm sido exploradas em LEDs, lasers, fotodetectores, emissores e dispositivos neuromórficos, destacando seu potencial para diferentes campos da eletrônica e fotônica.

A ajustabilidade de seu *band gap* também permite as perovskitas serem utilizadas como materiais isolantes ou dielétricos: certas perovskitas multiferroicas ou ferroelétricas exibem alta rigidez dielétrica, o que as torna úteis como isolantes em capacitores, dispositivos piezoelétricos ou em aplicações onde a estabilidade elétrica e a resposta a campos externos são críticas. Apesar desse leque de aplicações, a principal linha de investigação recai sobre as células solares de perovskita (PSCs), devido à sua altíssima eficiência, ao baixo custo de fabricação e à facilidade de processamento, fatores que as colocam como candidatas promissoras para a próxima geração de fotovoltaicos (ZHANG et al., 2023). A performance das PSCs pode ser comparada com outros materiais, como registrado na Tabela 2:

Tabela 2: Dados coletados de células solares de diferentes materiais.

Material	*Band gap* (eV)	V_{oc} (V)	J_{sc} (mA/cm^2)	FF (%)	PCE (%)	PCE^{SQ} (%)	PCE/PCE^{SQ} (%)
Perovskita	1,53	1,19	26,00	84,0	26,0	31,34	83,0
GaAs	1,43	1,13	29,78	86,7	29,1	32,54	89,4
c–Si	1,11	0,74	42,65	84,9	26,7	32,55	82,2

Fonte: (ZHANG et al, 2023).

A Tabela 2 nos indica vários dados relevantes sobre as células fotovoltaicas, dentre eles: V_{oc}, a tensão de circuito aberto; J_{sc}, a densidade de corrente de curto-circuito (quanto maior, melhor a captação de fótons); FF, o fator de preenchimento (fração da área máxima operacional da célula); PCE, a eficiência de conversão de potência e PCE^{SQ} a eficiência teórica máxima esperada de acordo com o material e limite de Shockley-Quessier, por volta da faixa de 1,3 eV (RÜHLE, 2016). Os valores coletados mostram que, em geral, a performance das perovskitas é comparável com a de outros materiais, inclusive o já consolidado no mercado, silício cristalino. Entretanto, ainda há desafios para a comercialização em larga escala, como questões de estabilidade estrutural e a toxicidade de chumbo presente em alguns tipos (ZHANG et al., 2023). Apesar disso, os dados evidenciam que as perovskitas, embora os estudos com células solares serem relativamente recentes, possuem a capacidade de ser uma opção no futuro mais barata e que possa reproduzir eficiência similar.

Além disso, as PSCs também possuem o potencial de ser uma das tecnologias fotovoltaicas mais baratas do mundo. A manufatura de módulos fotovoltaicos de perovskita apresenta vantagens estruturais em relação às tecnologias convencionais, especialmente pela combinação de deposição em baixa temperatura, alto coeficiente de absorção óptica, o que permite filmes ativos ultrafinos, e compatibilidade com processos de elevada taxa de produção. Esses fatores reduzem significativamente o consumo energético, a complexidade operacional e o capital necessário para implantação industrial, resultando em um custo direto de US$ 31,7/m², Preço Mínimo Sustentável (MSP) de US$ 0,41/Wp para módulos de 16% de eficiência e custo nivelado de energia (LCOE) variando entre 4,93 e 7,90 ¢/kWh, valores inferiores aos reportados para materiais comuns dos módulos fotovoltaicos como CdTe, CIGS e silício cristalino (c-Si) (SONG et al., 2017). Portanto, a rota de manufatura baseada em perovskitas não apenas reduz custos, mas reconfigura o paradigma produtivo do setor fotovoltaico, consolidando-se como uma candidata promissora para expansão sustentável e economicamente viável da tecnologia solar.

3.3 *MACHINE LEARNING* COMO COMPLEMENTO À DFT

Toda finalidade de se prever o *band gap* de uma perovskita só é possível de ser efetivamente operacionalizada a partir do advento de duas técnicas: os cálculos produzidos pela DFT e a potencialização deles via *Machine Learning*. Como discutido anteriormente, a DFT é uma ciência com base teórica no estado estacionário, ou seja, não é capaz de obter respostas precisas sobre a densidade eletrônica em estados de excitação. Técnicas para contornar esse problema até existem, porém, possuem um custo computacional que cresce proporcionalmente ao cubo do tamanho do sistema à medida que, em sistemas com mais de 1000 átomos, torna-se inviável de realizar o processamento (SCHLEDER et al., 2019). Logo, os algoritmos de *Machine Learning*, que em geral são bem menos custosos, entram como alternativa para realizar novos cálculos e predições em cima da base de dados já consolidada de perovskitas "The Perovskite Database Project", que segue os princípios FAIR: encontrável, acessível, interoperacional e reutilizável (JACOBSSON et al., 2022).

A seleção de descritores (features) constitui um passo fundamental no desenvolvimento de modelos de *Machine Learning* para previsão de propriedades eletrônicas de perovskitas, uma vez que determina a quantidade e a qualidade da informação disponível para o modelo. Em sistemas cristalinos como as perovskitas ABX_3 e suas variantes estruturais, descritores físico-químicos derivados dos elementos constituintes, como eletronegatividade, raio iônico, afinidade eletrônica e energia de ionização, frequentemente apresentam correlação direta com a estrutura eletrônica, contribuindo para a redução da dimensionalidade do problema e evitando redundâncias estatísticas (TALAPATRA et al., 2023). Além disso, a escolha adequada de features permite minimizar o risco de erros e prevenir *overfitting* (SHALEV-SHWARTZ; BEN-DAVID, 2014)., especialmente em bases de dados de tamanho moderado, onde muitos descritores podem capturar ruído ao invés de padrões físico-químicos relevantes. Dessa forma, a seleção criteriosa de descritores não só aprimora o desempenho preditivo dos modelos, mas também contribui para conectar a inferência estatística às relações físico-químicas que governam o comportamento das perovskitas.

Prosseguindo, os resultados já calculados da DFT representam, portanto, a etapa inicial do fluxo de emprego de *Machine Learning*, com método selecionado de acordo com a validação necessária, para a predição do *band gap*. A partir deles, é possível treinar o modelo em uma amostra inicial e depois, em um conjunto de dados separado, realizar o teste de performance preditiva do modelo e avaliar sua performance. A etapa mais custosa de todo esse processo é o

passo inicial da busca em conhecer o melhor algoritmo para atuar em cima do conjunto de dados selecionado. Após essa definição, as próximas fases são mais sublimes e, em caso de atualizações na base, basta retreinar o modelo. Por fim, é importante ressaltar que a avaliação de performance, devido à heterogeneidade dos dados, dos métodos de ML não traz uma resposta consensual sobre qual deles é a melhor ferramenta para prever propriedades, entretanto, para o *band gap*, podemos analisar os resultados obtidos em diferentes pesquisas e entender qual deles indica melhor desempenho (SCHLEDER et al, 2019). Nos parágrafos a seguir, a análise será feita com base em pesquisas recentes com aplicações de diversos modelos para um mesmo conjunto de dados: perovskitas ABX_3, perovskitas ABX_3 com discrição de *band gap*s diretos e indiretos. perovskitas duplas e 2D, e perovskitas de nitreto (ABN_3).

3.3.1 Predição de *Band gap* em Perovskitas ABX_3

A previsão de *band gap* em perovskitas foi realizada empregando seis modelos de *Machine Learning* diferentes, treinados em um conjunto de dados de 500 materiais perovskitas, como, por exemplo, $MAPbI_3$, $FAPbI_3$ e $CsSnI_3$, e testados em um subconjunto separado de 50 dados, indicou resultados de performance conforme a Tabela 3:

Tabela 3: Métricas de performance em perovskitas ABX_3.

Modelo de ML	Erro absoluto médio (MAE)	Erro quadrático médio (MSE)	Coeficiente de determinação (R^2)
Random forest	0.000775	0.00000920	0.9994
Gradient boosting regressor	0.0222	0.0041	0.8841
k-Nearest neighbors (KNN)	0.0365	0.0143	0.5914
AdaBoost	0.0283	0.0029	0.9163
Gaussian process regressor	0.0351	0.0082	0.7662
Bagging	0.0448	0.0162	0.6673

Fonte: Adaptado de (SAMANTARAY; SINGH; ANU TONK, 2024)

Entre os modelos testados, o Random Forest apresentou o melhor desempenho, com erros médios muito baixos e coeficientes de determinação próximos de 1. Métodos análogos ao *Boosting*, como *Gradient Boosting Regressor* (Regressor de Aumento de Gradiente) e *AdaBoost* também conquistaram o pódio de performance. Modelos mais simplórios, como o *k-Nearest neighbors* (k-vizinhos mais próximos) (MURPHY, 2012), apresentaram erros médios maiores e, perceptivamente, uma menor capabilidade preditiva generalizada, indicada pelo relativamente baixo coeficiente R^2. Embora esses resultados indiquem forte capacidade

preditiva para o domínio do conjunto de dados envolvido, é necessário cautela na conclusão pois, o conjunto de dados e descritores para este estudo ainda podem ser refinados para obter resultados mais acurados. Dessa forma, apesar da boa performance numérica, a robustez prática do modelo depende da validação sobre bases externas e mais diversificadas (SAMANTARAY; SINGH; ANU TONK, 2024).

3.3.2 Predição de *Band gap* em Perovskitas ABX_3 com Discrição de *Band gap*

O estudo envolvendo as perovskitas ABX_3 com discrição de *band gap* empregou seis modelos de ML para prever tanto o *band gap* direto quanto o indireto em uma base de dados de aproximadamente 200 compostos obtidos pelo funcional de DFT HSE06. A seleção de descritores foi utilizada com base na SHAP - Shapley Additive Explanation, uma metodologia interpretativa de ML que explica a saída de um modelo através da designação de um valor que representa sua contribuição com a predição (SHAP, 2018) e, por meio de uma série de cálculos, definiram como descritores mais impactantes as eletronegatividades dos elementos B, X e o raio covalente de X (OBADA et al., 2023). Na Tabela 4 a seguir, estão listados os valores obtidos para a performance dos modelos usados, separados na proporção 80-20 os subconjuntos de treinamento e teste, respectivamente:

Tabela 4: Métricas de performance para predição de *band gap* direto.

Modelos de ML	Treinamento (80%)			Teste (20%)		
	MAE	RMSE	R^2	MAE	RMSE	R^2
CatBoost	0.124 ± 0.003	0.166 ± 0.004	0.996 ± 3.656×10^{-4}	0.845 ± 0.094	1.390 ± 0.186	0.697 ± 0.099
XGBoost	0.003 ± 0.001	0.004 ± 0.001	0.999 ± 1.264×10^{-6}	0.810 ± 0.126	1.460 ± 0.177	0.664 ± 0.116
Random Forest	0.342 ± 0.011	0.507 ± 0.020	0.959 ± 0.004	0.963 ± 0.175	1.450 ± 0.264	0.665 ± 0.150
LightGBM (boosting)	0.720 ± 0.025	0.991 ± 0.066	0.843 ± 0.021	1.130 ± 0.106	1.610 ± 0.170	0.599 ± 0.101
CompoundNET (rede neural)	0.065 ± 0.027	0.129 ± 0.050	0.997 ± 0.002	0.991 ± 0.124	1.680 ± 0.252	0.557 ± 0.141
Árvore de Decisão	0.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	0.993 ± 0.142	1.790 ± 0.245	0.505 ± 0.124

Fonte: Adaptado de (OBADA et al., 2023).

Para este caso, métodos análogos ao *Boosting*, como *CatBoost* e *XGBoost* (*eXtreme Gradiente Boosting*), e *Random Forest* apresentaram erros relativamente menores e um melhor coeficiente de determinação, embora exista a possibilidade de estar ocorrendo um *overfitting* devido aos resultados de teste para todos os modelos serem bem discrepantes do que foi alcançado em treinamento. Uma das causas para este problema pode estar associada ao limitado número de amostras presentes na base de dados, que pode penalizar a capacidade generalizadora do algoritmo através da falta de informações dos descritores (SHALEV-SHWARTZ; BEN-DAVID, 2014). Esse fenômeno fica ainda mais perceptível analisando o desempenho da Árvore de Decisão, que durante o treinamento não apresentou erros e obteve um coeficiente de determinação perfeito, porém, ao ser testada, ficou enviesada nos ruídos e padrões dos dados de treinamento e não conseguiu reproduzir uma boa performance de predição. A seguir, são apresentados os dados de performance para predição de *band gap* indireto, sob mesmas condições, na Tabela 5:

Tabela 5: Métricas de performance para predição de *band gap* indireto.

Modelos de ML	Treinamento (80%)			Teste (20%)		
	MAE	RMSE	R^2	MAE	RMSE	R^2
CatBoost	0.116 ± 0.006	0.155 ± 0.008	0.996 ± 4.873×10^{-4}	0.795 ± 0.103	1.301 ± 0.226	0.699 ± 0.125
Random Forest	0.329 ± 0.012	0.495 ± 0.025	0.957 ± 0.005	0.909 ± 0.182	1.386 ± 0.310	0.650 ± 0.189
XGBoost	0.003 ± 0.006	0.004 ± 0.001	0.999 ± 2.006×10^{-6}	0.791 ± 0.127	1.397 ± 0.276	0.647 ± 0.178
LightGBM (boosting)	0.687 ± 0.029	0.965 ± 0.073	0.835 ± 0.025	1.073 ± 0.109	1.529 ± 0.192	0.591 ± 0.131
CompoundNET (rede neural)	0.053 ± 0.018	0.087 ± 0.037	0.998 ± 0.001	0.879 ± 0.146	1.521 ± 0.366	0.576 ± 0.237
Árvore de Decisão	0.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	0.812 ± 0.170	1.574 ± 0.365	0.557 ± 0.195

Fonte: Adaptado de (OBADA et al., 2023).

Comparando-se as Tabelas 4 e 5, percebe-se que os resultados entre elas não foram discrepantes, indicando que, no escopo dos descritores selecionados, a natureza do *band gap* não foi um fator impactante para a acurácia de sua predição. O estudo original não deixa claro uma justificativa para este acontecimento, entretanto, uma possível explicação para isso provém de o comportamento de alguns descritores relacionados às características estruturais do composto não serem capazes de representar detalhadamente propriedades composicionais, como, por exemplo, simetria (PILANIA et al., 2016). Outra análise que ambos os dados proporcionam é a validação de que o modelo *CompoundNET*, baseado em redes neurais, foi superior a um modelo considerado mais simples, como as Árvores de Decisão, mas ainda relativamente inferior aos modelos baseados em *Boosting* neste caso (OBADA et al., 2023). Por fim, o funcional HSE06 utilizado para obtenção dos dados de descritores é bom para o cálculo de *band gap*, mas pode encontrar desafios ao descrever alguns materiais em que ocorre o cruzamento entre *band gap*s direto e indireto (JAYAN; SEBASTIAN, 2019).

3.3.3 Predição de *Band gap* em Perovskitas Duplas e 2D

Neste estudo, a base de dados consiste em descritores de 1493 amostras de perovskitas duplas e 491 amostras de perovskitas 2D, extraídas do Repositório de Material Computacional (CMR), empregando os modelos de regressão de SVR (*Support Vector Regression*, derivado das SVM), RFR (*Random Forest Regressor*), GBR (*Gradient Boost Regressor*) e *XGBoost* sobre um conjunto de 40 descritores determinados via SHAP, em que o desvio padrão e a média de cargas na camada de valência foram verificados como os mais impactantes. Os funcionais de DFT selecionados foram os mais modernos, como o GLLB-SC (*Gritsenko, van Leeuwen, van Lenthe, and Baerends - Solid and Correlation*) e aproximação GW (MOEINI; TEHRANI; NAEIMI-SADIGH, 2024). Na Tabela 6 a seguir, são apresentados os dados de performance dos modelos para ambas as categorias de perovskitas:

Tabela 6: Métricas de performance em perovskitas duplas e 2D.

Bandgap	Modelo (tipo de perovskita)	R^2 treinamento (%)	R^2 teste (%)	MSE (eV)	MAE (eV)	Proporção teste/treinamento (%)
	SVR (dupla)	99	96	0.10	0.19	10/90

Direto	RFR (dupla)	98	93	0.16	0.29	10/90
	GBR (dupla)	99	94	0.14	0.23	10/90
	XGBoost (dupla)	99	96	0.10	0.22	5/95
	SVR (2D)	98	96	0.10	0.24	5/95
	RFR (2D)	97	80	0.53	0.49	10/90
	GBR (2D)	99	86	0.39	0.39	10/90
	XGBoost (2D)	99	85	0.40	0.45	15/85
Indireto	SVR (dupla)	99	96	0.10	0.19	10/90
	RFR (dupla)	98	93	0.17	0.29	10/90
	GBR (dupla)	98	95	0.12	0.26	10/90
	XGBoost (dupla)	99	95	0.11	0.23	5/95
	SVR (2D)	98	95	0.13	0.25	5/95
	RFR (2D)	97	82	0.45	0.48	10/90
	GBR (2D)	99	91	0.28	0.35	5/95
	XGBoost (2D)	99	90	0.29	0.39	5/95

Fonte: Adaptado de (MOEINI; TEHRANI; NAEIMI-SADIGH, 2024).

Dentre os algoritmos, o SVR apresentou melhor desempenho geral, tanto para *band gap* direto e indireto, quanto para perovskitas duplas ou 2D, apresentando valores mínimos de erro e um bom coeficiente de determinação. Em seguida, *XGBoost*, GBR e RFR também performaram muito bem, mas trouxeram erros mais sistemáticos ao preverem o *band gap* de

perovskitas 2D. A escolha de descritores é uma etapa importante da metodologia empregada, pois perovskitas de baixa simetria exibem variabilidade eletrônica maior para composições aparentemente semelhantes, de modo que modelos puramente composicionais tendem a perder parte do sinal se não se incluir informação que reflita distorções e ordenamento estrutural. Nesse contexto, a estratégia adotada, separar os conjuntos (dupla e 2D), empregar descritores estatísticos (variância e média) e testar inclusão de indicadores eletrônicos adicionais, como a descontinuidade derivada para perovskitas duplas, melhora a previsibilidade e reduz variância de predição em materiais estruturalmente complexos (MOEINI; TEHRANI; NAEIMI-SADIGH, 2024).

3.3.4 Predição de *Band gap* em Perovskitas de Nitreto

As perovskitas de nitreto ABN_3 foram analisadas utilizando um conjunto de dados amplo, composto por 1563 estruturas obtidas por DFT no rigor teórico dos funcionais PBE, HSE06 e aproximação GW e relacionados a um número total de 117 descritores. Os modelos de ML empregados foram MLP (*Multi-Layer Perceptron*, uma espécie de rede neural artificial), SVR, *Random Forest* e *Gradient Boosting* (GBDT), baseados em descritores de alta importância como eletronegatividade, elétrons de valência no orbital "d" e energia de formação (GHOSH; CHOWDHURY, 2024). Na Tabela 7, temos a avaliação de performance dos algoritmos:

Tabela 7: Métricas de performance para predição de *band gap* em perovskitas de nitreto.

Modelos de ML	**MAE**	**RMSE**	**R^2**
MLP	0.16	0.22	0.74
GBDT	0.10	0.15	0.90
SVR	0.08	0.13	0.91
RFR	0.03	0.11	0.94

Fonte: Adaptado de (GHOSH; CHOWDHURY, 2024).

A rede neural MLP não conseguiu reproduzir um bom resultado devido ao desempenho ruim durante a etapa de treinamento, apresentando muita variância, como é ilustrado na Figura 15:

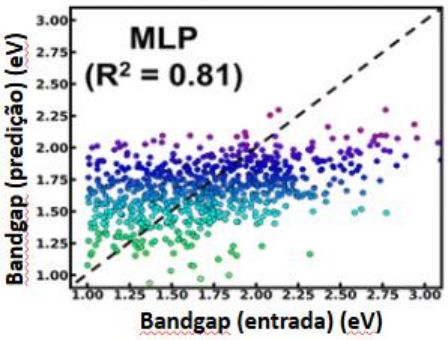


Figura 15: Performance do modelo MLP durante o treinamento. Fonte: Adaptado de (GHOSH; CHOWDHURY, 2024).

Em contraste, as melhores performances gerais foram perceptíveis nos modelos GBDT, SVR e RFR, com o método de Random Forest sendo o que teve a melhor capacidade preditiva apresentado erros sistemáticos mínimos. Essa performance geral pode ser salientada pelo cuidado e rigor na seleção de descritores e no grande volume de dados usado (GHOSH; CHOWDHURY, 2024).

Em síntese, todos os estudos nesse âmbito indicaram que os melhores modelos para predição de *band gap* em perovskitas foram em geral os derivados do método de *Boosting* e *Random Forest*, além de um destaque para o modelo SVR que demonstrou resultados excelentes em condições de alta complexidade envolvida, como no caso das perovskitas duplas e 2D. Os algoritmos de redes neurais *CompoundNET* e MLP não apresentaram resultados tão satisfatórios quanto os seus "concorrentes", o que pode ser explicado pelo tamanho relativamente pequeno do conjunto de dados empregado nas pesquisas (PALIWAL; KUMAR, 2011). Os resultados positivos reforçam que, apesar de bases de dados limitadas e sistemas estruturalmente diversos, os modelos de aprendizado se mostram capazes de lidar bem com variabilidade e evitar *overfitting*, podendo oferecer ainda mais robustez na análise caso ocorra um melhor preparo dos conjuntos coletados e reprodutibilidade, além de permitir que outros modelos, como redes neurais, alcancem resultados mais exemplares (SAMANTARAY; SINGH; ANU TONK, 2024). Portanto, estão enaltecidos os principais modelos que podem contribuir para o futuro e desbloqueio do potencial inerte das perovskitas através de predições acuradas de seu comportamento optoeletrônico.

4 CONCLUSÃO

Neste trabalho, foram abordados teoricamente temas fundamentais de Estado Sólido, Teoria do Funcional da Densidade (DFT) e *Machine Learning*. A premissa de verificar a importância de se prever o *band gap* é verdadeira, pois, como discutido, é uma etapa essencial para a triagem dos materiais. Tal afirmação é reforçada pelo potencial das perovskitas, que mostraram em aplicações como células fotovoltaicas a capacidade de reproduzir uma eficiência similar ao de líderes do mercado, como o silício cristalino, a um menor custo de sintetização. Dessa forma, prever o *band gap* de perovskitas é fundamental para agilizar metodologias de pesquisa com esse composto, em busca de compreender seu comportamento optoeletrônico e refinar seus atributos.

O entendimento de DFT mostrou como é possível descrever e calcular teoricamente as características estruturais da matéria sem se limitar a necessidade de métodos de caracterização experimental. Por meio de uma linha do tempo evolutiva, os funcionais teóricos aperfeiçoaram seus termos de aproximação da energia de troca e correlação a fim de trazer mais precisão aos cálculos. Entretanto, a DFT convencional falha como teoria exata por descrever apenas o estado estacionário e, embora métodos mais modernos consigam trazer o comportamento em estados excitados, eles requerem um custo computacional muito alto e são inviáveis de serem realizados para grandes escopos de sistemas atômicos.

Dessa forma, os algoritmos de *Machine Learning* visam mitigar essa deficiência, prevendo o *band gap* através de base de dados já consolidadas com cálculos de DFT. Os resultados de diferentes pesquisas apontam que os modelos de *Random Forest*, *Support Vector Regression* e derivados de *Boosting* apresentam as melhores métricas de performance para conseguir reproduzir um *band gap* acurado, indicando que, para as condições impostas durante estes estudos, eles foram capazes de descrever o comportamento do sistema corretamente. Ainda assim, em concordância com o Teorema da Inexistência de Almoço Grátis, essa análise não pode ser totalmente verificada de maneira universal, pois, cada pesquisa está sujeita a uma base de dados específica selecionada.

Como perspectivas futuras, deve-se fomentar a contínua construção de bases de dados mais robustas para melhor interpretação dos modelos de *Machine Learning* e entendimento dos descritores envolvidos. Essa visão está de acordo com o objetivo de aplicações envolvendo perovskitas, porque há a necessidade de se realizar mais estudos para verificação de estabilidade desses compostos, além de engenharia composicional para evitar a fabricação de compostos

tóxicos ao ambiente contendo a presença de chumbo. Logo, entende-se que a integração entre DFT e *Machine Learning* é essencial para o fluxo de estudo de materiais.

REFERÊNCIAS

AFROZ, M. et al. Perovskite solar cells: Progress, challenges, and future avenues to clean energy. **Solar Energy**, v. 287, p. 113205–113205, 25 dez. 2024.

ASHCROFT, N. W.; N. DAVID MERMIN. **Solid State Physics**. [s.l.] Cengage Learning, 1976.

AUGUSTINE, M. T. **A Survey on Universal Approximation Theorems**. 2024. Disponível em: https://arxiv.org/html/2407.12895v1. Acesso em: 15 nov. 2025.

BERA, S. et al. Review of defect engineering in perovskites for photovoltaic application. **Materials Advances**, v. 3, n. 13, p. 5234–5247, 4 jul. 2022.

CALLISTER, William D., Jr.; RETHWISCH, David G. **Ciência e engenharia de materiais: uma introdução**. Tradução de Sergio Murilo Stamile Soares. 9. ed. Rio de Janeiro: LTC, 2016. ISBN 978-85-216-3103-3.

CAPELLE, K. **A bird's-eye view of density-functional theory**. 2006. Disponível em: https://arxiv.org/abs/cond-mat/0211443. Acesso em: 15 nov. 2025.

GAO, L.; HU, P.; LIU, S. Low-dimensional perovskite modified 3D structures for higher-performance solar cells. **Journal of Energy Chemistry**, v. 81, p. 389–403, 16 fev. 2023.

GAO, Z. et al. Screening for lead-free inorganic double perovskites with suitable band gaps and high stability using combined machine learning and DFT calculation. **Applied Surface Science**, v. 568, p. 150916, dez. 2021.

GHOSH, S.; CHOWDHURY, J. Predicting band gaps of ABN3 perovskites: an account from machine learning and first-principle DFT studies. **RSC Advances**, v. 14, n. 9, p. 6385–6397, 2024.

ISAYEV, O. et al. Universal fragment descriptors for predicting properties of inorganic crystals. **Nature Communications**, v. 8, n. 1, 5 jun. 2017.

JACOBSSON, T. J. et al. An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. **Nature Energy**, v. 7, n. 1, p. 107–115, 1 jan. 2022.

JAYAN, K. D.; SEBASTIAN, V. A review on computational modelling of individual device components and interfaces of perovskite solar cells using DFT. **PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON ADVANCED MATERIALS: ICAM 2019**, 2019.

JONES, Frances. A corrida pelas células solares de perovskita. **Revista Pesquisa FAPESP**, São Paulo, 6 dez. 2023. Disponível em: https://revistapesquisa.fapesp.br/a-corrida-pelas-celulas-solares-de-perovskita/. Acesso em: 15 nov. 2025.

KHOSHVAGHT, H. et al. A critical review on selecting performance evaluation metrics for supervised machine learning models in wastewater quality prediction. **Journal of environmental chemical engineering**, v. 13, n. 6, p. 119675–119675, 9 out. 2025.

KITTEL, C. **Introduction to solid state physics**. Hoboken, Nj: Wiley, 2005.

KOHN, W. Nobel Lecture: Electronic structure of matter—wave functions and density functionals. **Reviews of Modern Physics**, v. 71, n. 5, p. 1253–1266, 1 out. 1999.

LI, Y.; YANG, K. High-throughput computational design of halide perovskites and beyond for optoelectronics. **WIREs Computational Molecular Science**, v. 11, n. 3, 18 ago. 2020.

MIAH, MD. H. et al. Band gap tuning of perovskite solar cells for enhancing the efficiency and stability: issues and prospects. **RSC Advances**, v. 14, n. 23, p. 15876–15906, 16 maio 2024.

MOEINI, A. S.; TEHRANI, F. S.; NAEIMI-SADIGH, A. Machine learning-enhanced band gaps prediction for low-symmetry double and layered perovskites. **Scientific Reports**, v. 14, n. 1, 5 nov. 2024.

MURPHY, K. P. **Machine learning: a probabilistic perspective**. Cambridge (Ma): Mit Press, 2012.

MYERS, H. P. **Introductory solid state physics**. London: Taylor & Francis, Repr, 2002.

OBADA, D. O. et al. Explainable machine learning for predicting the band gaps of ABX3 perovskites. **Materials Science in Semiconductor Processing**, v. 161, p. 107427–107427, 1 jul. 2023.

PALIWAL, M.; KUMAR, U. A. The predictive accuracy of feed forward neural networks and multiple regression in the case of heteroscedastic data. **Applied Soft Computing**, v. 11, n. 4, p. 3859–3869, 3 mar. 2011.

PERDEW, J. P.; BURKE, K.; ERNZERHOF, M. Generalized Gradient Approximation Made **Simple. Phys. Rev. Lett.**, v. 77, n. 18, p. 3865–3868, 28 out. 1996.

PERDEW, J. P. Jacob's ladder of density functional approximations for the exchange-correlation energy. **AIP Conference Proceedings**, 2001.

PILANIA, G. et al. Machine learning band gaps of double perovskites. **Scientific Reports**, v. 6, n. 1, 19 jan. 2016.

RÜHLE, S. Tabulated values of the Shockley–Queisser limit for single junction solar cells. **Solar Energy**, v. 130, n. 0038-092X, p. 139–147, jun. 2016.

SAMANTARAY, N.; SINGH, A.; ANU TONK. Identifying the best ML model for predicting the Band gap in a Perovskite Solar Cell. **RSC Sustainability**, 1 jan. 2024.

SCHLEDER, G. R. et al. From DFT to machine learning: recent approaches to materials science–a review. **Journal of Physics: Materials**, v. 2, n. 3, p. 032001, 16 maio 2019.

SCHLEDER, G. R.; FAZZIO, A. Machine Learning na Física, Química, e Ciência de Materiais: Descoberta e Design de Materiais. **Rev. Bras. Ensino Fís.** v. 43, n. suppl 1, 5 mar. 2021.

SHAP. **Welcome to the SHAP Documentation — SHAP latest documentation**. 2018. Disponível em: https://shap.readthedocs.io/en/latest/. Acesso em: 15 nov. 2025.

SHALEV-SHWARTZ, SHAI; BEN-DAVID, SHAI. **Understanding machine learning : from foundations to algorithms**. Cambridge Etc: Cambridge University Press, 2014.

SLAVNEY, A. H. et al. A pencil-and-paper method for elucidating halide double perovskite band structures. **Chemical Science**, v. 10, n. 48, p. 11041–11053, 1 jan. 2019.

SONG, Z. et al. A technoeconomic analysis of perovskite solar module manufacturing with low-cost materials and techniques. **Energy & Environmental Science**, v. 10, n. 6, p. 1297–1305, 2017.

TALAPATRA, A. et al. Band gap predictions of double perovskite oxides using machine learning. **Communications materials**, v. 4, n. 1, 10 jun. 2023.

YI, Z. et al. Will organic–inorganic hybrid halide lead perovskites be eliminated from optoelectronic applications? **Nanoscale Advances**, v. 1, n. 4, p. 1276–1289, 2019.

ZHANG, L. et al. Advances in the Application of Perovskite Materials. **Nano-micro Letters**, v. 15, n. 1, 10 jul. 2023.